Optimal colored perceptrons

D. Bollé*,† and P. Kozłowski†,‡
Instituut voor Theoretische Fysica, Katholieke Universiteit Leuven, B-3001 Leuven, Belgium
(Received 13 December 2000; published 26 June 2001)

Ashkin-Teller type perceptron models are introduced. Their maximal capacity per number of couplings is calculated within a first-step replica-symmetry-breaking Gardner approach. The results are compared with extensive numerical simulations using several algorithms.

DOI: 10.1103/PhysRevE.64.011915 PACS number(s): 87.10.+e, 02.50.−r, 64.60.Cn

I. INTRODUCTION

The perceptron that was first analyzed with statistical mechanics techniques in the seminal paper of Gardner [1] is by now a well-known and standard model in theoretical studies and practical applications in connection with learning and generalization [2–5]. A number of extensions of the perceptron model have been formulated, including many-state and graded-response perceptrons (e.g., [6–11]). Here we present some new extensions allowing for so-called colored or Ashkin-Teller type neurons, i.e., different types of binary neurons at each site possibly having different functions.

The idea of looking at such a model is based upon our recent work on Ashkin-Teller recurrent neural networks [12,13]. There we showed that for this model with two types of binary neurons interacting through a four-neuron term and equipped with a Hebb learning rule, both the thermodynamic and dynamic properties suggest that such a model can be more efficient than a sum of two Hopfield models. For example, the quality of pattern retrieval is enhanced through a larger overlap at higher temperatures and the maximal capacity is increased. For more details and an underlying neurobiological motivation for the introduction of different types of neurons we refer to [13].

In the light of these results an interesting question is whether such a colored perceptron can still be more efficient than the standard perceptron. In other words, can it have a larger maximal capacity than the one of a standard perceptron, which is known [1] to be $\alpha_c=2$ (for random uncorrelated patterns). It has been suggested that this number is characteristic for all binary networks independent of the multiplicity of the neuron interactions. Thereby, the capacity is defined as the thermodynamic limit of the ratio of the total number of bits per (input) neuron to be stored and the total number of couplings per (output) neuron [8]. We remark that ''input'' and ''output'' refer specifically to the perceptron case.

In the sequel the maximal capacity of colored perceptron models is studied using the Gardner approach [1,14]. The main advantage of this approach is that in order to determine this maximal capacity, there is no need to specify explicitly the *optimal* set of couplings for which, this maximum is reached. First-step replica-symmetry-breaking effects are evaluated and the analytic results are compared with extensive numerical simulations using various learning algorithms.

The rest of this paper is organized as follows. In Sec. II we introduce two Ashkin-Teller type perceptron models. Section III contains the replica theory and determines the maximal capacity by calculating the available volume in the space of couplings both in the replica-symmetric (Sec. III A) and the first-step replica-symmetry-breaking approximation (Sec. III B). Section IV describes the results of numerical simulations with algorithms obtained by generalizing various algorithms for simple perceptrons. In Sec. V we present our conclusions. Finally, two appendices contain some technical details of the derivations.

II. THE MODEL

Let us first formulate the colored perceptron models. We consider p input patterns $\boldsymbol{\zeta}^\mu=\{\zeta_i^\mu\}=\{\xi_i^\mu,\eta_i^\mu\}$, $i=1,\ldots,N$ consisting of two different types of patterns $\boldsymbol{\xi}^\mu=\{\xi_i^\mu\}$ and $\boldsymbol{\eta}^\mu=\{\eta_i^\mu\}$, and a corresponding set of outputs $\boldsymbol{\zeta}_0^\mu=\{\xi_0^\mu,\eta_0^\mu\}$ $\mu=1,\ldots,p$ that are determined by

$$\xi_0^\mu=\mathrm{sgn}(h_1^\mu+\eta_0^\mu h_3^\mu), \tag{1}$$

$$\eta_0^\mu=\mathrm{sgn}(h_2^\mu+\xi_0^\mu h_3^\mu), \tag{2}$$

$$\xi_0^\mu\eta_0^\mu=\mathrm{sgn}(\eta_0^\mu h_1^\mu+\xi_0^\mu h_2^\mu), \tag{3}$$

where h_r $(r=1,2,3)$ are the local fields acting on the patterns ξ, η, and their product $\xi\eta$, respectively,

$$h_1^\mu=\frac{1}{n_1}\sum_i J_i^{(1)}\xi_i^\mu, \qquad h_2^\mu=\frac{1}{n_2}\sum_i J_i^{(2)}\eta_i^\mu, \tag{4}$$

$$h_3^\mu=\frac{1}{n_3}\sum_i J_i^{(3)}\xi_i^\mu\eta_i^\mu, \qquad n_r^2=\sum_i (J_i^{(r)})^2, \quad r=1,2,3. \tag{5}$$

Both types of input patterns and their corresponding outputs are supposed to be independent identically distributed random variables taking the values $+1$ or -1 with probability 1/2.

