Optimal capacity of graded-response perceptrons

1993 J. Phys. A: Math. Gen. 26 3149

(http://iopscience.iop.org/0305-4470/26/13/019)

J. Phys. A: Math. Gen. 26 (1993) 3149–3158. Printed in the UK

Optimal capacity of graded-response perceptrons*

D Bollé†, R Kühn‡ and J van Mourik†

† Instituut voor Theoretische Fysica, K U Leuven, B-3001 Leuven, Belgium

‡ Institut für Theoretische Physik, Universität Heidelberg, W-6900 Heidelberg, Federal Republic of Germany

Received 18 September 1992

Abstract. Optimal capacities of perceptrons with graded input–output relations are computed within the Gardner approach. The influence of desired output precision, stability with respect to input errors, and output-pattern statistics are analysed and discussed.

1. Introduction

The collective properties of networks of formal two-state neuron-like elements are by now well understood [1–4]. Both retrieval properties of recurrent networks with specific prescriptions for the synaptic efficacies [5] and optimal capacities of perceptrons or recurrent networks [6, 7] have been analysed successfully by statistical mechanical approaches. More recently, questions regarding the behaviour of networks of multi-state [8–10] or graded-reponse [11–15] neurons have come to the fore—questions pertaining again to both retrieval properties of specific architectures [8, 11–15] and to optimal capacities of ensembles of networks designed to perform a given storage task [9, 10].

The purpose of the present paper is to extend previous analyses of optimal capacities of multi-state networks [9, 10] to the case of neurons with *graded* (continuous) input–output relations. This investigation is motivated by the fact that graded-response perceptrons constitute the basic building blocks of layered architectures trained by the backpropagation algorithm [16]. Such systems are to date the workhorses in practical applications of neural networks, and progress in the theoretical understanding of their capabilities and limitations should thus be welcome.

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

$$\zeta^\mu = g\left(\gamma \frac{1}{\sqrt{N}} \sum_j J_j \xi_j^\mu\right). \tag{1}$$

Here g is the input–output relation of the perceptron, which may be largely *arbitrary*. In particular, g need not be monotonic, non-decreasing or invertible for our general line of reasoning to be applicable. When studying specific examples, however, we shall specialize to monotonic, non-decreasing input–output relations. In (1), γ denotes a gain parameter, and the J_j are couplings of an architecture of perceptron type.

* Dedicated to Professor F Cerulus on the occasion of his 65th birthday.

0305-4470/93/133149+10$07.50

For the exact storage task (1), an explicit solution is known whenever $\alpha = p/N < 1$ [4], namely the pseudo-inverse solution

$$J_j^I = \frac{1}{\sqrt{N}} \sum_{\mu,\nu} h^\mu (C^{-1})_{\mu,\nu} \xi_j^\nu \tag{2}$$

where h^μ is chosen such that $\zeta^\mu = g(\gamma h^\mu)$, and where C is the correlation matrix of the input pattern set $\{\xi_i^\mu\}$, with elements $C_{\mu,\nu} = (1/N)\sum_i \xi_i^\mu \xi_i^\nu$. The solution (2) can be constructed independently of the pattern statistics as long as C^{-1} exists, which imposes an upper bound $\alpha_c = 1$ on the loading capacity α. *Within* the space spanned by the patterns, this solution is *unique* if g is *invertible*, since $h^\mu = (1/\gamma)g^{-1}(\zeta^\mu)$ is uniquely defined. If $\alpha < \alpha_c = 1$, equation (2) is *never* the only solution, though. One can add any $J^\perp$ orthogonal to the space spanned by the input patterns $\{\xi_i^\mu\}$ to $J^I = (J_j^I)$, and still satisfy (1). On the other hand, for non-invertible input–output relations there may be a *multiplicity* of choices for h^μ satisfying $\zeta^\mu = g(\gamma h^\mu)$, implying that even within the space spanned by the patterns the solution based on the correlation matrix is non-unique. This non-uniqueness raises the possibility of constructing solutions which are not based on the correlation matrix of the full input pattern set thereby achieving storage capacities larger than 1. For example, for $g(x) = \mathrm{sgn}(x)$ one has $\alpha_c = 2$ [6], *provided* the desired outputs ζ^μ take the values ± 1, whereas $\alpha_c = 0$, if $\zeta^\mu \notin \{\pm 1\}$. Thus, there is clearly an intimate relation between storage capacity, input–output relation, and output-pattern statistics. The aim of the present investigation is to elucidate this relation.

The paper is organized as follows. In section 2, we formulate the task set to graded-response perceptrons in a manner that allows us to use a Gardner-type analysis [6] in a relatively straightforward manner. In section 3, we present and discuss our main results for a variety of examples. Section 4, finally, contains a brief summary and an outlook on open questions.

