Learning by dilution in a neural network

1997 J. Phys. A: Math. Gen. 30 7753

(http://iopscience.iop.org/0305-4470/30/22/014)

J. Phys. A: Math. Gen. **30** (1997) 7753–7764. Printed in the UK PII: S0305-4470(97)86581-X

Learning by dilution in a neural network

B López and W Kinzel

Institut für Theoretische Physik, Universität Würzburg, Am Hubland, D-97074 Würzburg, Germany

Received 4 August 1997

Abstract. A perceptron with N random weights can store of the order of N patterns by removing a fraction of the weights without changing their strengths. The critical storage capacity as a function of the concentration of the remaining bonds for random outputs and for outputs given by a teacher perceptron is calculated. A simple Hebb-like dilution algorithm is presented which, in the teacher case, reaches the optimal generalization ability.

1. Introduction

Neural networks are able to learn from examples and to find an unknown rule. Storage capacities and generalization abilities have been calculated for a variety of network architectures within the framework of statistical mechanics [1]. Special interest has been devoted to diluted networks, where only a fraction of the neurons is connected [2, 3]. The most popular example of a diluted network which appears in nature is the human brain. Every neuron is connected to roughly 10 000 others, whereas their total number is about 10^7 times larger. Theoretical studies indeed show, that the effective storage capacity per neuron in diluted systems can be substantially larger than in undiluted networks [4].

So far, dilution of synapses has been considered in addition to the usual dynamical modification of the bonds, which takes place in the learning phase [3]. Motivated by biological observations [5], which indicate that at early stages of development of the brain, synapses are removed if their strength is not appropriate, we address the question whether it is possible to store patterns in a network with randomly chosen coupling strengths only by removing a fraction of these bonds without changing their strength. This is a nontrivial task, since given a specific set of patterns it is *a priori* not clear which of the bonds have to be removed. Previous studies [8] have considered a learning algorithm which removes weights that are frustrated in at least one of the patterns. However, this simple method removes too many weights, hence the storage capacity increases with $\log N$ only. In this paper we show that it is possible to learn of the order of N patterns perfectly and to generalize by removing a fraction of the weights. We focus on the perceptron, as it is the simplest network for which a tractable calculation is feasible. The same procedure, however, should be applicable to general network classes such as multi-layer perceptrons or attractor networks.

The paper is organized as follows. In section two we introduce the model and calculate the critical storage capacity for a random input–output relation following the standard statistical mechanics approach established by Gardner [6]. Section three examines the properties of a perceptron which learns from an undiluted teacher perceptron. A simple Hebb-like dilution algorithm that reaches the optimal generalization ability is presented in section four. In the last section we close with a summary.

0305-4470/97/227753+12$19.50

2. The model for random input–output relations

The diluted perceptron classifies an input pattern $\boldsymbol{\xi}^{\mu}$ according to

$$\sigma^{\mu} = \operatorname{sgn}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} c_i J_i \xi_i^{\mu}\right). \tag{1}$$

Where J_i are the components of the weight vector drawn at random from a distribution $P(J_i)$. The c_i are binary decision variables which can take the values 0 or 1 and determine whether the ith coupling J_i is removed or kept, respectively. For a given set of input–output pairs $\{\boldsymbol{\xi}^{\mu}, s^{\mu}\}_{\mu=1}^{p}$ the classification is correct if

$$s^{\mu} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} c_i J_i \xi_i^{\mu} \geqslant 0 \qquad \forall \mu = 1, \ldots, p. \tag{2}$$

We suppose that the inputs ξ_i^{μ} are drawn at random from a distribution with zero mean and unit variance and we choose $s_{\mu} = \pm 1$ with equal probability and independently of the inputs. The concentration of the remaining bonds is defined to be $c = N^{-1} \sum_{i=1}^{N} c_i$ and thus lies between 0 and 1.

We are interested whether the maximum number of patterns $p_{\max}$, which are correctly classified, can be of the order of their input dimension N, resulting in a critical storage capacity of $\alpha_c = p_{\max}/N$, for a fixed value of the concentration c of remaining bonds in the thermodynamic limit ($N \to \infty$). Let us first consider the extreme cases. For $c = 0$ all bonds have been removed and no classification is possible, so that $\alpha_c(c = 0) = 0$. For $c = 1$ all bonds are present and the classification is at random, so that $\alpha_c(c = 1) \to 0$ as $N \to \infty$. For intermediate values of c we shall calculate $\alpha_c(c)$ using Gardner's phase space approach [6, 7].

From a technical point of view the problem is related to the Ising perceptron [7, 16] and other discrete models [9], as well as to the knapsack problem [10, 11], where binary dynamical variables also appear. Note that in contrast to the common approach where the couplings J_i are the dynamical variables, here they represent, in addition to the patterns, a quenched disorder which has to be averaged out.

