On the generalization ability of diluted perceptrons

J. Phys. A: Math. Gen. 27 (1994) 3759–3774. Printed in the UK

On the generalization ability of diluted perceptrons

Peter Kuhlmann[†] and Klaus-Robert Müller[‡][§]

[†] Physikalisches Institut, Theoretische Physik III, Julius-Maximilians-Universität Würzburg, Am Hubland, 97074 Würzburg, Germany

[‡] GMD-Forschungsinstitut für Rechnerarchitektur und Softwaretechnik an der TU Berlin (FIRST), Rudower Chaussee 5, 12489 Berlin, Germany

Received 7 June 1993, in final form 21 January 1994

Abstract. A linearly separable Boolean function is learned by a diluted perceptron with optimal stability. A different level of dilution is allowed for teacher and student perceptron. The learning algorithms used were the optimal annealed dilution and Hebbian dilution. The generalization ability, i.e. the probability to recognize a pattern which has not been learned before, is calculated in replica symmetry.

1. Introduction

Neural networks are capable of learning how to recognize or classify patterns from given examples. They are adapting their synaptic couplings to the given problem by learning algorithms. This has been studied in detail using methods of statistical mechanics [1, 2, 4] or simulations [3].

Recently diluted neural architectures have received growning interest. Dilution gives rise to networks with fewer connections, which can be more efficient in solving tasks and which can be more easily implemented in parallel or in hardware [4, 5, 11, 12, 14, 16, 23]. Particularly, diluted perceptrons could possibly use their high generalization ability for feature extracting tasks, where not all dimensions of the incoming input patterns are of significance. Thus, a perceptron learning from examples being forced to maintain a certain dilution level is able to find out what parts of the input signal are unimportant by setting the respective weights to zero.

In this work we will study the generalization ability for the case, where a linearly separable Boolean function—given by a teacher perceptron—is learned by an optimal student perceptron with annealed dilution. Annealed dilution is a cutting algorithm which is learning a pattern set with both optimal stability and optimal dilution, i.e. optimal choice of the bonds. Since annealed dilution does not involve a practical algorithm, Hebbian dilution [11, 12] is also investigated. Hebbian dilution is an efficient approximation to annealed dilution.

Generalization problems have been analysed for various non-diluted systems previously (see e.g. [6–8, 10, 22, 24]).

2. The model

In our case we have a teaching perceptron defined by a normalized vector $\boldsymbol{B}$ which gives

[§] E-mail address: klaus@first.gmd.de

the output

$$s_0 = \operatorname{sign}\left(\sum_{j=1}^{N} B_j s_j\right) \tag{1}$$

for all inputs $s \in \{\pm 1\}^N$. The teacher

$$B = (1, \dots, 1, 0, \dots, 0)^T$$

has $N(1 - f_t)$ bits erased and is therefore projected onto a subspace of $\mathbb{R}^N$. The output function s_0 is taught to a diluted student perceptron, which adapts its synaptic weights $J \in \mathbb{R}^N$ according to the set of examples it is given by the teacher perceptron. Note that neither the degree of teacher dilution f_t nor the actual teacher distribution are known to the student.

The input set $\{\xi_i^\mu\}$ is chosen at random with Gaussian distributed patterns or $\xi_i^\mu = \pm 1$ and $i = 1, \dots, N$, $\mu = 1, \dots, p$. The outputs $\{\xi_0^\mu\}$ are generated by the teacher following (1). The student perceptron is an optimal perceptron starting in the full synaptic space searching for the teacher representation in its learning process. This is either done by cutting the 'unimportant' weights, so that both the stability and the degree of connectivity are optimized until the smaller space of dimension Nf_s is reached. Or Hebbian dilution [11, 12] is used, where we first calculate the Hebbian weights and then dilute all couplings with modulus below a certain threshold w. Note that the problem is *unlearnable* for $f_s < f_t$.

Learning and cutting are in a sense dual operations which yield an adjustment of the student to the example set given by the teacher on the reduced set of connections

$$s_0' = \operatorname{sign}\left(\sum_{j=1}^{N} c_j T_j s_j\right) \tag{1$'$}$$

where $J_j = c_j T_j$ with $c_j \in \{0, 1\}, \boldsymbol{T} \in \mathbb{R}^N$ and

$$\sum_{j=1}^{N} c_j = Nf_s \qquad f_s \in [0, 1] . \tag{2}$$

The student is constrained to a Gardner sphere [9]

$$\sum_j J_j^2 = \sum_j c_j T_j^2 = Nf_s . \tag{3}$$

We are interested in the question of how good the diluted student, who is trying to locate the teacher's missing dimensions, will generalize. The generalization ability $G(\alpha)$ is defined as the probability that a randomly chosen state s, which was not previously learned, is reproduced correctly, i.e. that the student gives the same answer as the teacher. $G(\alpha)$ is given [6] by a single parameter R defined as the cosine between teacher and student

$$R = \frac{1}{N\sqrt{f_t f_s}} \sum_j J_j B_j \tag{4}$$

and

$$G(\alpha) = 1 - \frac{1}{\pi} \cos^{-1} R . \tag{5}$$

In order to obtain the relevant model parameters analytically we perform the space of interaction calculation. For this we follow the lines of Bouten *et al* [5], Opper *et al* [6] and Kuhlmann *et al* [12], where we integrate over all T_j and include the c_j either optimally or in a quenched manner. The volume is constrained by (2) and (3). The integration over the diluted couplings T_j with $c_j = 0$ is done by using the factor

$$\exp\left[-\tfrac{1}{2}\sum_j (1-c_j)T_j^2\right] \tag{6}$$

similarly to Bouten *et al* [5]. A further constraint for the Gardner calculation is given by the input–output correlation of (1), i.e. the outputs

$$\xi_0^\mu = \mathrm{sign}\big(\sum_j B_j \xi_j^\mu\big) \tag{7}$$

are given by the teacher.

3. The calculation

3.1. Annealed dilution

The learning rule is the optimal perceptron algorithm, where the perceptron J_j is finding its maximal stability

$$\kappa = \min_\mu \left(\frac{\xi_0^\mu \sum_j c_j T_j \xi_j^\mu}{\sqrt{\sum_j c_j T_j^2}} \right) \tag{8}$$

subject to the constraints (2) and (3). So the perceptron J_j with maximal κ defines a unique hyperplane for a given set of random patterns.

We now consider the Gardner volume V subject to all constraints

$$V = \frac{1}{\mathcal{N}} \sum_{\{c_j\}} \left(\int \prod_{j=1}^N \frac{\mathrm{d}T_j}{\sqrt{2\pi}} \right) \exp\left[-\frac{1}{2}\sum_j (1-c_j)T_j^2 \right] \delta\left(\sum_j c_j T_j^2 - N f_s \right)$$
$$\times \delta_{Kr}\left[\sum_j c_j;\, f_s N \right] \prod_{\mu=1}^p \theta\left[\xi_0^\mu \sum_j c_j T_j \xi_j^\mu - \kappa\sqrt{N f_s} \right] \tag{9}$$

where $\mathcal{N}$ is the normalization. For $V \to 0$ the stability κ reaches its maximal value.