2. Formal analysis of the problem

In order to set up the formal framework of our investigation, our strategy is to require stability with respect to input-data errors *and* to allow a limited output precision in the mapping (1). More specifically, we require that

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

where h^μ is the local field generated by the input ξ^μ, i.e. $h^\mu = (1/\sqrt{N})\sum_j J_j \xi_j^\mu$, and where κ and ϵ denote the required input stability and the allowed output-error tolerance, respectively. The method we use is to compute the volume in the space of couplings [6] satisfying the modified version (3) of the mapping (1).

Note that for our problem, the concept of volume in J-space is well defined *without* additional scale constraints like the mean spherical constraint usually adopted in such investigations. The reason is that a global rescaling of the J_j will generally lead to violations of either lower or upper bounds set on the outputs for each of the patterns. In what follows, we shall nevertheless adopt such a constraint for two reasons. The first is to see how (3) compares—for special cases—with what is known about the standard perceptron with the same constraint [6,7]. The second more specific reason is to fix a scale for the gain parameters γ of the input–output relations we consider.

In the present paper, we restrict our attention to unbiased input patterns with $\langle \xi_i^\mu \rangle = 0$ and $\langle \xi_i^\mu \xi_j^\nu \rangle = d_{\mu,\nu} d_{i,j} \Delta_0$. Since the effect of Δ_0 in (1) can be absorbed in the gain parameter γ, we henceforth take $\Delta_0 = 1$. Neither the input–output relation g nor the statistics of the outputs ζ^μ need to be specified at this point, however. Of course, for each ζ^μ, $I_{\text{out}}(\zeta^\mu, \epsilon)$ should have a non-empty intersection with the *range* of g in order to have $\alpha_c > 0$.

To compute the available volume in J-space satisfying (3), we note that (3) can be rewritten

$$h^\mu \in I_\mu \equiv \{x;\ g(\gamma(x \pm \kappa)) \in I_{\text{out}}(\zeta^\mu, \epsilon)\} \qquad \mu = 1, \ldots, p. \tag{4}$$

This formulation indicates the most general input–output relation that can be handled by our approach. The transfer function g should be such that the sets I_μ defined in (4) are *measurable*. In general, the sets I_μ form a collection of intervals in $\mathbb{R}$, that is $I_\mu = \cup_\sigma I_\mu^\sigma = \cup_\sigma [l_\mu^\sigma, u_\mu^\sigma]$ where l_μ^σ and u_μ^σ denote lower and upper bounds of the subinterval I_μ^σ, respectively. In the case of monotonic, non-decreasing input–output relations, the sets I_μ are simply connected intervals, $I_\mu = [l_\mu, u_\mu]$ with lower and upper bounds l_μ and u_μ defined by

$$l_\mu = \inf_x \{x;\ g(\gamma(x-\kappa)) \geqslant \zeta^\mu - \epsilon\} \qquad \text{and} \qquad u_\mu = \sup_x \{x;\ g(\gamma(x+\kappa)) \leqslant \zeta^\mu + \epsilon\}. \tag{5}$$

For *invertible* g's, one has $l_\mu = (1/\gamma) g^{-1}(\zeta^\mu - \epsilon_-) + \kappa$ and $u_\mu = (1/\gamma) g^{-1}(\zeta^\mu + \epsilon_+) - \kappa$, where $\epsilon_- = \min\{\epsilon, \zeta^\mu - \inf_x g(x)\}$ and $\epsilon_+ = \min\{\epsilon, \sup_x g(x) - \zeta^\mu\}$ are chosen such that $\zeta^\mu \mp \epsilon_\pm$ are in the range of g^{-1}.

With these definitions, the fractional volume of J's satisfying (4) reads

$$V = \frac{\int \prod_j \mathrm{d}J_j \prod_\mu \chi^{I_\mu}(h^\mu) \delta(N - \sum_j J_j^2)}{\int \prod_j \mathrm{d}J_j\, \delta(N - \sum_j J_j^2)}. \tag{6}$$

Here the χ^{I_μ} are characteristic functions of the sets I_μ defined in (4), with integral representation $\chi^{I_\mu}(x) = \int_{I_\mu} \mathrm{d}y \int (\mathrm{d}\hat{y}/2\pi) \exp\{\mathrm{i}\hat{y}(y-x)\}$. Following Gardner [6], we use the replica technique to evaluate $v = \lim_{N\to\infty} N^{-1} \langle\langle \ln V \rangle\rangle$, where $\langle\langle \cdots \rangle\rangle$ denotes an average over the statistics of inputs $\{\xi_i^\mu\}$ and outputs $\{\zeta^\mu\}$. Assuming that replica symmetry is unbroken, we get the following result,

$$v = \alpha \left\langle \int \mathrm{D}z \ln \left[\sum_\sigma (H(L_\mu^\sigma) - H(U_\mu^\sigma)) \right] \right\rangle_{\zeta^\mu} + \frac{1}{2} \ln(1-q) + \frac{1}{2} \frac{q}{1-q} \tag{7}$$

