The effects of lesions on the generalization ability of a perceptron

1993 J. Phys. A: Math. Gen. 26 1847

(http://iopscience.iop.org/0305-4470/26/8/013)

J. Phys. A: Math. Gen. 26 (1993) 1847–1863. Printed in the UK

The effects of lesions on the generalization ability of a perceptron

D M L Barbato and J F Fontanari
Instituto de Física e Química de São Carlos, Universidade de São Paulo, Caixa Postal 369, 13560 São Carlos SP, Brazil

Received 27 September 1992

Abstract. We investigate the effects of dilution (lesions) on the memorization and generalization abilities of a single-layer perceptron whose non-zero weights are constrained to take on binary (± 1) values only. The diluted perceptron is trained to realize a Boolean linearly separable mapping generated by a fully connected perceptron. In the case where the training process is disturbed by noise and the vanishing weights are chosen so as to minimize the training error, we find that the dilution can improve the storage capacity and the generalization ability of the network. If the weights are cut randomly, however, the dilution will always degrade the network's performance. In this case we show that the main effect of dilution is to introduce an effective noise in the training examples.

1. Introduction

The study of the cognitive capabilities of patients who suffer from neurological damage has provided many clues to the understanding of the brain as, for instance, the localization of brain functions and the absence of a specific location for the memory. The design of artificial neural networks has long benefited from this neurological information, as attested by the associative memory models, where the memories are scattered throughout the network's synaptic weights (Hopfield 1982) in the hope of obtaining the robustness of the brain's memory system under destruction of neurons and synapses. On the other hand, the study of the effects of dilution (lesions) on artificial neural networks may indicate which properties of the brain are robust to details of model building. In this vein, Virasoro (1988) has shown that the random destruction of weights in a network that stores ultrametric memories produces a pattern similar to the prosopagnosia syndrome, which affects the capacity of recognizing individuals belonging to the same category. It seems then that both systems, the brain and artificial neural networks, employ the same principles to achieve categorization.

In this paper we use Gardner's (1988) statistical mechanics formalism to investigate how the elimination of a fixed fraction of synaptic weights affects the memorization and generalization performances of a single-layer perceptron whose non-zero weights are constrained to take on binary (± 1) values only. We consider two types of dilution which, following Bouten *et al* (1990a), we term annealed and quenched dilution. In the former case, the learning process determines which weights must be eliminated so as to minimize the effects of the lesion on the training error,

0305-4470/93/081847+17$07.50

while in the latter case the weights are cut randomly. Bouten *et al* (1990a, b) have studied the effects of both types of dilution on the storage capacity of networks of binary as well as real weights trained to realize random input/output mappings. The mapping we consider in this paper, however, is generated by a non-damaged reference perceptron (teacher perceptron), so it cannot be perfectly realized by the damaged network (student perceptron). Essentially, this is a version of the problem of learning unrealizable rules with perceptrons (Seung *et al* 1992, Meir and Fontanari 1992, Watkin and Rau 1992). To better appreciate the effects of dilution, we also study the problem of training the student perceptron with patterns corrupted by noise, showing that the annealed dilution can in fact improve the performance of the network in this case. Furthermore, we show that the main effect of quenched dilution is to introduce an effective noise in the training patterns.

In the case of binary-weights perceptrons, the problem of learning unrealizable rules, similarly to the random mapping problem, is not amenable to analysis within the canonical replica-symmetric formulation (Gardner and Derrida 1988, Krauth and Mezard 1989, Seung *et al* 1992), requiring a more elaborate framework, namely Parisi's replica symmetry breaking scheme (Parisi 1980, Mezard *et al* 1987). However, it was argued recently that the microcanonical replica-symmetric formulation provides a valuable approximation to study the thermodynamics of these models (Fontanari and Meir 1992), giving the exact solution for models that possess a frozen phase akin to the one present in the random-energy model (Derrida 1981) or in the simplest spin glass (Gross and Mezard 1984). Since a similar frozen phase appears in the random mapping problem (Krauth and Mezard 1989) as well as in the problem of learning realizable rules (Györgi 1990, Seung *et al* 1992), we believe that this phenomenon must also occur in the somewhat intermediate problem of learning unrealizable rules considered in this work. In this sense, we think that the results presented in this paper, obtained within the microcanonical replica-symmetric framework, are probably exact.

The remainder of this paper is organized as follows. In section 2 we describe the model and present the microcanonical version of the statistical mechanics of discrete-weights neural networks. The annealed and quenched dilutions are studied within the replica framework in sections 3 and 4, respectively. Finally, in section 5 we discuss our results and present some concluding remarks.

2. The model

The neural network we consider in this paper consists of N binary input units $S_i = \pm 1$ $(i = 1, \ldots, N)$, N synaptic weights W_i $(i = 1, \ldots, N)$, and a single output unit

$$\sigma = \operatorname{sgn}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_i S_i\right). \tag{1}$$

The task of the student perceptron is to realize the mapping between the 2^N possible input configurations $\{\xi\}$ and their respective outputs $\{t\}$ generated by the teacher perceptron,

$$t = \operatorname{sgn}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_i^0 \xi_i\right) \tag{2}$$

where $W_i^0 = \pm 1$ $(i = 1, \ldots, N)$ are the weights of the teacher perceptron. To achieve this task, the network is trained with $P = \alpha N$ input/output pairs (S^l, t^l) $(l = 1, \ldots, P)$ where t^l is the teacher's output to input ξ^l and each component S_i^l is drawn from the conditional probability distribution

$$P(S_i^l \mid \xi_i^l) = \frac{1+\gamma}{2}\delta(S_i^l - \xi_i^l) + \frac{1-\gamma}{2}\delta(S_i^l + \xi_i^l) \tag{3}$$

with ξ_i^l chosen randomly as ± 1 with equal probability. The input pattern S^l is thus a noisy version of the pure pattern ξ^l. We are modelling a situation where the communication between teacher and student during the training stage is disturbed by noise, whose strength is measured by parameter $\gamma \in [0,1]$. The case $\gamma = 0$ corresponds to the random-mapping problem (maximal noise) and $\gamma = 1$ to the problem of training with pure (noiseless) patterns.

