Canonical ensemble approach to graded-response perceptrons

D. Bollé[1,*] and R. Erichsen, Jr.[2]
[1]*Instituut voor Theoretische Fysica, KU Leuven, B-3001 Leuven, Belgium*
[2]*Instituto de Física, Universidade Federal do Rio Grande do Sul, Caixa Postal 15051, 91501-970 Porto Alegre, RS, Brazil*
(Received 20 March 1998)

Perceptrons with graded input-output relations and a limited output precision are studied within the Gardner-Derrida canonical ensemble approach. Soft non-negative error measures are introduced allowing for extended retrieval properties. In particular, the performance of these systems for a linear (quadratic) error measure, corresponding to the perceptron (adaline) learning algorithm, is compared with the performance for a rigid error measure, simply counting the number of errors. Replica-symmetry-breaking effects are evaluated, and the analytic results are compared with numerical simulations. [S1063-651X(99)04503-1]

PACS number(s): 87.10.+e, 64.60.Cn

I. INTRODUCTION

Graded-response perceptrons constitute the basic building blocks of layered architectures trained by the back propagation algorithm. This motivates the interest in these systems in recent years. Questions pertaining to retrieval properties of specific architectures [1–5], to optimal capacities of networks designed to perform a given storage task [6,7] and to generalization abilities [8] have been adressed by statistical mechanics approaches.

A problem still open for these graded-response perceptrons is the development of a Gardner-Derrida (GD) type analysis [9] in order to study the optimal storage properties when allowing errors. The solution of this problem is the purpose of the present paper. On the one hand, this extends our results [6,7] on the optimal capacity of graded-response perceptrons in the framework of the Gardner theory [10]. On the other hand, the relevant cost functions used in our analysis here define a perceptron and an adaline learning algorithm, which are both of special practical interest.

The underlying idea of the GD analysis is to view learning in these perceptrons as an optimization process in the space of couplings. By introducing soft non-negative error measures we investigate the canonical ensemble generated by the corresponding cost function in the space of couplings using the replica method. In this discussion we allow for a limited output precision in the storage task to be solved by the perceptron. In particular, a linear and a quadratic error measure are investigated. The corresponding cost functions define, respectively, a perceptron and an adaline learning algorithm through the method of gradient descent. Replica-symmetric (RS) and first-step replica-symmetry-breaking (RSB) solutions for the storage capacity, the average output error, and the local field distributions are studied. For comparison we also derive the results for the rigid GD error measure that simply counts the number of errors. One of the specific aims of this work is to determine which cost function is the most efficient one for learning with errors.

For the case of two-state attractor neural networks the canonical ensemble approach advocated in Ref. [9] has been streamlined and extended to other cost functions than the rigid one [11]. The methods and results obtained there are, of course, also relevant for perceptron networks. First-step RSB effects above the critical capacity have then been studied in [12] for binary perceptron networks with a GD cost function and have been extended to other cost functions [13,14]. Recently, it has been shown [15] for the GD cost function that in the region above the critical capacity full RSB is necessary for an exact solution. A direct evaluation of the two-step RSB solution has been performed in this case, yielding a minimum storage error only slightly greater than the one-step RSB. The conclusion was put forward that for most practical purposes one-step RSB will be adequate. In this study we also want to find out whether a similar conclusion is valid for the problem at hand.

The rest of this paper is organized as follows. In Sec. II we briefly review the canonical approach adapted to the graded-response perceptron and introduce the different cost functions we want to consider: the rigid one, the linear one, and the quadratic one. Section III contains the replica theory for these cost functions and determines the critical storage capacity, the distribution of the local fields, and the average output error. Both the RS approximation and the first-step RSB are treated for a general monotonic input-output relation. Section V describes the results of this theory applied to two specific, frequently used input-output relations, i.e., the hyperbolic tangent and the piecewise-linear one. These results are compared with some numerical simulations. In Sec. V the most important results are summarized. Finally, the Appendix contains the technical details of the derivations.

II. CANONICAL ENSEMBLE APPROACH

The task to be solved by the graded-response perceptron is to map a collection of input patterns $\{\xi_i^\mu; 1 \leqslant i \leqslant N\}$, $1 \leqslant \mu \leqslant p$, onto a corresponding set of outputs ζ^μ, $1 \leqslant \mu \leqslant p$, via

$$\zeta^\mu = g(\gamma h^\mu), \tag{1}$$

$$h^\mu = \frac{1}{\sqrt{N}} \sum_j J_j \xi_j^\mu . \tag{2}$$

*Also at Interdisciplinair Centrum voor Neurale Netwerken, K.U. Leuven, Leuven, Belgium.

Here g is the input-output relation of the perceptron, which is assumed to be a monotonic nondecreasing function. In Eq. (1) γ denotes a gain parameter, and h^μ is the local field generated by the inputs $\{\xi_i^\mu\}$ as specified in Eq. (2). The J_j are couplings of an architecture of perceptron type. We restrict our attention to general unbiased input patterns specified by $\langle \xi_i^\mu \rangle = 0$ and $\langle \xi_i^\mu \xi_j^\nu \rangle = \delta_{\mu,\nu}\delta_{i,j}C$. These two parameters are sufficient in specifying the input pattern distribution. Since the effect of C in Eq. (1) can be absorbed in the gain parameter we take $C=1$ in the sequel.

We explicitly allow a limited output precision in the mapping (1). In other words the output that results when the input layer is in the state $\{\xi_i^\mu\}$ is accepted if

$$g(\gamma(h^\mu)) \in I_{\rm out}(\zeta^\mu,\epsilon) \equiv [\zeta^\mu - \epsilon, \zeta^\mu + \epsilon], \quad \mu = 1, \ldots, p, \tag{3}$$

where ϵ denotes the allowed output-error tolerance.

The strategy of the canonical approach is to require the graded-perceptron network to go through a learning stage in the space of couplings in order to find for the absolute minima of a given cost function $E(\{h^\mu\},\{\zeta^\mu\})$ precisely networks with the properties (3). This cost function is assumed to be a sum of local terms for each pattern μ,

$$E(\{h^\mu\},\{\zeta^\mu\}) = \sum_\mu V(h^\mu,\zeta^\mu). \tag{4}$$

The different cost functions that will be studied here can be put into the form

$$V(h^\mu,\zeta^\mu) = W_s(\zeta^\mu - \epsilon - g(\gamma h^\mu)) + W_s(g(\gamma h^\mu) - \zeta^\mu - \epsilon), \tag{5}$$

where

$$W_s(x) = x^s \theta(x), \tag{6}$$

and $\theta(x)$ is the Heaviside step function. For $s=0$ we get the GD cost function, which simply counts the number of the errors, irrespective of their size. Moreover, we consider a linear cost function ($s=1$), where the errors are weighted proportionally to their magnitudes and a quadratic cost function ($s=2$) where the errors are weighted proportional to the square of their magnitudes. The relevance of this choice becomes clear when applying gradient descent dynamics to Eq. (4) with the result

$$\Delta J_j = \frac{s\gamma\delta}{\sqrt{N}} \sum_\mu \left(\xi_j^\mu - \frac{J_j h^\mu}{\sqrt{N}} \right) [W_{s-1}(\zeta^\mu - \epsilon - g(\gamma h^\mu)) + W_{s-1}(g(\gamma h^\mu) - \zeta^\mu - \epsilon)] g'(\gamma h^\mu), \tag{7}$$

where the prime denotes the derivative with respect to the argument. Taking $s=1$ ($s=2$) in this expression, we find the perceptron (adaline) learning algorithm with step size δ for the graded perceptron. The GD cost function does not correspond to any learning algorithm.

III. REPLICA THEORY

The physical properties of the graded-response perceptron network defined above are derived by investigating the canonical ensemble generated by the free energy

$$f(\beta) = -\lim_{N\to\infty} \frac{1}{N\beta} \ln Z, \tag{8}$$

where Z is the partition function

$$Z = \int \prod_j dJ_j \prod_j \delta\left(\sum_j J_j^2 - N \right) \exp[-\beta E(\{h^\mu\},\{\zeta^\mu\})]. \tag{9}$$

In Eq. (9) the mean spherical constraint $\Sigma_i J_i^2 = N$ is adopted to fix a scale for the gain parameter γ of the input-output relation. We are interested in the limit $\beta\to\infty$ in which the free energy gives information about the fraction of patterns that are stored incorrectly. In the usual way the free energy is assumed to be self-averaging with respect to the inputs $\{\xi^\mu\}$ and the outputs $\{\zeta^\mu\}$. This average, denoted by $\langle f(x) \rangle_{\{\xi^\mu\},\{\zeta^\mu\}} \equiv \langle f \rangle$, can be performed by applying the replica trick. The standard order parameter that appears in such a replica calculation is the overlap between two distinct replicas in coupling space,

$$q_{\lambda\lambda'} \equiv \frac{1}{N} \sum_{i=1}^N J_i^\lambda J_i^{\lambda'} \quad \lambda < \lambda', \quad \lambda,\lambda' = 1, \ldots, n. \tag{10}$$

Elsewhere we consider both the replica symmetry analysis and the one-step breaking effects (RSB1). We also suppress the index μ.

In the RS analysis we assume that

$$q_{\lambda\lambda'} = q, \quad \lambda < \lambda'. \tag{11}$$

The optimal capacity properties of the system are obtained in the limit $\beta\to\infty$, $q\to 1$, with $\beta(1-q)=x$ taking a finite value. In this limit, a standard calculation analogous to the binary perceptron problem [9,11] leads to the averaged free energy,

$$\langle f \rangle = \mathrm{extr}_x \left\{ -\frac{1}{2x} + \alpha \left\langle \int Dt \min_h [F_{\rm RS}(h,\zeta,x,t)] \right\rangle_{\{\zeta\}} \right\}, \tag{12}$$

with

$$F_{\rm RS}(h,\zeta,x,t) = V(h,\zeta) + \frac{(h-t)^2}{2x}, \tag{13}$$

and where $Dt = (dt/\sqrt{2\pi})\exp(-t^2/2)$, $\alpha = p/N$ denotes the storage capacity, and $\langle \cdots \rangle_{\{\zeta\}}$ indicates the average over the distribution of the output patterns.

