# Replica Symmetry Breaking and the Kuhn-Tucker Cavity Method in simple and multilayer Perceptrons *

F. GERL[1] and U. KREY[2]

[1] Institut für Theoretische Physik der Universität Göttingen, Bunsenstr. 9, D-30373 Göttingen

[2] Institut für Physik II der Universität Regensburg, Universitätsstr. 31, D-93040 F.R.G.

July 7, 1995

**Abstract**

Within a Kuhn-Tucker cavity method introduced in a former paper, we study optimal stability learning for situations, where in the replica formalism the replica symmetry can be broken, namely
(i) the case of a simple perceptron above the critical loading, and
(ii) the case of two-layer AND-perceptrons.
We find that the deviation of our cavity solution from the replica symmetric one is a clear indication of the necessity of replica breaking in both cases. However, in case (i) our cavity method seems to overestimate considerably the loading, which is necessary to minimize the errors.

## 1 Introduction

In a recent paper, [1], we have treated the learning problem for general Hopfield perceptrons with $Q$- and $Q'$ state Potts model input and output neurons using a general method based on a cavity formalism. In

*based on the PhD thesis of F. Gerl, Regensburg 1994 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, a number of exact results for learning with maximal stability has been obtained in [1] for Potts model networks, and in [2] for the Clock model case.

In the present paper, we extend our method to cases where in the replica formalism the replica symmetry is broken, namely (i) to perceptrons with Ising neurons above the critical loading and (ii) to two-layer AND perceptrons.

## 2 Simple perceptrons above the critical loading

We use the definitions as described in [1]: Our perceptron has $N$ input neurons $k = 1, ..., N$, whose possible states are $s_k = \pm 1$, and one output neuron $s'$, also with possible states $\pm 1$. The couplings leading from the input neurons to the output neuron are assumed to be real numbers, which are collected into the coupling vector $\underline{J} := (J_1, ..., J_N)$, the length of which is fixed (see below). The relation between input and output is

$$s' = \text{sign}\left(\sum_{k=1}^{N} J_k\, s_k\right) \tag{1}$$

which can be considered as a binary classification of the possible inputs, or as *answers* on *questions*.

Now we assume that there is a training set of $p$ input-output pairs, where the inputs are $\underline{\xi}^\mu := (\xi_1^\mu, ..., \xi_N^\mu)$ for $\mu = 1, ..., p$, and the corresponding desired outputs are $\zeta^\mu$. Here and in the following, unless otherwise stated, we assume that all input components and the outputs are independent random numbers, which take the values $\pm 1$ with equal probability.

The optimization task, which the perceptron then has to fulfill by adaptation of the couplings, is the minimization of a Hamiltonian, which performs a weighted count of bad classifications (see below) of the training examples. Among these Hamiltonians are that of Gardner and Derrida, [3], which simply counts the bad classifications, namely

$$\mathcal{H} := \sum_\mu V(E^\mu) := \sum_\mu \theta(\kappa - (E^\mu/L))\,. \tag{2}$$

Here $\theta(x) = 1$ for $x > 0$, $= 0$ else, and $E^\mu$ is defined as usual as the "oriented field"

$$E^\mu = \zeta^\mu \sum_{k=1}^{N} J_k \xi_k^\mu \,, \tag{3}$$

while $L = |\underline{J}|$. Bad classifications are those, where $E^\mu/L$ is $< \kappa$, i.e. for $\kappa > 0$ they are not necessarily wrong but lack a prescribed amount of *stability*, which is measured by $\kappa$.

There are $p =: \alpha \cdot N$ such "questions with prescribed answers", i.e. $\mu = 1, ..., p$, and it is assumed that the *loading parameter* $\alpha$ remains finite, while $N \to \infty$ and $p \to \infty$. Furthermore, the stability parameter $\kappa$ is to be maximized, if error-free classification (i.e. $\mathcal{H} = 0$) is possible.

