Generalization from educated teachers

1992 J. Phys. A: Math. Gen. 25 6649

(http://iopscience.iop.org/0305-4470/25/24/018)

J. Phys. A: Math. Gen. 25 (1992) 6649–6668. Printed in the UK

Generalization from educated teachers

L Reimers and A Engel

Institut für Theoretische Physik, Georg-August-Universität, Bunsenstrasse 9, D-3400 Göttingen, Federal Republic of Germany

Received 6 July 1992

Abstract. We study the generalization ability of a perceptron as a function of the structural similarity between teacher and student. To this end we consider a teacher perceptron $\boldsymbol{B}$ designed to implement a certain number of random input–output mappings and a student perceptron $\boldsymbol{J}$ trained with input–output examples generated by the teacher. If the pattern sets of teacher and student overlap, different combinations of teacher and student learning rules result in different generalization probabilities. Generalization can be facilitated or can be hindered depending on the combination of learning rules and the storage levels of teacher and student. Analytical calculations of the generalization ability for several combinations of teacher and student learning rules are in very good agreement with numerical simulations.

1. Introduction

One of the more interesting problems in the theory of networks of formal neurons concerns their ability to generalize. By this one has in mind that a network can infer a rule from input–output examples produced by this rule (for recent reviews see [1, 2]. A model situation amenable to a statistical mechanics analysis is provided by a perceptron $\boldsymbol{J}$, referred to as the student in the following, trained with input–output configurations which are generated by another perceptron $\boldsymbol{B}$, called the teacher [3, 1]. To be precise consider a set of $\alpha_s N$ N-bit patterns $\boldsymbol{\xi}^\mu = \{\xi_j^\mu\}$, $j = 1, \ldots, N$; $\mu = 1, \ldots, \alpha_s N$ generated at random with $\xi_j^\mu = \pm 1$ with equal probability. The teacher $\boldsymbol{B}$ associates with every input vector $\boldsymbol{\xi}^\mu$ an output η^μ according to

$$\eta^\mu = \operatorname{sign}(\boldsymbol{B} \cdot \boldsymbol{\xi}^\mu) = \operatorname{sign}\left(\sum_{j=1}^{N} B_j \xi_j^\mu\right) . \qquad (1)$$

Using a special learning rule the student $\boldsymbol{J}$ is then trained to reproduce these input–output relations $\{\boldsymbol{\xi}^\mu, \eta^\mu\}$. The generalization ability is defined as the probability (averaged over the statistics of the patterns) that for an *additional* random input vector $\boldsymbol{\xi}^{p_s+1}$ the output of teacher and student are the same:

$$g(\alpha_s) = \langle \operatorname{Prob}\{(\boldsymbol{B} \cdot \boldsymbol{\xi}^{p_s+1})(\boldsymbol{J} \cdot \boldsymbol{\xi}^{p_s+1}) > 0\} \rangle_{\boldsymbol{\xi}} \cdot \qquad (2)$$

Clearly $g(\alpha_s = 0) = 0.5$ which corresponds to a random guess and $g(\alpha_s \to \infty) \to 1$. Remarkably the detailed dependence of the generalization ability g on the number

0305-4470/92/246649+20$07.50

$\alpha_s N$ of training examples is independent of the particular teacher perceptron $\boldsymbol{B}$ as long as the B_i are uncorrelated with the patterns $\boldsymbol{\xi}^\mu$. In fact it turns out that $g(\alpha_s)$ depends only on the norm $\|\boldsymbol{B}\| = \sqrt{\sum_j B_j^2}$ of $\boldsymbol{B}$. Consequently the interest has focused on the properties of the *student* and $g(\alpha_s)$ has been determined for several interesting situations such as different learning rules of the student [4–8], perceptrons with binary synapses [6], 'intelligent' students selecting the examples according to their present state of knowledge [9, 10] and 'complex' students as exemplified by multilayer perceptrons [11, 12].

In the present paper we are interested in the complementary problem, namely the dependence of the generalization ability of the student on the *structure of the teacher*. To this end we assume that the teacher $\boldsymbol{B}$ is a perceptron also designed by a special learning rule to implement $\alpha_t N$ input–output relations $\{\boldsymbol{\zeta}^\nu, \tau^\nu\}$, where now the ζ_i^ν as well as the τ^ν are independent random variables. The so-constructed teacher then associates an output η^μ to all questions $\boldsymbol{\xi}^\mu$ of the student according to (1) and the generalization ability is again given by (2). It is clear that there will be no difference to the case of an uncorrelated teacher described above if the pattern sets $\{\boldsymbol{\zeta}^\nu\}$ and $\{\boldsymbol{\xi}^\mu\}$ are disjoint. If, however, part of the student's questions $\{\boldsymbol{\xi}^\mu\}$ belong to the patterns $\{\boldsymbol{\zeta}^\nu\}$ the teacher himself is trained with non-trivial correlations, with interesting consequences.

Our main interest will be to study whether the generalization ability is different for a teacher and a student using the same learning rule, i.e. being 'structured similarly', as compared with two perceptrons using different learning rules. It seems at first sight that it will always be advantageous to the student to ask questions out of the set $\{\boldsymbol{\zeta}^\nu\}$; this is, however, not the case. Moreover, it is interesting to investigate whether the very peculiar phenomenon of 'overfitting' as found for a student using the pseudo-inverse rule to generalize an uncorrelated teacher [5] can be compensated by using a teacher also designed by this rule.

The paper is organized as follows. In section 2 we briefly review some results on the generalization ability for an uncorrelated teacher relevant for a later comparison with our findings. Section 3 defines our model and sketches the determination of $g(\alpha_s)$ for several combinations of teacher and student learning rules. Some representative calculations are discussed in somewhat more detail in the appendices. In section 4 we discuss our analytical results and compare them with numerical simulations finding very good agreement. Finally section 5 contains our conclusions.

2. Generalization from uncorrelated teachers

In this section we briefly review the results obtained for the generalization ability in the case of an uncorrelated teacher which we need for later comparison. We mainly refer to the paper of Opper *et al* [5]. The teacher perceptron is given by an arbitrary vector $\boldsymbol{B} \in \boldsymbol{R}^N$ with norm $\|\boldsymbol{B}\| = \sqrt{N}$. The student's questions are random patterns $\boldsymbol{\xi}^\mu$, $\mu = 1, \ldots, \alpha_s N$ where the ξ_j^μ are independently chosen to be ± 1 with equal probability. It is essential that $\boldsymbol{B}$ is the same for all realizations $\{\boldsymbol{\xi}^\mu\}$ of the patterns, i.e. there are no correlations between the structure of the teacher and the questions of the student. The corresponding answers η^μ of the teacher are given by (1). We consider three learning rules for the student in detail:

(i) *The Hebb rule*. Here $\boldsymbol{J}$ is explicitly given by

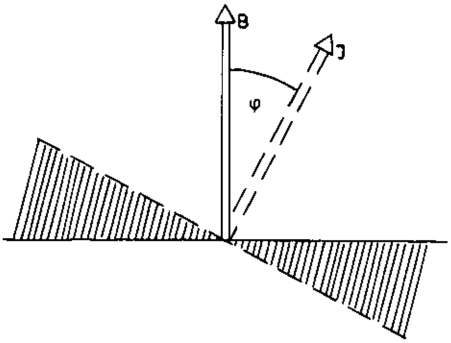

Figure 1. Two-dimensional hyperplane in the space of patterns $\{\xi^\mu\}$ spanned by the synaptic vectors $\boldsymbol{B}$ and $\boldsymbol{J}$. Teacher and student give different output for patterns with projection in the shaded region. For large N the generalization ability will thus approach $g = 1 - \varphi/\pi = 1 - (1/\pi)\cos^{-1}\rho$.

$$J_j = \frac{1}{\sqrt{\alpha_s N}} \sum_\mu \xi_j^\mu \eta^\mu . \tag{3}$$

(ii) *The pseudo-inverse rule.* $\boldsymbol{J}$ is defined as the vector minimizing

$$E(\boldsymbol{J}) = \sum_\mu \left(\eta^\mu - \frac{1}{\sqrt{N}} \sum_j J_j \xi_j^\mu \right)^2 . \tag{4}$$

For $\alpha_s < 1$ the minimum of E is zero and there are different vectors with this value of E. $\boldsymbol{J}$ is then defined as the vector with $E(\boldsymbol{J}) = 0$ and minimal norm. For $\alpha_s > 1$ (4) defines $\boldsymbol{J}$ uniquely.

(iii) *The optimal perceptron.* Following Gardner [13] $\boldsymbol{J}$ is defined as the vector with norm $\|\boldsymbol{J}\| = \sqrt{N}$ maximizing the so-called stability

$$\kappa = \min_\mu \frac{1}{\sqrt{N}} \sum_j J_j \xi_j^\mu \eta^\mu . \tag{5}$$

All the analytical calculations are done in the thermodynamic limit $N \to \infty$. The generalization ability $g(\alpha_s)$ as defined by (2) then depends only on the normalized overlap

$$\rho(\alpha_s) = \left\langle \frac{\boldsymbol{B} \cdot \boldsymbol{J}}{\|\boldsymbol{B}\| \, \|\boldsymbol{J}\|} \right\rangle_{\boldsymbol{\xi}} \tag{6}$$

between the synaptic vectors of teacher and student and is given by [5]

$$g(\alpha_s) = 1 - \frac{1}{\pi} \cos^{-1} \rho(\alpha_s) . \tag{7}$$

Figure 1 gives a geometrical interpretation of this relation. The overlap (6) can be calculated analytically for the three student learning rules defined by (3)–(5). The results for $g(\alpha_s)$ are summarized in figure 2. For $\alpha_s \leqslant 0.4$ the three learning rules give rather similar results for g whereas for large α_s the optimal perceptron generalizes best. Very remarkable is the minimum of $g(\alpha_s)$ at $\alpha_s = 1$ for the pseudo-inverse rule. The decrease of the generalization ability for $0.6 < \alpha_s < 1$ is called 'overfitting' and is due to the fact that the student tries to reproduce the examples too accurately [14].

