Optimal pruning in neural networks

Daniela M. L. Barbato*
Universidade Paulista, Avenida Comendador Enzo Ferrari 280, CEP 13043-055, Campinas, SP, Brazil

Osame Kinouchi
Departamento de Física e Matemática, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Avenida dos Bandeirantes 3900, CEP 14040-901, Ribeirão Preto, SP, Brazil
(Received 24 January 2000; revised manuscript received 12 June 2000)

We study pruning strategies in simple perceptrons subjected to supervised learning. Our analytical results, obtained through the statistical mechanics approach to learning theory, are independent of the learning algorithm used in the training process. We calculate the post-training distribution $\mathcal{P}(J)$ of synaptic weights, which depends only on the overlap ρ_0 achieved by the learning algorithm before pruning and the fraction κ of relevant weights in the teacher network. From this distribution, we calculate the optimal pruning strategy for deleting small weights. The optimal pruning threshold grows from zero as $\theta_{opt}(\rho_0,\kappa) \propto [\rho_0 - \rho_c(\kappa)]^{1/2}$ above some critical value $\rho_c(\kappa)$. Thus, the elimination of weak synapses enhances the network performance only after a critical learning period. Possible implications for biological pruning phenomena are discussed.

PACS number(s): 87.18.Sn, 05.20.−y, 87.10.+e, 07.05.Mh

I. INTRODUCTION

A very common but poorly understood developmental phenomenon, found widespread in vertebrate brains is the initial overproduction of neurons and synapses with posterior elimination of a large amount of these elements [1]. There is increasing evidence that, instead of being a simple maturational epiphenomenon, this ''pruning'' process has indeed developmental significance, being a selective/competitive procedure that eliminates the weaker synapses. Recent evidence points to the view that this selection process is not done over a random pool of synapses: in the experiments, during a previous developmental period, synapses both increase (LTP) and decrease (LTD) due to Hebbian/correlational mechanisms [2,3]. There is also evidence that the level of pruning experienced by different brain regions is not preprogrammed but reflects the variability and complexity of the environmental input to those areas [4]. Since the selection mechanism apparently works on directed, nonrandom variation, we refer to this view of the pruning process as a *selective trophism scenario* to differentiate it from the pure neural Darwinist account inspired by nondirected selection theories of the immune response [4,5].

A very common problem in statistical inference tasks is that simpler (smoother) functions, with fewer parameters, have better interpolation and extrapolation properties but are at risk of being too simple to reliably approximate the target function. Since one of the supposed roles of cortical networks is to provide similar input-output mappings, this kind of problem could also arise in the biological context. The conjectured role of biological pruning is to solve this problem by allowing the network architecture to be defined *a posteriori*, after some information about the needed complexity has been gathered.

Here we consider a simple learning process in formal neural networks where all these ingredients are present. We study the computational effect of pruning the weaker synapses in a single-layer perceptron. We find that pruning works only after clear differentiation between weak and strong synapses, a differentiation induced by learning. We find that the optimal pruning criteria (''optimal elimination threshold'') should not be fixed but must be time dependent, or, better, performance dependent. The optimal pruning threshold also depends on the complexity of the function to be implemented: networks that implement complex, multifactorial functions must be pruned only after a lot of learning.

It must be clear that, although inspired by biological experiments, we are not modeling some specific experimental situation. Our approach, instead, is to implement in a concrete although simplified way the selective trophism scenario, looking for possibly generic, robust features of the pruning process that are certainly present in machine learning and that could also be present in biological learning.

The computational advantage of pruning has been studied in artificial neural networks through numerical simulations (for a review, see [6]). However, up to now, there are few analytical results concerning pruning strategies. Some previous studies have concluded that pruning has a deleterious effect for the network performance. For example, in associative memory (Hopfield) networks, it has been shown that pruning always degrades the quality of memory retrieval, although it can save costly synapses [7]. In the case of feedforward networks, analytical studies have been done only for the single-layer perceptron, mainly for the capacity problem [8]. In this case also pruning is viewed as a mechanism to save synapses, at the cost of reducing the stability of the learned patterns.

There are also some results for the *teacher-student* scenario of supervised learning. The teacher-student scenario is a well studied paradigm for function approximation where the mapping to be implemented by a network (the student) is

*Present address: Faculdades COC, R. Abrao Issa Halack 980, CEP 14096-160, Ribeiro Preto, SP, Brazil.

defined by another network (the teacher), which may represent the regularities of the environment [10,11]. In this case, the natural performance measure is the teacher-student overlap. For example, Barbato and Fontanari have shown that pruning of trained networks always decreases the achieved teacher-student overlap if the distribution of teacher weights has the usual Gaussian form [9].

Pruning improves the performance in the teacher-student scenario if we consider the proper class of target functions. Here we consider the case where there exists only a fraction $\kappa<1$ of relevant weights in the target function (in single-layer perceptrons this is equivalent to the existence of a fraction $1-\kappa$ of irrelevant node inputs). We argue that this situation is much more common in real world problems than the $\kappa=1$ case. A practical instance where this happens is in the problem of sex classification based on a face recognition task, where most of the input pixels are irrelevant to the task [12]. The learning scenario with irrelevant teacher weights has been addressed previously by Kuhlmann and Müller [13], who studied a particular learning algorithm (the *maximum stability* perceptron). Here we extend their results by showing that the pruning performance depends only on the teacher complexity κ and the student-teacher overlap ρ_0 achieved before pruning, and can be studied without reference to learning algorithms.

The paper is organized as follows. In Sec. II, we present the learning scenario to be studied. Section III contains the derivation of the distribution of student weights as a function of prior teacher-student overlap ρ_0 and the parameter κ of the teacher weight distribution. This distribution does not depend on the learning algorithm used, which influences only the evolution of the overlap ρ_0 as a function of the number of examples. In that section, we also derive the optimal pruning strategy and present the phase diagram in the ρ_0 versus κ plane that shows the regime where pruning improves generalization ability. Then, as an example, we present in Sec. IV simulations of optimal pruning for simple Hebbian learning which confirm our analytical results. In Sec. V, we discuss the possible relevance of our results to biological pruning. We offer some conclusions in the final section.

