Generalization ability of perceptrons with continuous outputs

S. Bös, W. Kinzel, and M. Opper
Institut für Theoretische Physik, Justus-Liebig Universität, Heinrich Buff Ring 16, D-6300 Giessen, Germany

(Received 17 April 1992)

In this paper we examine the influence of different input-output relations on the generalization ability of a single-layer perceptron. The input-output relations can be linear, binary, or sigmoid. With this choice we take into account most of the cases which are of present interest. The generalization problem will be realizable or unrealizable if the input-output relations for teacher and student are identical or not. We show that sometimes it can have a positive effect on the generalization ability, if one learns with errors.

PACS number(s): 87.10.+e, 02.50.+s, 05.20.−y

I. INTRODUCTION

Generalization is a characteristic ability of feedforward networks. The rule, which connects an input to an output, can be learned by examples obeying the rule.

The process of generalization can be divided into two phases. During the training phase the network learns a set of input-output pairs $\{(\xi^\mu,\zeta^\mu);\mu=1,\ldots,p\}$. Usually its parameters (weights, thresholds, etc.) are adjusted in such a way that the distance between the desired output ζ_*^μ and the actual output ζ^μ is minimized. The learning error is the average of this distance over all presented examples ξ^μ,

$$E_L=\frac{1}{p}\sum_{\mu=1}^{p}\mathrm{dist}(\zeta_*^\mu,\zeta^\mu)\ . \tag{1}$$

The examples are presented repeatedly until the learning error fulfills a termination condition.

In the test or generalization phase one wants to know how well the rule has been learned. Therefore one compares the outputs of the network ζ and the correct outputs ζ_* for random inputs $\mathbf{S}$. The generalization error averages the distance between these outputs over the distribution of the inputs $\mathbf{S}$,

$$E_G=\langle \mathrm{dist}(\zeta_*(\mathbf{S}),\zeta(\mathbf{S}))\rangle_{\{\mathbf{S}\}}\ . \tag{2}$$

For theoretical purposes it is useful to represent the rule by a second network, which we shall call the teacher network. It yields per definition the correct outputs ζ_*. Its parameters are the suitable variables to describe the rule. An interesting consequence is the fact that a student network may not be able to learn the rule perfectly, if there are certain differences in the architecture of teacher and student. This leads to a main division of the rules in realizable and unrealizable rules.

Many aspects of generalization have been examined, such as the effect of binary and continuous weights, Boolean and continuous outputs, and different learning rules [1–3]. Intelligent students have been constructed, which choose the training examples in order to get the maximal information about the rule [4–6].

In the present paper we will confine ourselves to single-layer perceptrons [7] with N input nodes and one output node. The inputs shall be binary ($S_i=\pm1, i=1,\ldots,N$) but the outputs ζ can accept continuous values determined by the input-output (IO) relation g. Continuous outputs show their importance in multilayer networks, because many important learning rules are based on gradient descent methods (e.g., error backpropagation [8]). In this paper we will be interested in the influence of the input-output relations on the generalization ability of the perceptron.

	Teacher	Student
weights ($j=1,\ldots,N$)	W_j^*	W_j
local fields	$h_*=\frac{1}{\sqrt{N}}\mathbf{W}^*\cdot\mathbf{S}$	$h=\frac{1}{\sqrt{N}}\mathbf{W}\cdot\mathbf{S}$
output	$\zeta_*=g_*(h_*)$	$\zeta=g(h)$

The input-output relations g_* and g do not need to be identical. In the following they will be chosen independently out of the set $\mathcal{S}_{\mathrm{IO}}$ of input-output relations,

$$\mathcal{S}_{\mathrm{IO}}=\{\gamma x,\mathrm{sgn}(x),\tanh(\gamma x),f_{\mathrm{PL}}(\gamma x)\}\ , \tag{3}$$

where $f_{\mathrm{PL}}(\gamma x)$ is the piecewise linear function, $f_{\mathrm{PL}}(\gamma x)\equiv\min(|\gamma x|,1)\mathrm{sgn}(x)$, and γ is the gain factor of the functions. We selected for the set $\mathcal{S}_{\mathrm{IO}}$ some representative examples, the linear IO as the simplest, the sign function for binary outputs, and two sigmoid functions for bounded continuous outputs.

Outline

The paper is organized as follows. In Sec. II we introduce the method of our analysis. We define a suitable measure of the generalization error, describe the training phase as an optimization problem, and calculate the free energy. In Sec. III we derive order parameters for the linear student. A general examination of realizable cases is presented in Sec. IV. The next section deals with the piecewise linear student. In Sec. VI we discuss a learning strategy at finite temperatures. The last section reflects the results.

©1993 The American Physical Society

II. MEAN-FIELD CALCULATION

A. Generalization error

The generalization error measures the averaged difference between the desired output ζ_* and the actual output ζ of the network. Because the outputs can take continuous values, we measure the difference with the quadratic deviation,

$$E_G=\tfrac{1}{2}\langle[\zeta_*(\mathbf{S})-\zeta(\mathbf{S})]^2\rangle_{\{\mathbf{S}\}}$$
$$=\tfrac{1}{2}\langle[g_*(h_*)-g(h)]^2\rangle_{\{\mathbf{S}\}}\ . \qquad (4)$$