For a fixed concentration c, the number of allowed configurations according to (2) is given by

$$\mathcal{N}(c) = \sum_{\{c_i\}} \prod_{\mu=1}^{p} \Theta\left(s^{\mu} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} c_i J_i \xi_i^{\mu} - \kappa\right) \delta_{\mathrm{Kr}}\left(\sum_{i=1}^{N} c_i - cN\right). \tag{3}$$

We introduce, as usual, the stability parameter κ which should be positive. The corresponding entropy per bond of the microcanonical ensemble follows from

$$S(c) = \frac{1}{N} \langle \langle \ln \mathcal{N}(c) \rangle \rangle = \lim_{n \to 0} \frac{1}{nN} \ln \langle \langle \mathcal{N}^n(c) \rangle \rangle \tag{4}$$

where the last equality results from the replica trick. The quenched averages $\langle \langle \cdots \rangle \rangle$ have to be performed over the distributions of the patterns, outputs and couplings. Following the steps of the calculation by Gardner and Derrida [7] one can rewrite the replicated number of configurations using the integral representation of the theta function in (3). The averages over the pattern and output distributions lead to the exponential factor

$$\left\langle \prod_i \exp\left(-\frac{J_i^2}{2N} \sum_{\mu} \left(\sum_{\alpha} c_i^{\alpha} x_{\mu}^{\alpha}\right)^2\right)\right\rangle_{\{J\}}. \tag{5}$$

Here x_μ^α are the conjugate variables to the local fields λ_μ^α and α denotes the replica index running from 1 to n. The average over the couplings has still to be done. A straightforward evaluation, however, leads to an expression which cannot be rewritten in terms of an exponential, as it is convenient, for the argument of the exponent in (5) is not necessarily infinitesimal due to the sum over all patterns. Instead, we introduce at this point the order parameters

$$q^{\alpha\beta} = \frac{1}{N}\sum_{i=1}^{N} J_i^2 c_i^\alpha c_i^\beta \qquad (\alpha, \beta = 1, \ldots, n; \alpha < \beta) \tag{6}$$

$$Q^\alpha = q^{\alpha\alpha} = \frac{1}{N}\sum_{i=1}^{N} J_i^2 c_i^\alpha \qquad (\alpha = 1, \ldots, n) \tag{7}$$

and leave the average over the distribution of the couplings for the integral representations of the delta functions fixing $q^{\alpha\beta}$ and Q^α. In addition we have a third order parameter E^α which fixes the concentration c. We seek a replica symmetric solution, i.e. $q^{\alpha\beta} = q$, $Q^\alpha = Q$ and $E^\alpha = E$. Carrying out the sum over the c_i^α and using the saddle-point method we obtain for the entropy in the thermodynamic limit:

$$S(c) = \alpha \int \mathrm{D}t \ln \mathrm{H}\left(\frac{\kappa + \sqrt{q}t}{\sqrt{Q-q}}\right) + \frac{1}{2}Fq - \frac{1}{2}fQ + \frac{1}{2}cE$$
$$+ \int \mathrm{d}J\, P(J) \int \mathrm{D}t \ln(1 + \exp(\sqrt{F}|J|t + \tfrac{1}{2}(fJ^2 - FJ^2 - E))) \tag{8}$$

where F and f are the conjugate order parameters to q and Q respectively and we have used the notations:

$$\mathrm{D}t \equiv \mathrm{d}t \frac{\exp(-\frac{1}{2}t^2)}{\sqrt{2\pi}} \qquad \mathrm{H}(x) = \int_x^\infty \mathrm{D}t. \tag{9}$$

For the distribution of the couplings $P(J_i)$ we will focus on two cases: $|J_i| = 1$ (note, that the entropy (8) depends only on the absolute value of J_i) and J_i drawn from a normal distribution i.e. $\mathrm{d}J\,P(J) = \mathrm{D}J$. In the first case it follows that $Q = c$ from (7) and the definition of c. We solve the saddle-point equations $\partial S/\partial q = \partial S/\partial F = \partial S/\partial E = 0$. The critical storage capacity as determined by Gardner and Derrida [7] would be reached as q approaches c. This, however, leads to a negative entropy of the system, as frequently observed in discrete problems [13]. We therefore identify the critical storage capacity as the value of α at which the entropy (8) vanishes, as it has become the standard way by now [9, 11, 14]. The resulting curve for the critical storage capacity α_c is shown in figure 1 as a function of the concentration c. As expected α_c vanishes at both extremes of c. We observe a maximum of $\alpha_\mathrm{c} \approx 0.59$ at $c \approx 0.32$. This means, that about $\frac{2}{3}$ of the bonds have to be removed in order to reach the maximal value of α_c. It is somewhat surprising, since as a function of c, the maximal number of configurations $\{c_i\}$ lies at $c = 0.5$ and is exponentially larger in N than for any other value of the concentration. In section 4 we will come back to this point.

It is worth noting, that the case $|J_i| = 1$ can be mapped onto the Ising perceptron with couplings 0 or 1, as in (2) one can define the new patterns $\chi_i^\mu = J_i \xi_i^\mu$ and view the c_i as the couplings. The distribution of the χ_i^μ also has zero mean and unit variance. The critical storage capacity for the $(0, 1)$ Ising perceptron has been calculated by Gutfreund and Stein [9] with the zero-entropy (ZE) ansatz and was found to be 0.59 in agreement with the value found here at the maximum.