II. THE LEARNING SCENARIO

A. The teacher-student learning problem

In the teacher-student learning scenario the function to be approximated is represented by a given neural network (the teacher or target network). Another network (the student) tries to infer or approximate the parameters of the target by using only the information contained on a set of input-output pairs (*examples*). Here, both the teacher and the student are single-layer perceptrons with N inputs and a single scalar output. This case has been extensively studied as a paradigmatic scenario in the statistical physics approach to learning theory [10,11].

More pointedly, the neural network we consider in this paper consists of N binary input units $S_i=\pm 1$ $(i=1,\ldots,N)$ coupled to a single Boolean output unit σ through a set of real-valued synaptic weights J_i $(i=1,\ldots,N)$ according to the equation

$$\sigma=\mathrm{sgn}\left(\sum_{i=1}^{N} J_i S_i\right). \tag{1}$$

We observe that our results depend only on the first and second moments of the input distribution and will be valid also for real-valued inputs with zero centered Gaussian distributions.

The task of the student perceptron is to realize the mapping between the 2^N possible input configurations $\{\mathbf{S}\}$ and their respective outputs $\{\tau\}$ generated by the teacher perceptron:

$$\tau=\mathrm{sgn}\left(\sum_{i=1}^{N} B_i S_i\right), \tag{2}$$

where the weights B_i, $i=1,\ldots,N$, are statistically independent random variables distributed according to the probability distribution

$$\mathcal{P}_B(B)=\kappa_0\delta(B)+\kappa_1\delta(B-1)+\kappa_2\delta(B+1) \tag{3}$$

with $\kappa_0+\kappa_1+\kappa_2=1$. The motivation for choosing a teacher network with a fraction of null weights is to model the realistic situation in nature where most of the input components are completely irrelevant to the final outcome. In fact, exploring the effect of pruning in this more realistic setting is the main purpose of this paper.

Our results can be generalized in a straightforward manner to other teacher distributions since an arbitrary $\mathcal{P}_B(B)$ can be written as

$$\mathcal{P}_B(B)=\int_{-\infty}^{\infty}\kappa(x)\,\delta(B-x)\,dx \tag{4}$$

with an arbitrary density $\kappa(x)$. We expect that our conclusions will not change qualitatively if the teacher distribution continues to present a finite fraction κ_0 of zero weights.

To achieve its task, the student network is trained with $P=\alpha N$ examples, i.e., input-output pairs $(\mathbf{S}^\mu,\tau^\mu)$, $\mu=1,\ldots,P$, where τ^μ is the teacher's output to input $\mathbf{S}^\mu$ and each component S_i^μ is drawn independently from the probability distribution

$$\mathcal{P}_S(S_i^\mu)=\frac{1}{2}\,\delta(S_i^\mu-1)+\frac{1}{2}\,\delta(S_i^\mu+1). \tag{5}$$

In the present developmental context, it is better to interpret the teacher network not as an external supervisor but as representing an attractor state (the ''mature state''). The data or examples furnished by the teacher are supposed to be encoded, in a distributed way, in the genoma-environment interactions: the environment furnishes the possible inputs with a distribution $\mathcal{P}_S(\mathbf{S}^\mu)$ and an intrinsic recompense system furnishes the desired (teacher) outputs τ^μ. In other words, the presumed genetic information corresponds to the teacher outputs, but the actual teacher parameters (its architecture and weights) are not present. The teacher network represents an ideal or prototypical mature state partially realized by the student after the realization of the learning/developmental process. Here, development is thought of as the *transient* dynamical evolution of the immature network toward the

mature one, which constitutes a long lived metastable state (the true stable state is the dead one). So the natural measure of this process is the teacher-student overlap ρ (to be defined later), which may tend to but not achieve the ideal value $\rho=1$.

B. The statistical learning process

In the statistical physics approach to learning theory [10], the learning process is viewed as a search for the global minimum of a certain cost function, termed the training energy, usually assuming the form

$$E_{\mathcal{L}}(\mathbf{J})=\sum_{\mu=1}^{P} V(\lambda^{\mu}), \tag{6}$$

where $\lambda^{\mu}=\tau^{\mu}\mathbf{J}\cdot\mathbf{S}^{\mu}/\sqrt{N}$ is the *stability* of example μ and the potential $V(\lambda)$ defines the specific (training) algorithm used to explore the space of weights. We note that the stability λ^{μ} is positive only if the input $\mathbf{S}^{\mu}$ is associated with the correct output, namely, τ^{μ}. The diverse potentials $V(\lambda)$ proposed in the literature realize different ways of penalizing student vector that produce negative stabilities.

We start by considering the space of all networks with training energies $E_{\mathcal{L}}(\mathbf{J})$ subject to a stochastic minimization learning process under a spherical constraint in the weights. Note that this minimization is done with the set of variables $\mathcal{L}=\{S_i^{\mu},B\}$ quenched. This defines a post-training probability distribution on this space of networks, given by the canonical (Gibbs) distribution with the stochastic parameter (*temperature*) $T=1/\beta$,

$$\mathcal{P}(\mathbf{J}|\mathcal{L})=\frac{1}{Z_{\mathcal{L}}}\exp[-\beta E_{\mathcal{L}}(\mathbf{J})], \tag{7}$$

where $Z_{\mathcal{L}}$ is the partition function

$$Z_{\mathcal{L}}=\int_{-\infty}^{\infty} d\mu(\mathbf{J})\exp[-\beta E_{\mathcal{L}}(\mathbf{J})] \tag{8}$$

with $d\mu(\mathbf{J})=(2\pi eQ)^{-N/2}\Pi_i^N dJ_i\,\delta(NQ-\Sigma_i J_i^2)$ being the normalized prior student distribution with a spherical constraint (for details, see [10]).