where q must be chosen to satisfy the fixed point equation

$$\frac{q}{1-q} = \alpha \left\langle \int \mathrm{D}z \left[\sum_\sigma (H(L_\mu^\sigma) - H(U_\mu^\sigma)) \right]^{-1} \sum_\sigma \left[\frac{\exp(-L_\mu^{\sigma 2}/2)}{\sqrt{2\pi q}} (L_\mu^\sigma - l_\mu^\sigma \sqrt{1-q}) \right. \right.$$
$$\left. \left. - \frac{\exp(-U_\mu^{\sigma 2}/2)}{\sqrt{2\pi q}} (U_\mu^\sigma - u_\mu^\sigma \sqrt{1-q}) \right] \right\rangle_{\zeta^\mu}. \tag{8}$$

Here we have introduced the abbreviations $L_\mu^\sigma = (l_\mu^\sigma + \sqrt{q} z)/\sqrt{1-q}$, $U_\mu^\sigma = (u_\mu^\sigma + \sqrt{q} z)/\sqrt{1-q}$, and $H(x) = \int_x^\infty \mathrm{D}z$, with $\mathrm{D}z = (\mathrm{d}z/\sqrt{2\pi}) \exp(-z^2/2)$. In (7) and (8), averages over the output statistics remain to be done.

3. Results

Let us now turn to results. Specializing in what follows to the case of monotonic non-decreasing g's, and supposing that the maximal capacity $\alpha = \alpha_c$ in (7), (8) is signalled by the Gardner criterion [6] $q \to 1$, asymptotic analysis of (8) gives

$$\alpha_c^{-1} = \left\langle \int_{u_\mu}^{\infty} Dz(u_\mu - z)^2 + \int_{-l_\mu}^{\infty} Dz(l_\mu + z)^2 \right\rangle_{\zeta^\mu}. \tag{9}$$

The optimal capacity is a function of both ϵ and κ, because l_μ and u_μ are (see equations (4) and (5)). Moreover, α_c depends on the input–output relation g, on the gain parameter γ, and on the statistics of the desired outputs. Note that, formally, (9) resembles results previously obtained for networks of multi-state neurons [9, 10]. This is a consequence of our *choice* to formalize the graded-response problem at hand by introducing a desired input stability κ for the mapping (1) *alongside* with an admissible output tolerance ϵ, and a final specialization to monotonic non-decreasing g's. For continuous input–output relations, a non-vanishing input stability is not to be had without non-zero output-error tolerance. The fact that the input–output relations considered in the present paper are *graded* implies a number of differences from previous work. The ζ^μ are generally related in a non-linear way to the l_μ and the u_μ which—in contrast to previous work [9, 10]—is not everywhere singular but, in many cases of interest, even invertible. Consequently, the statistics of the ζ^μ translates non-linearly into a statistics of the l_μ and the u_μ. In contrast to multi-state perceptrons, the graded-response perceptron can realize mappings with outputs in a continuous output range. Accordingly, one of the questions about this system to be answered is how variations in output statistics affect the network performance, and to what extent they can be compensated by appropriate choices of gain functions or gain parameters.

4. Results for exact mapping

The first case to consider is where the inputs are mapped *exactly* onto the desired outputs, i.e. the case $\kappa = \epsilon = 0$. For invertible input–output relations, this implies $l_\mu = u_\mu = h^\mu = (1/\gamma)g^{-1}(\zeta^\mu)$ in (5)–(9). Hence (9) leads to the remarkably simple relation

$$\alpha_c^{-1} = \langle (h^\mu)^2 \rangle_{\zeta^\mu} + 1. \tag{10}$$

This result shows that we have an upper bound $\alpha_c \leqslant 1$ for exact mapping. The actual value of α_c depends only on the second moment of the local-field distribution required to produce the desired ζ^μ statistics, and it is strictly smaller than 1, if $\langle (h^\mu)^2 \rangle_{\zeta^\mu} \geqslant \delta$ for some $\delta > 0$.

At first sight, this result seems to be in conflict with the fact that the pseudo-inverse solution (2) exists up to $\alpha = 1$. However, closer investigation reveals that (2) does *not* generally satisfy the spherical constraint $\sum_j J_j^2 = N$. To satisfy it, we *must* use the freedom of adding a coupling vector orthogonal to the space spanned by the stored input patterns. Let $\hat{J}^\perp$ be any such vector of unit length, and let us adopt the convention

$$J_j = J_j^I + c_\perp \sqrt{\alpha N}\, \hat{J}_j^\perp \tag{11}$$

for a solution to (1) that *does* satisfy the sperical constraint

$$\sum_j J_j^2 = \alpha N (c_I^2 + c_\perp^2) = N \tag{12}$$

with

$$c_I^2 = \frac{1}{\alpha N} \sum_{\mu,\nu} h^\mu (C^{-1})_{\mu,\nu} h^\nu. \tag{13}$$