The set of three equations (1)–(3) defines a mapping of the inputs $\boldsymbol{\zeta}_i^\mu$ onto the corresponding outputs $\boldsymbol{\zeta}_0^\mu$. We call it model I. The specific form of these equations is related to the transition probabilities for a spin flip in the dynamics {see the expressions (9) in [12]}. Although for the Hebb learning rule

*Email address: desire.bolle@fys.kuleuven.ac.be
†Also at Interdisciplinair Centrum voor Neurale Netwerken, K. U. Leuven, Belgium.
‡Email address: piotr.kozlowski@fys.kuleuven.ac.be

the spin-flip dynamics defined by all three or only two of these transition probabilities lead to the same equilibrium properties, this is not necessarily the case for the optimal couplings. Therefore, a second model, denoted by II, is defined by considering only the two equations (1) and (2). When $|h_3|>|h_1|$ and $|h_3|>|h_2|$ then the relations (1)–(2) are satisfied by two (out of the four possible) values of the output $\boldsymbol{\zeta}_0$, otherwise model II gives the same output as model I. In other words, due to the presence of the η_0^μ and ξ_0^μ in the gain functions, model II contains more freedom and, strictly speaking, it is not a mapping. This additional freedom arises from the weaker constraints imposed on the couplings such that their available space is systematically larger. So, when the available volume of the coupling space for model I already shrinks to zero, the one for model II may still be finite. Therefore, we expect that model II allows for a bigger capacity.

At this point we remark that when all $J_i^{(3)}$ are equal to zero we find back two independent standard binary perceptron models. In the sequel we take the couplings to satisfy the spherical constraint $n_r=\sqrt{N}$.

III. REPLICA THEORY FOR THE MAXIMAL CAPACITY

The colored perceptron is trained to store correctly $p=\frac{3}{2}\alpha N$ patterns with α the loading capacity. The factor 3/2 follows naturally from the definition of capacity given in the introduction. A pattern is stored correctly when the so-called aligning field [15] is bigger than a certain constant $\kappa\geqslant 0$ whereby the latter indicates the stability. It is a measure for the size of the basin of attraction of that pattern. Specifically we require that

$$\lambda_\xi^\mu(\{J\})=\xi_0^\mu(h_1^\mu+\eta_0^\mu h_3^\mu)>\kappa_\xi\geqslant 0, \tag{6}$$

$$\lambda_\eta^\mu(\{J\})=\eta_0^\mu(h_2^\mu+\xi_0^\mu h_3^\mu)>\kappa_\eta\geqslant 0, \tag{7}$$

$$\lambda_{\xi\eta}^\mu(\{J\})=(\xi_0^\mu h_1^\mu+\eta_0^\mu h_2^\mu)>\kappa_{\xi\eta}\geqslant 0, \tag{8}$$

with $\{J\}=\{J_i^{(r)}\}$ denoting the configurations in the space of interactions. For $\kappa_\xi=\kappa_\eta=\kappa_{\xi\eta}=0$ all patterns that satisfy equations (1)–(3) also satisfy Eqs. (6)–(8). We remark that for model II the last inequality is superfluous.

The aim is then to determine the maximal value of the loading α for which couplings satisfying Eqs. (6)–(8) can still be found. In particular, the question whether this model can be more efficient than the existing two-state models is relevant.

Following Refs. [1,14] we formulate the problem as an energy minimization in the space of couplings with the formal energy function defined as

$$E(\{J\})=\sum_\mu\,(1-\Theta[\lambda_\xi^\mu(\{J\})-\kappa_\xi]\Theta[\lambda_\eta^\mu(\{J\})-\kappa_\eta]\,\Theta[\lambda_{\xi\eta}^\mu(\{J\})-\kappa_{\xi\eta}]). \tag{9}$$

We remark that for model II the third Θ factor is absent. The quantity above counts the number of weakly embedded patterns, i.e., the patterns with stability less than $\kappa_\xi,\kappa_\eta,\kappa_{\xi\eta}$. Therefore, the minimal energy gives the minimal number of patterns that are stored incorrectly. This number is zero below a maximal storage capacity $\alpha_c(\kappa_\xi,\kappa_\eta,\kappa_{\xi\eta})$.

The basic quantity to start from is the partition function

$$Z(\beta)=\langle\exp[-\beta E(\{J\})]\rangle_{\{J\}} \tag{10}$$

$$\langle\cdots\rangle_{\{J\}}=\int\prod_i\,\mathrm{d}J_i\prod_r\,\delta\left[\sum_i\,(J_i^{(r)})^2-N\right]\cdots, \tag{11}$$

with β the inverse temperature. As usual it is $\ln Z$, which is assumed to be a self-averaging extensive quantity [1,15]. The related free energy per site

$$f=-\lim_{N\to\infty}\frac{1}{N\beta}\ln Z(\beta) \tag{12}$$

is equal, in the limit $\beta\to\infty$, to

$$\frac{\langle E\rangle_\infty}{N}\equiv\lim_{\beta\to\infty}\frac{\langle E(\{J\})\exp[-\beta E(\{J\})]\rangle_{\{J\}}}{NZ(\beta)}, \tag{13}$$

which is the minimal fraction of wrong patterns [recall Eq. (9)].