Let us denote by $h_0(\zeta,x,t)$ the value of h that minimizes $F(h,\zeta,x,t)$. For a determined storage capacity α the variable x is given by the saddle-point equation $\partial\langle f \rangle/\partial x = 0$, which can be rewritten in the form

$$\alpha_{\rm RS}^{-1} = \left\langle \int Dt [h_0(\zeta,x,t) - t]^2 \right\rangle_{\{\zeta\}}. \tag{14}$$

We immediately remark that these results are not always stable against RSB. Following standard considerations [9,16,17] the stability condition reads

$$\alpha_{\mathrm{RS}}\left\langle \int Dt\left[\frac{d}{dt}[h_0(\zeta,x,t)-t]\right]^2\right\rangle_{\{\zeta\}}<1. \tag{15}$$

For the exact mapping task where $\epsilon=0$ the result found in [6] for the critical storage capacity corresponding to the GD cost function is retrieved when we take the limit $x\to\infty$ in Eq. (14)

$$\alpha_c^{-1}=1+\langle h_\zeta^2\rangle_{\{\zeta\}}, \tag{16}$$

with

$$h_\zeta=\frac{1}{\gamma}g^{-1}(\zeta). \tag{17}$$

Similar to binary networks [11,13], α_c is the same for all cost functions. Clearly, for $\alpha>\alpha_c$ errors will be introduced that depend both in quantity and in size on the specific cost function used. An interesting expression to look at in this respect is the distribution of local fields since it provides more information on the deviation of the errors from the correct output ζ. For a given desired output ζ, it is defined as

$$\rho(h|\zeta)=\left\langle \delta\left(h-\frac{1}{\sqrt{N}}\sum_{j=1}^{N}J_j\xi_j\right)\right\rangle_{\{J\},\{\xi\}}, \tag{18}$$

where the thermal average over J is taken subject to the mean spherical constraint introduced before. Following Kepler and Abbott [18], we find for the graded perceptron

$$\rho_{\mathrm{RS}}(h|\zeta)=\int Dt\,\delta(h-h_0(\zeta,x,t)). \tag{19}$$

An overall measure of the network performance is given by the average output error

$$\mathcal{E}=\langle\mathcal{E}(\zeta)\rangle_{\{\zeta\}}, \tag{20}$$

where the ζ-dependent output error $\mathcal{E}(\zeta)$ is given by

$$\mathcal{E}(\zeta)=\int dh\,\rho_{\mathrm{RS}}(h|\zeta)[W_1(\zeta-\epsilon-g(\gamma h)) + W_1(g(\gamma h)-\zeta-\epsilon)]. \tag{21}$$

From the results in the literature on the binary perceptron problem [12,13,15] and from our former studies on the graded perceptron system [6,7] we expect RSB effects. So we want to improve the RS results by applying the first step of Parisi's RSB scheme [19]. We, therefore, introduce the following order parameters:

$$q_{\lambda\lambda'}=q_{\beta_1\beta_2}^{\alpha_1\alpha_2}=\begin{cases}q_1, & \text{if } \alpha_1=\beta_1,\\ q_0, & \text{if } \alpha_1\neq\beta_1,\end{cases} \tag{22}$$

where $\alpha_1,\beta_1=1,\ldots,n/m;\alpha_2,\ \beta_2=1,\ldots,m$ and $1\leqslant m\leqslant n$. We remark that in the limit $n\to0$, $0\leqslant m\leqslant1$.

Similar to [13] we find after a standard but tedious calculation that in the limit $q_1\to1^-,m\to0$ and $0\leqslant q_0\leqslant q_1$ with $m/(1-q_1)=M$ a finite value and $x=\beta(1-q_1)$, the free energy averaged with respect to the inputs $\{\xi\}$ and the outputs $\{\zeta\}$ can be written as

$$\langle f\rangle=\lim_{\beta\to\infty}\max_{x,q_0,M}\left\{-\frac{1}{2Mx}\ln[1+M(1-q_0)] - \frac{q_0}{2x[1+M(1-q_0)]} - \frac{\alpha}{Mx}\left\langle\int Dt_0\ln\Psi(\zeta,x,q_0,M,t_0)\right\rangle_{\{\zeta\}}\right\} \tag{23}$$

with

$$\Psi(\zeta,x,q_0,M,t_0)=\int Dt_1\exp\{-Mx\min_h F_{\mathrm{RSB1}}(h,\zeta,x,q_0,t_0,t_1)\} \tag{24}$$

and

$$F_{\mathrm{RSB1}}(h,\zeta,x,q_0,t_0,t_1)=V(h,\zeta)+\frac{1}{2x}(h-t_0\sqrt{q_0}-t_1\sqrt{1-q_0})^2. \tag{25}$$

For a chosen storage capacity α, the variables x, q_0, and M are given by the saddle point equations $\partial\langle f\rangle/\partial x=0$, $\partial\langle f\rangle/\partial q_0=0$, and $\partial\langle f\rangle/\partial M=0$.

The first-step RSB distribution for the local fields corresponding to pattern ζ becomes

$$\rho_{\mathrm{RSB1}}(h,\zeta)=\int Dt_0\int Dt_1\times\frac{\exp[-MxF_{\mathrm{RSB1}}(h_0,\zeta,x,q_0,t_0,t_1)]\,\delta(h-h_0)}{\Psi(\zeta,x,q_0,M,t_0)}, \tag{26}$$

where $h_0=h_0(\zeta,x,q_0,t_0,t_1)$ is the value of h that minimizes $F_{\mathrm{RSB1}}(h,\zeta,x,q_0,t_0,t_1)$. The average output error in RSB1 approximation is obtained by replacing expression (19) by (26) in (21).

IV. RESULTS FOR SPECIFIC COST FUNCTIONS

The theory outlined in the last section has been applied to the specific cost functions defined in Eqs. (5) and (6). For the input-output relation g we have used both the hyperbolic tangent and the piecewise-linear function

$$g(x)=\begin{cases}x, & \text{for } |x|<1,\\ \mathrm{sgn}(x), & \text{elsewhere.}\end{cases} \tag{27}$$

A priori, our aim is not to compare the macroscopic properties of graded-perceptrons for the two different input-output relations since, in general, they are qualitatively the same. In

fact, the results obtained here are complementary. For the hyperbolic tangent input-output relation the RS solution is found to be stable over an important range of values for the parameters α and γ while in the case of the piecewise-linear input-output relation the RS solution is always unstable. However, from a more technical point of view in the case of the hyperbolic tangent function, the mimimization of $F_{\rm RS}$ in the corresponding averaged RS free energy with respect to the local field h [recall Eqs. (12) and (13)] only leads to an equation defining t as a function of the minimizing value h_0 (see the Appendix). This equation needs to be inverted but depending on the values of $\gamma^2 x$ and ζ the inverse function may be multiple valued and hence a (sometimes very tedious) Maxwell construction is required in order to make it single valued. Consequently, only the RS solution is studied in detail in this case. On the contrary, the piecewise-linear input-output relation permits an explicit calculation of the minimizing values $h_0(\zeta,x,t)$ and $h_0(\zeta,x,q_0,t_0,t_1)$ of the functions $F_{\rm RS}$ and $F_{\rm RSB1}$ in the corresponding averaged free energies. This, in turn, simplifies drastically the calculations and both the RS and RSB1 solutions are completely worked out in this case.

At this point, we remark already that the Maxwell construction in the hyperbolic tangent case gives a discontinuity in $h_0(t)$ having an effect on the stability of the RS solution. Similarly, due to the fact that the piecewise-linear input-output relation is not everywhere differentiable a gap structure in the distribution of the local fields emerges signaling the instability of the RS solution [17]. The effects of RSB for the cost functions (5) and (6) are found to be important.

Elsewhere we present the results of our calculations both for the hyperbolic tangent and the piecewise-linear input-output relations. In order not to interrupt the line of reasoning we refer all technical details of the calculations to the Appendix.

A. Hyperbolic tangent input-output relation

In this subsection we compare the performance of the three cost functions defined in Eqs. (5) and (6) by studying their average output error, $\mathcal{E}$ [recall Eq. (20)]. Our strategy is to consider a linear ($s=1$) and quadratic ($s=2$) ''entirely soft'' cost function versus a ''completely rigid'' one ($s=0$). Soft means that we do not fix the output-error tolerance ϵ, since some outputs might be far away from the correct output ζ. Entirely soft indicates that we work without tolerance at all by putting $\epsilon=0$. For the completely rigid cost function, ϵ was determined in the function of the loading capacity α, by solving (for ϵ) the optimal capacity for the graded perceptron in the microcanonical approach {recall Eq. (9) of Ref. [6]}. We remark that the rigid output error tolerance is almost constant for the whole γ interval considered. We have $\epsilon\simeq 0.105$, $\epsilon\simeq 0.225$, and $\epsilon\simeq 0.553$, respectively, for $\mathcal{E}=0.1$, $\mathcal{E}=0.2$, and $\mathcal{E}=0.4$. The output distribution is taken to be uniform in the interval $[-1,1]$.

The results are presented in Figs. 1 and 2. First, we show in Figs. 1(a)–1(c) the loading capacity α as a function of the gain parameter γ for a constant average output error $\mathcal{E}=0$, 0.1, and 0.2 in the case of the three cost functions. For the rigid cost function, we plot an additional curve for $\mathcal{E}=0.4$ to indicate that the capacity has a maximum for finite γ. For small $\mathcal{E}$ values, the peak is also present, but it is less pronounced and shifted towards higher γ values, similarly as in Ref. [6]. In the case of both the linear and the quadratic cost functions no maximum is found for a finite gain parameter.

Next, it is interesting to discuss the main features related to the de Almeida–Thouless (AT) line, $\alpha_{\rm AT}$, also shown in these figures for the three cost functions. As seen in Fig. 1(a) for the rigid cost function, $\alpha_{\rm AT}$ decreases monotonically with γ, similarly as in Ref. [6]. The RS solution is stable in the (connected) region below this line. For both the linear and the quadratic cost functions, $\alpha_{\rm AT}$ is no longer a monotonically decreasing function of γ. When α becomes larger than α_c, the patterns are stored in such a way that, as discussed below, a ''gap'' structure arises in the distribution of the local fields, for all values of γ. It is well known [17] that this gap destabilizes the RS solution, and consequently RSB is required immediatly above the line $\alpha_c(\gamma)$. For $\alpha<\alpha_c$, the network is not saturated, there is no gap in the distribution of the local fields, and the RS solution is stable. Furthermore, for α much larger than α_c, and sufficiently small γ, the gaps are being filled up, and the RS solution eventually becomes stable. Consequently, there is a RS-stable region also above α_c, but it is disconnected from that below α_c [see Figs. 1(b) and 1(c)]. It is important to repeat that the region of instability between the two stable regions has its origin in the gap structure. Conversely, we recall that for the rigid cost function there are no gaps. For all γ values, there is always a finite interval above α_c where the RS solution is stable. This is the reason why the stability region in this case is connected, with the $\alpha_{\rm AT}$ line extending up to $\gamma\to\infty$.

