A Kuhn-Tucker cavity method for generalization with applications to perceptrons with Ising and Potts neurons

J. Phys. A: Math. Gen. 28 (1995) 6501–6516. Printed in the UK

A Kuhn–Tucker cavity method for generalization* with applications to perceptrons with Ising and Potts neurons

F Gerl† and U Krey‡

† Institut für Theoretische Physik der Universität Göttingen, Bunsenstr. 9, D-30373 Göttingen, Germany

‡ Institut für Physik II der Universität Regensburg, Universitätsstr. 31, D-93040, Germany

Received 31 October 1994, in final form 7 August 1995

Abstract. Within the framework of statistical physics, we derive a cavity method for generalization by perceptrons, where the Kuhn–Tucker conditions for optimal stability are built into the cavity fields. In this way, the calculation of the generalization ability for learning processes leading to optimal stability is simplified. Within our approach, the degrees of freedom of the neurons can be rather arbitrary. For perceptrons with Ising neurons we relate our method to the traditional replica approach. New results are obtained for Q-state Potts model perceptrons, including the asymptotic behaviour for $\alpha \to \infty$ and general Q.

1. Introduction

In a recent paper, [1], we have treated the *learning problem* for Q-state Potts model perceptrons within the framework of statistical physics by a general method based on a cavity formalism. In this method, the Kuhn–Tucker conditions, which lead to optimal stability in AdaTron type learning processes, have been built into the *cavity formulation*. In this way, we obtained a number of exact results for learning with maximal stability for Potts model networks in [1], and in [2] for the clock model case.

In the present paper, we extend our Kuhn–Tucker cavity approach to the more complicated *generalization problem*. There one considers two automata, each possessing the same number and type of input units and one output unit. The inputs to the automata are termed 'questions', and the outputs 'answers'. One of these automata, 'the teacher', is fixed, and her answers are 'correct' per definition, whereas the second one, 'the student', has couplings, which are changed in the course of a training process during which she tries to minimize the number of errors of her answers with respect to a set of random 'questions'. Of course, this problem can be extended, e.g. to cases, where the 'teacher' gives the correct answer only with a certain probability. However, in the following, such extensions will not be considered.

An extensive description of generalization problems in the case of perceptrons with Ising neurons has been given by Watkin *et al* [3]. In section 2 we demonstrate our cavity method for this case and find that the results from our Kuhn–Tucker cavity approach agree with those obtained by the traditional replica approach, see [4], although our two self-consistency equations are different from those of [4] (see below). Moreover, in this section we present a detailed derivation and comparison of both approaches by cavity

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

0305-4470/95/236501+16$19.50 © 1995 IOP Publishing Ltd

arguments, which lead to insights allowing the treatment of more complex systems. The cavity arguments corresponding to the replica approach generalize ideas of Griniasty [5], and are equivalent to the assumption of replica symmetry, whereas our approach is more general, see [6]. Finally, at the end of section 2 we find it useful for numerical calculations, to combine equations from both approaches in a certain, non-trivial way.

In section 3, we extend these studies to the case of Potts perceptrons: we assume that both automata have the same number N of Q-state Potts input neurons and one Q'-state output neuron, and real couplings, see below. For this case, we obtain accurate results for the generalization ability $G(\alpha, Q')$ and for the information gain $\Delta I_{\rm rel}(\alpha, Q')$ for Q' ranging from 3 to 1000. To our knowledge the generalization ability of Potts perceptrons with optimal stability has up to now only been treated in the so-called annealed approximation, [7].

For both cases, i.e. for the perceptron with Ising neurons and for the Potts model case, both with real couplings subject to the usual spherical constraint, the desired maximal stability and good generalization properties can be obtained by the fast and efficient AdaTron learning processes defined in [8, 2] respectively. An even better generalization would be obtained with the so-called 'optimal (i.e. Bayesian) perceptron' as discussed by [3, 9]; however, the 'optimal perceptron' is not easily approximated and therefore not discussed in the following.

In [6], one of us has treated further problems with our Kuhn–Tucker cavity method, for which a replica-symmetric approach does not suffice, in contrast to the present situation. Those additional results of [6] will be published in a subsequent paper [10].

2. Generalization for perceptrons with Ising neurons

2.1. Basic equations

We consider perceptrons with N input neurons, $\underline{s} := (s_1, \ldots, s_N)$, and (for simplicity) one output neuron s_{out}. Here the s_i and s_{out} are binary variables, e.g. $s_i = \pm 1$ ('Ising neurons'). The output ('answer') generated by the input vector $\underline{s}$ ('question') is given by the usual rule for simple binary perceptrons

$$s_{out} = \mathrm{sign}\left[\sum_{k=1}^{N} J_k s_k\right] \tag{1}$$

which means that the perceptron classifies the inputs, or 'gives answers to input questions'. Now let us consider two perceptrons with couplings J_k^T and J_k, respectively, the 'teacher' and 'student' perceptrons, and a 'training set' of p questions $\underline{s} = \underline{\xi}^\mu := (\xi_1^\mu, \ldots, \xi_N^\mu)$ for $\mu = 1, \ldots, p$. The correct answers to any questions, particularly those to the 'training questions', ζ^μ, are given by the teacher perceptron ($\zeta^\mu \equiv \zeta_T^\mu$). Thus it is the student's task to adapt her couplings in such a way by a certain learning process (see below) that she gives the correct answers to the training questions.

The *measure of performance* of the student perceptron, after just having 'learnt' $p := \alpha \cdot N$ training questions $\underline{\xi}^\mu$ with answers $\zeta_T^\mu = \mathrm{sign}[\sum_{k=1}^N J_k^T \xi_k^\mu]$, is the generalization probability $G(\alpha)$, namely the probability to answer an additional random question in the same way as the teacher. (More generally this is the task to learn a given 'rule' from $p = \alpha N$ questions.)

In the following we assume for simplicity that the question bits are chosen randomly as ± 1 with probability $\frac{1}{2}$. Concerning the teacher we then only have to assume that the output is not dominated by very few couplings J_k^T, see below. The quantity $\alpha = p/N$ is usually called the 'loading parameter'.

Let us denote by

$$h_T^\mu := \sum_{k=1}^{N} J_k^T \xi_k^\mu \qquad \text{and} \qquad h^\mu := \sum_{k=1}^{N} J_k \xi_k^\mu \tag{2}$$

the so-called *presynaptic fields* at the output neuron of the teacher and the student perceptron, respectively, generated by the question $\underline{\xi}^\mu := (\xi_1^\mu, \ldots, \xi_N^\mu)$, which is drawn from the already mentioned 'training set' with $\mu = 1, \ldots, p(= \alpha N)$ questions. For simplicity, the length of the teacher's coupling vector $\underline{J}^T := (J_1^T, \ldots, J_N^T)$ is normalized as $|\underline{J}^T|^2 = 1$ (as already mentioned).

In contrast, for the student perceptron the length $L := |\underline{J}|$ is *minimized* in the course of the AdaTron training process, see below, i.e. the *stability* $\kappa := 1/L$ is maximized.

Also the presynaptic field h_T^μ at the teacher's output neuron is then (i.e. under the above-mentioned conditions concerning the couplings) for $N \gg 1$ a Gaussian random number with average 0 and variance 1, i.e. 'normally distributed'. Now the output of the teacher, i.e. 'the correct answer' to the 'question' $\underline{\xi}^\mu$, is $\zeta_T^\mu := \mathrm{sign}(h_T^\mu)$. Therefore, the so-called *re-oriented* field of the teacher, $E_T^\mu = \zeta_T^\mu h_T^\mu$, is a random variable x with the probability density $p(x) = (\Theta(x)/\sqrt{\pi}) \exp(-x^2/2)$, where $\Theta(x) = 1$ for $x > 0$, $= 0$ for $x < 0$.