C. Performance measures

The ultimate goal of the learning process is to produce a network capable of realizing an example not belonging to the training set. To measure this capability we introduce the generalization function

$$e_g(\mathbf{J},\mathbf{B})=\int d\mathbf{S}\,\mathcal{P}_S(\mathbf{S})\Theta(-\tau(\mathbf{B},\mathbf{S})\sigma(\mathbf{J},\mathbf{S})), \tag{9}$$

where $d\mathbf{S}\,P_S(\mathbf{S})\equiv\Pi_i dS_i\,\mathcal{P}_S(S_i)$ is the measure in the input space and $\Theta(x)$ is the Heaviside (step) function. Here σ and τ are the student's and teacher's outputs, respectively, to input $\mathbf{S}$. In the thermodynamic limit $N\to\infty$ the integration in Eq. (9) can be readily carried out, yielding [10]

$$e_g(\mathbf{J},\mathbf{B})=\frac{1}{\pi}\arccos\rho_0(\mathbf{J},\mathbf{B}),$$

$$\rho_0(\mathbf{J},\mathbf{B})\equiv\frac{R(\mathbf{J},\mathbf{B})}{\sqrt{Q(\mathbf{J})M(\mathbf{B})}}, \tag{10}$$

where $Q(\mathbf{J})=(1/N)\Sigma_i^N J_i^2$ is the squared norm of the student perceptron, $R(\mathbf{J},\mathbf{B})=(1/N)\Sigma_i^N B_i J_i$ is the overlap between student and teacher networks, and $M(\mathbf{B})=(1/N)\Sigma_i^N B_i^2$ is the squared norm of the teacher perceptron. We note that, in the thermodynamic limit, use of self-averaging yields $M=\kappa_1+\kappa_2=\kappa$, that is, the teacher norm equals the fraction κ of nonzero weights for our choice of $\mathcal{P}_B(B)$.

The relevant quantity in the statistical approach is the *average generalization error*

$$e_g(\alpha)=\langle\langle\langle e_g(\mathbf{J},\mathbf{B})\rangle_T\rangle\rangle, \tag{11}$$

where $\langle\cdots\rangle_T$ denotes the average over the post-training distribution $\mathcal{P}(\mathbf{J}|\mathcal{L})$ (the *thermal average*) and $\langle\langle\cdots\rangle\rangle$ stands for a *quenched average* over the random variables $\mathcal{L}=\{S_i^{\mu},B_i\}$. We note that after these averages are taken the generalization error depends only on the relative number of examples α.

Since the generalization error is a monotonic function of the teacher-student overlap, it is sometimes convenient to present the results as a function of the average overlap $\rho_0\equiv\langle\langle\langle\rho_0(\mathbf{J},\mathbf{B})\rangle_T\rangle\rangle$. In the following, we will call this overlap the *maturity of the network*.

III. ANALYTICAL RESULTS

A. Free energy

Following the standard prescription of taking averages over extensive quantities only [10] we define the average free energy density f by

$$-\beta f=\lim_{N\to\infty}\frac{1}{N}\langle\langle\ln Z_{\mathcal{L}}\rangle\rangle. \tag{12}$$

As usual, the quenched average can be calculated through the replica method, which consists of using the identity $\langle\langle\ln Z_{\mathcal{L}}\rangle\rangle=\lim_{n\to 0}n^{-1}\ln\langle\langle Z_{\mathcal{L}}^n\rangle\rangle$, evaluating $\langle\langle Z_{\mathcal{L}}^n\rangle\rangle$ for integer n, and then analytically continuing to $n=0$.

As the calculation of f in the thermodynamic limit is standard [10] and rather unilluminating, we present only the final result in replica symmetric (RS) approximation:

$$f=\mathcal{U}-T\mathcal{S} \tag{13}$$

with

$$\mathcal{S}=\frac{1}{2}\left[\ln\left(1-\frac{q}{Q}\right)+\frac{q-R^2/\kappa}{Q-q}\right], \tag{14}$$

$$\mathcal{U}=-2\frac{\alpha}{\beta}\int_{-\infty}^{\infty}Dz\int_{0}^{\infty}Dy\ln\int\frac{d\lambda e^{-\beta\mathcal{E}(\lambda)}}{\sqrt{2\pi(Q-q)}}, \tag{15}$$

$$\mathcal{E}(\lambda)=V(\lambda)+\frac{1}{2\beta(Q-q)}\left(\lambda-y\frac{R}{\sqrt{\kappa}}-z\sqrt{q-\frac{R^2}{\kappa}}\right)^2.$$

The physical order parameters in RS approximation,

$$q = q_{ab} = \frac{1}{N} \langle\langle \langle \mathbf{J}^a \cdot \mathbf{J}^b \rangle_T \rangle\rangle; \quad a < b, \tag{16}$$

$$R = R_a = \frac{1}{N} \langle\langle \langle \mathbf{J}^a \rangle_T \cdot \mathbf{B} \rangle\rangle, \tag{17}$$

measure the average overlap between two different solutions $\mathbf{J}^a$ and $\mathbf{J}^b$, and the average overlap between the typical student $\mathbf{J}^a$ and the teacher network $\mathbf{B}$, respectively. The saddle-point parameters (q,R) are obtained so as to extremize f. More specifically, due to the limit $n \to 0$, the parameter q maximizes the free energy, while R minimizes it, as usual. Notice that hitherto we have not specified the functional form of $V(\lambda)$.

B. Probability distribution of the weight entries

The equilibirum distribution $P(\mathbf{J}|\mathcal{L})$ minimizes the free energy, giving the probability of achieving a vector $\mathbf{J}$ after learning. Now our interest is to determine the probability distribution function that a given entry, say J_i, has the value J in the deterministic limit $T=0$. Clearly, this probability distribution is given by

$$\mathcal{P}_i(J) = \lim_{\beta \to \infty} \left\langle \left\langle \int d\mu(\mathbf{J}) \mathcal{P}(\mathbf{J}|\mathcal{L}) \, \delta(J_i - J) \right\rangle \right\rangle, \tag{18}$$

where $\mathcal{P}(\mathbf{J}|\mathcal{L})$ is the weight joint (Gibbs) probability distribution given by Eq. (7). Here the δ function guarantees that the entry J_i is not integrated out. Moreover, since this distribution is obviously independent of the particular chosen entry J_i, we can write $\mathcal{P}_i(J) = \mathcal{P}(J), \ \forall i$. This probability distribution can readily be evaluated by introducing an additional term to the energy function, $E_{aux}(\mathbf{J}) = E_{\mathcal{L}}(\mathbf{J}) + h \Sigma_i \delta(J_i - J)$, so that