In order to perform the average over the disorder in the input patterns $\boldsymbol{\zeta}^\mu$ and the corresponding outputs ζ_0^μ we employ the replica method. The calculations proceed in a standard way although the technical details are much more complex. Introducing the order parameters $q_{\gamma\tau}^{(r)}=(1/N)\Sigma_i J_i^{(r)\gamma}J_i^{(r)\tau}$, with $r=1,2,3$ and $\gamma,\tau=1,\ldots,n$ we write following [1],

$$\langle Z^n(\beta)\rangle=\int\prod_{r,\gamma,\tau>\gamma}\left(\frac{\mathrm{d}q_{\gamma\tau}^{(r)}\mathrm{d}\phi_{\gamma\tau}^r}{2\pi/N}\right)\prod_{r,\gamma}\frac{\mathrm{d}\epsilon_\gamma^r}{2\pi}\times\exp N\left\{\frac{3}{2}\alpha G_0(q_{\gamma\tau}^{(r)})+\sum_{r,\gamma,\tau>\gamma}iq_{\gamma\tau}^{(r)}\phi_{\gamma\tau}^r+G_1(\phi_{\gamma\tau}^r,\epsilon_\gamma^r)\right\}, \tag{14}$$

$$G_0=\ln\left\{\prod_\gamma\left(\left[e^{-\beta}\int\prod_{r'}\frac{\mathrm{d}\lambda^{r'\gamma}}{2\pi}+(1-e^{-\beta})\times\int_{\kappa_\xi}^\infty\int_{\kappa_\eta}^\infty\int_{\kappa_{\xi\eta}}^\infty\prod_{r'}\frac{\mathrm{d}\lambda^{r'\gamma}}{2\pi}\right]\int\prod_{r'}\mathrm{d}x^{r'\gamma}\times\exp\left\{\sum_{r'}ix^{r'\gamma}\lambda^{r'\gamma}-\frac{1}{2}[(x^{1\gamma}+x^{3\gamma})^2+(x^{2\gamma}+x^{3\gamma})^2+(x^{1\gamma}+x^{2\gamma})^2]-\sum_{\tau>\gamma}[(x^{1\gamma}+x^{3\gamma})(x^{1\tau}+x^{3\tau})q_{\gamma\tau}^{(1)}+(x^{2\gamma}+x^{3\gamma})(x^{2\tau}+x^{3\tau})q_{\gamma\tau}^{(2)}+(x^{1\gamma}+x^{2\gamma})\times(x^{1\tau}+x^{2\tau})q_{\gamma\tau}^{(3)}]\right\}\right)\right\}$$

$$G_1=\ln\left\{\int\prod_{r,\gamma}(\mathrm{d}J^{(r)\gamma})\exp\left[i\sum_{r,\gamma}\epsilon^r_\gamma((J^{(r)\gamma})^2-1)\right.\right.$$
$$\left.\left.-i\sum_{r,\gamma,\tau>\gamma}\phi^r_{\gamma\tau}J^{(r)\gamma}J^{(r)\tau}\right]\right\},$$

where $\langle\ldots\rangle$ denotes the average over the patterns, $r'=1,2,3$ for model I and 1,2 for model II. Because of the latter we remark that for model II the formula for G_0 can be simplified: the integrals with respect to $\lambda^{3\gamma}$ and $x^{3\gamma}$ are not present and thus $x^{3\gamma}$, $x^{3\tau}$, and $\lambda^{3\gamma}$ have to be set to zero. Because of this simplification we only outline explicitly the calculations for model II in the sequel. The corresponding formulas for model I can be found in Appendix B.

A. Replica symmetric ansatz

We continue by making the replica-symmetric (RS) ansatz $q^{(r)}_{\gamma\tau}=q^{(r)}, \phi^r_{\gamma\tau}=i\phi^r, \epsilon^r_\gamma=i\epsilon^r$. Moreover, for convenience, we set $q^{(1)}=q^{(2)}=q^{(3)}=q$. The latter is justified for model I because of the symmetry present in this model. Furthermore, since we are going to take all $q^{(r)}\to 1$ in the Gardner-Derrida analysis anyway, we keep this equality also for model II. Taking then the limits $\beta\to\infty$, $N\to\infty$, and $n\to 0$ we arrive, in the case of model II, at

$$v=\lim_{N\to\infty}\frac{1}{N}\langle\ln Z\rangle$$
$$=\frac{3}{2}\alpha\int \mathrm{D}[s_1(q/2)]\mathrm{D}[s_2(3q/2)]\ln\psi_{RS}(\kappa_\xi,\kappa_\eta,s_1,s_2,q)$$
$$+\frac{3}{2}\left[\ln(1-q)+\frac{1}{1-q}+\ln 2\pi\right] \tag{15}$$

with

$$\psi_{RS}(\kappa_\xi,\kappa_\eta,s_1,s_2,q)=\int_{l_1}^{\infty}\int_{l_2}^{l_3}\prod_\nu \mathrm{D}[s_\nu(1)] \tag{16}$$

where $\nu=1,2$, $\mathrm{D}[s(y)]=\mathrm{d}s\exp(-1/2ys^2)/\sqrt{2\pi y}$ is a modified Gaussian measure,

$$l_1=\frac{\sqrt{\frac{2}{3}}\left[\frac{1}{2}(\kappa_\xi+\kappa_\eta)-s_2\right]}{\sqrt{1-q}}, \tag{17}$$

$$l_2=\frac{\sqrt{2}(\kappa_\eta-s_2-s_1)}{\sqrt{1-q}}-u_2\sqrt{3}, \tag{18}$$

$$l_3=\frac{\sqrt{2}(-\kappa_\xi+s_2-s_1)}{\sqrt{1-q}}+u_2\sqrt{3}, \tag{19}$$

and q takes those values that minimize v, the available volume in the space of couplings. For the corresponding expression in the case of model I we refer to Appendix B.