It is seen that the rigid cost function has the worst performance for all values of γ. For both the linear and the quadratic cost function a monotonically increasing (but bounded) capacity α results. In general, the linear cost function has the best performance. This behavior of the graded perceptron network can be understood in terms of the ''strategy'' used by a specific cost function to arrange the local fields when learning the patterns. The rigid cost function puts all local fields in a connected interval, thereby minimizing its width. It does not try to optimize the learning *inside* the interval in order to decrease the average output error. However, the linear and quadratic soft cost functions do optimize their performance by penalizing the errors linearly or quadratically with their size. They try to arrange the local fields in a close region around the value h_ζ resulting in the correct output ζ under the action of the input-output relation. In both cases the resulting distribution of the local fields shows a sharp peak (a δ peak in the linear case) at h_ζ, and decreasing tails. A gap in between can occur. The tails of the quadratic cost function decrease faster than those of the linear cost function. We will present figures below for the case of the piecewise-linear input-output relation where a similar behavior has been found. It is worth remarking that the linear cost function shows a somewhat improved behavior compared to the quadratic one. This may be justified by recalling that, according to Eq. (21), the output error is the linear distance between the actual and the correct output, and this is exactly the quantity that is minimized by the linear cost function. In

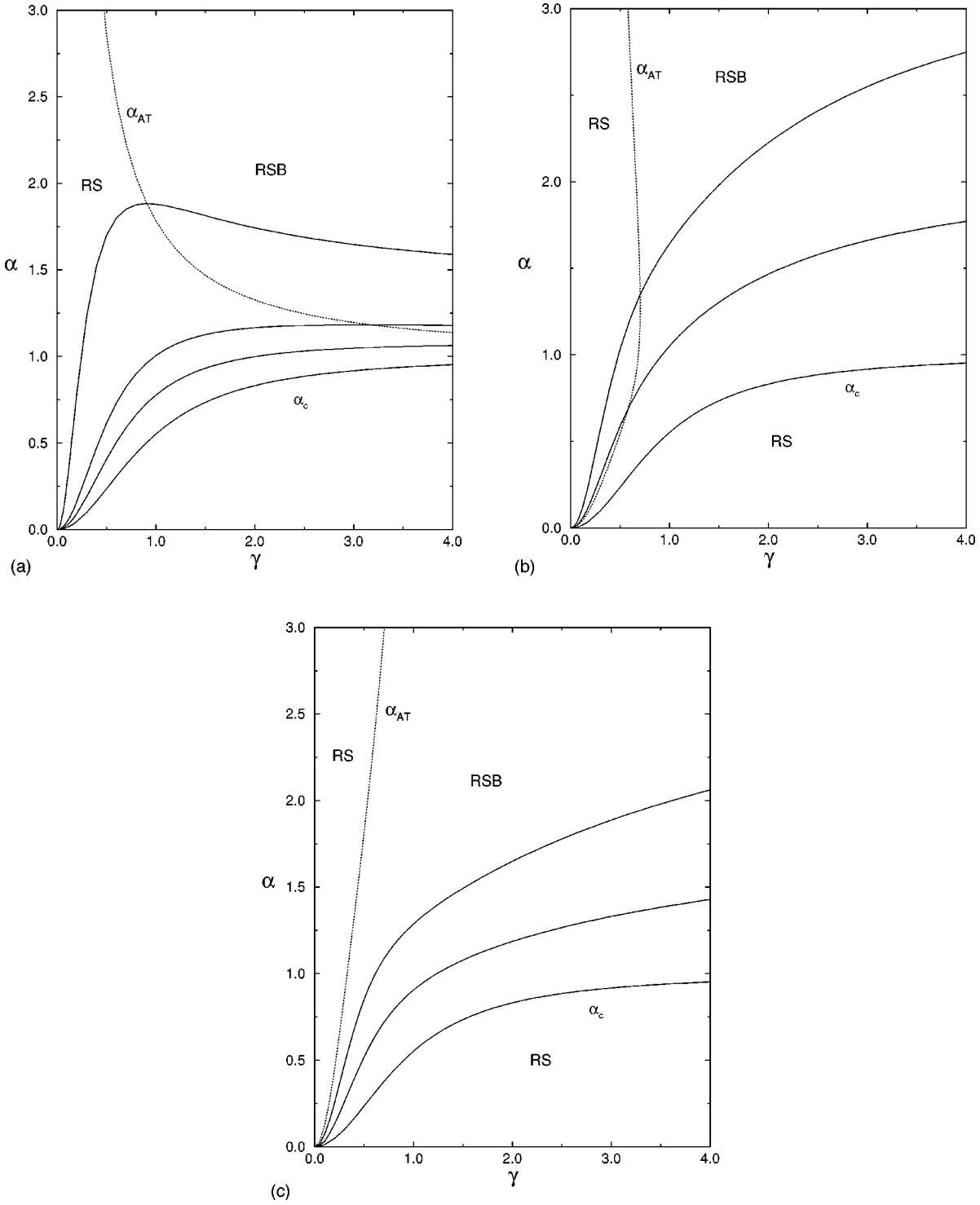

FIG. 1. Storage capacity α for the hyperbolic tangent input-output relation as a function of the gain parameters γ at constant total average output error $\mathcal{E}=0$ (lower curve), 0.1, 0.2, and 0.4 (upper curve) for the GD cost function (a), the linear cost function (b), and the quadratic cost function (c). In (b) and (c) the line for $\mathcal{E}=0.4$ is not shown. The dotted curve is the AT line. The curve α_c is the critical capacity.

other words, the linear cost function is suitably designed to minimize the output error.

Let us discuss then in more detail the gap structure of the local fields, revealed by the line α_g in Figs. 2(a) and 2(b). For the rigid cost function, no gaps are present, since the output tolerance ϵ is chosen such that *all* fields are inside a connected interval. For the linear and quadratic cost functions Figs. 2(a) and 2(b) present the relevant results in the $(\alpha\text{-}\gamma)$ plane. A gap is present in the region between the lines α_c (for $\mathcal{E}=0$) and α_g. For $\alpha<\alpha_c$, the perceptron is not saturated, i.e., $q<1$ and the present calculations do not cover this region. We notice that for small α the gap line lies very close to the AT line. A similar behavior has been noticed in binary networks trained with noisy patterns [20]. For growing $\alpha\geqslant\alpha_c$, the width of the gaps decreases from an infinite value at α_c to become zero as α approaches α_g. In the region between α_g and α_{AT} there are no gaps, but the RS solution remains unstable.

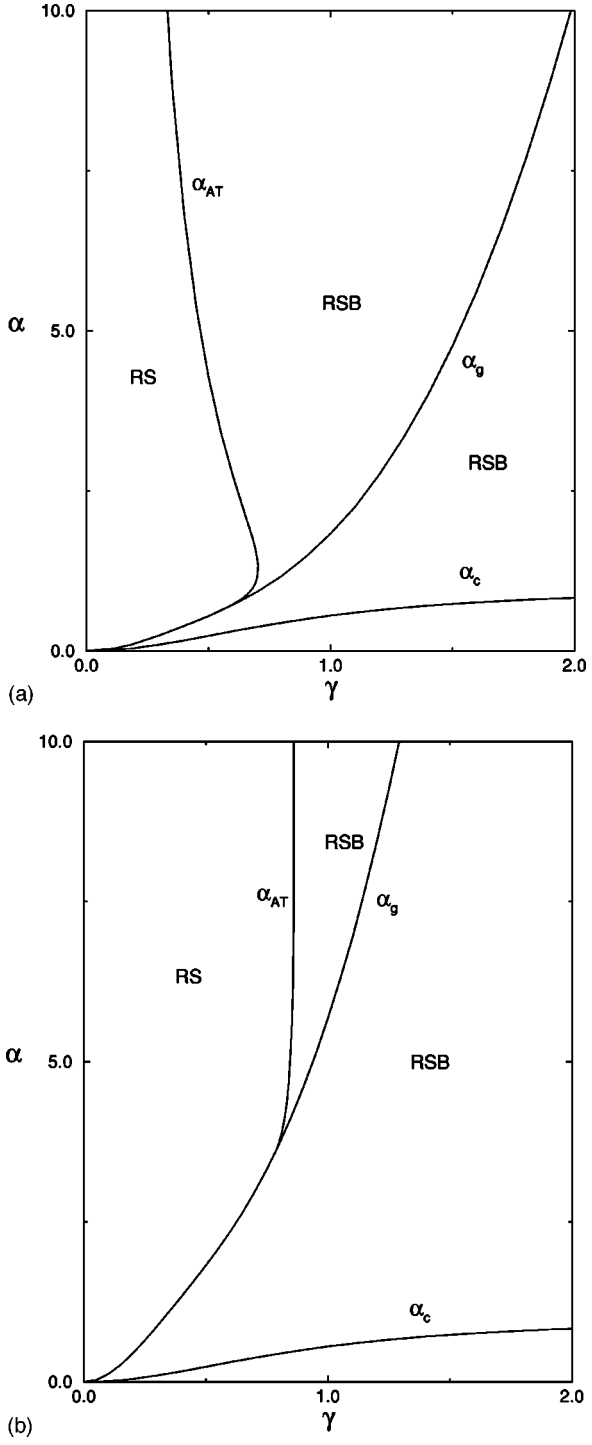

FIG. 2. The gap structure for the hyperbolic tangent input-output relation in the α-γ plane for the linear cost function (a) and the quadratic cost function (b). The curve α_c is the critical capacity, the curve α_g represents the gap line, and α_{AT} is the AT line.

Concerning the stability of our results with respect to RS breaking, we see that for the rigid cost function the curves for the capacity as a function of the gain parameter at constant average output error are ''stable'' starting from $\gamma=0$ up to the point where the curves reach their maximum (in agreement with the results of [6] for constant output-error tolerance). For the linear cost function, the RS curves are stable for small γ and not so small $\mathcal{E}$. However, for the quadratic cost function all the curves for the hyperbolic tangent input-output relation are RS unstable.