$$\mathcal{P}(J) = - \lim_{\beta \to \infty} \frac{1}{\beta N} \frac{\partial}{\partial h} \langle\langle \ln Z_{aux} \rangle\rangle |_{h=0}, \tag{19}$$

where Z_{aux} is the partition function of Eq. (8) with $E_{\mathcal{L}}$ replaced by E_{aux}. The advantage of this formulation is that the procedure used to find the average free energy in the previous section can be readily applied to evaluate Eq. (19) since the new term $h \Sigma_i \delta(J_i - J)$ affects only the entropic term Eq. (14). The final result is a superposition of Gaussian distributions centered at the different values of $\rho_0 B$:

$$\mathcal{P}(J) = \int dB \, \mathcal{P}_B(B) \frac{e^{-(J - \rho_0 B)^2/(2\chi^2)}}{\sqrt{2\pi\chi^2}}, \tag{20}$$

with $\chi^2 = \kappa(1 - \rho_0^2)$ and

$$\rho_0 = \frac{R}{\sqrt{QM}} = \frac{R}{\kappa}. \tag{21}$$

Here, ρ_0 is the normalized overlap between the student and teacher perceptrons $(-1 < \rho_0 < 1)$. Furthermore, to facilitate the comparison between $\mathcal{P}(J)$ and $\mathcal{P}_B(B)$, we have chosen the norm of the student perceptron so as to coincide with the norm of the teacher perceptron, i.e., $Q = M = \kappa$. In particular, for the distribution $\mathcal{P}_B(B)$ given by Eq. (3), we get

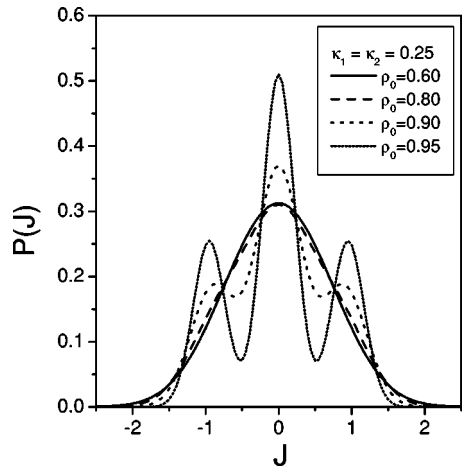

FIG. 1. Postlearning distribution of student weights $\mathcal{P}(J)$ as a function of the maturity ρ_0 given a teacher distribution with $\kappa_0 = 0.5$ and $\kappa_1 = \kappa_2 = 0.25$.

$$\mathcal{P}(J) = \frac{\kappa_0 e^{-J^2/2\chi^2}}{\sqrt{2\pi\chi^2}} + \frac{\kappa_1 e^{-(J-\rho_0)^2/2\chi^2}}{\sqrt{2\pi\chi^2}} + \frac{\kappa_2 e^{-(J+\rho_0)^2/2\chi^2}}{\sqrt{2\pi\chi^2}}. \tag{22}$$

Notice that the learning algorithm, i.e., the particular cost function $V(\lambda)$ used, does not appear explicitly in the expression for $\mathcal{P}(J)$. All information concerning the specific cost function is embodied in the value of the order parameter $\rho_0(\alpha)$. This means that ρ_0 can be viewed as an independent control parameter, whose physical realization may be achieved through a proper choice of the training algorithm as well as of the training set size α. In general, ρ_0 increases monotonically with α; in particular, $\rho_0 = 0$ for $\alpha = 0$ and $\rho_0 \to 1$ for $\alpha \to \infty$. However, we must note that for certain training tasks the regime of perfect learning $(\rho_0 = 1)$ may never be reached, even for infinite training set sizes. This occurs, for example, if the initial number of student weights is smaller than the number of teacher weights. This unrealizable case will not be considered here, since we are supposing that the synaptic overgrowth phase indeed leads to networks with initial synaptic number above that needed to perform the target function.

We show in Fig. 1 the dependence of $\mathcal{P}(J)$ on ρ_0 for $\kappa_0 = 0.5$, $\kappa_1 = 0.25$, and $\kappa_2 = 0.25$. Notice that the peaks around $J = 0$ and $J = \pm 1$ become more distinct as ρ_0 increases. In fact, $\mathcal{P}(J)$ reduces to the teacher probability distribution given in Eq. (3) in the limit $\rho_0 \to 1$. This result suggests that cutting weights with strength smaller than a certain threshold θ (i.e., with $|J| < \theta$) might be a good strategy to improve the generalization performance of the network.

C. Generalization error

Motivated by the past section we execute the pruning strategy, i.e., we cut off the weights that belong to the range $-\theta < J < \theta$. To implement this cutoff we introduce the prun-

ing function $\mathcal{F}^i_\theta = \Theta(|J_i| - \theta)$, so that the postpruning generalization error and the postpruning teacher-student overlap become [9]

$$e_g(\alpha,\theta) = \frac{1}{\pi}\arccos\rho(\rho_0(\alpha),\theta),$$

$$\rho(\rho_0,\theta) = \frac{S}{\sqrt{PM}},$$

$$P = \frac{1}{N}\sum_i \langle\langle\langle J_i^2 \mathcal{F}^i_\theta\rangle_T\rangle\rangle, \tag{23}$$

$$S = \frac{1}{N}\sum_i \langle\langle\langle B_i J_i \mathcal{F}^i_\theta\rangle_T\rangle\rangle.$$

