Generalization in the programed teaching of a perceptron

Imre Derényi* and Tamás Geszti†
Department of Atomic Physics, Eötvös University, H-1088 Budapest, Hungary

Géza Györgyi‡
Institute for Theoretical Physics, Eötvös University, H-1088 Budapest, Hungary
(Received 7 March 1994)

According to a widely used model of learning and generalization in neural networks, a single neuron (perceptron) can learn from examples to imitate another neuron, called the teacher perceptron. We introduce a variant of this model in which examples within a layer of thickness $2Y$ around the decision surface are excluded from teaching. That restriction transmits global information about the teacher's rule. Therefore for a given number $p = \alpha N$ of presented examples (i.e., those outside of the layer) the generalization performance obtained by Boltzmannian learning is improved by setting Y to an optimum value $Y_0(\alpha)$, which diverges for $\alpha \to 0$ and remains nonzero while $\alpha < \alpha_c \approx 5.7$. That suggests programed learning: easy examples should be taught first.

PACS number(s): 87.22.Jb, 05.20.−y

I. INTRODUCTION

The notion of generalization is central to much of the recent theoretical efforts devoted to neural networks [1–7]. It means learning a rule of input-output association from examples, i.e., generalizing a set of examples by extracting the rule behind them.

Generalization is usually studied in the context of feedforward networks that associate an output z to an input vector $\vec{x}$ of N components x_i $(i = 1, \ldots, N)$, entering the network through N input nodes. Neural networks learn to associate given outputs z^μ $(\mu = 1, \ldots, p)$ with p input vectors $\vec{\xi}^\mu$ of binary components $\xi_i^\mu = \pm 1$.

Such learning networks can be used for at least two different purposes: pattern storage or rule extraction. In the first case the associations are defined pattern by pattern. The task of learning them gets harder and harder with growing p, and — on the average over a distribution of the patterns — becomes unsolvable beyond some critical number p_c, which marks the limit of capacity of the network. On the other hand, if the learned associations are governed by a hidden rule, through learning the examples the network gradually learns to respond to a new input more and more in accordance with that rule. That is called rule extraction by generalization from examples.

The examples impose restrictions on the networks. Those compatible with the examples form the so-called version space. The error of generalization $\epsilon(p)$ can be defined [1] as the fraction of wrong answers given to any input (possibly excluding the p learned examples when testing). The process of learning a rule is associated with the decrease of that function.

Another suggestive measure of learning is a suitably defined entropy telling us how many networks are able to perform a given task. Depending on the type of the network, it can be chosen to be $\langle \ln N \rangle$, where $\mathcal{N}$ is the combinatorial number of possible ways to combine fixed Boolean gates into a network implementing a Boolean function [8,9], or $\langle \ln V \rangle$, where V is the volume available in the "phase space" of adaptable connection strengths of a neural network [10]. (Here $\langle\ \rangle$ denotes averaging over a distribution of random choices of examples.) The problem can be extended to a finite external noise level and treated by statistical physics, where the energy function is the training error [11,3,4]. Learning is accompanied by a reduction of entropy by the amount of information contained in the examples presented to the network. As suggested by Carnevali and Patarnello [8], one can characterize a rule — in their case a Boolean function — by the residual entropy measuring the freedom of choice that remains after learning it: $S_r = \ln \mathcal{N}_r$, where $\mathcal{N}_r$ is the number of different ways rule r can be implemented through a given kind of network. They argue that easily learnable rules are those of a high residual entropy and vice versa. Although learnability in practice depends on other factors as well [9], the Carnevali-Patarnello criterion still gives a useful and suggestive first orientation.

Recently considerable progress has been achieved in analyzing the way a simple perceptron, i.e., a single neuron of binary output

$$z = \mathrm{sgn}\left(\sum_{j=1}^{N} J_j x_j\right) = \mathrm{sgn}(\vec{J} \cdot \vec{x}) \ , \qquad (1)$$

learns from examples. The task is to imitate the linear classification of input vectors implemented by another perceptron of fixed connection strengths $J_j = B_j$ $(j = 1, \ldots, N)$, the so-called teacher perceptron [12,13]. According to the general philosophy of supervised learning, the teacher issues an error message whenever the two outputs differ for some example. Learning — not specified in its algorithm or dynamics — happens by modifying $\vec{J}$

*Electronic address: derenyi@hercules.phys.elte.hu
†Electronic address: geszti@hercules.phys.elte.hu
‡Electronic address: gyorgyi@hal9000.elte.hu

until the error messages are eliminated for each of the p examples. It turned out that this kind of learning process can be analyzed by the method of replicas [3–5]. Several versions have been investigated [6,7].

In the present paper we introduce an alternative version of the original model, in which training examples are selected by the teacher who knows the rule and deliberately excludes the hardest questions, i.e., the ones almost perpendicular to the teacher's connection vector $\vec{B}$. Such examples are found in a band of width $2Y$ around the decision surface (cf. Sec. II). That gives rise to a model solvable by the replica technique, furnishing a stable replica symmetric solution. It turns out that such a modification — in accordance with the everyday teachers' wisdom that you should teach simple things first, complicated ones afterwards — reduces the error of generalization, at least if the number of examples presented is not too large. The reason is that rule-guided pattern selection introduces anisotropy in the space of examples, marking the direction of the teacher's connection vector $\vec{B}$. Thereby it acts as an extra channel of transferring information about the rule itself [17]. Even if one excludes a finite fraction of hard examples from the process of learning, in the limit $N \to \infty$ the rule can be learned completely, although with slow (logarithmic) convergence to zero error. Let us notice that this is different from selecting examples in order to improve learning performances, studied by Kinzel and Ruján [14] and further analyzed by Watkin and Rau [15], in the context of an active student who has to guess the optimum choice on the basis of knowledge obtained from previous learning.