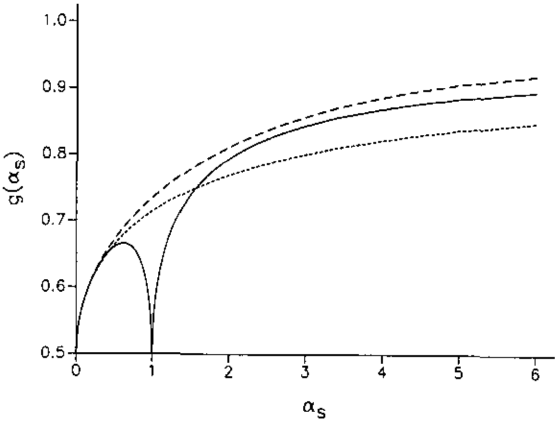

Figure 2. Generalization ability as a function of the relative number of training examples for an uncorrelated teacher and different student learning rules; full curve: pseudo-inverse rule; broken curve: optimal perceptron; dotted curve: Hebb rule (from [5]).

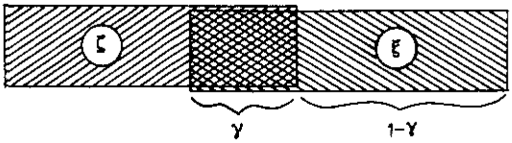

Figure 3. Partition of the set of examples in 'known' and 'unknown' questions.

3. Generalization from educated teachers

We now study the situation where the teacher has been designed himself by some learning rule. Accordingly the couplings $\boldsymbol{B}$ were adjusted to realize a given set of random input–output relations $\{\boldsymbol{\zeta}^{\nu}, \tau^{\nu}\}$, $\nu = 1, \ldots, \alpha_t N$. The student is then trained with examples $\{\boldsymbol{\xi}^{\mu}, \eta^{\mu}\}, \mu = 1, \ldots, \alpha_s N$, where the $\boldsymbol{\xi}^{\mu}$ are again chosen randomly and the η^{μ} are generated by the teacher according to $\eta^{\mu} = \mathrm{sign}(\boldsymbol{B} \cdot \boldsymbol{\xi}^{\mu})$. As explained in the introduction the interesting situation arises if the pattern sets $\{\boldsymbol{\xi}^{\mu}\}$ and $\{\boldsymbol{\zeta}^{\nu}\}$ do overlap. We therefore assume

$$\boldsymbol{\zeta}^{\mu'} = \boldsymbol{\xi}^{\mu'} \qquad \text{for} \quad \mu' = 1, \ldots, \gamma \alpha_s N \tag{8}$$

with $0 \leqslant \gamma \leqslant 1$. Hence the student asks $\gamma \alpha_s N$ questions 'known' and $(1-\gamma)\alpha_s N$ questions 'unknown' to the teacher. The relation between the pattern sets $\{\boldsymbol{\xi}^{\mu}\}$ and $\{\boldsymbol{\zeta}^{\nu}\}$ characterized by γ is schematically depicted in figure 3. Note that since α_t is fixed during a learning session of the student we have always $\alpha_s \leqslant \alpha_t/\gamma$. Hence the curves $g(\alpha_s)$ for fixed γ will not continue to $\alpha_s \to \infty$ but end at $\alpha_s = \alpha_t/\gamma$. It is impossible to continue learning beyond this point without reducing the percentage γ of known examples.

In order to calculate the generalization ability $g(\alpha_s)$ one has to determine the overlap

$$\rho = \left\langle \frac{\boldsymbol{B} \cdot \boldsymbol{J}}{\|\boldsymbol{B}\| \|\boldsymbol{J}\|} \right\rangle \tag{9}$$

where the average is over the distributions of $\boldsymbol{\zeta}^{\nu}$, τ^{ν} and $\boldsymbol{\xi}^{\mu}$. Note that for $\gamma = 0$ or $\alpha_t \to \infty$ we are dealing with an uncorrelated teacher and therefore have to reproduce the results discussed in section 2.

In the present section we discuss the calculation of $g(\alpha_s)$ for a Hebbian student generalizing a Hebbian teacher (case A) and for a Hebbian student generalizing a pseudo-inverse teacher (case B). We sketch the derivations for a pseudo-inverse student and a Hebbian teacher (case C) and finally a pseudo-inverse student with a pseudo-inverse teacher (case D). The latter two cases are explained in more detail in the appendices. The results are discussed together in the next section.

Case A. Let us first consider the simplest case where both teacher and student use the Hebb rule. We then have

$$B_j = \frac{1}{\sqrt{\alpha_t N}} \sum_{\nu=1}^{\alpha_t N} \tau^\nu \zeta_j^\nu \qquad (j = 1, \ldots, N) \tag{10}$$

$$J_j = \frac{1}{\sqrt{\alpha_s N}} \sum_{\mu=1}^{\alpha_s N} \mathrm{sign}(\boldsymbol{B} \cdot \boldsymbol{\xi}^\mu) \xi_j^\mu \qquad (j = 1, \ldots, N)\,. \tag{11}$$

For $N \to \infty$ the distributions of $R := \frac{1}{N}\boldsymbol{B} \cdot \boldsymbol{J}$, $\frac{1}{N}||\boldsymbol{B}||^2$ and $\frac{1}{N}||\boldsymbol{J}||^2$ become sharp and we have

$$\rho = \frac{\langle R \rangle}{\sqrt{\langle \frac{1}{N}||\boldsymbol{B}||^2 \rangle \langle \frac{1}{N}||\boldsymbol{J}||^2 \rangle}}\,. \tag{12}$$

From (10) we immediately find $\langle \frac{1}{N}||\boldsymbol{B}||^2 \rangle = 1$. To average R and $\frac{1}{N}||\boldsymbol{J}||^2$ we have to distinguish between known and unknown examples:

$$\begin{aligned} R = \frac{1}{N}\boldsymbol{B} \cdot \boldsymbol{J} &= \frac{1}{N\sqrt{\alpha_s}} \sum_{\mu'=1}^{\gamma\alpha_s N} \mathrm{sign}\left(\frac{1}{\sqrt{N}}\boldsymbol{B} \cdot \boldsymbol{\zeta}^{\mu'}\right) \frac{1}{\sqrt{N}}\boldsymbol{B} \cdot \boldsymbol{\zeta}^{\mu'} \\ &\quad + \frac{1}{N\sqrt{\alpha_s}} \sum_{\mu''=\gamma\alpha_s N+1}^{\alpha_s N} \mathrm{sign}\left(\frac{1}{\sqrt{N}}\boldsymbol{B} \cdot \boldsymbol{\xi}^{\mu''}\right) \frac{1}{\sqrt{N}}\boldsymbol{B} \cdot \boldsymbol{\xi}^{\mu''} = R_k + R_u\,. \end{aligned} \tag{13}$$

For fixed τ^ν and $\boldsymbol{\zeta}^\nu$, $\frac{1}{\sqrt{N}}\boldsymbol{B} \cdot \boldsymbol{\xi}^{\mu''}$ is a Gaussian random variable with mean 0 and variance $\langle \frac{1}{N}||\boldsymbol{B}||^2 \rangle$. This yields

$$\langle R_u \rangle = (1-\gamma)\sqrt{\frac{2}{\pi}\alpha_s}\,.$$

On the other hand $\frac{1}{\sqrt{N}}\boldsymbol{B} \cdot \boldsymbol{\zeta}^{\mu'}$ is for fixed τ^ν a Gaussian random variable with mean $\tau^{\mu'}/\sqrt{\alpha_t}$ and variance 1, giving

$$\langle R_k \rangle = \gamma\sqrt{\frac{\alpha_s}{\alpha_t}}\left[1 - 2H\left(\frac{1}{\sqrt{\alpha_t}}\right)\right] + \gamma\sqrt{\frac{2}{\pi}\alpha_s}\exp\left(-\frac{1}{2\alpha_t}\right)$$

where H is defined by $H(x) = \int_x^\infty \mathrm{d}t (1/\sqrt{2\pi}) \exp(-t^2/2)$.