For a fixed realization of the input/output pairs, the training process consists of a search on the space of networks for the global minima of the *training* energy, defined as

$$E(\boldsymbol{W}, \boldsymbol{S}, t) = \sum_{l=1}^{P} \Theta\left(-t^l \sigma\left(\boldsymbol{W}, \boldsymbol{S}^l\right)\right) \tag{4}$$

where $\Theta(x)$ is 1 for positive x and zero otherwise and $\sigma(\boldsymbol{W}, \boldsymbol{S}^l)$ is the student's response to noisy input $\boldsymbol{S}^l$. We note that the training energy does not give direct information about the network's performance on classifying correctly the P pure patterns associated with the noisy training examples (unless $\gamma = 1$, of course). In this paper we focus mainly on the network's performance on the whole domain of the mapping defined by (2). With this purpose we define the *generalization* function

$$E_g(\boldsymbol{W}) = 2^{-N} \sum_{\{\xi\}} \Theta\left(-t\sigma\left(\boldsymbol{W}, \xi\right)\right) \tag{5}$$

where $\sigma(\boldsymbol{W}, \xi)$ is the student's output to pure input ξ and the summation is over the 2^N possible input patterns. Thus, although the network is trained with noisy examples, its generalization ability is tested on the classification of pure patterns. The *generalization error* is obtained by averaging (5) over all possible realizations of the input/output mapping, i.e. over the 2^N possible teacher networks. We refer the reader to Györgi and Tishby (1989) and Seung *et al* (1992) for a more thorough discussion of the problem of learning from examples in neural networks.

The student network is damaged by setting a fraction $1 - Q$ (degree of dilution) of its weights to zero. The remaining NQ weights are constrained to assume the values ± 1 only. In the case of annealed dilution, the vanishing weights are chosen so as to minimize the training energy (equation (4)) implying thus that the dilution process depends on the particular realization of the input/output mapping. To study this problem, we consider the space of networks whose weights can take on the values $W_i = 0, \pm 1$ and obey the constraint

$$\frac{1}{N}\sum_{i=1}^{N} W_i^2 = Q \tag{6}$$

which ensures the correct degree of dilution. On the other hand, in the case of quenched dilution the weights are cut randomly, independently of the realization of the input/output mapping. Therefore, without loss of generality, we can set $W_i = 0$ $(i = NQ + 1, \ldots, N)$ so that (6) is satisfied automatically. In both cases, however, the generalization function (equation (5)) reduces to

$$E_g(\boldsymbol{W}) = \frac{1}{\pi}\cos^{-1}(\sqrt{Q}R) \tag{7}$$

where R is the overlap of the student network with the teacher network, i.e.

$$R = \frac{1}{NQ}\sum_{i=1}^{N} W_i^0 W_i \,. \tag{8}$$

Clearly, the best generalization performance is achieved for $R = 1$, independently of the type of dilution considered. However, we shall see that in the case of annealed dilution the student needs less training examples to approach this optimal regime than in the quenched case.

The main purpose of this paper is to study how the generalization error assigned to the global minima (ground states) of the training energy depends on the noise parameter γ, the connectivity Q, and the size of the training set α. Within the microcanonical formalism these minima can easily be characterized by computing $\mathcal{N}(E)$, the number of networks with energy $E \geqslant 0$. For a fixed realization of $\boldsymbol{S}^l$, $\boldsymbol{\xi}^l$ and $\boldsymbol{W}^0$, this quantity can be computed by defining the function $Y_{\boldsymbol{W}}$ which is 1 if $E(\boldsymbol{W}, \boldsymbol{S}, t) = E$ and zero otherwise, so that

$$\mathcal{N}(E) = \mathrm{Tr}_{\boldsymbol{W}} Y_{\boldsymbol{W}} \tag{9}$$

where $\mathrm{Tr}_{\boldsymbol{W}}$ is the summation over all allowed weight configurations (networks).

To rid our results of the dependence on the realization of the input/output mapping, we follow the standard prescription of taking averages over extensive quantities only, as they become self-averaging in the limit $N \to \infty$ (Binder and Young 1986), and define the average entropy density

$$s(E) = \frac{1}{N}\langle\langle \ln \mathcal{N}(E) \rangle\rangle \,. \tag{10}$$

Here, $\langle\langle \cdots \rangle\rangle$ stands for the averages over $\boldsymbol{S}^l$, $\boldsymbol{\xi}^l$ and $\boldsymbol{W}^0$. Since $\mathcal{N}(E)$ is a non-negative integer, $s(E)$ cannot take on finite negative values. Thus, the ground-state energy E_{gs} satisfies $s(E \geqslant E_{gs}) \geqslant 0$ and $s(E < E_{gs}) = -\infty$. The ground-state entropy density $s_{gs} = s(E = E_{gs})$ gives information about the ground-state degeneracy: in the case $s_{gs} > 0$, the number of ground states is of order $\exp(N s_{gs})$, while in the case $s_{gs} = 0$, it is of order N^x $(x > 0)$. Furthermore, E_{gs} is a good measure of the network's memorization capability, allowing us to define the network's *storage capacity* α_c as the ratio between the maximal number of input/output pairs for which $E_{gs} = 0$ and the number of input units N.

The average in (10) is evaluated within the replica framework, which consists basically of using the identity

$$\langle\langle \ln \mathcal{N}(E) \rangle\rangle = \lim_{n \to 0} \frac{1}{n} \ln \langle\langle (\mathcal{N}(E))^n \rangle\rangle \tag{11}$$

evaluating $\langle\langle (\mathcal{N}(E))^n \rangle\rangle$ for *integer* n and then analytically continuing to $n \approx 0$. Noting that the training energy (equation (4)) is a random variable distributed according to the probability distribution

$$P(E_t) = \langle\langle \delta(E_t - E(\boldsymbol{W}, \boldsymbol{S}, t)) \rangle\rangle \tag{12}$$

the calculation of the nth moment of $\mathcal{N}(E)$ becomes straightforward:

$$\langle\langle (\mathcal{N}(E))^n \rangle\rangle = \left\langle\left\langle \prod_{a=1}^{n} \mathrm{Tr}_{\boldsymbol{W}^a} Y_{\boldsymbol{W}^a} \right\rangle\right\rangle = \mathrm{Tr}_{\boldsymbol{W}^1} \ldots \mathrm{Tr}_{\boldsymbol{W}^n} P(E_t^1 = E, \ldots, E_t^n = E) \tag{13}$$

where $P(E_t^1 = E, \ldots, E_t^n = E)$ is the joint probability that networks $\boldsymbol{W}^1, \ldots, \boldsymbol{W}^n$ have energy equal to E. Therefore

$$\langle\langle (\mathcal{N}(E))^n \rangle\rangle = \left\langle\left\langle \prod_{a=1}^{n} \mathrm{Tr}_{\boldsymbol{W}^a} \delta(E - E(\boldsymbol{W}^a, \boldsymbol{S}, t)) \right\rangle\right\rangle . \tag{14}$$