The origin of instability against RS breaking fluctuations is relatively easy to understand in the region where gaps in the local field distribution are present [11,17,21]. One can argue that it is not possible to pass continuously from one replica of the system where a specific pattern is learned in one ''band'' of the local fields, to another replica where that pattern is learned in another band. The corresponding solutions are disconnected in the space of replicas, and the overlap between pairs of replicas cannot be the same for all pairs, contrary to the RS assumption. In the region where there are no gaps this argument is no longer valid. Here, one may argue that spreading the local fields over one single but wide band can also disrupt the space of replicas. Maybe the notion of critical bandwidth is relevant here. This could be an interesting subject for further study.

B. Piecewise-linear input-output relation

For the piecewise-linear input-output relation we do consider a nonzero output-error tolerance ϵ, i.e., all the inputs whose corresponding output lie inside the interval $[\zeta-\epsilon,\zeta+\epsilon]$ do not contribute to the average output error. As outlined before the numerical calculations are easier than those for the hyperbolic tangent, and the study of the RSB1 solution in some detail becomes feasible. Numerical results as well as simulations are discussed for $\epsilon=0.5$.

Before passing to these results, it is worth mentioning that the introduction of a fixed ϵ allows us to replace the study of the $s=0$ cost function with a completely rigid constraint discussed in Sec. IV A, by a true GD cost function [(5) and (6) for $s=0$].

In Figs. 3(a) and 3(b) we see both the RS and the RSB1 average output error $\mathcal{E}$ as a function of the loading capacity α for $\gamma=1$ for the three cost functions considered. Figure 3(a) concerns the GD (dotted curves) and quadratic (full curves) cost functions, Fig. 3(b) the linear one. For all cases the upper curves are the RSB1 results, so as expected, $\mathcal{E}_{\mathrm{RSB1}}>\mathcal{E}_{\mathrm{RS}}$ for all $\alpha>\alpha_c$. For comparison, simulations were performed for networks with $N=200$, 400 both for the linear and quadratic cost functions. At this point we remark that the GD cost function does not correspond to any learning algorithm. The circles (diamonds) denote the $N=200$ ($N=400$) results. The agreement with the analytic results is satisfactory in the sense that we know that the overall overestimate of the average output error is due to a very slow convergence to the minimum of the free energy, even using some straightforward techniques of simulated annealing. Much more sophisticated versions of the latter should be used but this is not the purpose of the present work.

In the region of the network parameters considered, the linear cost function gives the best performance. According to the RSB1 results, the least efficient is the quadratic cost function if $\alpha<4.8$, and the GD cost function elsewhere.

Figure 4 shows α as a function of γ at constant $\mathcal{E}$. For each cost function, the upper (lower) curve corresponds to the RS (RSB1) result. For all γ the highest capacity is given by the linear cost function and for $\gamma<\pm 2.5$, the quadratic cost function gives the lowest capacity.

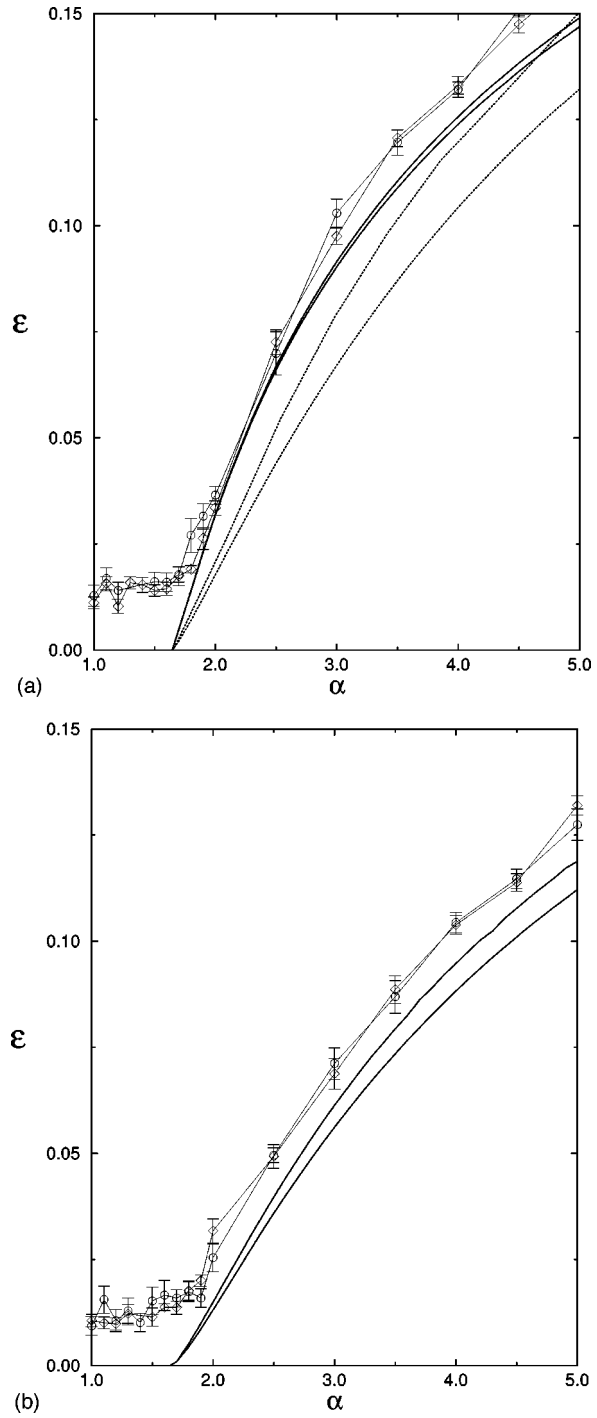

FIG. 3. The total average output error $\mathcal{E}$ as a function of the storage capacity α for output tolerance $\epsilon=0.5$ and gain parameter $\gamma=1$ in the case of the piecewise-linear input-output relation for (a) the GD cost function (dotted lines) and the quadratic cost function (solid lines), and (b) the linear cost function. For each case the upper curve is the RSB1 result, the lower one the RS result. Numerical simulations of the quadratic cost function in (a) and the linear cost function in (b) are indicated as circles ($N=200$) and diamonds ($N=400$). Error bars are shown.

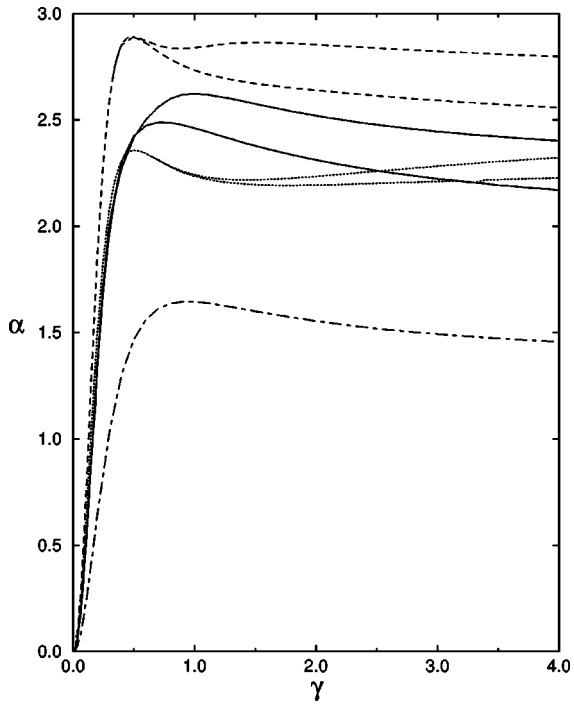

FIG. 4. Storage capacity α for the piecewise-linear input-output relation as a function of the gain parameter γ for an output tolerance $\epsilon=0.5$ and a total average output error $\mathcal{E}=0.05$ for the GD cost function (solid lines), the linear cost function (dashed curves), and the quadratic cost function (dotted curves). For each case the upper curve is the RS result, the lower one the RSB1 result. The dashed-dotted curve is the critical storage capacity.

The reason why the performance of the $s=2$ quadratic cost function is worse here is based on the fact that with a nonzero ϵ, the average output error decreases. The curves for the hyperbolic tangent input-output relation are all for $\mathcal{E} \leqslant 0.4$, while for the piecewise-linear input-output relation we have studied the case $\mathcal{E}=0.05$. In the latter, we are closer to the critical capacity. From these calculations one might conclude that the relative performance of the different cost functions depends also on the amount of errors. In other words, it matters how far one is beyond the critical capacity and the quadratic cost function performs better in the high-α regime.

In order to discuss in more detail the effects of RSB, we have studied the distribution of the local fields for the three cost functions. As an illustrative example, we present in Figs. 5(a)–5(c) the RS and RSB1 distributions for the specific parameters $\alpha=5$, $\gamma=1$, $\epsilon=0.5$, and $\zeta=0.6$. In general terms, the discussion above concerning the RS local field distribution for the hyperbolic tangent input-output relation remains valid. For the RSB1 distribution, the following has to be remarked. In the case of the GD and the linear cost function the coefficients of the δ-part in the RSB1 local field distributions become smaller. To give an idea about this change we mention that, e.g., for the GD cost function [recall Eqs. (A5) and (A12)], for the parameters mentioned above, at $h=0.1$ these coefficients are 0.415 for the RS solution versus 0.194 for the RSB solution. Similarly for the quadratic cost function the maximum in the distribution decreases. Furthermore for the three cost functions, the continuum part of the distribution is more populated for the RSB1 than for the RS solution, and the width of the gaps are smaller. Finally, RSB effects in the local field distribution are less pronounced for the quadratic cost function.