Notice that the dependence on the number of examples has been expressed in terms of the achieved maturity $\rho_0(\alpha)$. The order parameter P is calculated as $P = -(1/\beta N)(\partial/\partial h)\langle\langle \ln Z_{aux}\rangle\rangle|_{h=0}$ where the partition function involves the effective energy $E_{aux}(\mathbf{J}) = E_{\mathcal{L}}(\mathbf{J}) + h\Sigma_i J_i^2 \mathcal{F}^i_\theta$. In order to evaluate P, we use the replica method and take $h=0$. The final result is, in the RS approximation,

$$P = (\kappa\chi^2 + \kappa\rho_0^2)[H(A)+H(B)] + 2(1-\kappa)\chi^2 H(C) + \frac{\kappa\chi^2}{\sqrt{2\pi}}\left(Be^{-A^2/2} + Ae^{-B^2/2} + \frac{2(1-\kappa)}{\kappa}Ce^{-C^2/2}\right), \tag{24}$$

where $H(x) = \int_x^\infty Dx$, and $A = (\theta-\rho_0)/\chi$, $B = (\theta+\rho_0)/\chi$, and $C = \theta/\chi$. Notice that κ_1 and κ_2 appear solely as the combination $\kappa = \kappa_1 + \kappa_2$. The parameter S is calculated in the same way, by changing $J^2\mathcal{F}_\theta$ to $JB\mathcal{F}_\theta$ in the expression for E_{aux}, giving

$$S = \kappa\rho_0[H(A)+H(B)] + \frac{\kappa\chi}{\sqrt{2\pi}}(e^{-A^2/2} - e^{-B^2/2}). \tag{25}$$

As was suggested in Fig. 1, in order to minimize the generalization error we cut off weights with $|J| < \theta$. In Fig. 2 we show, for $\kappa = 0.25$, the postpruning overlap $\rho(\rho_0,\theta)$ as a function of the pruning threshold θ and the prepruning overlap ρ_0. Notice that at some critical value $\rho_c(\kappa)$ of ρ_0 the function $\rho(\rho_0,\theta)$ starts to have a maximum, i.e., for maturity levels with $\rho_0 > \rho_c$ one has $\theta_{opt} > 0$.

D. Optimal pruning

To find the optimal value θ_{opt} that leads to the best pruning performance, we calculate the derivative $d\rho/d\theta = 0$, which is equivalent to

$$2P\frac{dS}{d\theta} = S\frac{dP}{d\theta}, \tag{26}$$

with

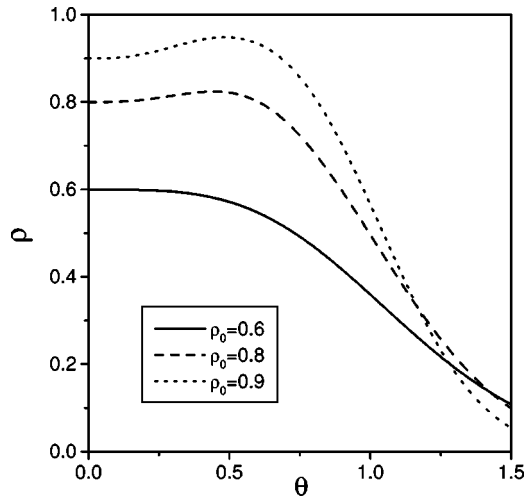

FIG. 2. The postpruning overlap $\rho(\theta)$ as a function of the pruning threshold θ, for different values of the maturity ρ_0 and a teacher distribution with $\kappa = 0.25$. Note that pruning improves the overlap ρ only after achieving a maturity level $\rho_0 > 0.64$ and that $\theta_{opt} \to 0.5$ when $\rho_0 \to 1$.

$$\frac{dP}{d\theta} = \frac{-\theta^2\kappa}{\sqrt{2\pi\chi^2}}(e^{-A^2/2} + e^{-B^2/2}) - \frac{2\chi(1-\kappa)C^2 e^{-C^2/2}}{\sqrt{2\pi}},$$

$$\frac{dS}{d\theta} = -\frac{\kappa\theta}{\sqrt{2\pi\chi^2}}(e^{-A^2/2} + e^{-B^2/2}). \tag{27}$$

Equation (26) has a solely numerical solution, so we illustrate the behavior of $\theta_{opt}(\rho_0,\kappa)$ through Fig. 3. As observed before, below some critical value $\rho_c(\kappa)$ for the prior overlap ρ_0, pruning always degrades the performance of the student network, so in this case the best choice is $\theta_{opt} = 0$. However, after the student has achieved some $\rho_0(\kappa) > \rho_c(\kappa)$, there exists a finite interval $\theta \in [0,\theta_{max}]$ that leads to $\rho(\theta) > \rho_0$, with some optimal pruning threshold $\theta_{opt}(\kappa)$ inside this interval.

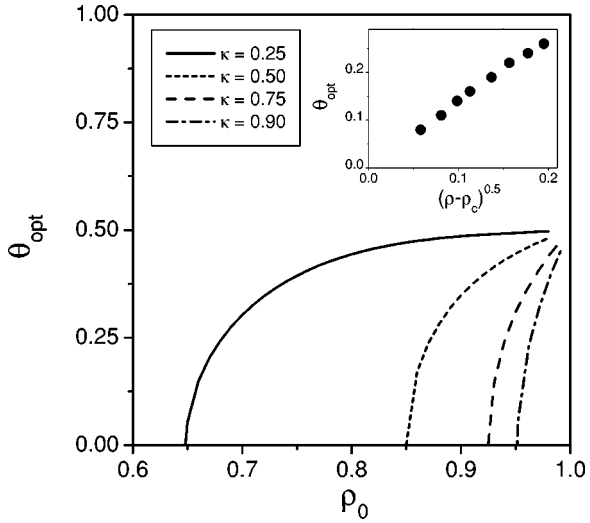

FIG. 3. The optimal pruning threshold θ_{opt} as a function of the maturity ρ_0 for different teacher complexity levels κ. Inset: Plot of θ_{opt} vs $(\rho-\rho_c)^{1/2}$ for $\kappa = 0.25$.