Like Gardner [9] we want to calculate $\langle \ln V \rangle$, where the angle brackets denote an averaging over the input patterns $\{\xi_i^\mu\}$. For this we consider the replicated partition function V^n and introduce suitable auxiliary variables. Subsequently, as usual, the average over the patterns $\{\xi_i^\mu\}$ and an integration over auxiliary variables is performed. We define the order parameters

$$Q_{\rho\sigma} = \frac{1}{N f_s} \sum_{j=1}^N c_j^\rho c_j^\sigma T_j^\rho T_j^\sigma \tag{10}$$

and

$$R_\rho = \frac{1}{N\sqrt{f_t f_s}} \sum_{j=1}^N B_j c_j^\rho T_j^\rho \,. \tag{11}$$

For a replica symmetric ansatz we now introduce conjugated parameters and integrate over the couplings, i.e. we sum over all possible c_j^a and integrate over T_j^a. After taking the replica limit $n \to 0$ we are left with three functions G_i for which we have to solve the saddle-point equations

$$\langle s \rangle = N^{-1}\langle \ln V \rangle = \mathrm{saddle}_\tau \sum_{i=1}^{3} G_i\,. \tag{12}$$

The expression saddle_τ means that the saddle point of the entropy s has to be taken with respect to the set of variables $\tau = \{q, R, \psi, G, E, F\}$. The functions G_i are given by

$$G_1 = \alpha \int \mathrm{D}u\mathrm{D}z \ln \Phi\left(-\frac{\kappa - z(q-R^2)^{1/2} - R|u|}{\sqrt{1-q}}\right) \tag{13}$$

$$G_2 = f_\mathrm{t} \int \mathrm{D}z\ \ln(1+\Xi) + (1-f_\mathrm{t}) \int \mathrm{D}z\ \ln(1+\Omega) \tag{14}$$

$$G_3 = \frac{f_\mathrm{s}}{2}\left(Fq + GR\sqrt{\frac{f_\mathrm{t}}{f_\mathrm{s}}} + E + \psi\right) \tag{15}$$

where Ξ and Ω are given by

$$\Xi = \frac{\mathrm{e}^{-\psi/2}}{\sqrt{E+F}} \exp \frac{(G - 2\sqrt{F}z)^2}{8(E+F)} \tag{16}$$

$$\Omega = \frac{\mathrm{e}^{-\psi/2}}{\sqrt{E+F}} \exp \frac{Fz^2}{2(E+F)}\,. \tag{17}$$

The function $\Phi(x) = \int_{-\infty}^{x} \mathrm{D}t$ is defined as usual and the Gaussian measure $\mathrm{D}t$ is given by $\mathrm{D}t = 1/\sqrt{2\pi} \exp(-1/2\ t^2)$. So the saddle-point equations with $\mathcal{G} = \sum_{i=1}^{3} G_i$ read, for instance,

$$\partial_\psi \mathcal{G} = 0 \to f_\mathrm{s} = f_\mathrm{t} \int \mathrm{D}z \frac{\Xi}{1+\Xi} + (1-f_\mathrm{t}) \int \mathrm{D}z \frac{\Omega}{1+\Omega}\,. \tag{18}$$

The first term in (18)

$$f_\mathrm{c} = f_\mathrm{t} \int \mathrm{D}z \frac{\Xi}{1+\Xi} \tag{19}$$

gives exactly the fraction of student weights coinciding with the non-zero teacher connections. With the help of the saddle-point equations we can derive an algebraic identity

$$E + Fq + \frac{1}{2}RG\sqrt{\frac{f_\mathrm{t}}{f_\mathrm{s}}} = 1 \tag{20}$$

similarly to the one found by Bouten *et al* [5], i.e. in the limit $G \to 0$ we find exactly the same identity as in [5]. The asymptotic expressions of the saddle-point equations in the limit $q \to 1$ are considered in the appendix and we finally arrive at three coupled nonlinear equations at the saddle points, which have to be solved numerically:

$$f_\mathrm{s} = f_\mathrm{t}\left(1 - \int_{z_1}^{z_2} \mathrm{D}z\right) + (1-f_\mathrm{t})(1 - \mathrm{erf}\eta) \tag{21}$$

$$\mathcal{A}_3 = \mathcal{A}_1 - R\mathcal{A}_2 \tag{22}$$

and

$$R\mathcal{A}_1 = \mathcal{A}_2\left[1 - \int_{z_1}^{z_2} \mathrm{D}z\right] + \left[\sqrt{\frac{\mathcal{A}_3}{2\pi f_\mathrm{t}}} \exp(-z^2/2)\right]_{z_1}^{z_2}\,. \tag{23}$$

The parameters $z_{1/2}$, η and $\mathcal{A}_1 - \mathcal{A}_3$ are discussed in appendix A. Equations (22) and (23) differ from the results by Opper *et al* and Bouten *et al* by an additional noise term u in $\mathcal{A}_2$, $\mathcal{A}_3$, the order parameter R responsible for the generalization part and the parameters η, f_t and f_s, which take care of the amount of dilution. Finally, after simplification of the double integrals the three equations are solved numerically using the Gauß–Kronrod method for numerical integration and the Powells hybrid method for root finding. The generalization is in our case a function in the three-dimensional parameter space of the teacher respectively student dilution f_t, f_s and α the number of examples presented, i.e. $G(f_t, f_s, \alpha)$.

3.2. Hebbian dilution

In Hebbian dilution a quenched dilution procedure takes place [12]. To determine the bonds to be removed the N Hebb-couplings

$$H_j = \frac{1}{\sqrt{N\alpha}} \sum_{\mu=1}^{p} \xi_0^\mu \xi_j^\mu \tag{24}$$

are calculated, where the outputs ξ_0^μ are given by the diluted teacher (7). Then all the bonds j with absolute values $|H_j|$ lower than a threshold value w are removed. On the remaining Nf_s bonds the perceptron of optimal stability is learned. So the learning proceeds after the

$$c_j = \Theta\left(|H_j| - w\right) \qquad j = 1, \ldots, N \tag{25}$$

have been determined. Hence the c_j are quenched variables. To obtain the threshold w as a function of the student dilution parameter f_s, we consider (2) and obtain

$$f_s = \frac{1}{N} \sum_{j=1}^{N} c_j = \frac{1}{N} \sum_{j=1}^{N} \Theta\left(|H_j| - w\right) . \tag{26}$$

By the law of large numbers, f_s is self-averaging in the limit $N \to \infty$,

$$\lim_{N\to\infty} f_s = \lim_{N\to\infty} \langle f_s \rangle_{\{\xi_j^\mu\}} . \tag{27}$$

For a teacher vector

$$\boldsymbol{B} = (1, \ldots, 1, 0, \ldots, 0)^T$$

where without loss of generality $N(1 - f_t)$ couplings have been set to zero, we have to insert

$$\xi_0^\mu = \operatorname{sign}\left(\frac{1}{\sqrt{Nf_t}} \sum_{j=1}^{Nf_t} B_j \xi_j^\mu \right) \qquad \forall \mu \tag{28}$$

in (27) and (24), respectively. Using the saddle-point method we obtain

$$f_s = f_{ch} + 2\,(1 - f_t)\,\Phi\,(-w) \tag{29}$$

where

$$f_{ch} = f_t \left(\Phi\left(-w + \sqrt{\frac{2\alpha}{\pi f_t}} \right) + \Phi\left(-w - \sqrt{\frac{2\alpha}{\pi f_t}} \right) \right) \tag{30}$$

denotes the contribution to f_s where the bonds have non-vanishing teacher connections†.