The distribution of the local fields resulting from simulations of networks with $N=200$ are also shown on these Figs. 5(b) and 5(c) for, respectively, the linear and quadratic cost

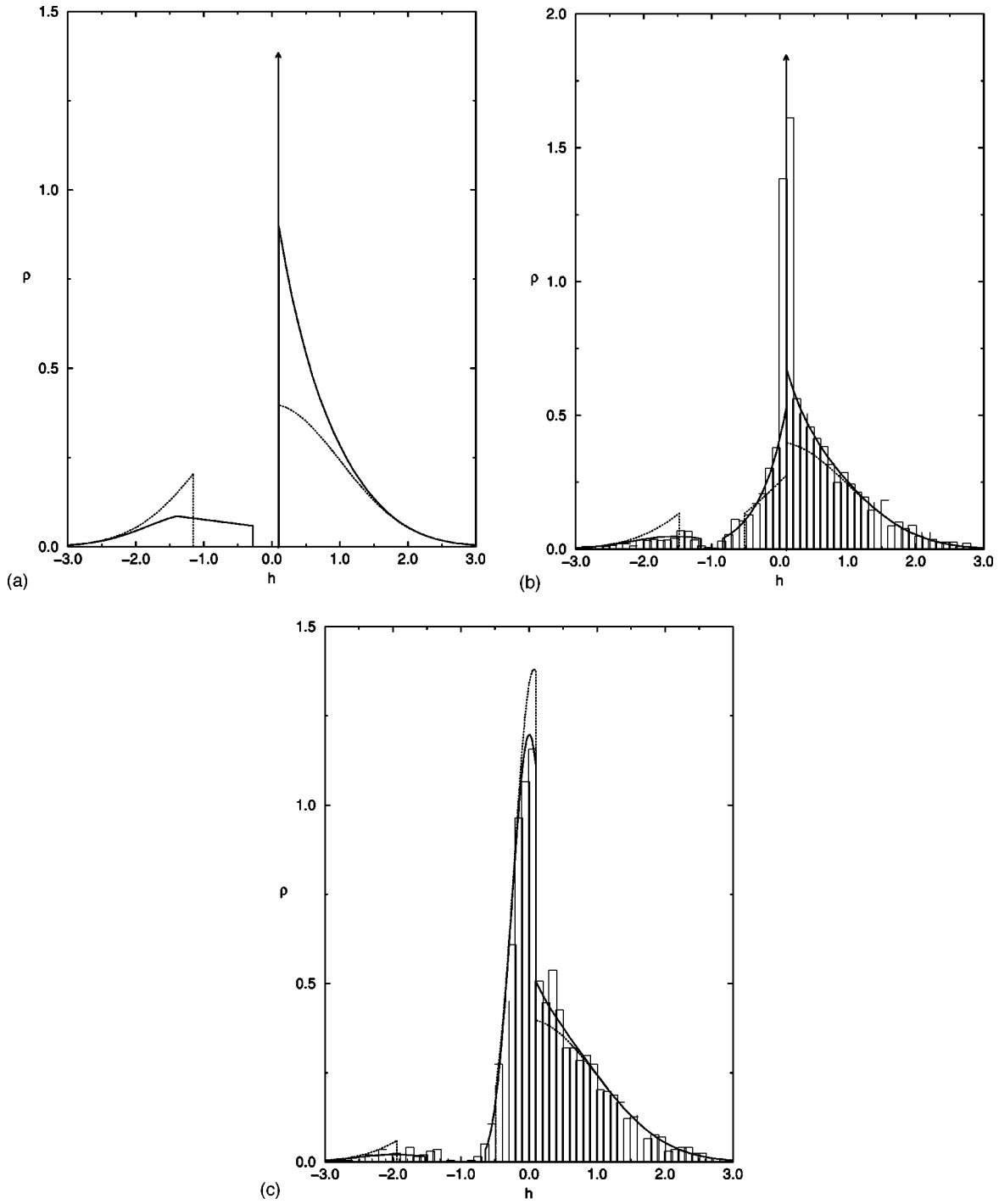

FIG. 5. Distribution of the local fields for the piecewise-linear input-output relation and $\alpha=5$, $\gamma=1$, $\epsilon=0.5$ and the correct output $\zeta=0.6$ for the GD cost function (a), the linear cost function (b), and the quadratic cost function (c). The dotted curves are the RS results, the solid lines the RSB1 results. In (b) and (c) the histograms are results from numerical simulations with ($N=200$).

functions. In both cases, a very good agreement with the RSB1 local field distributions obtained analytically is observed. This result extends the observation of [15] that an RSB1 treatment is adequate for the calculation of the storage error to the calculation of local field distributions.

V. CONCLUDING REMARKS

In this paper we have studied the canonical ensemble approach to the optimal capacity of graded-response perceptrons with a hyperbolic tangent and a piecewise-linear input-output relation for three different cost functions: the Gardner-Derrida cost function that simply counts the number of errors irrespective of their sizes, the linear cost function where the errors are weighted proportionally to their magnitudes and the quadratic cost function where the errors are weighted proportionally to the square of their magnitudes. Through the method of gradient descent the last two cost functions define, respectively, a perceptron and an adaline learning algorithm.

Results have been obtained for the storage capacity as a function of the gain parameter, for the distribution of the

local fields and for the average total output error above critical capacity in both RS and RSB1 approximations. Simulations for linear and quadratic cost functions have been performed. Here we summarize the main properties of the three cost functions.

In the whole range of parameters, the most efficient cost function is the linear one. This is not surprising, since it is most suited to minimizing linear deviations from the correct output. At small to intermediate loading α, the quadratic cost function is the least efficient. At high loading its efficiency increases in comparison with the GD cost function. The relative ordering between the linear and quadratic cost functions has been confirmed by numerical simulations.

In agreement with standard results (see, e.g., [17]) it is seen that whenever the distribution of the local fields displays a gap the RS saddle point is certainly unstable. A more refined discussion of this RS instability has been possible in the α-γ plane showing, e.g., that the instability stays in regions of the network parameters where no gap occurs and that RS stability can be restored for small γ.

First-step RSB results have been obtained and it is seen that already for a small, fixed average total output error (and an output tolerance of 0.5) the capacity is overestimated in RS by typically about 10%. Furthermore, also the local field distributions are changed considerably due to RSB1 effects, in agreement with earlier findings [7]. From Ref. [15] we know that although for the binary perceptron with the Gardner-Derrida cost function an exact solution for the storage error above critical capacity requires full RSB, one-step RSB may be considered sufficient. In a similar way we have shown here that also for the local field distributions the first-step RSB results are in remarkable agreement with the distributions obtained by numerical simulations. This is one of the important outcomes of this work which, we believe, may be extended to other systems.

ACKNOWLEDGMENTS

This work has been supported in part by the Research Fund of the KU Leuven (Grant No. OT/94/9). We are indebted to R. Kühn and J. van Mourik for stimulating discussions. D.B. thanks the Fund for Scientific Research Flanders (Belgium) for financial support. R.E. was supported in part by the CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), Brazil. He also acknowledges the kind hospitality of the Instituut voor Theoretische Fysica at the KU Leuven, where part of this work was done.

APPENDIX: THEORY FOR SPECIFIC COST FUNCTIONS

In this appendix we apply the general theory discussed in Sec. III to the specific cost functions (5) and (6). In particular we study the RS solutions for the hyperbolic tangent input–output relation and both the RS and RSB1 solutions for the piecewise-linear input-output relation defined in Eq. (27).

1. GD cost function

In the case of the GD cost function with output tolerance ϵ the results presented here are valid, of course, for both input-output relations considered, by taking in the end the relevant expression for $g^{-1}(\zeta-\epsilon)$. In the case of the RS treatment, the minimum in h of Eq. (13) is given by

$$
\begin{aligned}
&h_0=t, \quad F_{\mathrm{RS}}(h_0,\zeta,x,t)=1, \quad \text{for } -\infty<t<l-\sqrt{2x},\\
&h_0=l, \quad F_{\mathrm{RS}}(h_0,\zeta,x,t)=\frac{(l-t)^2}{2x}, \quad \text{for } l-\sqrt{2x}<t<l,\\
&h_0=t, \quad F_{\mathrm{RS}}(h_0,\zeta,x,t)=0, \quad \text{for } l<t<u,\\
&h_0=u, \quad F_{\mathrm{RS}}(h_0,\zeta,x,t)=\frac{(u-t)^2}{2x}, \quad \text{for } u<t<u+\sqrt{2x},\\
&h_0=t, \quad F_{\mathrm{RS}}(h_0,\zeta,x,t)=1, \quad \text{for } u+\sqrt{2x}<t<\infty,
\end{aligned}
\tag{A1}
$$

where

$$
l=\begin{cases}\dfrac{1}{\gamma}g^{-1}(\zeta-\epsilon), & \text{if } \zeta-\epsilon>-1,\\ -\infty, & \text{elsewhere}\end{cases}
\tag{A2}
$$

and

$$
u=\begin{cases}\dfrac{1}{\gamma}g^{-1}(\zeta+\epsilon), & \text{if } \zeta+\epsilon<1,\\ \infty, & \text{elsewhere.}\end{cases}
\tag{A3}
$$

From Eqs. (14) and (A1), we obtain the saddle-point equation

$$
\alpha_{\mathrm{RS}}^{-1}=\left\langle \int_{l-\sqrt{2x}}^{l} Dt(t-l)^2+\int_{u}^{u+\sqrt{2x}} Dt(t-u)^2 \right\rangle_{\zeta}.
\tag{A4}
$$

Combining Eqs. (19) and (A1), the distribution of local fields becomes

$$\rho(h,\zeta)=\frac{e^{-(h^2/2)}}{\sqrt{2\pi}}[\theta(l-\sqrt{2x}-h)+\theta(h-l)-\theta(h-u)+\theta(h-u-\sqrt{2x})]+\delta(h-l)\int_{l-\sqrt{2x}}^{l}Dt+\delta(h-u)\int_{u}^{u+\sqrt{2x}}Dt. \quad \text{(A5)}$$

The RS output error is obtained from Eqs. (21) and (A5):

$$\mathcal{E}(\zeta)=\gamma\left[\left(l+\frac{1}{\gamma}\right)\int_{-\infty}^{h_1}Dh+\int_{h_1}^{l-\sqrt{2x}}Dh(l-h)+\int_{u+\sqrt{2x}}^{h_2}Dh(h-u)+\left(\frac{1}{\gamma}-u\right)\int_{h_2}^{-\infty}Dh\right], \quad \text{(A6)}$$

where

$$h_1=\min\left(l-\sqrt{2x},-\frac{1}{\gamma}\right) \quad \text{(A7)}$$

and

$$h_2=\max\left(u+\sqrt{2x},\frac{1}{\gamma}\right). \quad \text{(A8)}$$

For the RSB1 solution we get $h_0(\zeta,x,q_0,t_0,t_1)$ and $F_{\mathrm{RSB1}}(h_0,\zeta,x,q_0,t_0,t_1)$ from $h_0(\zeta,x,t)$ and $F_{\mathrm{RS}}(h_0,\zeta,x,t)$, respectively, by substituting t by $t_0\sqrt{q_0}+t_1\sqrt{1-q_0}$ in Eq. (A1). The function $\Psi(\zeta,x,q_0,M,t_0)$ in Eq. (24) becomes

$$\Psi(\zeta,x,q_0,M,t_0)=e^{-Mx}\int_{-\infty}^{\Omega(l-\sqrt{2x},q_0,t_0)}Dt_1+\int_{\Omega(l-\sqrt{2x},q_0,t_0)}^{\Omega(l,q_0,t_0)}Dt_1\Phi(l,M,q_0,t_0,t_1)+\int_{\Omega(l,q_0,t_0)}^{\Omega(u,q_0,t_0)}Dt_1+\int_{\Omega(u,q_0,t_0)}^{\Omega(u+\sqrt{2x},q_0,t_0)}Dt_1\Phi(u,M,q_0,t_0,t_1)+e^{-Mx}\int_{\Omega(u+\sqrt{2x},q_0,t_0)}^{\infty}Dt_1, \quad \text{(A9)}$$

where

$$\Omega(\omega,q_0,t_0)=\frac{\omega-t_0\sqrt{q_0}}{\sqrt{1-q_0}} \quad \text{(A10)}$$

and

$$\Phi(\omega,M,q_0,t_0,t_1)=\exp\{-\tfrac{1}{2}M(1-q_0)[\Omega(\omega,q_0,t_0)-t_1]^2\}. \quad \text{(A11)}$$