Taking $\kappa_\xi=\kappa_\eta=\kappa$ and supposing that the maximal capacity, $\alpha_c=\alpha_{RS}$, is signaled by the Gardner-like criterion $q\to 1$, we obtain

$$\alpha_{RS}(\kappa)=\lim_{q\to 1}\left\{\frac{-\ln(1-q)-\frac{1}{1-q}-\ln 2\pi}{\int \mathrm{D}(s_1(q/2))\mathrm{D}(s_2(3q/2))\ln\psi_{RS}(\kappa,\kappa,s_1,s_2,q)}\right\}. \tag{20}$$

This maximal capacity as a function of κ is shown for both models in Figs. 1 and 2 as a full line. For model I we obtain, e.g., $\alpha_{RS}(\kappa=0)=1.92$, a value that is smaller than the Gardner capacity for the simple perceptron. For model II however, we get the interesting result that $\alpha_{RS}(\kappa=0)=2.74>2$.

B. First-step replica symmetry breaking

It is straightforward to show geometrically that learning almost antiparallel patterns, i.e., patterns satisfying $(\boldsymbol{\xi}^\mu\xi_0^\mu,\boldsymbol{\eta}^\mu\eta_0^\mu)\approx-(\boldsymbol{\xi}^\nu\xi_0^\nu,\boldsymbol{\eta}^\nu\eta_0^\nu)$ results in a splitting of the space of couplings into disconnected regions. This suggests that RS is broken and, consequently, the results for α_{RS} found in Sec. III A are only upperbounds for the true capacity. Therefore, we want to improve the RS results by applying the first step of Parisi's replica-symmetry-breaking (RSB) scheme (e.g., [21]). So, we assume that the $q^{(r)}_{\gamma\tau}$ in Eq. (14) have the following matrix block structure

$$q^{(r)}_{\gamma\tau}=\begin{cases}q_1^{(r)} & \text{if } \mathrm{int}\left(\frac{(\gamma-1)m}{n}\right)=\mathrm{int}\left(\frac{(\tau-1)m}{n}\right)\\ q_0^{(r)} & \text{otherwise,}\end{cases} \tag{21}$$

where n is the size of the matrix $q^{(r)}_{\gamma\tau}$, m is the number of diagonal blocks, and $\mathrm{int}(x)$ denotes the integer part of x.

For model II we take $q^{(1)}_{\gamma\tau}=q^{(2)}_{\gamma\tau}\neq q^{(3)}_{\gamma\tau}$ reflecting the symmetry of this model. For model I we repeat that all $q^{(r)}$'s can

be taken equal. We then consider the limits $q_1^{(r)} \to 1$ and $n \to 0$ in such a way that $m/(1-q_1)$, with $q_1^{(1)} = q_1^{(2)} = q_1^{(3)} = q_1$, remains finite. After a tedious calculation we arrive at the following expression for the RSB1 maximal capacity for model II

$$\alpha_{RSB1}(\kappa) = \min_{q_0^{(1)}, q_0^{(3)}, M} \left\{ \frac{-\frac{2}{3}\left[\ln(1+M) + \frac{q_0^{(1)} M}{(1+M)(1-q_0^{(1)})} + \frac{1}{2}\ln(1+M_3) + \frac{1}{2}\frac{q_0^{(3)} M}{(1+M_3)(1-q_0^{(1)})} \right]}{\int \mathrm{D}t_1 \mathrm{D}t_2 \ln \psi_{RSB1}(\kappa, t_1, t_2, q_0^{(1)}, q_0^{(3)}, M)} \right\} \tag{22}$$

with

$$r_3 = \frac{1-q_0^{(3)}}{1-q_0^{(1)}}, \quad M_3 = M r_3, \quad M = \frac{m(1-q_0^{(1)})}{1-q_1}, \tag{23}$$

and $\mathrm{D}t_i = \mathrm{d}t_i \exp[-(1/2)t_i^2]/\sqrt{2\pi}$ a Gaussian measure. The explicit form of the function $\psi_{RSB1}(\kappa, t_1, t_2, q_0^{(1)}, q_0^{(3)}, M)$ can be found in Appendix A. An analogous form for model I is written down in Appendix B.

The results are presented in Figs. 1 and 2 as full lines. As expected they lie below the RS results confirming the breaking of RS, e.g., $\alpha_{RSB1}(\kappa = 0) = 1.83$ for model I and 2.28 for model II. We remark that the breaking for model II is stronger than for model I, the reason being that model II allows more freedom as explained in the introduction. Finally, on the basis of results in the literature for the simple perceptron [15], [16] we expect that the RSB1 results are very close to the exact ones. This is further examined by performing numerical simulations as described in the following section.

IV. NUMERICAL SIMULATIONS

The idea of these simulations is to train the network with a certain learning algorithm in order to learn as many random patterns as possible. The main technical difficulties are to find an efficient algorithm and prove its convergence.

We have tried to generalize various algorithms proposed for simple perceptrons [17–20]. The most effective ones appeared to be some particular generalization of the adaptive Gardner algorithm [18] and the Adatron algorithm [19]. In the sequel we only report on the results obtained with these two algorithms. We remark that we have chosen $\kappa_\xi = \kappa_\eta = \kappa_{\xi\eta} = \kappa$ in all simulations.