† We have also calculated f_{ch} for different teacher distributions, e.g. Gaussian distribution with mean B_0 and standard deviation $\sqrt{1 - B_0^2}$, where we note a slower ascent in f_{ch} than for $B_0 = 1$, but qualitatively the same behaviour for all B_0.

As we know how to choose the threshold w for the Hebb couplings in order to obtain a dilution f_s we can now formulate the phase space volume V for the perceptron learning rule on the remaining bonds.

We relabel all non-zero bonds j ($j = 1, \dots, N$) by the index k ($k = 1, \dots, Nf_s$). The patterns ξ_j^μ on the remaining bonds are denoted by ϑ_k^μ. We consider

$$\begin{aligned} V &= \frac{1}{\mathcal{N}} \left(\prod_{k=1}^{Nf_s} \int_{-\infty}^{\infty} \frac{\mathrm{d}J_k}{\sqrt{2\pi}} \right) \delta \left(\sum_{k=1}^{Nf_s} J_k{}^2 - Nf_s \right) \prod_{\mu=1}^{p} \Theta \left(\frac{1}{\sqrt{Nf_s}} \sum_{k=1}^{Nf_s} J_k \xi_0^\mu \vartheta_k^\mu - \kappa \right) \\ &= \frac{1}{\mathcal{N}} \left(\prod_{j=1}^{N} \int_{-\infty}^{\infty} \frac{\mathrm{d}T_j}{\sqrt{2\pi}} \right) \delta \left(\sum_{j=1}^{N} c_j T_j{}^2 - Nf_s \right) \exp \left(-\frac{1}{2} \sum_{j=1}^{N} (1 - c_j)\, T_j{}^2 \right) \\ &\quad \times \prod_{\mu=1}^{p} \Theta \left(\frac{1}{\sqrt{Nf_s}} \sum_{j=1}^{N} c_j T_j \xi_0^\mu \xi_j^\mu - \kappa \right) \end{aligned} \qquad (31)$$

where c_j from (25) denotes again whether a site j is removed and $\mathcal{N}$ normalizes the integration over the Gardner sphere. Analogous to section 3.1 we proceed using the replica method [15]. During the calculation the order parameters

$$Q_{\rho\sigma} = \frac{1}{Nf_s} \sum_{j=1}^{N} \Theta\left(|H_j'| - w \right) T_j^\rho T_j^\sigma \qquad \rho \neq \sigma \qquad \rho, \sigma = 1, \dots, n \qquad (32)$$

and

$$R_\rho = \frac{1}{N\sqrt{f_t f_s}} \sum_{j=1}^{N} \Theta\left(|H_j'| - w \right) B_j T_j^\rho \qquad \rho = 1, \dots, n \qquad (33)$$

are introduced, where H_j' is the integration variable in the identity

$$f(H_j) = \int_{-\infty}^{\infty} \mathrm{d}H_j'\, f(H_j') \delta(H_j' - H_j) \,.$$

$Q_{\rho\sigma}$ denotes the overlap between two different solutions of the perceptron problem and R_ρ is the cosine between teacher and student as in (5). We further assume replica symmetry, e.g. $Q_{\rho\sigma} = q$ and $R_\rho = R$ and take the limit $n \to 0$. This results in the entropy

$$\begin{aligned} \langle s \rangle = N^{-1} \langle \ln V \rangle = \mathrm{saddle}_\tau & \frac{f_s}{2(1-q)} - \frac{f_s}{2} + R\hat{R} + h\hat{h} - \frac{1}{2\alpha} \hat{h}^2 (1-q) + \sqrt{\frac{2}{\pi}} \hat{h} R \\ & + \frac{f_s}{2} \ln(1-q) + \frac{f_t}{2} (1-q) \left(\frac{\hat{R}^2}{f_s f_t} I_1 + 2 \frac{\hat{R}\hat{h}}{f_s \sqrt{\alpha f_t}} I_2 + \frac{\hat{h}^2}{\alpha f_s} I_3 \right) \\ & + \frac{1 - f_t}{2} (1-q) \frac{\hat{h}^2}{\alpha f_s} I_4 \\ & + 2\alpha \int_0^\infty \mathrm{D}u \int_{-\infty}^{\infty} \mathrm{D}z \ln \Phi \left(- \frac{\kappa - h - uR + z\sqrt{q - R^2}}{\sqrt{1-q}} \right) . \end{aligned} \qquad (34)$$

Here the saddle point has to be taken with respect to the set of variables $\tau = \{q, h, \hat{h}, R, \hat{R}\}$. The quantities I_1–I_4 are given in appendix B.

Finally, we note that the student perceptron reaches its maximum stability κ in the limit $q \to 1$. The value of κ is obtained from the saddle-point equation with respect to q. The detailed saddle-point equations are given in the appendix.

4. Results and discussion

4.1. Annealed dilution

In figures 1–3 we show two surfaces in the four-dimensional parameter space of the annealed problem. First we keep f_t constant and study $G(f_s, \alpha)_{f_t}$ and second we fix $\alpha = p/N$ to investigate $G(f_t, f_s)_\alpha$.

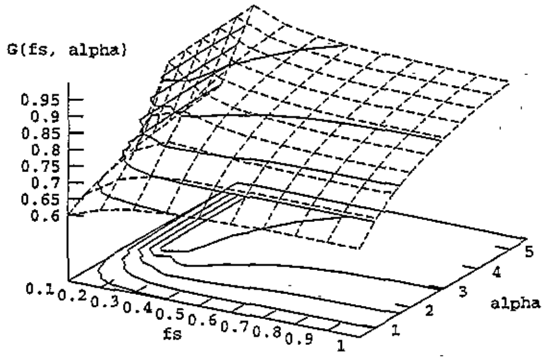

Figure 1. Generalization surface $G(f_s, \alpha)$ for $f_t = 0.2$. The changing of $G_{\max}$ can also be seen in the contour lines of the plots.