We argue that, with a spherical constraint thus imposed, the optimal capacity must indeed be expected to be a monotonic decreasing function of $\langle (h^\mu)^2 \rangle_{\zeta^\mu}$ as suggested by (10). To see this, note that C, hence C^{-1}, are positive definite matrices. As a consequence $c_I^2 > 0$ in (12) and (13) for $h^\nu \not\equiv 0$. Moreover, c_I^2 would be a monotonic increasing function of the h^ν scale necessary to produce the desired output statistics. That is, $c_I^2(\{\lambda h^\nu\}) = \lambda^2 c_I^2(\{h^\nu\})$. Thus, if the necessary h^ν scale becomes large due to, e.g., a small value for the gain parameter γ, then the multiplicity of solutions based on the pseudo-inverse solution (11), which can, in principle, exist up to $\alpha = 1$, is lost due to the spherical constraint when $c_I^2 = \alpha^{-1}$, enforcing $c_\perp^2 = 0$. At this point the opening angle φ of the cone of solutions (11), $\varphi = \tan^{-1}(c_\perp / c_I)$ will be zero, signifying that $\alpha = \alpha_c$ for the given set of input patterns (as embodied in the $C_{\mu,\nu}$), output pattern statistics, input–output relation and gain parameter γ which determine the h^ν scale. That is, α_c must be expected to be a decreasing function of the h^ν scale hence of $\langle (h^\mu)^2 \rangle_{\zeta^\mu}$, as announced. To get a rough order-of-magnitude check of (10) let us work with the *average* of C^{-1} (rather than with C^{-1} itself), namely with the unit matrix $\mathbb{1}$ in (13). Taking $C^{-1} \simeq \mathbb{1}$ gives $c_I^2 = \langle (h^\mu)^2 \rangle_{\zeta^\mu}$, implying $\alpha_c^{-1} = \langle (h^\mu)^2 \rangle_{\zeta^\mu}$ as long as the resulting α_c is less than 1, and otherwise 1. Salient features of the 'typical' result (10) are thus reproduced fairly well by this simple argument.

5. Non-zero output tolerance and finite input stability

If we allow a finite, i.e. non-zero output tolerance $\epsilon > 0$, then $\alpha_c(\epsilon, \kappa, \gamma)$ is found to be a monotonic non-decreasing function of ϵ. The behaviour of α_c as a function of γ, given ϵ and κ, is strongly dependent on the output statistics. It is important to remark immediately that the results obtained are not always stable against replica symmetry breaking (RSB). One can show that, to check this stability, it is sufficient to look at the sign of the product of the eigenvalues of the matrix of transverse fluctuations, the so-called 'replicon' eigenvalue λ_R. One needs to satisfy

$$\lambda_R \equiv \alpha(1-q)^2 \left\langle \int Dz ([x^2]_{\mu,z} - [x]_{\mu,z}^2)^2 \right\rangle_{\zeta^\mu} - 1 < 0 \tag{14}$$

where

$$[f(x)]_{\mu,z} = \frac{\int_{l_\mu}^{u_\mu} d\lambda \int (dx/2\pi) f(x) \exp[ix(\lambda + \sqrt{q}z) - \frac{1}{2}(1-q)x^2]}{\int_{l_\mu}^{u_\mu} d\lambda \int (dx/2\pi) \exp[ix(\lambda + \sqrt{q}z) - \frac{1}{2}(1-q)x^2]}.$$

For $\kappa = 0$, and for monotonic non-decreasing input–output relations—hence simply connected I_μ—a straightforward calculation leads to the following interesting identity $\mu = 1, \ldots, p$

$$\operatorname{sgn}[-\lambda_R(\epsilon, \kappa = 0, \gamma)] = \operatorname{sgn}\left[\frac{\partial \alpha_c(\epsilon, \kappa = 0, \gamma)}{\partial \gamma}\right] \tag{15}$$

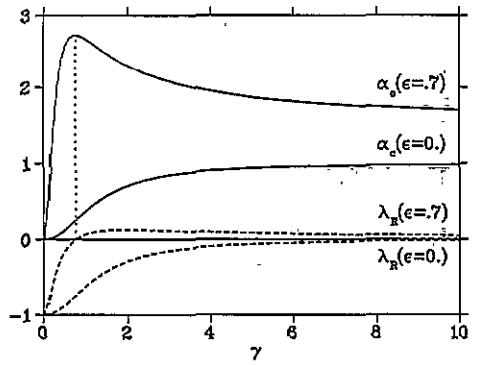

Figure 1. The optimal capacity α_c and the replicon eigenvalue λ_R as a function of the gain parameter γ for $\kappa = 0$ and for various values of ϵ. The output statistics is given by (16) with $d = 0$, i.e. it is uniform over the interval $[-1, 1]$.