We performed an analysis of the local stability of the replica symmetric (RS) solution according to de Almeida and Thouless (AT) [15] and obtained the curve denoted as AT line

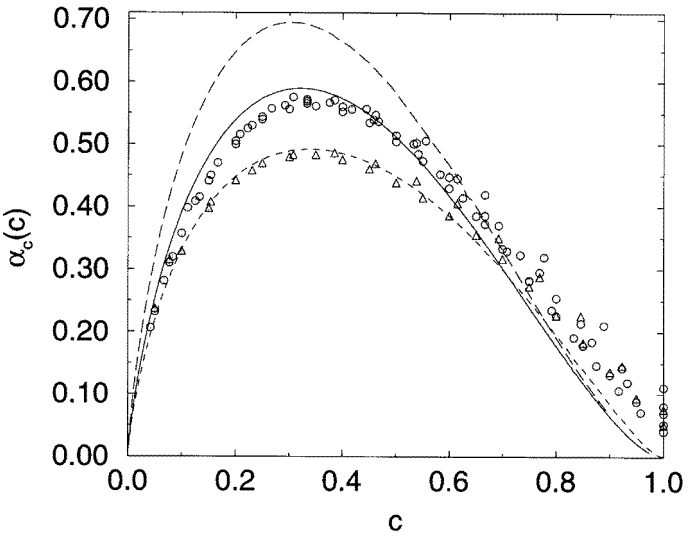

Figure 1. The critical storage capacity α_c as a function of the concentration c for $\kappa = 0$ and random outputs. The full line is the zero entropy RS–solution for $|J_i| = 1$ and the long broken curve is the corresponding AT–line beyond which it would become locally unstable. The broken curve is the RS–solution for J_i drawn from a normal distribution. Results from complete enumerations of systems with sizes $9 \leqslant N \leqslant 24$ are shown as circles and triangles, for both choices of $P(J_i)$ respectively. Standard errors are of symbol size.

in figure 1. For values of α that lie above this curve the RS solution is locally unstable. Our result for $\alpha_c(c)$ lies below the AT line for all c and is therefore locally stable. Nonetheless, global stability is not assured. For this reason we also performed complete enumerations of all possible dilution vectors for finite systems in the range $9 \leqslant N \leqslant 24$. For $c = 1$ finite-size effects lead to $\alpha_c \sim N^{-1}$, since the probability of classifying one pattern correctly by chance is $\frac{1}{2}$. In contrast, for values of c around the maximum, the numerical results seem to underestimate the theoretical values. A finite-size scaling analysis for $c = \frac{1}{3}$ gives the extrapolated value of $\alpha_c(\frac{1}{3}) = 0.586 \pm 0.004$ for $N \to \infty$ in agreement with the RS-solution (0.58935) at the same concentration. The general shape of the curve is well confirmed by the numerical results.

In the case where the J_i are drawn at random from a normal distribution the picture changes quantitatively. Now, for a fixed i, $\langle J_i \xi_i^\mu \rangle = 0$ as before, but $\langle (J_i \xi_i^\mu)^2 \rangle = J_i^2 \langle \xi_i^{\mu 2} \rangle = J_i^2$ which is in general different from unity as in the previous case. As a consequence Q is different from c and we have to solve the additional saddle-point equations $\partial S/\partial Q = \partial S/\partial f = 0$. The ZE condition yields for $\kappa = 0$ the critical storage capacity, also depicted in figure 1. Similarly, the curve has a maximum at $c \simeq 0.34$, the critical capacity however is lowered over a wide range of the concentration with respect to the binary case. Our interpretation of this effect is that in the Gaussian case the dilution variables are mainly used to remove the large couplings ($|J_i| > 1$) and only a few of them remain for learning. Therefore the storage capacity α_c is lower than for the binary weights. The order parameter Q measures the effective size of the remaining components J_i. For all c the RS solution gives $Q < c$, supporting the above argument. In addition we measured the probability distribution of the size of remaining couplings in complete enumerations and found that large couplings are likely to be removed. The values of α_c for finite N are displayed in figure 1 as well.

3. Learning with a teacher

A teacher perceptron $\boldsymbol{B}$ classifies a pattern $\boldsymbol{\xi}^\mu$ according to

$$s^\mu = \operatorname{sgn}\left(\frac{1}{\sqrt{N}}\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu\right). \tag{10}$$

We choose a teacher vector which is not diluted and has the normalization $\boldsymbol{B}^2 = N$. A transition to perfect generalization through dilution cannot be expected as in other discrete systems where the structure of teacher and student coincides [23]. The student perceptron with components $c_i J_i$ can only remove part of its weights in order to learn perfectly a set of examples given by (10), resulting in a finite storage capacity.

A straightforward evaluation of the entropy under the RS assumption yields

$$\begin{aligned} S(c) = 2\alpha \int \mathrm{D}t\,\mathrm{H}\left(\frac{R\sqrt{Q}}{\sqrt{q-QR^2}}t\right)\ln \mathrm{H}\left(\frac{\kappa+\sqrt{q}t}{\sqrt{Q-q}}\right) \\ +\frac{1}{2}Fq-\frac{1}{2}fQ+\frac{1}{2}cE+\frac{1}{2}GR\sqrt{Q}+\int \mathrm{d}B\,P_{\mathrm{T}}(B)\int \mathrm{d}J\,P(J) \\ \times\int \mathrm{D}t\ln\left(1+\exp\left(\sqrt{F}|J|t+\frac{1}{2}(fJ^2-FJ^2-E-GBJ)\right)\right) \end{aligned} \tag{11}$$

where the overlap $R = \sum_i c_i J_i B_i/(\sqrt{Q}N)$ between diluted student and teacher, and its conjugate G has been introduced. $P_{\mathrm{T}}(B)$ is the distribution of the teacher components.

