Dilution in a linear neural network

D. M. L. Barbato and J. F. Fontanari

Instituto de Física de São Carlos, Universidade de São Paulo, Caixa Postal 369, 13560-970 São Carlos, São Paulo, Brazil

(Received 28 November 1994; revised manuscript received 13 March 1995)

The effects of elimination of synaptic weights on the learning capability of a single-layer, feed-forward neural network composed of linear neurons are investigated within the equilibrium statistical mechanics framework of Gardner and co-workers [J. Phys. A **21**, 257 (1988); **21**, 271 (1988)]. A comparison between the performances of networks damaged by different types of dilution, which may occur either before or after the learning stage, shows that the strategy of minimizing the training error does not yield the best generalization performance. Moreover, this comparison also shows that, depending on the size of the training set and on the level of noise corrupting the training data, the smaller weights may become the determinant factors in the good functioning of the network. In particular, the larger the level of noise, the more important the contribution of the smaller weights to the generalization capability of the network.

PACS number(s): 87.22.Jb, 02.70.−c

I. INTRODUCTION

The main purpose of the study of dilution in artificial neural networks is the modeling and understanding of the effects of lesions in systems capable of learning. From both biological and psychological viewpoints, this enterprise is justified by the claim that the comparison between the behavioral patterns of damaged artificial neural networks and injured biological brains may elucidate the underlying mechanisms of functioning of the brain [1, 2]. From a more applied viewpoint, however, the study of lesions in artificial neural networks is attractive by itself, being important not only to probe the reliability of the system when partially damaged but also to single out its essential components that should be protected from damage. This is the trend we follow in the present paper, although the diverse types of lesions we consider are biologically motivated.

The neural network we consider in this paper is a single-layer, feed-forward neural network whose basic processing units (neurons) are linear elements. The so-called linear perceptron is probably the simplest nontrivial model of a learning system that can be solved exactly, either by a statistical dynamical approach [3] or by the equilibrium statistical mechanics framework of Gardner [4, 5] within the simplifying replica-symmetric assumption [6]. We work within the student-teacher scenario, which is the paradigm of the statistical mechanics approach to the problem of learning from examples in neural networks [7–9]. In this scenario, the input-output mapping or task in which the network is trained is generated by another neural network, not necessarily with the same architecture, termed the *teacher* network. The network trained to realize a subset of that mapping (training set) is termed the *student* network. In this study, both networks are linear perceptrons though only the student network is diluted. The task becomes then unrealizable as the computational power of the student network is insufficient to perfectly learn the rule supplied by the teacher. In fact, in this case there exists a finite training set size (storage capacity) beyond which the error made by the student perceptron in realizing the training set (training error) no longer vanishes.

Diluting or lesioning a neural network by cutting a certain fraction of its synaptic weights can be done by several ways. First, the lesion can occur *before* the learning process takes place. There are then two possibilities: the learning procedure specifies which weights must be eliminated so as to minimize the effect of the lesion on the training error, or the deleted weights are specified *a priori* without the possibility of changes during the learning stage. The first possibility is termed *annealed* dilution while the second is termed *quenched* dilution. Second, the lesion occurs *after* the learning stage has finished. In this case we consider three possibilities: only the smaller weights are eliminated, only the larger weights are eliminated, and the weights are eliminated randomly, independently of their magnitudes. Of course, the fraction of deleted weights is the same in all cases. The main goal of this paper is to compare the effects of these different types of dilution on the learning capability of the neural network. A remarkable by-product of this comparison is the result that the generalization performance obtained with the annealed dilution is not optimal, being overcome by the deletion of the smaller weights after learning. This is probably the simplest example where, even in the absence of any type of noise, the widespread strategy of optimizing the performance in the training set aiming at optimizing the generalization performance is not optimal.

Besides dilution, we also investigate the effects of static noise corrupting the input patterns presented to the student perceptron. In fact, there is evidence that dilution and noise may produce similar results. In particular, it was shown that the main effect of quenched dilution in a Boolean binary perceptron is to introduce an effective noise in the training patterns [10]. Moreover, some studies of lesions in perceptrons were carried out by adding

1063-651X/95/51(6)/6219(11)/$06.00

a static noise term to the internal fields of the neurons, rather than explicitly cutting synaptic weights [11]. We note that, similarly to dilution, the noise corrupting the input patterns makes the task of learning the training set unrealizable to the student perceptron.

The problem of learning unrealizable rules with real-weights *Boolean* perceptrons cannot be carried out within the replica-symmetric framework, because it yields locally unstable solutions for training set sizes larger than the storage capacity of the network. The exact solution to this problem is notoriously difficult since it demands the application of the full replica-symmetry breaking scheme of Parisi [12,13] (see [14,15] for an application of the first step of Parisi's scheme to this type of neural network). Nevertheless, the effects of noise [8] and dilution before learning [16] in the real-weights Boolean perceptron beyond its storage capacity were studied within the replica-symmetric assumption in the hope of obtaining a solution that might have at least the value of an approximation. This situation contrasts with that of the linear perceptron, for which the replica-symmetric assumption always yields locally stable solutions. The linear perceptron appears then as the sole real-weights neural network for which a study of unrealizable tasks can be fully and reliably carried out.

Dilution in neural networks was first considered in the context of the associative memory model proposed by Hopfield [17]. In this model all the memory patterns are embedded in the network at once through the Hebbian prescription for the synaptic weights. The equilibrium analysis of the retrieval properties of the randomly diluted Hopfield model, in the case that the connectivity of the network is of the same order of the number of neurons N, was carried out by Sompolinsky [18, 19]. It was shown that the effect of dilution is to add a Gaussian noise to the Hebbian synaptic weights: the larger the degree of dilution, the larger the noise. In a remarkable contribution, Derrida *et al.* [20] have solved analytically the dynamics of the Hopfield model in the limit of extreme dilution, i.e., when the connectivity is of order $\ln N$. It was found, however, that the degree of dilution or the connectivity parameter simply rescales the number of neurons, being of no importance to the retrieval properties of the network. In these studies there was no need to distinguish between dilution before or after learning since the Hebbian rule is purely local.

The seminal paper of Gardner [4] has deviated the attention from the Hopfield model to the neural network with optimal weights. Similarly to the Hopfield model, the dynamics of the optimal attractor neural network was solved analytically in the limit of extreme dilution [21]. The effect of dilution on the retrieval of hierarchically organized memories was also considered in this context [11, 22]. As the framework of Gardner readily applies to the study of the learning capabilities of perceptrons, the effect of dilution on the storage capacity of the Boolean perceptron was investigated by Bouten *et al.* [23], from whom we have borrowed the terms annealed and quenched. More recently, Kuhlmann and Müller [16] have considered the effect of dilution on the generalization performance of the Boolean perceptron. However, as mentioned before, the instability of the replica-symmetric ansatz employed in their analysis renders their results rather doubtful. The emphasis of these studies was on dilution before learning.

In the present paper, emphasis is given to the study of dilution *after* learning. This type of dilution is useful to identify the components whose destruction may affect more severely the performance of the network. Moreover, while previous studies have dealt mainly with the effect of dilution on the retrieval properties of attractor neural networks [18–20] or on the storage capacity of the Boolean perceptron [23], we focus mainly on the *generalization* performance of the linear perceptron. In particular, we have found that for $2 - 1/\gamma^2 < \alpha < 1/\gamma^2$, where α and γ are parameters measuring the training set size and the level of noise, respectively, the generalization performance is more sensitive to the deletion of the smaller weights. Moreover, for α within that range, we have found that the diluted network generalizes better than the nondiluted one. This finding lends support to the often employed strategy of deleting weights after learning in order to decrease overfitting in the case of training with noisy data.