So we get

$$\langle R\rangle = \sqrt{\frac{2}{\pi}\alpha_s}\left[1-\gamma+\gamma\exp\left(-\frac{1}{2\alpha_t}\right)\right]+\gamma\sqrt{\frac{\alpha_s}{\alpha_t}}\left[1-2H\left(\frac{1}{\sqrt{\alpha_t}}\right)\right]. \tag{14}$$

In order to calculate $\langle\frac{1}{N}\|\boldsymbol{J}\|^2\rangle$ we have to distinguish three cases of double sums since $\frac{1}{N}\|\boldsymbol{J}\|^2$ is quadratic in $\sum_{\mu=1}^{\alpha_t N}\operatorname{sign}(\frac{1}{\sqrt{N}}\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)\xi_j^\mu$. Averaging each term gives similarly

$$\begin{aligned}\left\langle\frac{1}{N}\|\boldsymbol{J}\|^2\right\rangle = 1+\alpha_s\Bigg\{&\frac{2}{\pi}(1-\gamma)^2+2\gamma(1-\gamma)\left[\sqrt{\frac{2}{\pi\alpha_t}}\left[1-2H\left(\frac{1}{\sqrt{\alpha_t}}\right)\right]\right.\\ &\left.+\frac{2}{\pi}\exp\left(-\frac{1}{2\alpha_t}\right)\right]+\gamma^2\left[2\sqrt{\frac{2}{\pi\alpha_t}}\left[1-2H\left(\frac{1}{\sqrt{\alpha_t}}\right)\right]\exp\left(-\frac{1}{2\alpha_t}\right)\right.\\ &\left.+\frac{2}{\pi}\exp\left(-\frac{1}{\alpha_t}\right)\right]\Bigg\}.\end{aligned} \tag{15}$$

For $\gamma=0$ or $\alpha_t\to\infty$ we find from (14) and (15)

$$\langle R\rangle=\sqrt{\frac{2}{\pi}\alpha_s}\qquad\left\langle\frac{1}{N}\|\boldsymbol{J}\|^2\right\rangle=1+\frac{2}{\pi}\alpha_s \tag{16}$$

i.e. the well known results for an uncorrelated teacher [4, 5].

Case B. Next we consider a Hebbian student generalizing a pseudo-inverse teacher. The student's couplings $\boldsymbol{J}$ are still given by (11). As in (13) we write $R=R_k+R_u$. The calculation of R_u is almost identical to the previous case and gives $\langle R_u\rangle=(1-\gamma)\sqrt{(2/\pi)\alpha_s}\langle\frac{1}{N}\|\boldsymbol{B}\|^2\rangle$. Restricting ourselves to $\alpha_t<1$ we have $\frac{1}{\sqrt{N}}\boldsymbol{B}\cdot\boldsymbol{\zeta}^{\mu'}=\tau^{\mu'}$ and hence $\langle R_k\rangle=\gamma\sqrt{\alpha_s}$. In order to calculate $\langle\frac{1}{N}\|\boldsymbol{B}\|^2\rangle$ and $\langle\frac{1}{N}\|\boldsymbol{J}\|^2\rangle$ we represent the pseudo-inverse rule by an integral over the space of couplings $\boldsymbol{B}$ [13]. In this way we get $\langle\frac{1}{N}\|\boldsymbol{B}\|^2\rangle=\langle\langle\frac{1}{N}\|\boldsymbol{B}\|^2\rangle_B\rangle$ and $\langle\frac{1}{N}\|\boldsymbol{J}\|^2\rangle=\langle\langle\frac{1}{N}\|\boldsymbol{J}\|^2\rangle_B\rangle$ with

$$\langle f(\boldsymbol{B})\rangle_B=\lim_{Q_t\to\min}\frac{\int\prod_{j=1}^N\mathrm{d}B_j\prod_{\nu=1}^{\alpha_t N}\delta(\tau^\nu-\frac{1}{\sqrt{N}}\boldsymbol{B}\cdot\boldsymbol{\zeta}^\nu)\delta(\|\boldsymbol{B}\|^2-NQ_t)f(\boldsymbol{B})}{\int\prod_{j=1}^N\mathrm{d}B_j\prod_{\nu=1}^{\alpha_t N}\delta(\tau^\nu-\frac{1}{\sqrt{N}}\boldsymbol{B}\cdot\boldsymbol{\zeta}^\nu)\delta(\|\boldsymbol{B}\|^2-NQ_t)}. \tag{17}$$

The limit of minimal norm $Q_t\to\min$ eliminates all components of $\boldsymbol{B}$ orthogonal to the subspace spanned by the patterns $\{\boldsymbol{\zeta}^\nu\}$. Using standard techniques ([13], see also appendix A) we get $\langle\frac{1}{N}\|\boldsymbol{B}\|^2\rangle=\alpha_t/1-\alpha_t$. A similar calculation can be performed to determine $\langle\frac{1}{N}\|\boldsymbol{J}\|^2\rangle$. We finally find with (12)

$$\rho=\frac{(1-\gamma)\sqrt{\frac{2}{\pi}\alpha_s}+\gamma\sqrt{\frac{\alpha_s}{\alpha_t}(1-\alpha_t)}}{\left(1+\frac{2}{\pi}(1-\gamma)^2\alpha_s+2\sqrt{\frac{2}{\pi}}\gamma(1-\gamma)\alpha_s\sqrt{\frac{1-\alpha_t}{\alpha_t}}\right)^{1/2}}\qquad(\alpha_t<1). \tag{18}$$

Again $\gamma=0$ gives the result (16) for an uncorrelated teacher.

Case C. In the complementary case of a pseudo-inverse student generalizing a Hebbian teacher the teacher's rule is explicitly given by (10). It is again convenient to represent the pseudo-inverse rule by an integral over the space of couplings $\boldsymbol{J}$. Let us first consider the case $\alpha_s < 1$. Similarly to (17) we then have to calculate $\langle \rho \rangle_J$ where

$$\langle g(\boldsymbol{J}) \rangle_J = \lim_{Q_s \to \min} \frac{\int \prod_{j=1}^{N} \mathrm{d}J_j \prod_{\mu=1}^{\alpha_s N} \delta(\eta^\mu - \frac{1}{\sqrt{N}} \boldsymbol{J} \cdot \boldsymbol{\xi}^\mu) \delta(\|\boldsymbol{J}\|^2 - N Q_s) g(\boldsymbol{J})}{\int \prod_{j=1}^{N} \mathrm{d}J_j \prod_{\mu=1}^{\alpha_s N} \delta(\eta^\mu - \frac{1}{\sqrt{N}} \boldsymbol{J} \cdot \boldsymbol{\xi}^\mu) \delta(\|\boldsymbol{J}\|^2 - N Q_s)} . \tag{19}$$

The pattern average can be performed using the replica trick. The resulting integrals are disentangled introducing order parameters and the remaining integrals are calculated by the saddle-point method. Some details of the calculation can be found in appendix A. Assuming replica symmetry ρ can be determined from the following saddle-point equations:

$$\begin{gathered} \rho = \frac{R}{\sqrt{Q_s}} \qquad R = \left(1 - \gamma \frac{\alpha_s}{\alpha_t}\right) M + \gamma \frac{\alpha_s}{\sqrt{\alpha_t}} \left[1 - 2H\left(\frac{1}{\sqrt{\alpha_t}}\right)\right] \\ Q_s = \frac{\alpha_s - \left(1 - \gamma \frac{\alpha_s}{\alpha_t}\right) M^2 - 2\gamma \frac{\alpha_s}{\sqrt{\alpha_t}} M \left[1 - 2H\left(\frac{1}{\sqrt{\alpha_t}}\right)\right]}{1 - \alpha_s} \end{gathered} \tag{20}$$

with

$$M = \sqrt{\frac{2}{\pi}} \alpha_s \left[(1 - \gamma) + \gamma \exp\left(-\frac{1}{2\alpha_t}\right)\right] .$$

Again $\gamma = 0$ or $\alpha_t \to \infty$ lead to the results for an uncorrelated teacher [5]

$$R = \sqrt{\frac{2}{\pi}} \alpha_s \qquad Q_s = \frac{\alpha_s - R^2}{1 - \alpha_s} \tag{21}$$

For $\alpha_s > 1$, $\langle g(\boldsymbol{J}) \rangle_J$ is not well defined by (19) and we have to use instead

$$\langle g(\boldsymbol{J}) \rangle_J = \lim_{\beta \to \infty} \frac{\int \prod_{j=1}^{N} \mathrm{d}J_j \exp\left[-\beta \sum_{\mu=1}^{\alpha_s N} (\eta^\mu - \frac{1}{\sqrt{N}} \boldsymbol{J} \cdot \boldsymbol{\xi}^\mu)^2 \right] g(\boldsymbol{J})}{\int \prod_{j=1}^{N} \mathrm{d}J_j \exp\left[-\beta \sum_{\mu=1}^{\alpha_s N} \left(\eta^\mu - \frac{1}{\sqrt{N}} \boldsymbol{J} \cdot \boldsymbol{\xi}^\mu\right)^2 \right]} . \tag{22}$$

This form takes into consideration the fact that for $\alpha_s > 1$ no vector $\boldsymbol{J}$ exists which makes the quadratic form in the exponent exactly zero. A calculation almost identical to $\alpha_s < 1$ now yields

$$\begin{gathered} R = \left(1 - \frac{\gamma}{\alpha_t}\right) M + \frac{\gamma}{\sqrt{\alpha_t}} \left[1 - 2H\left(\frac{1}{\sqrt{\alpha_t}}\right)\right] \\ Q_s = \frac{1 + 2(\alpha_s - 1) M R - (1 - \frac{\gamma}{\alpha_t}) M^2 - 2\frac{\gamma}{\sqrt{\alpha_t}} M \left[1 - 2H(\frac{1}{\sqrt{\alpha_t}})\right]}{\alpha_s - 1} \end{gathered} \tag{23}$$

with

$$M = \sqrt{\frac{2}{\pi}} \left[(1-\gamma) + \gamma \exp\left(-\frac{1}{2\alpha_t} \right) \right] .$$

As before we find the correct behaviour for $\gamma = 0$ and $\alpha_t \to \infty$ [5]:

$$R = \sqrt{\frac{2}{\pi}} \qquad Q_s = \frac{1 + R^2(\alpha_s - 2)}{\alpha_s - 1} . \tag{24}$$

Note that for $\alpha_s \to 1$ from above as well as from below Q_s diverges, implying $g(\alpha_s) = 0.5$.