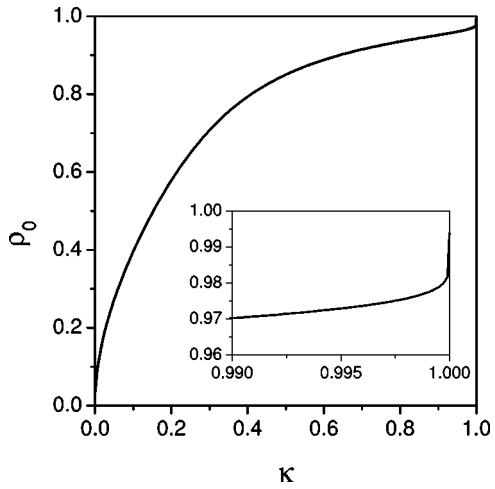

FIG. 4. Pruning phase diagram. Above the line $\rho_c(\kappa)$ pruning improves the final overlap ρ: $\rho(\theta_{opt})>\rho_0$. Inset: phase diagram near $\kappa=1$. Note that $\rho_c(\kappa)$ goes to $\rho_c(1)=1$ continuously but in a highly nonlinear manner.

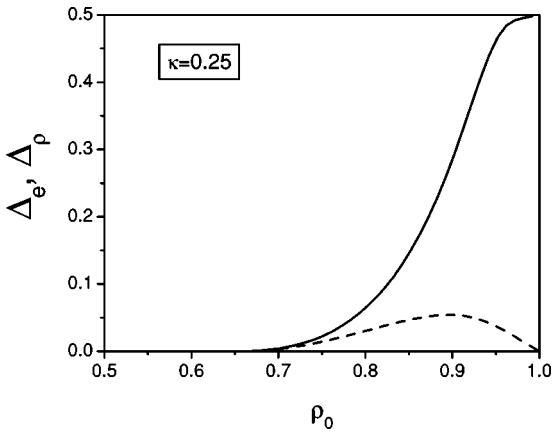

FIG. 5. Relative error gain $\Delta_e=(e_g^0-e_g^{opt})/e_g^0$ and relative overlap gain $\Delta_\rho=(\rho_{opt}-\rho_0)/\rho_0$ as a function of the maturity level ρ_0, for $\kappa=0.25$. Note that, although $\theta_{opt}>0$ after $\rho_c\approx 0.64$, the optimal epoch for pruning occurs when the maturity error is $\rho_0^*\approx 0.9$ (to maximize Δ_ρ) or after $\rho_0^*=0.95$ (to maximize Δ_e).

Near the critical value ρ_c, we have verified numerically (see inset of Fig. 3) that the optimal threshold starts growing as

$$\theta_{opt}(\rho_0,\kappa)\propto[\rho_0-\rho_c(\kappa)]^{1/2}. \tag{28}$$

Notice also that in the limit $\rho_0\to 1$ one has $\theta_{opt}\to 0.5$.

In order to find the minimal maturity level $\rho_c(\kappa)$ for which pruning improves the generalization error we calculate $d\rho/d\theta|_{\theta=0}$. As suggested by Fig. 2, we would naively expect to determine ρ_c as the point for which the derivative $d\rho/d\theta|_{\theta=0}$ turns out positive. However, at the critical value, both the first and the second derivative calculated in $\theta=0$ are null for any ρ_0, so the critical line $\rho_c(\kappa)$ will be determined by the condition that the third derivative starts to be positive. The condition $d^3\rho/d\theta^3|_{\theta=0}=0$ leads (after some algebra) to the transcendental equation

$$\left(\frac{2}{1-\rho_c^2}-\kappa\right)\exp\left(-\frac{\rho_c^2}{2\kappa(1-\rho_c^2)}\right)=1-\kappa, \tag{29}$$

where we have used that $P=\kappa$ and $S=\kappa\rho_0$ at $\theta=0$.

This equation gives the phase boundary $\rho_c(\kappa)$ shown in Fig. 4. Above $\rho_c(\kappa)$, pruning is effective (if $\theta<\theta_{max}$), but below this line we have $\rho(\theta)<\rho(0)\equiv\rho_0$ and ($\theta_{opt}=\theta_{max}=0$) so it is better not to prune. Curiously, near $\kappa=1$, the convergence of $\rho_c(\kappa)$ toward $\rho_c(1)=1$ is highly nonlinear (see inset in Fig. 4).

There is also an optimal pruning epoch $\rho_0^*(\kappa)>\rho_c$ such that the relative gain $\Delta_\rho=[\rho(\theta_{opt})-\rho_0^*]/\rho_0^*$ is maximized (see Fig. 5). In this figure we also present the generalization error gain $\Delta_e=[e_g^0-e_g(\theta_{opt})]/e_g^0$, which perhaps is a more meaningful property. This gain grows monotonically until the 50% level, with an inflection point at ρ^*. The sigmoidal character of Δ_e leads to a natural separation between two regimes, the last one favorable to pruning.

For a given κ, the optimal pruning epoch ρ^* naturally defines two learning phases: in the first one there is a huge amount of connection, and not only the synapses but also the network architecture is plastic; in the second one pruning occurs with advantage and, after pruning, the architecture is structured in an irreversible way, plasticity being restricted only to the fine tuning of the surviving synapses. It is tempting to identify the first plastic phase with a large number of synapses ($\rho_0<\rho_0^*$), as the usual ''critical learning period'' commonly found in developmental studies.

IV. SIMULATION RESULTS

To illustrate the validity of our calculations we present simulation results for a particular learning algorithm, the so called *simple Hebb rule*. For this learning procedure, the change due to the μth example is written as

$$\Delta J_i\propto S_i^\mu\tau^\mu \tag{30}$$

and the training potential is $V(\lambda)=-\lambda$.

We first check the validity of Eq. (20), presenting results for the student distribution of weights $P(J)$ produced by the simple Hebb rule when the teacher distribution has two peaks,

$$\mathcal{P}_B(B)=\kappa\delta(B-1)+(1-\kappa)\delta(B), \tag{31}$$

that is, with parameters ($\kappa_1=\kappa,\kappa_2=0,\kappa_0=1-\kappa$). The peak at zero represents the fraction of possible inputs available to the neuron at the developmental stage but which should not remain in the mature network. The peak at $B=1$ corresponds to the fraction of mature synapses to be present in the target (mature) network. The simulation, with a network with $N=4000$ and $\kappa=0.25$, gives a very good confirmation of the analytical curve (see Fig. 6).