The remainder of the paper is organized as follows. In Sec. II we describe the architecture of the linear perceptron and define the quantities employed to measure its performance. Section III is devoted to the study of the case when dilution occurs before the learning process takes place and Sec. IV to the case when dilution occurs after the learning stage has finished. The results are then compared and analyzed in Sec. V. Finally, in Sec. VI we present some concluding remarks.

II. THE MODEL

The neural network we consider in this paper consists of N binary input units $S_i = \pm 1$ $(i = 1, \ldots, N)$, N real-valued synaptic weights J_i $(i = 1, \ldots, N)$ and a single linear output unit

$$\sigma = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i S_i. \tag{2.1}$$

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

$$t = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i^0 \xi_i, \tag{2.2}$$

where the weights J_i^0 $(i = 1, \ldots, N)$ are statistically independent random variables of zero mean and variance M. The specific probability distribution is not important because only the first two moments enter the calculations presented in the next sections. To achieve its task, the student network is trained with $P = \alpha N$ input-output pairs $(\vec{S}^l, t^l)$ $(l = 1, \ldots, P)$ where t^l is the teacher's output to input $\vec{\xi}^l$ and each component S_i^l is drawn from the conditional probability distribution

$$P(S_i^l \mid \xi_i^l) = \frac{1+\gamma}{2}\,\delta(S_i^l - \xi_i^l) + \frac{1-\gamma}{2}\,\delta(S_i^l + \xi_i^l) \qquad (2.3)$$

with

$$P(\xi_i^l) = \tfrac{1}{2}\,\delta(\xi_i^l - 1) + \tfrac{1}{2}\,\delta(\xi_i^l + 1). \qquad (2.4)$$

Thus, not only the student network has access to a very small fraction of the total number of input-output pairs of the mapping as its input patterns $\vec{S}^l = \left(S_1^l, \dots, S_N^l\right)$ are noisy versions of the pure patterns $\vec{\xi}^l = \left(\xi_1^l, \dots, \xi_N^l\right)$. The noise parameter $\gamma \in [0,1]$ determines the Hamming distance between these two patterns. In particular, $\gamma = 1$ characterizes the problem of learning from pure (noiseless) examples, while $\gamma = 0$ is the random mapping problem. Actually, the study of the case $\gamma = 0$ was carried out recently in a rather different context: it is the classical combinatorial optimization problem of finding the minimum weight solution, i.e., the solution with the minimum number of nonvanishing entries, to randomly generated linear equations [24, 25].

The learning process consists of a search for the global minima of the training energy, defined as

$$E\left(\{J_i\}\right) = \frac{1}{2} \sum_{l=1}^{P} \left(t^l - \sigma^l\right)^2, \qquad (2.5)$$

where $\sigma^l = \sigma(\{J_i\}, \vec{S}^l)$ is the student's response to noisy input $\vec{S}^l$ and t^l is the teacher's output to pure input pattern $\vec{\xi}^l$. The training energy is then a measure of the performance of the network in realizing the $P = \alpha N$ input-output pairs of the training set. The storage capacity α_c is defined as the ratio between the maximal training set size that the network can realize perfectly and the number of input neurons N.

The ultimate goal of the learning process is to generate a network capable of realizing correctly an input-output pair not belonging to the training set. To measure this capability we introduce the generalization function

$$E_g\left(\{J_i\}\right) = \frac{1}{2} \int d\nu(\vec{S})\,\left(t - \sigma(\{J_i\}, \vec{S})\right)^2, \qquad (2.6)$$

where

$$d\nu(\vec{S}) = \prod_i d\xi_i dS_i P(S_i \mid \xi_i) P(\xi_i) \qquad (2.7)$$

is the measure in input space. Here σ is the student's response to noisy input $\vec{S}$ and t is the teacher's output to the randomly chosen pure input pattern $\vec{\xi}$. In the thermodynamic limit $N \to \infty$ the integrations in Eq. (2.6) can be easily carried out yielding

$$E_g\left(\{J_i\}\right) = \tfrac{1}{2}\left(Q + M - 2\gamma R\right), \qquad (2.8)$$

where Q is the squared norm of the student perceptron

$$Q = \frac{1}{N} \sum_i^N J_i^2, \qquad (2.9)$$

R is the overlap between student and teacher,

$$R = \frac{1}{N} \sum_i^N J_i J_i^0, \qquad (2.10)$$

and M is the squared norm of the teacher perceptron

$$M = \frac{1}{N} \sum_i^N \left(J_i^0\right)^2. \qquad (2.11)$$

Due to the self-averaging property [12, 13], M coincides with the variance of the random variable J_i^0 in the thermodynamic limit. For a more thorough discussion of the problem of learning from examples in neural networks we refer the reader to Refs. [8, 9].

As the focus of this paper is the equilibrium properties of the ensemble of weights that minimize the training energy, Eq. (2.5), we can directly apply the standard statistical mechanics techniques to characterize its ground states (global minima). In particular, the appropriately normalized average training error is given by

$$\epsilon_t = \frac{1}{\alpha} \lim_{\beta \to \infty} \frac{\partial\,(\beta f)}{\partial \beta}, \qquad (2.12)$$

where f is the average free-energy density

$$-\beta f = \lim_{N \to \infty} \frac{1}{N}\,\langle\!\langle \ln Z \rangle\!\rangle \qquad (2.13)$$

and Z is the partition function

$$Z = \mathrm{Tr}\,\exp\left[-\beta E\left(\{J_i\}\right)\right]. \qquad (2.14)$$

The notation $\langle\!\langle\,\rangle\!\rangle$ stands for a quenched average over the statistically independent random variables S_i^l, ξ_i^l, and J_i^0, while Tr indicates an integration over all allowed configurations $\{J_i\}$. Such configurations must satisfy a normalization constraint, Eq. (2.9), and a dilution constraint, namely, that $\{J_i\}$ possesses only κN nonvanishing components. The specific way these constraints are implemented depends on the type of dilution considered. There are, however, some general remarks we can make about the normalization constraint. In contrast with the real-weights Boolean perceptron for which the value of Q is irrelevant, the choice of the normalization is germane to the thermodynamic analysis of the linear perceptron: Q must be chosen so as to minimize ϵ_t [6]. In the regime $\alpha < \alpha_c$, where there are an infinity of values of Q that yield $\epsilon_t = 0$, we will choose the smallest one that corresponds then to the solution of minimal norm, so-called pseudoinverse [26].

The average generalization error is similarly defined as

$$\epsilon_g = \lim_{\beta \to \infty} \langle\!\langle \left\langle E_g\left(\{J_i\}\right)\right\rangle_T \rangle\!\rangle, \qquad (2.15)$$

where $\langle\,\rangle_T$ stands for a thermal average taken with the Gibbs probability distribution $\exp\left(-\beta E\right)/Z$. The zero-temperature limit ($\beta \to \infty$) ensures that only configurations that minimize the training energy will contribute to this average.