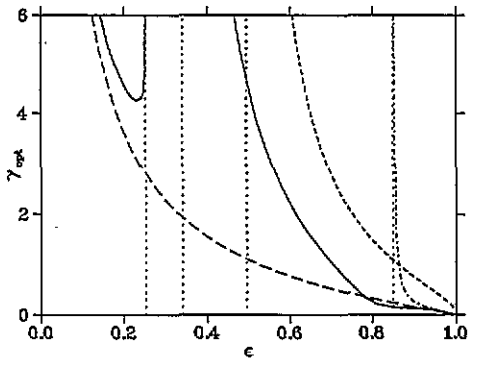

Figure 2. The optimal gain parameter γ_{opt} as a function of the output tolerance ϵ for $\kappa = 0$ and for different output statistics: equation (16) with $d = 0$ (long-dashes); equation (16) with $d = 0.5$ and $x = 0.75$ (full curve); equation (16) with $d = 0.5$ and $x = 1.0$ (short-dashes); equation (17) with $y = 0.7$ (chain curve). The dotted vertical lines indicate the asymptotes.

relating the sign of the replicon eigenvalue with the sign of the derivative of α_c with respect to the gain parameter γ. This relation is true for all ϵ, and independent of the statistics of the ζ^μ. Equation (15) implies the existence of an optimal gain parameter γ_{opt} such that, given ϵ and the output statistics, α_c is maximal at $\gamma = \gamma_{opt}$, and replica symmetry is unbroken for all γ satisfying $\gamma < \gamma_{opt}$. The optimal gain parameter γ_{opt} can be either finite or infinite, depending on the distribution of outputs ζ^μ and on the output-error tolerance ϵ.

In figures 1–3, we illustrate this behaviour, showing some representative examples for the input–output relation $g(x) = \tanh(x)$, which is widely used for networks trained by the back-propagation algorithm and for the following two types of output statistics,

$$P(\zeta^\mu) = \tfrac{1}{2}(1-d) + \tfrac{1}{2}d[\delta(\zeta^\mu - x) + \delta(\zeta^\mu + x)] \qquad d, x \in [0,1] \qquad (16)$$

and

$$P(\zeta^\mu) = \frac{\Theta(1-(\zeta^\mu)^2)}{2(1-y)}[\Theta(\zeta^\mu - y) + \Theta(-\zeta^\mu - y)] \qquad y \in [0,1]. \qquad (17)$$

The first distribution, (16), consists of a contribution which is constant over the interval $[-1, 1]$ and a pair of δ peaks at $\pm x$, with adjustable relative weights for the smooth and the singular part of $P(\zeta^\mu)$. The second distribution, (17), is made of two constant contributions extending inward from the boundaries of the interval $[-1, 1]$, symmetric about zero and with equal weight.

In figure 1, we show the behaviour of α_c as a function of γ for the output distribution (16) with $d = 0$, i.e. for a constant homogeneous distribution on the interval $[-1, 1]$. As long as $\epsilon < 1$, the critical storage capacity α_c approaches a constant smaller than or equal to 2, as $\gamma \to \infty$. The replicon eigenvalue (14) is also depicted in figure 1, illustrating the relation (15) between the sign of λ_R and the slope of the α_c against γ curve.

Figures 2 and 3 show γ_{opt} and $\alpha_c(\gamma_{opt})$ as a function of ϵ for $\epsilon \in [0, 1]$. The most interesting features are the following. For a constant homogeneous distribution of the outputs, i.e. for (16) with $d = 0$, γ_{opt} decreases monotonically from ∞ to 0, while $\alpha_c(\gamma_{opt})$

increases monotonically from 1 to ∞, as ϵ varies between 0 and 1. If one allows δ peaks in the output distribution (16), this behaviour changes depending on the position x of these peaks. For $d = 0.5$ and $x = 1$, i.e. for δ peaks at the boundaries of the interval $[-1, 1]$, γ_{opt} is infinite for $\epsilon \in [0, \frac{1}{2}]$, whereas it is finite for $\epsilon > \frac{1}{2}$ and decreases to zero, as ϵ approaches 1. The corresponding $\alpha_c(\gamma_{opt})$ increases monotonically from the value 1.33 up to 4. For $d = 0.5$ and $x = 0.75$, i.e. with δ peaks shifted to the interior of $[-1, 1]$, γ_{opt} is infinite only in the interval $\epsilon \in [0.25, 0.34]$; outside this interval, γ_{opt} is finite, and it diverges if ϵ approaches either zero or the boundaries of this interval. For $\epsilon > 0.34$, γ_{opt} is monotonic decreasing to zero, as $\epsilon \to 1$. The corresponding $\alpha_c(\gamma_{opt})$ shows a finite jump discontinuity at $\epsilon = 0.25$, the lower bound of the interval in which $\gamma_{opt} = \infty$. No anomaly of α_c could be detected at the upper bound of this interval.