It is known [10] that after the presentation of αN examples, the student-teacher overlap achieved by the simple Hebb rule is $\rho_0(\alpha)=(1+\pi/2\alpha)^{-1/2}$. In Fig. 7 we show the performance before and after pruning the weights with $|J|<\theta_{opt}(\rho_0,\kappa)$ for $\kappa=0.25$. The curve $e_g^{opt}(\alpha)$ (solid) gives the theoretical lower bound for any pruning strategy applied to simple Hebbian learning. Notice that the critical maturity ρ_c translates to a critical number of examples

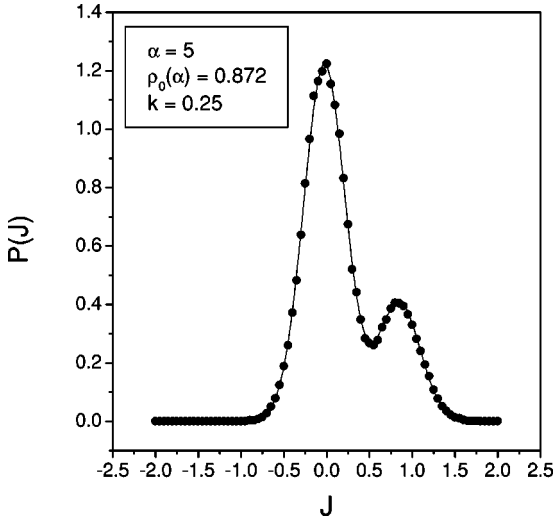

FIG. 6. Post-training distribution $\mathcal{P}(J)$ produced by the simple Hebb rule when the teacher distribution is $\mathcal{P}_B(B)=(1-\kappa)\delta(B)+\kappa\delta(B-1)$ with $\kappa=0.25$. Simulation results (circles) compared to the theoretical curve (solid) for $\alpha=5$, that is, $\rho_0(\alpha)\approx 0.872$.

$$\alpha_c=\frac{\pi}{2}\frac{\rho_c^2}{1-\rho_c^2}\approx 1.01, \tag{32}$$

before which the optimal choice is not to prune. After $\alpha \approx 15$, the error gain stabilizes around $\Delta_e=0.5$, so we may pick this value as a good pruning epoch.

V. DISCUSSION

There exists an intense debate in the literature about the meaning of synaptic pruning, which reflects the controversy between instructionist, selectionist, and nativist theories of development (for a review, see [4,5]). Instructionists emphasize the role of environmental factors in directing the development of neural synapses. Selectionists view the role of the environment as selecting synapses from a primary repertoire with the initial value of weights set by intrinsic (nondirected) factors. Nativists would see synaptic pruning as an example of ''programmed death'' of structures within a maturational schedule.

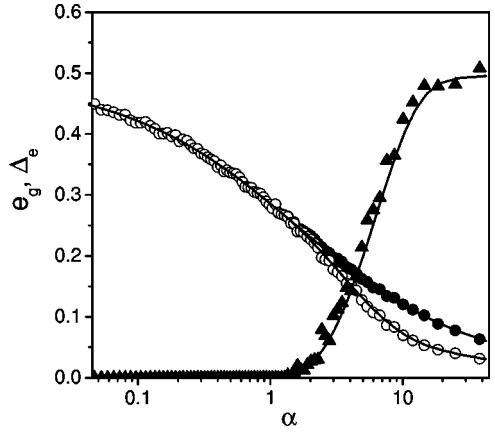

FIG. 7. Generalization error $e_g(\alpha)$ for simple Hebb algorithm, in log-linear scale, without pruning (above) and with the optimal pruning threshold $\theta_{opt}(\rho_0(\alpha))$ (below) for teacher complexity $\kappa=0.25$. Data from simulation (circles) with $N=400$ and theoretical curves (solid). Also shown is the relative error gain Δ_e (triangles). Note that the critical maturity ρ_c is achieved at $\alpha_c\approx 1.01$ but the optimal pruning epoch is near $\alpha=15$.

It is interesting that all these different ideas can be implemented, analyzed, and compared within the same perceptron student-teacher scenario. For example, López and Kinzel studied a pure selectionist process [14]. In that case, Hebbian information is not used for incremental modification of the student weights but only to set the pruning criterion. They found that the performance depends strongly on the initial distribution of student weights and its correlation to the teacher weight distribution. Since the diluted teacher scenario was not analyzed in that paper, only partial overlap with the teacher vector could be achievable in the limit $\alpha \to \infty$. To our knowledge, López and Kinzel performed the single analytical study of pure neural Darwinism in perceptrons and further comparison with our diluted teacher results would be desirable.

In the model studied here, synaptic selection, although computationally relevant, is a secondary process. Error correction (directed change in the synaptic weights) is the primary process, being also essential for discovering which synapses should be pruned. Without sufficient learning (that is, when $\rho<\rho_c$), there is no clear differentiation in synaptic magnitudes, and no hint as to which synapses should be eliminated. Optimization of the pruning criterion inevitably leads to a very rich scenario. If optimization principles are relevant to biological processes (this is not a consensual idea), and if some generic or robust behaviors found in perceptron theory can be translated to the biological context, then the model studied here could provide interesting suggestions.

First, our results suggest that the level of pruning could depend on the complexity of the task to be performed: in our simple scenario, on the fraction κ of relevant inputs. So we expect that the pruning level will vary enormously for different brain regions, depending on the complexity of the function to be implemented by these networks. This indeed has been found in some experimental studies [1,2,4].

Second, the fraction of eliminated synapses could depend on time not directly (being not simply a maturational schedule) but indirectly through the time evolution of the performance, that is, on some measure similar to our ''maturity'' level ρ_0, which depends on the number of learning instances α that occur up to time t. This indeed is coherent with experiments where Hebbian learning (LTP) is chemically blocked, that is, $\alpha(t)$ is slowed down: in that case, the pruning process is also retarded [1,2]. The dependence of optimal strategies on the overlap ρ_0 has been previously found in the realm of optimal generalization algorithms [11] and is a robust feature preserved in optimal multilayer networks and Bayesian approaches to learning.