Case D. In order to calculate the overlap ρ between teacher and student for the case in which both use the pseudo-inverse rule we write ρ as an integral over the coupling spaces of teacher *and* student:

$$\rho = \left\langle \left\langle \left\langle \frac{\boldsymbol{B} \cdot \boldsymbol{J}}{\|\boldsymbol{B}\| \|\boldsymbol{J}\|} \right\rangle_J \right\rangle_B \right\rangle_{\xi, \zeta, \tau} \tag{25}$$

where the averages are defined in (17), (19) and (22). In order to do the pattern average in (25) we introduce replica-indices $a = 1, \ldots, m$ for the B_j and $\alpha = 1, \ldots, n$ for the J_j. Note that the average over $\boldsymbol{J}$ in (19) depends via η^μ parametically on $\boldsymbol{B}$, hence the $\boldsymbol{J}$-average has to be done before the average over $\boldsymbol{B}$. Accordingly the limit $n \to 0$ has to be taken before the limit $m \to 0$. This will be important to get the correct saddle-point equations for the order parameters (see appendix B). These are, for $\alpha_s < 1$,

$$\begin{aligned} R &= \frac{bQ_t - \gamma\alpha_s b(1+Q_t) + \gamma\alpha_s}{1 - \gamma\alpha_s} \qquad & b &= \frac{(1-\gamma)\alpha_s \sqrt{\frac{2}{\pi Q_t}}}{1 - \gamma\alpha_s} \\ Q_s &= \frac{\alpha_s - b^2[Q_t - \gamma\alpha_s(1+Q_t)] - 2\gamma\alpha_s b}{1 - \alpha_s} \qquad & Q_t &= \frac{\alpha_t}{1 - \alpha_t} \end{aligned} \tag{26}$$

and for $\alpha_s > 1$

$$\begin{aligned} R &= \frac{\sqrt{\frac{2Q_t}{\pi}} + \gamma\left(1 - (1+Q_t)\sqrt{\frac{2}{\pi Q_t}}\right)}{1 - \gamma} \qquad Q_t = \frac{\alpha_t}{1 - \alpha_t} \\ Q_s &= \frac{1 - \frac{2}{\pi}\alpha_s + 2(\alpha_s - 1)R\sqrt{\frac{2}{\pi Q_t}} + \gamma\frac{2}{\pi}(1 + \frac{1}{Q_t}) - 2\gamma\sqrt{\frac{2}{\pi Q_t}}}{\alpha_s - 1} \end{aligned} \tag{27}$$

with

$$\rho = \frac{R}{\sqrt{Q_t Q_s}} .$$

For $\gamma = 0$ we get back the results for a pseudo-inverse rule generalizing an uncorrelated teacher (see (21) and (24)).

4. Results

Let us first consider the case where both teacher and student are designed by the Hebb rule. In figure 4 we plot the generalization ability g versus α_s for $\alpha_t = 0.1$ and different values of the parameter γ describing the overlap between the pattern sets of teacher and student. The curves show the expected behaviour. With increasing γ the student is more and more trained with the same patterns as the teacher and the generalization ability grows. Note, however, that even for $\gamma = 1$ the curve does not end at $g(\alpha_s = \alpha_t) = 1$ because the teacher made a small percentage of mistakes in learning his patterns [15]) so that the student does not, in all cases, get the answers the teacher was trained with.

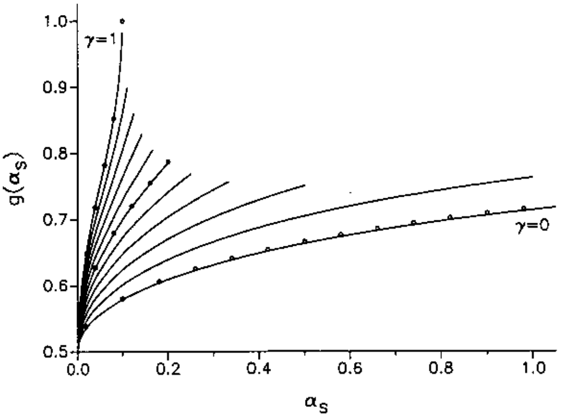

Figure 4. Generalization ability versus relative number of training examples for a Hebbian student generalizing a Hebbian teacher with $\alpha_t = 0.1$. The parameter γ has values $1, 0.9, \ldots, 0$ (from top to bottom). The circles are simulation results for system size $N = 50$ for $\gamma = 1, 0.5$ and 0 respectively averaged over typically 20 000 samples of random patterns.

It is somewhat surprising that the value of g at the endpoints of the curves decreases with increasing α_s. Consider, for example, the two curves for $\gamma = 1$ and $\gamma = 0.9$. The endpoint of the $\gamma = 0.9$ curve corresponds to a student who has asked $0.11N$ questions $0.1N$ of which were those the teacher had learnt himself. Since the succession of questions is irrelevant we can imagine that the student first asks these questions. His generalization ability is then given by the endpoint of the $\gamma = 1$ curve, i.e. $g \simeq 0.98$. Asking now the remaining $0.01N$ questions g decreases to $g \simeq 0.90$. Naively one would expect g to continue to increase since the additional questions yield additional information about the teacher to the student. Although this is true the Hebbian student is not able to use this information properly. In fact he overestimates it thereby reducing his similarity with the teacher. An extreme example is given by two identical perceptrons $\boldsymbol{B} = \boldsymbol{J}$, i.e. g exactly equal to 1. If the student nevertheless poses an additional question $\boldsymbol{\xi}^\mu$ he gets an answer η^μ and has to increment his couplings by $\Delta \boldsymbol{J} \sim \boldsymbol{\xi}^\mu \eta^\mu$. Therefore teacher and student now differ from each other and g will decrease from 1.

Figure 5 shows that the behaviour remains qualitatively the same for larger values of α_t. Still it is advantageous to the student to ask those questions the teacher was designed with although he will already get approximately 14% wrong answers for $\alpha_t = 1$. For $\alpha_t \to \infty$ all curves collapse to the $\gamma = 0$ curve identical to the Hebb-rule result of figure 2.

Consider now the case where a Hebbian student generalizes a teacher designed by the pseudo-inverse rule. As shown in figure 6 the behaviour is rather similar to

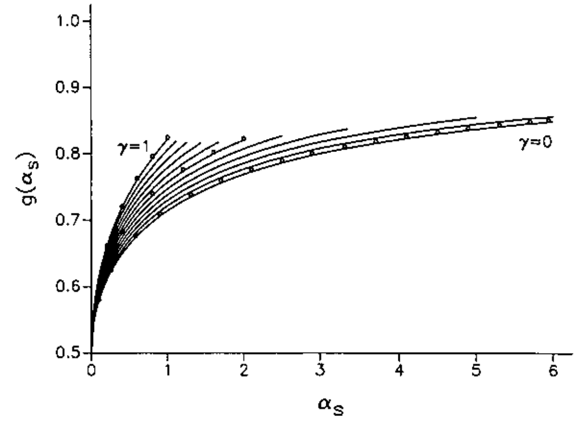

Figure 5. Same as figure 4 for $\alpha_t = 1$.

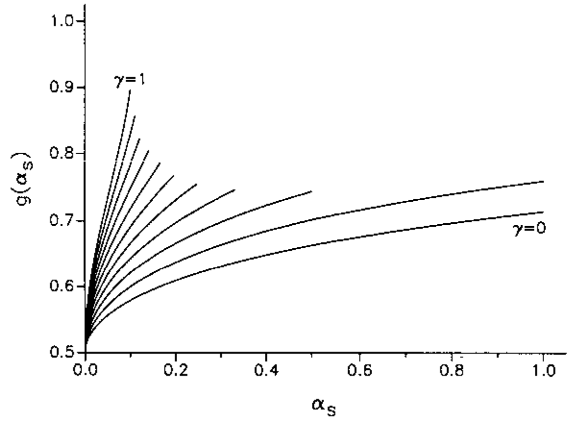

Figure 6. Same as figure 4 for a Hebbian student generalizing a teacher designed by the pseudo-inverse rule with $\alpha_t = 0.1$.

the case discussed above if α_t is small. This was to be expected since for small α_t the synaptic vectors $\boldsymbol{B}$ produced by the Hebb rule and by the pseudo-inverse rule are not too different. So again it is advantageous to the student to ask the teacher the known questions. The resulting values of g are always slightly below those of the previous case with the largest deviation occurring at the end point of the $\gamma = 1$ curve where now only the value $g \simeq 0.90$ is reached.

The situation is markedly differrent for larger values of α_t. Figure 7 shows the results for $\alpha_t = 0.9$. For a given value of α_s the generalization ability is now decreasing with increasing γ. Hence it is now *disadvantageous* to the student to ask those questions the teacher has learnt himself. The reason for this is that the synaptic vector $\boldsymbol{B}$ produced by the pseudo-inverse rule at sufficiently large values of α_t is rather different from the one resulting from the Hebb rule. The best strategy for the student to generalize is therefore now to ask questions the teacher has never heard of. Then the generalization ability is the same as for an uncorrelated teacher. Any overlap with the questions the teacher has learnt himself will reduce the generalization ability since teacher and student use rather different internal structures to store the required input–output relations.