In figures 1 and 2 we see that the generalization ability is an increasing function in the number of presented examples α, as usual. An important observation is that $G(f_s, \alpha)$ reaches maxima $G_{\max}$ in f_s depending on the amount of teacher dilution f_t and α. The maxima in G are converging towards f_t for large α. This means that a student with the same number of connections can asymptotically generalize best. This principal behaviour is observed e.g. in figure 1, where the maximum is taken first for $f_s \sim 0.4$ at small values of α and for $\alpha > 2.5$ the quantity $G_{\max}$ tends towards $f_t = f_s = 0.2$. Comparing figure 1 and 2 we note, that the larger f_t the larger we have to choose α, in order to get $G_{\max}$ for $f_s = f_t$. For few learning examples a student with $f_s' > f_t$ is superior to a student with $f_s = f_t$. For a higher α we observe strong 'overfitting' effects for this student ($f_s' > f_t$) yielding $G(f_{s'}, \alpha) < G_{\max}$ (cf figure 1 and 2). The term 'overfitting' is used in analogy to function approximation problems. We now take a closer look at the overfitting effects occurring. Considering figure 4 where the student dilution is kept fixed at $f_s = 0.7$ and the teacher dilution is $f_t = 0.2$, 0.7, we find for $\alpha > 5.7$

$$G(f_t = 0.7,\ f_s = 0.7) > G(f_t = 0.2,\ f_s = 0.7)$$

which could be interpreted as the crossover point, where overfitting occurs for $G(f_t = 0.2,\ f_s = 0.7)$. It is clear that the more degrees of freedom the system has, the more probable is an overfitting effect due to the useless dimensions. Nevertheless, overfitting is dependent on α, i.e. for a small number of training examples the strongly diluted teacher ($f_t = 0.2$) is approximated better by a student with $f_s = 0.7$, than a teacher with $f_t = 0.7$ (cf figures 1, 2 and 4).

As mentioned above, the quantity f_c from (19) describes how well the student has found the relevant, non-vanishing sites in the teacher couplings. For $f_c = f_t$ the student has identified the positions of the non-zero teacher couplings. Obviously the student with ($f_t = 0.2,\ f_s = 0.7$) guessed the relevant teacher connections correctly after few examples per neuron ($\alpha > 1.5$) yielding $f_c = f_t = 0.2$ (cf curve '$f_t = 0.2, f_c$' in figure 4). For ($f_t = 0.7,\ f_s = 0.7$) the student needs up to $\alpha \sim 5.7$ examples per neuron to find the non-trivial teacher weights (cf curve '$f_t = 0.7, f_c$' in figure 4).

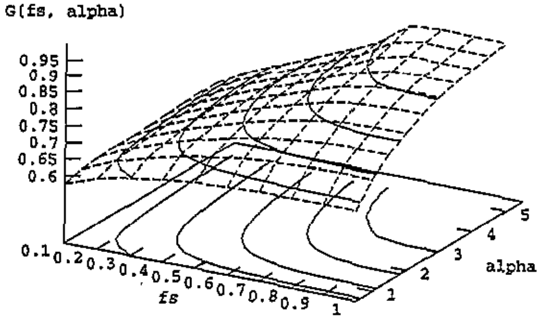

Figure 2. Generalization surface $G(f_s, \alpha)$ for $f_t = 0.7$.

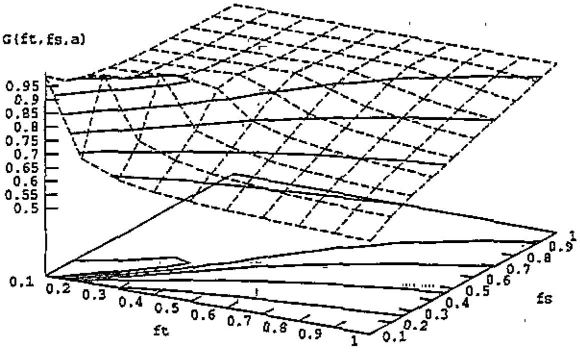

Figure 3. Generalization surface $G(f_s, f_t)$ for $\alpha = 3.5$, α is denoted by a in the diagramm.

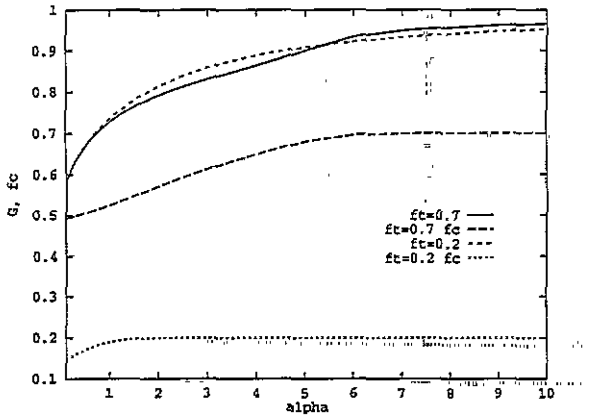

Figure 4. Generalization ability $G(f_s = 0.7, f_t)$ for different values of α with $f_t = 0.2, 0.7$ from bottom to top. f_c is the fraction of couplings for which both teacher and student have non-vanishing components (cf text).

For $\alpha > 5.7$ the student–teacher combination ($f_t = f_s = 0.7$) generalizes better—as expected—since the student with ($f_t = 0.2$, $f_s = 0.7$) has too many degrees of freedom. Asymptotically, the student with ($f_t = 0.2$, $f_s = 0.7$) will be able to reach the generalization of ($f_t = f_s = 0.7$), because the algorithm allows one to choose weights at vanishing teacher couplings to be small and asymtotically to be 0.

Comparing a diluted student J^A to a non-diluted student J^B with $f_s^B = 1$ in the range of higher α, we observe that dilution gives a significant improvement ΔG in generalization ability. This is illustrated for $\alpha = 4$, where we find

$$\Delta G(f_t = 0.7,\ f_s = 0.8,\ f_s^B = 1.0) \sim 0.004$$

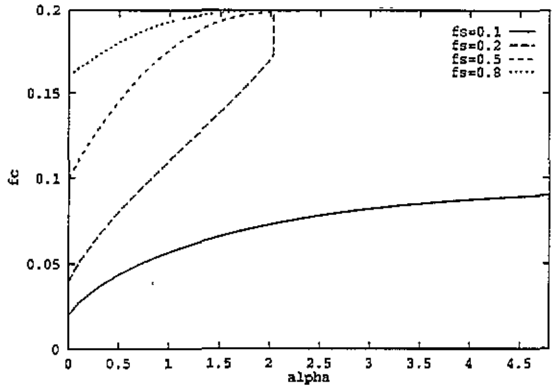

Figure 5. f_c, the fraction of couplings for which both teacher and student have non-vanishing components for $f_t = 0.2$ and different values of f_s. Note the discontinuity near $\alpha = 2.034$ in the case of $f_s = 0.2$ (for discussion see text).

$$\Delta G(f_t = 0.2,\ f_s = 0.2,\ f_s^B = 1.0) \sim 0.074\,.$$

Comparing figures 1 and 2 we see that $G(f_s = 1,\ \alpha)$ is independent of f_t. So, if the student has all his degrees of freedom $f_s = 1$, a dilution rate f_t of the teacher is of no importance. The generalization rate then coincides with the results given by Opper *et al* [6]. This is to be expected since the calculation in [6] holds for arbitrary teachers B_j for $f_s = 1$, i.e. also for teachers containing $N(1 - f_t)$ zeros.