The main technical difficulty in carrying out the calculations delineated above is the evaluation of the quenched average in Eq. (2.13). This can be accomplished by the replica method which consists of using the identity

$$\langle\!\langle \ln Z \rangle\!\rangle = \lim_{n\to 0} \frac{1}{n} \ln \langle\!\langle Z^n \rangle\!\rangle , \tag{2.16}$$

evaluating $\langle\!\langle Z^n \rangle\!\rangle$ for *integer* n and then analytically continuing to $n = 0$. As for the quenched average in Eq. (2.15), it can be evaluated in a similar way by adding the term $hE_g(\{J_i\})$ to the training energy $E(\{J_i\})$ and then calculating the resulting average free energy. The derivative with respect to the auxiliary field h taken at $h = 0$ will give the desired result.

III. DILUTION BEFORE LEARNING

In this section we discuss the case when the lesion occurs before the learning process takes place. The two possibilities we consider in the following are inspired by the clinical observation of patients who received severe injury to only one of the brain's hemispheres [27]. If the patient is at an early enough age, the other side of the brain can take over, compensating for the damage. This flexibility is modeled by the annealed dilution, which leaves to the learning process the decision of which weights to cut in order to attenuate the effect of the lesion on the training error. In this sense, this dilution process depends on the particular realization of the training set. As the brain grows older, it looses the flexibility and the lesion is best described by the quenched dilution, where the deleted weights are chosen randomly and held fixed during the learning stage.

A. Annealed dilution

In this case we define a new set of real-valued weights $\{W_i\}$ so that $J_i = c_i W_i$ $(i = 1, \ldots, N)$ where the binary variables $c_i = 0, 1$ are needed to enforce the correct degree of dilution

$$\frac{1}{N} \sum_{i=1}^{N} c_i = \kappa. \tag{3.1}$$

In terms of these new variables the training energy is rewritten as

$$E_\kappa(\{W_i\}, \{c_i\}) = \frac{1}{2} \sum_{l=1}^{P} \left(t^l - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} c_i W_i S_i^l \right)^2 . \tag{3.2}$$

The minimization of E_κ involves then a simultaneous search in the discrete space of c_i and in the continuous space of W_i. The search in the discrete space is the main reason why the statistical dynamics approach [3] is not well suited to the analysis of this problem. The partition function Eq. (2.14) becomes

$$Z = \sum_{\{\vec{c}\}} \delta_{\mathrm{Kr}} \left(\sum_i c_i, \kappa N \right) \int_{-\infty}^{\infty} \prod_i dW_i \times \exp\left[-\beta E_\kappa(\{W_i\}, \{c_i\})\right] , \tag{3.3}$$

where δ_{Kr} is the Kronecker delta. To avoid divergences when carrying out the integrals over W_i we must impose two normalization constraints,

$$Q = \frac{1}{N} \sum_{i=1}^{N} c_i W_i^2 \tag{3.4}$$

and

$$Q^0 = \frac{1}{N} \sum_{i=1}^{N} (1 - c_i) W_i^2 , \tag{3.5}$$

which guarantee the convergence of all integrals. Note that constraint (3.4) is identical to constraint (2.9) since $c_i^2 = c_i$. Clearly, our results must not depend on Q^0 as it is the squared norm of the subset of weights that do not contribute to the training energy.

The calculation of the average free-energy density Eq. (2.13) in the thermodynamic limit is standard [4, 5] so we present the final result only,

$$-\beta f^{ann} = \lim_{n\to 0} \operatorname{ext} \frac{1}{n} \left\{ - \sum_{a<b}^{n} q_{ab}\, \hat{q}_{ab} + \sum_{a}^{n} (\kappa \hat{c}_a + Q^0 \hat{Q}_a^0 - \tfrac{1}{2} Q \hat{Q}_a - R_a \hat{R}_a) + G_0(\hat{q}_{ab}, \hat{c}_a, \hat{Q}_a, \hat{Q}_a^0, \hat{R}_a) + \alpha G_1(q_{ab}, R_a) \right\}, \tag{3.6}$$

where

$$G_0 = \ln \sum_{\{c^a = 0,1\}} \int \prod_{a=1}^{n} dW^a \exp\left\{ - \sum_a^n [\hat{c}_a c^a + \hat{Q}_a^0 (1 - c^a)(W^a)^2 - \tfrac{1}{2} \hat{Q}_a c^a (W^a)^2] + \sum_a^n c^a \hat{R}_a W^a J^0 + \sum_{a<b}^{n} \hat{q}_{ab}\, c^a c^b\, W^a W^b \right\} \tag{3.7}$$

and

$$G_1 = \ln \int \prod_{a=1}^{n} \frac{dy_a}{\sqrt{2\pi}} \exp\left\{ -\tfrac{1}{2} \sum_a^n y_a^2 [1 + \beta(Q + M - 2\gamma R_a)] - \beta \sum_{a<b}^{n} y_a y_b (q_{ab} + M - 2\gamma R_a) \right\}. \tag{3.8}$$

The extremum in Eq. (3.6) is taken over all saddle-point parameters $(\hat{c}_a, \hat{q}_{ab}, \hat{Q}_a, \hat{Q}_a^0, \hat{R}_a, q_{ab}, R_a)$. The physical order parameters

$$q_{ab} = \frac{1}{N} \sum_{i=1}^{N} c_i^a c_i^b W_i^a W_i^b , \qquad a < b \tag{3.9}$$

and

$$R_a = \frac{1}{N} \sum_{i=1}^{N} c_i^a W_i^a J_i^0 \tag{3.10}$$

measure the overlap between two different global minima $\{J_i^a\}$ and $\{J_i^b\}$, and the overlap between the global minimum $\{J_i^a\}$ and the teacher network $\{J_i^0\}$, respectively.

The search for the extremum of f^{ann} will be restricted to a very particular subspace — the replica-symmetric subspace — where the values of the saddle-point parameters are independent of their replica indices: $q_{ab} = q$, $\hat{q}_{ab} = \hat{q}$ $\forall a < b$ and similarly for $\hat{c}_a$, $\hat{Q}_a$, R_a, $\hat{R}_a$, and $\hat{Q}^0_a$. Although this ansatz is clearly correct for parameters that possess only one replica index [12], its use for parameters that possess two replica indices must be justified by a stability analysis. Evaluation of Eqs. (3.7) and (3.8) with the replica-symmetric ansatz is straightforward, resulting in the following expression for the replica-symmetric average free-energy density:

$$-\beta f^{ann}_{rs} = \tfrac{1}{2}q\hat{q} - \tfrac{1}{2}Q\hat{Q} + Q^0\hat{Q}^0 - R\hat{R} - \frac{1-\kappa}{2}\ln\hat{Q}^0 + \kappa\hat{c}' + \frac{\kappa}{2}\ln 2 + \tfrac{1}{2}\ln\pi - \frac{\alpha}{2}\ln\left[1+\beta\left(Q-q\right)\right]$$
$$-\frac{\alpha\beta}{2}\frac{q+M-2\gamma R}{1+\beta\left(Q-q\right)} + \int Dz\ln\left\{1+\frac{\exp[-\hat{c}' + z^2(\hat{q}+M\hat{R})/2(\hat{q}-\hat{Q})]}{\sqrt{\hat{q}-\hat{Q}}}\right\}, \tag{3.11}$$

where $Dz = dz/\sqrt{2\pi}\exp(-z^2/2)$ is the Gaussian measure. We have introduced the parameter $\hat{c}' = \hat{c} - \frac{1}{2}\ln 2\hat{Q}^0$ which allows for the complete decoupling between $\hat{Q}^0$ and the remaining saddle-point parameters which are relevant to the characterization of the ground states. The replica-symmetric saddle-point parameters $(\hat{c}, \hat{q}, \hat{Q}, \hat{Q}^0, \hat{R}, q, R)$ are obtained by extremizing f^{ann}_{rs} which gives rise to a set of six coupled equations, as the equation for $\hat{Q}^0$ does not involve the other parameters. To take the zero-temperature limit, one must be careful to distinguish between two regimes that arise naturally from the analysis of the average training error, Eq. (2.12), which within the replica-symmetric assumption is given by

$$\epsilon^{ann}_t = \lim_{\beta\to\infty}\frac{1}{2}\,\frac{M+Q-2\gamma R+\beta\left(Q-q\right)^2}{\left[1+\beta\left(Q-q\right)\right]^2}. \tag{3.12}$$