The complementary case is given by a student using the pseudo-inverse rule to

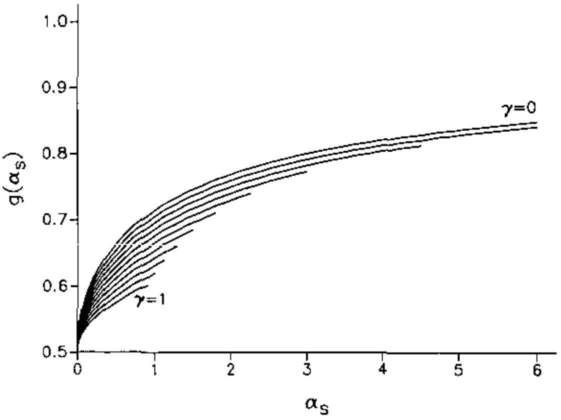

Figure 7. Same as figure 4 for a Hebbian student generalizing a teacher designed by the pseudo-inverse rule with $\alpha_t = 0.9$. The values of γ are $0, 0.1, \ldots, 1$ (from top to bottom).

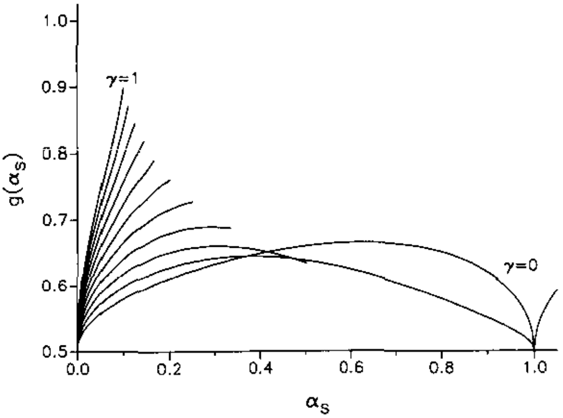

Figure 8. Same as figure 4 for a student using the pseudo-inverse rule to generalize a Hebbian teacher, $\gamma = 1, 0.9, \ldots, 0$ (left from top to bottom).

generalize a Hebbian teacher. As can be seen in figure 8 it is, for small α_s, again advantageous to the student to ask the teacher the known questions. However, with increasing values of α_s the student uses an internal structure becoming more and more different from the one of the teacher and accordingly becomes more confused than enlightened from the answers of the teacher to these questions. So if the student is allowed to pose only a rather limited number of questions he should choose the questions the teacher has learnt himself. If, however, he may ask many questions he should avoid these and try to make up his own mind. Note that the results of the last two cases are consistent with each other although they have been obtained by using rather different techniques.

Finally we discuss the case of two perceptrons using both the pseudo-inverse rules. In particular we are interested to see whether it is possible to compensate the decrease of g around $\alpha_s = 1$ by using a large enough value of γ. For small values of α_t we get results similar to figure 8 as expected. For $\alpha_t = 0.9$ the behaviour of $g(\alpha_s)$ is shown in figure 9. One clearly sees that the overfitting phenomenon persists and in particular that always $g(\alpha_s = 1) = 0.5$ as can also be seen analytically from (26) and (27). With increasing γ the maximum of $g(\alpha_s)$ is shifted towards larger values of α_s, so there is at least a tendency to reduce overfitting. For $\gamma = 1$ we have $\alpha_s \leqslant \alpha_t = 0.9$

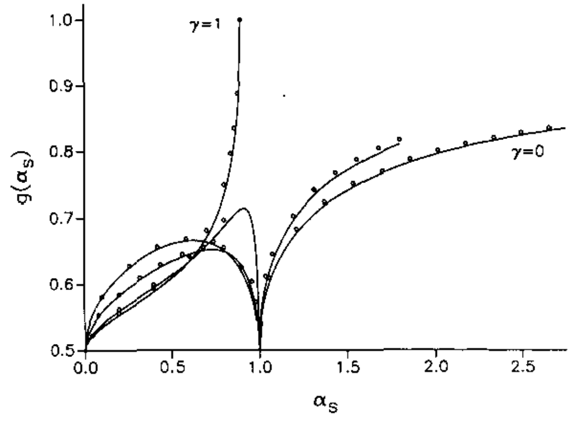

Figure 9. Same as figure 4 for both teacher and student using the pseudo-inverse rule and $\alpha_t = 0.9$. $\gamma = 0, 0.5, 0.9$ and 1.0 (left from top to bottom). Simulation results are for $N = 50$ averaged over typically 3000 samples of random patterns.

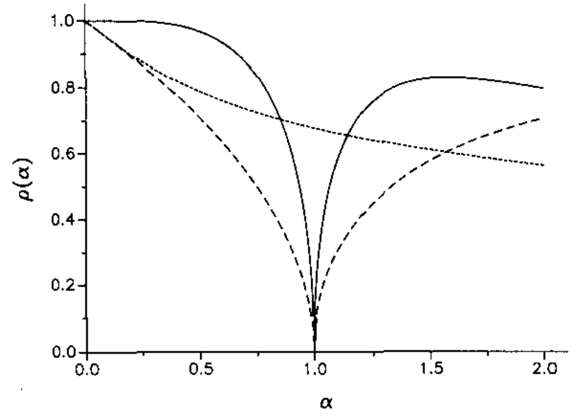

Figure 10. Normalized overlap ρ between two perceptrons using different learning rules for storing the same patterns as a function of the storage ratio α. Full curve: pseudo-inverse rule versus optimal perceptron; broken line: pseudo-inverse rule versus Hebb-rule; dotted curve: Hebb rule versus optimal perceptron.

and hence no overfitting occurs at all. Note that now it is disadvantageous for *small* values of α_s to ask questions the teacher has learnt himself and becomes increasingly helpful for larger values of α_s. This can be clearly seen from a comparison of the curves for $\gamma = 0$ and $\gamma = 1$.

We have also studied the cases where a teacher designed by the optimal perceptron rule is generalized by a Hebbian or a pseudo-inverse student as well as the complementary cases of a student using the optimal perceptron rule to generalize a Hebbian or a pseudo-inverse teacher [16]. The analytical calculations are rather similar to the cases where instead of the optimal perceptron rule the pseudo-inverse rule is used. As long as α is not too near to 1 the results are qualitatively the same (cf also figure 10; see below). If both teacher and student use the optimal perceptron rule the calculations can also be performed, although the resulting saddle-point equations become more complicated.

5. Summary

In the present paper we have analysed the generalization ability $g(\alpha_s)$ of a student perceptron $\boldsymbol{J}$ on the basis of input–output examples $\{\xi^\mu, \eta^\mu\}$ generated by a teacher perceptron B. Contrary to most previous studies which concentrated on the dependence of $g(\alpha_s)$ on the properties of the student alone we were interested in the influence of the structural similarity between teacher and student. Somewhat related problems that were studied in the literature concern the generalization ability for a binary teacher $B_j = \pm 1$ and a student with continuous synapses [7] as well as the dependence of $g(\alpha_s)$ on different continuous input–output characteristics of teacher and student [17].

In the present study we have assumed that the teacher perceptron is designed using some learning rule to store a given number of random input–output mappings $\{\zeta^\nu, \tau^\nu\}$. If now a fraction γ of the student's questions ξ^μ are identical to some of the patterns ζ^ν which the teacher was trained with himself, interesting correlations between the learning rules of teacher and student arise. As representative examples for learning rules we have used the Hebb rule, the pseudo-inverse rule and the perceptron rule of optimal stability. We have determined the dependence of the generalization ability $g(\alpha_s)$ on the fraction γ of identical patterns in the training sets of teacher and student for all possible combinations of the above-mentioned learning rules except for the case where both teacher and student use the perceptron rule.

Our main result is that overlapping pattern sets, i.e. $\gamma > 0$, can be both advantageous and disadvantageous for optimal generalization. They are advantegeous if teacher and student are structured similarly, as when they use the same learning rule and have similar storage ratios $\alpha_t \cong \alpha_s$. On the other hand they can affect generalization adversely if different rules are used and/or the storage ratios are markedly different. The reason for this is that using different learning rules for storing the *same* set of patterns may result in rather different synaptic vectors for teacher and student. This has also been discussed recently by Wong *et al* [18]. In figure 10 we plot the normalized overlap between teacher and student as a function of the storage ratio α for different combinations of learning rules storing exactly the same set of patterns. For small values of α the different synaptic vectors are still rather similar but become more and more different with increasing α. One also clearly sees that the pseudo-inverse rule produces for values of α around 1 a very special synaptic vector [14]. If now a Hebbian student tries to generalize a pseudo-inverse teacher the same patterns are stored internally in a rather different fashion and it is hence questionable whether overlapping pattern sets are helpful. In fact our results show that for not too small values of α_t the student should avoid questions out of the pattern set the teacher learnt himself. The generalization ability is higher if he adapts to the answers η^μ the teacher gives on questions ξ^μ that do not belong to his own training set. For given values of α_t and α_s, the storage ratios of teacher and student respectively, we have never found an optimal value of γ somewhere between the extreme cases $\gamma = 0$ and $\gamma = \gamma_{\max}$. That means that for getting a high generalization probability one should, depending on α_t and α_s, either avoid completely any correlation between the pattern sets of teacher and student or use them as much as possible.