Essentially the same calculations can be given a different interpretation: instead of excluding hard examples, we can consider the error message policy of accepting any answer to the hardest questions. That philosophy presents our model as an interpolation between supervised learning in the usual sense and reinforcement learning [16]. More generally, it can be regarded as an illustration to the important fact that neither biological nor machine learning would fully fix all connection strengths of the student network, leaving a residual freedom after learning. In all those cases learning may end up in a variety of learned states, giving the same answers in the range where the teacher had given definite association, and different answers where the teacher gave no hint. In view of that residual freedom the analysis connects our calculations to the Carnevali-Patarnello approach.

The most practical consequence — already mentioned — is the enhanced generalization ability obtained from rule-guided restriction of the training set. From that result an objective for further analysis emerges: the development of optimal teaching schedules in the sense of achieving minimum error on presenting a given number of examples in a programmed nonuniform distribution. The question is reasonably posed for neural networks of practical architectures; however, its solution is hard even for the simple perceptron, and in the present paper only a restricted question will be answered explicitly: given the total number $p = \alpha N$ of examples that can be presented in a given period of teaching, one calculates the optimal *constant* value $Y = Y_0(\alpha)$ of the width of the excluded band that gives the minimum error of generalization.

Our model is defined and analyzed by means of replicas in Sec. II. The relation to the Carnevali-Patarnello approach and related questions are discussed in Sec. III. Section IV takes up the analysis with answering the simplest question about optimal teaching schedules: determining the optimal *fixed* width of the band of hard examples that should be excluded from learning in order to achieve minimum error on exploiting a fixed number of examples. In Sec. V we briefly describe some variants of the model, differing in detail but not in the essential content nor in complexity from the one presented in the main text, and give an outlook. Finally, Appendix A contains details about the evaluation of the asymptotics of the generalization ability and Appendix B presents the derivation of an analytical result concerning the replica stability matrix.

II. REPLICA ANALYSIS

To analyze the model we use the well-known replica technique in its zero-temperature version [10,3], evaluating the entropy per input channel $s = S/N$ connected to the problem-solving volume V through

$$s = N^{-1}\left\langle \ln \frac{V}{V_0} \right\rangle = N^{-1} \frac{\partial}{\partial n}\bigg|_{n=0} \left\langle \frac{V^n}{V_0^n} \right\rangle, \tag{2}$$

where V^n is interpreted as the joint problem-solving volume for n replicas of the student perceptron, receiving the same randomly chosen examples and being supervised by the same teacher, but otherwise independent of each other in the course of learning. V_0 is the available phase-space volume restricted by normalization of $\vec{J}$ only, not by examples. We choose the normalization $\sum_{j=1}^{N} |J_j^a|^2 = \sum_{j=1}^{N} |B_j|^2 = N$, where replicas are labeled by $a = 1, \ldots, n$; then $V_0 = (2\pi e)^{N/2}$. We notice that Eq. (2) gives the zero-temperature limit of $-\beta f$, where f is the free energy per synapsis, if there is no extensive ground state energy.

A threshold of instruction Y can be introduced to restrict teaching to examples $\vec{\xi}^{\mu}$ with projection onto the direction of $\vec{B}$ larger in modulus than Y (Fig. 1), i.e.,

$$\left| N^{-1/2} \sum_{j=1}^{N} B_j \xi_j^{\mu} \right| > Y. \tag{3}$$

Then the excluded equatorial band of width $2Y$ occupies a finite fraction $1 - 2H(Y)$ (for $N \to \infty$) of the surface of the hypersphere of radius $\sqrt{N}$ in N dimensions. Here the standard notations $Dx = (dx/\sqrt{2\pi}) \exp(-x^2/2)$, and $H(z) = \int_z^{\infty} Dx$ have been used.

More generally, and anticipating the case of a tunable threshold $Y(\alpha)$, what we have to handle here is a uniaxial anisotropy in the distribution of learned examples, with the teacher's connection vector $\vec{B}$ as the rotation axis. The easiest way to do replica calculations for such

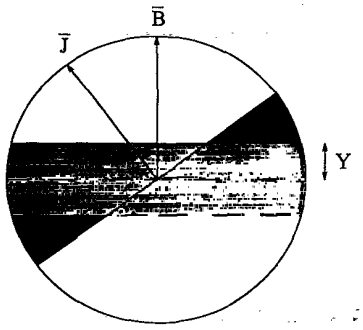

FIG. 1. Restricting the examples used for teaching: $\vec{B}$ and $\vec{J}$ are the connection vectors of the teacher and the student, respectively. Training examples are taken from outside of the light gray region. The student $\vec{J}$ shown here correctly classifies examples in the white region and misclassifies those in the dark gray region.

cases is to take into account that for a "typical" direction of $\vec{B}$ (i.e., not parallel to any of the coordinate axes) an unrestricted distribution of examples, uniform on the vertices of the N-dimensional unit cube $\xi_i^\mu = \pm 1$, can be approximated as a uniform distribution on the surface of an N-dimensional sphere of radius $\sqrt{N}$. The central limit theorem, appearing in some explicit or implicit way in all replica calculations, is now contained in the fact that the projection of such a distribution onto the $(\vec{B}, \vec{J})$ plane has mean 0 and dispersion 1 in each direction, being therefore strongly centered around the origin where it is Gaussian to $O(N^{-1})$.