Using the integral representation of the delta function and defining the *training error* $\epsilon = E/\alpha N \in [0,1]$ we find

$$\langle\langle (\mathcal{N}(\epsilon))^n \rangle\rangle = \int \prod_{a=1}^{n} \frac{\mathrm{d}\hat{\epsilon}_a}{2\pi \mathrm{i}} \exp N \left(\alpha\epsilon \sum_{a=1}^{n} \hat{\epsilon}_a + \frac{1}{N} \ln \langle\langle (Z(\hat{\epsilon}, \boldsymbol{S}, t))^n \rangle\rangle \right) \tag{15}$$

where

$$Z(\hat{\epsilon}, \boldsymbol{S}, t) = \mathrm{Tr}_{\boldsymbol{W}} \exp(-\hat{\epsilon} E(\boldsymbol{W}, \boldsymbol{S}, t)) \tag{16}$$

is the canonical partition function with $\hat{\epsilon}$ playing the role of the inverse temperature. The connection with the canonical formulation is made through the thermodynamic relationship

$$\frac{\partial(Ns)}{\partial E} = \frac{\partial s}{\partial(\alpha\epsilon)} = \frac{1}{T} . \tag{17}$$

In the next two sections we evaluate the average entropy density for the cases of annealed and quenched dilutions. As mentioned before, these problems differ only in the way the summation over the allowed weight configurations ($\mathrm{Tr}_{\boldsymbol{W}}$) is performed.

3. Annealed dilution

In this case, the weights are allowed to take on the values $0, \pm 1$ and the constraint (6) must be enforced by a Kronecker delta. Using standard techniques (Gardner 1988, Gardner and Derrida 1988) we obtain, in the thermodynamic limit,

$$s(\epsilon)/Q = \lim_{n \to 0} \mathrm{extr} \left\{ \alpha'\epsilon \sum_a \hat{\epsilon}_a - \sum_{a<b} q_{ab}\hat{q}_{ab} - \sum_a R_a \hat{R}_a \right.$$
$$\left. - \sum_a \hat{Q}_a + \frac{1}{Q} G_0(\hat{q}_{ab}, \hat{R}_a, \hat{Q}_a) + \alpha' G_1(q_{ab}, R_a, \hat{\epsilon}_a) \right\} \tag{18}$$

where

$$G_0 = \ln\left\{\prod_{a=1}^{n}\sum_{\{W^a=0,\pm 1\}}\exp\left(\sum_{a<b}\hat{q}_{ab}W^aW^b + \sum_a \hat{Q}_a(W^a)^2 + \sum_a \hat{R}_aW^aW^0\right)\right\} \tag{19}$$

and

$$G_1 = \ln\int \mathrm{D}y\int\prod_a \frac{\mathrm{d}x_a\mathrm{d}\hat{x}_a}{2\pi}\exp\left[-\sum_a \hat{\epsilon}_a\Theta(-yx_a) - \frac{1}{2}\sum_a \hat{x}_a^2(1-\gamma'^2R_a^2)\right.$$
$$\left. - \sum_{a<b}\hat{x}_a\hat{x}_b(q_{ab}-\gamma'^2R_aR_b) + \mathrm{i}\sum_a \hat{x}_a(x_a - y\gamma'R_a)\right] \tag{20}$$

with the notation

$$\mathrm{D}y = \frac{\mathrm{d}y}{\sqrt{2\pi}}\mathrm{e}^{-y^2/2}. \tag{21}$$

Here, $\alpha' \equiv \alpha/Q$ and $\gamma' \equiv \gamma\sqrt{Q}$. The extremum in (18) is taken over all order parameters $(\hat{\epsilon}_a, \hat{q}_{ab}, \hat{R}_a, \hat{Q}_a, q_{ab}, R_a)$. The physical order parameters

$$q_{ab} = \frac{1}{NQ}\sum_{i=1}^{N} W_i^aW_i^b \qquad a \neq b \tag{22}$$

and

$$R_a = \frac{1}{NQ}\sum_{i=1}^{N} W_i^aW_i^0 \tag{23}$$

measure the overlap between two different networks with training error ϵ and the overlap between a network with training error ϵ and the teacher network, respectively.

To proceed further we make the replica-symmetric ansatz, i.e. we assume that the values of the order parameters are independent of their replica indices,

$$\begin{aligned} q_{ab} &= q \quad \text{and} \quad \hat{q}_{ab} = \hat{q} \quad \forall a<b \\ R_a &= R \quad \text{and} \quad \hat{R}_a = \hat{R} \quad \forall a \\ \hat{Q}_a &= \hat{Q} \quad \text{and} \quad \hat{\epsilon}_a = \hat{\epsilon} \quad \forall a\,. \end{aligned} \tag{24}$$