2.2. *The re-oriented field of a new pattern*

If now an additional pattern with index $\mu = 0$ is added to the training set, the re-oriented presynaptic field (i.e. re-oriented with the teacher's answer) of the student is

$$\tilde{E}^0 := \mathrm{sign}(h_T^0) \sum_{k=1}^{N} J_k \xi_k^0 \tag{3}$$

where the $\tilde{}$ is introduced to indicate that one is dealing with the 'bare' re-oriented cavity field, i.e. before any additional training. Crucial in constructing this *non*-Gaussian random number is the *overlap* R between the teacher and student perceptrons

$$R := \frac{\underline{J}^T \cdot \underline{J}}{|\underline{J}^T||\underline{J}|} \tag{4}$$

which also characterizes the statistical physics of our learning process.

The coupling vector of the student perceptron with length $L := |\underline{J}|$ can be decomposed into a component of length $R \cdot L$ parallel to the coupling vector of the teacher perceptron and a part of length $\sqrt{1 - R^2} \cdot L$ perpendicular to it. We introduce two normally distributed random numbers u_1 and u_2, which characterize the local fields generated by the normalized coupling vectors in these directions according to (3). We then get for the re-oriented fields generated by the new patterns at the teacher's and student's output neuron, t_1 and t_2, respectively,

$$t_1 := |u_1| \qquad t_2 := L(R|u_1| + \sqrt{1 - R^2} u_2) . \tag{5}$$

The probability density for the pair (t_1, t_2) is

$$\begin{aligned} P(t_1, t_2) &= \iint \mathcal{D}u_1 \, \mathcal{D}u_2 \, \delta(t_1 - |u_1|) \delta(t_2 - L(R|u_1| + \sqrt{1 - R^2} u_2)) \\ &= \frac{2\Theta(t_1)}{2\pi} \frac{1}{L\sqrt{1 - R^2}} \exp\left(- \frac{t_1^2 - 2Rt_1t_2/L + (t_2/L)^2}{2(1 - R^2)} \right) \end{aligned} \tag{6}$$

where $\mathcal{D}x = \exp(-x^2/2)\mathrm{d}x/\sqrt{2\pi}$. The probability density for the bare re-oriented presynaptic field $\tilde{E}^0 = t_2$ at the student's output neuron simplifies to

$$P(t_2) = \int_0^\infty \mathrm{d}t_1 P(t_1, t_2) = \frac{1}{L\sqrt{2\pi}} \exp\left(-\frac{t_2^2}{2L^2}\right) 2\Phi\left(\frac{Rt_2}{L\sqrt{(1-R^2)}}\right) \tag{7}$$

with $\Phi(x) = \int_{-\infty}^x \mathcal{D}z$. Correct classification implies $t_2 > 0$, which leads for given overlap R to the *generalization ability*

$$G(R) := \int_0^\infty \mathrm{d}t_2\, P(t_2) = 1 - \frac{1}{\pi} \arccos(R) . \tag{8}$$

These, of course, are known results of [4], and only repeated here for later purposes. For the Potts perceptron, the equations are considerably more complex (see section 3).

2.3. System response for optimal stability

In the following we calculate the overlap R under the condition of optimal stability as a function of α. With equation (8) one can then derive G as a function of α. Here we use the Kuhn–Tucker conditions (see below) and cavity arguments as in [1] concerning the necessary response of the system to maintain these conditions for the already stored patterns in the presence of the newly added pattern, which must also be stored.

The couplings of a perceptron trained for optimal stability can always be expressed in the form [8]

$$J_k = \frac{1}{N} \sum_\mu x^\mu \zeta_T^\mu \xi_k^\mu \tag{9}$$

with the so-called 'embedding strengths' $x^\mu \geqslant 0$. As can be shown using Lagrangian multipliers [8, 1], these embedding strengths have to fulfil the so-called Kuhn–Tucker conditions, see below. Without restriction of generality, these are usually formulated by fixing the length L of the coupling vector $\underline{J}$ in such a way that the stability limit for $\kappa > 0$ corresponds to $E^\mu = 1$, i.e. $L = \kappa^{-1}$. With this convention, which we always use in the following, unless otherwise stated, the Kuhn–Tucker conditions are

$$\text{either} \quad (x^\mu > 0 \quad \text{and} \quad E^\mu = 1) \quad \text{or} \quad (x^\mu = 0 \quad \text{and} \quad E^\mu > 1) . \tag{10}$$

In fact the AdaTron algorithm (without overrelaxation) of Anlauf and Biehl [8],

$$\delta x^\mu = \max(-x^\mu, 1 - E^\mu) \qquad \text{(sequentially or in parallel)} \tag{11}$$

simply fixes the x^μ repeatedly to values which fulfil the Kuhn–Tucker conditions (10). If the algorithm converges, the conditions are therefore automatically obeyed.

Using the 'oriented correlation' matrix

$$B^{\mu\nu} = \zeta_T^\mu \zeta_T^\nu \sum_{k=1}^N N^{-1} \xi_k^\mu \xi_k^\nu \tag{12}$$

and the definition (2), we can write for the re-oriented field E^μ

$$E^\mu := \zeta_T^\mu \sum_k J_k \xi_k^\mu = \sum_\nu B^{\mu\nu} x^\nu . \tag{13}$$

With the Kuhn–Tucker conditions we have finally

$$L^2 = \sum_{\mu,\nu} x^\mu B^{\mu\nu} x^\nu = \frac{1}{N} \sum_\mu x^\mu . \tag{14}$$

As in [1], we now add a new 'question' $\underline{\xi}^0$ to the training set, with correct answer ζ_T^0 given by the teacher. If the new training pattern ('question') is not incidentally answered correctly and with sufficient stability with unchanged J_k, i.e. with $\tilde{E}^0 \geqslant 1$, then one *tries* to embed it with $x_0 = 1 - \tilde{E}^0 > 0$, see equation (11). Here $\tilde{E}^0$ has been defined in (3), and as already mentioned, the ˜ only reminds us of the fact that one is dealing with the re-oriented field *before* embedding of the pattern $\underline{\xi}^0$. The change of the couplings due to the introduction of x_0 leads to a perturbation of order $\simeq \mathcal{O}(1/\sqrt{N})$ of the presynaptic fields generated by the other stored patterns, so that the Kuhn–Tucker conditions of (10) are now generally violated.

The Kuhn–Tucker conditions are then restored by a parallel AdaTron step (see below). This restoration corresponds to a 'macroscopic response' of the systems (i.e. the couplings). As in [1], all further restoring steps lead only to corrections of order $\mathcal{O}(1/\sqrt{N})$ in the response and are therefore neglected in the thermodynamic limit $N \to \infty$. Additionally, for the desired accuracy $\simeq \mathcal{O}(1/\sqrt{N})$ it can be assumed that the x^μ and the matrix elements $B^{\mu\nu}$ are uncorrelated.

