Storage capacity and optimal learning of Potts-model perceptrons by a cavity method

1994 J. Phys. A: Math. Gen. 27 7353

(http://iopscience.iop.org/0305-4470/27/22/012)

J. Phys. A: Math. Gen. 27 (1994) 7353–7372. Printed in the UK

Storage capacity and optimal learning of Potts-model perceptrons by a cavity method*

F Gerl† and U Krey

Institut für Physik III, Universität Regensburg, Universitätsstrasse 31, D-93040, Germany

Received 24 August 1994, in final form 30 October 1994

Abstract. By means of a general formulation for the optimal learning capacity of perceptrons with multi-state neurons and real-valued couplings with spherical constraints, which we derive by a cavity method, we calculate the optimal learning capacity $\alpha_c(Q', \kappa) := p_{max}/[N(Q-1)]$ for perceptrons with a Q- resp. Q'-state Potts-model *input* resp. *output* neurons as a function of Q' and the stability parameter κ. Among other results, the asymptote for $Q' \to \infty$ is found, and it is shown that for $\kappa = 0$ the information gain per coupling, $\Delta I = (\alpha_c \ln Q')/(Q'-1)$, converges slowly to $\frac{1}{2}$ in this limit. Moreover, for $Q' \to \infty$ the same asymptotics also apply for the simple case of Hebbian learning.

1. Introduction

The learning capacity $\alpha_c(\kappa)$ of simple perceptrons with neurons of the Ising type, and with essentially unconstrained real couplings, has already been calculated exactly for general values of the stability parameter κ (see below) in 1987 in the seminal paper of Gardner [1]. In this paper, the so-called replica trick was used to calculate the expectation value of the function $\ln V$, where V is the relative phase-space volume of coupling vectors $J := (J_1, \ldots, J_N)$, which solve the given classification task under the constraint of a fixed norm ($J^2 = N$) in the limit $N \to \infty$ for constant $\alpha := p/N$, where p is the number of patterns. The replica method, which is borrowed from statistical mechanics, is already well known (see, for example, [2–4]). However, for systems with more complicated neural degrees of freedom, e.g. Q-state clock neurons (see, for example, [5, 6] and references therein), or for Q-state Potts neurons [7, 8] the formulation and evaluation of the Gardner formalism already becomes a rather complex task. In fact, concerning the behaviour of the Q-state Potts-model perceptron in the limit of large Q, the results obtained up to now are partially contradictory [7, 8], see below.

On the other hand, in [5, 6] we have already found a general formulation both for the evaluation of α_c, and also for the AdaTron training algorithms, by which the optimal learning capacity of the system can be implemented in a reasonable time.

These formulations do not cover just the usual case of perceptrons with Ising neurons, but also the clock-model case, and lend themselves (as already stated in [5]) to a natural generalization for more complex neural architectures. In the present paper, we formulate this approach, which is based on a cavity method employing explicitly the Kuhn–Tucker conditions, for perceptrons with Q-state Potts neurons, i.e. we evaluate the learning capacity

* Based on the PhD thesis of F Gerl, Regensburg 1994.

† Present address: Institut für Theoretische Physik der Universität Göttingen, Bunsenstrasse 9, D-37073 Göttingen.

0305-4470/94/227353+20$19.50

$\alpha_c(Q, \kappa)$ of such systems and give an explicit prescription for the corresponding AdaTron algorithm. The result for α_c can be expressed by multi-dimensional integrals, which we have evaluated directly for $Q \leqslant 10$, and indirectly, through the implementation of a certain Gaussian process, for larger values. Actually, we have performed the evaluation up to extremely large Q-values, e.g. for values much larger than $Q = 10^5$.

In this way, we obtain a large number of precise results, including the asymptotic behaviour for $Q \to \infty$, which is approached extremely slowly. Among other consequences, these results include the resolution of the above-mentioned controversial results; i.e. our results show that the information gain per coupling in the limit of $Q \to \infty$ remains finite, namely $= \frac{1}{2}$, whereas by former authors it was predicted to diverge [7] or vanish [8] respectively.

As already noted by [8], the value of Q mentioned above is that of the output neuron, which henceforth will be called Q', whereas for the random patterns considered, the Q-value of the input neuron does not matter (see below).

This paper is organized as follows. In section 2 we describe the model. In section 3 we formulate an optimization problem with Kuhn–Tucker conditions leading to maximal stability and derive from it our AdaTron training algorithm. In section 4 we describe our cavity method for the derivation of α_c and present results for $\kappa = 0$. Additionally, we show in this section that, for the present model, Hebbian learning leads to the same information gain in the limit $Q' \to \infty$ as AdaTron learning with $\kappa = 0$. In section 5 we discuss the optimal storage for $\kappa > 0$ and present two general equations for the evaluation of α_c for positive κ. Finally, section 6 presents the conclusions.

2. Model definitions and some known results

We consider feedforward networks with N input neurons, namely Q-state Potts neurons (see below), and with N' output neurons, which are Q'-state Potts neurons. In the following, for simplicity, $N' = 1$ is assumed, unless otherwise stated. However, usually we do not assume $Q' = Q$.

Thus input neuron number j can be found in one of Q different states $\sigma_j \in \{1, \ldots, Q\}$, which are described by the vectors $\boldsymbol{m}_{\sigma_j} = (m_{\sigma_j}(1), \ldots, m_{\sigma_j}(Q))$ spanning an equilateral triangle for $Q = 3$, a regular tetrahedron for $Q = 4$, etc, with

$$m_{\sigma_j}(s) := Q\delta_{\sigma_j,s} - 1 . \tag{1}$$

This implies

$$\boldsymbol{m}_{\sigma_i} \cdot \boldsymbol{m}_{\sigma_j} := \sum_s m_{\sigma_i}(s) m_{\sigma_j}(s) = Q(Q\delta_{\sigma_i,\sigma_j} - 1) . \tag{2}$$

For a simple perceptron, the input state generates a 'presynaptic output field' $\boldsymbol{h} = (h(1), \ldots, h(Q'))$ with

$$h(s') := \sum_{j=1}^{N} \sum_{s=1}^{Q} J_j(s', s)\, m_{\sigma_j}(s) . \tag{3}$$

Here, the synaptic matrix $J_j(s', s)$ determines the input–output coupling. From the presynaptic output field $\boldsymbol{h}$, the state σ' of the Q'-state Potts-model output neuron is determined as that value σ' which maximizes the overlap $\boldsymbol{m}_{\sigma'}^T \boldsymbol{h} := \sum_{s'=1}^{Q'} m_{\sigma'}(s') h(s')$. In the thermodynamic limit $N \to \infty$, σ' is uniquely defined with probability 1.

As is well known (see [8]) the synaptic matrix $J_j(s', s)$ can be constrained to fulfill the following gauge conditions:

$$\sum_{s=1}^{Q} J_j(s', s) = 0 \qquad \forall s' \tag{4}$$

$$\sum_{s'=1}^{Q'} J_j(s', s) = 0 \qquad \forall s . \tag{5}$$

Thus, h simplifies to

$$h(s') = Q \sum_{j,s} J_j(s', s)\, \delta_{\sigma_j, s} . \tag{6}$$

To a given set of p random input states $\mu = 1, \ldots, p$, with $\sigma_j^\mu = n_j^\mu$, with $j = 1 \ldots N$, $\mu = 1 \ldots p$ and $n_j^\mu \in \{1, \ldots, Q\}$, one now assigns random 'desired outputs' $\sigma'^\mu = n'^\mu$ ranging between 1 and Q'. The learning task is then to find synaptic couplings for which the actual output of the system is the desired one.

A simple prescription fulfilling this task under certain conditions is the so-called Hebbian rule [7]

$$J_j(s', s) = J_j^{\text{Hebb}}(s', s) := \frac{1}{N\, Q(Q-1)} \sum_{\mu} m_{n'^\mu}(s') m_{n_j^\mu}(s) . \tag{7}$$

In the following we try to find the maximum number $p_{\max}$ of independently and randomly chosen input–output pairs (n^μ, n'^μ) which can be learned by our system for given values of Q and Q' by a suitable choice of the couplings. Following [8], we define the critical loading parameter

$$\alpha_c := \frac{p_{\max}}{N\,(Q-1)} . \tag{8}$$

As we will see below, under the randomness assumptions made above, α_c depends on Q' but *not* on Q.

For an auto-associative system with symmetric Hebbian couplings (i.e. $Q = Q'$ and $N = N'$) and random sequential updating, Kanter [7] has analysed the retrieval dynamics along the lines of Amit *et al* [9], and from the results obtained with the replica method for $Q = 3, 4, 5$ and 9 he estimated

$$\alpha_c \simeq 0.138\, \frac{Q}{2} . \tag{9}$$

However, Kanter could not determine whether at α_c the retrieval quality R, i.e. the overlap of the stable final state $\sigma^{\mu *}$ with the corresponding input n^μ,

$$R := \frac{1}{N\, Q(Q-1)} \sum_{j,s} m_{\sigma_j^{\mu *}}(s) m_{n_j^\mu}(s) \tag{10}$$

vanishes for $Q \to \infty$, or whether it remains finite.