Evaluation of equations (19) and (20) with this ansatz is straightforward, resulting in the following expression for the replica-symmetric average entropy density:

$$s_{\mathrm{rs}}(\epsilon)/Q = \alpha'\epsilon\hat{\epsilon} + q\hat{q}/2 - R\hat{R} - \hat{Q} + 2\alpha'\int \mathrm{D}y H(\xi_1)\ln[\mathrm{e}^{-\hat{\epsilon}} + (1-\mathrm{e}^{-\hat{\epsilon}})H(\xi_2)]$$
$$+ \frac{1}{Q}\int \mathrm{D}y \ln[1 + 2\exp(\hat{Q} - \hat{q}/2)\cosh(\hat{R} + y\sqrt{\hat{q}})] \tag{25}$$

where

$$\xi_1 = y\sqrt{\frac{\gamma'^2 R^2}{q - \gamma'^2 R^2}} \tag{26}$$

$$\xi_2 = y\sqrt{\frac{q}{1-q}} \tag{27}$$

and

$$H(x) = \int_x^\infty \mathrm{D}y\,. \tag{28}$$

As expected, the thermodynamic relationship (17) gives $\hat{\epsilon} = 1/T$. The replica-symmetric order parameters $(\hat{\epsilon}, \hat{q}, \hat{R}, \hat{Q}, q, R)$ are given by the saddle-point equations

$$Qq = \int \mathrm{D}y \left[\frac{2\exp(\hat{Q} - \hat{q}/2)\sinh(\hat{R} + y\sqrt{\hat{q}})}{1 + 2\exp(\hat{Q} - \hat{q}/2)\cosh(\hat{R} + y\sqrt{\hat{q}})}\right]^2 \tag{29}$$

$$QR = \int \mathrm{D}y \frac{2\exp(\hat{Q} - \hat{q}/2)\sinh(\hat{R} + y\sqrt{\hat{q}})}{1 + 2\exp(\hat{Q} - \hat{q}/2)\cosh(\hat{R} + y\sqrt{\hat{q}})} \tag{30}$$

$$Q = \int \mathrm{D}y \frac{2\exp(\hat{Q} - \hat{q}/2)\cosh(\hat{R} + y\sqrt{\hat{q}})}{1 + 2\exp(\hat{Q} - \hat{q}/2)\cosh(\hat{R} + y\sqrt{\hat{q}})} \tag{31}$$

$$\hat{R} = -\frac{2\alpha' q}{\sqrt{2\pi}R(q - \gamma'^2 R^2)} \int \mathrm{D}y\xi_1 \mathrm{e}^{-\xi_1^2/2} \ln[(\mathrm{e}^{\hat{\epsilon}} - 1)^{-1} + H(\xi_2)] \tag{32}$$

$$\hat{q} = \frac{R\hat{R}}{q} + \frac{2\alpha'}{\sqrt{2\pi}q(1-q)} \int \mathrm{D}y\xi_2 \frac{\mathrm{e}^{-\xi_1^2/2} H(\xi_1)}{(\mathrm{e}^{\hat{\epsilon}} - 1)^{-1} + H(\xi_2)} \tag{33}$$

$$\epsilon = 2(\mathrm{e}^{\hat{\epsilon}} - 1)^{-1} \int \mathrm{D}y H(\xi_1) \frac{1 - H(\xi_2)}{(\mathrm{e}^{\hat{\epsilon}} - 1)^{-1} + H(\xi_2)}\,. \tag{34}$$

For ϵ, α, γ and Q fixed, this system of coupled equations is solved numerically and its solution inserted into (25). As a function of the training error ϵ, the average entropy density increases with increasing ϵ, reaches its maximum value $Q \ln 2$ at $\epsilon = \frac{1}{2}$, and then decreases as ϵ increases further towards 1. Furthermore, since the training energy defined by (4) satisfies $E(-W) = 1 - E(W)$, the entropy density must be invariant with respect to reflection about the point $\epsilon = \frac{1}{2}$. We note, however, that the region $\frac{1}{2} < \epsilon \leqslant 1$ corresponds to a regime of negative temperatures and refer the reader to Landau and Lifshitz (1980) for a physical interpretation of a similar phenomenon in the context of solid state physics.

As we are interested in the ground-state properties, we must look for the lowest value of ϵ for which the entropy is positive. In contrast with the exact entropy density $s(\epsilon)$ which cannot assume finite negative values, the replica-symmetric entropy density $s_{\mathrm{rs}}(\epsilon)$ can become negative for certain values of the parameters α, Q and γ. We define then the replica-symmetric estimate of the ground-state training error as the lowest value of $\epsilon \geqslant 0$ for which $s_{\mathrm{rs}}(\epsilon)$ is positive and denote it by ϵ_{t}. Clearly, if

$\epsilon_t = 0$ then $s_{rs}(\epsilon) \geqslant 0$ for all ϵ and the replica-symmetric theory is exact, provided the replica-symmetric saddle-point is locally stable. On the other hand, in the case $\epsilon_t > 0$, the replica-symmetric theory predicts that $s_{rs}(\epsilon)$ vanishes at $\epsilon = \epsilon_t$ and is negative for $\epsilon < \epsilon_t$. This is a clear indication that the ansatz (24) does not describe correctly the structure of the order parameters for $\epsilon < \epsilon_t$. However, Krauth and Mezard (1989) have shown that the condition $s_{rs}(\epsilon) = 0$ actually determines the exact ground state of the training energy for the random-mapping problem ($Q = 1$ and $\gamma = 0$). Due to the similarity between the models, we believe that the replica-symmetric estimate ϵ_t determines the exact ground state of the diluted models too, provided again that the replica-symmetric saddle-point is locally stable. In any event, we mention that, even for models that possess a more complex ground-state structure, the microcanonical replica-symmetric theory gives estimates for the ground-state training energy which are comparable with the estimates of the canonical one-step replica symmetry breaking theory (Fontanari and Meir 1993). We note that for fixed α, γ and Q, the knowledge of ϵ_t suffices for specifying the values of all order parameters that characterize the ground state. For instance, the ground-state generalization error is given by (5) with R replaced by its saddle-point value calculated at $\epsilon = \epsilon_t$.