We focus again on the two cases where $|J_i| = |B_i| = 1$ or both chosen independently at random from a normal distribution. The corresponding critical storage capacity α_{c} determined with the ZE condition is shown in figure 2 as a function of the concentration of remaining bonds. Once more, we observe a maximum of α_{c} at $c \approx \frac{1}{3}$. The critical storage capacity is higher than for random outputs indicating that the problem, although unlearnable, is easier with examples from a teacher.

An important point to note is that the generalization ability defined as the probability to classify a new unseen pattern correctly (as the teacher) is poor compared with the value which could be achieved by *intelligent* dilution. For $c = 0.3$, $R \simeq 0.32$ in the binary case $|J_i| = |B_i| = 1$, whereas an overlap of $R = \sqrt{0.3} \simeq 0.55$ could be possible according to the following argument. The product $B_i J_i$ should be $+1$ for as many sites as possible in order to maximize R. Since B_i and J_i are drawn at random they will coincide in $N/2$ of the cases for $N \to \infty$. For $c \leqslant 0.5$ we choose all those $c_i = 1$ for which $B_i J_i = +1$ up to a total number of cN, so that $R_{\max} = cN/(\sqrt{c}N) = \sqrt{c}$. If c is larger than 0.5 then we also have to add up $(cN - N/2)$ times a value of -1 and $R_{\max} = (N/2 - cN + N/2)/(\sqrt{c}N) = (1-c)/\sqrt{c}$. Perfect storage without errors lowers the overlap R, an effect known as *overfitting*, which can be overcome by allowing a finite training error (see section 4).

Up to now, we have assumed, that student $\boldsymbol{J}$ and teacher $\boldsymbol{B}$ are uncorrelated before the dilution. In biological systems however, we would rather expect to find synaptical structures, which are already *prepared* for a specific task before the learning process starts. In our model we can mimic it by allowing an initial positive overlap R_0, and therefore certain similarity, between teacher and student. Let us choose

$$P(J) = \frac{1+R_0}{2}\delta(J-B) + \frac{1-R_0}{2}\delta(J+B) \tag{12}$$

with $0 \leqslant R_0 \leqslant 1$. In $(1+R_0)N/2$ of the cases the components of teacher and student will coincide and in $(1-R_0)N/2$ they will be opposed. Since we are not allowed to change the

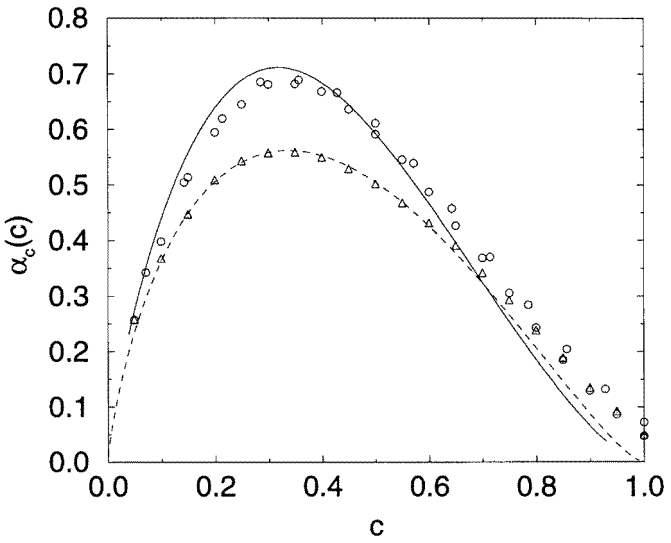

Figure 2. The critical storage capacity α_c as a function of the concentration c for $\kappa = 0$ and outputs from an undiluted teacher perceptron. The upper curve is for $|J_i| = |B_i| = 1$ and the lower for J_i and B_i both Gaussian. The circles and triangles are the corresponding numerical results from complete enumerations of system with sizes up to $N = 20$, standard errors are of symbol sizes.

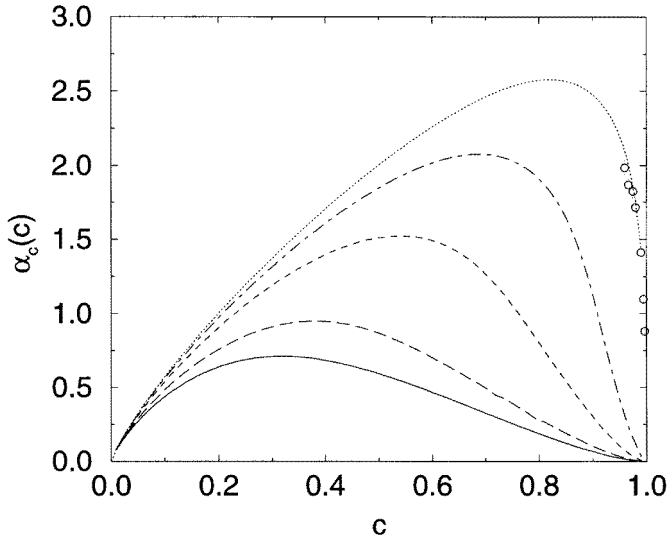

Figure 3. The critical storage capacity α_c as a function of the concentration c for $\kappa = 0$ and outputs from a teacher with $|B_i| = 1$ for different values of the initial overlap R_0 between $\boldsymbol{J}$ and $\boldsymbol{B}$. $R_0 = 1.0, 0.9, 0.7, 0.3, 0.0$ (from top to bottom). The circles are numerical results from complete enumerations of systems with sizes $25 \leqslant N \leqslant 400$ and c close to 1.

values of J_i, but at best to remove the bond, we will not reach perfect generalization even for large R_0. The same would hold even if we chose a diluted teacher.