In order to formulate example selection by the teacher, we introduce a weight function $f(y)$. It depends on the projection $y = \vec{B} \cdot \vec{\xi}/\sqrt{N}$ of example $\vec{\xi}$ onto $\vec{B}$, it is even in y, and it is normalized to unity over the Gaussian measure as $\int Dy f(y) = 1$. Then the distribution of examples reads

$$P(\vec{\xi}) = (2\pi)^{-\frac{N}{2}} \exp\left(-\frac{\vec{\xi}^2}{2}\right) f\left(\frac{\vec{B}\cdot\vec{\xi}}{\sqrt{N}}\right). \tag{4}$$

In the special case of excluding a band of width $2Y$ according to Eq. (3), we have

$$f(y) = \left[2H(Y)\right]^{-1} \Theta(|y| - Y). \tag{5}$$

The expression to be calculated by means of replicas is

$$\langle V^n \rangle = \int \left(\prod_{j=1}^{N} \prod_{a=1}^{n} dJ_j^a \right) \prod_{a=1}^{n} \delta\left(\vec{J}^a \cdot \vec{J}^a - N\right) \times \prod_{\mu=1}^{p} 2\left\langle \Theta\left(\vec{B}\cdot\vec{\xi}^\mu/\sqrt{N}\right) \times \prod_{a=1}^{n} \Theta(\vec{J}^a \cdot \vec{\xi}^\mu/\sqrt{N}) \right\rangle_{\{f(y)\}}. \tag{6}$$

The factor 2 accounts for the regions obtained from the ones presented explicitly by the reflexion $\vec{\xi}^\mu \to -\vec{\xi}^\mu$, giving equal contribution to the volume after averaging over an even distribution of examples.

The evaluation of the above expression follows the usual course of replica calculations. By introducing the Fourier integral representation of Θ functions, one observes that the replicated connection strengths J_j^a, the examples ξ_j^μ, and the conjugate Fourier variables are combined into random variables with Gaussian distribution. In the course of averaging second moments emerge, containing the fundamental order parameters of the problem: the student-student replica overlaps (for $a \neq b$)

$$q_{ab} = N^{-1} \sum_j J_j^a J_j^b \tag{7}$$

and the student-teacher overlaps

$$R_a = N^{-1} \sum_j B_j J_j^a. \tag{8}$$

Then for $N \gg 1$, $p \gg 1$ [$\alpha = p/N = \mathcal{O}(1)$] saddle points are sought in the multiple integrals and the replica symmetric ansatz is introduced: $q_{ab} = q\ \forall a, b\ (a \neq b)$, $R_a = R\ \forall a$. In the limit $n \to 0$ through some lengthy but standard calculation one obtains the entropy in the form

$$s = \frac{1}{2}\left(\ln(1-q) + \frac{q - R^2}{1-q}\right) + 2\alpha \int_0^\infty Dy\, f(y) \int Dt \ln H(z), \tag{9}$$

where the variable

$$z = \frac{t\sqrt{q - R^2} - yR}{\sqrt{1-q}} \tag{10}$$

has been introduced. The maximum of s can be found by solving the final equations determining q and R:

$$q\sqrt{\frac{q-R^2}{1-q}} = \alpha\sqrt{\frac{2}{\pi}} \int_0^\infty Dy\, f(y) \int_{-\infty}^\infty Dt \frac{\exp\left(-\frac{z^2}{2}\right)}{H(z)} t, \tag{11}$$

$$R\sqrt{1-q} = \alpha\sqrt{\frac{2}{\pi}} \int_0^\infty Dy\, f(y) \int_{-\infty}^\infty Dt \frac{\exp\left(-\frac{z^2}{2}\right)}{H(z)} y. \tag{12}$$

The numerical solution of Eqs. (11) and (12) gives R and q as functions of α for a given $f(y)$. Substituting the results into Eqs. (9) and (10), we obtain the corresponding values of the entropy.

As to the measure of performance of the trained network, we focus on the generalization error evaluated over test examples with Gaussian distribution, given by the standard formula [3]

$$\epsilon = (1/\pi)\cos^{-1} R. \tag{13}$$

It is only in the learning period that we try to gain ex-

tra information through example selection by the filter function $f(y)$, while in the test period $f(y) \equiv 1$ is used. In that way the generalization error has a common measure, which enables us to compare the performances of different teaching programs. We also mention here the more traditional case when the distributions of both the training and test sets are identical and given by Eq. (4), when the generalization error becomes

$$\tilde{\epsilon}(\alpha) = 2 \int_0^\infty Dy\ f(y) H\left(y \frac{R}{\sqrt{1-R^2}} \right). \qquad (14)$$

Let us turn to the evaluation of the equations of state. Equations (9)–(12) go over into the result of Ref. [3] with $f(y) \equiv 1$ and then the two generalization measures (13) and (14) coincide. In the rest of this paper we restrict ourselves to the special case corresponding to Eq. (5), i.e., teacher-guided pattern selection using criterion (3).

The entropy function Eq. (9) is displayed for generic, fixed α and Y in Fig. 2. The region of real s is delimited by the curve $q = R^2$ and the lines $R = 0$ and $q = 1$. Here, apparently a single saddle point exists, which we identify with the physically valid solution as presented hereafter in this paper. Note that while the saddle point is a local maximum in R, as usual for the most probable state in ordered systems, it is a local minimum in q, which is characteristic to disorder, as observed already in the case of the Kirkpatrick-Sherrington model [18]. The saddle point is actually the continuation in the $n \to 0$ limit of the maximum found for positive integer n. If R is considered as a quantity to which a fixed value can be assigned for any given α, a thermodynamical meaning can be attributed to the minima of $s(q, R)$ with respect to q [19]. We note that $s(q, R = 0)$ is independent of the pattern distribution $f(y)$. It corresponds to the case of random pattern storage investigated by Gardner [10]. There the local minimum as a function of q appears only for $\alpha < 2$, i.e., when error-free storage is possible.