Third, the optimal pruning threshold θ_{opt} is not stationary. Pruning is effective only after gross differentiation between synaptic strengths, promoted by learning, is achieved. This is what has been observed in the neuromuscular junction [3]: pruning is most observed after the magnitude ratio between strong and weak synapses is near 4. But our results suggest that the pruning criterion θ_{opt} should not be fixed but could change with time: initially only the very weak synapses are

eliminated, but, after more learning, even medium size synapses could be pruned. We think that the gradual increase of θ_{opt} could be implemented by a maturational decrease in the abundance of neural growth factors so that small synapses start first to fall into starvation [2]. We do not know if this nonstationary pruning threshold has already been measured in experiments, so it can be viewed as an independent suggestion of our model.

From the model an optimal epoch ρ_0^* for massive pruning also emerges naturally, namely, the epoch where the gains Δ_ρ and Δ_e are maximized. We suggest that the more plastic phase before massive pruning, when the architecture is not well defined and $\rho_0 < \rho^*$, should be correlated to the critical learning period usually observed in biological learning [1].

Finally, we observe that our calculations refer to a *quenched pruning scenario* [8,9,13,14], that is, to the problem of what is the better pruning strategy after the network has seen αN examples (or, equivalently, after it has achieved the maturity level ρ_0). After pruning, connections cannot reappear or be learned any longer in this scenario. One must recognize, however, that since new examples certainly arrive after this epoch, it would be better to permit the formation of new synapses and also new pruning phases. At the present it is not clear how to optimize this process, and it must be stressed that our results concern only optimization of a single pruning phase. The scenario of successive overgrowth/pruning waves suggested by some authors (open review in [4]) appears naturally in this context, being a possible mechanism to overcome the limitations of quenched pruning.

VI. CONCLUSIONS

We have studied analytically a simple model of pruning in artificial neural networks with similar features to those present in the experiments of Colman *et al.* [3]: pruning by deletion of the weaker synapses after strength differentiation induced by Hebbian learning. We have found a phase diagram in the space of environmental complexity κ and prior maturity level ρ_0, with two different regimes. If the maturity level is insufficient $[\rho_0 < \rho_c(\kappa)]$, pruning is deleterious and should not be performed $(\theta_{opt}=0)$. After a critical maturity level $\rho_c(\kappa)$, pruning enhances the network performance, the optimal pruning threshold is different from zero, and grows as $\theta_{opt}(\kappa) \propto [\rho_0 - \rho_c(k)]^{-1/2}$. Our results are valid for a large class of learning algorithms since the way the prior learning level ρ_0 is achieved is not of primary importance. We also expect that optimization of θ will lead to similar features in multilayer networks.

Since pruning is a very complex, environmentally dependent, and nonstationary process even in the simple perceptron learning scenario, we conjecture that similar effects could also be present in biological pruning. In contrast to pure selectionism or genetic nativism, we studied the case where synaptic pruning is thought of as a process analogous to branch and leaf selection during the growth of trees and similar biological structures (the *selective trophism scenario*). Like a tree (in Greek, ''dendron''), the detailed dendritic architecture most probably is not genetically prewired nor does it simply represent the outcome of pruning a randomly generated graph. Growth by environmentally directed processes (trophism) and competition for nerve-growth factors (selectionism) will lead to a strongly nonlinear outcome of genetic and environmental factors.

ACKNOWLEDGMENTS

We have the pleasure to acknowledge J. F. Fontanari for his advice during the elaboration of this work. Silvia M. Kuva aided us in revising the paper. D.M.B. acknowledges research support from UNIP and O.K. from FAPESP.

[1] M. C. Brown, W. G. Hopkins, and R. J. Keynes, *Essentials of Neural Development* (Cambridge University Press, Cambridge, 1991); P. Rakic, J. P. Bourgeois, and P. S. Goldman-Rakie, Prog. Brain Res. **102**, 227 (1994); G. M. Innocenti, TINS **18**, 397 (1995).

[2] E. R. Kandel and T. J. O'Dell, Science **258**, 243 (1992); W. Singer, *ibid.* **270**, 758 (1995).

[3] H. Colman, J. Nabekura, and J. W. Lichtman, Science **275**, 356 (1997); E. Frank, *ibid.* **275**, 324 (1997).

[4] S. R. Quartz and T. J. Sejnowski, Behav. Brain Sci. **20**, 537 (1997). See also the open review commentary that follows this article.

[5] J-P. Changeaux and A. Danchin, Nature (London) **264**, 705 (1976); J-P. Changeaux, *The Neuronal Man* (Oxford University Press, Oxford, 1985).

[6] R. Reed, IEEE Trans. Neural Netw. **4**, 740 (1993).

[7] G. Chechik, I. Meilijson, and E. Ruppin, Neural Comput. **10**, 1759 (1998).

[8] K. Y. M. Wong and M. Bouten, Europhys. Lett. **16**, 525 (1991); P. Kuhlmann, R. Garcés, and H. Eißfeller, J. Phys. A **25**, L593 (1992).

[9] D. M. L. Barbato and J. F. Fontanari, Phys. Rev. E **51**, 6219 (1995); J. Phys. A **29**, 7003 (1996).

[10] H. S. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992); T. L. H. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys. **65**, 599 (1993); M. Bouten, J. Schietse, and C. Van den Broeck, Phys. Rev. E **52**, 1958 (1995); M. Opper and W. Kinzel, in *Models of Neural Networks III*, edited by E. Domany *et al.* (Springer-Verlag, Berlin, 1995).

[11] O. Kinouchi and N. Caticha, J. Phys. A **25**, 6243 (1992); C. W. H. Mace and A. C. C. Coolen, Stat. Comput. **8**, 55 (1998); R. Vicente, O. Kinouchi, and N. Caticha, Mach. Learning **32**, 179 (1998); N. Caticha and O. Kinouchi, Philos. Mag. B **77**, 1565 (1998); in *Online Learning in Neural Networks*, edited by D. Saad (Cambridge University Press, Cambridge, 1999).

[12] M. S. Gray, D. T. Lawrence, B. A. Golomb, and T. J. Sejnowsky, Neural Comput. **7**, 1160 (1995).

[13] P. Kuhlmann and K-R. Müller, J. Phys. A **27**, 3759 (1994).

[14] B. López and W. Kinzel, J. Phys. A **30**, 7753 (1997).