We can see in figures 1 and 2 that $G_{\max}(f_t = 0.2) \geqslant G_{\max}(f_t = 0.7)$. This behaviour is similar to the one found for the Hebb rule at $f_s = 1$, where the generalization ability of the 'easy' case $\boldsymbol{B} = (1, 0, \ldots, 0)^{\mathrm{T}}$ is higher than the generalization ability of the 'hard' case $\boldsymbol{B} = \boldsymbol{1}$.

Now we consider the second surface $G(f_t,\ f_s)_\alpha$, where α is kept constant. From figure 3 we note a strong ridge along the diagonal of the plot for $\alpha = 3.5$, telling us that the highest generalization rates are indeed found for $f_s = f_t$. Furthermore, the smaller f_t, the stronger is the benefit taken from a diluted student and the clearer is the maximum of the generalization surface.

This behaviour is to be expected for diluted perceptrons, since the fine-tuning of the student hyperplane is of course dependent on the fact that the student has found the correct non-zero teacher sites and this process takes a different number of patterns presented for different levels of dilution as we conclude from figure 5. In figure 5 we also observe a discontinuity in f_c in the case $f_s = f_t = 0.2$. The reason for this effect is that in our calculation a new solution occurs which has a higher stability κ and a higher overlap R than the old one. This discontinuity is always present at values of $f_s \geqslant f_t$ in the vicinity of f_t. For $f_s \sim f_t$ and increasing α the student eventually accepts the teacher's couplings, because there are no alternative sites left to maximize the stability instead of the generalization rate.

In figure 6, κ is plotted as a function of α. For the parameters $f_t = 0.5$ and $f_s = 0.3$ or $f_s = 0.4$ we find $\kappa < 0$ beyond

$$\alpha_u(f_t = 0.5, f_s = 0.3) \sim 2.8 \qquad \alpha_u(f_t = 0.5, f_s = 0.4) \sim 4.2\,.$$

As discussed above, here we also encouter the problem of unlearnability [10] if $f_s < f_t$. In figure 6 we find positive stabilities for few examples even if the problem is unlearnable, obviously the student can embed the patterns correctly up to a certain number of training examples, although with low generalization ability. From a certain finite value $\alpha_u > 0$ on we find negative stabilities ($\kappa < 0$). The higher the difference between f_s and f_t the more negative is the respective stability.

For $\alpha < \alpha_u(f_s,\ f_t)$ the student network is able to give a rough first approximation of the

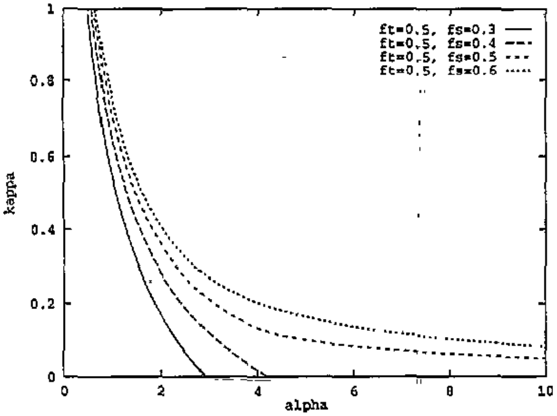

Figure 6. Stability κ as a function of α for unlearnable configurations $f_t = 0.5$, $f_s = 0.3$, 0.4 and learnable configurations $f_t = 0.5$, $f_s = 0.5$, 0.6.

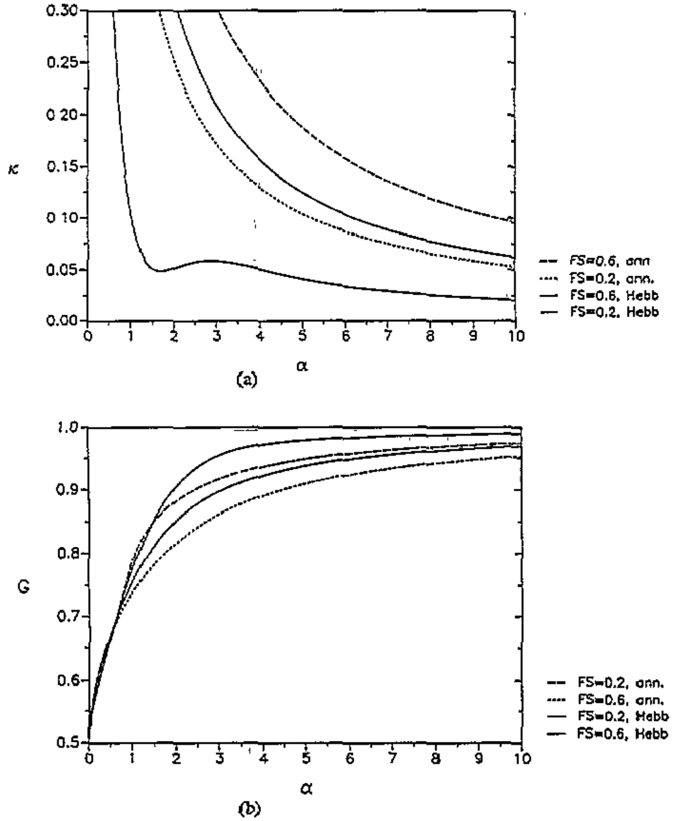

Figure 7. (a) Comparison of the stability curves $\kappa(\alpha)$ for Hebbian dilution (Hebb) and annealed (ann.) dilution, for $f_t = 0.1$ and $f_s = 0.2$, 0.6. (b) Generalization curves $G(\alpha)$ for the same parameter set.

teacher—still with positive stabilities—although it could never reach perfect generalization for $f_s < f_t$, since the degrees of freedom are not sufficient. For more examples, i.e. $\alpha > \alpha_u$ the student network is not able to meet the constraints imposed by the example set and the patterns are embedded with negative stabilities. The same diagram 6 also shows the

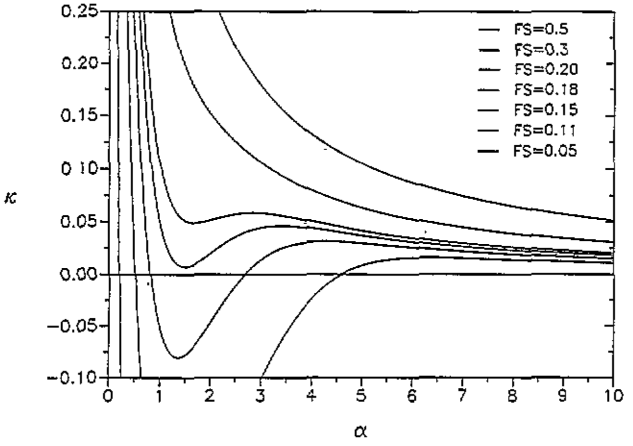

Figure 8. Stability curves $\kappa(\alpha)$ for Hebbian dilution. The parameters are $f_t = 0.1$ and different values of f_s. The $f_s = 0.5$ line gives the highest values of κ while for $f_s = 0.05$ the problem is unlearnable and κ stays negative. Note that for the learnable problem of $f_s = 0.11$ and $f_s = 0.15$ we also find a range, with $\kappa < 0$.

learnable combination $f_t = 0.5$ and $f_s = 0.5$ resp $f_s = 0.6$.