The averaged free-energy is obtained by plugging this expression into Eq. (23). Expression (26) then leads to the ζ-dependent distribution of local fields,

$$\rho(h,\zeta)=\int\frac{Dt_0}{\Psi(\zeta,x,q_0,M,t_0)}\times\left\{\frac{\exp\left[-\frac{1}{2}\Omega^2(h,q_0,t_0)\right]}{\sqrt{2\pi(1-q_0)}}[e^{-Mx}\theta(l-\sqrt{2x}-h)+[\theta(h-l)-\theta(h-u)]+e^{-Mx}\theta(h-u-\sqrt{2x})]+\delta(h-l)\int_{\Omega(l-\sqrt{2x},q_0,t_0)}^{\Omega(l,q_0,t_0)}Dt_1\Phi(l,M,q_0,t_0,t_1)+\delta(h-u)\int_{\Omega(u,q_0,t_0)}^{\Omega(u+\sqrt{2x},q_0,t_0)}Dt_1\Phi(u,M,q_0,t_0,t_1)\right\}. \quad \text{(A12)}$$

Finally, the ζ-dependent RSB1 output error is given by combining Eqs. (26) and (A12):

$$\mathcal{E}(\zeta)=\int\frac{Dt_0}{\Psi(\zeta,x,q_0,M,t_0)}\frac{\gamma e^{-Mx}}{\sqrt{2\pi(1-q_0)}}\times\left\{\left(l+\frac{1}{\gamma}\right)\int_{-\infty}^{h_1}dh+\int_{h_1}^{l-\sqrt{2x}}dh(l-h)+\int_{u+\sqrt{2x}}^{h_2}dh(h-u)+\left(\frac{1}{\gamma}-u\right)\int_{h_2}^{\infty}dh\right\}\times\exp[-\tfrac{1}{2}\Omega^2(h,q_0,t_0)], \quad \text{(A13)}$$

with h_1 and h_2 defined in Eqs. (A7) and (A8), respectively.

2. Linear cost function

Let us start by considering the RS approximation first. For the piecewise-linear input-output relation the minimization in h of Eq. (13) can be done explicitly leading to the following result:

$$h_0=t,\quad F_{\rm RS}(h_0,\zeta,x,t)=\gamma\left(l+\frac{1}{\gamma}\right),\quad \text{for } -\infty<t<h_1,$$

$$h_0=\gamma x+t,\quad F_{\rm RS}(h_0,\zeta,x,t)=\gamma\left(l-\frac{\gamma x}{2}-t\right),\quad \text{for } h_1<t<h_2,$$

$$h_0=l,\quad F_{\rm RS}(h_0,\zeta,x,t)=\frac{(l-t)^2}{2x},\quad \text{for } h_2<t<l,$$

$$h_0=t,\quad F_{\rm RS}(h_0,\zeta,x,t)=0,\quad \text{for } l<t<u,$$

$$h_0=u,\quad F_{\rm RS}(h_0,\zeta,x,t)=\frac{(u-t)^2}{2x},\quad \text{for } u<t<h_3,$$

$$h_0=-\gamma x+t,\quad F_{\rm RS}(h_0,\zeta,x,t)=\gamma\left(-u-\frac{\gamma x}{2}+t\right),\quad \text{for } u<t<h_4,$$

$$h_0=t,\quad F_{\rm RS}(h_0,\zeta,x,t)=\gamma\left(\frac{1}{\gamma}-u\right),\quad \text{for } h_4<t<\infty, \tag{A14}$$

where l and u are again given by the formulas (A2) and (A3). The variables h_1, h_2, h_3, and h_4 are defined as follows:

$$h_1=\begin{cases} l-\sqrt{2\gamma x\left(l+\frac{1}{\gamma}\right)}, & \text{if } l<-\frac{1}{\gamma}+\frac{\gamma x}{2},\\ -\frac{1}{\gamma}-\frac{\gamma x}{2}, & \text{elsewhere,}\end{cases} \tag{A15}$$

$$h_2=\begin{cases} l-\sqrt{2\gamma x\left(l+\frac{1}{\gamma}\right)}, & \text{if } l<-\frac{1}{\gamma}+\frac{\gamma x}{2},\\ l-\gamma x, & \text{elsewhere,}\end{cases} \tag{A16}$$

$$h_3=\begin{cases} u+\sqrt{2\gamma x\left(\frac{1}{\gamma}-u\right)}, & \text{if } u>\frac{1}{\gamma}-\frac{\gamma x}{2},\\ u+\gamma x, & \text{elsewhere,}\end{cases} \tag{A17}$$

$$h_4=\begin{cases} u+\sqrt{2\gamma x\left(\frac{1}{\gamma}-u\right)}, & \text{if } u>\frac{1}{\gamma}-\frac{\gamma x}{2},\\ \frac{1}{\gamma}+\frac{\gamma x}{2}, & \text{elsewhere.}\end{cases} \tag{A18}$$

The RS saddle-point equation is obtained from Eqs. (14) and (A14):

$$\alpha_{\rm RS}^{-1}=\left\langle \gamma^2x^2\int_{h_1}^{h_2}Dt+\int_{h_2}^{l}Dt(t-l)^2+\int_{u}^{h_3}Dt(t-u)^2+\gamma^2x^2\int_{h_3}^{h_4}Dt\right\rangle_\zeta . \tag{A19}$$

Using Eqs. (19) and (A14), the RS ζ-dependent distribution of local fields becomes

$$\begin{aligned}\rho(h,\zeta)=&\frac{\exp\left[-\frac{h^2}{2}\right]}{\sqrt{2\pi}}\\ &\times[\theta(h_1-h)+\theta(h-l)-\theta(h-u)+\theta(h-h_4)]\\ &+\frac{\exp\left[-\frac{(h-\gamma x)^2}{2}\right]}{\sqrt{2\pi}}[\theta(h-h_2')-\theta(h-l)]\\ &+\delta(h-l)\int_{h_2}^{l}Dt+\delta(h-u)\int_{u}^{h_3}Dt\\ &+\frac{\exp\left[-\frac{(h+\gamma x)^2}{2}\right]}{\sqrt{2\pi}}[\theta(h-u)-\theta(h-h_3')],\end{aligned} \tag{A20}$$

where

$$h_2' = \begin{cases} l, & \text{if } l < -\dfrac{1}{\gamma} + \dfrac{\gamma x}{2}, \\ -\dfrac{1}{\gamma} + \dfrac{\gamma x}{2}, & \text{elsewhere} \end{cases} \tag{A21}$$

and

$$h_3' = \begin{cases} u, & \text{if } u > \dfrac{1}{\gamma} - \dfrac{\gamma x}{2}, \\ \dfrac{1}{\gamma} - \dfrac{\gamma x}{2}, & \text{elsewhere.} \end{cases} \tag{A22}$$

The RS average output error is obtained from Eqs. (21) and (A20):

$$\begin{aligned} \mathcal{E}(\zeta) = \gamma \Bigg[& \left(l + \frac{1}{\gamma} \right) \int_{-\infty}^{h_1} \frac{dh}{\sqrt{2\pi}} e^{-(h^2/2)} + \int_{h_2'}^{l} dh \sqrt{2\pi} \\ & \times \exp\left[-\frac{(h-\gamma x)^2}{2} \right] (l-h) \\ & + \int_{u}^{h_3'} \frac{dh}{\sqrt{2\pi}} \exp\left[-\frac{(h+\gamma x)^2}{2} \right] (h-u) \\ & + \left(\frac{1}{\gamma} - u \right) \int_{h_4}^{-\infty} \frac{dh}{\sqrt{2\pi}} e^{-(h^2/2)} \Bigg]. \end{aligned} \tag{A23}$$

For the hyperbolic tangent input-output relation, h_ζ given by Eq. (17) is always a local minimum of Eq. (13). Other local minima of Eq. (13) are defined as solutions of

$$\left(\frac{\partial F(h,\zeta,x,t)}{\partial h} \right)_{h=h_0} = 0 \tag{A24}$$

and they can no longer be determined analytically. The equation

$$t(h_0) = h_0 + \gamma x g'(\gamma h_0) \operatorname{sgn}[g(\gamma h_0) - \zeta] \tag{A25}$$

needs to be inverted in order to find $h_0 = h_0(t)$ (the prime denotes the derivative with respect to h). Depending on the value of $\gamma^2 x$, $t(h_0)$ is a monotonic function or not, and consequently invertible or not. The onset of nonmonotonicity is given by the system of equations

$$\frac{dt}{dh_0} = 0,$$

$$\frac{d^2 t}{dh_0^2} = 0. \tag{A26}$$

If monotonicity holds, h_0 is a solution of Eq. (A25) for $t < t_\zeta^-$ or $t > t_\zeta^+$, where $t_\zeta^\pm = h_\zeta \pm \gamma x g'(\gamma h_\zeta)$. If nonmonotonicity holds, $h_0(t)$ has one or two jumps at $t = t_1$ and/or $t = t_2$, whereby we assume that $t_1 < t_2$. The values of t_1 and t_2 are then determined using a Maxwell construction in the function $t(h_0)$. The number of jumps depends on the value of $\gamma^2 x$ and ζ.