Gardner and Derrida, [3], evaluated not only the number of errors $\mathcal{H}$ above the critical loading $\alpha_c(\kappa)$, where for positive $\kappa$ error-free classification is no longer possible, but they also tried to evaluate the so-called *Almeida-Thouless*-line $\alpha_{AT}(\kappa)$, above which within the replica formalism replica symmetry breaking (RSB) is necessary, and found surprisingly that $\alpha_{AT} > \alpha_c$ for $\kappa > 0$. However, this was due to a subtle integration error recently discovered by Bouten, [4], who has proven that $\alpha_{AT}(\kappa) = \alpha_c(\kappa)$, as expected. Bouten also showed that replica symmetry is always broken, if the distribution of local fields possesses a gap. So these results, on which we will comment later, provide a test for our cavity approach.

Other Hamiltonians, on which we comment at a later stage, are (see [5, 6]) the Perceptron function and the AdaTron function with

$$V(E^\mu) = [\kappa - (E^\mu/L)]^x \cdot \theta[\kappa - (E^\mu/L)], \tag{4}$$

with $x = 1$ and $x = 2$, respectively.

### 2.1 Representation of optimal couplings

Instead of minimizing the weighted error rate $f(\alpha, \kappa) := \mathcal{H}/p$ one can of course also maximize $\kappa(\alpha, f)$ for the given training set. Since we are above the critical loading $\alpha_c(\kappa)$, $f \cdot p$ of the training examples are badly classified. There is however an exponentially large number $\mathcal{N}$ of partitions of the training set into a "good fraction" and a "waste fraction of size $(1-f)\cdot p$ and $f\cdot p$, respectively, from which one has to choose the optimal one. Namely, $\mathcal{N} \simeq \exp(c\cdot p)$, with $c = -f \ln f - (1-f) \ln(1-f)$. We nevertheless assume, that every one of these combinations has been trained for optimal stability, e. g. with the AdaTron algorithm [7]. The optimal perceptron with error rate $f$ is then given by the partition leading to maximal $\kappa$.

The couplings of a perceptron trained for optimal stability can always be expressed in the form [7]

$$J_k = \frac{1}{N} \sum_{\mu \in \{(1-f)p\}} x^\mu \zeta^\mu \xi_k^\mu \tag{5}$$

with the so called "embedding strengths" $x^\mu$ of patterns which do not belong to the $fp$ badly classified patterns. As can be shown using Lagrangian multipliers [7, 1], these embedding strengths have to fulfill 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

$$\text{either} \quad (x^\mu > 0 \quad \text{and} \quad E^\mu = 1) \quad \text{or} \quad (x^\mu = 0 \quad \text{and} \quad E^\mu > 1)\,. \tag{6}$$

In fact the AdaTron algorithm (without overrelaxation)

$$x^\mu = \max(x^\mu, 1 - E^\mu) \qquad \text{(sequentially or in parallel)} \tag{7}$$

simply fixes the $x^\mu$ repeatedly to values which fulfill (6). If it converges the conditions are necessarily obeyed.

Using the "oriented correlation" matrix

$$B^{\mu\nu} = \zeta^\mu \zeta^\nu \sum_{k=1}^{N} N^{-1} \xi_k^\mu \xi_k^\nu\,. \tag{8}$$

and the definitions (3) and(5), we can write for the oriented field $E^\mu$

$$E^\mu = \sum_\nu B^{\mu\nu} x^\nu\ . \tag{9}$$

With the Kuhn-Tucker conditions we have finally

$$L^2 \quad = \quad \sum_{\mu,\nu} x^\mu B^{\mu\nu} x^\nu = \frac{1}{N}\sum_\mu x^\mu \tag{10}$$

As in [1], we now add a new "question with desired answer" $(\underline{\xi}^0, \zeta^0)$ to the training set, assuming one simple groundstate. We note that the distribution of the oriented fields $E^0$ acting on it, before any further adaptation has been performed, is Gaussian with average 0 and variance $|\underline{J}|^2 = L^2 = \kappa^{-2}$. Of course the error rate has to remain constant. As will be seen below, the best strategy is, to give up and add the new pattern to the "waste fraction of the training set", if the oriented field $E^0$ acting on $\underline{\xi}^0$ is smaller than a certain number $Z < 1$. For self-consistency, $Z$ is of course determined by