The first regime of interest, characterized by a nonzero training error, occurs only if $q \to Q$ so that $x \equiv \beta(Q-q)$ is finite. In this regime and choosing Q so as to minimize f^{ann}_{rs}, which contributes with an additional equation $\partial f^{ann}_{rs}/\partial Q = 0$, the task of solving the saddle-point equations is greatly simplified resulting in the following expressions for the relevant order parameters:

$$x = \frac{\Lambda_\kappa}{\alpha - \Lambda_\kappa}, \tag{3.13}$$

$$q = Q = \frac{M\Lambda_\kappa}{\alpha-\Lambda_\kappa}\,\left[1+\gamma^2\left(\alpha - 2\Lambda_\kappa\right)\right], \tag{3.14}$$

and

$$R = M\gamma\Lambda_\kappa, \tag{3.15}$$

where

$$\Lambda_\kappa = 2\int_{\lambda_\kappa}^{\infty} Dz\; z^2 \tag{3.16}$$

and λ_κ is the unique solution of

$$\kappa = 2\int_{\lambda_\kappa}^{\infty} Dz. \tag{3.17}$$

Hence the average training error becomes

$$\epsilon^{ann}_t/M = \frac{1}{2}\,\left(1-\gamma^2\Lambda_\kappa\right)\left(1-\frac{\Lambda_\kappa}{\alpha}\right), \tag{3.18}$$

from which we conclude that the regime of nonzero training error occurs for $\alpha > \alpha^{ann}_c = \Lambda_\kappa$. The dependence of α^{ann}_c on the connectivity parameter κ is depicted in Fig. 1. Note that $\Lambda_\kappa \geq \kappa$ for all κ and $\Lambda_1 = 1$. For $\alpha > \alpha^{ann}_c$ the average generalization error is given by

$$\epsilon^{ann}_g/M = \frac{1}{2}\,\frac{\alpha\left(1-\gamma^2\Lambda_\kappa\right)}{\alpha - \Lambda_\kappa} \tag{3.19}$$

which, in the limit of large α, can be rewritten as

$$\epsilon^{ann}_g/M = \tfrac{1}{2}\left(1-\gamma^2\Lambda_\kappa\right)\left(1+\frac{\Lambda_\kappa}{\alpha}\right) + O(\alpha^{-2}). \tag{3.20}$$

To investigate the second regime of interest, characterized by a vanishing training error, we must first solve the saddle-point equations for Q fixed *a priori*. The strategy of determining Q by minimizing the free energy with respect to this parameter fails in this case because, at zero temperature, there are an infinity of values of Q consis-

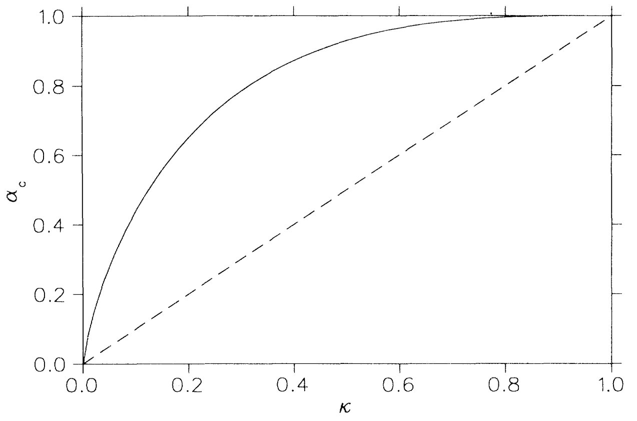

FIG. 1. Storage capacity α_c of the linear perceptron as a function of the connectivity κ for the annealed dilution (solid curve) and the quenched dilution (broken curve).

tent with $\epsilon_t^{ann} = 0$. More specifically, the analysis of the solution of the saddle-point equations for fixed Q shows that any $Q > Q^P$, where

$$Q^P = \frac{M}{2} \frac{\alpha\left(1-\gamma^2\alpha\right)}{\Lambda_\kappa - \alpha}, \tag{3.21}$$

yields $\epsilon_t^{ann} = 0$ for $\alpha \leq \alpha_c^{ann}$ (see [6] for a similar but more detailed analysis for the nondiluted problem). The choice $Q = Q^P$ corresponds to the pseudoinverse solution [26]. Within this framework we find $q = Q^P$ and $R = M\gamma\alpha$. Although the training error is zero, the generalization error does not vanish, being given by

$$\epsilon_g^{ann}/M = \frac{1}{2} \frac{\Lambda_\kappa - \gamma^2\alpha\left(2\Lambda_\kappa - \alpha\right)}{\Lambda_\kappa - \alpha}. \tag{3.22}$$

To conclude the analysis of the annealed dilution we discuss the stability of the replica-symmetric ansatz employed in the calculation of the average training and generalizations errors. The condition for the local stability of this ansatz is given by [28]

$$\alpha\gamma_0\gamma_1 < 1, \tag{3.23}$$

where γ_0 and γ_1 are the transverse eigenvalues of the matrices of second derivatives of G_0 and G_1 with respect to $\hat{q}_{ab}$ and q_{ab}, respectively. Following the analysis of Ref. [5] we find that this condition reduces to

$$\frac{\kappa}{\alpha} < 1 \qquad \text{if} \qquad \alpha > \Lambda_\kappa \tag{3.24}$$

and

$$\frac{\alpha\kappa}{2\Lambda_\kappa^2} < 1 \qquad \text{if} \qquad \alpha \leq \Lambda_\kappa, \tag{3.25}$$

which are always satisfied since $\Lambda_\kappa \geq \kappa$ as shown in Fig. 1. It is interesting to note that both the storage capacity α_c^{ann} and the local stability conditions do not depend on the noise parameter γ.