If the inputs are distributed independently with mean zero and variance one, the local fields are sums of independent random numbers and in the limit $N\rightarrow\infty$ they become Gaussian distributed. We describe them by Gaussian variables x and y which are correlated through the couplings $\mathbf{W}^*$ and $\mathbf{W}$. The correlations are expressed by the values of the following order parameters:

$$\langle(x)^2\rangle=\frac{1}{N}\sum_{j=1}^{N}(W_j^*)^2=1\ , \qquad (5)$$

$$\langle xy\rangle=\frac{1}{N}\sum_{j=1}^{N}W_j^*W_j\equiv R\ , \qquad (6)$$

$$\langle(y)^2\rangle=\frac{1}{N}\sum_{j=1}^{N}(W_j)^2\equiv Q\ . \qquad (7)$$

The first can be chosen to be equal to one, the second measures the overlap between the weights of teacher and student, and the third one is the norm of the students' weights. The generalization error can be expressed as an integral over the two Gaussian variables x and y with a two-dimensional density function $F(x,y)$ which has a covariance matrix given by (5)–(7),

$$E_G=\int_{-\infty}^{\infty}dx\int_{-\infty}^{\infty}dy\,F(x,y)\tfrac{1}{2}[g_*(x)-g(y)]^2\ . \qquad (8)$$

It can be transformed into a more convenient form,

$$E_G=\int_{-\infty}^{\infty}Dx\int_{-\infty}^{\infty}Dy\tfrac{1}{2}\times\{g_*(x)-g[(Q-R^2)^{1/2}y+Rx]\}^2\ , \qquad (9)$$

where Dx is the Gaussian measure

$$Dx\equiv\frac{dx}{\sqrt{2\pi}}e^{-x^2/2}\ . \qquad (10)$$

The behavior of the generalization error is completely determined by the order parameters Q and R.

B. Learning algorithms

The training phase can be formulated as an optimization problem. The p input-output pairs provide p equations $\zeta^\mu=g(h^\mu(\mathbf{W}))$ and one has to optimize the N weights W_j of the student in order to minimize learning error E_L.

If the minimal learning error is zero, the optimal couplings can be expressed analytically. From $E_L=0$ it follows that $\zeta^\mu=\zeta_*^\mu$ for all μ, which yields a condition for h^μ, namely, $g(h^\mu)=\zeta_*^\mu$. The solution for the couplings with minimal norm Q has a pseudoinverse form [9],

$$W_j=\frac{1}{\sqrt{N}}\sum_{\mu,\nu=1}^{p}h^\mu(C^{-1})_{\mu\nu}\xi_j^\nu\ ,\quad \text{with } g(h^\mu)=\zeta_*^\mu\ , \qquad (11)$$

where $(C^{-1})_{\mu\nu}$ is the inverted correlation matrix $C_{\mu\nu}=(1/N)\xi^\mu\cdot\xi^\nu$.

Obviously one gets $\zeta^\tau=g(h^\tau)=\zeta_*^\tau$. We will call a case learnable if the learning error can be minimized to zero.

Here one should notice that the gain factor γ_S of the student IO is not an independent parameter. From the condition $g(h^\mu)=\zeta_*^\mu$ it follows that $W_j\propto h^\mu\propto\gamma_S^{-1}$, therefore the product γ_S^2Q is constant. The norm of the couplings determines the effective gain of the student. Later we will decide which of the two parameters we will choose as independent.

If the minimal learning error is larger than zero (this includes learnable cases above $\alpha=1$), one can apply an iterative learning procedure called adaline (adaptive linear) [10,11] to find the optimal solution. It is a gradient procedure similar to the backpropagation algorithm [8], which finds the optimum with respect to the quadratic deviation. For $\alpha<1$ it yields the pseudoinverse solution. Here we will only mention how the couplings are changed by this procedure,

$$\Delta W_j\propto\frac{\partial E_L^\mu}{\partial W_j}\propto\sum_{\mu=1}^{p}[g_*(h_*^\mu)-g(h^\mu)]g'(h^\mu)\xi_j^\mu\ . \qquad (12)$$

A more complete description can be found in the literature, for example in [12].

Beside these two methods, standard numerical algorithms for quadratic optimization can be used to solve the problem.

C. Free energy

If we want to know the evolution of the generalization error with the number of examples we have to calculate the values of the order parameters Q and R.

The above-mentioned learning strategy will be to minimize the learning error by adjusting the weight vector $\mathbf{W}$. More generally, we will assume a stochastic training algorithm that results in a Gibbs distribution of the couplings $\mathbf{W}$. As a normalization condition we use the spherical constraint which selects all weights with the norm Q. This results in a partion function

$$Z=\int_{-\infty}^{\infty}\prod_{j=1}^{N}dW_j\exp\left[-\beta\sum_{\mu=1}^{p}E_\mu(\mathbf{W})\right]\times\delta\left[\sum_{j=1}^{N}(W_j)^2-NQ\right]\ , \qquad (13)$$

where E_μ is the training energy, which is defined as the learning error of example ξ^μ,

$$E_\mu(\mathbf{W})=\tfrac{1}{2}\left[g_*\left[\frac{1}{\sqrt{N}}\sum_{j=1}^{N}W_j^*\xi_j^\mu\right]-g\left[\frac{1}{\sqrt{N}}\sum_{j=1}^{N}W_j\xi_j^\mu\right]\right]^2\ . \qquad (14)$$

As usual the order parameters will be found from the free energy, which has to be averaged over all input patterns. This can be done using the replica trick,

$$-\beta f=\frac{1}{N}\langle \ln Z\rangle=\frac{1}{N}\lim_{n\to 0}\frac{\langle Z^n\rangle-1}{n} \;. \tag{15}$$

The free energy will give us the averaged learning error,

$$E_L=\frac{1}{\alpha}\frac{\partial(\beta f)}{\partial\beta} \;. \tag{16}$$

To shorten the calculation we can again introduce Gaussian variables u_μ and v_μ^ρ which approach the local fields h_*^μ and h_ρ^μ for fixed weights $\mathbf{W}$ in the limit $N\to\infty$. Note that $\rho=1,\ldots,n$ is the replica index. As in the case of the generalization error above, they are correlated through the weights:

$$\langle u_\mu u_\nu\rangle=\delta_{\mu\nu}\;,\quad \langle v_\mu^\rho v_\nu^\rho\rangle=\delta_{\mu\nu}Q_\rho\;, \tag{17}$$

$$\langle v_\mu^\rho v_\nu^\sigma\rangle=\delta_{\mu\nu}q_{\rho\sigma}\;,\quad \langle u_\mu v_\nu^\rho\rangle=\delta_{\mu\nu}R_\rho\;. \tag{18}$$