For the output distribution (17) with $y = 0.7$, γ_{opt} is infinite throughout the interval $\epsilon \in [0, 0.85]$. The corresponding optimal capacity increases quickly from $\alpha_c = 1$ (at $\epsilon = 0$) to $\alpha_c = 2$ (at $\epsilon = 0.3$), and stays at $\alpha_c = 2$ for the remainder of the ϵ interval in which $\gamma_{opt} = \infty$, i.e. for all ϵ satisfying $g(\pm\infty) \in I_{out}(\pm y, \epsilon)$. Qualitatively the same behaviour is found for the piecewise-linear input–output relation $g(x) = \mathrm{sgn}(x) \min\{|x|, 1\}$.

In general, we expect $\gamma_{opt}(\epsilon)$ to be different from zero and the corresponding $\alpha_c(\gamma_{opt}, \epsilon)$ to be finite, as long as ϵ is such that $g(0) \notin I_{out}(\pm 1, \epsilon)$. Furthermore, $\gamma_{opt}(\epsilon)$ is infinite if in the distribution of outputs the total weight of the ζ^μ for which $g(0) \notin I_{out}(\zeta^\mu, \epsilon)$ and $g(\pm\infty) \in I_{out}(\zeta^\mu, \epsilon)$ *dominates* the weight of the ζ^μ for which $g(0) \in I_{out}(\zeta^\mu, \epsilon)$.

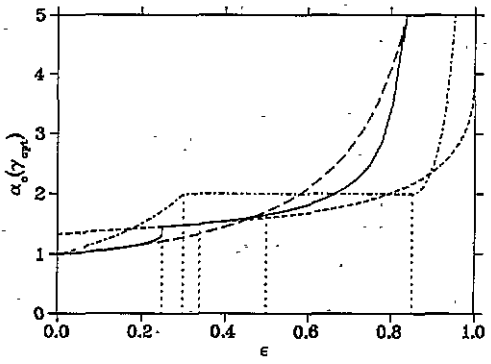

Figure 3. The optimal capacity α_c at γ_{opt} as a function of the output tolerance ϵ for $\kappa = 0$ and for the same output statistics as in figure 2.

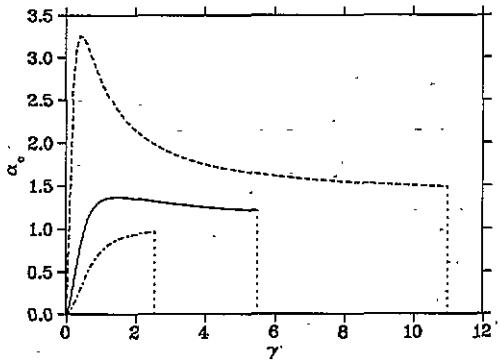

Figure 4. The optimal capacity α_c as a function of the gain parameter γ for $\kappa = 0.1$ and for various values of ϵ. The output statistics is given by (16) with $d = 0$, i.e. it is uniform over the interval $[-1, 1]$. The curves shown correspond to $\epsilon = 0.25$ (chain curve), $\epsilon = 0.5$ (full curve), and $\epsilon = 0.80$ (broken curve), respectively. The dotted vertical lines indicate the value of γ beyond which α_c is strictly zero.

In brief, if we demand correctness only of the *sign* of the output, we do recover the $\alpha_c = 2$ result of Cover [17] and Gardner [6]. By increasing the output tolerance, we eventually come to a point where the requirement on the mapping is weaker than the demand for a correct sign of the output, so that we can have $\alpha_c > 2$ for sufficiently large ϵ. Where exactly this occurs, and how large the ϵ that can sensibly be tolerated actually are, will naturally depend on g, on γ, and on the desired output statistics, as illustrated by figures 2 and 3.

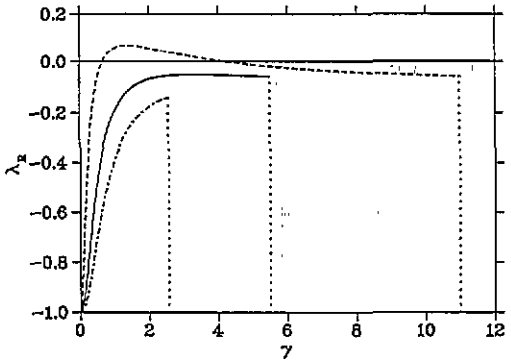

Figure 5. The same as figure 4, but for the replicon eigenvalue λ_R. Note the restoration of stability against RSB for large values of γ in the upper curve, corresponding to $\epsilon = 0.8$.