From Eq. (19) and from the inversion of Eq. (A25), we obtain the following expression for the distribution of the local fields:

$$\begin{aligned} \rho(h|\zeta) = & \frac{dt}{dh} \frac{\exp\left[-\dfrac{t^2(h)}{2} \right]}{\sqrt{2\pi}} + \frac{1}{2} \left[\operatorname{erf}\left(\frac{t_2}{\sqrt{2}} \right) - \operatorname{erf}\left(\frac{t_1}{\sqrt{2}} \right) \right] \\ & \times \delta(h - h_\zeta). \end{aligned} \tag{A27}$$

If monotonicity holds, then $t_1 = t_\zeta^-$ and $t_2 = t_\zeta^+$.

Due to the fact that h_ζ is a global minimum of Eq. (13) in the interval $t_1 < t < t_2$, $t(h_0)$ always displays a jump in $h = h_\zeta$. This jump gives rise to the second term in the right-hand side (r.h.s.) of Eq. (A27). If nonmonotonicity holds, $t(h_0)$ shows plateaus at t_1 and t_2, leading to a gap structure in the distribution of the local fields. The resulting discontinuity in dh_0/dt causes a divergence of the l.h.s. of Eq. (15). This means that when nonmonotonicity holds, the RS solution is always unstable. In the case of monotonicity, the stability condition reads

$$\begin{aligned} \alpha_{\mathrm{RS}} \Bigg\langle & \int_{-\infty}^{t_\zeta^-} Dt \left(\frac{1}{\gamma^2 x g''(\gamma h_0)} - 1 \right)^{-2} \\ & + \int_{t_\zeta^+}^{\infty} Dt \left(\frac{1}{\gamma^2 x g''(\gamma h_0)} + 1 \right)^{-2} \Bigg\rangle_{\{\zeta\}} < 1. \end{aligned} \tag{A28}$$

Next we consider first step RSB for the piecewise-linear input-output relation. Similarly to the GD cost function we substitute t by $t_0 \sqrt{q_0} + t_1 \sqrt{1 - q_0}$ in Eq. (A14) in order to obtain $h_0(\zeta, x, q_0, t_0, t_1)$ and $F_{\mathrm{RSB1}}(h_0, \zeta, x, q_0, t_0, t_1)$. The free energy is obtained from Eq. (23) with the function $\Psi(\zeta, x, q_0, M, t_0)$ [recall Eq. (24)] given by

$$\Psi(\zeta,x,q_0,M,t_0)=\exp\left[-M\gamma x\left(l+\frac{1}{\gamma}\right)\right]\int_{-\infty}^{\Omega(h_1,q_0,t_0)}Dt_1+\exp\left[-M\gamma x\left(l-\frac{\gamma x}{2}-t_0\sqrt{q_0}\right)\right]$$
$$\times\int_{\Omega(h_1,q_0,t_0)}^{\Omega(h_2,q_0,t_0)}Dt_1\exp[M\gamma x t_1\sqrt{(1-q_0)}]+\int_{\Omega(h_2,q_0,t_0)}^{\Omega(l,q_0,t_0)}Dt_1\Phi(l,M,q_0,t_0,t_1)+\int_{\Omega(l,q_0,t_0)}^{\Omega(u,q_0,t_0)}Dt_1$$
$$+\int_{\Omega(u,q_0,t_0)}^{\Omega(h_3,q_0,t_0)}Dt_1\Phi(u,M,q_0,t_0,t_1)+\exp\left[-M\gamma x\left(-u-\frac{\gamma x}{2}+t_0\sqrt{q_0}\right)\right]\int_{\Omega(h_3,q_0,t_0)}^{\Omega(h_4,q_0,t_0)}Dt_1$$
$$\times\exp[-M\gamma x t_1\sqrt{(1-q_0)}]+\exp\left[-M\gamma x\left(\frac{1}{\gamma}-u\right)\right]\int_{\Omega(h_4,q_0,t_0)}^{\infty}Dt_1. \tag{A29}$$

The RSB1 ζ-dependent distribution of the local fields (26) becomes

$$\rho(h,\zeta)=\int\frac{Dt_0}{\Psi(\zeta,x,q_0,M,t_0)}\left\{\frac{\exp\left[-M\gamma x\left(l+\frac{1}{\gamma}\right)-\frac{1}{2}\Omega^2(h,q_0,t_0)\right]}{\sqrt{2\pi(1-q_0)}}\theta(h_1-h)\right.$$
$$+\frac{\exp\left[-M\gamma x\left(l+\frac{\gamma x}{2}-h\right)-\frac{1}{2}\Omega^2(h-\gamma x,q_0,t_0)\right]}{\sqrt{2\pi(1-q_0)}}[\theta(h-h_2')-\theta(h-l)]\theta\left(l+\frac{1}{\gamma}-\frac{\gamma x}{2}\right)+\delta(h-l)\int_{\Omega(h_2,q_0,t_0)}^{\Omega(l,q_0,t_0)}$$
$$\times Dt_1\Phi(l,M,q_0,t_0,t_1)+\frac{\exp\left[-\frac{1}{2}\Omega^2(h,q_0,t_0)\right]}{\sqrt{2\pi(1-q_0)}}[\theta(h-l)-\theta(h-u)]+\delta(h-u)\int_{\Omega(u,q_0,t_0)}^{\Omega((h_3,q_0,t_0)}Dt_1\Phi(u,M,q_0,t_0,t_1)$$
$$+\frac{\exp\left[-M\gamma x\left(h-u+\frac{\gamma x}{2}\right)-\frac{1}{2}\Omega^2(h+\gamma x,q_0,t_0)\right]}{\sqrt{2\pi(1-q_0)}}[\theta(h-u)-\theta(h-h_3')]\theta\left(\frac{1}{\gamma}-\frac{\gamma x}{2}-u\right)$$
$$\left.+\frac{\exp\left[-M\gamma x\left(\frac{1}{\gamma}-u\right)-\frac{1}{2}\Omega^2(h,q_0,t_0)\right]}{\sqrt{2\pi(1-q_0)}}\theta(h-h_4)\right\}. \tag{A30}$$

Finally, the ζ-dependent RSB1 average output error reads

$$\mathcal{E}(\zeta)=\int\frac{Dt_0}{\Psi(\zeta,x,q_0,M,t_0)}\frac{\gamma}{\sqrt{2\pi(1-q_0)}}\left\{\left(l+\frac{1}{\gamma}\right)\exp\left[-M\gamma x\left(l+\frac{1}{\gamma}\right)\right]\int_{-\infty}^{h_1}dh\exp\left[-\frac{1}{2}\Omega^2(h,q_0,t_0)\right]\right.$$
$$+\int_{h_2'}^{l}dh(l-h)\exp\left[-\frac{1}{2}\Omega^2(h-\gamma x,q_0,t_0)-M\gamma x\left(l+\frac{\gamma x}{2}-h\right)\right]\theta\left(l+\frac{1}{\gamma}-\frac{\gamma x}{2}\right)$$
$$+\int_{u}^{h_3'}dh(h-u)\exp\left[-\frac{1}{2}\Omega^2(h+\gamma x,q_0,t_0)-M\gamma x\left(h-u+\frac{\gamma x}{2}\right)\right]\theta\left(\frac{1}{\gamma}-\frac{\gamma x}{2}-u\right)$$
$$\left.+\left(\frac{1}{\gamma}-u\right)\exp\left[-M\gamma x\left(\frac{1}{\gamma}-u\right)\right]\int_{h_4}^{\infty}dh\exp\left[-\frac{1}{2}\Omega^2(h,q_0,t_0)\right]\right\}. \tag{A31}$$

3. Quadratic cost function

Again we look at the RS treatment first. We start by defining

$$h_1=l-\sqrt{1+2\gamma^2 x}\left(l+\frac{1}{\gamma}\right), \tag{A32}$$

$$h_2 = l - \frac{l + \frac{1}{\gamma}}{\sqrt{1+2\gamma^2 x}}, \tag{A33}$$

$$h_3 = u + \frac{\frac{1}{\gamma} - u}{\sqrt{1+2\gamma^2 x}}, \tag{A34}$$

$$h_4 = u + \sqrt{1+2\gamma^2 x}\left(\frac{1}{\gamma} - u\right). \tag{A35}$$

Using these definitions, the result of the minimization in h of F_{RS} [Eq. (13)] becomes

$$h_0 = t, \quad F_{\mathrm{RS}}(h_0,\zeta,x,t) = \gamma^2\left(l + \frac{1}{\gamma}\right)^2, \quad \text{for } -\infty < t < h_1,$$

$$h_0 = \frac{2\gamma^2 x l + t}{1+2\gamma^2 x}, \quad F_{\mathrm{RS}}(h_0,\zeta,x,t) = \gamma^2 \frac{(l-t)^2}{1+2\gamma^2 x}, \quad \text{for } h_1 < t < l,$$

$$h_0 = t, \quad F_{\mathrm{RS}}(h_0,\zeta,x,t) = 0, \quad \text{for } l < t < u,$$

$$h_0 = \frac{2\gamma^2 x u + t}{1+2\gamma^2 x}, \quad F_{\mathrm{RS}}(h_0,\zeta,x,t) = \gamma^2 \frac{(u-t)^2}{1+2\gamma^2 x}, \quad \text{for } u < t < h_4,$$

$$h_0 = t, \quad F_{\mathrm{RS}}(h_0,\zeta,x,t) = \gamma^2\left(\frac{1}{\gamma} - u\right)^2, \quad \text{for } h_4 < t < \infty. \tag{A36}$$

The RS saddle-point equation is obtained from Eqs. (14) and (A36):

$$\alpha_{\mathrm{RS}}^{-1} = \left(\frac{2\gamma^2 x}{1+2\gamma^2 x}\right)\left\langle \int_{h_1}^{l} Dt(t-l)^2 + \int_{u}^{h^4} Dt(t-u)^2 \right\rangle_{\zeta}. \tag{A37}$$

From Eqs. (19) and (A36), the RS ζ-dependent distribution of the local fields becomes

$$\begin{aligned}\rho(h,\zeta) = {} & \frac{\exp\left[-\frac{h^2}{2}\right]}{\sqrt{2\pi}}[\theta(h_1-h) + \theta(h-l) - \theta(h-u) \\ & + \theta(h-h4)] + (1+2\gamma^2 x) \\ & \times \left\{ \frac{\exp\left\{-\frac{1}{2}[(1+2\gamma^2 x)h - 2\gamma^2 x l]^2\right\}}{\sqrt{2\pi}} \right. \\ & \times [\theta(h-h_2) - \theta(h-l)] \\ & + \frac{\exp\left[-\frac{1}{2}[(1+2\gamma^2 x)h - 2\gamma^2 x u]^2\right]}{\sqrt{2\pi}} \\ & \left. \times [\theta(h-hu) - \theta(h-h_3)] \right\}. \end{aligned} \tag{A38}$$