$$f = \int_{-\infty}^{Z} \mathrm{d}E^0 \, \frac{\exp[-(E^0)^2/2L^2]}{\sqrt{2\pi L^2}} = \int_{-\infty}^{z} \mathrm{d}t \frac{\exp(-t^2/2)}{\sqrt{2\pi}} \,, \tag{11}$$

where $t := E^0/L$ and $z := Z/L$. On the other hand, if $E^0$ is $> 1$, then it is not necessary to embed $\underline{\xi}^0$ in the couplings, since it can be added to the set of those correctly classified training patterns, which need no explicit embedding (see [8]) by the Adatron algorithm, [7]. Thus, only for $Z < E^0 \leq 1$, i.e. $z < t \leq \kappa$, the new pattern must be embedded in the couplings, and the implementation strength $x^\mu$ of the other patterns must be corrected by $\delta x^\mu$ (see below), to compensate for the influence of the new pattern. We proceed as in [1]:

As we have just seen, the AdaTron algorithm would try to embed $\underline{\xi}^0$, if necessary, with the "bare" embedding $x^0 = 1 - E^0$, and the other patterns have to respond by $\delta x^\mu = -B^{\mu 0} x^0$ to the disturbation by $x^0$, because the Kuhn-Tucker conditions still have to be fulfilled. Thus, the Onsager reaction cavity field acting on the new pattern is $\delta x^0 = g\, x^0$, with

$$g = -\sum_{\mu>0\ (x^\mu>0)} |B^{0\mu}|^2 =: -\alpha_{\mathrm{eff}} \equiv -\alpha P(Z < E^0 < 1) = -\alpha \int_z^\kappa \mathcal{D}t \,, \tag{12}$$

where $\mathcal{D}t = (2\pi)^{-1/2}\exp(-t^2/2)\mathrm{d}t$. This derivation, and the following reasoning, are as in section 5 in [1]. In particular, we have assumed as in [1] that $x^\mu$ and the $B^{\mu\nu}$ are statistically independent to the necessary accuracy, i.e. $\sim 1/\sqrt{N}$.

With our approach, the final embedding strength for $\underline{\xi}^0$ is given by $(1 - E^0)/(1 + g)$. Furthermore, we now identify the distribution of embedding strengths of all patterns with that of pattern $\underline{\xi}^0$. Putting the Gaussian distribution of the field $E^0$ before training into (10) gives

$$L^2 \;=\; \alpha L \int_z^\kappa \mathcal{D}t \frac{\kappa - t}{1+g}\,. \tag{13}$$

With $L = \kappa^{-1}$, after multiplication with $1 + g$, this finally leads to

$$\begin{aligned} 1 \;&=\; \alpha \int_z^\kappa \mathcal{D}t + \alpha\kappa \int_z^\kappa \mathcal{D}t\,(\kappa - t) & (14) \\ &=\; \alpha\,(1+\kappa^2) \int_z^\kappa \mathcal{D}t + \frac{\alpha\kappa}{\sqrt{2\pi}}(e^{-\kappa^2/2} - e^{-z^2/2})\,. & (15) \end{aligned}$$

This result will now be compared with the cavity approach of Griniasty, [9], which is equivalent to the replica formalism in the replica-symmetric approximation, see [9] and [1], so that we use the suffix "RS" for the corresponding results. We have found in [1] that below $\alpha_c$, where replica-symmetry is exact, our Kuhn-Tucker cavity approach and the approach of Griniasty agree; however in the present situation, where $\alpha$ is $> \alpha_c$, they disagree:

We remind that, when Griniasty derives the constant of integration in [9], the reaction factor $g$ is assumed to vanish, while at the same time also the $B^{\mu\nu}$ with $\mu \neq \nu$ are neglected. Thus we have instead of eqn. (15):

$$\begin{aligned} 1 \;&=\; \frac{1}{N}\,\vec{x}_{\rm RS}^{\,T}\,\overleftrightarrow{B}_{\rm RS}\,\vec{x}_{\rm RS} \;=\; \frac{1}{N}\sum_\mu (x^\mu_{\rm RS})^2 = \alpha_{\rm RS}\int_z^\kappa \mathcal{D}t\,(\kappa - t)^2 \\ &=\; \alpha_{\rm RS}(1+\kappa^2)\int_z^\kappa \mathcal{D}t \;+\; \frac{\alpha_{\rm RS}\kappa}{\sqrt{2\pi}}(e^{-\kappa^2/2} - e^{-z^2/2}) + \underbrace{\frac{\alpha_{\rm RS}}{\sqrt{2\pi}}(z-\kappa)e^{-z^2/2}}_{=:M}\,. \quad (16) \end{aligned}$$