If both ϵ and κ are non-zero, then there is an upper bound on γ beyond which $\alpha_c(\epsilon, \kappa, \gamma) = 0$. The reason is that for ζ^μ in the range of steep parts of g, there will be an upper bound on γ beyond which the set $I_\mu = \{h^\mu;\ g(\gamma(h^\mu - \kappa)) \geqslant \zeta^\mu - \epsilon,\ g(\gamma(h^\mu + \kappa)) \leqslant \zeta^\mu + \epsilon\}$ is *empty*, signalling that the required output precision and the desired input stability are no longer compatible. The limiting γ will depend on the input–output relation g, on the *support* of the distribution of outputs and, naturally, on κ and ϵ. In figures 4 and 5 we illustrate this by showing the typical behaviour of α_c and λ_R as functions of γ, given the output statistics $P(\zeta^\mu)$ as well as ϵ and a non-zero κ. The figures represent results for the hyperbolic tangent input–output relation and for the constant output distribution (16) with $d = 0$. Again, the piecewise linear input–output relation leads to similar results.

In particular, we see that α_c drops to zero discontinuously at $\gamma = \tan^{-1}(\epsilon)/\kappa$. For sufficiently large ϵ, a maximum of α_c as a function of γ may occur, and RSB may be observed for sufficiently large γ. However, relation (15) equating the position of the α_c maximum with the boundary of stability against RSB is no longer valid in the present $\kappa \neq 0$ case. In all cases we have studied, though, replica symmetry was found unbroken up to and usually even beyond the position of the α_c maximum whenever such a maximum existed. RSB was found to occur at intermediate values of γ beyond the position of the maximum, the replica symmetric results becoming stable again for still higher values of γ, usually including the γ value at which the discontinuous breakdown of α_c occurs. The behaviour of γ_{opt} and $\alpha_c(\gamma_{opt})$ as functions of ϵ is not fundamentally different from that at $\kappa = 0$ as illustrated in figures 2 and 3.

There are two further directions in which the previous considerations allow immediate generalization. One is *annealed* dilution [18], where the perceptron is chosen to organize itself in a manner that only a fraction f of its N bonds remain; this fraction, however, being optimally adapted to the storage task. With an appropriately adapted spherical constraint on the couplings [18] we find, again in the replica symmetric limit,

$$\alpha_c^{-1} = \frac{\left\langle \int_{u_\mu}^{\infty} Dz(u_\mu - z)^2 + \int_{-l_\mu}^{\infty} Dz(l_\mu + z)^2 \right\rangle_{\zeta^\mu}}{f + 2u\exp(-u^2)/\sqrt{\pi}} \tag{18}$$

where u and f are related through $f = \mathrm{erfc}(u)$. The result (9) is recovered in the limit $f \to 1$. For any $f < 1$, replica symmetry is found to be broken on the Gardner–Derrida line [6, 7].

The second possibility is to turn to the Gardner–Derrida [7] ensemble by introducing a soft non-negative error measure $\epsilon(\zeta^\mu, h^\mu) \geqslant 0$, taken to be zero only if $\zeta^\mu = g(\gamma h^\mu)$, and to investigate the canonical ensemble generated by $E = \sum_\mu \epsilon(\zeta^\mu, h^\mu)$ at a given inverse temperature β. The free energy in the replica symmetric limit is given by

$$-\beta f(\beta) = \alpha \left\langle \int \mathrm{D}z \ln \int \frac{\mathrm{d}\lambda}{\sqrt{2\pi(1-q)}} \exp\left\{ -\frac{1}{2}\left(\frac{\lambda + \sqrt{q}z}{\sqrt{1-q}}\right)^2 - \beta\epsilon(\zeta^\mu, \lambda)\right\}\right\rangle_{\zeta^\mu} + \frac{1}{2}\ln(1-q) + \frac{1}{2}\frac{q}{1-q} + \frac{1}{2}(1 + \ln 2\pi) \tag{19}$$

where q is chosen to make the right-hand side of (19) stationary. Here one might investigate the behaviour of the average energy as a function of temperature to assess network performance. So far, however, we have not investigated this case in any detail.

A third possible topic, the generalization ability of graded-reponse perceptrons has recently been investigated by Bös *et al* [19].

6. Summary and outlook

In summary, we have computed optimal storage capacities of graded-response perceptrons within the Gardner approach to neural networks. Since optimal capacities are found to depend mainly on local-field distributions necessary to produce the desired output statistics (see, e.g. equation (10)), optimal capacities are strongly influenced by input–output relations and, more pronounced even, by gain parameters. Gain parameters can therefore be used fairly effectively to compensate for the effects of changes in ouput statistics on optimal capacities. However, if a certain non-zero input stability is demanded, there are upper limits to the gain parameter γ beyond which the required input stability is no longer compatible with the desired output precision so that the task set to the graded-response perceptron becomes unsolvable. In the case of sigmoid input–output relations such as $g(x) = \tanh(x)$, the problem arises first for outputs (thus local fields) near zero, where $\tanh(x)$ has its maximum slope. Incidentally, a related phenomenon is also observed in recursive networks of graded-reponse neurons [15]. At high gains, the local field distributions of stationary configurations develop a gap around zero, implying that in recursively stable situations small fields, thus small outputs, are systematically avoided, because at small fields the combination of input stability and output precision necessary to produce recursively stable configurations are incompatible for high gains. Which ϵ–κ combinations are generally required to achieve recursive stability near the optimal capacity of networks of graded response neurons is an interesting question which we are, as yet, unable to answer and which certainly deserves further study.