The overlap R is shown in Fig. 3 for various values of

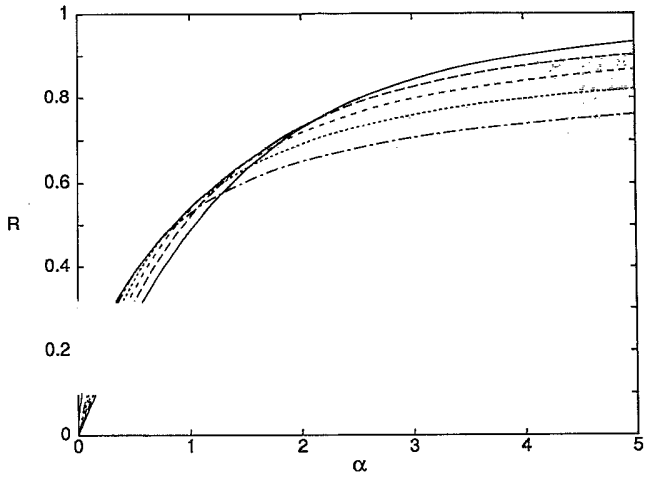

FIG. 3. The evolution of the teacher-student overlap $R(\alpha, Y)$ as a function of the number of examples presented ($\alpha = p/N$ for p examples and N input synapses), for various values of the exclusion threshold: $Y = 0$ (lower solid line), 0.25, 0.52, 0.84, and 1.28, chosen so as to give, for the respective retained fractions, $2H(Y) = 1$, 0.8, 0.6, 0.4, and 0.2. The upper solid line is the envelope. The same values of Y appear in Figs. 4–8.

Y as function of α. The curves start linearly as

$$R = \frac{1}{\pi} e^{-\frac{Y^2}{2}} \left[H(Y) \right]^{-1} \alpha, \qquad (15)$$

however, their envelope behaves as $\sqrt{\alpha}$ for small α. The corresponding numerical values for $\epsilon(\alpha, Y)$ are presented in Fig. 4.

The crossing curves of Figs. 3 and 4 indicate that at the initial stage of the learning process, and in any case if learning is restricted to a relatively small number αN of examples, it is advantageous to restrict learning to easy examples by introducing a finite threshold Y. With

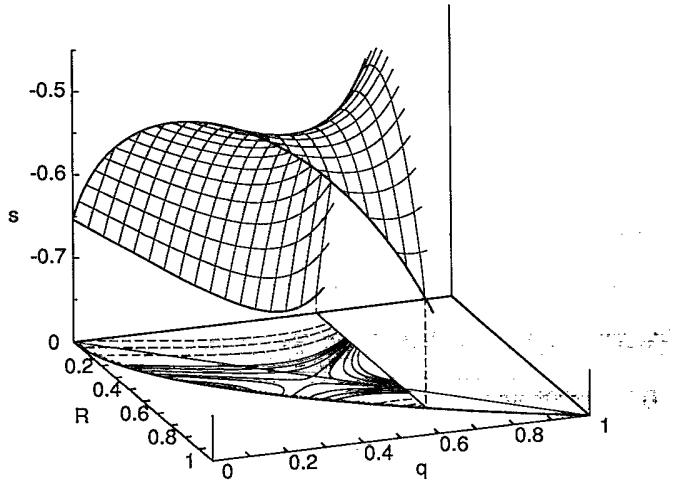

FIG. 2. Generic entropy surface for $\alpha = 0.94$ and $Y = 0.25$. There is a single saddle point, maximum in R and minimum in q, which is identified with the physically relevant solutions as presented throughout the paper.

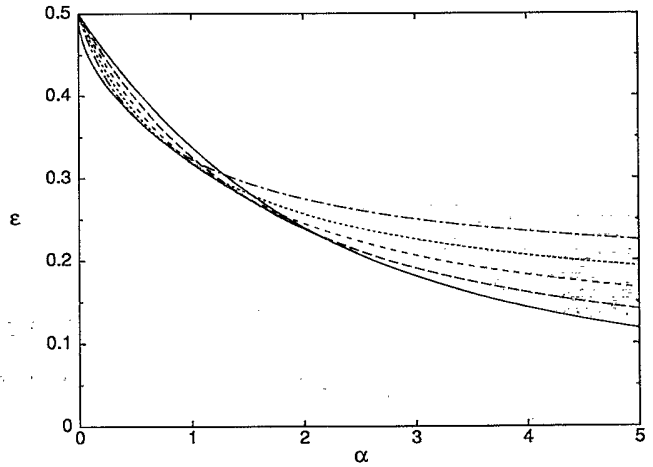

FIG. 4. The error $\epsilon(\alpha, Y)$ of generalization evaluated on unrestricted examples, after teaching αN examples which are restricted by the condition $\left| N^{-1/2} \sum_{j=1}^{N} B_j \xi_j^\mu \right| > Y$.

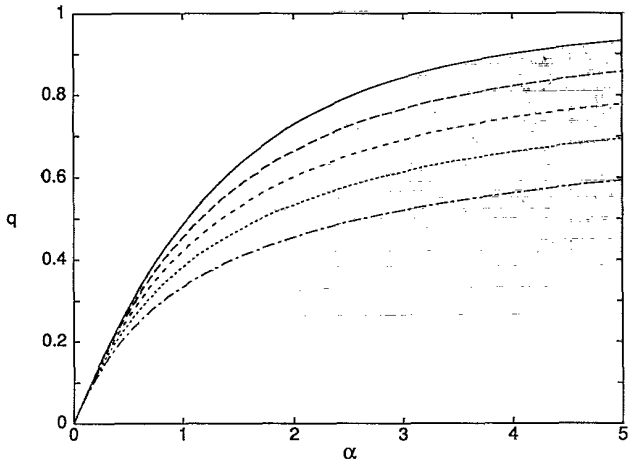

FIG. 5. The Edwards-Anderson parameter $q(\alpha, Y)$ measuring the average overlap between two different replicas.

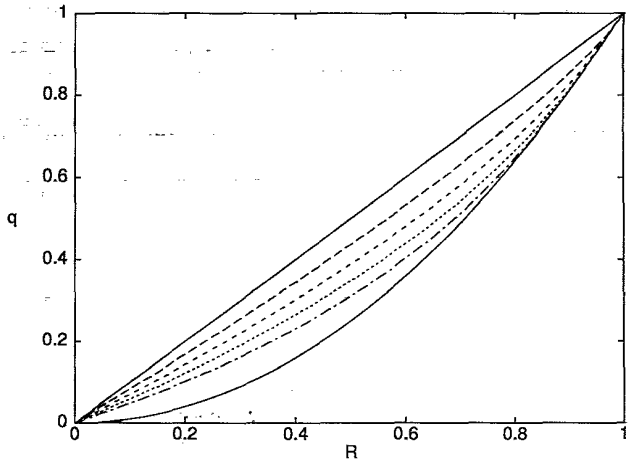

FIG. 7. The student-student overlap q as a function of the student-teacher overlap R, for various values of the exclusion threshold Y. Upper solid line, $Y = 0$; lower solid line, $Y \to -\infty$.

growing α the advantage turns into a disadvantage. That feature is further discussed in Sec. IV.