Another interesting result concerns the case where both teacher and student use the same learning rule. Posing questions out of the teacher's training set the generalization ability increases rather quickly. If now some questions not contained in the teacher's pattern set are added the generalization ability can again decrease. This seems strange since any input–output pair produced by the teacher yields additional information about the teacher to the student. Nevertheless it is difficult for the student to use this information properly. As we have shown it is rather likely that he overestimates the information in the additional examples, which results in an effective synaptic noise that reduces the generalization ability.

Finally we were interested to see whether the very peculiar phenomenon of overfitting of the pseudo-inverse rule could be compensated by using a teacher also designed by this rule. The result is negative. Although it is possible to narrow the interval of α_s-values where $g(\alpha_s)$ is small by using large values of γ we find always $g(\alpha_s = 1) = 0.5$. The reason for this is the singular norm of the synaptic vector produced by the pseudo-inverse rule for $\alpha \to 1$ (cf also figure 10).

Acknowledgments

We would like to thank Annette Zippelius and Reiner Kree for helpful discussions.

Appendix A

In this appendix we perform the replica calculation for a pseudo-inverse student generalizing a Hebbian teacher. For $\alpha_s < 1$ we start with (19) which, introducing replicas, can be written as

$$\rho = \lim_{Q_s \to \min} \lim_{n \to 0} \left\langle \int \prod_{\alpha=1}^{n} \prod_{j=1}^{N} \mathrm{d}J_j^\alpha \prod_{\mu=1}^{\alpha_s N} \prod_{\alpha=1}^{n} \delta\left(\mathrm{sign}(\boldsymbol{B} \cdot \boldsymbol{\xi}^\mu) - \frac{1}{\sqrt{N}} \boldsymbol{J}^\alpha \cdot \boldsymbol{\xi}^\mu \right) \right.$$
$$\left. \times \prod_{\alpha=1}^{n} \delta(\|\boldsymbol{J}^\alpha\|^2 - NQ_s) \frac{\boldsymbol{B} \cdot \boldsymbol{J}^1}{\|\boldsymbol{B}\| \sqrt{NQ_s}} \right\rangle . \tag{A1}$$

In order to perform the average over $\tau^\nu, \zeta_j^\nu, \xi_j^\mu$ we introduce

$$B_j = \frac{1}{\sqrt{\alpha_t N}} \sum_{\nu=1}^{\alpha_t N} \tau^\nu \zeta_j^\nu \qquad (j = 1, \ldots, N)$$
$$u^\mu = \frac{1}{\sqrt{N}} \boldsymbol{B} \cdot \boldsymbol{\xi}^\mu \qquad (\mu = 1, \ldots, \alpha_t N)$$

via δ-functions and their Fourier representations

$$\rho = \lim_{Q_s, n} \int \prod_{\alpha j} \mathrm{d}J_j^\alpha \int \prod_\alpha \frac{\mathrm{d}e^\alpha}{4\pi} \exp\left(\frac{\mathrm{i}}{2} NQ_s \sum_\alpha e^\alpha - \frac{\mathrm{i}}{2} \sum_\alpha e^\alpha \sum_j (J_j^\alpha)^2 \right)$$
$$\times \int \prod_j \frac{\mathrm{d}B_j \mathrm{d}A_j}{2\pi} \exp\left(\mathrm{i} \sum_j A_j B_j \right)$$
$$\times \int \prod_{\alpha\mu} \frac{\mathrm{d}x^{\mu\alpha}}{2\pi} \int \prod_\mu \frac{\mathrm{d}u^\mu \mathrm{d}s^\mu}{2\pi} \exp\left(\mathrm{i} \sum_{\alpha\mu} x^{\mu\alpha} \mathrm{sign}(u^\mu) + \mathrm{i} \sum_\mu u^\mu s^\mu \right)$$
$$\times \left\langle \exp\left(- \frac{\mathrm{i}}{\sqrt{N}} \sum_{\alpha\mu j} x^{\mu\alpha} J_j^\alpha \xi_j^\mu - \frac{\mathrm{i}}{\sqrt{N}} \sum_{\mu j} s^\mu B_j \xi_j^\mu \right.\right.$$
$$\left.\left. - \frac{\mathrm{i}}{\sqrt{\alpha_t N}} \sum_{\nu j} A_j \tau^\nu \zeta_j^\nu \right) \right\rangle \frac{\boldsymbol{B} \cdot \boldsymbol{J}^1}{\|\boldsymbol{B}\| \sqrt{NQ_s}} . \tag{A2}$$

We now average over ζ_j^ν and ξ_j^μ and introduce the order parameters

$$\begin{aligned} q_s^{\alpha\beta} &:= \frac{1}{N}\sum_j J_j^\alpha J_j^\beta \qquad (\alpha,\beta = 1,\ldots,n \quad \alpha \neq \beta) \\ R^\alpha &:= \frac{1}{N}\sum_j B_j J_j^\alpha \qquad (\alpha = 1,\ldots,n) \\ M^\alpha &:= \frac{1}{N}\sum_j A_j J_j^\alpha \qquad (\alpha = 1,\ldots,n) \\ y &:= \frac{1}{N}\sum_j A_j B_j\,. \end{aligned} \tag{A3}$$

We then obtain

$$\begin{aligned} \rho = \lim_{Q_s,n} &\int \frac{\mathrm{d}y\mathrm{d}z}{2\pi/N}\int\prod_{\alpha<\beta}\frac{\mathrm{d}q_s^{\alpha\beta}\mathrm{d}f^{\alpha\beta}}{2\pi/N}\int\prod_\alpha \frac{\mathrm{d}e^\alpha}{4\pi}\frac{\mathrm{d}R^\alpha\mathrm{d}G^\alpha}{2\pi/N}\frac{\mathrm{d}M^\alpha\mathrm{d}K^\alpha}{2\pi/N} \\ &\times \frac{R^1}{\sqrt{Q_s}}\exp\Bigg[N\Bigg(-\mathrm{i}yz + \frac{\mathrm{i}}{2}Q_s\sum_\alpha e^\alpha - \mathrm{i}\sum_{\alpha<\beta} q_s^{\alpha\beta} f^{\alpha\beta} \\ &- \mathrm{i}\sum_\alpha R^\alpha G^\alpha - \mathrm{i}\sum_\alpha M^\alpha K^\alpha + G_2(\boldsymbol{e},\boldsymbol{f},\boldsymbol{G},\boldsymbol{K},z) \\ &+ (1-\gamma)\alpha_s G_1(\boldsymbol{q}_s,\boldsymbol{R}) + \gamma\alpha_s G_1^*(\boldsymbol{q}_s,\boldsymbol{R},\boldsymbol{M},y)\Bigg)\Bigg] \end{aligned} \tag{A4}$$

with

$$\begin{aligned} G_2 = \log &\int\frac{\mathrm{d}A\mathrm{d}B}{2\pi}\int\prod_\alpha \mathrm{d}J^\alpha \exp\Bigg(\mathrm{i}(1+z)AB - \frac{A^2}{2} - \frac{\mathrm{i}}{2}\sum_\alpha e^\alpha (J^\alpha)^2 \\ &+ \mathrm{i}\sum_{\alpha<\beta} f^{\alpha\beta}J^\alpha J^\beta + \mathrm{i}B\sum_\alpha G^\alpha J^\alpha + \mathrm{i}A\sum_\alpha K^\alpha J^\alpha\Bigg) \\ G_1 = \log &\int\prod_\alpha\frac{\mathrm{d}x^\alpha}{2\pi}\int\frac{\mathrm{d}u\mathrm{d}s}{2\pi}\exp\Bigg(\mathrm{i}us - \frac{s^2}{2} - s\sum_\alpha R^\alpha x^\alpha + \mathrm{i}\,\mathrm{sign}(u)\sum_\alpha x^\alpha \\ &- \frac{Q_s}{2}\sum_\alpha (x^\alpha)^2 + \frac{1}{2}\sum_{\alpha\neq\beta} q_s^{\alpha\beta}x^\alpha x^\beta\Bigg) \\ G_1^* = \log &\int\prod_\alpha\frac{\mathrm{d}x^\alpha}{2\pi}\int\frac{\mathrm{d}u\mathrm{d}s}{2\pi}\exp\Bigg(\mathrm{i}us - \frac{s^2}{2} - s\sum_\alpha R^\alpha x^\alpha + \mathrm{i}\,\mathrm{sign}(u)\sum_\alpha x^\alpha \\ &- \frac{Q_s}{2}\sum_\alpha (x^\alpha)^2 + \frac{1}{2}\sum_{\alpha\neq\beta} q_s^{\alpha\beta}x^\alpha x^\beta - \frac{1}{\sqrt{\alpha_t}}\sum_\alpha M^\alpha x^\alpha - \frac{y}{\sqrt{\alpha_t}}s\Bigg)\,. \end{aligned} \tag{A5}$$