B. Quenched dilution

Since in this case the $(1-\kappa)N$ deleted weights are chosen randomly, we can set $J_i = 0$ ($i = \kappa N + 1, \ldots, N$), without lack of generality, so as to automatically satisfy the dilution constraint. In fact, an explicit calculation of the distribution of probability of a given weight, say J_i, taking on the value J shows that this distribution is a Gaussian of zero mean and variance Q, discarding thus the existence of a singularity at $J = 0$ [6]. As the binary auxiliary variables c_i are no longer necessary to enforce the dilution constraint, the calculation of the average free energy is much simplified for the quenched dilution. Similarly to the analysis of the annealed dilution, we find

$$-\beta f^q = \lim_{n\to 0} \text{ext} \frac{1}{n} \left\{ -\sum_{a<b}^{n} q_{ab}\, \hat{q}_{ab} - \sum_{a}^{n} (\tfrac{1}{2} Q\hat{Q}_a + R_a\hat{R}_a) + \kappa\, G_0(\hat{q}_{ab}, \hat{Q}_a, \hat{R}_a) + \alpha\, G_1\left(q_{ab}, R_a\right) \right\}, \tag{3.26}$$

where

$$G_0 = \ln \int \prod_{a=1}^{n} dJ^a \exp\left\{ \frac{1}{2} \sum_{a}^{n} \hat{Q}_a (J^a)^2 + \sum_{a}^{n} \hat{R}_a J^a J^0 + \sum_{a<b}^{n} \hat{q}_{ab}\, J^a J^b \right\} \tag{3.27}$$

and G_1 is given in (3.8). Using the replica-symmetric ansatz yields

$$-\beta f_{rs}^q = \tfrac{1}{2} q\hat{q} - \tfrac{1}{2} Q\hat{Q} - R\hat{R} + \frac{\kappa}{2} \ln 2\pi - \frac{\alpha}{2} \ln[1 + \beta(Q - q)] - \frac{\alpha\beta}{2} \frac{q + M - 2\gamma R}{1 + \beta\left(Q - q\right)} - \frac{\kappa}{2} \ln\left(\hat{q} - \hat{Q}\right) + \frac{\kappa}{2} \frac{\hat{q} + M\hat{R}^2}{\hat{q} - \hat{Q}}. \tag{3.28}$$

The average training error is again given by Eq. (3.12), except that now the saddle-point parameters q, Q, and R extremize the free energy (3.28). Solving the saddle-point equations in both regimes $\epsilon_t > 0$ and $\epsilon_t = 0$, following the same procedure used before, we find a strikingly formal similarity between the equations describing the equilibrium properties of the annealed and quenched dilutions. More specifically, to obtain the physical saddle-point parameters as well as the average training and generalization errors for the quenched dilution we must only replace Λ_κ by κ in the corresponding equations of the annealed dilution. In particular, the storage capacity of the quenched diluted network is $\alpha_c^q = \kappa$ while, for $\alpha > \kappa$, the average training and generalization errors are given by

$$\epsilon_t^q/M = \frac{1}{2} \left(1 - \gamma^2\kappa\right)\left(1 - \frac{\kappa}{\alpha}\right) \tag{3.29}$$

and

$$\epsilon_g^q/M = \frac{1}{2} \frac{\alpha\left(1 - \gamma^2\kappa\right)}{\alpha - \kappa}, \tag{3.30}$$

respectively. For $\alpha \leq \kappa$ we find $\epsilon_t^q = 0$ and

$$\epsilon_g^q/M = \frac{1}{2} \frac{\kappa - \gamma^2\alpha\left(2\kappa - \alpha\right)}{\kappa - \alpha}. \tag{3.31}$$

The formal similarity between the equations describing the annealed and quenched dilutions is probably a peculiarity of the linear perceptron, since an equally thorough comparison between these two types of dilutions in the Boolean binary perceptron has not indicated anything of this sort [10].

We have also verified that the replica-symmetric ansatz is locally stable, in the sense of satisfying inequality (3.23), for all values of the control parameters α, γ, and κ.

IV. DILUTION AFTER LEARNING

We consider now the more interesting problem when the lesion occurs after the learning process has finished. In this case, the student network can learn the training set using its full capabilities, as there are no constraints on the number of nonvanishing weights during the learning stage. The equilibrium properties of the ensemble of networks or weight configurations $\{J_i\}$ so generated are obtained by setting $\kappa = 1$ in the equations for the order parameters, training and generalization errors given in Sec. III. Of course, both annealed and quenched dilution give the same results for $\kappa = 1$. A particularity of the noiseless ($\gamma = 1$), nondiluted ($\kappa = 1$) limit is the existence of a continuous transition to a regime of perfect generalization at $\alpha = 1$ [3, 6, 9]. Thus, for $\alpha > 1$ the unique ground state (global minimum) of the training energy is $\{J_i = J_i^0\}$. A similar phenomenon occurs in the Boolean binary perceptron, though the transition is discontinuous in that case [29]. We note that, in the real-weights Boolean perceptron, the regime of perfect generalization is reached only in the limit $\alpha \to \infty$ [8]. The question we address in this section is how the training and generalization errors are affected by setting to zero $(1-\kappa)N$ components of the weight configurations that minimize $E_{\kappa=1}$. As the results must necessarily depend on the criterion we employ to choose which weights to eliminate, we consider in the following three rather natural possibilities: only the smaller weights are deleted, only the larger weights are deleted and the weights are deleted randomly.

A. Deletion of the smaller weights

Once the learning stage is finished, we set to zero all weights J_i such that $|J_i| < \omega$, where the threshold ω is chosen so as to guarantee that the fraction of nonvanishing weights equals κ. The performances of the damaged network in realizing the training set and a new input-output pair are measured by the training and generalization errors defined as

$$\epsilon_t^s = \frac{1}{\alpha N} \lim_{\beta\to\infty} \langle\!\langle \langle E(\{J_i\Theta(|J_i| - \omega)\}) \rangle_T \rangle\!\rangle \quad (4.1)$$

and

$$\epsilon_g^s = \lim_{\beta\to\infty} \langle\!\langle \langle E_g(\{J_i\Theta(|J_i| - \omega)\}) \rangle_T \rangle\!\rangle , \quad (4.2)$$

where E and E_g are given in (2.5) and (2.6), respectively. Here $\Theta(x) = 1$ if $x > 0$ and 0 otherwise, and the thermal average $\langle\,\rangle_T$ is taken with the Boltzmann weights $\exp[-\beta E_{\kappa=1}(\{J_i\})]$, as discussed in the beginning of this section. The relation between the threshold ω and the connectivity κ can be obtained by calculating explicitly the fraction of weights such that $|J_i| > \omega$, i.e.,

$$\kappa = \left\langle\!\!\left\langle \left\langle \frac{1}{N}\sum_i^N \Theta(|J_i| - \omega) \right\rangle_T \right\rangle\!\!\right\rangle . \quad (4.3)$$

The evaluation of the averages in Eqs. (4.1), (4.2), and (4.3) is rather involved but can be carried out straightforwardly by following the same procedure mentioned in Sec. II for the calculation of the average generalization error. Performing the calculations we find the following simple equation relating ω and κ:

$$\kappa = 2\int_{\omega/\sqrt{Q_1}}^{\infty} Dz, \quad (4.4)$$

where Q_1, the squared norm of the weight configurations that minimize the cost function $E_{\kappa=1}$, is obtained by setting $\kappa = 1$ in Eqs. (3.14) or (3.21) depending on whether $\alpha > 1$ or $\alpha \le 1$, respectively. The result for the average training error is

$$\epsilon_t^s/M = \frac{1-\Lambda_\kappa}{2}\left(1 - \Lambda_\kappa + \frac{\alpha\Lambda_\kappa\left(1-\gamma^2\alpha\right)}{1-\alpha}\right), \quad \alpha \le 1 \quad (4.5)$$

and

$$\epsilon_t^s/M = \frac{\alpha^2\left(1-\gamma^2\Lambda_\kappa\right) - \alpha\left[1+\Lambda_\kappa\left(1-2\gamma^2\right)\right] + \Lambda_\kappa\left(2-\Lambda_\kappa\right)\left(1-\gamma^2\right)}{2\alpha\left(\alpha-1\right)}, \quad \alpha > 1. \quad (4.6)$$