We have calculated the critical storage capacity for $|B_i| = 1$ as a function of R_0 and c and find the results plotted in figure 3. For $R_0 = 0$ we recover the uncorrelated case, whereas with increasing R_0 the storage capacity is enhanced for all c. At the same time the maximum of the curve moves towards higher values of the concentration, as less bonds have to be removed in order to mimic the teacher. For $R_0 = 1$ teacher and student are identical before dilution. Although this situation is of less relevance from a biological point of view, it offers an interesting physical solution. If a fraction of the bonds is now removed, we obtain a finite storage capacity. For increasing concentration c, we would expect the capacity to increase at the same time. Above $c \simeq 0.82$, however, we find that it decreases and finally tends to zero for $c \to 1$. At $c = 1$ we have $\alpha_c = \infty$ per definition, hence $c = 1$ is a singular point. This surprising behaviour may be understood if we look at the annealed approximation [22] for the entropy

$$S_{\text{ann}}(c) = \frac{1}{N} \ln \langle\langle \mathcal{N}(c) \rangle\rangle. \tag{13}$$

For $R_0 = 1$ we have $R = \sqrt{c}$, thus the probability that a pattern is classified correctly is $(1 - \pi^{-1} \arccos \sqrt{c})$. Since all possible dilution vectors have for fixed c the same overlap R with the teacher, the averaged number $\langle\langle \mathcal{N}(c) \rangle\rangle$ of allowed configurations simply factorizes into the the total number $\binom{N}{cN}$ and the probability that p patterns are classified correctly:

$$\langle\langle \mathcal{N}(c) \rangle\rangle = \binom{N}{cN} \left(1 - \frac{1}{\pi} \arccos \sqrt{c}\right)^p. \tag{14}$$

Using the ZE condition we obtain in the thermodynamic limit for the critical storage capacity in the annealed approximation

$$\alpha_{\text{ann}}(c) = \frac{c \ln c + (1 - c) \ln(1 - c)}{\ln\left(1 - \frac{1}{\pi} \arccos \sqrt{c}\right)} \tag{15}$$

which for $c \to 1$ results in

$$\alpha_{\text{ann}}(c \to 1) \to \lim_{c \to 1} -4\pi \sqrt{1 - c} \ln \sqrt{1 - c} \to 0. \tag{16}$$

Since $\alpha_{\text{ann}}(c)$ is an upper bound for $\alpha_c(c)$, the critical storage capacity has to decrease to zero as well when c tends to 1. From (14) we see, that although the probability of classifying one pattern correctly tends to one for $c \to 1$, at the same time the total number of dilution vectors decreases rapidly, such that the averaged number of allowed configurations is no longer exponentially large in N. Perfect storage of patterns is different from optimal generalization, which in this case becomes better the closer c is to 1. In figure 3 we also included results from complete enumerations of systems with $25 \leqslant N \leqslant 400$ for $R_0 = 1$ and c close to 1. They confirm that α_c decreases in this region.

4. A simple Hebb-like dilution algorithm

As we have seen in the previous sections, Gardner's method is very powerful when asking if there exists, on average, a set of c_i such that all perceptron conditions (2) are satisfied, but it does not provide us with the corresponding dilution vector for a specific set of patterns $\boldsymbol{\xi}^\mu$, outputs s^μ and couplings J_i. The development of a learning, or in our case dilution algorithm, is an independent task, which in the case of binary variables $c_i = 0, 1$ becomes extremely difficult and compares with the binary perceptron problem or the knapsack problem. In the worst case, the number of computational steps towards the optimal solution scales exponentially with the size N of the system. The most successful approaches try to

find the global minimum of a properly defined energy function, which penalizes the violation of constraints, by using sequential descent [17] or simulated annealing [18] strategies. Although an algorithm based on mean-field annealing has proven to be very effective in finding solutions to the knapsack problem [10], none of the known techniques yields the critical values for the storage capacity predicted by Gardner calculations. Typically, in large systems, the solutions still violate a finite fraction of the imposed constraints.

In view of these general difficulties, we cannot expect to find the optimal dilution vector $\boldsymbol{c}$, which allows us to store perfectly $\alpha_c N$ patterns for large N, within a reasonable time. Rather we present here a simple dilution algorithm, which gives us an insight into the basic properties of the solutions and has optimal generalization ability for $\alpha \to \infty$.