Since we will be interested only in the replica symmetric case, we assume replica symmetry for the order parameters, i.e., $Q_\rho=Q$, $q_{\rho\sigma}=q$ $(\rho\neq\sigma)$, and $R_\rho=R$. Then we can get rid of the correlations, if we transform the two variables into three uncorrelated normal distributed variables, i.e., x_μ, y_μ, and z_μ^ρ,

$$u_\mu\equiv x_\mu\;, \tag{19}$$

$$v_\mu^\rho\equiv(q-R^2)^{1/2}y_\mu+Rx_\mu+\sqrt{Q-q}\,z_\mu^\rho\;. \tag{20}$$

This leaves us with the averaged replicated partition function

$$\langle Z^n\rangle=\int_{-\infty}^{\infty}\prod_\mu Dx_\mu\int_{-\infty}^{\infty}\prod_\mu Dy_\mu\int_{-\infty}^{\infty}\prod_{\mu,\rho}Dz_\mu^\rho\exp\{NG_n(Q,q,R)\}$$
$$\times\exp\left[\frac{\beta}{2}\sum_{\mu,\rho}\{g_*(x_\mu)-g[(q-R^2)^{1/2}y_\mu+Rx_\mu+\sqrt{Q-q}\,z_\mu^\rho]\}^2\right]\;. \tag{21}$$

The factor including G_n, which measures the phase space volume for fixed Q, q, and R, does not need any further examination. It is not affected by the IO relations and can be taken from previous works [3]. In the limit $n\to0$ it is

$$G_n(Q,q,R)\simeq n\left[\frac{1}{2}\frac{q-R^2}{Q-q}+\tfrac{1}{2}\ln(Q-q)\right]\equiv nG_0'(Q,q,R)\;. \tag{22}$$

Evaluating the $n\to0$ limit in the remaining terms produces the free energy

$$-\beta f=\alpha\int_{-\infty}^{\infty}Dx\int_{-\infty}^{\infty}Dy\,\ln\left[\int_{-\infty}^{\infty}Dz\,\exp\left\{-\frac{\beta}{2}[g_*-g]^2\right\}\right]+G_0'(Q,q,R)\;, \tag{23}$$

with

$$g_*=g_*(x)\quad\text{and}\quad g=g[(q-R^2)^{1/2}y+Rx+\sqrt{Q-q}\,z]\;. \tag{24}$$

The three integrals can be solved analytically only in the case of a linear student. Later this will prove to be not as severe as it now seems.

III. LINEAR STUDENT

The case, which can be calculated analytically, is the one with the linear student $g(x)=\gamma_S x$. We will solve this case in a general form for an unspecified teacher $g_*(x)$. Included are the cases with the linear teacher, which was examined by Krogh and Hertz [13,14], and the binary teacher, where $g_*(x)=\operatorname{sgn}(x)$, which was solved by Opper *et al.* [3].

Before we evaluate the integrals in (23), it is useful to introduce the following two abbreviations:

$$\langle g_*^2\rangle\equiv\int_{-\infty}^{\infty}Dx\,g_*^2(\gamma_T x)\;,$$
$$\langle g_*x\rangle\equiv\int_{-\infty}^{\infty}Dx\,g_*(\gamma_T x)x\;. \tag{25}$$

With these definitions, we get the following expression for the free energy:

$$-\beta f=\frac{1}{2}\frac{q-R^2}{Q-q}+\tfrac{1}{2}\ln(Q-q)-\frac{\alpha}{2}\ln[1+\beta\gamma_S^2(Q-q)]$$
$$-\frac{\alpha\beta}{2}\frac{\langle g_*^2\rangle-2\gamma_S R\langle g_*x\rangle+\gamma_S^2 q}{1+\beta\gamma_S^2(Q-q)}\;. \tag{26}$$

The desired order parameters R and q are the values which make the free energy stationary,

$$R=\frac{1}{\gamma_S}\langle g_*x\rangle\frac{\alpha}{a}\;,$$
$$q=\frac{1}{\gamma_S^2}\frac{\alpha}{a^2-\alpha}\left[\langle g_*^2\rangle-\alpha\frac{2-a}{a}\langle g_*x\rangle^2\right]\;, \tag{27}$$

with the abbreviation

$$a\equiv1+[\beta\gamma_S^2(Q-q)]^{-1}\;. \tag{28}$$

The generalization error and the learning error can be expressed in terms of the same abbreviations,

$$E_G=\tfrac{1}{2}(\langle g_*^2\rangle-2\gamma_S R\langle g_*x\rangle+\gamma_S^2 Q)\;, \tag{29}$$

$$E_L=\tfrac{1}{2}(\langle g_*^2\rangle-2\gamma_S R\langle g_*x\rangle+\gamma_S^2 q)\left[\frac{a-1}{a}\right]^2+\frac{1}{2\beta a}\;. \tag{30}$$

If we want to find the absolute minimum of the learning error we have to go to the limit $\beta\rightarrow\infty$. Before this can be done we have to distinguish two cases, the case with α smaller than 1 and the one with α larger than 1.

$\alpha<1$: Remembering the interpretation of learning as an optimization problem (see learning algorithms), the condition $\alpha<1$ means that one has fewer equations than parameters W_j. That is why many solutions exist which have different norms Q of the weights. The factor $\beta\gamma_S^2(Q-q)$ diverges with $\beta\rightarrow\infty$. After the evaluation of the limit $\beta\rightarrow\infty$, one can choose the solution with the minimal norm $Q=q$, which corresponds to the pseudoinverse solution (11).