To be specific, the 'embedding perturbation' y^μ, which is generated by the newly added pattern 0 and afflicts the pattern μ with $x^\mu > 0$, is $y^\mu = B^{\mu 0} x^0$. Therefore it necessitates the correction $\delta x^\mu = -B^{\mu 0} x^0$. In all, these corrections for $\mu = 1, \ldots, p$ generate at pattern 0 a response field

$$g x^0 = \sum_{\mu,(x^\mu>0)} B^{0\mu} \delta x^\mu = - \sum_{\mu,(x^\mu>0)} (B^{0\mu})^2 x^0 \tag{15}$$

which *reduces* the effect of the AdaTron step with x^0. Therefore, one has to enhance $x^0 = 1 - \tilde{E}^0$ by an *amplification factor* $1/(1+g)(>1)$.

Now the $(B^{0\mu})^2$ are $1/N$ on average, see (12). Therefore one gets immediately

$$\sum_{\mu,(x^\mu>0)} (B^{0\mu})^2 = \alpha P(x^\mu > 0) =: \alpha_{\rm eff} \tag{16}$$

where $\alpha_{\rm eff}$ is the percentage of exhausted degrees of freedom, i.e. if pattern 0 is as typical as the other random patterns $\mu = 1, \ldots, p$, one has to postulate

$$g = -\alpha_{\rm eff} = -\alpha \cdot P(t_2 < 1) \quad \overset{!}{\geqslant} (-1)\,. \tag{17}$$

With these results, using $E^\mu = \sum_\nu B^{\mu\nu} x^\nu$ and $L^2 = N^{-1} \vec{x}^T \overleftrightarrow{B} \vec{x}$, we can now calculate self-consistently all desired quantities by our Kuhn–Tucker cavity method.

2.4. *Self-consistency equations with the cavity method*

We proceed as in section 5 of [1], starting with the Kuhn–Tucker conditions and then using the probability density $w(x)$ for the embedding strengths:

$$\begin{aligned} L^2 &= \frac{1}{N} \sum_{\mu\nu} x^\mu B^{\mu\nu} x^\nu = \frac{1}{N} \sum_{\mu} x^\mu = \alpha \int \mathrm{d}x\, x w(x) \\ &= \frac{\alpha}{1+g} \int_{-\infty}^{1} \mathrm{d}t_2\, P(t_2)(1 - t_2) \\ &= \frac{2L\alpha}{1+g} \int_{-\infty}^{\kappa} \mathcal{D}\tilde{t}_2\, \Phi\left(\frac{R\tilde{t}_2}{\sqrt{1-R^2}}\right)(\kappa - \tilde{t}_2)\,. \end{aligned} \tag{18}$$

Here, in agreement with the considerations in the last subsection, we have written the embedding strength as $x = \Theta(1-t_2)(1-t_2)/(1+g)$. Furthermore, $\kappa L = 1$ and $\tilde{t}_2 := t_2/L$. Finally, replacing L by $1/\kappa$ and substituting g from (17), we get

$$1 + g = 2\alpha\kappa \int\limits_{-\infty}^{\kappa} \mathcal{D}\tilde{t}_2 \, \Phi\left(\frac{R\tilde{t}_2}{\sqrt{1-R^2}}\right)(\kappa - \tilde{t}_2)$$

or (19)

$$1 = 2\alpha \int\limits_{-\infty}^{\kappa} \mathcal{D}\tilde{t}_2 \, \Phi\left(\frac{R\tilde{t}_2}{\sqrt{1-R^2}}\right)(1 + \kappa(\kappa - \tilde{t}_2))$$

$$= \alpha \int\!\!\int_{\mathcal{S}} \mathcal{D}u_1 \, \mathcal{D}u_2 [1 + \kappa(\kappa - R|u_1| - \sqrt{1-R^2}u_2)] \, . \tag{20}$$

In equation (20) the integration region is

$$\mathcal{S} = \left\{(u_1, u_2) | R|u_1| + \sqrt{1-R^2}u_2 \leqslant \kappa \right\} . \tag{21}$$

From equation (20) one can derive a set of formal solutions $\alpha_{\mathrm{cav}}(R, \kappa)$, depending on R; however, of all these functions only one is relevant: namely, R has to fulfil a certain self-consistency requirement (see equation (23) below). Together, the two equations (20) and (23) fix the desired result $\alpha(R) := \alpha(R(\kappa), \kappa)$, which can be inverted to yield $R(\alpha)$.

Mathematically, the just-announced self-consistency equation (23) for the overlap R between student and teacher couplings can be derived as follows: a typical pattern, which generates a reduced presynaptic re-oriented field $\tilde{t}_2 := \tilde{E}/L < \kappa$ at the *student's* output neuron, requires an (enhanced) embedding strength $x = (\kappa - \tilde{t}_2)/(1+g)$. If it simultaneously generates the re-oriented pre-synaptic field t_1 at the teacher's output neuron, this leads to a contribution xt_1 in the direction of the fixed teacher's coupling vector. Summing up all these contributions leads to

$$\alpha \int_0^{\infty} \mathrm{d}t_1 \int_{-\infty}^{\kappa} \mathrm{d}\tilde{t}_2 \, \tilde{P}(t_1, \tilde{t}_2) t_1 \frac{\kappa - \tilde{t}_2}{1+g} = R \, . \tag{22}$$

Here $\tilde{P}(t_1, \tilde{t}_2)$ is obtained from $P(t_1, t_2)$ in (6) by formally putting $L = 1$ there.

Multiplying again with $(1+g)$, inserting (17), and integrating finally over t_1, we get the announced self-consistency equation complementing (20), namely

$$2\alpha \int\limits_{-\infty}^{\kappa} \mathcal{D}\tilde{t}_2 \left\{ \frac{\sqrt{1-R^2}}{\sqrt{2\pi}} \exp\left(\frac{-R^2\tilde{t}_2^2}{2(1-R^2)}\right)(\kappa - \tilde{t}_2) + R\Phi\left(\frac{R\tilde{t}_2}{\sqrt{1-R^2}}\right)(1 + \kappa\tilde{t}_2 - \tilde{t}_2^2) \right\} = R \, . \tag{23}$$

Together equations (20) and (23) determine $R(\alpha)$, and therefore $G(\alpha)$ as well, see equation (8). Examples will be given below in figure 1; those results are special cases $Q' = 2$ of the corresponding figures for the general Potts model case in the next section.

2.5. *Heuristic derivation of the replica-symmetric results*

We now give a short heuristic derivation of the set of formal solutions $\alpha_{\mathrm{RS}}(R, \kappa)$ derived from the *replica theory* under the assumption of replica symmetry, i.e. from the first self-consistency equation (21) in [4]. As we will see immediately, $\alpha_{\mathrm{RS}}(R, \kappa)$ differs from $\alpha_{\mathrm{cav}}(R, \kappa)$, except at the 'stationary value' $R = R_0(\kappa)$, where (23) is fulfilled with $\alpha = \alpha_{\mathrm{cav}}(R_0, \kappa)$. Moreover, this requirement for $R = R_0$, namely $\alpha_{\mathrm{RS}}(R, \kappa) = \alpha_{\mathrm{cav}}(R, \kappa)$,