One of the algorithms that has demonstrated its efficiency and for which convergence has been shown in the case of the standard perceptron is given in Ref. [18]. It is an adaptive version of the original algorithm proposed by Gardner [1]. Using heuristic arguments presented in [18] we have constructed for the coloured perceptron model II the following analogous learning rule

$$J_i^{(1)} \to J_i^{(1)} + \xi_0^\mu \xi_i^\mu \tfrac{1}{2}(\kappa_\xi - \lambda_\xi^\mu)\Theta(\kappa_\xi - \lambda_\xi^\mu), \tag{24}$$

$$J_i^{(2)} \to J_i^{(2)} + \eta_0^\mu \eta_i^\mu \tfrac{1}{2}(\kappa_\eta - \lambda_\eta^\mu)\Theta(\kappa_\eta - \lambda_\eta^\mu), \tag{25}$$

$$J_i^{(3)} \to J_i^{(3)} + \xi_0^\mu \eta_0^\mu \xi_i^\mu \eta_i^\mu \tfrac{1}{2}[(\kappa_\xi - \lambda_\xi^\mu)\Theta(\kappa_\xi - \lambda_\xi^\mu) + (\kappa_\eta - \lambda_\eta^\mu)\Theta(\kappa_\eta - \lambda_\eta^\mu)]. \tag{26}$$

The form of the algorithm for model I is a bit different and given in Appendix B. This algorithm should be carried out sequentially over the patterns and sequentially or parallel over the couplings as long as one of the arguments of the Θ functions is positive. It appears to have the characteristics of the most efficient, nonlinear algorithm discussed in [18].

Using this learning rule we have trained networks of sizes $50 \leqslant N \leqslant 1000$ sites (depending on the value of κ) in order to store perfectly as many randomly chosen patterns as possible. For each value of κ we have calculated the maximal capacity for different N and extrapolated the results to $N = \infty$. Results for a given value of κ and N are averages over 1000 samples. As shown in Figs. 1 and 2 this algorithm performs especially well for small values of κ for both the models I and II.

The second algorithm we report on is the Adatron algorithm [19] that works in a different way. Instead of searching the maximal capacity for a given stability it tries to find the maximal stability for a given capacity. The derivation of this algorithm and a proof of its convergence are based upon the assumption that the problem can be formulated as a quadratic optimization with linear constraints [19,7]. Such a formulation cannot be given for the colored perceptron model, because the three different types of couplings have to be normalized independently and because the stability conditions (6)–(7) are more complex. Hence, a straightforward generalization similar to the one for the Potts model [7] is not possible. Below we describe a learning rule that tries to incorporate the ideas of the Adatron approach. We assume that the couplings can be written in the form (cfr., [19] and references therein)

$$J_i^{(1)} = \frac{1}{N}\sum_\mu x_1^\mu \xi_0^\mu \xi_i^\mu, \quad J_i^{(2)} = \frac{1}{N}\sum_\mu x_2^\mu \eta_0^\mu \eta_i^\mu,$$

$$J_i^{(3)} = \frac{1}{N}\sum_\mu x_3^\mu \xi_0^\mu \eta_0^\mu \xi_i^\mu \eta_i^\mu, \tag{27}$$

where x_r^μ $(r = 1,2,3)$ are the so-called embedding strengths of pattern μ. Then, in the case of model II the couplings are updated by modifying x_r^μ with the following increments

$$\delta x_1^\mu = \tfrac{1}{2}\max\{-x_1^\mu - x_3^\mu, \gamma(1 - n_1 \lambda_\xi^\mu)\}, \tag{28}$$

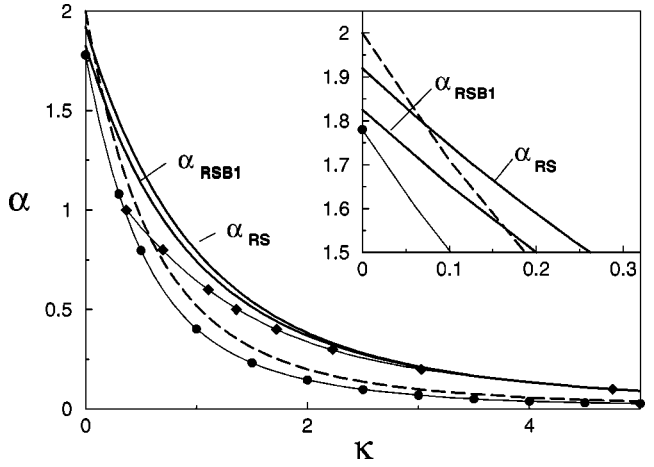

FIG. 1. The maximal capacity of the colored perceptron model I as a function of κ. Theoretical results for α_{RS} and α_{RSB1} are indicated by the thick solid lines. The circles are the results of the simulations for the adaptive Gardner algorithm, the diamonds for the Adatron algorithm. The error bars are smaller than the size of the symbols (not in the inset). The solid thin lines are polynomial fits to these results. The maximal capacity of a simple perceptron is indicated with a broken line.

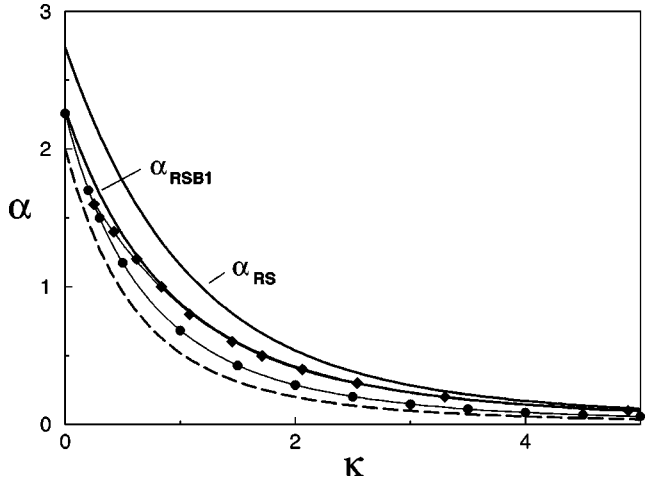

FIG. 2. The maximal capacity of the colored perceptron model II as a function of κ. The meaning of the symbols is as in Fig. 1.