Our aim is to fulfil all constraints (2) by removing a fraction $(1-c)N$ of the bonds J_i. If we think of the terms $J_i \xi_i^\mu s^\mu$ as matrix elements $a_{i\mu}$ of an $N \times p$ matrix, then we want all p vertical sums $\sum_{i=1}^N a_{i\mu} (\mu = 1, \ldots, p)$ to be positive. For this purpose, we are allowed to remove $(1-c)N$ rows of the matrix. The idea is to remove those, which contain many negative elements $a_{i\mu}$, since these contribute in many vertical sums negatively. Let us take away all rows i with horizontal sums $\sum_{\mu=1}^p a_{i\mu}$ smaller than a threshold h, so that

$$c_i = \theta\left(\frac{1}{\sqrt{N}} \sum_{\mu=1}^{p} J_i \xi_i^\mu s^\mu - h \right). \tag{17}$$

The larger h, the more c_i will be zero, leading to a lower concentration c. From a different perspective one can view (17) as comparing the Hebb couplings $H_i = \sum_{\mu=1}^p \xi_i^\mu s^\mu / \sqrt{N}$ of the problem [19] with J_i, the ones imposed at random. If their product $H_i J_i$ is larger than the threshold h, then J_i is accepted as the coupling strength. As h becomes more and more positive, this is only the case if H_i and J_i agree in their sign.

Let us now give a simple derivation of the critical storage capacity for random input–output pairs and $|J_i| = 1$, which results from (17) by allowing a certain percentage of errors. For simplicity, suppose that $\xi_i^\mu = \pm 1$ with equal probability. Then, $a_{i\mu} = \pm 1$ with equal probability and the horizontal and vertical sums have, in the limit $N \to \infty$, a Gaussian distribution with zero mean and variance p and N, respectively. According to (17) we remove all horizontal sums which are smaller than h. The resulting concentration is $c = \int_{h/\sqrt{\alpha}}^{\infty} \mathrm{D}z = \mathrm{H}(h/\sqrt{\alpha})$ and the new mean of the horizontal sums is $\langle \mathrm{HS} \rangle = \sqrt{p}\exp(-\frac{1}{2}h^2/\alpha)/\sqrt{2\pi c^2}$. The new vertical sums still have a Gaussian distribution, but now with mean $\langle \mathrm{VS} \rangle = c\langle \mathrm{HS} \rangle/\alpha$ and variance cN for $N \to \infty$. The fraction of errors is thus equal to the integral over the Gaussian tail below zero:

$$\text{learning error} \equiv \epsilon_\mathrm{L} = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{cN}} \exp\left(-\frac{1}{2} \frac{(z - \langle \mathrm{VS} \rangle)^2}{cN} \right) = \mathrm{H}\left(\frac{\langle \mathrm{VS} \rangle}{\sqrt{cN}} \right). \tag{18}$$

For a fixed error $\epsilon_\mathrm{L} = \mathrm{H}(A)$ and fixed concentration c, we obtain for the storage capacity

$$\alpha(c, A) = \frac{\exp(-(\mathrm{H}^{-1}(c))^2)}{2\pi c A^2} \tag{19}$$

with $\mathrm{H}^{-1}(x)$ the inverse function of $\mathrm{H}(x)$. Figure 4 shows the resulting α for $A = 1$ ($\epsilon_\mathrm{L} \simeq 15.9\%$) as a function of c. The maximum of α is reached at $c \approx 0.27$, which is not too far from 0.32, the concentration at which the maximal α_c was obtained according to the Gardner calculation with the ZE condition (see figure 1). Also, the shape of the curve is similar, for different values of ϵ_L (or A), $\alpha(c, A)$ is simply rescaled. At first sight one would expect that as h increases, the mean of the vertical sums $\langle \mathrm{VS} \rangle$ increases, leading to a lower learning error. This however is prevented by two effects. First, as $\langle \mathrm{HS} \rangle$ increases, $\langle \mathrm{VS} \rangle$ does not necessarily increase, since $\langle \mathrm{VS} \rangle \sim c\langle \mathrm{HS} \rangle$ and c is lowered dramatically with

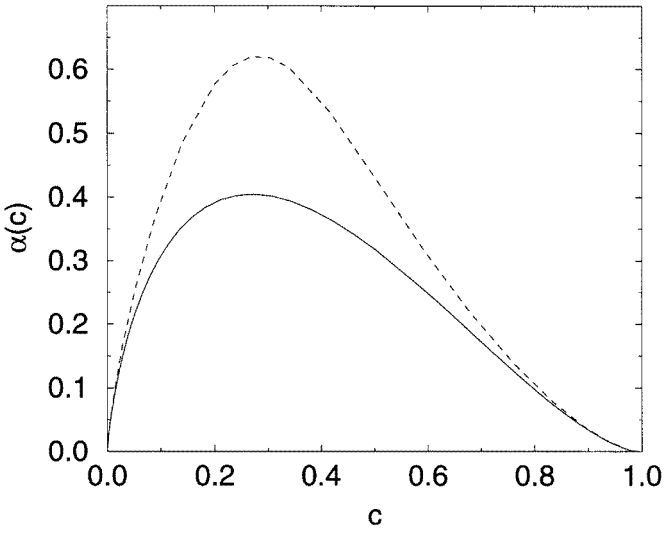

Figure 4. The storage capacity α as a function of the concentration c for $|J_i| = 1$, $\kappa = 0$ and $A = 1(\epsilon_L \simeq 15.9\%)$. The lower curve is for randomly chosen outputs and the upper curve for outputs from a teacher with $|B_i| = 1$.

increasing h. As a result, the maximal storage capacity would be at $c = 0.5$ for fixed ϵ_L. The second effect is that the width of the distribution of vertical sums is proportional to $\sqrt{c}$, lowering the learning error for smaller values of c. If c is too small, however, the gain is compensated by the exponential factor in $\langle HS \rangle$, which tends to zero. As a consequence, the maximum storage capacity for fixed ϵ_L lies at a value of c somewhat smaller than 0.5.