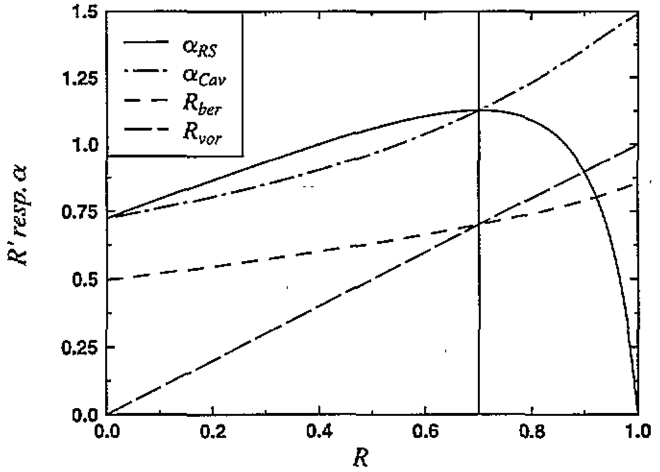

Figure 1. *Graphical representation of the calculation of the desired relation* $\alpha(R(\kappa),\kappa)$ between the reduced number $\alpha := p/N$ of correctly learned 'questions with answers' and the related stationary value $R(\kappa)$. Here R is the scalar product between teacher and student coupling unit-vectors. In our Kuhn–Tucker cavity approach, R is changed to $R'(R)$, if for given R and κ at first α is calculated from (11) and then R' from (14). In contrast, R remains unchanged ($R' = R$) by the replica approach. For $\kappa = 0.720037$ (leading to $R_0 := R(\kappa) = 0.7$ and $\alpha(R(\kappa),\kappa) = 1.12678$) the behaviour of our cavity method and of the replica-symmetric approach is compared: The two respective upper curves, $\alpha_{\text{cav}}(R,\kappa)$ and $\alpha_{\text{RS}}(R,\kappa)$, agree only for $R = 0$ and for the stationary value R_0, where $\alpha(R,\kappa)$ is maximal. This value $R_0 = R(\kappa)$ yields the desired monotonic relation $\alpha(R_0)$, and just there R is unchanged by our cavity learning, as shown in the lower part of the figure.

can be replaced by the postulate that for given κ the function $f(R) := \alpha_{\text{RS}}(R,\kappa)$, should have a maximum at R_0, which leads to the second self-consistency equation (21) in [4]. In contrast, $\alpha_{\text{cav}}(R,\kappa)$ is *not* maximal at $R = R_0$.

In [5], Griniasty has already developed a cavity theory of perceptron learning (but not of generalization), which is different from ours and completely *equivalent* to replica calculations in the replica-symmetric approximation. In his derivation, Griniasty concentrates on minimizing cost functions instead of using a learning algorithm. In section 5 of his paper he develops a simplified version (which, in our opinion, nevertheless catches the spirit of his method) to derive a certain constant of integration, namely, he shows that one can arrive at the replica results, if one makes the incorrect, but consistent assumptions that both the reaction field gx^0 and the correlations $B^{\mu\nu} = N^{-1}\sum_{k=1}^{N}\zeta_T^\mu\xi_k^\mu\xi_k^\nu\zeta_T^\nu$ for $\mu \neq \nu$ vanish.

With these two assumptions, in our case one can obtain the final embedding strength x_{RS}^μ (which corrects t_2, the field before training, whose probability distribution is given in (7)) by a simple AdaTron learning step (11). We then calculate the length of the coupling vector, which again has to be consistent. Since we now cannot determine R from the linear combinations of the patterns as in (22), we have to actively fix the overlap R with the teacher. This implies that now the relevant length of the coupling vector is not L^2 as in [1], but $L^2 \cdot (1 - R^2)$. Therefore, with the just-mentioned assumptions we get

$$L^2(1-R^2) = \frac{1}{N}\vec{x}_{\text{RS}}^T\,\overset{\leftrightarrow}{B}_{\text{RS}}\,\vec{x}_{\text{RS}} = \frac{1}{N}\sum_\mu (x_{\text{RS}}^\mu)^2 \qquad (24)$$

which implies

$$\begin{aligned}(1-R^2) &= \alpha_{\mathrm{RS}} \int_{-\infty}^{\kappa} \mathrm{d}\tilde{t}_2\, \tilde{P}(\tilde{t}_2)(\kappa-\tilde{t}_2)^2 \\ &= \alpha_{\mathrm{RS}} \int_{-\infty}^{\kappa} \mathcal{D}\tilde{t}_2\, \Phi\left(\frac{R\tilde{t}_2}{\sqrt{1-R^2}}\right)(\kappa-\tilde{t}_2)^2 \\ &= \alpha_{\mathrm{RS}} \int\!\!\int_S \mathcal{D}u_1\, \mathcal{D}u_2 (\kappa - R|u_1| - \sqrt{1-R^2}u_2)^2 \,. \end{aligned} \tag{25}$$

Equation (25), which determines our $\alpha_{\mathrm{RS}}(R,\kappa)$, is identical with the first self-consistency equation (21) in the replica calculation of [4]. The second self-consistency equation of (21) in [4] was already mentioned: among the possible values of R, the optimal student chooses that one, which maximizes the number of patterns which can be stored with the given stability. Together with (25), this implies

$$-2R = \alpha_{\mathrm{RS}} \frac{\partial}{\partial R} \int_{-\infty}^{\kappa} \mathrm{d}\tilde{t}_2 P(\tilde{t}_2)(\kappa-\tilde{t}_2)^2 \,. \tag{26}$$

Interestingly, instead of calculating R_0 from (20) and (23) (cavity approach), or from (25) and (26) (replica calculation), one can also combine the 'best of either set', namely (20) and (25), see below.

2.6. *Results for perceptrons with Ising neurons*

The different philosophies of the two approaches become clear, as we compare the results obtained with our Kuhn–Tucker cavity approach, i.e. equations (20) and (23), with those obtained with the 'replica formalism', i.e. (25) and (26) in figure 1.

As already mentioned, and as can be seen in figure 1, the results agree at the extremal overlap value $R = R_0$, which corresponds to the uniquely determined saddle point of the RS solution. The disagreement for $R \neq R_0$, however, does not imply contradictory results. Instead, what happens is that the extremal point is approached along different paths. In the lower part of figure 1, we present the function $R'(R)$. This is obtained, if for given R and κ, at first α_{cav} is calculated from (20) and then, with $\alpha = \alpha_{\mathrm{cav}}$, R' from (23). As we increase R the patterns become progressively easier to store and α_{cav} grows monotonically (see the chain curve in figure 1). Only for $R = R'$, however, is the solution valid.

We now discuss the full curve in figure 1, representing α_{RS}. As mentioned before, this is the storage capacity, if the student has a certain *fixed* overlap R with the teacher. For $R = 0$ the problem reduces to storing randomly oriented patterns, and α_{RS} is equal to α_{cav}. With increasing $R \to R_0$ the patterns become *easier* to store, i.e. α_{RS} increases, until at R_0 the maximum is obtained, where again $\alpha_{\mathrm{RS}} = \alpha_{\mathrm{cav}}$. Then, for $R > R_0$, the fixation of the overlap R with the teacher drastically *lowers* the number of patterns $p = N \cdot \alpha_{\mathrm{RS}}$, which can be stored, as $R \to 1$. Finally, for $R = 1$ no extensive number of patterns can be stored (i.e. $\alpha_{\mathrm{RS}} = 0$), because the teacher herself has stability $\kappa = 0$ for random patterns.

As stated above, for $R = R_0$ one can use either the cavity equations (20) and (23) or the RS equations (25) and (26); but an easier and numerically more accurate way is to combine the simplest equations from both methods, i.e. equations (20) and (25), into a single equation for R, namely

$$\int_{-\infty}^{\kappa} \mathcal{D}\tilde{t}_2 \, \Phi\left(\frac{R\tilde{t}_2}{\sqrt{1-R^2}}\right) [\tilde{t}_2^2 - (1+R^2)\kappa\tilde{t}_2 + R^2\kappa^2 - (1-R^2)] = 0 \, . \tag{27}$$

The loading parameter $\alpha(R)$ as a function of the overlap R is then obtained, e.g. from equation (20).