The condition for the local stability of replica-symmetric saddle point (de Almeida and Thouless 1978) is given by

$$\alpha\gamma_0\gamma_1 < 1 \tag{35}$$

where γ_0 and γ_1 are the transverse eigenvalues of the $\frac{1}{2}n(n+3)$-dimensional matrices of second derivatives of G_0 and G_1 with respect to $\hat{q}_{ab}$ and q_{ab}, respectively. Following the analysis of Gardner and Derrida (1988) we find

$$\gamma_0 = 4\exp(2\hat{Q} - \hat{q}) \int \mathrm{D}y \frac{[2\exp(\hat{Q} - \hat{q}/2) + \cosh(\hat{R} + y\sqrt{\hat{q}})]^2}{[1 + 2\exp(\hat{Q} - \hat{q}/2)\cosh(\hat{R} + y\sqrt{\hat{q}})]^4} \tag{36}$$

and

$$\gamma_1 = 2\int \mathrm{D}y H(\xi_1)\left(\langle x^2\rangle - \langle x\rangle^2\right)^2 \tag{37}$$

where

$$\langle x\rangle = \frac{\mathrm{i}}{\sqrt{2\pi(1-q)}} \frac{\mathrm{e}^{-\xi_2^2/2}}{(\mathrm{e}^{\hat{\epsilon}} - 1)^{-1} + H(\xi_2)} \tag{38}$$

$$\langle x^2\rangle = -\frac{1}{\sqrt{2\pi}(1-q)} \frac{\xi_2 \mathrm{e}^{-\xi_1^2/2}}{(\mathrm{e}^{\hat{\epsilon}} - 1)^{-1} + H(\xi_2)} \tag{39}$$

with ξ_1 and ξ_2 given in (26) and (27) respectively.

Henceforth we shall refer to the replica-symmetric ground-state training and generalization errors as simply the training and generalization errors, denoted by ϵ_t and ϵ_g respectively. With regard to the computation of the network's storage capacity α_c we must seek the larger value of α for which $\epsilon_t = 0$. In this case, the saddle-point equations simplify considerably, since (34) implies that $\hat{\epsilon} \to \infty$. In fact, setting

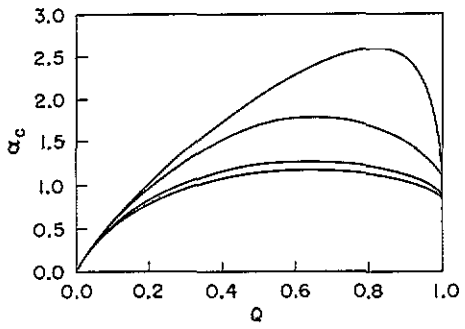

Figure 1. The storage capacity α_c for the annealed dilution as function of the connectivity Q for (from top to bottom) $\gamma = 1$, 0.9, 0.5 and 0.

$\epsilon = \gamma = 0$ we recover the results of Bouten *et al* (1990b) for the random-mapping problem.

In figure 1 we show α_c as a function of Q for several values of γ. The upper curve ($\gamma = 1$) tends to $\alpha = 1.23$ as Q approaches 1. In this limit, the training error vanishes for all α and the value $\alpha = 1.23$ signals a transition to a regime of perfect generalization (Györgi 1990, Seung *et al* 1992). This figure shows that a moderate cutting of weights actually improves the network's storage capacity, as seen by the increase of α_c as Q departs from 1. It seems then that, for small α, the enlargement of the weight space compensates for the damaging effects of dilution embodied in constraint (6). The value of Q corresponding to the maximum of α_c for a fixed γ can be calculated by extremizing s_{rs} with respect to Q. For $\gamma = 0$ this maximum is $\alpha_c = 1.17$ obtained for $Q = 0.63$ (Bouten *et al* 1990b), while for $\gamma = 1$ it is $\alpha_c = 2.58$ obtained for $Q = 0.82$. In fact, allowing Q to be an order parameter, determined by the saddle-point equation $\partial s_{rs}/\partial Q = 0$, results in a version of the problem of learning *over-realizable* rules (Meir and Fontanari 1992), since in this case the computational power of the student network exceeds that of the teacher network. The limit $\gamma = 0$ (random mapping) of this problem was studied by Gutfreund and Stein (1990) who have also obtained the maximal value of the storage capacity given above. We note that in the case of real-weights networks trained to realize a random input/output mapping the dilution process always deteriorates the network's performance (Bouten *et al* 1990a).

As α increases beyond α_c, the destructive effects of dilution on the network's memorization ability become more pronounced as depicted in figure 2, where we show the training error as function of α for $\gamma = 1$ and $Q = 0.5$, 0.8 and 0.98. In fact, it seems that the larger the lesion (the smaller the Q), the more rapidly the network's performance degrades with increasing α. Leaving aside the case $Q = 1$, we note that for each α there exists a particular value of Q for which ϵ_t is minimal. The same behaviour occurs for $\gamma < 1$, except that in this case the fully-connected network ($Q = 1$) can perform worse than the diluted ones. In the limit of large α, it can easily be shown that

$$\epsilon_t \approx \frac{1}{\pi}\cos^{-1}(\gamma\sqrt{Q}) \qquad \alpha \to \infty \tag{40}$$

so the larger Q is, the better is the performance.

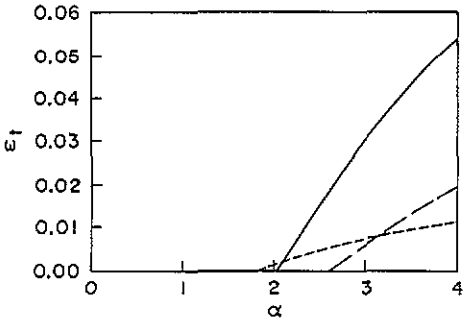

Figure 2. Training error ϵ_t for the annealed dilution as function of α for $Q = 0.5$ (full curve), 0.8 (broken curve) and 0.98 (short-broken curve). For the fully-connected ($Q = 1$) network, we find $\epsilon_t = 0$ for all α. The noise parameter is $\gamma = 1$.

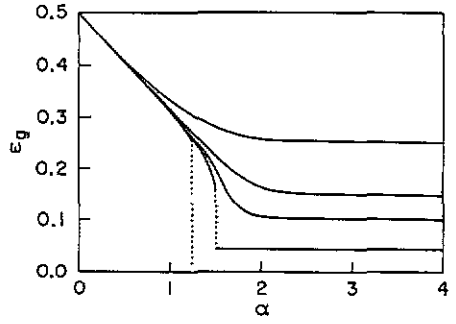

Figure 3. Generalization error ϵ_g for the annealed dilution as function of α for (from top to bottom) $Q = 0.5$, 0.8, 0.9, 0.98 and 1. The noise parameter is $\gamma = 1$.