In eqn. (16) we have the result of Gardner and Derrida [3, 5] from the replica-symmetric approximation, which is identical with that obtained by Griniasty in [9] by his cavity method; the expression abbreviated by $M$ in eqn. (16) yields the difference between $\alpha_{\mathrm{Cav}}$ and $\alpha_{\mathrm{RS}}$. Because of $z < \kappa$, it is $M \leq 0$ and therefore $\alpha_{\mathrm{RS}} \geq \alpha_{\mathrm{Cav}}$. For $z \to -\infty$, i.e. for $f \to 0$, one has $M = 0$, and thus for $\alpha \leq \alpha_c$ it is $\alpha_{\mathrm{RS}} = \alpha_{\mathrm{Cav}}$, as already mentioned.

In Fig. 1, for error rates of $f = 0.2$ and $0.02$, the learning capacities $\alpha(\kappa, f) = \alpha_{\mathrm{Cav}}$ and $\alpha_{\mathrm{RS}}$, respectively, as obtained from our Kuhn-Tucker cavity theory, eqn. (14), and with the replica-symmetric approximation (16) respectively, are plotted against the stability $\kappa$. Obviously, the learning capacity obtained with the Kuhn-Tucker cavity theory is lower, particularly at small values of $\alpha$, i.e. for $\kappa = 0$ and $f = 0.2$, $\alpha_{\mathrm{Cav}}$ is only 10/3, whereas $\alpha_{\mathrm{RS}}$ is 15.53. This means at the same time that for given $\kappa$ and $\alpha$, our error rate $f$ would be larger than that obtained with eqn. (16). This is a strong hint that the assumption of a unique optimum for the distribution of those patterns, which are put to waste, is wrong. Therefore, one cannot say in advance, which approach is better: Rather what has been gained is the following: Since both approaches agree below $\alpha_c(\kappa, f = 0)$, but not for $f > 0$, we can use the different results as a criterion for the necessity of replica-symmetry breaking for $f > 0$, which agrees with the recent rigorous proof of [4].

### 2.2 Comparison with a One-Step-RSB calculation

Majer *et al.*, [6], have performed a calculation above $\alpha_c(\kappa, f = 0)$. From eqn. (2), they obtained the following result for the free energy $\langle f_{\mathrm{E}} \rangle = -\alpha f$:

$$\begin{aligned} \langle f_{\mathrm{E}} \rangle \;&=\; \min_{x, q_0, w} \Big\{ \frac{q_0}{2x(1 + w\Delta q)} + \frac{\log(1 + w\Delta q)}{2wx} \\ &\quad + \frac{\alpha}{wx} \int \mathcal{D}z_0 \, \ln \Big[ \int_{\frac{A-2x}{\sqrt{\Delta q}}}^{\frac{A}{\sqrt{\Delta q}}} \mathcal{D}z_1 \exp\Big\{ -\frac{w}{2}(A - z_1\sqrt{\Delta q}\,)^2 \Big\} \\ &\quad + \exp\{-wx\}\Phi(\frac{A - 2x}{\sqrt{\Delta q}}) + \Phi(-\frac{A}{\sqrt{\Delta q}}) \Big] \Big\}, \end{aligned} \qquad (17)$$

with $A = \kappa - z_0\sqrt{q_0}$ and $\Delta q = 1 - q_0$. The parameters $q_0$, $w$ and $x$ have to be chosen such that the number of errors is maximized. In the limit $q_0 \to 0$ one regains the RS result.

In Fig. 2 the stability $\kappa$ is plotted against the error rate $f$ for three different values of $\alpha$. Obviously there is only a small difference between the results of the RS and the 1-step RSB calculation, in contrast to our cavity results, which differ considerably from both replica calculations, i.e. with our estimate, the stability increases much more slowly with increasing $\alpha > \alpha_c$.