In figure 5 the learning error $\epsilon_L(c, \alpha)$ is plotted as a function of α for $h = 0 (\Leftrightarrow c = 0.5)$. For small α, ϵ_L is small, as is typical for the Hebb couplings and tends to 0.5 for $\alpha \to \infty$.

A more interesting quantity is the generalization error ϵ_G, defined as the probability of correctly classifying a new pattern $\boldsymbol{\xi}^0$, which does not belong to the training set. For a random input–output relation $\epsilon_G = 0.5$, since the classification s^0 of $\boldsymbol{\xi}^0$ is at random. In the presence of a teacher $\boldsymbol{B}$, however, we can expect to reach a lower generalization error. A straightforward evaluation (see e.g. [20]) yields for $|J_i| = 1$

$$\epsilon_L(c, \alpha) = 2 \int_0^\infty \mathrm{D}s \mathrm{H}\left(\frac{sR\sqrt{c} + q}{\sqrt{c}\sqrt{1 - R^2}} \right) \tag{20}$$

$$\epsilon_G(c, \alpha) = \frac{1}{\pi} \arccos R \tag{21}$$

with

$$c = \tfrac{1}{2} \langle \mathrm{H}(h_-) + \mathrm{H}(h_+) \rangle_{B_i} \tag{22}$$

$$R = \frac{1}{2\sqrt{c}} \langle B_i (\mathrm{H}(h_-) - \mathrm{H}(h_+)) \rangle_{B_i} \tag{23}$$

$$q = \frac{1}{2\sqrt{2\pi\alpha}} \left\langle \exp\left(-\frac{1}{2} h_-^2 \right) + \exp\left(-\frac{1}{2} h_+^2 \right) \right\rangle_{B_i} \tag{24}$$

where

$$h_+ = \frac{h}{\sqrt{\alpha}} + B_i \sqrt{\frac{2\alpha}{\pi}} \qquad h_- = \frac{h}{\sqrt{\alpha}} - B_i \sqrt{\frac{2\alpha}{\pi}}. \tag{25}$$

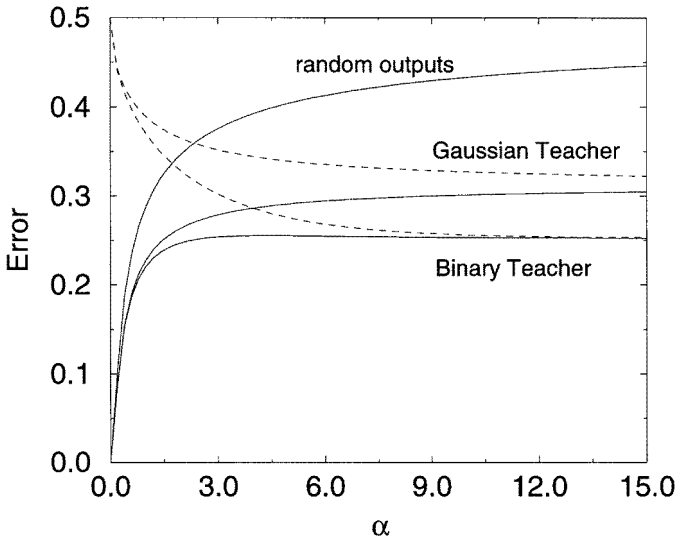

Figure 5. Learning error ϵ_L (full lines) and generalization error ϵ_G (broken lines) as a function of α for $h = 0 (\Leftrightarrow c = 0.5)$ and $|J_i| = 1$ for random outputs, a binary teacher and a Gaussian teacher.

The average is to be performed over the distribution $P_T(B_i)$ of the teacher components B_i. As before, c is the concentration of the remaining bonds and R the normalized overlap between diluted student and teacher. The parameter q (not to be confused with q defined by (6)) does not seem to have a direct physical meaning, but in a certain way it does take into account the randomness which is still inherent for R less than 1. In the extreme case where we ignore the teacher by setting all $B_i = 0$, we obtain $R = 0$, $q = \exp(-\frac{1}{2}h^2/\alpha)/\sqrt{2\pi\alpha}$ and recover for ϵ_L the expression for random outputs (18). The above result for ϵ_L and ϵ_G is similar to the one obtained with the clipped-Hebb algorithm for the perceptron [21]. This comes as no surprise, as our prescription (17) is also in some sense a way of *clipping* the bonds.