For given R, the probability P_0 that the final state of an output neuron agrees with the desired value is $P_0 := [R(Q-1)+1]/Q$. Then, the information gain per coupling element ΔI, relative to $R = 0$ (i.e. $P_0 = 1/Q$), at α_c is (see [9])

$$\Delta I = \frac{\alpha_c(Q)}{Q-1} \left(\ln Q + P_0 \ln P_0 + (1 - P_0) \ln \left(\frac{1 - P_0}{Q-1} \right) \right) . \tag{11}$$

Thus, ΔI would diverge $\propto \ln Q$ with $Q \to \infty$ if, in this limit, the quantity R remained finite, whereas ΔI would remain finite or vanish if, in this limit, R vanished $\propto 1/\ln Q$ or faster, respectively.

On the other hand, Shim *et al* [10] considered layered feedforward Potts-model neural networks with Hebbian couplings between successive layers and found for $Q \gg 1$ that

$$\alpha_c(Q) \sim Q^{0.85}. \tag{12}$$

This would imply that the information gain ΔI of (11) vanishes for $Q \to \infty$ as $Q^{-0.15} \ln Q$. Although these two suggestions for the learning capacity, (9) and (12), should not be compared directly, (12) may be taken as a hint that a divergence of the information gain ΔI for $Q \to \infty$ should not be expected. We will comment on these results later.

However, our main interest is not Hebbian coupling, but the Potts perceptron with couplings leading to optimal stability. Nadal and Rau [8] have calculated its learning capacity, using separate spherical constraints for $s' = 1, \ldots, Q$:

$$\sum_{j,s} (J_j(s', s))^2 = N\gamma \tag{13}$$

where γ is an arbitrary constant. They predicted that $\alpha_c(Q')$ should vary slowly with Q' and should remain bounded from above, which would lead for $Q' \to \infty$ to an information gain vanishing as

$$\Delta I(Q') = \alpha_c(Q') \frac{\ln Q'}{Q' - 1}. \tag{14}$$

All these results look questionable, as has already been noted by [11], e.g. naively one would guess that for any given Q' the (bounded) optimal learning capacity calculated by Nadal and Rau should be larger than that calculated by Shim *et al* for Hebbian feedforward nets (see equation (12)) which again should be larger than that calculated by Kanter for the fully coupled auto-associative Hebbian system (9). This naive expectation is not fulfilled by the just-mentioned results.

In the following sections we derive a cavity approach by which we calculate $\alpha_c(Q', \kappa)$ exactly, including the symptotics for $Q' \to \infty$.

3. Couplings of maximal stability

3.1. The optimization problem

To simplify our formalism we use, in the following, the symbols J_j as an abbreviation for the $Q' \times Q$-matrices $J_j(s', s)$, and also the symbols h^μ and x^μ for the Q'-vectors with components $h^\mu(s')$ and $x^\mu(s')$, respectively. Here $x^\mu(s')$ play the role of 'embedding strengths' (see below).

In the following, summation with respect to $s = 1, \ldots, Q$ will usually be abbreviated as a dot product as in (2), whereas for Q'-vectors, i.e. with respect to $s' = 1, \ldots, Q'$, we will write the scalar product as $x^T y$. Furthermore, for Q'-vectors we continue the components cyclically by assuming $x(Q' + n) = x(n)$ for any integer n. Thus, we can define the shift operator $\mathcal{P}$, which shifts the component cyclically by one unit to the right:

$$\mathcal{P}: \quad (x(1), \ldots, x(Q')) \longrightarrow (x(Q'), x(1), \ldots, x(Q' - 1)) \tag{15}$$

and its natural extensions $\mathcal{P}^n$ and $\mathcal{P}^{-n}$, where n is any integer. Finally, we use the abbreviations $\vec{x} = (x^1, x^2, \ldots, x^p)$ and $\underline{J} = (J_1, J_2, \ldots, J_N)$. Thus we can abbreviate

(3) as

$$h^\mu := \sum_j J_j \cdot m_{n_j^\mu} . \tag{16}$$

With the shift operator, we can define the *re-oriented field* E^μ of a pattern:

$$E^\mu := \mathcal{P}^{1-n'^\mu} h^\mu = \mathcal{P}^{1-n'^\mu} \sum_j J_j \cdot m_{n_j^\mu} . \tag{17}$$

Here, for simplicity, the first component $s' = 1$ has been distinguished: a pattern is stored if this component of the re-oriented field is the largest one.

The *stability* κ of a pattern set is defined to be similar, as in [8] to

$$\kappa := c/L \equiv \min_{\mu, s'(>1)} \{E^\mu(1) - E^\mu(s')\}/L \tag{18}$$

with

$$L^2 = |\underline{J}|^2 = \sum_{j,s,s'} J_j(s', s)^2 =: \operatorname{tr}_{s,s'} \Big[\sum_j J_j^T J_j \Big] . \tag{19}$$

This definition of the stability is equivalent to using a global spherical constraint.

To maximize the stability, we make the *ansatz*

$$J_j = \frac{1}{N\, Q(Q-1)} \sum_\mu (\mathcal{P}^{n'^\mu - 1} x^\mu) \otimes m_{n_j^\mu} \tag{20}$$

where $\otimes$ means a dyadic $Q' \times Q$ product, i.e.

$$J_j(s', s) = \frac{1}{N\, Q(Q-1)} \sum_\mu x^\mu(n'^\mu - 1 + s')\, m_{n_j^\mu}(s) . \tag{21}$$

This ansatz automatically fulfills the gauge (4). The second gauge condition (5) is also fulfilled if

$$\sum_{s'} x^\mu(s') = 0 \qquad \forall \mu . \tag{22}$$

Here, the quantities $m_{n_j^\mu}(s)$ and $x^\mu(s')$ play the same role as the product $\xi^\mu \xi_j^\mu$ and the embedding strengths x^μ in the formula $J_j = N^{-1} \sum_\mu x^\mu \xi^\mu \xi_j^\mu$ for the perceptron with Ising neurons and input–output pairs $\{(\xi_1^\mu, \ldots, \xi_N^\mu), \xi^\mu)\}$, see [12].

According to (20), the coupling matrix J_j is a linear combination of contributions from different patterns with the $x^\mu(s')$ as cofficients, i.e. it belongs to a subspace spanned by the patterns. Every contribution orthogonal to this subspace would leave the oriented presynaptic output field h^μ unchanged, as can be seen from (16), whereas according to (19) the length L of the coupling vector J_j would be enhanced and thus the stability reduced according to (18).

With a similar argument, one can show that the gauge condition (22) will also be automatically fulfilled (see below). This implies that the ansatz (20) is no restriction, and at the same time it shows that the calculation of the capacity $\alpha_c(Q', \kappa)$ can be formulated as an optimization task, where L has to be minimized, see below. Also the Hebb rule (7), see [7], can be obtained from the ansatz (20) with $x^\mu(s') = Q'\delta_{1,s'} - 1$. Additionally, Nadal and Rau [8] stress that a perceptron algorithm, equation (31) in [8] which allows us to generate a solution, whenever there is at least one, can be used with any gauge, in particular with (4) and (5). The resulting couplings will be of the form (20).

The task of finding the optimal perceptron can now be formulated as follows:

Find x^μ such that $\min_{\mu,s'(>1)}\{E^\mu(1) - E^\mu(s')\} \geqslant \kappa$ with maximal κ for given random input–output pairs $\{n^\mu, n'^\mu\}$, with $\mu = 1, \ldots, p$, while $|\underline{J}|^2 = 1$.

With (18), i.e. $\kappa = c/L$, this is equivalent to:

Minimize $L^2 := |\underline{J}|^2$ for $\min_{\mu,s'(>1)}\{E^\mu(1) - E^\mu(s')\} \geqslant c$.

For the norm L of the couplings, we have, with (19),

$$|\underline{J}|^2 = \frac{1}{(N\,Q(Q-1))^2}\mathrm{tr}_{s,s'}\Big[\sum_j \Big(\sum_\mu \mathcal{P}^{n'^\mu - 1} x^\mu \otimes m_{n_j^\mu}\Big)^T \Big(\sum_\nu \mathcal{P}^{n'^\nu - 1} x^\nu \otimes m_{n_j^\nu}\Big)\Big]$$

$$= \frac{1}{(N\,Q(Q-1))^2}\sum_{\mu,\nu}\Big(\sum_{s'} x^\mu(s')\, x^\nu(n'^\nu - n'^\mu + s')\Big)\Big(\sum_{j,s} m_{n_j^\mu}(s) m_{n_j^\nu}(s)\Big)$$

$$\geqslant 0\,. \tag{23}$$

Analogously to the model with Ising spins [12], we define the pattern correlation $p \times p$-matrix $\overleftrightarrow{C}$, and the oriented correlation matrix operator $\overleftrightarrow{B}$,

$$C^{\mu\nu} := \frac{1}{N\,Q(Q-1)}\sum_{j,s} m_{n_j^\mu}(s)\, m_{n_j^\nu}(s) = \frac{1}{N(Q-1)}\Big(Q\sum_j \delta_{n_j^\mu, n_j^\nu} - N\Big)$$

$$B^{\mu\nu} := C^{\mu\nu}\,\mathcal{P}^{n'^\nu - n'^\mu} = \frac{1}{N\,Q(Q-1)}\sum_j m_{n_j^\mu}\cdot m_{n_j^\nu}\,\mathcal{P}^{n'^\nu - n'^\mu}\,. \tag{24}$$