$$\delta x_2^\mu = \frac{1}{2}\max\{-x_2^\mu - x_3^\mu, \gamma(1-n_2\lambda_\eta^\mu)\}, \tag{29}$$

$$\delta x_3^\mu = \frac{1}{4}(\max\{-x_1^\mu - x_3^\mu, \gamma(1-n_3\lambda_\xi^\mu)\} + \max\{-x_2^\mu - x_3^\mu, \gamma(1-n_3\lambda_\eta^\mu)\}). \tag{30}$$

This is done sequentially over the patterns. We remark that again the algorithm for model I is somewhat different (see Appendix B). For each value of the capacity we have considered system sizes $50 \leqslant N \leqslant 500$ and extrapolated the results to $N=\infty$. The best results were obtained for a learning rate $\gamma \in (0,2)$. Results for each size are averages over 1000 samples. For small values of the capacity the algorithm gives better results, both in the case of models I and II than the first algorithm we have discussed, as shown in Figs. 1 and 2. For larger values of the capacity, however, it performs worse. The results for the Adatron algorithm are displayed only in the region where they are better than the results for the Gardner algorithm. We remark that the numerical simulations with the different algorithms give different results and that we have not shown their convergence analytically such that, in principle, the values for α_c obtained here are lower bounds.

Looking at Figs. 1 and 2 in more detail we see that for the whole range of κ the values of the maximal capacity in model II are larger than those of a standard binary perceptron. For $\kappa=0$, e.g., the simulations give $\alpha_c = 2.26 \pm 0.01$, which is bigger than the maximal capacity of the binary perceptron model [1] and the binary many-neuron interaction model [8], both of which have $\alpha_c = 2$. For model I the maximal capacity at $\kappa=0$ found by simulations is 1.78 ± 0.01.

V. CONCLUDING REMARKS

In this work we have calculated the maximal capacity per number of couplings for two colored perceptron models. Compared with the standard perceptron these models have two neuronal variables per site and a local field that contains higher order neuron terms. The method used is a generalization of the Gardner approach and both the RS and RSB1 results have been discussed. We expect that the latter give very close upperbounds for the exact values.

Extensive numerical simulations have been performed for finite systems and extrapolated to $N=\infty$. The adaptive Gardner algorithm and the Adatron algorithm give the best, but different results. Hence, the results of the simulations can be considered only as lower bounds for the exact maximal capacity. Additional work looking for improved algorithms would be welcome.

Comparing both the RSB1 results and the results from numerical simulations we conclude that they are in good agreement. For bigger values of κ they even completely coincide. For model I we find that at $\kappa=0$ the maximal capacity satisfies $1.78 \leqslant \alpha_c \leqslant 1.83$. This suggests that it is equal to the maximal capacity of the $Q=$four-Potts perceptron, i.e., $\alpha_c = 1.83$ (after appropriate rescaling of the latter [7]). This would parallel the situation for Hebb learning [13]. For model II we have for $\kappa=0$ that $2.26 \leqslant \alpha_c \leqslant 2.28$, which is larger than the maximal capacity of the standard binary perceptron. Furthermore, as anticipated, the maximal capacity of model II is larger than that of model I for all values of κ.

ACKNOWLEDGMENTS

The authors would like to thank M. Bouten and J. van Mourik for critical discussions.

APPENDIX A: TECHNICAL DETAILS FOR MODEL II

The function $\psi_{RSB1}(\kappa,t_1,t_2,q_0^{(1)},q_0^{(3)},M)$ in formula (22) reads

$$\psi_{RSB1}(\kappa,t_1,t_2,q_0^{(1)},q_0^{(3)},M)=\frac{1}{2c_1}e^{\varepsilon_3}\int_{-\infty}^{c1/c(u_1+\delta_3)}\mathrm{Ds}\left[1+\mathrm{erf}\left\{\sqrt{\frac{3r}{2c^2}}\left(\frac{x_3}{\sqrt{3r}}-\delta_3+\frac{c}{c_1}s\right)\right\}\right]+\frac{1}{2c_1}e^{\varepsilon_2}\int_{-\infty}^{c1/c(u_1-\delta_2)}$$
$$\times\mathrm{Ds}\left[1+\mathrm{erf}\left\{\sqrt{\frac{3r}{2c^2}}\left(-\frac{x_2}{\sqrt{3r}}+\delta_2+\frac{c}{c_1}s\right)\right\}\right]+\frac{1}{2c_2}e^{\phi_2}\int_{-\infty}^{-c2/c(u_1-\gamma_2)}\mathrm{Ds}\left[1+\mathrm{erf}\left\{\sqrt{\frac{3r}{2c^2}}\right.\right.$$
$$\left.\left.\times\left(\frac{x_2}{\sqrt{3r}}-\gamma_2+\frac{c}{c_2}s\right)\right\}\right]+\frac{1}{2c_2}e^{\phi_3}\int_{-\infty}^{-c2/c(u_1-\gamma_3)}\mathrm{Ds}\left[1+\mathrm{erf}\left\{\sqrt{\frac{3r}{2c^2}}\left(-\frac{x_3}{\sqrt{3r}}-\gamma_3\right.\right.\right.$$
$$\left.\left.\left.+\frac{c}{c_2}s\right)\right\}\right]+\frac{1}{2c'}e^{d_1}\int_{-\infty}^{-u_1/c'}\mathrm{Ds}\left[\mathrm{erf}\left\{\sqrt{\frac{3r}{2}}\left(\frac{x_3}{\sqrt{3r}}-b_1-\frac{1}{c'}s\right\}\right)+\mathrm{erf}\left\{\sqrt{\frac{3r}{2}}\left(-\frac{x_2}{\sqrt{3r}}-b_1\right.\right.\right.$$
$$\left.\left.\left.-\frac{1}{c'}s\right)\right\}\right]+\frac{1}{2}\int_{-\infty}^{u_1}\mathrm{Ds}\left[\mathrm{erf}\left\{\sqrt{\frac{3r}{2}}\left(\frac{x_2}{\sqrt{3r}}-s\right)\right\}+\mathrm{erf}\left\{-\sqrt{\frac{3r}{2}}\left(\frac{x_3}{\sqrt{3r}}+s\right)\right\}\right],$$