Neither the Edwards-Anderson parameter q (Fig. 5) nor the entropy s (Fig. 6) shows the crossing behavior: both the growth of q towards 1 and the approach of s to $-\infty$ with growing α become slower with the introduction of a positive Y.

To understand the reason why R and ϵ behave differently from q and s, it is instructive to consider q as a function of R (Fig. 7). In the original model with $f(y) \equiv 1$ ($Y = 0$) one has $q = R$. For a finite value of Y this is modified to

$$q = 2\, e^{\frac{Y^2}{2}} H(Y)\, R + O(R^2) \approx \frac{2}{\pi}\, \alpha \tag{16}$$

[cf. Eq. (15)] which holds for small R and crosses over for $R \to 1$ into

$$q \approx R^2, \tag{17}$$

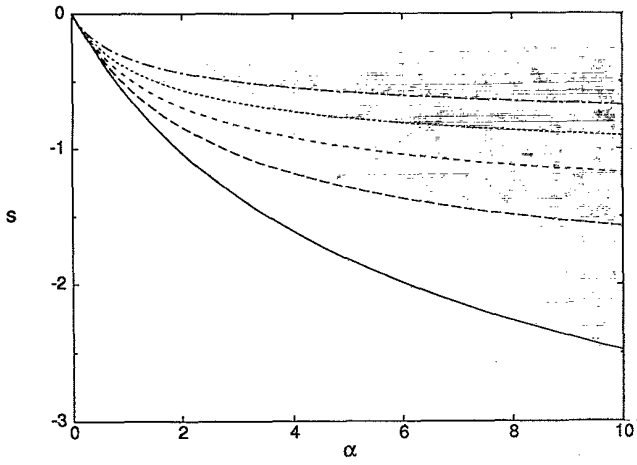

FIG. 6. The change of the entropy s with learning, for different values of Y.

the sooner the higher Y is chosen. In particular, for $Y \to \infty$ the quadratic dependence of Eq. (17) holds in the whole range of R.

The above results can be given a geometrical interpretation by considering the decomposition

$$q_{ab} = N^{-1}\langle \vec{J}^a_{\parallel} \cdot \vec{J}^b_{\parallel} \rangle + N^{-1}\langle \vec{J}^a_{\perp} \cdot \vec{J}^b_{\perp} \rangle, \tag{18}$$

where $\vec{J}^a_{\parallel}$ and $\vec{J}^a_{\perp}$ are the projections of $\vec{J}^a$ parallel and orthogonal to $\vec{B}$, respectively. The first longitudinal term of Eq. (18) is equal to R^2 by definition. Therefore, apparently $q = R^2$ corresponds to a version space with axial symmetry around the teacher's connection vector $\vec{B}$, in which the second transverse term of Eq. (18) vanishes. In the original situation with $Y = 0$, quenched fluctuations of the actual choice of the training set strongly break that axial symmetry, increasing q by $R - R^2$ due to the nonvanishing transverse term. It is that kind of symmetry-breaking fluctuation which is gradually suppressed by increasing Y or by increasing α with $Y \neq 0$.

The insight emerging from the above reasoning is this: learning more and more examples has a double effect. First, the version space shrinks, which is measured by the growth of q and the decrease of s; second, on the average it becomes more and more parallel to B, which is reflected on the growth of R and the decrease of ϵ. At the initial stage of the learning process, while R is small, hard examples contribute efficiently only to the first effect, whereas the practical aim is connected to the second. That is cured by the introduction of a positive Y, excluding those examples which are not immediately useful at the initial stage of learning. In that way one obtains faster initial learning, although slower shrinking of the version space.

As usual, the asymptotic behavior of the error of generalization for $\alpha \to \infty$ and $Y \neq 0$ can be extracted from Eqs. (10)–(12), as described in detail in Appendix A.

Unlike in the original case [3], if hard examples are excluded according to Eq. (5) with $Y > 0$, in the asymptotic range the error probability approaches zero as

$$\epsilon \approx Y(2\pi^2 \ln \alpha)^{-1/2} \quad \text{for } \alpha \to \infty. \tag{19}$$

We see that the rule can be fully taught through the restricted training set; however, hard examples would become more and more beneficial as learning advances; in need of them, convergence to zero error becomes logarithmically slow.

We give here also the asymptotics of the alternative error measure (14) $\tilde{\epsilon}(\alpha)$ evaluated over examples outside of the excluded band

$$\tilde{\epsilon} \approx (2\alpha \ln \alpha)^{-1} \tag{20}$$

(see Appendix A), a result which does not depend on Y. It can be shown that this error measure does not show improvement due to programmed learning for any α.

Let us mention that the situation handled here is different from the case of Parrondo and Van den Broeck [17], who consider teaching an unrestricted set of examples, however evaluate the generalization error by excluding the hardest ones — in that case, defined by the student, not by the teacher. In our case the restriction pertains basically to teaching, not to testing.

The replica symmetric solution is stable over the whole range of α, as expected from Gardner's argument [10], implying here that a learnable task is implemented by the simple perceptron on a connected region of the phase space of connections, excluding the formation of different thermodynamical states. Our model satisfies that expectation. Excluding hard examples makes error-free processing easier and indeed the stability parameter Γ (see Appendix B) grows with increasing Y for all values of α (Fig. 8).