The diagonal elements of $\overleftrightarrow{C}$ are equal to 1, whereas for random patterns and $N \gg 1$ the non-diagonal elements are Gaussian random numbers with average 0 and variance $1/(N(Q-1))$. Now, $L = |\underline{J}|$ simplifies from (23) and (24) to the compact result

$$L^2 = \frac{1}{N\,Q(Q-1)}\sum_{\mu\nu} x^{\mu\,T} B^{\mu\nu} x^\nu =: \frac{1}{N\,Q(Q-1)}\,\vec{x}^T\,\overleftrightarrow{B}\,\vec{x}\,. \tag{25}$$

So, the final formulation of the optimization task with embedding strengths fullfilling the gauge condition (22) is

Minimize $f(\vec{x}) = \sum_{\mu,\nu} x^{\mu\,T} B^{\mu\nu} x^\nu$ under the constraints

$$E^\mu(1) - E^\mu(s') \geqslant c \qquad \forall\mu,\ \forall s' > 1\,. \tag{26}$$

Also the oriented field E^μ can simply be written as

$$E^\mu = \mathcal{P}^{1-n'^\mu}\sum_j J_j \cdot m_{n_j^\mu}$$

$$= \frac{1}{N\,Q(Q-1)}\mathcal{P}^{1-n'^\mu}\sum_{j,s,\nu}\big(\mathcal{P}^{n'^\nu - 1}x^\nu\big) m_{n_j^\nu}\cdot m_{n_j^\mu}$$

$$= \sum_\nu B^{\mu\nu}x^\nu = x^\mu + \sum_{\nu(\neq\mu)} B^{\mu\nu}x^\nu \tag{27}$$

or in compact form

$$\vec{E} = \overleftrightarrow{B}\,\vec{x}\,. \tag{28}$$

3.2. Kuhn–Tucker conditions and the AdaTron algorithm

Thus, with (26) we have a quadratic optimization problem with $p(Q'-1)$ linear constraints, to which one can assign the following Lagrangian:

$$\mathcal{L}(\vec{x},\vec{\lambda}) = \tfrac{1}{2}\,\vec{x}^T \overleftrightarrow{B}\,\vec{x} - \sum_{\mu=1}^{p}\sum_{s'=2}^{Q'} \lambda^\mu(s')(E^\mu(1) - E^\mu(s') - c) \tag{29}$$

with $p(Q'-1)$ Lagrange multipliers $\{\lambda^\mu(s')|\,\mu = 1,\ldots,p;\ s' = 2,\ldots,Q'\}$ for the constraints (26).

Now we can apply the theorem of Kuhn and Tucker [13]: let $\vec{x}^*$ be a local minimum of (26), then multipliers $\vec{\lambda}^*$ exist, with

$$\vec{\nabla}_x \mathcal{L}(\vec{x}^*,\vec{\lambda}^*) = \vec{0} \tag{30}$$

$$E^{*\mu}(1) - E^{*\mu}(s') \geqslant c \qquad \forall\mu,\ \forall s' > 1 \tag{31}$$

$$\lambda^{*\mu}(s') \geqslant 0 \qquad \text{if} \quad E^{*\mu}(1) - E^{*\mu}(s') = c \tag{32}$$

$$\lambda^{*\mu}(s')[E^{*\mu}(1) - E^{*\mu}(s') - c] = 0 \qquad \forall\mu, \forall s' > 1\,. \tag{33}$$

Thus, with (28), the Lagrangian takes the compact form

$$\mathcal{L}(\vec{x},\vec{\lambda}) = \tfrac{1}{2}\,\vec{x}^T \overleftrightarrow{B}\,\vec{x} + \tilde{\vec{\lambda}}^T \overleftrightarrow{B}\,\vec{x} + c\vec{1}^T\vec{\lambda} \tag{34}$$

where we have used the abbreviations $\tilde{\lambda}^\mu(1) := -\sum_{s'=2}^{Q'}\lambda^\mu(s')$, $\tilde{\lambda}^\mu(s') = \lambda^\mu(s')$ for $s' = 2,\ldots,Q'$, and $\vec{1}^T\vec{\lambda} := \sum_\mu \sum_{s'=2}^{Q'}\lambda^\mu(s')$. Thus, minimizing (34) with respect to $\vec{x}$ one finds in case of an invertible $\overleftrightarrow{B}$ the unique solution

$$\begin{aligned} x^{*\mu}(1) &= \lambda^{*\mu}(2) + \lambda^{*\mu}(3) + \cdots \\ x^{*\mu}(2) &= -\lambda^{*\mu}(2) \\ x^{*\mu}(3) &= -\lambda^{*\mu}(3) \\ &\vdots \end{aligned} \tag{35}$$

The gauge condition (22) is thus automatically fulfilled. Furthermore, for the recognition direction, $s' = 1$, the embedding strengths are $\geqslant 0$, whereas for $s' > 1$ they are $\leqslant 0$.

If $\overleftrightarrow{B}$ is not invertible, the solution (35) of (34) also applies, since one can prove that any other KT-point, i.e. a solution of the Kuhn–Tucker equations, leads to the same couplings. For networks with Ising neurons this has already been shown in [12].

Now we can already define our AdaTron algorithm, which generalizes that defined in [12] for the perceptron with Ising neurons. In the $(Q'-1)$-dimensional hyperplane through $\mathbf{0}$, which is orthogonal to the vector $(1,\ldots,1)$ in Q'-space, the stability region is a segment spanned by $Q'-1$ hyperplanes, as sketched in figure 1 for $Q' = 3$.

The vector $v_{s'}$, with

$$\begin{aligned} v_{s'} &:= \frac{1}{Q'}(m_1 - m_{s'}) \\ &= \frac{1}{Q'}\,[(Q'-1,-1,\ldots,-1) - (-1,-1,\ldots,Q'-1,-1,\ldots,-1)] \\ &= (1,0,\ldots,0,-1,0,\ldots,0) \end{aligned} \tag{36}$$

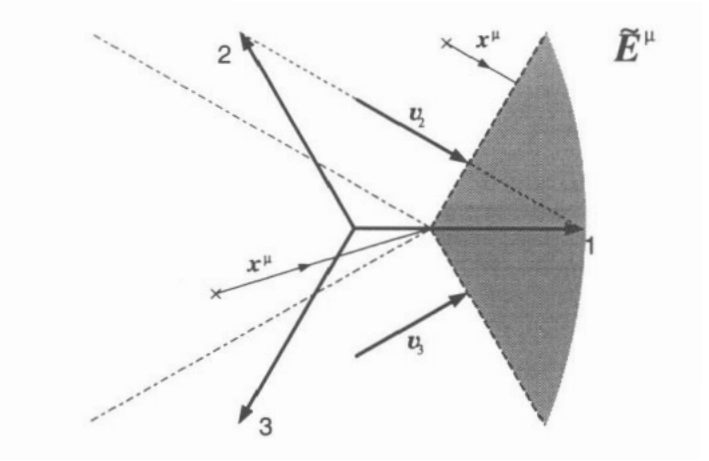

Figure 1. This figure visualizes for $Q' = 3$ the results from our optimization theory, which lead to the construction of our AdaTron algorithm for Potts perceptrons. Everything happens in the $(1, 1, 1)$-plane through the origin. The embedding strengths x^μ are linear combinations of the learning directions $v_{s'}$, the Lagrangian multiplicators $\lambda^\mu(s')$ are the coefficients; see (36) and (37). The x^μ counteract the noise field $\widetilde{E}^\mu$ (equation (38)) which is represented for two cases by the crosses (×). The vectors **1**, **2**, and **3** represent $m_1 := (2, -1, -1)$, $m_2 := (-1, 2, -1)$ and $m_3 := (-1, -1, 2)$. The x^μ are the distance vectors from the fields (×) to the hatched stability region. The distance vector of this region from the origin is, according to (18), given by $c\, m^1/Q'$, where c is related to to the stability κ and the quantity $L = |\vec{J}|$ by $\kappa = c/L$. For $\kappa > 0$ it is convenient to assume $c = 1$ (which has been used in the figure), whence $\kappa = 1/L$, whereas for $\kappa = 0$, L can be fixed to an arbitrary constant, while $c = 0$.

is orthogonal to the hyperplane separating the attraction regions of state s' and state 1. Comparison with (35) shows that the $\lambda^\mu(i)$, with $i = 2, \ldots, Q'$, are the coefficients belonging to these learning directions:

$$x^\mu = \sum_{s'=2}^{Q'-1} \lambda^\mu(s')\, v_{s'} . \tag{37}$$

The weights x^μ obviously compensate the noise $\widetilde{E}^\mu$ originating from the other patterns, (compare (27)):

$$\widetilde{E}^\mu := \sum_{\nu(\neq\mu)} B^{\mu\nu} x^\nu \Longrightarrow E^\mu = x^\mu + \widetilde{E}^\mu . \tag{38}$$

If for a certain pattern μ exactly one $\lambda^\mu(s')$ is positive, then x^μ points to the corresponding hyperplane. If two or more constraints are active, i.e. the corresponding λ-coefficients are positive, then x^μ is the shortest vector pointing from $\widetilde{E}^\mu$ to the intersecting manifold. Thus x^μ is simply the distance vector from $\widetilde{E}^\mu$ to the stability region.