In the limit of large α, this last equation can be written as

$$\epsilon_t^s/M = \frac{1-\gamma^2\Lambda_\kappa}{2} - \frac{\Lambda_\kappa\left(1-\gamma^2\right)}{2\alpha} + O(\alpha^{-2}). \quad (4.7)$$

The average generalization error is given by

$$\epsilon_g^s/M = \frac{\alpha^2\Lambda_\kappa\gamma^2 - \alpha\left[1-\Lambda_\kappa\left(1-2\gamma^2\right)\right] + 1}{2\left(1-\alpha\right)}, \quad \alpha \le 1 \quad (4.8)$$

and

$$\epsilon_g^s/M = \frac{1}{2} + \frac{\Lambda_\kappa\left(1-\gamma^2\alpha\right)}{2\left(\alpha-1\right)}, \quad \alpha > 1 \quad (4.9)$$

which, for large α can be written as

$$\epsilon_g^s/M = \frac{1-\gamma^2\Lambda_\kappa}{2} + \frac{\Lambda_\kappa\left(1-\gamma^2\right)}{2\alpha} + O(\alpha^{-2}). \quad (4.10)$$

B. Deletion of the larger weights

In this case we set to zero all weights J_i such that $|J_i| > \omega$, where again ω must be chosen so as to guarantee that the fraction of nonvanishing weights equals κ. The definitions of the average training and generalization errors, ϵ_t^l and ϵ_g^l, are obtained simply by replacing $\Theta(x)$ by $1 - \Theta(x)$ in Eqs. (4.1) and (4.2). As a result, the equations for ϵ_t^l and ϵ_g^l are given by their counterparts in the case of the deletion of the smaller weights with Λ_κ replaced by $1-\Lambda_{1-\kappa}$, where

$$\Lambda_{1-\kappa} = 2\int_{\lambda_{1-\kappa}}^{\infty} Dz\; z^2 \qquad (4.11)$$

and $\lambda_{1-\kappa}$ is the unique solution of

$$1-\kappa = 2\int_{\lambda_{1-\kappa}}^{\infty} Dz. \qquad (4.12)$$

A numerical analysis of these equations yields $\Lambda_\kappa \geq \kappa \geq 1-\Lambda_{1-\kappa}$.

C. Random deletion of weights

Finally, we consider now the third possibility of deleting weights after learning, which consists of setting to zero $(1-\kappa)N$ randomly chosen weights J_i. The training and generalization performances of the damaged network are measured, respectively, by the quantities

$$\epsilon_t^r = \frac{1}{\alpha N}\lim_{\beta\to\infty}\langle\langle\langle\langle E(\{c_iJ_i\})\rangle_T\rangle\rangle\rangle_c \qquad (4.13)$$

and

$$\epsilon_g^r = \lim_{\beta\to\infty}\langle\langle\langle\langle E_g(\{c_iJ_i\})\rangle_T\rangle\rangle\rangle_c, \qquad (4.14)$$

where $\langle\,\rangle_c$ stands for the quenched average over the statistically independent random variables c_i distributed according to

$$P(c_i) = \kappa\,\delta(c_i-1) + (1-\kappa)\,\delta(c_i). \qquad (4.15)$$

Thus, the dilution constraint is fulfilled in the average only. In the thermodynamic limit, however, the probability that this constraint is violated can be safely neglected. As $E_{\kappa=1}$, which is implicit in the thermal average, does not depend on c_i, the averages over these random variables can be readily performed yielding

$$\epsilon_t^r = \frac{1}{\alpha N}\lim_{\beta\to\infty}\left\langle\!\!\left\langle\left\langle E(\{\kappa J_i\}) - \frac{1}{2}\,\alpha\kappa(1-\kappa)NQ\right\rangle_T\right\rangle\!\!\right\rangle \qquad (4.16)$$

and

$$\epsilon_g^r = \lim_{\beta\to\infty}\left\langle\!\!\left\langle\left\langle\frac{1}{2}(M+\kappa Q-2\gamma\kappa R)\right\rangle_T\right\rangle\!\!\right\rangle, \qquad (4.17)$$

where Q, R, and M are given in (2.9), (2.10), and (2.11), respectively. The evaluation of the quenched $\langle\langle\,\rangle\rangle$ and the thermal $\langle\,\rangle_T$ averages follows the same procedure mentioned before. The result for the average training error is

$$\epsilon_t^r/M = \frac{1-\kappa}{2}\left(1-\kappa+\frac{\alpha\kappa(1-\gamma^2\alpha)}{1-\alpha}\right), \quad \alpha\leq 1 \qquad (4.18)$$

and

$$\epsilon_t^r/M = \frac{\alpha^2(1-\kappa\gamma^2) - \alpha\left[1+\kappa(1-2\gamma^2)\right] + \kappa(2-\kappa)(1-\gamma^2)}{2\alpha(\alpha-1)}, \quad \alpha>1, \qquad (4.19)$$

which, for large α, becomes

$$\epsilon_t^r/M = \frac{1-\gamma^2\kappa}{2} - \frac{\kappa(1-\gamma^2)}{2\alpha} + O(\alpha^{-2}). \qquad (4.20)$$

The average generalization error is given by

$$\epsilon_g^r/M = \frac{\alpha^2\gamma^2\kappa - \alpha\left[1-\kappa(1-2\gamma^2)\right]+1}{2(1-\alpha)}, \quad \alpha\leq 1 \qquad (4.21)$$

and

$$\epsilon_g^r/M = \frac{1}{2} + \frac{\kappa(1-\gamma^2\alpha)}{2(\alpha-1)}, \quad \alpha>1, \qquad (4.22)$$

whose asymptotic behavior is

$$\epsilon_g^r/M = \frac{1-\gamma^2\kappa}{2} + \frac{\kappa(1-\gamma^2)}{2\alpha} + O(\alpha^{-2}). \qquad (4.23)$$

We note again the remarkable formal similarity between the above equations and the ones describing the deletion of the smaller weights: the description of the random cutting of weights could be obtained by simply replacing Λ_κ by κ in the equations describing the deletion of the smaller weights. Another point worth emphasizing is that for $\gamma = 1$ and $\alpha > 1$ the generalization error coincides with the training error, being independent of α. This is true for all types of dilution after learning. In fact, this result was expected since in this regime the global minimum $\{J_i = J_i^0\}$ does not depend on α.

V. ANALYSIS OF THE RESULTS

A peculiarity of the linear perceptron is the occurrence of divergences at the storage capacity α_c of the network. In fact, as signaled by the divergence of Q, some weights become arbitrarily large at α_c, and any external disturbance, whether due to noise or dilution, may cause the training and generalization errors to diverge. In the case of dilution before learning, the training error is always finite; the generalization error, however, diverges at $\alpha = \alpha_c$ for all γ and $\kappa < 1$. In the case of dilution after learning, both the training and generalization errors diverge at $\alpha = 1$, the storage capacity of the nondiluted network, for $\gamma < 1$ and $\kappa < 1$. We note that α_c is independent of the noise parameter γ. In fact, the storage capacity of the linear perceptron is solely determined by the breakdown of the linear independence condition between the rows and columns of the $P \times N$ matrix composed of the random variables S_i^l. As the choice of $\gamma < 1$ does not affect the statistical independence of the random variables S_i^l, this parameter cannot change the value of α_c. Since the only effect of the variance of the weights of the teacher network is to rescale the training and generalization errors, we will set $M = 1$ in the following.