For $f \ll 1$, we can compare these different estimates with simple simulations: For $\alpha < 2$ one can train perceptrons to optimal stability by means of the AdaTron algorithm. If one then skips the pattern with the largest embedding strength, i.e. the one which was most difficult to store, and re-learns the remaining patterns, one gets an enhanced stability, which agrees within the error limit with the replica calculation, and not with the cavity results, as we found in the simulations.

Thus, for $f > 0$, our Kuhn-Tucker cavity method yields a non-sufficient approximation for $\kappa(\alpha, f)$. In the following we try to find the reasons for the discrepancy and to estimate at the same time the quality of the 1-step RSB calculation. For this purpose, we need the distribution of the oriented fields $t$, which according to Majer et al., [6], is

$$\varrho(t) = \int \mathcal{D}z_0 \frac{\int \mathcal{D}z_1 \exp\left\{-wx\Big(V(\lambda_0) + \frac{(\lambda_0 - z_0\sqrt{q_0} - z_1\sqrt{\Delta q})^2}{2x}\Big)\right\} \delta(t - \lambda_0)}{\int \mathcal{D}z_1 \exp\left\{-wx\Big(V(\lambda_0) + \frac{(\lambda_0 - z_0\sqrt{q_0} - z_1\sqrt{\Delta q})^2}{2x}\Big)\right\}} . \tag{18}$$

where $V$ has been defined for the Gardner-Derrida rule in (2), and where $\lambda_0$ has to be chosen such that the exponent is minimized.

The parameters $q_0$, $w$ and $x$ are determined for given $\alpha$ and $\kappa$ from eqn. (17). With eqn. (18) the probability distribution of the local fields can be determined; in particular a $\delta$-peak at $t = 1$ is obtained, which represents the patterns, which are embedded explicitly by the training. However, our interest is in the field distribution *before* the additional training.

This distribution can be obtained by extending Griniasty's interpretation from the RS results in [9] to the present case.

According to [9], after addition of the pattern $\underline{\xi}^0$, the quantities $z_0\sqrt{q_0} - z_1\sqrt{\Delta q}$ and $\lambda_0$ are the local oriented fields before and after the additional training, respectively.

The exact value of $\lambda_0$ results from a compromise between the increase of the energy $V(\lambda_0)$ for patterns, which are badly classified by the perceptron, and the term $(\lambda_0 - z_0\sqrt{q_0} - z_1\sqrt{\Delta q})^2/2x$ representing the increase in energy of the $(1-f)p$ patterns, which had been stored before the addition of the new pattern.

Thus we can determine the field before the corrections by simply eliminating $V(\lambda_0)$ and $\lambda_0$ in (18): With the abbreviations $u_0 = z_0\sqrt{q_0}$ and $u_1 = z_1\sqrt{\Delta q}$, the integrand in eqn. (18) is

$$\begin{aligned}
&\frac{1}{\mathcal{N}}\int \mathcal{D}z_1\Big[\exp\left\{-\frac{w}{2}(\kappa - u_0 - u_1)^2\right\}\theta(\kappa - u_0 - u_1)\theta(u_0 + u_1 - (\kappa - \sqrt{2x})) \\
&\quad + \exp(-wx)\,\theta(\kappa - \sqrt{2x} - (u_0+u_1)) \; + \; \theta(u_0 + u_1 - \kappa)\Big]\delta(t - (u_0+u_1)) \\
= \; & \frac{1}{\mathcal{N}}\exp\left\{\frac{(t-u_0)^2}{2\Delta q}\right\}\Big[\exp\{-\frac{w}{2}(\kappa - t)^2\}\theta(\kappa - t)\theta(t - (\kappa - \sqrt{2x})) \\
&\qquad + \exp\{-wx\}\theta(\kappa - \sqrt{2x} - t) \; + \; \theta(t-\kappa)\Big] \quad .
\end{aligned} \tag{19}$$

Here the first term describes the patterns, which have been successfully embedded (i.e. with $V = 0$), the second term those, which are badly classified and put to the waste (i.e. with $V = 1$ and $\lambda_0 - z_0\sqrt{q_0} - z_1\sqrt{q_1} = 0$), and the final term the patterns which are stored without embedding. The denominator $\mathcal{N}$ in (19) is the integral over all fields $t$ in (19) and serves for normalization. Thus, we have no longer a Gaussian field-distribution before learning.