Although the learning directions $v_{s'}$ are not orthogonal, one can give a simple prescription for the construction of the embedding strengths x^μ. To this end, an additional oriented field $\widehat{E}^\mu$ is defined as

$$\widehat{E}^\mu := \widetilde{E}^\mu - (c, 0, \ldots, 0) . \tag{39}$$

Then the AdaTron algorithm for Potts model perceptrons is the following.

- **The components of $\widehat{E}^\mu$ should be brought into descending order, i.e. a vector ϱ should** be determined, such that $\widehat{E}^\mu(\varrho(1))$ is largest, $\widehat{E}^\mu(\varrho(2))$ second-largest, etc.

- If $\varrho(1) = 1$, then $\widehat{E}^{\mu}(1)$ is already the largest component of $\widehat{E}^{\mu}$. So pattern μ is stored automatically, and we are ready with this pattern. Otherwise, the smallest integer $k > 1$ has to be found such that

$$a := \frac{1}{k+1}\left(\widehat{E}^{\mu}(1) + \sum_{s'=1}^{k} \widehat{E}^{\mu}(\varrho(s'))\right) \geqslant \widehat{E}^{\mu}(\varrho(s')) \qquad \forall s' > k\,. \tag{40}$$

  That is, the average calculated out of $\widehat{E}^{\mu}(1)$ and the k largest components $\widehat{E}^{\mu}(\varrho(s'))$ should be larger than every remaining component.
- These first k elements correspond to the explicitly learned 'active directions', i.e. one uses

$$\lambda^{\mu}(\varrho(s')) = \widehat{E}^{\mu}(\varrho(s')) - a \qquad 1 \leqslant s' \leqslant k\,. \tag{41}$$

- The new embedding strengths are now

$$x^{\mu}_{\text{new}} = \sum_{s'=1}^{k} \lambda^{\mu}(\varrho(s'))\, v_{\varrho(s')} \tag{42}$$

  i.e.

$$x^{\mu}_{\text{new}}(\varrho(s')) = a - \widehat{E}^{\mu}(\varrho(s')) \qquad 1 \leqslant s' \leqslant k \tag{43}$$

$$x^{\mu}_{\text{new}}(1) = a - \widehat{E}^{\mu}(1) \tag{44}$$

$$x^{\mu}_{\text{new}}(s') = 0 \qquad \text{otherwise}\,. \tag{45}$$

- This algorithm is iterated successively for all μ, and the whole procedure is repeated, until the changes $\delta x^{\mu} := x^{\mu}_{\text{new}} - x^{\mu}_{\text{old}}$ are sufficiently small. The procedure can be speeded up by using an *over-relaxation* factor ω with $0 < \omega < 2$, where for $\alpha > 0$ one should take $\omega > 1$, with $\omega \to 2$ for $\alpha \to \alpha_c$. This is for sequential learning of the patterns, whereas with parallel learning ω must be chosen sufficiently small.

By construction, a KT-point is a fixed-point of this algorithm and *vice versa*.

For perceptrons with Ising neurons Anlauf and Biehl [12] proved that the AdaTron algorithm converges exponentially fast as long as $\alpha < \alpha_c$, and the characteristic convergence time has been shown to diverge $\propto (\alpha_c - \alpha)^{-2}$ for $\alpha \to \alpha_c$, as long as $\alpha < \alpha_c$. In our case, we do not have a proof, but numerically we have obtained the same result. Moreover, it is clear that our formalism applies to all perceptrons with neurons having well defined convex stability regions.

4. Optimal learning capacity by a cavity method

In the following, we derive the optimal learning capacity by a cavity method; our approach is inspired by ideas expressed in a review by Kinzel and Opper [14]. It is also related to the cavity approach for spin glasses as described in the book of Mézard *et al* [15], although we do not use the additional spin (= neuron) of those authors, which has also been used by Mézard in a paper for perceptrons with Ising spins and Ising couplings [16]. As in a paper by Griniasty [17] we start by assuming one simple ground state. Our approach, which leads to these new results, is based on the Kuhn–Tucker conditions described in the preceding section, and in contrast to [17] it also calculates explicitly the response of the system to newly added patterns. More details and extensions can be found in [18].

4.1. The response on a new pattern

We add a new pattern, $\mu = 0$. The noise field acting on this pattern, resulting from these patterns which are already implemented explicitly, is according to (38) given by

$$\widetilde{E}^0 = \sum_{\mu>0} B^{0\mu} x^\mu . \tag{46}$$

For random patterns $B^{0\mu}$ and x^μ are uncorrelated, therefore $\widetilde{E}^0$ is a random vector with length

$$\begin{aligned}\langle|\widetilde{E}^0|^2\rangle &= \Big\langle \Big(\sum_{\mu>0} B^{0\mu} x^\mu\Big)^T \Big(\sum_{\nu>0} B^{0\nu} x^\nu\Big)\Big\rangle \\ &= \frac{1}{N^2\, Q^2 (Q-1)^2} \sum_{\mu\nu} x^{\mu T} \mathcal{P}^{n'^\nu - n'^\mu} x^\nu \Big\langle \sum_{j,k} m_{n_j^\mu} \cdot m_{n_j^0} m_{n_k^0} \cdot m_{n_k^\nu} \Big\rangle . \end{aligned} \tag{47}$$

Here, the definition of $\overleftrightarrow{B}$ in (24) was used. The last bracket in (47) gives a non-vanishing result only for $j = k$, and can be simplified to $Q\, m_{n_j^\mu} \cdot m_{n_j^\nu} [= Q^2(Q\delta_{n_j^\mu, n_j^\nu} - 1)]$ (see equation (2)).

Thus, from (47), and with (24), (25) and (19), we get

$$\begin{aligned}\langle|\widetilde{E}^0|^2\rangle &= \frac{1}{N^2\, Q(Q-1)^2} \sum_{\mu\nu} x^{\mu T} \mathcal{P}^{n'^\nu - n'^\mu} x^\nu \sum_j m_{n_j^\mu} \cdot m_{n_j^\nu} \\ &= \frac{1}{N\,(Q-1)} \sum_{\mu\nu} x^{\mu T} B^{\mu\nu} x^\nu = Q|\underline{J}|^2 = Q\, L^2 . \end{aligned} \tag{48}$$

The vector $\widetilde{E}^0$ lies in the $(Q'-1)$-dimensional hyperplane spanned by the Potts vectors and is orthogonal to $(1, \ldots, 1)$. In an orthonormal basis lying in the hyperplane, from (48), the $Q'-1$ components are each Gaussian random numbers with average 0 and variance $L^2 Q/(Q'-1)$. Adding to $\widetilde{E}^0$ an additional random component with the same properties in the normal direction leads to a spherical Gaussian in $\mathbb{R}^{Q'}$. Therefore, it is practical to generate $\widetilde{E}^0$ by random numbers from a spherically symmetric Gaussian distribution with variance $L^2\, Q/(Q'-1)$ for all Q' components in $\mathbb{R}^{Q'}$, subtracting the component parallel to $(1, \ldots, 1)$ afterwards. Actually, as stated in the preceding section, for the construction of the weights x^μ through the AdaTron algorithm this component is arbitrary. Therefore, we often omit this subtraction.

Now, the added pattern is either already stored incidentally without implementation, or it must be stored *by force*, i.e. with explicit embedding. The first-mentioned case happens for $\kappa \to 0$ only with probability $1/Q'$. So with probability $1 - (1/Q')$, and for positive κ with even higher probability, one has to ensure storing by a weight x^0, and eventually also the other embedding strengths x^μ have to be changed. Then, instead of $\widetilde{E}^0$, from (46) one gets E^0, with

$$E^0 = \sum_{\mu=0}^{P} B^{0\mu} \check{x}^\mu = \widetilde{E}^0 + x^0 + G x^0 . \tag{49}$$

Here we have $\check{x}^0 = x^0$, whereas for $\mu > 0$ it is $\check{x}^\mu = x^\mu + \delta x^\mu$, where $\delta x^\mu = \mathcal{O}(1/\sqrt{N})$ is a reaction of the already stored patterns on the implementation of the new pattern in the

couplings, i.e. on x^0. This reaction couples back to pattern 0, leading to an enhancement of the necessary implementation strength through the term

$$Gx^0 := \sum_{\mu>0} G^\mu x^0 := \sum_{\mu>0} B^{0\mu} \left(\frac{\delta x^\mu}{\delta x^0}\right) x^0 . \tag{50}$$

Thus, $(\delta x^\mu/\delta x^0)$ is the *Jacobian matrix* describing the necessary response of the implementation strengths x^μ on the introduction of x^0: this response is necessary to keep the Kuhn–Tucker conditions fulfilled. In the *mean-field limit* we assume that G is self-averaging, and it turns out that $Gx^0 = gx^0$, with a negative real number g, see below.