We solve the integral by the saddle-point method and use a replica-symmetric ansatz for the order parameters. This gives

$$\rho = \lim_{Q_s,n} \frac{R}{\sqrt{Q_s}} \exp\Bigg\{ N\frac{n}{2}\,\mathrm{extr}_{y,z,e,f,q_s,R,G,M,K}\Bigg[\frac{2}{n}[-\mathrm{i}yz - \log(1+z)]$$

$$+\mathrm{i}Q_s e + \mathrm{i}q_s f - 2\mathrm{i}RG - 2\mathrm{i}MK + \log\left(\frac{2\pi}{\mathrm{i}(e+f)}\right) + \frac{f - 2\frac{KG}{1+z} + \frac{\mathrm{i}G^2}{(1+z)^2}}{e+f}$$

$$-\alpha_s\frac{1+Q_s}{Q_s - q_s} - \alpha_s\log[2\pi(Q_s - q_s)] + 2(1-\gamma)\alpha_s\sqrt{\frac{2}{\pi}}\frac{R}{Q_s - q_s}$$

$$-\frac{\gamma\alpha_s}{Q_s - q_s}\left(-2\sqrt{\frac{2}{\pi}}R\exp\left(\frac{y^2}{2\alpha_t}\right) + \frac{1}{\alpha_t}(\mathrm{i}M)^2\right.$$

$$\left.+\frac{2}{\sqrt{\alpha_t}}\mathrm{i}M\left[1 - 2H\left(\frac{-\mathrm{i}y}{\sqrt{\alpha_t}}\right)\right]\right)\Bigg]\Bigg\}. \tag{A6}$$

Since all saddle-point equations are algebraic we can solve them analytically. Note that the derivative $\partial\{\cdots\}/\partial y = 0$ yields $z = O(n)$ and therefore $\log(1+z) = z + O(n^2)$, $[1/1+z] = 1 - z + O(n^2)$ and $[1/(1+z)]^2 = 1 - 2z + O(n^2)$. Taking finally the limit of minimal norm in the form $q_s \to Q_s$ we find (20). For $\alpha_s > 1$ the calculation is almost identical except for Q_s becoming a saddle-point variable. The saddle-point equations result in (23).

Appendix B

In this appendix we perform the replica calculation for a pseudo-inverse student generalizing a pseudo-inverse teacher for $\alpha_{s,t} < 1$. We start from (25):

$$\rho = \lim_{Q_t\to\min}\lim_{Q_s\to\min}\lim_{m\to 0}\lim_{n\to 0}\Bigg\langle \int\prod_{a=1}^{m}\prod_{j=1}^{N}\mathrm{d}B_j^a \int\prod_{\alpha=1}^{n}\prod_{j=1}^{N}\mathrm{d}J_j^\alpha$$

$$\times\prod_{\nu=1}^{\alpha_t N}\prod_{a=1}^{m}\delta\left(\tau^\nu - \frac{1}{\sqrt{N}}\boldsymbol{B}^a\cdot\boldsymbol{\zeta}^\nu\right)\prod_{a=1}^{m}\delta(\|\boldsymbol{B}^a\|^2 - NQ_t)$$

$$\times\prod_{\mu=1}^{\alpha_s N}\prod_{\alpha=1}^{n}\delta\left(\mathrm{sign}(\boldsymbol{B}^1\cdot\boldsymbol{\xi}^\mu) - \frac{1}{\sqrt{N}}\boldsymbol{J}^\alpha\cdot\boldsymbol{\xi}^\mu\right)\prod_{\alpha=1}^{n}\delta(\|\boldsymbol{J}^\alpha\|^2 - NQ_s)$$

$$\times\frac{\boldsymbol{B}^1\cdot\boldsymbol{J}^1}{N\sqrt{Q_tQ_s}}\Bigg\rangle. \tag{B1}$$

We introduce $u^{\mu''} := \frac{1}{\sqrt{N}}\boldsymbol{B}^1\cdot\boldsymbol{\xi}^{\mu''}$ for the unknown examples ($\mu'' = \gamma\alpha_s N + 1,\ldots,\alpha_s N$) only since $\mathrm{sign}(\boldsymbol{B}\cdot\boldsymbol{\xi}^{\mu'}) = \tau^{\mu'}$ for the known examples due to $\alpha_t < 1$ ($\boldsymbol{\xi}^{\mu'} = \boldsymbol{\zeta}^{\mu'} : \mu' = 1,\ldots,\gamma\alpha_s N$). Using Fourier representations of the

δ-functions we get

$$
\begin{aligned}
\rho = \lim_{Q_t, Q_s, m, n} & \int \prod_{aj} \mathrm{d}B_j^a \int \prod_{\alpha j} \mathrm{d}J_j^\alpha \int \prod_a \frac{\mathrm{d}E^a}{4\pi} \int \prod_\alpha \frac{\mathrm{d}e^\alpha}{4\pi} \\
& \times \exp\Big(\frac{\mathrm{i}}{2} N Q_t \sum_a E^a - \frac{\mathrm{i}}{2} \sum_a E^a \sum_j (B_j^a)^2 \\
& + \frac{\mathrm{i}}{2} N Q_s \sum_\alpha e^\alpha - \frac{\mathrm{i}}{2} \sum_\alpha e^\alpha \sum_j (J_j^\alpha)^2 \Big) \\
& \times \int \prod_{a\nu} \frac{\mathrm{d}X^{\nu a}}{2\pi} \exp\Big(\mathrm{i} \sum_{a\nu} X^{\nu a} \tau^\nu \Big) \int \prod_{\alpha\mu'} \frac{\mathrm{d}x^{\mu'\alpha}}{2\pi} \exp\Big(\mathrm{i} \sum_{\alpha\mu'} x^{\mu'\alpha} \tau^{\mu'} \Big) \\
& \times \int \prod_{\alpha\mu''} \frac{\mathrm{d}x^{\mu''\alpha}}{2\pi} \int \prod_{\mu''} \frac{\mathrm{d}u^{\mu''} \mathrm{d}s^{\mu''}}{2\pi} \\
& \times \exp\Big(\mathrm{i} \sum_{\alpha\mu''} x^{\mu''\alpha} \mathrm{sign}(u^{\mu''}) + \mathrm{i} \sum_{\mu''} u^{\mu''} s^{\mu''} \Big) \\
& \times \Big\langle \exp\Big(-\frac{\mathrm{i}}{\sqrt{N}} \sum_{a\nu j} X^{\nu a} B_j^a \zeta_j^\nu - \frac{\mathrm{i}}{\sqrt{N}} \sum_{\alpha\mu' j} x^{\mu'\alpha} J_j^\alpha \zeta_j^{\mu'} \\
& - \frac{\mathrm{i}}{\sqrt{N}} \sum_{\alpha\mu'' j} x^{\mu''\alpha} J_j^\alpha \xi_j^{\mu''} - \frac{\mathrm{i}}{\sqrt{N}} \sum_{\mu'' j} s^{\mu''} B_j^1 \xi_j^{\mu''} \Big) \Big\rangle \frac{\boldsymbol{B}^1 \cdot \boldsymbol{J}^1}{N \sqrt{Q_t Q_s}} .
\end{aligned}
\tag{B2}
$$

We introduce the order parameters

$$
\begin{aligned}
q_t^{ab} &:= \frac{1}{N} \sum_j B_j^a B_j^b \qquad (a, b = 1, \ldots, m \quad a \neq b) \\
q_s^{\alpha\beta} &:= \frac{1}{N} \sum_j J_j^\alpha J_j^\beta \qquad (\alpha, \beta = 1, \ldots, n \quad \alpha \neq \beta) \\
R^{a\alpha} &:= \frac{1}{N} \sum_j B_j^a J_j^\alpha \qquad (a = 1, \ldots, m \quad \alpha = 1, \ldots, n) .
\end{aligned}
\tag{B3}
$$

This yields

$$
\begin{aligned}
\rho = \lim_{Q_t, Q_s, m, n} & \int \prod_a \frac{\mathrm{d}E^a}{4\pi} \int \prod_\alpha \frac{\mathrm{d}e^\alpha}{4\pi} \int \prod_{a<b} \frac{\mathrm{d}q_t^{ab} \mathrm{d}F^{ab}}{2\pi/N} \int \prod_{\alpha<\beta} \frac{\mathrm{d}q_s^{\alpha\beta} \mathrm{d}f^{\alpha\beta}}{2\pi/N} \\
& \times \int \prod_{a\alpha} \frac{\mathrm{d}R^{a\alpha} \mathrm{d}G^{a\alpha}}{2\pi/N} \frac{R^{11}}{\sqrt{Q_t Q_s}} \exp\Big[N \Big(\frac{\mathrm{i}}{2} Q_t \sum_a E^a + \frac{\mathrm{i}}{2} Q_s \sum_\alpha e^\alpha \\
& - \mathrm{i} \sum_{a<b} q_t^{ab} F^{ab} - \mathrm{i} \sum_{\alpha<\beta} q_s^{\alpha\beta} f^{\alpha\beta} - \mathrm{i} \sum_{a\alpha} R^{a\alpha} G^{a\alpha} + G_2(\boldsymbol{E}, \boldsymbol{F} \boldsymbol{e}, \boldsymbol{f}, \boldsymbol{G}) \\
& + (\alpha_t - \gamma\alpha_s) G_0(\boldsymbol{q}_t) + (1-\gamma)\alpha_s G_1(\boldsymbol{q}_s, \boldsymbol{R}) + \gamma\alpha_s G_1^*(\boldsymbol{q}_t \boldsymbol{q}_s, \boldsymbol{R}) \Big) \Big]
\end{aligned}
\tag{B4}
$$