The generalization error ϵ_g as function of α is shown in figure 3 for $\gamma = 1$ and several values of Q. As mentioned before, a discontinuous transition to a phase of perfect generalization ($\epsilon_g = 0$) takes place at $\alpha = 1.23$ for $Q = 1$. The discontinuous transition persists for Q near 1, though the jump becomes smaller as Q decreases. The discontinuity for $Q < 1$ occurs for $\alpha < \alpha_c$. As Q decreases further, reaching the value $Q \approx 0.94$, the generalization error becomes a smooth function of α, approaching its asymptotic value,

$$\epsilon_g \approx \frac{1}{\pi}\cos^{-1}(\sqrt{Q}) \qquad \alpha \to \infty \tag{41}$$

continuously with increasing α. As a consequence of our definition of the generalization function (equation (5)) the training error can become larger than the generalization error for $\gamma < 1$, as can be seen from (40) and (41). In contrast to the memorization performance, in the case of training with pure patterns ($\gamma = 1$) cutting weights always results in degradation of the generalization performance of the network, independently of the value of α. In the case of training with noisy patterns ($\gamma < 1$) however, the generalization performance for a given size of the training set can be improved by elimination of a small fraction of weights, as depicted

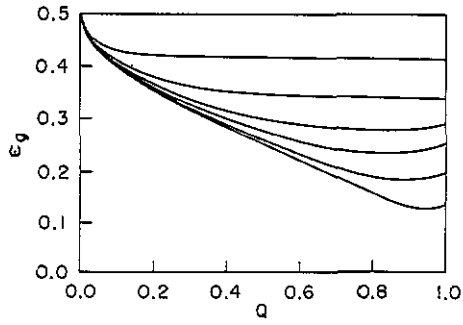

Figure 4. Generalization error ϵ_g for the annealed dilution as function of the connectivity Q for (from top to bottom) $\alpha = 0.5$, 1.0, 1.5, 2.0, 3.0 and 5.0. The noise parameter is $\gamma = 0.9$.

in figure 4, where we show ϵ_g against Q for $\gamma = 0.9$ and various α. For small α, the generalization error is practically insensitive to the cutting of weights, the best performance being achieved for $Q = 1$. As α increases, the value of Q that gives the minimal generalization error starts to decrease and then turns to increase again, tending to 1 for large α in agreement with (41). This behaviour is reminiscent of that found on the analysis of the problem of training real-weights networks with noisy examples at *non-zero* temperatures (Györgi and Tishby 1989), with T playing a role analogous to the connectivity Q.

We have verified that the replica-symmetric saddle-point is locally stable, in the sense of (35), for all $\epsilon \geqslant \epsilon_t$.

4. Quenched dilution

In this case we set $W_i = 0$ $(i = NQ + 1, \ldots, N)$ and allow the remaining NQ weights to take on the values ± 1 only so that constraint (6) is satisfied automatically, without the need of being enforced by a Kronecker delta. Similarly to the analysis of the annealed dilution we find

$$s(\epsilon)/Q = \lim_{n\to 0} \operatorname{extr}\Big\{ \alpha'\epsilon \sum_a \hat{\epsilon}_a - \sum_{a<b} q_{ab}\hat{q}_{ab} - \sum_a R_a \hat{R}_a + G_0(\hat{q}_{ab}, \hat{R}_a) + \alpha' G_1(q_{ab}, R_a, \hat{\epsilon}_a) \Big\} \tag{42}$$

where

$$G_0 = \ln\Big\{ \prod_{a=1}^{n} \sum_{\{W^a=\pm 1\}} \exp\Big(\sum_{a<b} \hat{q}_{ab} W^a W^b + \sum_a \hat{R}_a W^a W^0 \Big) \Big\} \tag{43}$$

and G_1 is given in (20). At this point we note that (42) gives the entropy density of a perceptron of $N' = QN$ input units trained with $P = \alpha' N'$ input/output pairs (S^l, t^l) where the noisy input patterns are drawn from the conditional probability distribution (3) with γ replaced by $\gamma' = \gamma\sqrt{Q}$. Thus the main effect of quenched

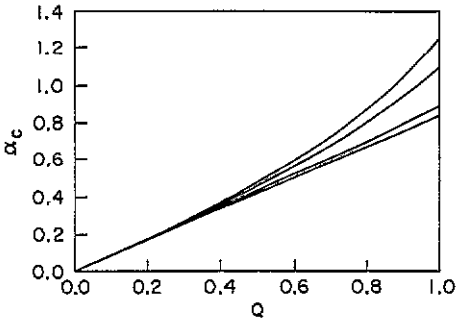

Figure 5. Same as figure 1, but for the quenched dilution.

dilution, besides the obvious rescaling of N, is to introduce an effective noise in the training process.

Using the replica-symmetric ansatz (equation (24)) we obtain

$$s_{rs}(\epsilon)/Q = \alpha'\epsilon\hat{\epsilon} - \frac{\hat{q}}{2}(1-q) - R\hat{R} + \int \mathrm{D}y \ln[2\cosh(\hat{R} + y\sqrt{\hat{q}})]$$
$$+ 2\alpha' \int \mathrm{D}y H(\xi_1) \ln[\mathrm{e}^{-\hat{\epsilon}} + (1 - \mathrm{e}^{-\hat{\epsilon}}) H(\xi_2)] \tag{44}$$

with ξ_1 and ξ_2 given in (26) and (27), respectively. The derivatives of s_{rs} with respect to R, q and $\hat{\epsilon}$ result in the saddle-point equations (32), (33) and (34), while the derivatives with respect to $\hat{q}$ and $\hat{R}$ yield

$$q = \int \mathrm{D}y \tanh^2(\hat{R} + y\sqrt{\hat{q}}) \tag{45}$$

and

$$R = \int \mathrm{D}y \tanh(\hat{R} + y\sqrt{\hat{q}}) \tag{46}$$

respectively. The condition for the local stability of the replica-symmetric saddle point is again given by (35), with γ_1 as in (37) and γ_0 given by

$$\gamma_0 = \int \mathrm{D}y [1 - \tanh^2(\hat{R} + y\sqrt{\hat{q}})]^2 . \tag{47}$$