A small perturbation y^μ of a stored pattern μ necessitates a correction of the embedding strength x^μ. With (43) one obtains for the k^μ directions, which have been learned explicitly,

$$\delta x^\mu(\varrho^\mu(s')) = -y(\varrho^\mu(s')) + \frac{1}{k^\mu+1}\left[y(1) + \sum_{s'=1}^{k^\mu} y(\varrho^\mu(s'))\right] . \tag{51}$$

The perturbation of pattern μ and its contribution to the response of the system arise through the random correlation of patterns μ and 0,

$$y^\mu = B^{\mu 0}x^0 \qquad \Longrightarrow \qquad \text{response:} \quad G^\mu x^0 = B^{0\mu}\delta x^\mu . \tag{52}$$

With the definition (24) of $\overleftrightarrow{B}$ and with $x^\mu(Q'+n) = x^\mu(n)$ we get

$$(G^\mu x^0)(s) = C^{0\mu 2}\left(-x^0(s) + \frac{1}{k^\mu+1}\left[x^0(n'^\mu - n'^0) + \sum_{s'=1}^{k^\mu} x^0(n'^\mu - n'^0 + \varrho^\mu(s'))\right]\right)$$
$$\times \sum_{s'=1}^{k^\mu} \delta_{s,n'^\mu - n'^0 + \varrho^\mu(s')} . \tag{53}$$

At this point the following remarks on the averaging process are in order. The embedding strengths are usually of the order $x^\mu = \mathcal{O}(1)$. Since they compensate the influence of all other patterns, one expects $\langle B^{\mu\nu}x^\nu\rangle = \mathcal{O}(1/N)$ for $\nu \neq \mu$. On the other hand, $B^{\mu\nu}x^\nu$ itself should be $= \mathcal{O}(1/\sqrt{N})$. Since we only calculate to order $\mathcal{O}(1/\sqrt{N})$, the x^μ and $B^{\mu\nu}$ can thus be treated as *on average uncorrelated*. Then, in the thermodynamic limit, for our averages we do not make a mistake by assuming that all these expressions can be averaged separately.

We now average over all values of n'^μ and all equally probable combinations of ϱ^μ. At first, only patterns with $k^\mu = k$ are considered. The probability that such a pattern generates a 'response' of the kind discussed above, is then $(k+1)/Q'$. In the square brackets $[\cdots]$ in (53), where the averaging takes place, $x^0(s)$ always appears, namely either in the leading term or in the sum. Of the remaining $Q'-1$ components, only k contribute, i.e. each of them with probability $k/(Q'-1)$. With the gauge (22), their sum yields $-(k/(Q'-1))x^0(s)$, which leads finally to the response

$$\sum_{\mu,(k^\mu=k)} (G^\mu x^0)(s) = x^0(s)\left(-1 + \frac{1}{k+1}\left[1 - \frac{k}{Q'-1}\right]\right)\frac{k+1}{Q'}\sum_\mu C^{0\mu 2}$$
$$= -x^0(s)\frac{k}{Q'-1}\sum_\mu C^{0\mu 2} =: x^0(s)g . \tag{54}$$

Thus, G can be represented by a negative real number g, representing the resistance of the system against the implementation of the additional pattern. Furthermore, it is

$$\left\langle \sum_{\mu} C^{0\mu 2} \right\rangle = \frac{1}{(N\,Q(Q-1))^2} \left\langle \sum_{\mu,\,j,k} m_{n_j^\mu} \cdot m_{n_j^0} m_{n_k^0} \cdot m_{n_k^\mu} \right\rangle$$

$$= \frac{1}{(N\,Q(Q-1))^2} \sum_{\mu,\,j} \left\langle (m_{n_j^\mu} \cdot m_{n_j^0})^2 \right\rangle \tag{55}$$

$$= \frac{1}{(N\,Q(Q-1))^2}\, pN(Q^2(Q-1)) \equiv \alpha\,. \tag{56}$$

Here we have used simplifications which have already been described, e.g. that only $j = k$ contributes, as well as (2) and the definition $\alpha = p/(N(Q-1))$, used already in (8).

In fact, with $\bar{k} := 1/p \sum_\mu k^\mu = \sum_k P(k^\mu = k)\,k$ one gets

$$g x^0 := G x^0 = -\alpha \frac{\bar{k}}{Q'-1} x^0\,. \tag{57}$$

Thus, due to the response of the patterns x^μ, the embedding strength x^0 is not given by a simple AdaTron learning step, but must be enhanced by a factor $1/(1+g) > 1$.

Additionally, g measures to what extent the inequalities for the coupling set (equation (18); see also (17)) are fulfilled. The coupling set has, in total, $N(Q-1)(Q'-1)$ degrees of freedom; $p \cdot \bar{k} = \alpha N(Q-1)\bar{k}$ of the inequalities are 'critical', i.e. fulfilled as equalities, these belong to the patterns and directions which have to be explicitly embedded, whereas the remaining $\alpha N(Q-1)(Q'-\bar{k})$ patterns and directions are stored automatically, i.e. with higher stability than necessary; $|g|$ is therefore the exhausted proportion of the number of degrees of freedom. For $g < -1$ the set of inequalities is over-determined.

4.2. Calculation of the maximal learning capacity

Now, g can be calculated self-consistently, equating pattern 0 in a statistical sense with the other patterns μ. The optimal learning capacity $\alpha_c(Q')$ for $\kappa = 0$ is then obtained with $g = -1$ (for $\kappa > 0$ see section 5).

It was shown at the beginning of the section that the field $\tilde{E}^0$ is a vector in $\mathbb{R}^{Q'}$, whose components are Gaussian random numbers around 0 with variance $|\tilde{J}|^2 Q/(Q'-1)$. As already stated, it is not necessary for the results to subtract the component $\propto (1, 1, \ldots, 1)$. If we fix the couplings such that $L := |J| \equiv L_0$, then according to (18) the distance vector from $\mathbf{0}$ to the hatched stability region in figure 1 is $c\,m_1/Q'$, i.e. its length is $c\sqrt{(Q'-1)/Q'}$, with $c = \kappa L_0$. For $\kappa > 0$, we usually set $c = 1$, i.e. $\kappa = 1/L$, however for $\kappa = 0$, L_0 can be chosen arbitrarily, e.g. $L_0 = \sqrt{(Q'-1)/Q}$ such that the above-mentioned variances become $= 1$, whereas now $c = 0$.

For $\kappa = 0$, $g(Q', \kappa)$ can thus be calculated as follows. For all random Gaussian vectors in $\mathbb{R}^{Q'}$ with components drawn from a Gaussian with zero average and variance 1, one determines the probabilities that $k = 0, 1, \ldots$ directions are active, while the pattern is stored with offset $c = 0$ in figure 1. Then, for each random vector, the integer k is determined according to (40); averaging the integers k over all vectors, one gets the expectation value $\bar{k}$. In this way, the maximal stability $\alpha_c(Q') = \alpha(Q', 0)$ can be determined from (57). For Q' up to 10, we have been able to formulate this as a Q'-dimensional integral, which could be evaluated with the NAG routine D01FCF. For $Q' \leqslant 5$, with geometric considerations, even closed expressions for $\bar{k}$ and α have been obtained.

With the abbreviation $\mathcal{D}t = (2\pi)^{-1/2}\,\mathrm{d}t\,\exp\left(-t^2/2\right)$ one gets

$$\bar{k}_c(3) = 2\int_{-\infty}^{\infty}\mathcal{D}t_1\int_{t_1}^{\infty}\mathcal{D}t_2\left(\int_{-\infty}^{t_2}\mathcal{D}t_3+\int_{(t_1+t_2)/2}^{t_2}\mathcal{D}t_3\right)$$

$$= 2\int_{0}^{\infty}\mathcal{D}t_1\left(\int_{-\infty}^{t_1/\sqrt{3}}\mathcal{D}t_2+\int_{-\infty}^{-\sqrt{3}t_1}\mathcal{D}t_2\right)=\frac{5}{6}$$

$$\bar{k}_c(4) = 3\int_{-\infty}^{\infty}\mathcal{D}t_1\int_{t_1}^{\infty}\mathcal{D}t_2\left(\int_{-\infty}^{t_2}\mathcal{D}t_3\int_{-\infty}^{t_3}\mathcal{D}t_4+\cdots\right)=\frac{3\cos^{-1}((1-2\sqrt{6})/6)}{2\pi}$$

$$\vdots.$$

The result $\alpha_c = \frac{12}{5}$ for $Q' = 3$ agrees with that of the clock model (see [5]) as it should, whereas the authors of [8] obtained the value 2.320. This is probably due to a hidden mistake in the calculation of [8], since in our formalism the additional constraints (13) used in [8] are automatically fulfilled in the limit $N \to \infty$, i.e. the components $s' = 1, \ldots, Q'$ are all equivalent, as can already be seen from (20) for unbiased random patterns. For $Q' = 4$ and 5 we get

$$\alpha_c(4) = \frac{2\pi}{\cos^{-1}\left((1-2\sqrt{6})/6\right)} \simeq 2.7580 \tag{58}$$

$$\alpha_c(5) = \left(\frac{3}{10}+\frac{\cos^{-1}\left((1+\sqrt{15})/4\sqrt{2}\right)}{2\pi}-\frac{\cos^{-1}\left(\frac{1}{6}+\left(\frac{5}{6}\right)^{3/2}\right)}{2\pi}\right)^{-1} \simeq 3.0887\,. \tag{59}$$

In all, the values for $\alpha_c(Q')$ up to $Q' = 10$ are for $Q' = 2, 3, 4, \ldots, 10$: $\alpha_c(Q')$ = 2, 2.4, 2.7580, 3.0887, 3.3996, 3.6954, 3.9791, 4.2527 and 4.5179. These results are presented in figure 2.