Most interesting is the behaviour for $R \to 1$, i.e. $\kappa \to 0$ or $\alpha \to \infty$. In this case, the Gaussian measure $\mathcal{D}\tilde{t}_2$ in (27) is practically stationary in the relevant region, so that the integrals over $\mathrm{d}\tilde{t}_2$ can be evaluated. With $u := \kappa/\sqrt{1-R^2}$, one then obtains the implicit equation $2\pi^{-1/2}(1-1/u^2)\exp\left(-u^2/2\right)+u\Phi(u) = 0$, which yields $u = 0.638\,833\,215\,8\ldots$.

Introducing this into (25) for the loading parameter α, one obtains for perceptrons with Ising neurons, with $c := 1.998\,046\,18\ldots$ asymptotically for $\alpha \to \infty$ the overlap

$$R(\alpha) \simeq 1 - \frac{1}{2}\frac{\pi^2}{c^2\alpha^2} = 1 - \alpha^{-2} \times 1.236\,114\,51\ldots \tag{28}$$

and with (8) the generalization probability

$$G(\alpha) \simeq 1 - \frac{1}{c\alpha} = 1 - \alpha^{-1} \times 0.500\,488\,93\ldots . \tag{29}$$

Finally, the optimal stability $\kappa(\alpha)$ is for given α

$$\kappa(\alpha) \simeq \frac{u\pi}{c\alpha} = \alpha^{-1} \times 1.004\,458\,132\ldots . \tag{30}$$

Further, for $\alpha \to \infty$, the asymptotic behaviour of g is that of $(-R(\alpha))$. Thus the results from our cavity approach agree completely with those of the replica calculation of [4].

3. The generalization ability of the Potts perceptron with optimal stability

What has been gained in the preceding section are two simple cavity methods, which can even be combined with some care, to calculate the generalization ability for perceptrons with Ising neurons and real couplings under the usual spherical constraints. The methods are (i) our Kuhn–Tucker cavity method of section 2.3, which (as we will see in a forthcoming paper) can also be applied to situations, where replica symmetry is broken (although in that case it is no longer exact), and (ii) the cavity approach of section 2.4 based on 'noise arguments' *à la Griniasty*, see [5], which is equivalent to the replica calculation for *non*-broken replica symmetry ('RS approach'). In the preceding section, we have demonstrated the similarities and differences between these two approaches, which are equivalent as long as the RS approach is correct. The essential point, however, is that both cavity approaches are easily applicable to more complicated models.

In the present section, using these cavity approaches, we will get new results, namely, the generalization ability of Potts perceptrons with general values Q and Q' of the number of states of the input and output neurons, respectively.

3.1. Basic equations for the Potts case

For Potts model perceptrons, the input neurons at a site k have Q different states (i.e. $n_k = 1, \ldots, Q$ below), which are vectors $\boldsymbol{m}_{n_k}$ with Q components,

$$m_{n_k}(s) := Q\delta_{s,n_k} - 1 \tag{31}$$

for $s = 1, \ldots, Q$. For the output neuron, the corresponding quantities are Q' and s', i.e. Q' and Q can be different. Then the couplings J_k also become more complicated, namely one gets $s' \times s$ matrices $J_k(s', s)$ for the student and $J_k^T(s', s)$ for the teacher, respectively, which are abbreviated as $\underline{J}$ (with norm $L^2 = |\underline{J}|^2$) and $\underline{J}^T$, respectively (with $|\underline{J}^T|^2 = 1$). Their mutual overlap $R = L^{-1}(\underline{J}^\dagger \underline{J}^T)$ is defined as

$$R := \frac{1}{L}\mathrm{tr}_{s,s'}\sum_{k=1}^{N}(J_k^\dagger J_k^T) = \frac{1}{L}\sum_{k=1}^{N}\sum_{s=1}^{Q}\sum_{s'=1}^{Q'} J_k(s', s) J_k^T(s', s) \tag{32}$$

where $\mathrm{tr}_{s,s'}$ means 'trace' and † means 'transposed'.

The sum over s resp. s' of the coupling matrices can be assumed to vanish for both perceptrons. These are the usual gauge conditions for Potts systems (see, e.g. (4) and (5) in [1]).

It has been shown in [1] that the re-oriented presynaptic field at the student's output neuron can be determined for a newly added random pattern by drawing the Q' components from a Gaussian distribution with average 0 and variance $L^2 Q/(Q' - 1)$ (for the teacher, $L = 1$). Furthermore, due to the gauge freedom, the component in $(1, \ldots, 1)$ direction in Q'-space is arbitrary as in [1]. At the teacher perceptron a random pattern μ generates the (non-re-oriented) presynaptic output field

$$h_T^\mu = \sum_k J_k^T \cdot \boldsymbol{m}_{n_k^\mu} \qquad \text{i.e.} \quad h_T^\mu(s') = \sum_k \sum_s J_k^T(s', s) m_{n_k}^\mu(s)\,. \tag{33}$$

Finally, the *re-oriented* presynaptic field $\boldsymbol{E}_T^\mu$ of the *teacher's* output neuron is

$$\boldsymbol{E}_T^\mu = \mathcal{P}^{n'^\mu - 1} \boldsymbol{h}_T^\mu \tag{34}$$

where $n'^\mu = \sec(\boldsymbol{h}_T^\mu)$. Here 'sec' is simply the function (defined already in [2]), which determines the output value $n' = 1, \ldots, Q'$ of the perceptron from the presynaptic field according to the phase space *section*, to which the field belongs. Similarly, $\mathcal{P}^{n'^\mu - 1}$ is the cyclical shift operator, which shifts the maximal component n'^μ to the first place, see, e.g. equation (17) of [1].

3.2. *Self-consistency equations for the Potts case*

In contrast, $\tilde{\boldsymbol{E}}$, the *re-oriented* field of the *student*, i.e. re-oriented with respect to the output of the teacher, can be constructed for a newly added random pattern $\mu = 0$ as in the preceding section by a decomposition of the coupling vector of the student perceptron into *two* perpendicular components, where the first component is parallel to the teacher's coupling vector and has the length $R\ L$, whereas the second one of length $L\sqrt{1 - R^2}$ is perpendicular to it. Thus we can write down the generalization probability $G(R)$ that for such a newly added random pattern (= 'question') the student produces the same output (= 'answer') as the teacher: With the integer $n_1' := \sec(\boldsymbol{u}_1)$ and the Kronecker symbol $\delta\{n_1, n_2\} := 1$ for $n_1 = n_2$, :=0 else, we get

$$G(R) = \iint \mathcal{D}\boldsymbol{u}_1\, \mathcal{D}\boldsymbol{u}_2\, \delta\left\{1, \sec\left(R\mathcal{P}^{1-n_1'}\boldsymbol{u}_1 + \sqrt{1 - R^2}\boldsymbol{u}_2\right)\right\} \tag{35}$$

where the Q' components of the vectors $\boldsymbol{u}_i$ are normally distributed.

Here again, learning is a prerequisite for generalization. Therefore, since the learning task for the student perceptron with optimal stability has already been treated in [1], we can be rather sketchy in the following. As in equation (9) above, there is again a representation with embedding strengths $\boldsymbol{x} := (x(1), \ldots, x(Q'))$, and again there exists the optimization problem equation (26) in [1], which leads to the Kuhn–Tucker conditions for the solution, i.e. equations (30)–(33) in [1], which can be solved by the same AdaTron learning algorithm for Potts perceptrons as described there.