An interesting and, indeed, unexpected result is the restoration of local stability against RSB at the Gardner–Derrida line for large values of the gain parameter γ at non-zero κ, as illustrated in figure 5. As yet, it is unclear whether this stable phase at large γ is unique or whether it coexists with other stable phases which do exhibit RSB, and, if so, which of them would describe the proper physics of the problem.

Note also that our results concerning stability with respect to RSB in general shed some light on a popular argument which states that convexity of a solution space implies connectedness and hence precludes RSB. For the systems with monotonic non-decreasing input–output relations we have investigated quantitatively, the intervals I_μ *are* simply connected, implying that the set of couplings satisfying (4) *is* convex. This convexity

in $\mathbb{R}^N$ does, indeed, imply connectedness in $\mathbb{R}^N$, but not necessarily within the subset of J_j's which also satisfy the spherical constraint. Thus RSB is not precluded, and is indeed observed, despite convexity of the set of couplings satisfying (4).

In the present paper we have not evaluated the case of perceptrons with non-monotonic input–output relations in any detail, though we do consider such systems worth studying. In particular, we expect them to give rise to new and enhanced processing capabilities. To mention just one example, a perceptron with an input–output relation given by $g(x) = \chi_{[1/2,3/2]}(x)$—the characteristic function of the interval $[\frac{1}{2}, \frac{3}{2}]$—*is* able to solve the notorious XOR problem *without* additional hidden units: Take $J_1 = J_2 = 1$ and a 0–1 representation of inputs. This example, however simple, might indicate that considerable processing power, or—from a complementary point of view—considerable design simplifications, are to be gained by employing non-monotonic input–output relations when realizing, for instance, Boolean logic in networks of 'simple' processing elements. A systematic investigation of perceptrons with non-monotonic input–output relations is currently under way [20].

Acknowledgments

This work has been supported in part by the Research Fund of the K U Leuven (grant no. OT/91/13). One of us (RK) would like to thank the neural networks group at the Instituut voor Theoretische Fysica of the K U Leuven for the kind hospitality extended to him during one month's stay in Leuven where this work was done. One of us (DB) is indebted to the Belgian National Fund for Scientific Research for support as a Research Director.

References

[1] Amit D J 1989 *Modeling Brain Function—The World of Attractor Neural Networks* (Cambridge: Cambridge University Press)
[2] Müller B and Reinhard J 1990 *Neural Networks—An Introduction* (Berlin: Springer)
[3] Domany E, van Hemmen J L and Schulten K (ed) 1991 *Models of Neural Networks* (Heidelberg: Springer)
[4] Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Redwood City: Addison-Wesley)
[5] Amit D J, Gutfreund H and Sompolinsky H 1987 *Ann. Phys., NY* **173** 30
[6] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[7] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[8] Rieger H 1990 *J. Phys. A: Math. Gen.* **23** L1273
[9] Mertens S, Köhler H M and Bös S 1991 *J. Phys. A: Math. Gen.* **24** 4941
[10] Bollé D, Dupont P and van Mourik J 1991 *Europhys. Lett.* **15** 893; 1992 *Physica* A **185** 357
[11] Treves A 1990 *Phys. Rev.* A **42** 2418; 1990 *J. Phys. A: Math. Gen.* **23** 2631
[12] Marcus C M and Westervelt R M 1989 *Phys. Rev.* A **40** 501
Marcus C M, Waugh F M and Westervelt R M 1990 *Phys. Rev.* A **41** 3355
[13] Kühn R 1990 *Statistical Mechanics of Neural Nerworks (Proc. XI. Sitges Conference: Springer Lecture Notes in Physics 368)* ed L Garrido (Berlin: Springer) p 19
Kühn R, Bös S and van Hemmen J L 1991 *Phys. Rev.* A **43** 2084
[14] Shiino M and Fukai T 1990 *J. Phys. A: Math. Gen.* **23** L1009
[15] Kühn R and Bös S 1993 *J. Phys. A: Math. Gen.* **26** 831
[16] Rumelheart D E and McClelland J L 1986 *Parallel Distributed Processing* vol 1 (Cambridge, MA: MIT Press)
[17] Cover T M 1965 *IEEE Trans. Electron Comput.* **14** 326
[18] Bouten M, Engel A, Komoda A and Serneels R 1990 *J. Phys. A: Math. Gen.* **23** 4643
[19] Bös S, Kinzel W and Opper M 1993 *Phys. Rev.* E **47** 1384
[20] Bollé D, Kühn R and van Mourik J in preparation