Above $Q' = 10$, $\alpha_c(Q')$ has been calculated by Monte Carlo simulation. For $\ln Q' \gg 1$, the simulation (see above) can be speeded up considerably: since $\bar{k} \simeq 2\ln Q'$ (see below), it is not necessary to actually generate $Q' - 1$ random numbers. With a variant of the

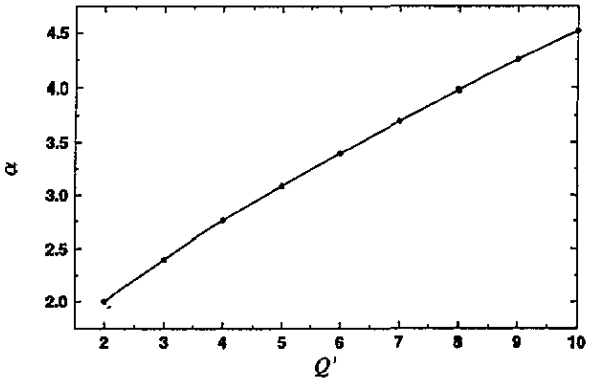

Figure 2. Optimal storage capacity α_c for $\kappa = 0$ of the Potts perceptron as a function of the number of states Q' of the output neuron for $Q' \leqslant 10$.

polar method [19] Gaussian random values above a certain threshold can be generated. The number of these values is obtained from a Poisson distribution characterized by a parameter which is identical to the average number of elements above the threshold. More details can be found in [19].

4.3. The behaviour for large values of Q'

Because of the contradictory results mentioned in section 2, we now concentrate on the limit $Q' \to \infty$. In this limit, the number $\bar{k}$ of active directions also diverges. Therefore, the exact value of the field component in the desired direction, i.e. the first component of the oriented field, is unimportant, and we replace it by 0 (see below).

If the average of the field after learning is u, then one gets for the average number $\bar{k}$ of active directions

$$\bar{k} = \frac{Q'-1}{\sqrt{2\pi}} \int_u^\infty \mathrm{e}^{-t^2/2}\,\mathrm{d}t\,. \tag{60}$$

As already mentioned, the average value t, obtained from the k active directions and the value 0 of the first component, should be larger than any of the remaining components. This implies

$$\frac{(Q'-1)(1/\sqrt{2\pi})\int_u^\infty t\exp(-t^2/2)\,\mathrm{d}t}{(Q'-1)(1/\sqrt{2\pi})\int_u^\infty \exp(-t^2/2)\,\mathrm{d}t + 1} \geqslant u \tag{61}$$

and with the definition of the error integral $\Phi(z) = \int_{-z}^{\infty} \mathcal{D}t$, and replacing the inequality by an equality, one gets the approximation

$$\frac{\exp(-u^2/2)}{u\sqrt{2\pi}} - \Phi(-u) = \frac{1}{Q'-1}\,. \tag{62}$$

The results from this approximation for the reduced learning capacity $\Delta I(Q') = (\alpha_c \ln Q')/(Q'-1)$, i.e. the information gain per coupling element (14) obtained from (11) for $P_0 = 1$, are presented in figure 3 as a chain curve, together with the exact results (full curve), and the results from the Hebb rule (broken curve, see below). The asymptotic value, to which all these curves converge extremely slowly, is shown in the following to be $= \frac{1}{2}$.

In fact, the well known asymptotic expansion of $\Phi(u)$ leads to

$$\frac{\mathrm{e}^{-u^2/2}}{\sqrt{2\pi}u^3} = \frac{1}{Q'} \quad \Longrightarrow \quad u^2 = 2\ln Q' - \mathcal{O}(\ln(\ln Q'))\,. \tag{63}$$

From this we obtained, with (60),

$$\bar{k} = \frac{Q'}{\sqrt{2\pi}}\frac{\mathrm{e}^{-u^2/2}}{u} = u^2 = 2\ln Q' - \mathcal{O}(\ln(\ln Q')) \tag{64}$$

and thus

$$\alpha_c = \frac{Q'-1}{\bar{k}} = \frac{Q'-1}{2\ln Q'}\left[1 + \mathcal{O}\left(\frac{\ln(\ln Q')}{\ln Q'}\right)\right] \tag{65}$$

or

$$\Delta I(Q') = \frac{\alpha_c \ln Q'}{Q'-1} = \frac{1}{2} + \mathcal{O}\left(\frac{\ln(\ln Q')}{\ln Q'}\right)\,. \tag{66}$$

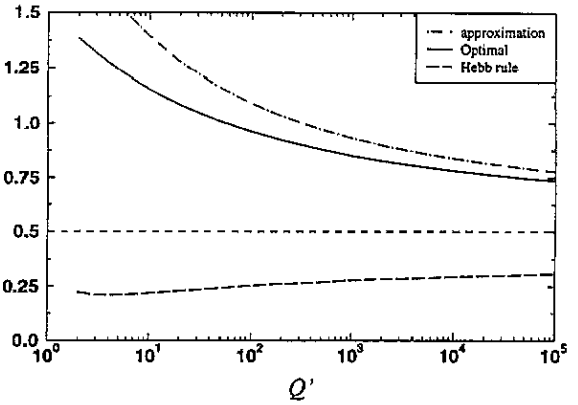

Figure 3. The information per coupling, ΔI from (14), at $\alpha = \alpha_c(Q')$ is plotted over Q' for the optimal Q'-output state Potts perceptron with $\kappa = 0$ (full curve). The approximation of (62) is given additionally (chain curve). As one can see, convergence towards the limiting value $\frac{1}{2}$ is extremely slow. For comparison, the information content (11) is also plotted for the Hebb rule for the particular case that the output neuron has to give the right answer for 90% of the input patterns. The limiting value in this case is again $\frac{1}{2}$.

Thus, the information gain $\Delta I(Q')$ per matrix element decreases extremely slowly against $\frac{1}{2}$ and thus remains finite, in contrast to what Nadal and Rau have predicted in [8], who obtained $\Delta I \to 0$.

At this point, we stress that for $\kappa = 0$ we have extended our Monte Carlo integration up to $Q' = 10^{1600}$, whereas in figure 3 we have plotted results only up to $Q' = 10^5$. For $Q' = 10^{1600}$ we have found $\Delta I = 0.5019$, in agreement with our asymptotic estimate of (66).

4.4. Results for the Hebb rule

The results for the Hebbian case, i.e. the broken line of figure 3, are evaluated from (11) for $P_0 = 0.9$, i.e. $\alpha_c(Q')$ has been evaluated with (7) for the case that 90% of the patterns $\mu = 1, \ldots, p$ are classified correctly. According to figure 3, even for this case our asymptotic estimate (66) seems to apply; however, since for Hebbian couplings (7) we have not been able to go beyond $Q' = 10^5$, it seems necessary to support this suggestion by an analytical estimate: for the Hebb rule, the oriented field is

$$\boldsymbol{E}^\mu = \boldsymbol{x}^\mu + \tilde{\boldsymbol{E}}^\mu = \boldsymbol{m}^1 + \tilde{\boldsymbol{E}}^\mu . \tag{67}$$

Here, the noise field $\tilde{\boldsymbol{E}}^\mu$ is again a random vector. Since for the Hebb rule embedding strengths and correlations between different patterns are independent, the variance can be calculated as

$$\left\langle \left| \tilde{\boldsymbol{E}}^\mu \right|^2 \right\rangle = \left\langle \frac{1}{N(Q-1)} \sum_{\varrho\nu} \boldsymbol{x}^{\varrho T} \boldsymbol{B}^{\varrho\nu} \boldsymbol{x}^\nu \right\rangle = \frac{p}{N(Q-1)} Q'(Q'-1)$$
$$= \alpha\, Q'(Q'-1) . \tag{68}$$

For random patterns, only the diagonal elements of the oriented correlation matrix $\overleftrightarrow{\boldsymbol{B}}$ in (68) should contribute, which are unit matrices. The probability $P(\alpha, Q')$ that a perceptron, having implemented $\alpha N(Q-1)$ patterns with the Hebb rule, will classify, for example, the first pattern correctly, can be obtained from the following probabilistic experiment: it

is tested whether the signal term, i.e. the sum $g_1 + Q'$, where g_1 is a Gaussian random number with average 0 and variance $\alpha Q'$, is larger than the noise resulting from the other patterns, i.e. the sum of $Q' - 1$ similar random numbers $g_2, \ldots, g_p$. With $u := \sqrt{Q'/\alpha}$ and $t_i := g_i/\sqrt{\alpha Q'}$ one gets

$$P(\alpha, Q') = \int_{-\infty}^{\infty} \mathcal{D}t_1 \left(\int_{-\infty}^{u+t_1} \mathcal{D}t_2 \right)^{Q'-1} = \int_{-\infty}^{\infty} \mathcal{D}t_1 (\Phi(u + t_1))^{Q'-1} \tag{69}$$

which generalizes the corresponding well known expression for $Q' = 2$ (see [21]).