4.2. Hebbian dilution

We now compare the results found above with the results for a perceptron that has been trained by the Hebbian dilution algorithm. As expected, it can be seen from figure 7(*a*) that the stability is lower than in the optimal case, of course a higher value of f_s always yields a higher stability κ. Nevertheless the generalization rate can be higher in the Hebbian case. In figure 7(*b*) we observe that for higher α the Hebbian dilution algorithm yields higher generalization rates than the optimal algorithm. The reason for this are two effects: 'overfitting' and optimization of stability. Note, that both values $f_s = 0.2,\ 0.6$ in figure 7(*b*) give rise to strong 'overfitting' effects for $f_t = 0.1$. For $f_s = 0.6$ we find an intersection of the generalization curves at $\alpha_i \sim 0.5$ with a higher generalization rate for annealed dilution below α_i and higher generalization rate for Hebbian dilution above α_i. The same effect occurs for $f_s = 0.2$ at a higher value of α_i. For $f_s \to f_t$ we expect $\alpha_i \to \infty$, i.e. a generalization rate for annealed dilution above Hebbian dilution. In figure 11 we observe that the maximum generalization rate of annealed dilution is higher than in the Hebbian case, as expected. Moreover, the 'overfitting' effects are seen to become stronger with growing α, for $\alpha = 6$ we note an intersection of both generalization curves similar as in figure 7(*b*). We can conclude Hebbian dilution is less sensitive to 'overfitting' effects than annealed dilution.

The annealed dilution algorithm chooses its sites in order to maximize the stability, but it loses track of the problem to generalize optimally. Contrarily, the Hebbian dilution algorithm does concentrate on the task of generalization, but it yields low stabilities for small values of α.

From figure 8 it can be seen that negative stabilities can even occur, if f_s is greater than, but close to f_t. The reason for this behaviour is that the student has not yet determined the teacher's relevant sites. For higher α the stability κ increases again, reaches a positive maximum and decreases to zero for $\alpha \to \infty$. For higher values of f_s the stability will stay positive for all α. In a medium range of f_s the stability κ can still have a local minimum with respect to α.

If we consider the stability parameter κ, we find that κ is a monotonously increasing function in f_s reaching a maximum for $f_s = 1$. As also known from undiluted models stabilities are high for a few examples and decrease monotonously for increasing α.

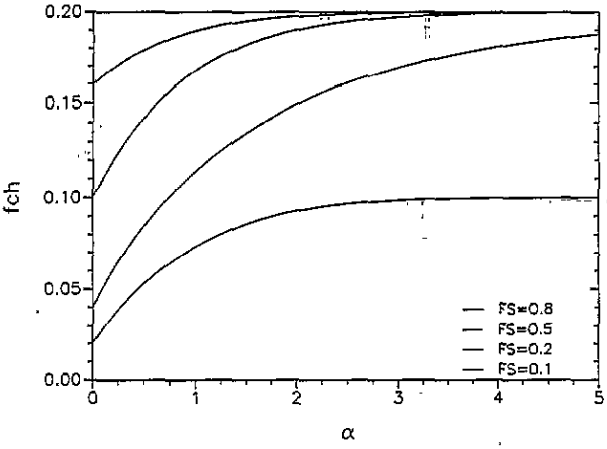

Figure 9. The fraction of student sites f_{ch} coinciding with the non-vanishing teacher sites for Hebbian dilution as discussed in the text ($f_{\mathrm{t}} = 0.2$).

So during the learning procedure there are two competing effects. On one side the perceptron tries to learn an increasing number of patterns, and on the other side it chooses the sites on which the teacher is expected to be active. Hence it is conceivable that there is a medium range of α in which the choice of the active sites is most important.

For the unlearnable problem $f_s < f_t$ negative stabilities occur, if α increases. As in the optimal case, the student can embed a few patterns correctly, whereas from a value α_u on we find $\kappa < 0$.

Figure 9 shows that f_{ch} from (30) increases with α. In the limit $\alpha \to \infty$ the active sites will be determined eventually if $f_s > f_t$. By an expansion for high α in (29) we obtain $f_{\mathrm{ch}} \to f_t$ for $\alpha \to \infty$ in this case. If $f_s < f_t$ the teacher cannot be learned for higher α. Nevertheless the student only acts on the teacher's active sites, i.e. $f_{\mathrm{ch}} \to f_s$ for $\alpha \to \infty$.

If we examine the dependence of the generalization rate on the student dilution f_s we observe an 'overfitting' effect as in the annealed dilution case (see figures 10 and 11). We also note that the maximum of the generalization rate with respect to the student dilution f_s converges to f_t as α increases. As described above the student gets more and more information about the teacher's relevant sites as α increases. If the student's objective is a high generalization ability he will have to operate at a low value of f_s. But for a practical algorithm like the Hebbian dilution method positive stabilities κ must be reached. Consequently, in order to obtain a positive κ the student is always required to work on the right-hand side of the dotted $\kappa = 0$ curve in figure 10.

5. Conclusions

We have shown how well a student perceptron generalizes, which is forced to a certain degree of dilution f_s. It was demonstrated that a diluted student can yield a strong improvement of generalization as compared with the non-diluted student, if the teacher is also diluted to a degree f_t.

The theoretical annealed dilution algorithm and the practical Hebbian dilution algorithm appeared to behave similarly with respect to overfitting. To make use of the higher generalization ability at high values of α, one has to choose the student dilution f_s at the lowest possible value that still guarantees a positive stability κ. Thus the overfitting effect will be avoided.

It can be seen that the smaller f_t the easier the teacher hyperplane can be approximated

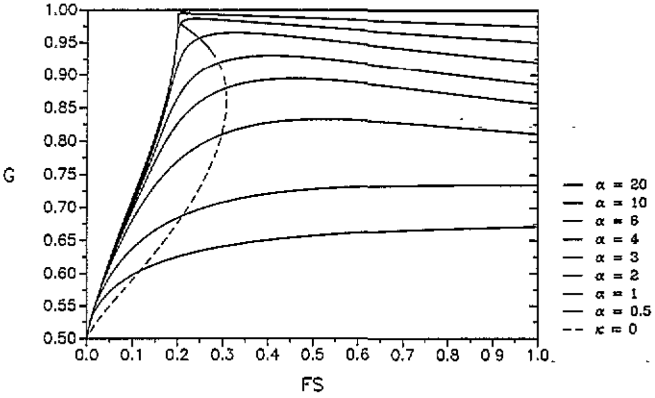

Figure 10. Comparison of generalization curves $G(f_s)$ for $f_t = 0.2$ and different values of α. The value $\alpha = 20$ gives the highest generalization ability. The broken curve shows the transition line of $\kappa = 0$, on the left of this line we find $\kappa < 0$ and on the right $\kappa > 0$.