But the subsequent calculation leading to the *response* of the couplings on a newly added pattern is now slightly more complicated than the corresponding calculation in [1]. (As in the preceding section, this response is necessary to fulfil the Kuhn–Tucker conditions for the new pattern and the already stored ones as well.) The complication is that now the *re-oriented field* $\tilde{\boldsymbol{t}}$ of the student (see below) must be decomposed into the already mentioned components parallel (resp. perpendicular) to the coupling vector of the teacher perceptron. For the reaction factor we again obtain $g = -\alpha\bar{k}/(Q' - 1)$, where $\bar{k}$ is the average of the number k of active directions (see equation (57) in [1]). However, the equivalent to the integral (79) in [1] is now the double integral

$$L^2 = \frac{\alpha}{\sqrt{Q(Q'-1)}}\frac{L}{1+g}\iint \mathcal{D}\boldsymbol{u}_1\,\mathcal{D}\boldsymbol{u}_2\,x_{\kappa'}\{\tilde{\boldsymbol{t}}\}(1) \tag{36}$$

with the abbreviation

$$\tilde{\boldsymbol{t}} := R\mathcal{P}^{1-n_1'}\boldsymbol{u}_1 + \sqrt{1-R^2}\boldsymbol{u}_2 \tag{37}$$

for the re-oriented field of the student. In equation (36) we have distinguished the (1)-component of the embedding vector as that one corresponding the 'recognition section' of the phase space, to which the 're-oriented field' must belong, and we have used $\tilde{\boldsymbol{t}} := \boldsymbol{t}\sqrt{(Q'-1)/(QL^2)}$, $x_{\kappa'}\{\tilde{\boldsymbol{t}}\}(1) := x\{\boldsymbol{t}\}(1)\sqrt{(Q'-1)/(QL^2)}$, and $c = \kappa L = 1$ for the quantity c appearing in the Kuhn–Tucker conditions and the Adatron process in [1], see, e.g. equations (29)–(39) there. If we then set $L = \sqrt{(Q'-1)/Q}$, we get $c = \kappa L = \kappa' := \kappa\sqrt{(Q'-1)/Q}$ as our *reduced stability measure*, and additionally $\boldsymbol{t} = \tilde{\boldsymbol{t}}$. The following steps are similar to those leading to (83) in [1] and lead to the result

$$1 = \frac{\alpha}{Q'-1}\iint \mathcal{D}\boldsymbol{u}_1\,\mathcal{D}\boldsymbol{u}_2\left(k\{\boldsymbol{t}\} + \kappa'\{\boldsymbol{t}\}\cdot\boldsymbol{x}_{\kappa'}\{\boldsymbol{t}\}\right) \tag{38}$$

which corresponds to (20) for the case of Ising neurons (see above). From this equation, $\alpha_{\mathrm{cav}}(R,\kappa)$ can be calculated. (The determination of $k\{\boldsymbol{t}\}$ and $\kappa'\{\boldsymbol{t}\}$ is described in [1].)

As in the preceding section, we need a second condition fixing $R = R_0(\kappa)$. One can use, for example, the self-consistency condition for the overlap

$$\begin{aligned} R &= \frac{1}{L}\mathrm{tr}_{s,s'}\sum_k(\boldsymbol{J}_k^\dagger\boldsymbol{J}_k^T) \\ &= \frac{1}{LNQ(Q-1)}\mathrm{tr}_{s,s'}\Big[\sum_k\sum_\mu(\mathcal{P}^{1-n_1'^\mu}\boldsymbol{x}^\mu\otimes\boldsymbol{m}_{n_k^\mu})^\dagger\boldsymbol{J}_k^T\Big] \\ &= \frac{1}{N(Q-1)\sqrt{Q(Q'-1)}}\sum_\mu\boldsymbol{x}_{\kappa'}^\mu\cdot\boldsymbol{E}_T^\mu \\ &= \frac{\alpha}{Q'-1}\frac{1}{1+g}\iint\mathcal{D}\boldsymbol{u}_1\,\mathcal{D}\boldsymbol{u}_2\,\boldsymbol{x}_{\kappa'}\{\tilde{\boldsymbol{t}}\}\cdot\mathcal{P}^{1-n_1'}\boldsymbol{u}_1\,. \end{aligned} \tag{39}$$

Here equation (20) of [1] has been used for the couplings $\boldsymbol{J}$ of the student perceptron with optimal stability; $\boldsymbol{E}_T^\mu$ is the re-oriented presynaptic output field of the teacher perceptron, see equation (34); finally, the generation of the presynaptic output fields of teacher and student from a Gaussian has been used, as described above. Performing similar steps as above, multiplying (39) with $1+g$ and inserting the expression for g from above, one thus obtains

$$R = \frac{\alpha}{Q'-1}\iint\mathcal{D}\boldsymbol{u}_1\,\mathcal{D}\boldsymbol{u}_2\big(Rk\{\boldsymbol{t}\} + \boldsymbol{x}_{\kappa'}\{\boldsymbol{t}\}\cdot\mathcal{P}^{1-n_1'}\boldsymbol{u}_1\big). \tag{40}$$

To conclude, we stress that (38) generalizes (20) for the determination of $\alpha_{\rm cav}(R,\kappa)$ from the Ising case to the Potts case, while (40) corresponds to the additional condition (23), i.e. the self-consistency for the critical overlap R. Both equations have been derived with the present Kuhn–Tucker cavity method.

On the other hand, the generalization of the 'noise arguments' *à la Griniasty* [5], is also possible and leads without difficulties to

$$1 - R^2 = \frac{\alpha}{Q'-1} \iint \mathcal{D}u_1\, \mathcal{D}u_2\, (x_{\kappa'}\{t\})^2 \,. \tag{41}$$

This equation generalizes (25) and must again be augmented by the condition that for given κ, $R = R_0$ should be chosen such that $\alpha_{\rm RS}(R,\kappa)$, as determined from (41), is maximal. As mentioned in [1] and sketched in [6], it can be shown by partial integrations that both sets of equations are equivalent, but only as long as the replica symmetry is guaranteed, since in case of replica-symmetry breaking our approach leads to different results compared with the RS approximation, in contrast to Griniasty's approach (see [5, 6]).

3.3. *Evaluation of the integrals and results for the Potts case*

The evaluation of the self-consistency equations for Potts perceptrons requires numerical results for the corresponding multi-dimensional integrals. For the special case $Q' = 2$, after a suitable rescaling from κ' to κ, we regain the results of section 2. The results for $Q' = 3$ are identical to those for *clock* perceptrons with $Q = 3$, which have been published in [9].

For general Q, the integrals can be evaluated quite accurately and efficiently by the following *Monte Carlo process*. Choose a value R between 0 and 1. Then generate the Q' components of the vector u_1 as independent random numbers drawn from a 'normal distribution', i.e. a Gaussian with average 0 and variance 1. The largest of these Q' components should be shifted to the first place, so that $\mathcal{P}^{1-n'_1}u_1$ represents the re-oriented presynaptic field generated by a random input vector at the teacher's output neuron. Then a second vector u_2 is generated, again with independently and normally distributed components, and the re-oriented presynaptic field $t = \tilde{t}$ at the student's output neuron is evaluated according to (37).

With u_1 and t, all relevant integrals can be evaluated. For the generalization ability one has to check for each event, whether the student's 'answer' agrees with that of the teacher. For $x_{\kappa'}\{t\}$ and $k\{t\}$ an AdaTron training step has to be performed with the scaled stability $c = \kappa'$, by which then the contribution to the integrals (38), (39) and (41) is determined.