Let us now assume that α_c grows much more slowly than Q', i.e. with $u := \sqrt{Q'/\alpha}$ one has $u \to \infty$ for $Q' \to \infty$. The probability $P(\alpha, Q')$ in (69) shall now be abbreviated as P_0, see (11), and assumed to be finite ($0 < P_0 < 1$). The function $f := \Phi^{Q'-1}$ in (69) is a strictly monotonic function of t_1, which increases from 0 to 1 when t_1 increases from $-\infty$ to ∞. The learning capacity can thus be estimated from above (from below, respectively) by the following simple consideration: let us choose at first a threshold t_a such that the Gaussian random number t_1 is less than t_a with probability $1 - (P_0/2)$.

The least favourable estimate, which can be made for f, is that $f(t) = 1$ for $t > t_a$ and $f(t) = z$ (= constant) for $t \leqslant t_a$. If the integral still yields P_0, then $z = P_0/(2 - P_0)$. Thus, we have

$$\int_{-\infty}^{t_a} \mathcal{D}t = 1 - \tfrac{1}{2}P_0 \quad \Longrightarrow \quad (\Phi(u + t_1))^{Q'-1} \geqslant \frac{P_0}{2 - P_0} =: C_a \,. \tag{70}$$

Analogously one has

$$\int_{-\infty}^{t_b} \mathcal{D}t = \tfrac{1}{2}(1 - P_0) \quad \Longrightarrow \quad (\Phi(u + t_1))^{Q'-1} \leqslant \frac{2P_0}{1 + P_0} =: C_b \,. \tag{71}$$

We now treat both cases simultaneously, writing t for t_a (t_b respectively) and C for C_a (C_b respectively). For the following it is only important that both C and t do not depend on Q'. Since $0 < P_0 < 1$, C is also between 0 and 1. Since $u \to \infty$ for $Q' \to \infty$, while t remains finite, we again apply the asymptotic expansion of $\Phi(x)$ and obtain

$$\left(1 - \frac{1}{\sqrt{2\pi}(u + t)} \exp\left(-\tfrac{1}{2}(u + t)^2\right)\right)^{Q'-1} \simeq C \tag{72}$$

$$\Longleftrightarrow \ln\left(1 - \frac{1}{\sqrt{2\pi}\,(u + t)} \exp\left(-\tfrac{1}{2}(u + t)^2\right)\right) \simeq \frac{\ln C}{Q' - 1} \tag{73}$$

$$\Longleftrightarrow \frac{1}{\sqrt{2\pi}\,(u + t)} \exp\left(-\tfrac{1}{2}(u + t)^2\right) \simeq \frac{-\ln C}{Q' - 1} \,. \tag{74}$$

Again replacing $Q' - 1$ by Q', we obtain

$$\tfrac{1}{2}(u + t)^2 \simeq \ln Q' - \ln(-\ln C) - \ln(\sqrt{2\pi}) - \ln(u + t) \tag{75}$$

$$\Longrightarrow \quad u = \sqrt{2 \ln Q'} \left(1 - \mathcal{O}\left(\frac{\ln \ln Q'}{\sqrt{\ln Q'}}\right)\right) - t \,. \tag{76}$$

This result justifies the assumption made above with respect to u. With the definition of u one gets for α_c, in the limit $Q' \gg 1$,

$$\alpha_c = \frac{Q'}{u^2} = \frac{Q'}{2 \ln Q'} \left(1 + \mathcal{O}\left(\frac{\ln \ln Q'}{\ln Q'}\right)\right) . \tag{77}$$

This result does not depend on P_0, as long as P_0 is > 0, but < 1. For $Q' \to \infty$ we thus obtain the same information gain ΔI (equation (66)), as for the optimal perceptron.

In contrast, with similar techniques we could prove in [18] that for perceptrons which have been forced to implement with the Hebb prescription (7) as many patterns as correspond to $\alpha_c(Q') = b\,Q'$, with $b > 0$, the value of P_0 would *not* remain finite, but converge to 0 for $Q' \to \infty$ so fast that $\Delta I \to 0$. This solves the question left open by Kanter [7] (see section 2): as mentioned there, Kanter has obtained results for Hebbian auto-associative Potts perceptrons with $Q' \leqslant 9$ which apparently favoured $\alpha_c \simeq 0.069\,Q'$, but the question remained whether sensible recognition would be possible for these α_c even in the limit $Q' \to \infty$. Since $P_0 \to 0$, this is not the case, whereas for our smaller asymptotic values, $\alpha_c(Q') = \frac{1}{2} \ln Q'/(Q'-1)$, a Hebbian Potts perceptron does perform sensible classifications in this limit, and with the same information gain as the AdaTron algorithm for $\kappa = 0$.

For the *layered feedforward* model of Shim *et al* in [10] we have no such general statements. However, we have seen that the learning capacity of the Potts perceptron only converges extremely slowly to the final limit, both for the Hebb rule and also for the case of optimal stability. In [10] it was found that, for $Q \leqslant 10^4$, $\alpha_c(Q)$ behaved $\sim Q^{\Delta}$, with $\Delta \simeq 0.85$. Since this behaviour was derived from a log–log plot, it may be reasonable to also expect for this model that the true behaviour might be similar, as just seen, namely $\alpha_c \simeq c(Q)Q/\ln Q$, where $c(Q)$ may vary slowly with Q. With a pocket calculator one can convince oneself that for $Q \leqslant 10^4$ these two suggestions can hardly be distinguished. With the last-mentioned suggestion, the information gain per coupling would again remain finite for $Q \to \infty$.

5. Optimal storage capacity for $\kappa > 0$: results and two general formulae

The case of finite stability κ is, of course, more realistic for applications than the marginal case $\kappa = 0$ considered above.

With the formulae for L^2 in (25), and with (27) and (5) and the Kuhn–Tucker conditions with $c = 1$ for non-vanishing $x^\mu(s)$ with $s > 1$, we get

$$\begin{aligned} L^2 &= \frac{1}{N\,Q(Q-1)} \sum_{\mu,\nu} x^{\mu T} B^{\mu\nu} x^\nu = \frac{1}{N\,Q(Q-1)} \sum_\mu \sum_{s=1}^{Q'} x^\mu(s) E^\mu(s) \\ &= \frac{1}{N\,Q(Q-1)} \sum_\mu \left[x^\mu(1) E^\mu(1) + \sum_{s=2}^{Q'} x^\mu(s)(E^\mu(1) - 1) \right] \\ &= \frac{1}{N\,Q(Q-1)} \sum_\mu \left[x^\mu(1) \cdot 1 + (E^\mu(1) - 1) \sum_{s=1}^{Q'} x(s) \right] \\ &= \frac{1}{N\,Q(Q-1)} \sum_\mu x^\mu(1) = \frac{\alpha}{Q} \int x(1) w(x(1))\,\mathrm{d}x(1)\,. \end{aligned} \tag{78}$$

Here $w(x(1))\,\mathrm{d}x(1)$ is the probability to find a weight in $[x(1), x(1)+\mathrm{d}x(1)]$. As mentioned above, if we normalize the couplings such that $c := \min_{\mu,s'(>1)}\{E^\mu(1) - E^\mu(s')\} = 1$ in (18), then $\kappa^{-1} = L = |\vec{J}|$, and $w(x(1))$ results from a Gaussian distribution with average zero and variance per component $L^2\,Q/(Q'-1)$ for a Q'-dimensional vector, the oriented field t, from which an AdaTron learning step $x\{t\}$ is performed in the direction of the stability region. As already mentioned, the response of the other patterns necessitates an enhancement of the embedding strengths by the factor $1/(1+g)$. Then one gets from the

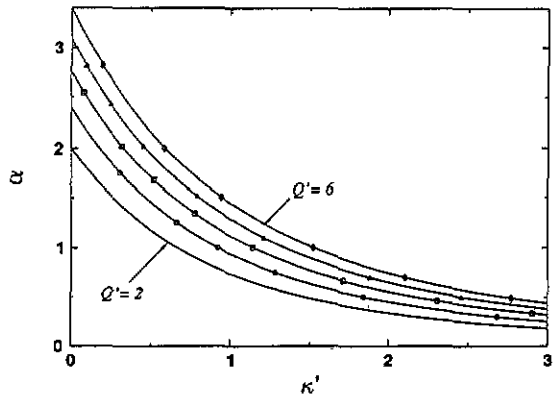

Figure 4. The storage capacity $\alpha(Q', \kappa')$ as a function of the rescaled stability parameter $\kappa' = \kappa\sqrt{(Q'-1)/Q}$, where κ is defined in (18), is plotted against κ' for $Q' = 2, 3, \ldots, 6$. Rescaling κ' we arrive at the well known results for $Q' = 2$ (see [1]) and $Q' = 3$ (see [5]). The symbols represent direct simulations using random patterns stored with our AdaTron algorithm. These simulations yield results in agreement with the analytical calculations. Standard error bars are smaller than the size of the symbols.

final expression in (78)

$$L^2 = \frac{\alpha}{\sqrt{Q(Q'-1)}} \frac{L}{1+g} \int \mathcal{D}\tilde{t}\, x_{\kappa'}\{\tilde{t}\}(1)\,. \tag{79}$$

In (79) we have used $\tilde{t} := t\sqrt{(Q'-1)/(QL^2)}$ and $x_{\kappa'}\{\tilde{t}\}(1) := x\{t\}\sqrt{(Q'-1)/(QL^2)}$. Here we still have $c = \kappa L = 1$.