At the end of this section we make a proposal how these formulae can be interpreted and how one can gain a cavity approach for RSB.

In Fig. 3, for $\kappa = 1$ and $\alpha = 1$, the local fields are presented for the three approximations mentioned, namely for our cavity method, and for the RS and RSB1 approximations. One can see that in RSB1, the distribution of the local fields before learning the additional pattern, although strongly non-Gaussian, is still everywhere continuous, and for $t = \kappa$ even continuously differentiable.

Comparing the results of our cavity theory with the RS approximation, one can see that for increasing $\alpha > \alpha_c$

- on the one hand, the *error rate* $f$ obtained with the RS theory is much better than that obtained with our cavity method (if the error rate obtained with RSB1 is considered as target approximation), see Fig. 2; whereas

- on the other hand, the value of the *limiting negative field value* $t := t_0 = z/\kappa$ ($\approx -0.5$ in Fig.3), below which patterns are no longer learnable, is almost exactly the same with our simple cavity approach and the much more complicated RSB1 calculation.

If we accept the local fields from RSB1 as a good approximation, then again the Kuhn-Tucker conditions (6) must be fulfilled after learning, and with (14) we can calculate the loading $\alpha_{\text{RSB-KT}}$ from the RSB1 field distribution calculated with the parameters $q, w$ and $x$. To this purpose, in the above-mentioned formula one only has to replace the Gaussian measure $\mathcal{D}t$ by the RSB1 field distribution of those patterns, which are expicitly embedded, i.e. from $t_0$ to $\kappa(=1)$ in Fig. 3. This is related to $\alpha_{\text{RSB1}}$ in a similar way as $\alpha_{\text{Cav}}$ is related to $\alpha_{\text{RS}}$.

For three values of $\kappa$, the result $f_{\text{KT}} := f_{\text{RSB1-KT}}$ of such a calculation is presented in Fig. 4. For comparison also the results of the two simple approximations, $f_{\text{Cav}}$ and $f_{\text{RS}}$ are presented, together with the RSB1 replica result $f_{\text{RSB1}}$. Obviously, the improved cavity result $f_{\text{KT}}$ is only slightly higher than $f_{\text{RSB1}}$, which is already a criterion for the quality of the RSB1 calculation compared with the RS theory. One expects therefore that the rigorous result should lie between our $f_{\text{KT}}$ as an upper and $f_{\text{RSB1}}$ as a lower bound, so that further replica breaking steps should give only a slight improvement. Precisely, we expect both a slow monotoneous *increase* of the error rate $f_{\text{RSBn}}$ with increasing $n$ in a RSBn calculation, [10], and a slow monotoneous *decrease* of $f_{\text{RSBn-KT}}$, and at the same time a decrease of the relative number of explicitly embedded patterns and of the averaged embedding strength. As a consequence, $\alpha_{\text{RSBn-KT}}$, which is calculated from a formula analogous to (14), increases slowly. For $n \to \infty$, the RSBn replica result, and the RSBn-KT cavity result derived from the RSBn field distribution, should agree.

We have thus found a method leading to an upper bound for the error rate $f$ as a function of $\alpha$ and $\kappa$, if the local fields before relearning are known. In [10], Fontanari and Theumann derive an upper bound for $f$ at $\kappa$=0, evaluated in RS approximation at the AT line for finite temperatures: For $\alpha \lesssim 50$, our RSB1-KT upper bound is lower, whereas for $\alpha \gtrsim 50$ the bound in Fig. 2 of [10] is lower.

That the RSB1 itself is not yet sufficient, has already been shown by Majer und Engel [6]). Moreover, also the fact that after our RSB1-

KT cavity learning there is still a *gap* in the field distribution, which is according to Bouten [4] responsible for RSB, points to this fact. Additionally in this approximation the condition is violated that the number of explicitly embedded patterns should not exceed the number $N$ of coupling degrees of freedom, see eqn. (19).