with Ds a Gaussian measure and

$$c=\sqrt{1+M},\quad c'=\sqrt{1+M_1},\quad c_1=\sqrt{1+M(1+3r)},$$

$$c_2=\sqrt{M_1c^2+c_1^2},\quad M_1=rM,\quad r=\frac{1-q_1'}{1-q_0^{(1)}},$$

$$x_2=\sqrt{3r}u_1-t_2\sqrt{\frac{q_0^{(1)}}{1-q_0^{(1)}}},\quad x_3=-\sqrt{3r}u_1-t_2\sqrt{\frac{q_0^{(1)}}{1-q_0^{(1)}}},$$

$$\varepsilon_2=-\frac{1}{2}\frac{Mx_2^2}{c_1^2},\quad \varepsilon_3=-\frac{1}{2}\frac{Mx_3^2}{c_1^2},\quad d_1=\frac{1}{2}u_1b_1,$$

$$\phi_2=-\frac{1}{2c_2^2}[Mx_2^2(c')^2+M_1u_1^2c_1^2-2MM_1\sqrt{3r}u_1x_2],$$

$$\phi_3=-\frac{1}{2c_2^2}[Mx_3^2(c')^2+M_1u_1^2c_1^2+2MM_1\sqrt{3r}u_1x_3],$$

$$\delta_2=\frac{\sqrt{3r}Mx_2}{c_1^2},\quad \delta_3=\frac{\sqrt{3r}Mx_3}{c_1^2},\quad b_1=-\frac{M_1u_1}{(c')^2},$$

$$\gamma_2=\frac{1}{c_2^2}[M_1u_1c^2+M\sqrt{3r}x_2],$$

$$\gamma_3=\frac{1}{c_2^2}[M_1u_1c^2-M\sqrt{3r}x_3],$$

$$u_1=-\frac{\sqrt{\frac{2}{3}}\kappa+\sqrt{q_1'}t_1}{\sqrt{1-q_1'}},\quad q_1'=\frac{1}{3}q_0^{(1)}+\frac{2}{3}q_0^{(3)},$$

$$\kappa=\kappa_\xi=\kappa_\eta.$$

APPENDIX B: FORMULA FOR MODEL I

For model I the calculations are very similar. Some resulting expressions, however, have a somewhat different structure. For completeness we write down these expressions here.

For the available space of couplings we get in the RS approximation [compare Eq. (15)]

$$v=\tfrac{3}{2}\alpha\int\prod_r \mathrm{D}[s_r(q)]\ln[\psi_{RS}(\kappa_\xi,\kappa_\nu,\kappa_{\xi\nu},s_1,s_2,s_3,q)]$$
$$-\tfrac{3}{2}\alpha\ln 4+\tfrac{3}{2}\left[\ln(1-q)+\frac{1}{1-q}+\ln 2\pi\right] \tag{B1}$$

with

$$\psi_{RS}(\kappa_\xi,\kappa_\nu,\kappa_{\xi\nu},s_1,s_2,s_3,q)$$
$$=\left(\int_{-\infty}^{l_1}du_1\int_{l_4}^{\infty}du_2\int_{l_5}^{\infty}du_3+\int_{-\infty}^{l_2}du_2\int_{l_6}^{\infty}du_1\int_{l_7}^{\infty}du_3\right.$$
$$+\int_{-\infty}^{l_3}du_3\int_{l_8}^{\infty}du_1\int_{l_9}^{\infty}du_2$$
$$\left.+\int_{l_1}^{\infty}du_1\int_{l_2}^{\infty}du_2\int_{l_3}^{\infty}du_3\right)\prod_r\frac{e^{-(1/2)u_r^2}}{\sqrt{2\pi}} \tag{B2}$$

where

$$l_i=\frac{L_i+s_i}{\sqrt{1-q}},\quad i=1,2,3,$$

$$l_4=\frac{L_1+L_2+s_1+s_2}{\sqrt{1-q}}-u_1,\ \ l_6=l_4+u_1-u_2,$$

$$l_5=\frac{L_1+L_3+s_1+s_3}{\sqrt{1-q}}-u_1,\ \ l_8=l_5+u_1-u_3,$$

$$l_7=\frac{L_2+L_3+s_2+s_3}{\sqrt{1-q}}-u_2,\ \ l_9=l_7+u_2-u_3,$$

$$L_1=\frac{1}{2}(\kappa_\xi-\kappa_\eta+\kappa_{\xi\eta}),\ \ L_2=\frac{1}{2}(-\kappa_\xi+\kappa_\eta+\kappa_{\xi\eta}),$$