The storage capacity as a function of Q is shown in figure 5 for various γ. The curve for $\gamma = 0$ is $\alpha_c = \alpha_c(0)Q$ where $\alpha_c(0) = 0.83$ is the storage capacity of the fully-connected network. As in the case of annealed dilution, the curve for $\gamma = 1$ tends to $\alpha = 1.23$ as $Q \to 1$. This value of α is not the network's storage capacity ($\epsilon_t = 0$ for all α in this limit) but signals a transition to a phase of perfect generalization. In contrast to the annealed dilution, the storage capacity is degraded for any degree of dilution. In figures 6 and 7 we show the training and the generalization errors, respectively, as functions of α for $\gamma = 1$ and different values of Q. It is interesting to compare these figures with their counterparts for the

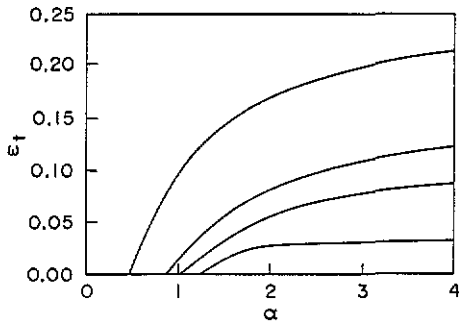

Figure 6. Training error ϵ_t for the quenched dilution as function of α for (from top to bottom) $Q = 0.5$, 0.8, 0.9 and 0.99. The noise parameter is $\gamma = 1$.

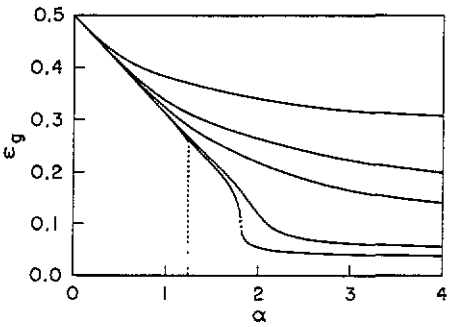

Figure 7. Generalization error ϵ_g for the quenched dilution as function of α for (from top to bottom) $Q = 0.5$, 0.8, 0.9, 0.98, 0.99 and 1. The noise parameter is $\gamma = 1$.

annealed dilution, figures 2 and 3. Although the asymptotic limits, ((41) and (40)) are the same for both types of dilution, the generalization error for the annealed dilution approaches its minimal value much faster than for the quenched dilution. The situation is reversed for the training error. The discontinuity on the generalization error for $Q < 1$ occurs for $\alpha > \alpha_c$, and disappears for values of Q larger than in the annealed case.

To better appreciate the effects of the different types of dilution on the network's memorization and generalization abilities, we present in figures 8 and 9 the training and the generalization errors, respectively, as functions of Q for $\alpha = 3.0$ and $\gamma = 1$. With these parameters we find $\epsilon_t = \epsilon_g = 0$ for the fully connected ($Q = 1$) network. The low sensitivity of the annealed training error to dilution is not surprising, since the learning process is designed to minimize this quantity. The generalization error, however, increases rapidly as Q departs from 1. As expected, the overall performance of the network is less affected in the case of annealed dilution, although the gain on the generalization error, as compared with the quenched case, is not as pronounced as the gain on the training error.

As in the annealed case, the replica-symmetric saddle point is locally stable for all $\epsilon \geqslant \epsilon_t$.

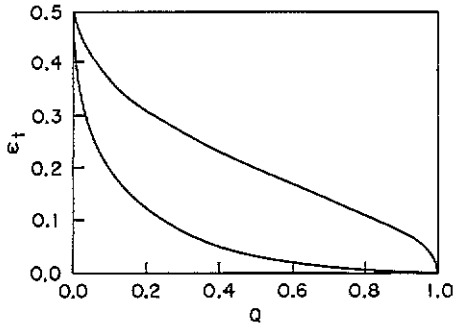

Figure 8. Training error ϵ_t as function of the connectivity Q for the annealed (lower curve) and quenched (upper curve) dilutions. The training set size is $\alpha = 3.0$ and the noise parameter is $\gamma = 1$.

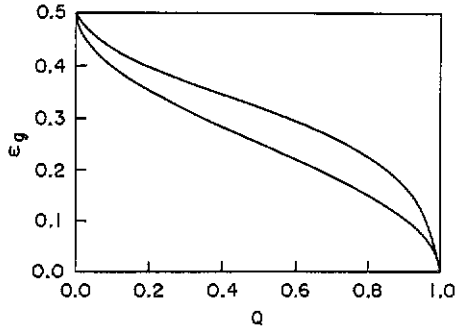

Figure 9. Same as figure 8, but for the generalization error ϵ_g.

5. Discussion

We have studied the effects of two types of dilution (lesions), annealed and quenched, on the memorization and generalization capabilities of a single-layer perceptron of binary (± 1) weights. In both cases the network is damaged *before* the learning process takes place. Since cutting weight W_i is equivalent to killing input neuron S_i, the annealed dilution can be thought of as a process where some other input neuron S_j takes over the function of neuron S_i, with the loss of its own function, so the overall performance of the network is less affected by the lesion. The quenched dilution, however, models a process where this flexibility does not exist. Actually, both processes have been observed in children who received severe injury on only one of the brain's hemispheres (Lindsay and Norman 1977). If the patient is at an early enough age, the other side of the brain can take over, compensating for the damage. It is impossible, however, to know whether or not this compensation is complete, since it would be necessary to know how the person would have developed had the brain been normal. Clearly, this is the type of process that the annealed dilution is aimed at modelling. As the organism grows older, the brain loses this flexibility, the process being then best described by the quenched dilution. One of the main advantages of studying lesions on artificial neural networks is that we have access to the performance of the non-damaged (fully connected) network, which can

then be used to single out the effects of the lesion.

We have also considered the problem of training the student perceptron with noisy input patterns S^l which fluctuate around the pure input patterns ξ^l with variance $1-\gamma^2$. This problem proved to be closely related to the dilution problem: a decreasing of the network's connectivity corresponding to an increasing of the noise's variance.

Our analytical results were all obtained within the microcanonical replica-symmetric framework, which we believe gives exact results for the diluted discrete-weights neural networks considered in this paper. Our main results are the following.

(i) The annealed dilution increases the storage capacity of the network provided the connectivity is not too low. The gain on α_c becomes more pronounced as the noise's strength decreases (i.e. $\gamma \to 1$).