This can be performed with little additional effort simultaneously for different values of κ'. If one knows the average a from (40) in [1] and the active directions for a certain κ', a can then be calculated for a larger κ' in one additional step. One only has to look whether for that κ' an additional active direction has to be added to the list. Thus the generation of new random numbers and the sorting of components is avoided.

Figures 2–4 present results (i) for the generalization probability $G(R)$ as a function of the overlap R, (ii) for $R(\alpha)$ (where α is the loading parameter), and (iii) for $G(\alpha)$, for the values $Q' = 2$, 3, 4, 5, 6, 10, 100 and 1000. To produce the curves, we generated 5×10^8 random numbers for every point in the plot. For given R, at first a κ' was guessed, and the integrals were evaluated for 60 different κ' in the vicinity of the first guess, usually with full accuracy only in a second iteration. We have found in this way within our statistical accuracy that for all points for given R at the intersection of $\alpha_{\rm cav}$ from (38) with $\alpha_{\rm RS}$ from (41) also the self-consistency condition (39) was fulfilled. With condition (39), the desired values $\kappa'(R)$, and thus $\alpha(R,\kappa'(R))$, can be determined more accurately and can better be interpolated from neighbouring values.

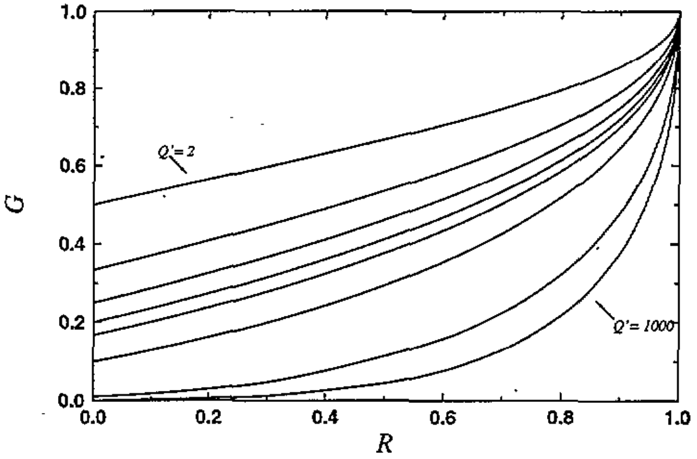

Figure 2. For perceptrons with Potts Q'-state output neurons, the generalization ability $G(\alpha, Q')$ is plotted as a function of the overlap R for $Q' = 2, 3, 4, 5, 6, 10, 100, 1000$. Particular values are $G(0, Q') = 1/Q'$ and $G(1, Q') = 1$. The full curves have been determined for $Q' \geqslant 4$ by the Monte Carlo integration algorithm described in the text.

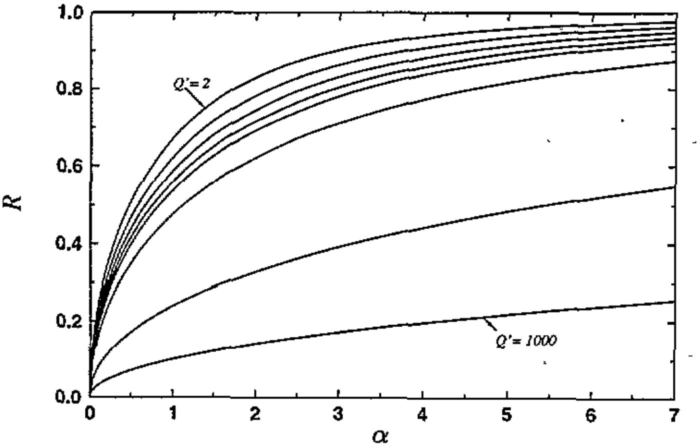

Figure 3. The stationary overlap $R = R_0(\alpha, Q')$ of the student's and the teacher's coupling vectors is plotted against the reduced number $\alpha = p/N$ of learned 'questions with answers' for $Q' = 2, 3, 4, 5, 6, 10, 100, 1000$.

It is interesting to compare the results of the figures 2–4 with those obtained for the *clock* model in [9]. For the clock perceptron, as a function of α, in the limit of $Q' \to \infty$ one gets a second-order 'phase transition' at $\alpha = 2$ from a phase with no generalization for $\alpha \leqslant 2$ to a 'generalizing phase' at $\alpha > 2$. More precisely, one has for $\alpha > 2$ the exact inequality $G(\alpha) \geqslant 1 - 2/\alpha$ for all values of Q'. In contrast, for the present case of the Potts perceptron the generalization ability vanishes at $Q' = \infty$ for all finite values of α, and not only for $\alpha \leqslant 2$ as for the clock perceptron. The reason for this different behaviour is that for the Potts model, there are many more degrees of freedom to be fixed.

Thus, when a problem allows the application of both Potts and clock perceptrons with

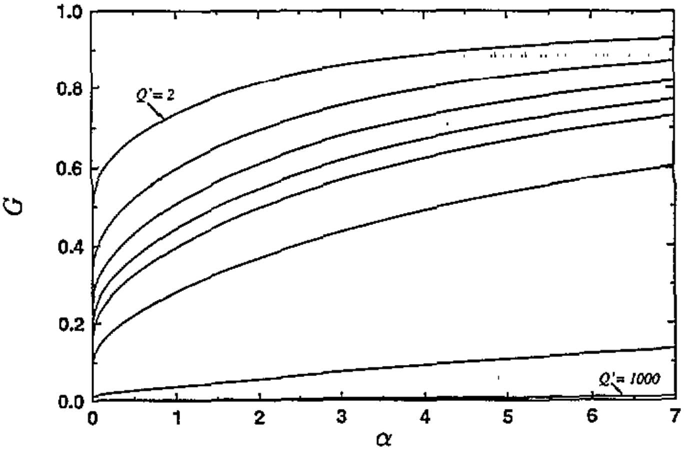

Figure 4. The generalization ability $G(\alpha, Q')$ of the student is plotted as a function of the reduced number $\alpha = p/N$ of learned 'questions with answers' for $Q' = 2, 3, 4, 5, 6, 10, 100, 1000$. For $\alpha = 0$ it is $G = 1/Q'$; this corresponds to random guessing. For $\alpha \to \infty$ one confirms the suggestion $G(\alpha, Q') \simeq 1 - (Q' - 1) \times 0.500\,488\,93\alpha^{-1}$.

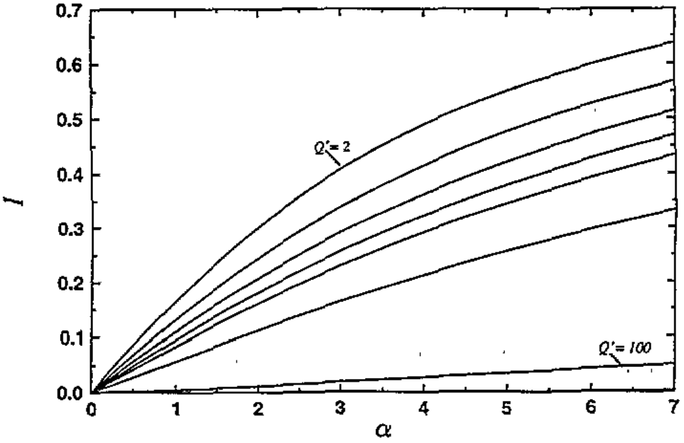

Figure 5. The relative information gain $I := \Delta I_{\mathrm{rel}}$, see the text, is presented as a function of the reduced number $\alpha := p/N$ of learned 'questions with answers' for $Q' = 2, 3, 4, 5, 6, 10, 100$.

large Q', it is more advantageous, concerning the generalization properties, to choose the Clock model.