For all α and c we find $\epsilon_L < \epsilon_G$ and for $\alpha \to \infty$ and c fixed, $\epsilon_L \to \epsilon_G$, as it should. The most important feature is, however, that in this same limit the optimal generalization error is reached. For $\alpha \to \infty$ and c fixed by choosing h appropriately, we obtain from (23) for $|B_i| = 1$:

$$R = \begin{cases} \sqrt{c} & \text{for } c \leqslant \frac{1}{2} \\ \dfrac{1-c}{\sqrt{c}} & \text{for } c > \frac{1}{2} \end{cases} \tag{26}$$

which is exactly the maximal overlap that can be achieved by removing $(1-c)N$ bonds (see section 3). For a teacher with Gaussian B_i we find in the same limit:

$$R = \frac{1}{\sqrt{2\pi c}} \exp\left(-\frac{1}{2}(\mathrm{H}^{-1}(c))^2\right) \tag{27}$$

which is also the optimal overlap for this case, as can be shown easily.

5. Conclusion

We have shown, that it is possible to store information in a neural network solely by the dilution of synapses. Using Gardner's phase-space approach the critical storage capacity of a perceptron with random coupling vector was calculated as a function of the concentration of remaining bonds for random input–output relations. We found a maximum $\alpha_c \simeq 0.6$ of the capacity at $c \simeq \frac{1}{3}$, i.e. after $\frac{2}{3}$ of the bonds have been removed. Similar results are obtained if the desired outputs are generated by an undiluted teacher perceptron, whose coupling vector is uncorrelated to the initial student vector. In this case perfect learning is possible up to a critical capacity $\alpha_c(c)$, only. If the initial network has some prior knowledge, i.e. if there is a nonzero overlap between the teacher and the initial student vector we find that the maximum of the capacity moves towards higher concentrations c.

The problem of finding the subset of couplings which have to be removed is extremely difficult and is comparable with problems which belong to the NP-complete class. Nevertheless, properties of a Hebb-like learning algorithm, which allows for a finite fraction of errors in the training set, were calculated. The algorithm for $\alpha \to \infty$ reaches the maximal overlap between the diluted random student and the undiluted teacher and thus the lowest possible generalization error. As the Hebb-rule, it is a local algorithm that accumulates information about the training set and decides at the end of this batch process which of the couplings are removed. It would be more desirable to find a prescription that removes disturbing couplings *on-line*. In contrast to common on-line learning algorithms [12], where infinitesimal changes of the coupling vector are performed in every time step, here we would remove single couplings. Whether this procedure can give satisfactory results, similar to those obtained in the batch process, also for unlearnable rules, remains open and should be studied in the future.

Acknowledgments

We would like to thank M Opper and A Mietzner for useful discussions. This work has been supported by the Deutsche Forschungsgemeinschaft.

References

[1] Watkin T L H, Rau A and Biehl M 1993 *Rev. Mod. Phys.* **65** 499
Opper M and Kinzel W 1995 *Models of Neural Networks III* ed E Domany *et al* (Berlin: Springer) p 151
[2] Sompolinsky H 1987 *Heidelberg Colloquium on Glassy Dynamics* ed J L van Hemmen and I Morgenstern (Berlin: Springer) p 485
Canning A and Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 3275
Bollé D and van Mourik J 1994 *J. Phys. A: Math. Gen.* **27** 1151
Kree R and Zippelius A 1991 *Models of Neural Networks* ed E Domany *et al* (Berlin: Springer) p 193
[3] Bouten M, Engel A, Komoda A and Serneels R 1990 *J. Phys. A: Math. Gen.* **23** 4643
[4] Kuhlmann P, Garcés R and Eissfeller H 1992 *J. Phys. A: Math. Gen.* **25** L593
[5] Scheich H *et al* 1991 *Memory, Organization and Locus of Change* (Oxford: Oxford University Press)
[6] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[7] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[8] Kinzel W 1985 *Z. Phys.* B **60** 205
van Hemmen J L and van Enter A C D 1986 *Phys. Rev.* A **34** 2509
Jonker H J J and Coolen A C C 1993 *J. Phys. A: Math. Gen.* **26** 563
[9] Gutfreund H and Stein Y 1990 *J. Phys. A: Math. Gen.* **23** 2613
[10] Ohlsson M, Peterson C and Söderberg B 1993 *Neural Comput.* **5** 808
[11] Korutcheva E, Opper M and López B 1994 *J. Phys. A: Math. Gen.* **27** L645

[12] Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Redwood, CA: Addison-Wesley)
[13] Iwanski J, Schietse J and Bouten M 1995 *Phys. Rev.* E **52** 888
[14] Urbanczik R 1994 *Europhys. Lett.* **26** 233
[15] de Almeida J R L and Thouless D J 1978 *J. Phys. A: Math. Gen.* **11** 983
[16] Krauth W and Mézard M 1989 *J. Physique* **50** 3057
[17] Köhler H, Diederich S, Opper M and Kinzel W 1990 *Z. Phys.* B **78** 333
[18] Kirkpatrick S, Gelatt C D and Vecchi M P 1983 *Science* **220** 671
Horner H 1992 *Z. Phys.* B **86** 291
Patel H K 1993 *Z. Phys.* B **91** 257
[19] Hopfield J J 1982 *Proc. Natl Acad. Sci., USA* **79** 2554
[20] Vallet F 1989 *Europhys. Lett.* **8** 747
Vallet F and Cailton J G 1990 *Phys. Rev.* A **41** 3059
[21] Van den Broeck C and Bouten M 1993 *Europhys. Lett.* **22** 223
[22] Seung H S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056
[23] Györgyi G 1990 *Phys. Rev.* A **41** 7097