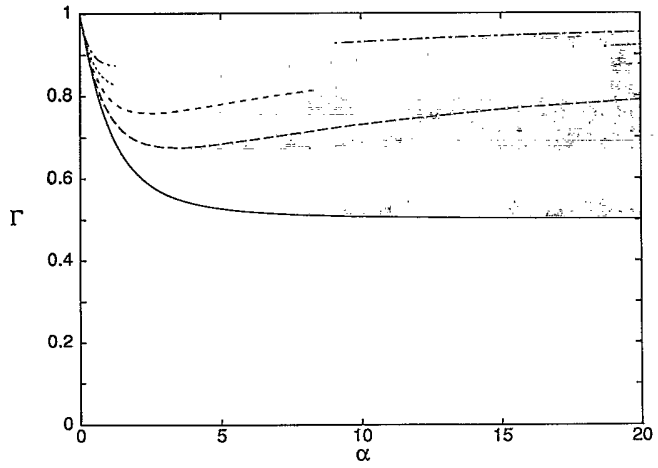

FIG. 8. The replica stability parameter Γ. Its vanishing — not the case here — would indicate instability of the replica symmetric solution.

III. RESIDUAL ENTROPY

In this section we consider a reinterpretation of the above model. First, the distribution of examples is taken to be pure Gaussian [$f(y) \equiv 1$ in Eq. (4)]. On the other hand, Y enters through the error message policy: outside of the equatorial band of width $2Y$ (called the "band of tolerance" in this section) the usual error message is taken, whereas inside that band the error is zero. In other words, within a range of input questions — in the present case within the band of tolerance — any answer is a good answer, whereas for the rest of inputs a well-defined output is required. Combined rules of that kind are further discussed in Sec. V.

The problem of fully imitating a teacher perceptron is the hardest task in the sense of the Carnevali-Patarnello approach [8]: it can be solved in a single way, by adjusting the connection strengths of the student perceptron exactly to the values the teacher has. That corresponds to zero residual entropy in the case of binary connection strengths and to an entropy of negative infinity for a continuous but spherically normalized vector of connection strengths.

The introduction of a band of tolerance into the rule does not change the structure of the student, who still classifies outputs according to Eq. (1). One might expect that in that case the student has a residual freedom in choosing its connection vector $\vec{J}$, with a corresponding residual entropy. Contrary to that expectation, we find that for $\alpha \to \infty$, like in the no-tolerance case [3], one has $q = R = 1$, which entails that $s \to -\infty$: our model has no *extensive* residual entropy.

The mathematical reason is clear: the radius of the base of the conelike region available for $\vec{J}$ after learning the restricted training set is $Y = O(1)$ with respect to N; therefore for large but finite N the residual entropy [cf. Eq. (2)] can be estimated as

$$s \approx N^{-1} \ln\left(\frac{Y^N}{\sqrt{N}^N}\right) = \ln Y - \frac{1}{2}\ln N, \tag{21}$$

approaching $-\infty$ as $N \to \infty$, in accordance with the replica result. Nevertheless, for large but finite N the above expression is a well-defined Carnevali-Patarnello entropy of a rule with a band of tolerance.

An extensive residual entropy is obtained by assigning an N dependence to the tolerance threshold, writing $Y\sqrt{N}$ instead of Y. Then the above diverging term drops out and the residual entropy remains finite. In that case, however, the Gaussian approximation, fundamental to all applications of the replica technique, breaks down since the Gaussian centers of the random variables appearing in the Fourier representation are cut out. An unattractive property of that variant is that the examples remaining outside of the band of tolerance, and therefore still carrying information, would be restricted to a vanishing part of the hypershpere for $N \to \infty$.

Turning back to the case of finite N, a related feature of our model is the fact that the student whose possibilities are limited to choosing a connection vector $\vec{J}$

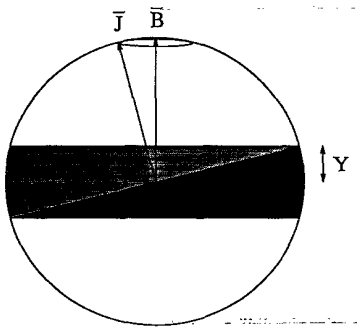

FIG. 9. The creation of pseudorules on learning with a threshold of tolerance: the student assigns a definite classification to inputs about which the teacher has no definite opinion.

cannot implement the freedom contained in the rule. Instead, as shown in Fig. 9, one stops learning on reaching a connection strength vector $\vec{J}$ on the nappe of a cone around the teacher's connection strength vector $\vec{B}$. That state of mind of the student breaks the rotation symmetry around $\vec{B}$ in such a way that the student will give a definite classification to any input vector, including those in the band of tolerance. By that the student has created a *pseudorule* extending the range of action of the learned rule beyond its competence revealed by the teacher and restricting the freedom of decision admitted by the teacher's tolerance. It is a rather common experience that such pseudorules are created as a consequence of low-level learning.

While the volume of the residual version space vanishes in the thermodynamical limit, signatures of the finite residuum for finite N can be seen within the replica framework. First, the entropy has a slower divergence to $-\infty$ for increasing Y (see Fig. 6), reflecting the tendency in Eq. (21). Second, the order parameter q decreases, demonstrating the increase of the version space volume. Those effects are generally associated with an increase in the stability parameter Γ of the replica symmetric solution and indeed that happens for increasing Y; see Fig. 8.

IV. PROGRAMMED LEARNING

From the point of view of example selection by the teacher the simple perceptron can behave in an extreme manner. If just the "easiest" example, the one parallel to $\vec{B}$, is presented, then both in one-shot Hebbian learning and in Bayesian learning one arrives at perfect generalization. For the Hebbian case that can be easily seen from the construction [20]. Bayesian learning gives a connection vector $\vec{J}$ pointing to the center of mass of the version space, which after presenting the easiest example covers the hemisphere with $\vec{B}$ as its rotation axis.