(ii) In the case of training a highly connected ($Q \approx 1$) perceptron with pure examples ($\gamma = 1$), the generalization error as a function of the training set size presents a discontinuity at some value of α which is smaller than α_c in the annealed case, but greater than α_c in the quenched case. The training error, however, is continuous for all α. This behaviour resembles the discontinuous transition to a phase of perfect generalization that occurs for $Q = 1$.

(iii) In the case of training with noisy patterns ($\gamma < 1$), the annealed diluted networks can achieve the best generalization performance, depending on value of the training set size, α. In the case of training with pure patterns ($\gamma = 1$) the best generalization performance is achieved by the fully-connected ($Q = 1$) perceptron.

(iv) The effect of the quenched dilution is to rescale the number of input units $N' = QN$ and the noise parameter $\gamma' = \gamma\sqrt{Q}$. The network's performance is always degraded for this type of dilution.

(v) In the limit $\alpha \to \infty$, both types of dilution give identical results: the training error approaches its maximal value while the generalization error approaches its minimal value given in (40) and (41), respectively. The difference is that, as α increases, the annealed diluted networks reach the asymptotic limit of the generalization error much faster than the quenched diluted networks.

We should emphasize that the performance improvements obtained in the case of annealed dilution are probably an artefact of the restriction to binary weights, since for such networks the dilution can actually increase the number of possible separating hyperplanes. In fact, Bouten *et al* (1990a) have shown that, in the case of real-weights networks trained to realize random mappings, the annealed dilution decreases the storage capacity.

We have performed a similar analysis using the the canonical replica-symmetric formulation where the ground-state properties are obtained in the limit of zero temperature. In fact, the canonical saddle-point equations are identical to the microcanonical ones, except that $\hat{\epsilon}$ is replaced by the fixed parameter $\beta = 1/T$ so that (34) must be discarded. It is well known that this formulation is not appropriate to describe unrealizable (or random) rules, since it overestimates the network's storage capacity, obtained by taking the limit $q \to 1$ in the saddle-point equations. As a result, the dependence of the training error on α is totally different from that obtained within the microcanonical formulation. Surprisingly, we have found that the dependence of the generalization error on α for $\gamma = 1$ (figure 3) is indistinguishable in both formulations. For $\gamma < 1$, however, the canonical replica-symmetric theory predicts an *increase* of the generalization error for α between the microcanonical and the canonical estimates of α_c (the entropy is negative in this region), which is not seen in the microcanonical results. The reason why the two ensembles give different

results is that the temperature associated with the ground state in the microcanonical ensemble, obtained through relationship (17), is not zero for $\epsilon_t > 0$. In the region $\alpha < \alpha_c$ where $\epsilon_t = 0$, both ensembles give identical results since (34) implies $\hat{\epsilon} = 1/T \to \infty$.

An interesting issue, which we have not pursued in this paper, is the effects of cutting weights *after* the learning process has finished. This problem can be studied by looking at the distribution of the stabilities for the ground-state configuration W,

$$\Delta^l = \frac{t^l}{\sqrt{N}} \sum_{i=1}^{QN} W_i \xi_i^l \tag{48}$$

as function of Q (Kepler and Abbott 1988, Bouten *et al* 1990a). Clearly, the larger Δ^l, the more stable pattern ξ^l is against destruction of weights.

It would also be interesting to study the effects of dilution on the generalization ability of real-weights perceptrons, since these networks are somewhat more realistic than the binary-weights perceptrons considered in this paper. Moreover, we could check the theoretical predictions through numerical simulations, employing, for instance, Rosenblatt's algorithm (Rosenblatt 1962) to find the ground-state weight configurations for a given realization of the input/output mapping. Unfortunately, a numerical verification of the results presented in this paper is not possible, since the problem of finding a ground-state configuration for the binary-weights perceptron is equivalent to the integer programming problem, and therefore belongs to the *NP*-complete class (Garey and Johnson 1979). The best heuristic we know, the directed-drift algorithm (Venkatesh 1991, Fontanari and Meir 1991), becomes useless for networks of size larger than $N = 30$, while our analytical results were obtained in the thermodynamic limit, $N \to \infty$.

Acknowledgment

This work was supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

Binder K and Young A P 1986 *Rev. Mod. Phys.* **58** 801
Bouten M, Engel A, Komoda A and Serneels R 1990a *J. Phys. A: Math. Gen.* **23** 4643
—— 1990b *J. Phys. A: Math. Gen.* **1990** 2605
de Almeida J R and Thouless 1978 *J. Phys. A: Math. Gen.* **11** 983
Derrida B 1981 *Phys. Rev.* B **24** 2613
Fontanari J F and Meir R 1991 *Network* **2** 353
Fontanari J F and Meir R 1993 *J. Phys. A: Math. Gen.* **26** 1077–89
Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
Garey M R and Johnson D S 1979 *Computers and Intractability: A Guide to the Theory of NP-completeness* (San Francisco: Freeman)
Gross D J and Mezard M 1984 *Nucl. Phys.* B **240** 431
Gutfreund H and Stein Y 1990 *J. Phys. A: Math. Gen.* **23** 2613
Györgi G 1990 *Phys. Rev.* A **41** 7097
Györgi G and Tishby N 1989 *Neural Networks and Spin Glasses* ed W K Theumann and R Köberle (Singapore: World Scientific)

Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554
Kepler T B and Abbott L F 1988 *J. Physique* **49** 1657
Krauth W and Mezard M 1989 *J. Physique* **50** 3057
Landau L D and Lifshitz E M 1980 *Statistical Physics* (New York: Pergamon)
Lindsay P H and Norman D A 1977 *Human Information Processing* (New York: Academic)
Meir R and Fontanari J F 1992 *J. Phys. A: Math. Gen.* **25** 1149
Mezard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
Parisi G 1980 *J. Phys. A: Math. Gen.* **13** 1101
Rosenblatt F 1962 *Principles of Neurodynamics* (Washington DC: Spartan)
Seung S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056
Sherrington S and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792
Venkatesh S 1991 *J. Comput. Sci. Syst.* in press
Virasoro M A 1988 *Europhys. Lett.* **7** 293
Watkin T L H and Rau A 1992 *Phys. Rev.* A **45** 4102