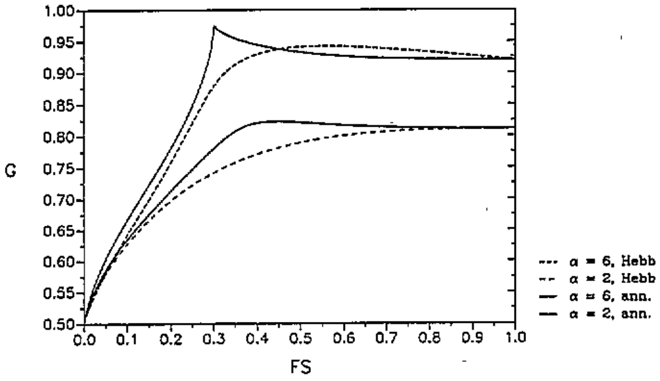

Figure 11. Comparison of generalization curves $G(f_s)$ for $f_t = 0.3$ and $\alpha = 2$, 6 for Hebbian dilution and annealed dilution. We observe strong 'overfitting' effects for $\alpha = 6$ (upper two curves), where the Hebbian dilution curve is intersecting with the annealed dilution curve.

by a given degree f_s of the student dilution. The quantity f_c demonstrates how fast the student learns the relevant teacher connections, i.e. how fast he knows the relevant subspace spanned by the teacher vector $\boldsymbol{B}$. A well performing annealed dilution algorithm would make this result interesting for applications, because for higher values of α a diluted system works more efficiently than a fully connected system.

In our analytical calculation of the generalization rate in the annealed case we made the replica symmetric approximation, since we were only interested in the qualitative behaviour of this algorithm. Nevertheless the discontinuity in f_c in figure 5 deserves further investigation in the context of a replica-symmetry breaking calculation of first order.

The Hebbian dilution algorithm is a good approximation—although not optimal—to annealed dilution. The results show that this practical algorithm has qualitatively the same behaviour as the optimal algorithm, and is even less sensitive to overfitting as we observed in figure 7(*b*). For the Hebbian dilution we do not have replica-symmetry breaking effects for positive stabilities, because the c_j are quenched and therefore the Hebbian case is similar to the standard perceptron problem, i.e. the phase space is simply connected and only one—replica symmetric—solution exists [17]. While replica symmetry breaking is known to occur in the annealed dilution case for all stabilities κ [12], it should only occur for unlearnable problems, i.e. negative κ, for quenched dilution algorithms [17].

Nevertheless the qualitative behaviour of the learning error (see [6, 17]) of the diluted perceptron may be of importance for the construction of new learning and generalization strategies. However, in the context of generalization, one could conceive an approach similar to [23], where the learning error has been studied for several models of diluted networks. In particular, the learning error could be used for comparison of two dilution algorithms that have nearly the same generalization rates and the same stabilities.

In summary, we conclude that it is worthwhile using systematically diluted networks in order to obtain an increased generalization behaviour. Therefore a diluted perceptron could be used practically as a basic element for feature extraction, since it can learn to detect wildcards, i.e. unimportant channels in input signals.

Another possible learning algorithm to be analysed would be: perceptron learning, removal of all weights below a threshold w and perceptron learning on the remaining couplings [11, 12, 16]. We would expect the same qualitative behaviour for this algorithm. Furthermore, it would be interesting to investigate multilayer diluted perceptrons arranged as a commitee or parity machine [18–21], here we would also expect a gain in efficiency.

Acknowledgments

K-RM acknowledges partial financial support by Landesgraduiertenförderung Baden-Württemberg and A Glenz. PK is supported by DFG. This work is part of the PhD thesis of K-RM and PK. It is a pleasure to thank Reimer Kühn, Manfred Opper, W Kinzel and Werner M Seiler for stimulating discussions.

Appendix A. Annealed dilution

If we consider the asymptotic expressions of the saddle-point equations (12) in the limit $q \to 1$, we find for $E+F$

$$\mathcal{A}_1 = \lim_{q\to1} f_s(E+F)(1-q) = f_t \int \mathrm{D}z \frac{\Xi}{1+\Xi}\left(z^2 - z\frac{G}{2\sqrt{F}}\right) + (1-f_t)\int \mathrm{D}z \frac{\Omega}{1+\Omega} z^2 \tag{A1}$$

the same $(1-q)^{-1}$ behaviour which also holds for G from $\partial_R \mathcal{G} = 0$.

$$\mathcal{A}_2 = \lim_{q\to1} -\tfrac{1}{2}\sqrt{f_t f_s}\, G(1-q) = \alpha\, \partial_R \int_{\mathcal{D}} \mathrm{D}u\, \mathrm{D}z\, \left(\kappa - z(1-R^2)^{1/2} - R|u|\right)^2 \tag{A2}$$

whereas F from $\partial_q \mathcal{G} = 0$ shows a $(1-q)^{-2}$ asymptotic behaviour:

$$\mathcal{A}_3 = \lim_{q\to1} f_s F(1-q)^2 = \alpha \int_{\mathcal{D}} \mathrm{D}u\, \mathrm{D}z\, \left(\kappa - z(1-R^2)^{1/2} - R|u|\right)^2 . \tag{A3}$$

The integration in (A2) and (A3) has to be done over the two-dimensional domain $\mathcal{D}$ defined by $\mathcal{D} = \{\kappa - z(1-R)^{1/2} - R|u| > 0\}$. Using (20) we derive the asymptotic identity

$$\lim_{q\to1} f_s F(1-q)^2 = \lim_{q\to1} f_s(1-q)\left(E + F + \frac{1}{2} G R \sqrt{\frac{f_t}{f_s}}\right) . \tag{A4}$$

On our way to (21)–(23) we had to evaluate expressions containing Ξ and Ω, e.g.

$$\int \mathrm{D}z \frac{\Xi}{1+\Xi} = 1 - \int_{z_1}^{z_2} \mathrm{D}z . \tag{A5}$$

This is done by considering the asymptotic properties of $(G-2\sqrt{F}z)^2/(8(E+F))$ and $\mathrm{e}^{-\psi/2}/(\sqrt{E+F})$ discussed in (A1)–(A3). For $q \to 1$ the quantity Ξ either goes to ∞ or 0, so the integral (A5) only contributes for $\Xi^{-1} \to 0$. Using the abreviations

$$\mathrm{e}^{\sigma} = \mathrm{e}^{\psi/2}\sqrt{E+F} \qquad a = \frac{F}{E+F}$$

and $\eta^2 = \sigma/a$, we get

$$\Xi^{-1} = \exp\left(-\frac{a}{2}\left[z^2 - \frac{G}{\sqrt{F}}z + \frac{G^2}{4F} - 2\eta^2\right]\right). \tag{A6}$$