Boltzmannian learning [3] is different in that respect: aimed at just reaching the margin of the version space, it is far from guessing the rule from a single example. At the beginning of learning, $\vec{J}$ is typically in a narrow equatorial band of directions rather orthogonal to $\vec{B}$. If examples are taken from the same band, they only rotate $\vec{J}$ around $\vec{B}$ instead of turning it closer to $\vec{B}$. Therefore it is advantageous indeed to exclude such examples at the initial stage of learning and include them but gradually, as learning proceeds. That tendency is expressed in the crossing curves of Figs. 3 and 4.

To find an optimal schedule for such programmed learning is a harder task that is beyond the scope of this paper. Here we choose the simpler case of fixing the total relative number α of examples and looking for the optimal width $Y = Y_0(\alpha)$ of a band around the decision surface of the teacher — the hardest examples — that should be excluded from learning, keeping the distribution of active examples uniform on the remaining part of the unit cube. In the Gaussian representation described in Sec. II that corresponds to the weight function given by Eq. (5). Of course, as emphasized above, the generalization error is evaluated on the full unrestricted set of examples.

Our aim is now to find the optimum $Y_0(\alpha)$ that gives the largest value of R and consequently the minimum error of generalization. That can be done by solving the saddle-point equations (11) and (12) with weight function (5) for various values of Y and using numerical interpolation to find the value giving the largest R.

The results are presented in Fig. 10. Below a threshold $\alpha_c = 5.7$, $R(\alpha, Y)$ develops a single maximum as a function of Y at some $Y_0(\alpha) > 0$. Since $R(\alpha, Y_0(\alpha)) > R(\alpha, 0)$, programmed learning achieves better generalization. For $\alpha > \alpha_c$ we have $\left[\partial R(\alpha, Y)/\partial Y\right]_{Y=0} < 0$ and the global maximum of R in terms of Y is at $Y_0 = 0$: the teacher should use all examples if one can teach more than $\alpha_c N$ of them. In that restricted choice of teaching schedules therefore the exclusion of hard examples is ad-

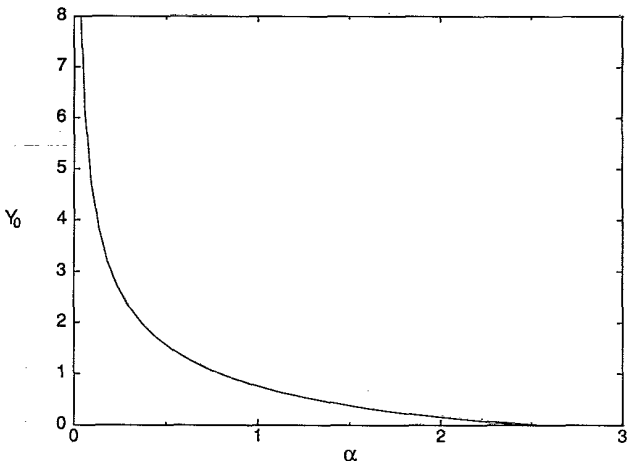

FIG. 10. The optimum value $Y_0(\alpha)$ yielding minimum error of generalization. On the scale of the figure, Y_0 is indistinguishable from zero already at $\alpha \approx 3$; however, an accurate numerical evaluation shows that the curve intercepts the α axis at $\alpha_c = 5.7$ (beyond the range of the figure).

vantageous only if learning can utilize but a restricted number of examples.

V. DISCUSSION

Excluding a band of hard examples from teaching can be supported by various human attitudes on behalf of the teacher. Indeed, beside the active teacher who aims at optimizing the efficiency of teaching by selecting easy examples, one can think of a teacher who also includes hard questions but either (i) accepts any answer to them from the student; or (ii) does not accept any answer to them from the student. Obviously both of them agree with the active teacher in giving no information to the student about the correct answer to the hard questions nor — in a broader context —about whether a single correct answer exists at all. Indeed if it does not exist, their attitude (full freedom or full rejection) is part of the rule which is then no simple input-to-output fuctional relationship anymore. In real-life applications such cases abound; an obvious example is the freedom present in grammar of human languages.

From the technical viewpoint, case (i), when the teacher accepts any answer to hard questions, is the same as the situation discussed in the previous sections; however, in (i) more examples are used up to achieve the same generalization performance. Namely, in all formulas and diagrams of Secs. II – IV, α is devaluated and should be replaced by $2H(Y)\alpha$.

The situation is similar for case (ii), when the teacher rejects any answer to hard questions; however, then the entropy formula Eq. (2) gives $-\infty$ for any α because of the appearance of an extensive positive training error $\epsilon_t > 0$, implying a positive ground state energy per synapsis. The singularity can be eliminated through the inclusion of thermal noise in the model. The ground state entropy is then $s = \lim_{\beta\to\infty}\beta(\epsilon_t - f)$, when we recover Eq. (9) with α scaled by $2H(Y)$. Thus the saddle point values of the observables R and q are the same as in case (i), and so is the generalization error.

Selecting examples according to difficulty by the teacher who knows the rule is only one of the possible contexts which leave residual freedom to the student after learning the task. Another one, of greater practical significance for network learning, although probably not for a single neuron, is the case when examples are presented on a fraction of the input channels only, leaving complete freedom in responding to stimuli arriving through the remaining channels, referring to some other subject.

Multidimensionality of the space of examples is an important feature of the learning process, which gives much weight to hard examples of little use at the initial stage of learning. Our approach, unlike most of the applications of machine learning, addresses conscious teachers. We hope that if extended to more realistic layered networks to gain more experience about less symmetric cases, it can be developed into a way of statistically analyzing the efficiency of teaching in classroom situations and developing strategies for teaching.

ACKNOWLEDGMENTS

We are indebted to C. Van den Broeck, Manfred Opper, and Sebastian Seung for enlightening discussions during the meeting on Statistical Mechanics of Generalization (Alden Biesen, Belgium). One of the authors (T. G.) is indebted to M. A. Virasoro for kindly calling his attention to Ref. [19]. This work has been partially supported by the Hungarian Research Foundation (Grants Nos. OTKA I/3-2179 and F-4489) and the PHARE program (Grant No. H 9112-0378). One of the authors (G. G.) is grateful for hospitality at Forschungszentrum Jülich, where part of this work was done, and for support by the Hungarian-German research exchange program (Grant No. OMFB 62).