The order parameters R and Q take the form $(a=1)$

$$R=\frac{1}{\gamma_S}\langle g_*x\rangle\alpha\ ,\quad Q=\frac{1}{\gamma_S^2}\frac{\alpha}{1-\alpha}[\langle g_*^2\rangle-\alpha\langle g_*x\rangle^2] \tag{31}$$

and the generalization and learning errors are

$$E_G=\frac{1}{2}\frac{1}{1-\alpha}[\langle g_*^2\rangle-\alpha(2-\alpha)\langle g_*x\rangle^2]\ ,\quad E_L=0\ . \tag{32}$$

$\alpha>1$: This case is overdetermined, one has more equations than parameters $(p>N)$. Not all equations can be fulfilled exactly, but one can find an optimal solution. Therefore one has to find a minimum of the free energy with respect to the norm Q. This variation provides a value for the factor $\beta\gamma_S^2(Q-q)$,

$$\beta\gamma_S^2(Q-q)=(\alpha-1)^{-1}\ , \tag{33}$$

which is finite even in the limit $\beta\rightarrow\infty$.

Now the order parameters become $(a=\alpha)$

$$R=\frac{1}{\gamma_S}\langle g_*x\rangle\ ,$$
$$Q=\frac{1}{\gamma_S^2}\frac{1}{\alpha-1}[\langle g_*^2\rangle-(2-\alpha)\langle g_*x\rangle^2]\ , \tag{34}$$

and thus the generalization and the learning error are

$$E_G=\frac{1}{2}\frac{\alpha}{\alpha-1}[\langle g_*^2\rangle-\langle g_*x\rangle^2]\ ,\quad E_L=\left[\frac{\alpha-1}{\alpha}\right]^2E_G\ . \tag{35}$$

Here it can be seen again that γ_S is determined by the norm Q of the weights. Only the product γ_S^2Q has a physical meaning. One of the two can be fixed to 1. Later we will see that it is more instructive to keep Q fixed to 1 as we demanded for the teacher. This defines new order parameters $\tilde{\gamma}_S$ and $\tilde{R}$. The first describes the effective gain of the student due to the adaptation of the weights W_j, the other the normalized overlap between the two weight vectors $\mathbf{W}^*$ and $\mathbf{W}$,

$$\tilde{\gamma}_S\equiv\gamma_S\sqrt{Q}\ ,\quad \tilde{R}\equiv\frac{R}{\sqrt{Q}}\ . \tag{36}$$

The above equations simplify in a special case where the teacher is also linear.

A. Realizable case: Linear teacher

An interesting special case is the one with identical input-output relations for the teacher and the student. Then the student should be able to realize the target rule perfectly. We will call such a case realizable. A case is unrealizable if the generalization error will always be larger than zero. For the case at hand the two abbreviations (25) reduce to $\langle g_*^2\rangle=\langle g_*x\rangle^2=\gamma_T^2$, which simplifies the equations for the order parameters and the errors. For $\alpha<1$

$$\tilde{\gamma}_S=\gamma_T\sqrt{\alpha}\ ,\quad \tilde{R}=\sqrt{\alpha}\ ,\quad E_G=\tfrac{1}{2}\gamma_T^2(1-\alpha)\ ,\quad E_L=0\ .$$

For $\alpha>1$

$$\tilde{\gamma}_S=\gamma_T\ ,\quad \tilde{R}=1\ ,\quad E_G=0,\quad E_L=0\ .$$

E_G decreases linearly from its maximal value at $\alpha=0$ to zero at $\alpha=1$. If $p=N$, there are enough equations to determine the N variables W_j completely. The student is an exact copy of the teacher. Many extensions of this case have been studied by Krogh and Hertz [13,14].

B. Unrealizable cases: Other teachers

In these cases the IO relations of the teacher and the student are different. The student will never be able to learn the rule exactly.

If $g_*(x)=\text{sgn}(x)$, this is the binary teacher, which was examined by Opper *et al.* [3]. The order parameters can be compared, if one uses the values of the abbreviations (25), $\langle g_*^2\rangle=1$ and $\langle g_*x\rangle=\sqrt{2/\pi}$.

As an example we will evaluate the case where $g_*(x)=\tanh(\gamma_Tx)$. Figure 1 shows the behavior of the generalization error for this case.

The piecewise linear teacher will show qualitatively the same behavior. Both cases include the sign teacher as a special case in the limit $\gamma_T\rightarrow\infty$.

If one wants to understand the behavior of E_G the order parameters $\tilde{\gamma}_S$ and $\tilde{R}$ introduced above show their usefulness. There are three interesting limits to consider, namely, $\alpha\rightarrow0$, 1, and ∞. For $\alpha\rightarrow0$

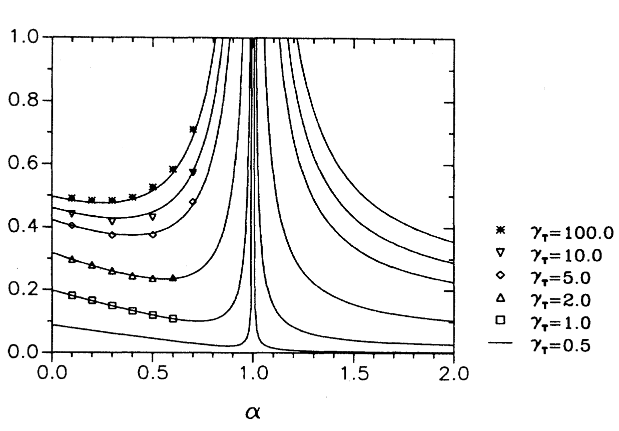

FIG. 1. Generalization error for a typical unrealizable case, linear student learns tanh teacher. Several teacher gains are shown. The dots mark the results of the simulations $(N=100)$.

$$\tilde{\gamma}_S \to 0\ ,\quad \tilde{R} \to 0\ ,\quad E_G = \tfrac{1}{2}\langle g_*^2 \rangle,\quad E_L = 0\ .$$

For $\alpha \to 1$

$$\tilde{\gamma}_S \to \infty\ ,\quad \tilde{R} \to 0\ ,\quad E_G \to \infty\ ,\quad E_L = 0\ .$$

For $\alpha \to \infty$

$$\tilde{\gamma}_S \to \langle g_* x \rangle\ ,\quad \tilde{R} \to 1\ ,$$

$$E_G = \tfrac{1}{2}(\langle g_*^2 \rangle - \langle g_* x \rangle^2)\ ,\quad E_L = E_G\ .$$

At $\alpha = 0$ both $\tilde{\gamma}_S$ and $\tilde{R}$ are zero. The student has not learned anything yet. The generalization error is a Gaussian average of the teacher output.