The minimal training error is, of course, always obtained for the annealed dilution, since in this case the deleted weights are chosen so as to attenuate the effect of dilution on that error. The comparison between the training errors obtained for different types of dilution for the noiseless ($\gamma = 1$) case is presented in Figs. 2 and 3 for $\alpha = 0.5$ and $\alpha = 2.0$, respectively. We note that the deletion of the larger weights is the type of dilution that causes more damage to the training error.

In the noiseless case and for $\alpha > 1$, the deletion of the smaller weights actually yields the optimal generalization performance, since in this case the generalization function Eq. (2.6) is rewritten as

$$E_g\left(\{J_i = J_i^0\}\right) = \frac{1}{N}\sum_i \left(J_i^0\right)^2, \tag{5.1}$$

where the summation is restricted to the set of deleted weights only. Thus, in order to minimize E_g we must minimize the norm of the deleted weights, which can be achieved by deleting the smaller weights. For $\alpha < 1$, however, we were unable to prove the optimality of the deletion of the smaller weights as the microscopic configurations of the global minima are not known. Nevertheless, at least for the types of dilution we have considered, the strategy of deleting the smaller weights yields the best generalization performance. This optimality is illustrated in Figs. 4 and 5, where the generalization error is shown as function of κ for $\alpha = 0.5$ and $\alpha = 2.0$, respectively.

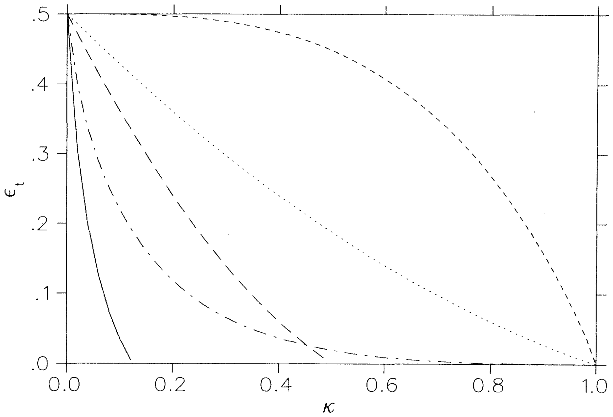

FIG. 2. Training error ϵ_t as a function of the connectivity κ for the annealed dilution (solid curve), quenched dilution (long broken curve), deletion of the smaller weights (chain curve), deletion of the larger weights (short broken curve), and random deletion of weights (dotted curve). The parameters are $\gamma = 1$ and $\alpha = 0.5$.

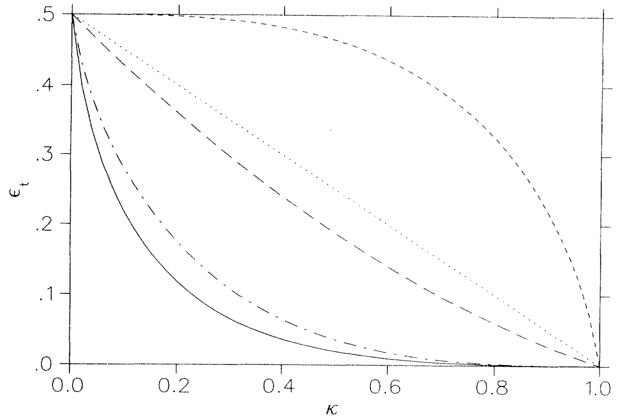

FIG. 3. Training error ϵ_t as a function of the connectivity κ for $\gamma = 1$ and $\alpha = 2.0$. The convention is the same used in Fig. 2.

Figures 2–5 allow us to compare the effect of dilution on the storage performance with the effect on the generalization ability of the perceptron. While the dilution after learning affects the training and the generalization errors in a similar way, the effect of dilution before learning is much more pronounced on the generalization error, which actually diverges at α_c, than on the storage performance.

In the noisy case $\gamma < 1$, the strategy of deleting the smaller weights after learning no longer yields the optimal generalization performance. In particular, we find $\epsilon_g^s > \epsilon_g^r > \epsilon_g^l$ if $2 - 1/\gamma^2 < \alpha < 1/\gamma^2$, otherwise the inequalities are simply reversed. Thus, we are lead to

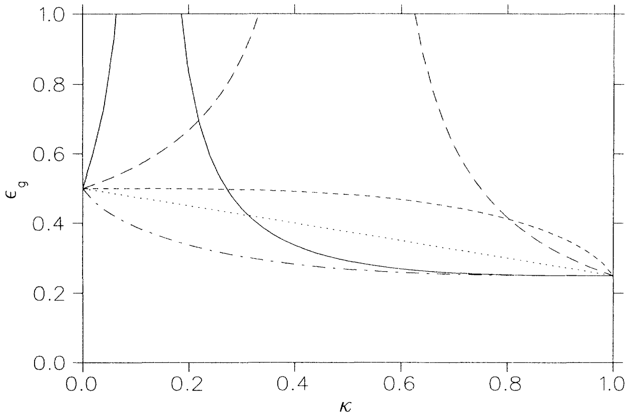

FIG. 4. Generalization error ϵ_g as a function of the connectivity κ for $\gamma = 1$ and $\alpha = 0.5$. The divergences occur at $\kappa = 0.12$ and $\kappa = 0.5$ for the annealed and quenched dilutions, respectively. The convention is the same used in Fig. 2.

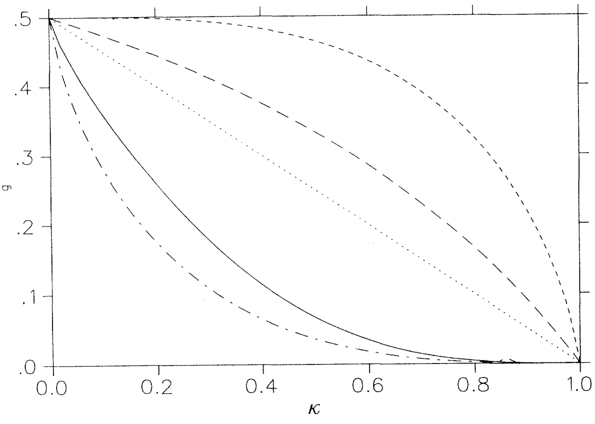

FIG. 5. Generalization error ϵ_g as a function of the connectivity κ for $\gamma = 1$ and $\alpha = 2.0$. The convention is the same used in Fig. 2.