Since we know that $a \to \infty$ holds for $q \to 1$, the polynomial in z in the exponent of (A6) has to be positive in order to yield $\Xi^{-1} \to 0$. This condition is fulfilled in both intervals $[z_2, \infty]$ and $[-\infty, z_1]$. The roots of the polynomial in (A6) are given as

$$z_{1/2} = \mp\sqrt{2}\eta + \frac{G}{2\sqrt{F}}. \tag{A7}$$

Appendix B. Hebbian dilution

We first define the quantities I_1–I_4 used for the entropy in (34):

$$I_1 = \Phi(w_1) + \Phi(-w_2) \tag{B1}$$

$$I_2 = \sqrt{\frac{2\alpha}{\pi f_t}}\,(\Phi(w_1) + \Phi(-w_2)) + \frac{1}{\sqrt{2\pi}}\left(\exp\left(-\tfrac{1}{2}w_2^2\right) - \exp\left(-\tfrac{1}{2}w_1^2\right)\right) \tag{B2}$$

$$\begin{aligned} I_3 = {} & \left(1 + \frac{2\alpha}{\pi f_t}\right) I_1 + \sqrt{\frac{4\alpha}{\pi^2 f_t}}\left(\exp\left(-\tfrac{1}{2}w_2^2\right) - \exp\left(-\tfrac{1}{2}w_1^2\right)\right) \\ & + \sqrt{\frac{1}{2\pi}}\left(w_2 \exp\left(-\tfrac{1}{2}w_2^2\right) - w_1 \exp\left(-\tfrac{1}{2}w_1^2\right)\right) \end{aligned} \tag{B3}$$

$$I_4 = 2\left(\Phi(-w) + \frac{w}{\sqrt{2\pi}}\exp\left(-\tfrac{1}{2}w^2\right)\right) \tag{B4}$$

and

$$w_1 = -w - \sqrt{\frac{2\alpha}{\pi f_t}} \tag{B5}$$

$$w_2 = w - \sqrt{\frac{2\alpha}{\pi f_t}}. \tag{B6}$$

To obtain the detailed saddle-point equations, we first define auxiliary variables

$$\hat{S} = (1-q)\hat{R} = -\frac{f_s R + I_2\sqrt{(f_t/\alpha)}D}{I_1 + I_2\sqrt{(f_t/\alpha)}C} \tag{B7}$$

where

$$C = -\sqrt{\frac{f_t}{f_s^2\alpha}}I_2 \frac{1}{-1/\alpha + f_t I_3/\alpha f_s + (1-f_t)I_4/\alpha f_s} \tag{B8}$$

$$D = \left(-h - \sqrt{\frac{2}{\pi}}R\right) \frac{1}{-1/\alpha + f_t I_3/(\alpha f_s) + (1-f_t)I_4/(\alpha f_s)} \tag{B9}$$

$$\hat{G} = (1-q)\hat{h} = C\hat{S} + D\,. \tag{B10}$$

We introduce the abbreviation for integrals over polynomials and Gauss functions

$$\mathcal{I}(s,u,v,w) = \int_{-s}^{\infty} \mathrm{D}z(uz^2 + vz + w)\,. \tag{B11}$$

Now the saddle-point equations in the limit $q \to 1$ read

$$-\hat{S} = \sqrt{\frac{2}{\pi}}\hat{G} + 2\alpha \int_0^{\infty} \mathrm{D}u\, \mathcal{I}\left(t, R, (Rt + u\sqrt{1-R^2}), ut\sqrt{1-R^2}\right) \tag{B12}$$

$$-\hat{G} = 2\alpha \int_0^{\infty} \mathrm{D}u\, \mathcal{I}\,(t, 0, 1, t)\,\sqrt{1-R^2} \tag{B13}$$

$$\begin{aligned}-f_s = \frac{1}{\alpha}\hat{G}^2 - \frac{\hat{S}^2}{f_s}\,I_1 - 2\hat{S}\hat{G}\frac{I_2}{f_s}\sqrt{\frac{f_t}{\alpha}} - \frac{f_t\hat{G}^2}{\alpha f_s}\,I_3 \\ + \;-(1-f_t)\frac{I_4}{\alpha f_s}\,\hat{G}^2 - 2\alpha(1-R^2)\int_0^{\infty} \mathrm{D}u\, \mathcal{I}\left(t, 1, 2t, t^2\right)\end{aligned} \tag{B14}$$

where $t = (\kappa - h - uR)/\sqrt{1-R^2}$ is a function of the integration variable u. For a given α the three saddle-point equations are solved for κ, h and R.

References

[1] Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev. Lett.* 55 1530; 1985 *Phys. Rev.* A **32** 1007; 1987 *Ann. Phys., NY* 173 30

[2] Kinzel W and Opper M 1991 Dynamics of learning *Physics of Neural Networks* ed E Domany, J L van Hemmen and K Schulten (Berlin: Springer)

[3] Forrest B M 1988 *J. Phys. A: Math. Gen.* **21** 245
Kepler T B and Abbott L F 1988 *J. Physique* 49 1657

[4] Canning A and Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 3275

[5] Bouten M, Engel A, Komoda A and Serneels R 1990 *J. Phys. A: Math. Gen.* **23** 4643

[6] Opper M, Kinzel W, Kleinz J and Nehl R 1990 *J. Phys. A: Math. Gen.* **23** L581

[7] Vallet F 1989 *Europhys. Lett.* **8** 3824

[8] Györgyi G and Tishby N 1990 *Proc. STATPHYS-17 Workshop on Neural Networks and Spin Glasses* ed W K Theumann and R Köberle p 3 (Singapore: World Scientific)

[9] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257; 1989 *J. Phys. A: Math. Gen.* **22** 1969

[10] Watkin T L H, Rau A and Biehl M 1993 *Rev. Mod. Phys.* **65** 499

[11] Müller K-R 1992 *PhD Thesis* Fakultät für Informatik, University Karlsruhe, appeared as GMD-Bericht 218 (in German) R Oldenburg Verlag München (1994)
Müller K-R 1991 *Proc. ISCIS VI conf.* ed M Barray and B Özgüç (New York: Elsevier) p 845
Stiefvater T, Müller K-R and Janßen H 1993 *Network: Comput. Neural Syst.* **4** 313

[12] Kuhlmann P 1993 *PhD Thesis* Fakultät für Physik, Justus–Liebig–Universität Gießen (in German)
Kuhlmann P, Garces R and Eissfeller H 1992 *J. Phys. A: Math. Gen.* **25** L593
Garces R, Kuhlmann P and Eissfeller H 1992 *J. Phys. A: Math. Gen.* **25** L1335

[13] Vallet F and Cailton J-G 1990 *Phys. Rev.* A **41** 3059

[14] Jackel L D *et al* 1986 *The Physics and Fabrication of Microstructures* pp 453

[15] van Hemmen J L and Palmer R G 1979 *J. Phys. A: Math. Gen.* **12** 563

[16] Janowsky S A 1989 *Phys. Rev.* A 39 6600

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

[18] Biehl M and Opper M 1991 *Phys. Rev.* A **44** 6888

[19] Mézard M aand Nadal J–P 1989 *J. Phys. A: Math. Gen.* **22** 2191

[20] Rujan P 1990 *Statistical Mechanics of Neural Networks* ed Luis Garrido (Berlin: Springer)

[21] Schmitz H J *et al* 1990 *J. Physique* **51** 167

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

[23] Wong K Y M and Bouten M 1991 *Europhys. Lett.* **16** 525

[24] Wong K Y M and Sherrington D 1990 *J. Phys. A: Math. Gen.* **23** 4659