At $\alpha = 1$ the network is just at the limit of its perfect learning capacity. It has learned the maximal number of examples without error. This leads to a divergence of the norm of the couplings Q. But the student has completely misunderstood the rule ($\tilde{R} = 0$). The generalization error goes to infinity. This phenomenon is known as overfitting [15].

As $\alpha \to \infty$ the generalization error approaches its minimal value. The weights of the student W_j approach the teacher's weights W_j^*; this is expressed by $\tilde{R} = 1$. The gain factor $\tilde{\gamma}_S$ takes its optimal value $\langle g_* x \rangle$.

The asymptotic behavior can be understood without knowledge of the order parameters. From the assumption that the best generalization can only be achieved if the local fields are identical, one can infer the identity of the weights ($\tilde{R} = 1$). This implies an asymptotic generalization error

$$E_G(\infty) = \tfrac{1}{2}\int_{-\infty}^{\infty} Dx\,[g_*(\gamma_T x) - \tilde{\gamma}_S x]^2$$
$$= \tfrac{1}{2}(\langle g_*^2 \rangle - 2\tilde{\gamma}_S \langle g_* x \rangle + \tilde{\gamma}_S^2)\ . \qquad (37)$$

The gain of the student can be optimized to minimize $E_G(\infty)$.

$$\frac{\partial E_G}{\partial \tilde{\gamma}_S} = 0 \Longrightarrow \tilde{\gamma}_S^{\text{opt}} = \langle g_* x \rangle\ . \qquad (38)$$

These results are in complete agreement with the limits of the order parameters.

IV. OTHER REALIZABLE CASES

Now we want to solve the general realizable case, where the IO relations of the teacher and the student are identical, i.e., $g_*(x) = g(x)$ and $g^{-1}(\zeta_*)$ exists for all teacher outputs ζ_*. Then the prescription for the pseudoinverse couplings (11) simplifies.

$$g(h^\mu) = \zeta_*^\mu \Longrightarrow h^\mu = g^{-1}(\zeta_*^\mu) = g^{-1}(g_*(h_*^\mu)) = h_*^\mu\ . \qquad (39)$$

The weights W_j are thus independent of the IO relation.

The independence can also be seen in the free energy (23) in the limit $\beta \to \infty$,

$$\exp\left\{-\frac{\beta}{2}[g_* - g]^2\right\} \propto \delta(\zeta_* - g(h))\ ,\quad \beta \to \infty$$
$$\propto \delta(g^{-1}[g_*(h_*)] - h),\quad \exists g^{-1}(\zeta_*)$$
$$\propto \delta(h_* - h)\ ,\quad g = g_* \qquad (40)$$

where, in the second line, we have exploited the existence of $g^{-1}(\zeta_*)$.

This means that training provides for all realizable cases the same order parameters. There is no need to calculate the order parameters; we can directly take the ones of the linear realizable case, $\tilde{\gamma}_S = \sqrt{\alpha}\gamma_T$ and $\tilde{R} = \sqrt{\alpha}$.

In the particular case with $g_*(x) = g(x) = \tanh(\gamma_T x)$, the generalization error is for $\alpha < 1$,

$$E_G = \tfrac{1}{2}\int_{-\infty}^{\infty} Dx \int_{-\infty}^{\infty} Dy[\tanh(\gamma_T x)$$
$$- \tanh(\gamma_T \sqrt{\alpha}\sqrt{1-\alpha} y$$
$$+ \gamma_T \alpha x)]^2\ . \qquad (41)$$

Above $\alpha = 1$ it is zero (see Fig. 2). These analytical results are supported by simulations.

Increase of E_G for small α

Although the network is learning examples its understanding about the rule can decrease. This is a phenomenon of a bounded teacher IO relation. In a general realizable case we can approximate the student for small α by a linear function,

$$E_G(\alpha \ll 1) = E_G(\alpha = 0) - \tfrac{1}{2}\alpha\gamma_T(2\langle g_* x \rangle - \gamma_T)\ . \qquad (42)$$

As long as $2\langle g_* x \rangle$ is larger than γ_T, the generalization error will decrease. This is always fulfilled in the linear case, where $\langle g_* x \rangle = \gamma_T$. But with a bounded IO, $\langle g_* x \rangle$ is also bounded and γ_T can become larger than $2\langle g_* x \rangle$. In the case of $g_*(x) = \tanh(\gamma_T x)$ the critical gain factor is $\gamma_T^c = 1.33$.

One can understand the effect if one remembers that learning means optimization of gain factor $\tilde{\gamma}_S$ and couplings W_j. Since it seems that the student approximates the teacher first linearly, the increase of the gain γ_S can be too fast if the teacher has a bounded IO relation.