As already mentioned, the result of the cavity method is unsufficient above $\alpha_c$, since the combinatorial possibilities to select the "waste patterns" are not considered. However, it should also be possible to get equations, which are equivalent to RSB1, without using the replica trick, i.e. as in the seminal book [11] on spin glasses.

To achieve this, one considers a multitude of ground states, which are ordered in an ultrametric structure. If one decreases the stability constraints, the number of different ground states is assumed to increase exponentially. If one now adds a pattern to this ensemble, a lower embedding strength is therefore favoured, which gives a higher storage capacity compared to the Gaussian one of eqn. (14). Work on such an intrinsic cavity approach for $\alpha > \alpha_c$ is in progress.

# 3 The AND-machine

For multilayer perceptrons, replica breaking is a general phenomenon, when optimal learning capacity is required. As a consequence, the optimal capacity is not easily estimated. Griniasty and Grossman [12] have calculated the storage capacity of the AND-machine and gave arguments, which made a suppression of replica symmetry breaking credible. However, with our method we are able to check their proposal and find that replica symmetry is broken.

## 3.1 Model description

The AND-*machine* is a simple example for a *multi-layer* perceptron. Multilayer perceptrons have been introduced to overcome the limitations of single layer perceptrons, which are limited to *linearly separable classifications* [13], whereas with sufficiently many units in the additional layer(s) between the input and output units, every Boolean function can be implemented. However, at the same time, the analytical treatment of the models becomes much more complicated, and in general, the space of solutions is no longer simply connected. The RS approximation yields then results which contradict the exact bounds derived by Mitchison and Durbin [14]. However, the necessary RSB calculation is rather complicated for such models. A multilayer model, which has been studied in this way, is the so-called *committee machine* with 3 (or any other odd number $N_h$ of) *hidden units* [15, 16]. Here the intermediate *hidden layer* consists of three neurons, and the output is given by the *majority vote* of these *hidden neurons*. The maximal capacity of the system decreases from $\alpha_c \simeq 4.02$ for the RS approximation to $\alpha_c \simeq 3.0$ for the RSB-*ansatz*. This is for the case of non-overlapping receptive fields, see fig. 5.

In contrast to the *committee* machine, for the AND machine the number $N_h$ of intermediate neurons is arbitrary ($\geq 2$); here the output unit gives a positive vote iff all intermediate neurons vote with $+1$. This AND machine was at first treated by Griniasty and Grossman, [12], both with non-overlapping and also with overlapping receptive fields (NRF and ORF cases, respectively). Within a replica-symmetric approach, the authors find for the case of an equal number of patterns with output $+1$ and $-1$ for $N = 2$ intermediate neurons a critical capacity of $\alpha_c \simeq 3.5$ ($\alpha_c \simeq 3.3$ with simulations), for the case of overlapping receptive fields, and $\alpha_c \simeq 3.66$ for the NRF case, [12]. At the end of their paper, Griniasty und Grossman discuss the validity of their RS approximation and find support for their assumption that only a single minimum contributes to their solution. Griniasty [9] repeated the calculation with his cavity method and got the same results. In the following, we use our different cavity method.

### 3.2 Calculation of the learning capacity

As we have seen, our method yields a convenient way to recognize the necessity of RSB. Additionally one gets a rough quantitative estimate, to which extent the actual solution is approximated. At first we study the AND machine with $N_h = 2$ and non-overlapping receptive fields (NRF case). Since this does not necessitate an additional effort and the argumentation is simplified, we assume that both subnetworks are trained with optimal stability, as long as the optimal capacity $\alpha_c$ is not yet reached.

We assume $p_+ = \alpha_+ N$ and $p_- = \alpha_- N$ patterns with positive resp. negative output. Then the parameter $b$ is defined as follows

$$\alpha_{\pm} \;=\; \frac{1 \pm b}{2}\,\alpha \quad . \tag{20}$$

The (+)-patterns must be trained in both subnets, since both intermediate neurons must vote positively, whereas for the (−)-patterns one can choose at will a subnet with a negative vote. From this fact, Griniasty and Grossman [12] derive the sharp bounds

$$\frac{6}{2+b} \;\leq\; \alpha_c \;\leq\; \frac{4}{1+b} \tag{21}$$

for the maximal capacity $\alpha_c$, which also applies with overlapping receptive fields (ORF case). Moreover, recently Wendemuth [17] was able to generalize the method of Mitchison and Durbin [14] and obtained for the AND-Machine as a lower bound the sharper condition

$$\frac{8}{4-(1-b)\log_2(3)} \geq \alpha_c \quad , \tag{22}$$

which also applies both to the NRF and the ORF case.