For $Q' = 2$ the behaviour of $G(\alpha, Q')$ for $\alpha \to \infty$, i.e. $R \to 1$, is contained in equation. (29); for the Clock model with $Q' = 3$ one has from [9] the asymptotic result for $\alpha \to \infty$: $G(\alpha, Q' = 3) \simeq 1 - 2 \times 0.500\,488\,03\alpha^{-1}$, i.e. for $Q' = 3$ the prefactor in front of α^{-1} exceeds that for $Q' = 2$ by a factor of 2. This leads to the suggestion that for the *general case* the asymptotic behaviour for $\alpha \to \infty$ is

$$G(\alpha, Q') \simeq 1 - (Q' - 1) \times 0.500\,488\,93\alpha^{-1}\,. \tag{42}$$

Within the numerical accuracy, this result is in fact observed in our calculations.

A further quantity, which allows a comparison of Potts perceptrons with different Q', is the *information gain*: before training, all 'answers' $s' = 1, \ldots, Q'$ on a 'question' are equally probable, i.e. the information entropy is $I_0 = \ln Q'$, whereas afterwards it is $I(R) = -\sum_{s'=1}^{Q'} P_{s'}(R) \ln P_{s'}(R)$. Thus, the information gain is

$$\Delta I(R) = \log Q' + \sum_{s'=1}^{Q'} P_{s'}(R) \cdot \ln P_{s'}(R)$$
$$= \log Q' + G \ln G + (1-G) \ln\left(\frac{1-G}{Q'-1}\right). \tag{43}$$

In figure 5, the *relative information gain* $\Delta I_{\mathrm{rel}}(R) := \Delta I(R)/\ln Q'$ is plotted against α for $Q' = 2, 3, 4, 6, 10, 100$. Here again, if one compares with the Clock perceptron, a qualitatively different behaviour is observed (cf figure 9 in [9]).

4. Conclusions

In this paper, within the framework of statistical physics, we have derived a 'Kuhn–Tucker cavity method' for the generalization ability of perceptrons which have been trained to optimal stability. The approach simplifies the calculation of the generalization probability, compared with the technically more complicated replica calculation, to which our cavity method is equivalent as long as replica symmetry holds. For the present applications this is the case, i.e. here our cavity approach is exact.

At first, we have exemplified our method for the case of perceptrons with *Ising neurons*, where we can compare directly with the known results from the replica approach. Then we applied it to the more complicated generalization properties of perceptrons with *Potts neurons*, where we obtained new results, since there no replica calculation exists and would be extremely complicated. In both cases, we calculated the generalization ability $G(\alpha)$, i.e. the probability for the event that a student perceptron, after having learned 'with optimal stability' (e.g. with the AdaTron process of [2, 8]) to give the same answers as the teacher to a certain 'training set' of $p = \alpha N$ questions ξ^{μ} for $\mu = 1, \ldots, p$, gives again the 'correct answer' (i.e. that of the teacher) to an additional random question ξ^0.

The essence of our method is the *reaction strength* $-gx_0$, which acts against the *trial implementation* of ξ^0 with a 'bare' embedding strength x^0. This 'bare embedding' x^0 would correspond to a simple sequential AdaTron learning step which does not take into account the $p = \alpha N$ 'questions with correct answers' already stored. The *reaction* is a necessary consequence of the fact that those of the already 'stored patterns' ξ^{μ}, with $\mu = 1, \ldots, p$ $(= \alpha N)$, which are at the limit of stability, are perturbed. To counteract this perturbation, the embedding strengths x^{μ} must be changed by a certain δx^{μ}, which in turn acts against the original attempt to store pattern 0. This necessitates an *enhancement* of x^0 by the factor $1/(1+g)$. At the limit $\kappa \to 0$, where all coupling degrees of freedom are fixed, one has for $\alpha \to \infty$ asymptotically $g(\alpha) \simeq -R(\alpha) \to -1$.

In section 2, we first showed the equivalence of this approach with the known exact replica-symmetric solution (see [4]) for the particular case of Ising neurons, i.e. $Q = Q' = 2$ with real-valued couplings fulfilling the usual spherical constraint. Precisely, we found that both approaches are equivalent just at that value $R = R_0$ of the overlap $L^{-1}(J^{\dagger} J^{T})$ of student and teacher coupling vector, which leads to the maximal value of α for given κ. We even profited from combining different equations from the two approaches.

We also showed that the *replica* approach to generalization, as long as replica symmetry is not broken, is equivalent to a different cavity theory for generalization, which for the *learning* paradigm would reduce to the simplified cavity theory of Griniasty, see [5].

However, as soon as replica symmetry is broken, the 'Griniasty-cavity approach' differs from our Kuhn–Tucker cavity method, see [6, 10]. (In the present paper, where replica symmetry prevails, both approaches are equivalent. However, Griniasty concentrates on minimization of a cost function, whereas our Kuhn–Tucker cavity approach concentrates on learning algorithms.)

In section 3, both cavity methods have been extended and combined to the case of *Potts* perceptrons with general values of Q and Q', and by a careful and very accurate Monte Carlo implementation of the multi-dimensional integrations involved, we obtained the desired results for $G(\alpha, Q')$ and for the related information gain $\Delta I(\alpha, Q')$. However, the present results for the *Potts models* are simply monotonically increasing in $1/Q'$ and α without any kinks in the derivatives, and $G(\alpha, Q')$ vanishes for $Q' = \infty$ for *all* values of α, whereas for the *Clock* model case it has been shown in [9] that for $Q' = \infty$ there is a second-order phase transition from a 'non-generalizing phase' with $G(\alpha) = 0$ for $\alpha \leqslant 2$ to a 'generalizing phase' with $G(\alpha) \geqslant 1 - (2/\alpha)$ at $\alpha > 2$.

For the present case of Potts model output neurons with finite Q', we have obtained for $\alpha \to \infty$ the asymptotic result $G(\alpha, Q') \simeq 1 - (Q' - 1) \times 0.500\,488\,93\alpha^{-1}$.

As already mentioned, we have found that our Kuhn–Tucker theory can also be used for problems, where replica symmetry is broken, and in this case it yields results which differ from the replica-symmetric calculations. This will be discussed in a following paper, [10].

Acknowledgments

The authors would like to thank G Pöppel, M Opper, A Scharnagl and B Schottky for valuable discussions. The computations have been performed at the computing centers of the university at Regensburg, the LRZ in Munich and the HLRZ in Jülich.

References

[1] Gerl F and Krey U 1994 *J. Phys. A: Math. Gen.* **27** 7353

[2] Gerl F, Bauer K and Krey U 1992 *Z. Phys.* B **88** 339

[3] Watkin T L H, Rau A and Biehl M 1993 *Rev. Mod. Phys.* **65** 65

[4] Opper M, Kinzel W, Kleinz J and Nehl R 1990 *J. Phys. A: Math. Gen.* **23** 23

[5] Griniasty M 1993 *Phys. Rev.* E **47** 4496

[6] Gerl F *PhD Thesis* University of Regensburg

[7] Watkin T L H, Rau A, Bollé D and van Mourik J 1992 *J. Physique I* **2** 167

[8] Anlauf J K and Biehl M 1989 *Europhys. Lett.* **10** 687

[9] Schottky B, Gerl F and Krey U 1995 *Z. Phys.* B at press

[10] Gerl F and Krey U to be published