V. PIECEWISE LINEAR STUDENT

Learning with a piecewise IO causes problems which we can understand immediately, if we look at the learn-

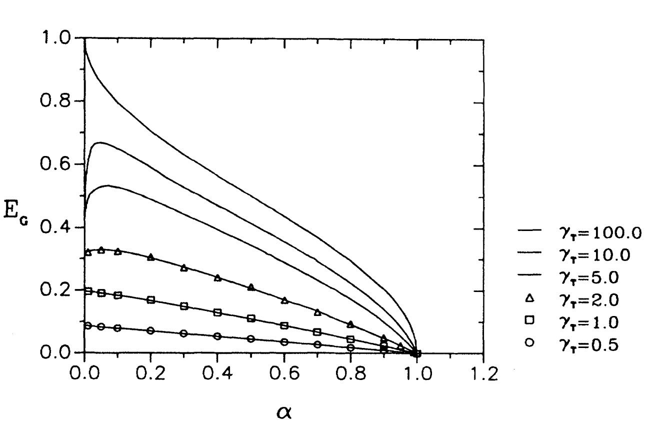

FIG. 2. Generalization error for the realizable case, where a tanh student learns a tanh teacher. Above $\alpha = 1$ it is zero. For small α and high teacher gains γ_T the generalization error can increase while α is increasing.

ing algorithms. The pseudoinverse solution (11) demands an invertible student function, which is not given because of the constant part of the function $f_{\rm PL}$. Moreover, the adaline learning rule (12) changes the weights only if the derivative g' is nonzero, which is also not fulfilled by the constant part of $f_{\rm PL}$.

That is why we will try to find a learning strategy for the PL student (i.e., $g=f_{\rm PL}$). We will restrict learning to the linear part of g. If we try to learn all the outputs of the teacher, we will end with the results of the linear student. Therefore we have to restrict the teacher too. For a PL teacher (i.e., $g_*=f_{\rm PL}$) the restriction is obvious; it means that only patterns with an output smaller than 1 will be learned. For other teachers we need a more general condition.

To find a threshold for the outputs of an unspecified teacher, we can look at the asymptotics. The asymptotic generalization error can be calculated without knowledge of the order parameters, see (37). It can be used to set up a condition for the gain factor of the PL student, which learns a teacher g_*,

$$E_G(\infty)=\tfrac{1}{2}\int_{-\infty}^{\infty}Dx[g_*(\gamma_T x)-f_{\rm PL}(\tilde{\gamma}_S x)]^2 \ , \tag{43}$$

$$\frac{\partial E_G}{\partial\tilde{\gamma}_S}=0\Longrightarrow\tilde{\gamma}_S^{\rm opt}({\rm PL}) \ . \tag{44}$$

In our new learning method we learn only outputs which correspond to local fields of teacher and student in a restricted interval $[-x_0,x_0]$, where x_0 has to be determined. We calculate again the asymptotic generalization error, but now in the restricted interval,

$$E_G(\infty)=\tfrac{1}{2}\int_{-x_0}^{x_0}Dx[g_*(\gamma_T x)-\tilde{\gamma}_S x]^2 \ , \tag{45}$$

$$\frac{\partial E_G}{\partial\tilde{\gamma}_S}=0\Longrightarrow\tilde{\gamma}_S^{\rm opt}({\rm linear},x_0) \ . \tag{46}$$

If we demand that this effective linear student in the restricted interval should have the same gain as the true PL student, we get a condition for the boundary x_0,

$$\tilde{\gamma}_S^{\rm opt}({\rm linear},x_0)=\tilde{\gamma}_S^{\rm opt}({\rm PL})\Longrightarrow x_0 \ . \tag{47}$$

The boundary x_0 for the local fields is connected via the IO relation g_* with a threshold for the outputs $g_*(\gamma_T x_0)$.

We can test the method at the PL teacher. The optimal gain is obviously γ_T, therefore the interval must be in the linear part of the teacher, i.e., $x_0=\gamma_T^{-1}$, the threshold is, as expected, smaller than 1. The method is consistent and makes sense.

Now we will use it for the tanh teacher. Table I shows the numerical results for some examples of teacher gains γ_T.

Naturally the restricted learning is of practical interest only if the threshold is more or less independent of the unknown teacher's gain. So it is quite surprising that this is also fulfilled for the tanh teacher; see the last line of Table I. To restrict the learning on outputs with an amount smaller than 0.85 is a good strategy to solve the case with the tanh teacher.

TABLE I. Numerical determination of the output threshold for the case PL student learns from tanh teacher.

γ_T	0.50	1.00	2.00	5.00	10.00	100.0
$\gamma_S^{\rm opt}$	0.43	0.80	1.55	3.85	7.70	76.9
x_0	2.33	1.25	0.64	0.26	0.13	0.013
$\tanh(\gamma_T x_0)$	0.82	0.85	0.86	0.86	0.86	0.86

A. Order parameters

Now it should be obvious how the replica calculation has to be changed to fit for this strategy. The x integration in (23) must be confined to the interval $[-x_0,x_0]$. The restriction means also that not every pattern is learned. Therefore it is useful to define an effective fraction of patterns $\alpha_{\rm eff}$,

$$\alpha_{\rm eff}=\alpha\int_{-x_0}^{x_0}Dx=\alpha\,{\rm erf}(x_0/\sqrt{2}) \ . \tag{48}$$

This leads to effective abbreviations $\langle g_*^2\rangle_{\rm eff}$, $\langle g_* x\rangle_{\rm eff}$, and $\langle x^2\rangle_{\rm eff}$ of the form

$$\langle x^2\rangle_{\rm eff}=\Big(\int_{-x_0}^{x_0}Dx\,x^2\Big)\Big(\int_{-x_0}^{x_0}Dx\Big)^{-1} \ . \tag{49}$$

In the limit $\beta\longrightarrow\infty$ we get the order parameters R and Q.