with

$$\begin{aligned}
G_2 &= \log \int \prod_a \mathrm{d}B^a \int \prod_\alpha \mathrm{d}J^\alpha \exp\Big(-\frac{\mathrm{i}}{2}\sum_a E^a (B^a)^2 - \frac{\mathrm{i}}{2}\sum_\alpha \mathrm{e}^\alpha (J^\alpha)^2 \\
&\qquad + \mathrm{i}\sum_{a<b} F^{ab} B^a B^b + \mathrm{i}\sum_{\alpha<\beta} f^{\alpha\beta} J^\alpha J^\beta + \mathrm{i}\sum_{a\alpha} G^{a\alpha} B^a J^\alpha \Big) \\
G_0 &= \log \int \prod_a \frac{\mathrm{d}X^a}{2\pi} \exp\Big(\mathrm{i}\sum_a X^a - \frac{Q_t}{2}\sum_a (X^a)^2 - \frac{1}{2}\sum_{a\neq b} q_t^{ab} X^a X^b \Big) \\
G_1 &= \log \int \prod_\alpha \frac{\mathrm{d}x^\alpha}{2\pi} \int \frac{\mathrm{d}u\mathrm{d}s}{2\pi} \exp\Big(\mathrm{i}us - \frac{Q_t}{2}s^2 - s\sum_\alpha R^{1\alpha} x^\alpha \\
&\qquad + \mathrm{i}\,\mathrm{sign}(u)\sum_\alpha x^\alpha - \frac{Q_s}{2}\sum_\alpha (x^\alpha)^2 - \frac{1}{2}\sum_{\alpha\neq\beta} q_s^{\alpha\beta} x^\alpha x^\beta \Big) \qquad \text{(B5)} \\
G_1^* &= \log \int \prod_a \frac{\mathrm{d}X^a}{2\pi} \int \prod_\alpha \frac{\mathrm{d}x^\alpha}{2\pi} \exp\Big(\mathrm{i}\sum_a X^a - \frac{Q_t}{2}\sum_a (X^a)^2 - \frac{1}{2}\sum_{a\neq b} q_t^{ab} X^a X^b \\
&\qquad + \mathrm{i}\sum_\alpha x^\alpha - \frac{Q_s}{2}\sum_\alpha (x^\alpha)^2 - \frac{1}{2}\sum_{\alpha\neq\beta} q_s^{\alpha\beta} x^\alpha x^\beta - \sum_{a\alpha} R^{a\alpha} X^a x^\alpha \Big) .
\end{aligned}$$

The order parameter integrals in (B4) can be solved by the saddle-point method. Since $\boldsymbol{B}^1$ plays a special role we use the modified replica-symmetric ansatz

$$R^{a\alpha} = \begin{cases} R_0 & (a=1) \\ R_1 & (a\neq 1) \end{cases} \qquad G^{a\alpha} = \begin{cases} G_0 & (a=1) \\ G_1 & (a\neq 1) . \end{cases} \qquad \text{(B6)}$$

We then find

$$\begin{aligned}
\rho = \lim_{Q_t, Q_s, m, n} \frac{R_0}{\sqrt{Q_t Q_s}} \exp\Big(\frac{N}{2} \mathrm{extr}\,[m f_t(E, F, q_t) \\
+ n f_s(e, f, q_s, R_0, R_1, G_0, G_1; E, F, q_t) + \mathrm{O}(m^2, n^2, mn)] \Big) \qquad \text{(B7)}
\end{aligned}$$

with

$$\begin{aligned}
f_t &= \mathrm{i}Q_t E + \mathrm{i}q_t F + \log\frac{2\pi}{\mathrm{i}(E+F)} + \frac{F}{E+F} - \alpha_t \log 2\pi (Q_t - q_t) - \alpha_t \frac{1+q_t}{Q_t - q_t} \\
f_s &= \mathrm{i}Q_s e + \mathrm{i}q_s f + \log\frac{2\pi}{\mathrm{i}(e+f)} + \frac{f}{e+f} - \alpha_s \log 2\pi (Q_s - q_s) - \alpha_s \frac{1+q_s}{Q_s - q_s} \\
&\quad - 2\mathrm{i}R_0 G_0 + 2\mathrm{i}R_1 G_1 + \Big(\frac{1}{E+F} + \frac{F}{(E+F)^2} \Big) \frac{(G_0 - G_1)^2}{e+f} \qquad \text{(B8)} \\
&\quad + 2\frac{G_1(G_0 - G_1)}{(E+F)(e+f)} - \gamma\alpha_s \frac{(1+Q_t)(R_0 - R_1)^2}{(Q_s - q_s)(Q_t - q_t)^2} \\
&\quad + 2\gamma\alpha_s \frac{(1+R_0)(R_0 - R_1)}{(Q_s - q_s)(Q_t - q_t)} + 2(1-\gamma)\alpha_s \sqrt{\frac{2}{\pi Q_t}} \frac{R_0}{Q_s - q_s} .
\end{aligned}$$

Now we first have to solve the saddle-point equations for the order parameters $e, f, q_s, R_0, R_1 G_0, G_1$ from $\mathrm{d} f_s = 0$. This yields

$$
\begin{aligned}
&\mathrm{i}(e+f) = \frac{1}{Q_s - q_s} \\
&\mathrm{i}f = \frac{q_s}{(Q_s - q_s)^2} - \frac{\mathrm{i}(E+2F)}{(E+F)^2}(G_0 - G_1)^2 - 2\frac{\mathrm{i}G_1(G_0 - G_1)}{E+F} \\
&\mathrm{i}(G_0 - G_1) = (R_0 - R_1)\frac{\mathrm{i}(E+F)}{Q_s - q_s} \\
&\mathrm{i}G_1 = R_0\frac{\mathrm{i}(E+F)}{(Q_s - q_s)} - (R_0 - R_1)\frac{\mathrm{i}(E+2F)}{Q_s - q_s} \\
&\mathrm{i}(G_0 - G_1) = \frac{\gamma\alpha_s(R_0 - R_1)}{(Q_t - q_t)(Q_s - q_s)} + \frac{(1-\gamma)\alpha_s}{Q_s - q_s}\sqrt{\frac{2}{\pi Q_t}} \\
&\mathrm{i}G_1 = \gamma\alpha_s\frac{(1+q_t)(R_0 - R_1)}{(Q_t - q_t)^2(Q_s - q_s)} - \gamma\alpha_s\frac{1+R_1}{(Q_t - q_t)(Q_s - q_s)} \\
&\mathrm{i}f(Q_s - q_s)^2 = \alpha_s(1+q_s) + \gamma\alpha_s\frac{(1+Q_t)(R_0 - R_1)^2}{(Q_t - q_t)^2} - 2\gamma\alpha_s\frac{(1+R_0)(R_0 - R_1)}{Q_t - q_t}\,.
\end{aligned}
\tag{B9}
$$

Hence the saddle-point values of these order parameters still depend on E, F and Q_t. Next we take the limit $n \to 0$ and determine the saddle-point values of E, F and Q_t from the $\mathrm{O}(m)$-terms alone. This procedure corresponds to the appropriate order of limits (first $n \to 0$, then $m \to 0$) and reflects the fact that the teacher is not influenced by the student but vice versa. Accordingly we get for E, F and Q_t the standard saddle-point equations for the pseudo-inverse rule

$$
\mathrm{i}(E+F) = \frac{1}{Q_t - q_t} \qquad \mathrm{i}F = \frac{q_t}{(Q_t - q_t)^2} \qquad q_t = \frac{\alpha_t}{1-\alpha_t}\,. \tag{B10}
$$

Plugging this into (B9) we get (26). Note that because of $R_0 - R_1 \sim Q_t - q_t$ we have $R_0 - R_1 \to 0$ for $q_t \to Q_t$. Taking, however, $R_0 = R_1 = R$ already in the ansatz (B6) makes f_s independent of R and hence gives no equation for R.

In the case of $\alpha_s > 1$ the calculation is almost the same. The most important difference is that $Q_s \longrightarrow Q_s^\alpha := \frac{1}{N}\sum_j (J_j^\alpha)^2$ becomes a saddle-point variable, the saddle-point equation of which results from the limit $\beta \to \infty$. We finally get (27).

References

[1] Györgyi G and Tishby N 1990 *Neural Networks and Spin Glasses* ed W K Theumann and R Köberle (Singapore: World Scietific)

[2] Watkin T, Rau A and Biehl M Statistical mechanics of learning a rule *Rev. Mod. Phys*. in press

[3] E Gardner and B Derrida 1988 *J. Phys. A: Math. Gen.* **21** 257

[4] Vallet F 1989 *Europhys. Lett.* **8** 747

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

[6] Györgyi G 1990 *Phys. Rev. Lett.* **64** 2957

[7] 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** 4056

[8] Opper M and Haussler D 1991 *Phys. Rev. Lett.* **66** 2677

[9] Kinzel W and Rujan P 1990 *Europhys. Lett.* 13 473

[10] Watkin T L H and Rau A 1992 *J. Phys. A: Math. Gen.* **25** 113

[11] Schwarze H, Opper M and Kinzel W 1991 Generalization in a two-layer neural network *Preprint* University of Giessen

[12] Schwarze H, Hertz J A 1992 Generalization in a large committy-machine, *Nordita Preprint*

[13] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257

[14] Vallet F, Cailton J-G and Ph Refregier 1989 *Europhys. Lett.* **9** 315

[15] Amit D J, Gutfreund H and Sompolinsky H 1987 *Ann. Phys., NY* **173** 30

[16] Reimers L 1992 *Diploma-Thesis* University of Göttingen

[17] Bös S, Kinzel W and Opper M 1992 The generalization ability of perceptrons with continuous outputs *Preprint* University of Giessen

[18] Wong K Y M and Rau A D 1992 *Europhys. Lett.* **19** 559