Consequently, the RS ζ-dependent output error becomes

$$\begin{aligned}\mathcal{E}(\zeta) = \gamma \Bigg\{ & \left(l + \frac{1}{\gamma}\right) \int_{-\infty}^{h_1} \frac{dh}{\sqrt{2\pi}} \exp\left[-\frac{h^2}{2}\right] + (1+2\gamma^2 x) \\ & \times \int_{h_2}^{l} dh \sqrt{2\pi}\, \exp\left[\frac{[(1+2\gamma^2 x)h - 2\gamma^2 x l]^2}{2}\right](l-h) \\ & + (1+2\gamma^2 x) \int_{u}^{h_3} dh \sqrt{2\pi} \\ & \times \exp\left[-\frac{[(1+2\gamma^2 x)h - 2\gamma^2 x u]^2}{2}\right](h-u) \\ & + \left(\frac{1}{\gamma} - u\right) \int_{h_4}^{-\infty} \frac{dh}{\sqrt{2\pi}} \exp\left[-\frac{h^2}{2}\right] \Bigg\}. \end{aligned} \tag{A39}$$

For the hyperbolic tangent input-output relation, h_ζ (recall Eq. (17)) is no longer a local minimum of Eq. (13). The minima are always given by the solutions of Eq. (A24), which define $t(h_0)$ as

$$t(h_0) = h_0 + 2\gamma x g'(\gamma h_0)[g(\gamma h_0) - \zeta]. \tag{A40}$$

Depending on the values of $\gamma^2 x$ *and* ζ, $t(h_0)$ is a monotonic function or not. The onset to nonmonotonicity is given by the system of Eqs. (A26). If monotonicity holds, $h_0(t)$ is continuous, and is obtained by inverting Eq. (A40). Otherwise, $h_0(t)$ displays one or two jumps at $t = t_1$ and/or $t = t_2$, whose values are obtained by a Maxwell construction.

The distribution of local fields is obtained directly from Eq. (19),

$$\rho(h|\zeta)=\frac{dt}{dh}\,\frac{\exp\left[-\frac{t^2(h)}{2}\right]}{\sqrt{2\pi}}. \tag{A41}$$

When nonmonotonicity holds, the jumps in $h_0(t)$ give rise to a gap structure in the distribution of the local fields and the RS solution becomes unstable. In the monotonic case the stability condition (15) reads

$$\alpha_{\rm RS}\left\langle \int_{-\infty}^{+\infty} Dt\left(\frac{1}{2\gamma^2 x[(g'(\gamma h_0))^2+(g(\gamma h_0)-\zeta)g''(\gamma h_0)]}+1\right)^{-2}\right\rangle_{\{\zeta\}}<1. \tag{A42}$$

Let us finally turn to the RSB1 treatment. Again, in order to obtain $h_0(\zeta,x,q_0,t_0,t_1)$ and $F_{\rm RSB1}(h_0,\zeta,x,q_0,t_0,t_1)$ we substitute t by $t_0\sqrt{q_0}+t_1\sqrt{1-q_0}$ in Eq. (A14). The free energy is obtained from Eq. (23), whereby the function $\Psi(\zeta,x,q_0,M,t_0)$, given by Eq. (24), becomes

$$\begin{aligned}\Psi(\zeta,x,q_0,M,t_0)=&\exp\left[-M\gamma^2x\left(l+\frac{1}{\gamma}\right)^2\right]\int_{-\infty}^{\Omega(h_1,q_0,t_0)}Dt_1+\int_{\Omega(h_1,q_0,t_0)}^{h_2,q_0,t_0)}Dt_1\Phi\left(l,\frac{2M\gamma^2x}{1+2\gamma^2x},q_0,t_0\right)+\int_{\Omega(l,q_0,t_0)}^{\Omega(u,q_0,t_0)}Dt_1\\&+\int_{\Omega(u,q_0,t_0)}^{h_4,q_0,t_0)}Dt_1\Phi\left(u,\frac{2M\gamma^2x}{1+2\gamma^2x},q_0,t_0\right)+\exp\left[-M\gamma^2x\left(\frac{1}{\gamma}-u\right)^2\right]\int_{\Omega(u,q_0,t_0)}^{\infty}Dt_1.\end{aligned} \tag{A43}$$

For the RSB1 ζ-dependent distribution of the local fields we obtain

$$\begin{aligned}\rho(h,\zeta)=\int\frac{Dt_0}{\Psi(\zeta,x,q_0,M,t_0)}\Bigg\{&\frac{\exp\left[-M\gamma x\left(l+\frac{1}{\gamma}\right)-\frac{1}{2}\Omega^2(h,q_0,t_0)\right]}{\sqrt{2\pi(1-q_0)}}\theta(h_1-h)\\&+\frac{\exp\left[-M\gamma x\left(l+\frac{\gamma x}{2}-h\right)-\frac{1}{2}\Omega^2(h-\gamma x,q_0,t_0)\right]}{\sqrt{2\pi(1-q_0)}}[\theta(h-h_2')-\theta(h-l)]+\delta(h-l)\int_{\Omega(h_2,q_0,t_0)}^{\Omega(l,q_0,t_0)}Dt_1\Phi(M,l,q_0,t_0,t_1)\\&+\frac{\exp[-\frac{1}{2}\Omega(h,q_0,t_0)]}{\sqrt{2\pi(1-q_0)}}[\theta(h-l)-\theta(h-u)]+\delta(h-u)\int_{\Omega(u,q_0,t_0)}^{\Omega(h_3,q_0,t_0)}Dt_1\Phi(M,u,q_0,t_0,t_1)\\&+\frac{\exp\left[-M\gamma x\left(h-u+\frac{\gamma x}{2}\right)-\frac{1}{2}\Omega(h+\gamma x,q_0,t_0)\right]}{\sqrt{2\pi(1-q_0)}}[\theta(h-u)-\theta(h-h_3')]\\&+\frac{\exp\left[-M\gamma x\left(\frac{1}{\gamma}-u\right)-\frac{1}{2}\Omega(h,q_0,t_0)\right]}{\sqrt{2\pi(1-q_0)}}\theta(h-h_4)\Bigg\}.\end{aligned} \tag{A44}$$

The ζ-dependent RSB1 average output error becomes

$$\begin{aligned}\mathcal{E}(\zeta)=\int\frac{Dt_0}{\Psi(\zeta,x,q_0,M,t_0)}\frac{\gamma}{\sqrt{2\pi(1-q_0)}}\Bigg\{&\left(l+\frac{1}{\gamma}\right)\exp\left[-M\gamma^2x\left(1+\frac{1}{\gamma}\right)^2\right]\int_{-\infty}^{h_1}dh\,\exp[-\tfrac{1}{2}\Omega^2(h,q_0,t_0)]+(1+2\gamma^2x)\\&\times\int_{h_2}^{1}dh(1-h)\exp\{-\tfrac{1}{2}\Omega^2[(1+2\gamma^2x)(h-l)+l,q_0,t_0]-M\gamma^2x(1+2\gamma^2x)(h-l)^2\}+(1+2\gamma^2x)\int_{u}^{h_3}dh(h-u)\\&\times\exp\{-\tfrac{1}{2}\Omega^2[(1+2\gamma^2x)(h-u)+u,q_0,t_0]-M\gamma^2x(1+2\gamma^2x)(h-u)^2\}+\left(\frac{1}{\gamma}-u\right)\exp\left[-M\gamma^2x\left(\frac{1}{\gamma}-u\right)^2\right]\\&\times\int_{h_4}^{\infty}dh\exp[-\tfrac{1}{2}\Omega^2(h,q_0,t_0)]\Bigg\}.\end{aligned} \tag{A45}$$

[1] A. Treves, Phys. Rev. A **42**, 2418 (1990); J. Phys. A **23**, 2631 (1990).
[2] C. M. Marcus and R. M. Westervelt, Phys. Rev. A **40**, 501 (1989); C. M. Marcus, F. M. Waugh, and R. M. Westervelt, *ibid.* **41**, 3355 (1990).
[3] R. Kühn, in *Statistical Mechanics of Neural Networks*, Proceedings of the XIth Sitges Conference, edited by L. Garrido, Springer Lectures Notes in Physics Vol. 368 (Springer, Berlin, 1990), p. 19; R. Kühn, S. Bös, and J. L. van Hemmen, Phys. Rev. A **43**, 2084 (1991).
[4] M. Shiino and T. Fukai, J. Phys. A **23**, L1009 (1990).
[5] R. Kühn and S. Bös, J. Phys. A **26**, 831 (1993).
[6] D. Bollé, R. Kühn, and J. van Mourik, J. Phys. A **26**, 3149 (1993).
[7] D. Bollé and R. Erichsen, Jr., J. Phys. A **29**, 2299 (1996).
[8] S. Bös, W. Kinzel, and M. Opper, Phys. Rev. E **47**, 1384 (1993).
[9] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).
[10] E. Gardner, Europhys. Lett. **4**, 481 (1987); J. Phys. A **21**, 257 (1988).
[11] M. Griniasty and H. Gutfreund, J. Phys. A **24**, 715 (1991).
[12] R. Erichsen, Jr. and W. K. Theumann, J. Phys. G **26**, L61 (1993).
[13] P. Majer, A. Engel, and A. Zippelius, J. Phys. A **26**, 7405 (1993).
[14] W. Whyte, D. Sherrington, and K. Y. M. Wong, J. Phys. A **28**, 7105 (1995).
[15] W. Whyte and D. Sherrington, J. Phys. A **29**, 3063 (1996).
[16] J. R. de Almeida and D. Thouless, J. Phys. A **11**, 983 (1978).
[17] M. Bouten, J. Phys. A **27**, 6021 (1994).
[18] T. B. Kepler and L. F. Abbott, J. Phys. (France) **49**, 1657 (1988).
[19] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[20] K. Y. M. Wong and D. Sherrington, Phys. Rev. E **47**, 4465 (1993).
[21] K. Y. M. Wong and D. Sherrington, J. Phys. A **23**, 4659 (1990).