If $\alpha<1$,

$$R=\frac{1}{\gamma_S}\alpha_{\rm eff}\frac{\langle g_* x\rangle_{\rm eff}}{1-\alpha[1-\langle x^2\rangle_{\rm eff}]} \ , \tag{50}$$

$$Q=\frac{1}{\gamma_S^2}\frac{\alpha_{\rm eff}}{1-\alpha_{\rm eff}}[\langle g_*^2\rangle_{\rm eff}-2\gamma_S R\langle g_* x\rangle_{\rm eff}+\gamma_S^2R^2\langle x^2\rangle_{\rm eff}]+R^2 \ , \tag{51}$$

and if $\alpha>1$,

$$R=\frac{1}{\gamma_S^2}\frac{\langle g_* x\rangle_{\rm eff}}{\langle x^2\rangle_{\rm eff}} \ , \tag{52}$$

$$Q=\frac{1}{\gamma_S^2}\frac{1}{\alpha_{\rm eff}-1}[\langle g_*^2\rangle_{\rm eff}-2\gamma_S R\langle g_* x\rangle_{\rm eff}+\gamma_S^2R^2\langle x^2\rangle_{\rm eff}]+R^2 \ . \tag{53}$$

They have a form similar to those of the linear student; this leads to the same behavior. The case in which teacher and student are both piecewise linear should again be realizable.

B. Realizable case: PL teacher

With $x_0=\gamma_T^{-1}$ the constants can be expressed in terms of $\langle x^2\rangle_{\rm eff}$,

$$\langle g_*^2\rangle_{\rm eff}=\gamma_T^2\langle x^2\rangle_{\rm eff} \ \text{and} \ \langle g_* x\rangle_{\rm eff}=\gamma_T\langle x^2\rangle_{\rm eff} \ . \tag{54}$$

This changes the behavior of the order parameters in the case of $\alpha_{\rm eff}<1$. Q does not diverge if $\alpha_{\rm eff}$ approaches 1. Above $\alpha_{\rm eff}=1$ we get

$$Q=R^2=\frac{\gamma_T^2}{\gamma_S^2}\Longrightarrow\tilde{\gamma}_s=\gamma_T \ , \quad \tilde{R}=1 \ , \tag{55}$$

which yields a vanishing generalization error.

Figure 3 shows the behavior of E_G. The result is

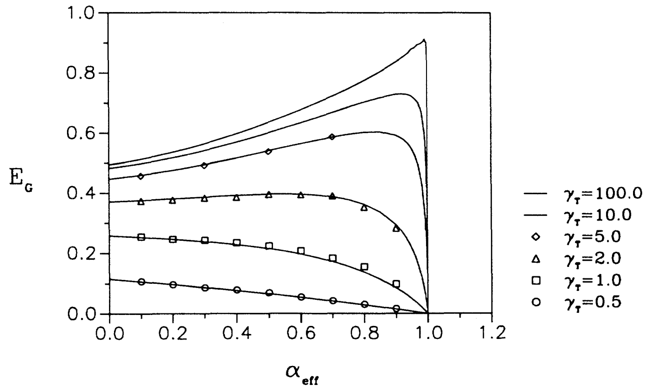

FIG. 3. Generalization error of an effective piecewise linear student, which learns a PL teacher.

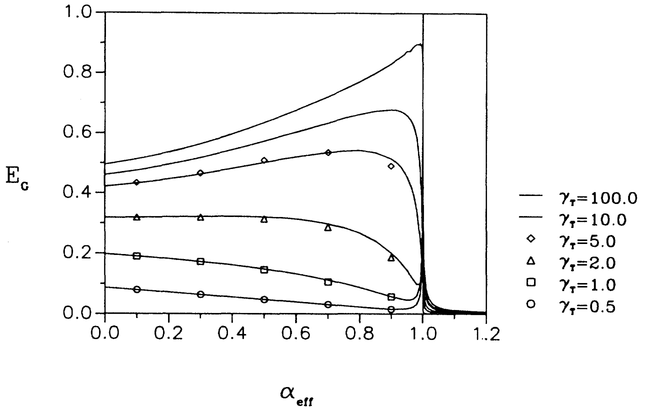

FIG. 4. Generalization error of an effective piecewise linear student which learns a tanh teacher.

confirmed by simulations. The simulations show a strong finite size effect if x_0 becomes very small.

C. Unrealizable case: tanh teacher

This unrealizable case shows the same characteristic behavior as the unrealizable cases of the linear student (see Fig. 4). In the different limits for α_{eff} the order parameters take the same values. As a consequence of the boundness of both IO relations the generalization does not diverge. With the greater similarity of the IO relations (note that they become identical if $\gamma_T \to \infty$) the asymptotic errors are much smaller.

VI. LEARNING WITH A FINITE ERROR

Until now our strategy was to find the absolute minimum of the learning error hoping that this will lead to the best generalization ability. The overfitting phenomenon in the unrealizable cases was an unwanted result of this strategy. Next we will show that sometimes it can be better to learn less than optimally. Overfitting can be diminished if we terminate the training process at a learning error larger than the minimal one,

$$E_L(t) \le E_L^{\min} + \epsilon \ , \quad \text{with } \epsilon > 0 \ . \tag{56}$$

For an illustration we go back to the linear student. For a given learning algorithm this strategy is simply implemented by minimizing with reduced accuracy. Within an analytical framework this approach can be modeled at least qualitatively by integrating over the volume of all couplings which yield a fixed positive learning error E_L. This ansatz is equivalent to learning at a positive temperature in (13).

If we insert the values of the order parameters R and q (27) in Eqs. (29), generalization and the learning error can be expressed at finite temperature in the form

$$E_G = A(a,\alpha) + \frac{1}{2\beta(a-1)} \ , \tag{57}$$

$$E_L = A(a,\alpha)\left[\frac{a-1}{a}\right]^2 + \frac{1}{2\beta a} \ , \tag{58}$$

where

$$A(a,\alpha) = \frac{1}{2}\frac{1}{a^2-\alpha}[a^2\langle g_*^2\rangle - \alpha(2a-\alpha)\langle g_* x\rangle^2] \ , \tag{59}$$

and

$$a \equiv 1 + [\beta\gamma_s^2(Q-q)]^{-1} \ . \tag{60}$$

[Equations (32) and (35) describe the errors in the case of $\beta \to \infty$, where a was 1 for $\alpha > 1$, and $a = \alpha$ for $\alpha > 1$.]