Now we set $L = \sqrt{(Q'-1)/Q}$, i.e. the variance of the components of t $(= \tilde{t})$ is scaled to 1. As a consequence, $c := \kappa L$ is now identical to $\kappa' := \kappa\sqrt{(Q'-1)/Q}$, which we use as a scaled stability measure. Thus, if a pattern originally possessing the oriented field t is *embedded* with this scaled stability $c = \kappa'$, for $L = \sqrt{(Q'-1)/Q}$, this means that by the training process the field is shifted beyond the limit of correct recognition by the additional amount

$$\kappa'(t) := \frac{1}{k+1}\left(k\kappa',\, 0, \ldots, -\kappa', 0, \ldots\right) \tag{80}$$

where $-\kappa'$ only appears for the k 'active directions'. Now, from the gauge condition (22) it follows that

$$\kappa' x_{\kappa'}\{t\}(1) = \kappa'\{t\} \cdot x_{\kappa'}\{t\}\,. \tag{81}$$

If additionally we substitute $L^{-1} = \kappa = \kappa'\sqrt{Q/(Q'-1)}$ in (79) and use (81), we obtain

$$1 + g = \frac{\alpha}{Q'-1} \int \mathcal{D}t\, \kappa'\{t\} \cdot x_{\kappa'}\{t\}\,. \tag{82}$$

Finally, we replace the response g of the system for a small perturbation by (57) and use the number k of active directions, which can be expressed with Heavisides' function $\theta[x]$ as $k = \sum_{s=2}^{Q'} \theta[x_{\kappa'}\{t\}(s)]$. This leads to

$$1 = \frac{\alpha}{Q'-1} \int \mathcal{D}t\, k\{t\} + \frac{\alpha}{Q'-1} \int \mathcal{D}t\, \kappa'\{t\} \cdot x_{\kappa'}\{t\}\,. \tag{83}$$

For $Q' = 2, 3, \ldots, 6$, from an evaluation of the integrals in (83), in figure 4 we present the dependence of the capacity $\alpha_c(Q', \kappa')$ on the scaled stability κ', together with results obtained from direct simulations, and obtain convincing agreement.

Interestingly, one can show by partial integration that (83) is equivalent to

$$1 = \frac{\alpha}{Q'-1} \int \mathcal{D}t\,(x_{\kappa'}\{t\})^2 . \tag{84}$$

Particular cases of this equation are the well known formula of Gardner [1] for perceptrons with Ising neurons

$$1 = \alpha \int_{-\kappa}^{\infty} \mathcal{D}t\,(t+\kappa)^2 \tag{85}$$

and the corresponding equations for perceptrons with clock neurons in [5]. However, from a recent paper of Griniasty [17] it follows that approaches of the type of (84) or (85) are equivalent to the replica-symmetric approximations, whereas our formula (83) yields different results (see [18]) when replica symmetry is broken, i.e. when there is no longer a unique ground state nor a Gaussian distribution of oriented fields.

Futhermore, as already stated by [17], equation (84) can formally be derived by assuming that simultaneously (i) $g = 0$, and (ii) only the diagonal elements of B contribute. Neither (i) nor (ii) are true, as shown above and as can also be seen in our simulations, however, as long as one is in the replica-symmetric phase, (i) and (ii), apparently 'conspire' to the correct result: namely, assuming that the norm of the couplings, L (=1), is determined self-consistently, we obtain immediately from (25)

$$\frac{Q'-1}{Q} = \frac{1}{N\,Q(Q-1)} \vec{x}_{RS}^{\,T} \overleftrightarrow{B}_{RS}\, \vec{x}_{RS} \tag{86}$$

$$= \frac{1}{N\,Q(Q-1)} \sum_{\mu} |x^{\mu}|^2 = \frac{\alpha}{Q} \int \mathcal{D}t\, x_{\kappa'}^2 . \tag{87}$$

Here the subscript RS means 'replica symmetric'.

6. Conclusions

We have studied the problem of maximal storage capacity α_c (8) and optimal learning processes of the AdaTron-type for Potts model perceptrons, i.e. with Q-state Potts input and Q'-Potts output neurons and real couplings constrained by one global spherical condition. Concerning the patterns and the input–output relation, we have assumed $\mu = 1, \ldots, p$ random input vectors, and random outputs with no correlation to the inputs, although our methods and the AdaTron algorithm would also apply to more general situations. Under the above-mentioned 'random conditions', the results do not depend on Q but only on Q' and on the stability κ. For $\kappa = 0$, we have obtained exact results for $2 \leqslant Q' \leqslant 10$, where α_c could be evaluated from multi-dimensional integrals. With the accurate and fast simulation of a certain Gaussian process, α_c could also be obtained for much larger values up to as large as $Q' = 10^{1600}$. Moreover, we were able to derive analytically the asymptotic behaviour for $Q' \to \infty$, namely $\alpha_c(Q') := p_{max}/[N(Q-1)] \to (Q'-1)/(2 \ln Q')$, which is, however, only reached extremely slowly, namely with an error vanishing for $Q' \to \infty$ as $[\ln(\ln Q')]/\ln Q'$. For AdaTron learning, i.e. optimal stability, the information gain per coupling, $\Delta I(Q') := (\alpha_c \ln Q')/(Q'-1)$, then approaches the value $\frac{1}{2}$ slowly from above, whereas for Hebbian learning, with the allowance of only a finite percentage of errors, the same limit is obtained from below, in both cases extremely slowly, as just mentioned. In this way, certain open problems from the literature (e.g. [7, 8, 10]) have been solved. The reason for the efficiency of the Hebb rule for Potts perceptrons in the limit $Q' \to \infty$, which should

be contrasted with the different behaviour of the *clock* model perceptron as discussed in [5, 6], is probably related to the fact that for Potts neurons, due to the dimensional increase, the phase-space segments leading to recognition increase strongly in volume with increasing Q', whereas for *clock* model neurons they become smaller and smaller angular segments in two dimensions.

Finally, we also obtained results for finite κ, including two general formulae (83) and (84), derived by cavity arguments, which put our results into a more general context. These formulae will be extended and applied in a forthcoming paper to the problem of generalization, and in a second forthcoming paper to situations where the replica-symmetry is broken, namely (i) for perceptrons above α_c and (ii) for the AND-machine.

Acknowledgments

Valuable hints and discussions with K Bauer, D Bollé, M Bouten, A Engel, W Kinzel, O Krisement, R Kühn, M Opper, B Schottky, J Winkel and A Zippelius are gratefully acknowledged by the authors. The computations have been performed at the computing centres of the university of Regensburg, the LRZ at Munich, and the HLRZ at Jülich.

References

[1] Gardner E 1987 *Europhys. Lett.* 4 481
[2] Hertz J, Krogh A and Palmer R G 1994 *Introduction to the Theory of Neural Computation* (Redwood, CA: Addison-Wesley)
[3] Müller B and Reinhardt J 1991 *Neural Networks* (Berlin: Springer)
[4] Peretto P 1992 *An Introduction to the Modeling of Neural Networks* (Cambridge: Cambridge University Press)
[5] Gerl F, Bauer K and Krey U 1992 *Z. Phys.* B **88** 339
[6] Schottky B, Gerl F and Krey U 1994 *Z. Phys.* B at press
[7] Kanter I 1988 *Phys. Rev.* A **37** 2739
[8] Nadal J P and Rau A 1991 *J. Physique I* 1109
[9] Amit D J, Gutfreund H and Sompolinsky H 1987 *Ann. Phys.* **173** 30
[10] Shim G M, Kim D and Choi M Y 1992 *Phys. Rev.* A **45** 1348
[11] Vogt R and Zippelius A 1992 *J. Phys. A: Math. Gen.* **25** 2209
[12] Anlauf J K and Biehl M 1989 *Europhys. Lett.* **10** 687
[13] Fletcher R 1988 *Practical Methods of Optimization* (New York: Wiley)
[14] Kinzel W and Opper M 1989 Models of neural networks ed E Domany, L van Hemmen and K Schulten *Physics of Neural Networks* (Berlin: Springer)
[15] Mézard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
[16] Mézard M 1989 *J. Phys. A: Math. Gen.* **22** 2181
[17] Griniasty M 1993 *Phys. Rev.* E **47** 4496
[18] Gerl F 1994 *PhD thesis* University of Regensburg (unpublished)
Gerl F and Krey U to be published
[19] Knuth D E 1969 *The Art of Computer Programming* vol 2 (Reading, MA: Addison-Wesley) ch 3.4.1
[20] Watkin T L H, Rau A, Bollé Dand van Mourik J 1992 *J. Physique I* **2** 167
[21] Amit D J 1989 *Modelling Brain Function: The World of Attractor Neural Networks* (Cambridge: Cambridge University Press)