$$L_3=\frac{1}{2}(\kappa_\xi+\kappa_\eta-\kappa_{\xi\eta})$$

and q taking those values that minimizes v. Thus, for $\kappa=\kappa_\xi=\kappa_\eta=\kappa_{\xi\eta}$ the maximal capacity in the RS approximation can be written as

$$\alpha_{RS}(\kappa)=\lim_{q\to 1}\left\{\frac{-\ln(1-q)-\frac{1}{1-q}-\ln 2\pi}{\int\prod_r \mathrm{D}(s_r(q))\,\psi_{RS}(\kappa,\kappa,\kappa,s_1,s_2,s_3,q)-\ln 4}\right\}.$$

For the RSB1 approximation with the form of the order parameters given by Eq. (21) the maximal capacity reads

$$\alpha_{RSB1}(\kappa)=\min_{q_0,M}\left\{\frac{-\ln(1+M)-\frac{q_0M}{(1+M)(1-q_0)}}{\int\prod_r \mathrm{D}t_r\ln\,\psi_{RSB1}(\kappa,t_1,t_2,t_3,q_0,M)}\right\}$$

with $\psi_{RSB1}(\kappa,t_1,t_2,t_3,q_0,M)$ a linear combination of thirty-four, mostly double, integrals over error functions. An interested reader can find a complete formula for $\psi_{RSB1}(\kappa,t_1,t_2,t_3,q_0,M)$ in [22].

Finally, the learning algorithms for model I differ in the way that the couplings $J^{(1)}$ and $J^{(2)}$ are updated. We have for the adaptive Gardner algorithm

$$J_i^{(1)}\to J_i^{(1)}+\xi_0^\mu\xi_i^\mu\tfrac{1}{2}[(\kappa_\xi-\lambda_\xi^\mu)\Theta(\kappa_\xi-\lambda_\xi^\mu)+(\kappa_{\xi\eta}-\lambda_{\xi\eta}^\mu)\Theta(\kappa_{\xi\eta}-\lambda_{\xi\eta}^\mu)],$$

$$J_i^{(2)}\to J_i^{(2)}+\eta_0^\mu\eta_i^\mu\tfrac{1}{2}[(\kappa_\eta-\lambda_\eta^\mu)\Theta(\kappa_\eta-\lambda_\eta^\mu)+(\kappa_{\xi\eta}-\lambda_{\xi\eta}^\mu)\Theta(\kappa_{\xi\eta}-\lambda_{\xi\eta}^\mu)]$$

instead of Eqs. (24) and (25) and for the Adatron algorithm we take

$$\delta x_1^\mu=\tfrac{1}{4}[\max\{-x_1^\mu-x_3^\mu,\gamma(1-n_1\lambda_\xi^\mu)\}+\max\{-x_1^\mu-x_2^\mu,\gamma(1-n_1\lambda_{\xi\eta}^\mu)\}],$$

$$\delta x_2^\mu=\tfrac{1}{4}(\max\{-x_2^\mu-x_3^\mu,\gamma(1-n_2\lambda_\eta^\mu)\}+\max\{-x_1^\mu-x_2^\mu,\gamma(1-n_2\lambda_{\xi\eta}^\mu)\}),$$

instead of Eqs. (28) and (29).

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

[2] J. Hertz, A. Krogh, and R.G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, 1991).

[3] B. Müller, J. Reinhardt, and M.T. Strickland, *Neural Networks: An Introduction* (Springer, Berlin, 1995).

[4] M. Opper and W. Kinzel, *Models of Neural Networks III*, edited by E. Domany, J.L. van Hemmen, and K. Schulten (Springer, New York, 1996), Vol. 151.
[5] *On-line Learning in Neural Networks*, edited by D. Saad (Cambridge University Press, Cambridge, 1998).
[6] J.P. Nadal and A. Rau, J. Phys. I **1**, 1109 (1991).
[7] F. Gerl and U. Krey, J. Phys. A **27**, 7353 (1994).
[8] D.A. Köhring, J. Phys. (France) **51**, 145 (1990).
[9] S. Mertens, H.M. Köhler, and S. Bös, J. Phys. A **24**, 4941 (1991).
[10] D. Bollé, P. Dupont, and J. van Mourik, Europhys. Lett. **15**, 893 (1991).
[11] D. Bollé, R. Kühn, and J. van Mourik, J. Phys. A **26**, 3149 (1993).
[12] D. Bollé and P. Kozłowski, J. Phys. A **31**, 6319 (1998).
[13] D. Bollé and P. Kozłowski, J. Phys. A **32**, 8577 (1999).
[14] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).
[15] W. Whyte and D. Sherrington, J. Phys. A **29**, 3063 (1996).
[16] D. Bollé and R. Erichsen, Jr., Phys. Rev. E **59**, 3386 (1999).
[17] W. Krauth and M. Mézard, J. Phys. A **20**, L745 (1987).
[18] L.F. Abbott and T.B. Kepler, J. Phys. A **22**, L711 (1989).
[19] J.K. Anlauf and M. Biehl, Europhys. Lett. **10**, 687 (1989).
[20] J. Imhoff, J. Phys. A **28**, 2173 (1995).
[21] M. Mézard, G. Parisi and M.A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[22] P. Kozłowski, Ph.D. thesis, Katholieke Universiteit Leuven, Belgium 2001.