If we assume β to be large but finite, we can neglect the last term in both equations. Then the generalization error and the learning error are connected via the variable a. If we choose a finite ϵ in (56), the expression (58) gives us the appropriate value $a(\alpha,\epsilon)$. Then the first expression yields the generalization error $E_G(\alpha,\epsilon)$ with respect to α and ϵ.

Figure 5 shows the evolution of E_G with α for different values of ϵ in the case where a linear student learns tanh teacher with a gain $\gamma_T = 5.00$.

The curve with error-free learning ($\epsilon = 0.0$) is nearly reached by $\epsilon = 0.0001$; larger ϵ show a reduced overfitting. For $\epsilon = 0.1$ the overfitting has vanished, but this large ϵ nearly doubles the asymptotic generalization error, which is $E_G^{\min}(\infty) = 0.112$. Therefore an ϵ between 0.01 and 0.1 will be optimal.

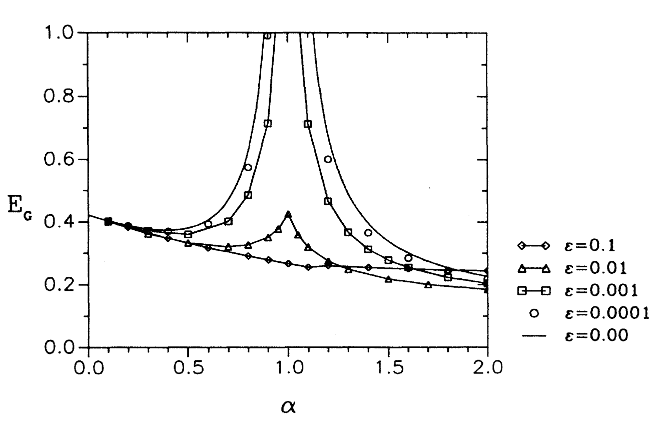

FIG. 5. Generalization error for learning with a finite learning error E_L.

It is interesting to compare our results to a recent study of Krogh and Hertz [13], who found that the overfitting arising from noisy teacher outputs can be suppressed by learning strategies which allow learning with errors. A related result was obtained by Opper and Haussler [16] within the framework of Bayesian learning. Here a noise in the teacher outputs naturally leads to learning at a finite temperature. This suggests that whenever teacher and student will not properly match, due to noise or inappropriate network architecture, learning will benefit from errors in the training process.

VII. CONCLUSION

In the present paper we investigated the influence of different input-output relations on the generalization ability of a single-layer perceptron. It turns out that this problem exhibits many characteristic phenomena of generalization. Choosing different input-output relations independently for teacher and student we got realizable cases, if the IO's are identical, and unrealizable cases, if they are not. For a gradient descent learning rule it was important, whether the student's IO is invertible for all teacher outputs or not. As long as we learned with minimal learning error, an invertible student was able to learn a realizable teacher exactly after N examples had been presented.

On the other hand, if the number of examples is small compared to N it seems that the student effectively approximates the teacher by a linear function. This leads to an increase of the generalization error in the realizable case with a bounded IO for high enough teacher gains γ_T.

An unrealizable case can never be learned exactly. It even shows a strong increase in the generalization error if α approaches 1. This is a characteristic consequence of the exact learning of the pseudoinverse learning rule, called overfitting [15]. If learning is not exact, that means if the learning process is terminated at a learning error larger than the minimal one, the overfitting can be reduced, which leads to an improved generalization ability in the surroundings of $\alpha=1$. We expect that increasing the generalization ability by learning with errors is relevant for many practical applications and is still an interesting question (see [17] and references therein).

ACKNOWLEDGMENTS

The authors thank R. Kühn, A. Scharnagl, and R. Garcés for useful hints concerning the presentation. S.B. acknowledges financial support from Stiftung Volkswagen and thanks Andreas Herz for the hospitality at Caltech during the completion of this work.

[1] G. Györgyi and N. Tishby, in *Neural Networks and Spin Glasses,* edited by R. Theumann and R. Koeberle (World Scientific, Singapore, 1990).

[2] H. S. Seung, H. Somplinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992).

[3] M. Opper, W. Kinzel, J. Kleinz, and R. Nehl, J. Phys. A **23**, L581 (1990).

[4] W. Kinzel and P. Rujàn, Europhys. Lett. **13**, 473 (1991).

[5] T. L. H. Watkin and A. Rau, J. Phys. A **25**, 113 (1992).

[6] H. S. Seung, M. Opper and H. Sompolinsky (private communication); and in *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory (COLT 92)* (ACM, New York, 1992).

[7] F. Rosenblatt, *Principles of Neurodynamics* (Spartan, New York, 1962).

[8] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing* (MIT, Cambridge, MA, 1986), Vols. 1 and 2.

[9] I. Kanter, and H. Sompolinsky, Phys. Rev. A **35**, 380 (1987).

[10] T. Kohonen, *Associative Memory,* (Springer Verlag, Berlin, 1977).

[11] S. Diederich and M. Opper, Phys. Rev. Lett. **58**, 949 (1987).

[12] J. A. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, CA, 1991).

[13] A. Krogh and J. Hertz, J. Phys. A **25**, 1135 (1992); and in *Advances in Neural Information Processing Systems III* (Morgan Kaufmann, San Mateo, 1991).

[14] A. Krogh, J. Phys. A **25**, 1119 (1992).

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

[16] M. Opper and D. Haussler, in *Proceedings of the Fourth Annual ACM Workshop on Computational Learning Theory (COLT 91),* edited by M. K. Warmuth and L. G. Valiant (Morgan Kaufmann, San Mateo, 1991).

[17] L. Holmström and P. Koistinen, IEEE Trans. Neural Networks **3**, 24 (1992).