APPENDIX A: ASYMPTOTIC ESTIMATES

In order to derive the asymptotic estimates (19) and (20) for the generalization error, we first recall that axial symmetry about the teacher vector is characterized by the relation

$$q = R^2 \tag{A1}$$

while the equilibrium solution satisfies

$$q = R. \tag{A2}$$

The values obtained from numerical solution for any fixed $Y > 0$ and presented in Fig. 7 behave similarly to Eq. (A2) for small α [cf. Eq. (16)] but asymptotically approach Eq. (A1) for $\alpha \to \infty$, so that relationship is correct for the high-α asymptotic region. Then using also $R \approx 1$, Eq. (12) reduces to

$$\sqrt{1-R^2} = \frac{\alpha}{2H(Y)}\sqrt{\frac{2}{\pi}}\int_Y^\infty Dy\, y \times \exp\left[-\frac{1}{2}\left(\frac{y}{\sqrt{1-R^2}}\right)^2\right]. \tag{A3}$$

Since $R \approx 1$, the exponential factor cuts the Gaussian measure short and one can write $Dy \approx dy/\sqrt{2\pi}$. Then evaluating the integral we have

$$\frac{\alpha}{2H(Y)}\sqrt{\frac{2}{\pi}}\sqrt{1-R^2}\exp\left[-\frac{1}{2}\left(\frac{Y}{\sqrt{1-R^2}}\right)^2\right] = 1. \tag{A4}$$

Taking the logarithm and selecting the term most strongly diverging with $R \to 1$, we obtain

$$\ln\alpha = \frac{Y^2}{2(1-R^2)}, \tag{A5}$$

which, using $\epsilon \approx \pi^{-1}\sqrt{1-R^2}$ valid in the present limit, gives Eq. (19) of the main text.

To obtain Eq. (20) for the error evaluated on taught examples only, the same estimates should be carried out on Eq. (14). Partial integration gives

$$\tilde{\epsilon} = \frac{1}{2H(Y)}\sqrt{\frac{2}{\pi}}\sqrt{1-R^2}\frac{\exp\left[-\frac{1}{2}\left(\frac{Y}{\sqrt{1-R^2}}\right)^2\right]}{\left(\frac{Y}{\sqrt{1-R^2}}\right)^2}. \quad \text{(A6)}$$

Using Eqs. (A4) and (A5), one obtains Eq. (20) of the main text.

APPENDIX B: REPLICA STABILITY

The stability of the replica symmetric solution with respect to symmetry breaking perturbations — replicon modes — can be determined by using the arguments of de Almeida and Thouless [21] and Gardner [10]. The calculation follows Ref. [3], with minor modifications, resulting in the stability parameter

$$\Gamma = 1 - 2\alpha \int_0^\infty Dy\, f(y) \int_{-\infty}^\infty Dt \left(\frac{\partial^2 \ln H(z)}{\partial z^2}\right)^2, \quad \text{(B1)}$$

where z is given by Eq. (10). Γ is the product of the only two potentially dangerous eigenvalues, those of the replicon modes. For $\alpha \to 0$ the replica symmetric solution is stable, both eigenvalues are positive, and so destabilization occurs at the smallest α where $\Gamma = 0$.

In the canonic case $Y = 0$, it can be shown that $1 \geq \Gamma \geq 1/2$ and the lower limit is reached for $\alpha \to \infty$. We observe the inequality $\Gamma(\alpha, Y) > \Gamma(\alpha, Y')$ if $Y > Y'$, so the replica symmetric solution is stable for all $Y \geq 0$ and $\alpha \geq 0$. Those features are illustrated by Fig. 8.

[1] T.L.H. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys. **65**, 499 (1993).
[2] J.S. Denker, D. Schwartz, B. Wittner, S.A. Solla, R.E. Howard, L.D. Jackel, and J.J. Hopfield, Complex Syst. **1**, 877 (1987).
[3] G. Györgyi and N. Tishby, in *Neural Networks and Spin Glasses,* edited by W.K. Theumann and R. Köberle (World Scientific, Singapore, 1990).
[4] G. Györgyi, Phys. Rev. Lett. **64**, 2957 (1990).
[5] H.S. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992).
[6] S. Bös, W. Kinzel, and M. Opper, Phys. Rev. E **47**, 1384 (1993).
[7] T.L.H. Watkin and A. Rau, Phys. Rev. A **45**, 4102 (1992).
[8] P. Carnevali and S. Patarnello, Europhys. Lett. **4**, 1199 (1987).
[9] C. Van Den Broeck and R. Kawai, Phys. Rev. A **42**, 6210 (1990).
[10] E. Gardner, J. Phys. A **21**, 257 (1988).
[11] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).
[12] E. Gardner and B. Derrida, J. Phys. A **22**, 1983 (1989).
[13] F. Vallet, Europhys, Lett. **8**, 747 (1989).
[14] W. Kinzel and P. Ruján, Europhys. Lett. **13**, 473 (1990).
[15] T.L.H. Watkin and A. Rau, J. Phys. A **25**, 113 (1992).
[16] J.A. Hertz, A. Krogh, and R.G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, CA, 1991).
[17] That observation has been kindly made by C. Van den Broeck.
[18] S. Kirkpatrick and D. Sherrington, Phys. Rev. B **17**, 4384 (1978).
[19] S. Franz, G. Parisi, and M.A. Virasoro, J. Phys. (Paris) **2**, 1869 (1992).
[20] J.M.R. Parrondo and C. Van den Broeck, Europhys. Lett. **22**, 319 (1993).
[21] J.R. de Almeida and D.J. Thouless, J. Phys. A **11**, 983 (1978).