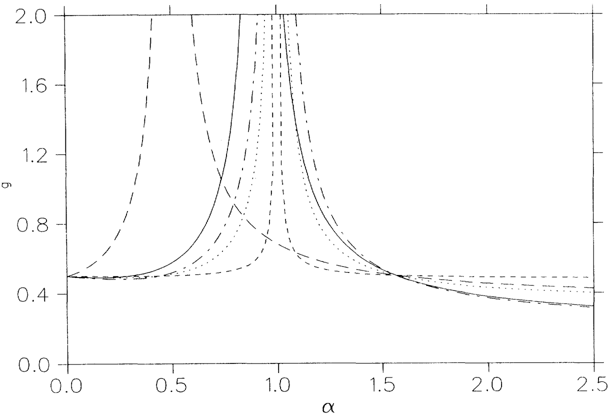

FIG. 7. Generalization error ϵ_g as a function of the training set size α for $\gamma = 0.8$ and $\kappa = 0.5$. The divergences occur at $\alpha = 0.93$ and $\alpha = 0.5$ for the annealed and quenched dilutions, respectively. The convention is the same used in Fig. 2.

the rather odd conclusion that the larger the noise, the more important the role of the small weights. When the network is near its storage capacity $\alpha = 1$ the generalization performance is more sensitive to the deletion of the smaller weights, even in the case of small noise $\gamma \approx 1$. These results are illustrated in Figs. 6 and 7 which present the training and generalization errors, respectively, as functions of α for $\gamma = 0.8$ and $\kappa = 0.5$. Independently of the type of dilution, we find $\epsilon_g = 1/2$ for $\alpha = 1/\gamma^2$. The curves of the generalization error for the dilution after learning intersect again at $\alpha = 2 - 1/\gamma^2$, corresponding to $\epsilon_g = 1/2$. It is interesting to note that for $2 - 1/\gamma^2 < \alpha < 1/\gamma^2$ the diluted network ($\kappa < 1$) generalizes better than the nondiluted one ($\kappa = 1$), independently of the magnitude of the deleted weights.

In the asymptotic regime, $\alpha \to \infty$, the optimal generalization error must tend, from above, to the optimal training error, which we know to be given by the annealed dilution $\epsilon_t^{ann} \to \left(1 - \gamma^2 \Lambda_\kappa\right)/2$. This result, however, coincides with the training error obtained for the deletion of the smaller weights Eq. (4.7), so that both types of dilution give the correct leading term of the optimal generalization error, though ϵ_g^s tends to this limiting value faster than ϵ_g^{ann}.

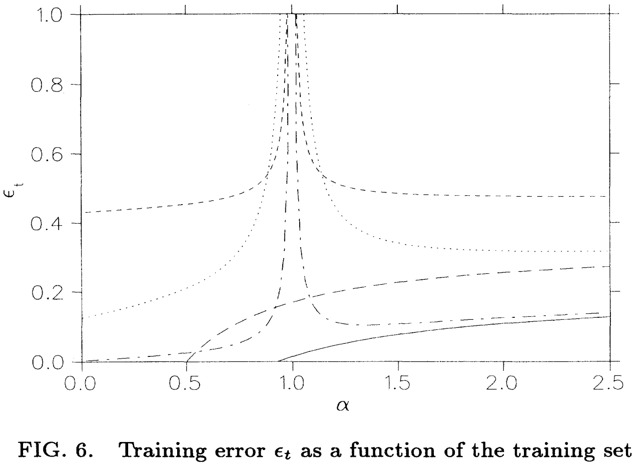

FIG. 6. Training error ϵ_t as a function of the training set size α for $\gamma = 0.8$ and $\kappa = 0.5$. The convention is the same used in Fig. 2.

VI. CONCLUSION

In this paper we have investigated the effects of different types of dilution of synaptic weights on the learning capability of the linear perceptron. Although the problem of lesioning a neural network *before* the learning process takes place has received considerable attention recently [10, 16, 23], the problem of lesioning the network *after* the learning stage has finished has remained practically untouched. The main motivation for this type of analysis is the identification of the components whose destruction may affect more severely the generalization performance of the neural network. In the case of the linear perceptron, the relative importance of the roles played by weights of different magnitudes depends on the training set size α and on the noise parameter γ. More specifically, the generalization performance is more sensitive to deletion of the smaller weights if $2 - 1/\gamma^2 < \alpha < 1/\gamma^2$. Outside this range, the larger weights give the more important contribution to the generalization performance. It would be interesting to know whether these results hold, at least qualitatively, for the real-weights Boolean perceptron as well.

For finite training set sizes α, the widespread strategy of choosing the weights to be deleted so as to lessen the effect of the lesion on the training error (annealed dilution) does not yield the best generalization performance. In the noiseless case, for instance, it is overcome by the deletion of the smaller weights after learning, which actually yields the optimal performance for $\alpha > 1$.

Besides the local stability of the replica-symmetric solution, another important advantage of the linear perceptron as compared with the Boolean perceptron is the possibility of solving analytically the saddle-point equations and hence obtaining neat equations for the training and generalization errors. To conclude, we should mention that the exactness of the results presented in this paper depends on the global stability of the replica-symmetric solution. To prove this type of stability we should, for instance, show that the replica symmetric is the unique solution to the full replica-symmetry breaking saddle-point equations. We are content, however, with the proof of the local stability of our solution. Additional evidence for the exactness of our results is provided by the rigorous solution of a very similar model, the spherical model of a spin glass, which was shown to coincide with the replica-symmetric solution [30].

ACKNOWLEDGMENT

This work was supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

[1] G. E. Hinton, D. C. Plaut, and T. Shallice, Sci. Am. **269**, 76 (1993).

[2] D. C. Plaut and T. Shallice, Cognitive Neuropsych. **10**, 377 (1993).

[3] A. Krogh and J. A. Hertz, J. Phys. A **25**, 1135 (1992).

[4] E. Gardner, J. Phys. A **21**, 257 (1988).

[5] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).

[6] J. F. Fontanari, J. Phys. A **26**, 6147 (1993).

[7] E. Gardner and B. Derrida, J. Phys. A **22**, 1983 (1989).

[8] G. Györgyi and N. Tishby, in *Neural Networks and Spin Glasses*, edited by W. K. Theumann and R. Köberle (World Scientific, Singapore, 1990), pp. 3–36.

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

[10] D. M. L. Barbato and J. F. Fontanari, J. Phys. A **26**, 1847 (1993).

[11] M. A. Virasoro, Europhys. Lett. **7**, 293 (1988).

[12] K. Binder and A. P. Young, Rev. Mod. Phys. **58**, 801 (1986).

[13] M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[14] R. Erichsen, Jr. and W. K. Theumann, J. Phys. A **26**, L61 (1993).

[15] P. Majer, A. Engel, and A. Zippelius, J. Phys. A **26**, 7405 (1993).

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

[17] J. J. Hopfield, Proc. Natl. Acad. Sci. USA **79**, 2554 (1982).

[18] H. Sompolinsky, Phys. Rev. A **34**, 2571 (1986).

[19] H. Sompolinsky, in *Heidelberg Colloquium on Glassy Dynamics*, edited by J. L. van Hemmen and I. Morgenstern, Lecture Notes in Physics Vol. 275 (Springer, Berlin, 1987), pp. 485–527.

[20] B. Derrida, E. Gardner, and A. Zippelius, Europhys. Lett. **4**, 167 (1987).

[21] E. Gardner, J. Phys. A **22**, 1969 (1989).

[22] N. Brunel, J. Phys. I (France) **3**, 1693 (1993).

[23] M. Bouten, A. Engel, A. Komoda, and R. Serneels, J. Phys. A **23**, 4643 (1990).

[24] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, San Francisco, 1979).

[25] D. M. L. Barbato and J. F. Fontanari, J. Phys. A **27**, 8029 (1994).

[26] T. Kohonen, *Self-Organization and Associative Memory* (Springer-Verlag, Berlin, 1984).

[27] P. H. Lindsay and D. A. Norman, *Human Information Processing* (Academic, New York, 1977).

[28] J. R. de Almeida and D. J. Thouless, J. Phys. A **11**, 983 (1978).

[29] G. Györgyi, Phys. Rev. A **41**, 7097 (1990).

[30] J. M. Kosterlitz, D. J. Thouless, and R. C. Jones, Phys. Rev. Lett. **36**, 1217 (1976).