Let us consider fictitiously all possibilities to distribute the responsibility for the (−)-patterns among the $N_h$ subnetworks; these patterns, including the (+)-patterns, shall then be trained to optimal stability. Afterwards we select that distribution, which leads to maximal stability $\kappa = \min\{\kappa_1, \kappa_2\}$.

In both subnetworks $i = 1, 2$ we have $N$ input neurons, and in both subnetworks the couplings are defined with embedding strengths $x_i^\mu$ as

$$J_{ik} \;=\; \frac{1}{N}\sum_{\mu} x_i^\mu \zeta^\mu \xi_{ik}^\mu \quad . \tag{23}$$

Here $\zeta^\mu$ is the desired final output. As before, we define the oriented correlation matrix $\overleftrightarrow{B}$, the length $L$ of the coupling vectors and the oriented local fields $\vec{E}$

$$B_i^{\mu\nu} \;:=\; \frac{1}{N}\,\zeta^\mu\zeta^\nu \sum_k \xi_{ik}^\mu \xi_{ik}^\nu \tag{24}$$

$$L_i^2 \;:=\; \underline{J}_i^T \underline{J}_i \;=\; \frac{1}{N}\sum_{\mu,\nu} x_i^\mu B_i^{\mu\nu} x_i^\nu \tag{25}$$

$$E_i^\mu \;:=\; \sum_{\nu} B_i^{\mu\nu} x_i^\nu \quad . \tag{26}$$

The embedding strengths $x_i^\mu$, for $i = 1, 2$ and all $\mu$, together with the $E_i^\mu$, fulfill again the KT conditions

$$\text{either} \quad (x_i^\mu > 0 \;\text{ and }\; E_i^\mu = 1) \quad \text{or} \quad (x_i^\mu = 0 \;\text{ and }\; E_i^\mu > 1) \quad . \tag{27}$$

Additionally we know that for our optimal choice, a $(-)$-pattern can have a positive embedding strength only at one of the subnets, since otherwise the less stable subnet could enhance the stability by reducing its unnecessarily positive $x_i^\mu$ to 0 and relearning the other patterns. Furthermore, in the thermodynamic limit, both stabilities should be equal, i.e. $L_1^2 = L_2^2 =: L^2$. Let us assume again that there is only one groundstate, and add a new pattern. Again this feels a normally distributed random oriented field $\widetilde{E}_i^0$ with variance $L^2$ in each subnet. A $(+)$-pattern must be classified correctly in both subnets. So we need in case of $\widetilde{E}_i^0 < 1$ positive embedding strengths $x_i^0 = (1 - \widetilde{E}_i^0)/(1 + g_i)$, where the response factors $g_i$ have still to be determined. In contrast, for a $(-)$-pattern we can choose the subnet with negative intermediate output. As we will see below, it is best to embed the $(-)$-pattern in that subnet, where the oriented field is larger, so that the embedding strength can be smaller. The field distribution $P(t)$ for the larger oriented field $t$, with normalized couplings, $(\widetilde{E}_{\text{max}}^0 = Lt)$, can be calculated from the Gaussian distribution of the fields $(t_1, t_2)$ of the two sublattices as follows:

$$\begin{aligned} P(t) &= 2\,P(t_1 = t, t_2 < t) = \frac{2}{\sqrt{2\pi}}\, e^{-t^2/2} \int_{-\infty}^{t} \mathcal{D}t_2 \\ &= \frac{2}{\sqrt{2\pi}}\, e^{-t^2/2}\, \Phi(t) \quad . \end{aligned} \tag{28}$$

If again we assume that $x^\mu$ and $B^{\mu\nu}$ are uncorrelated, then the answer $g$ of the patterns, which had already been implemented and now must keep the KT conditions, is similar to eqn. (12), namely

$$- \sum_{\substack{\mu \\ (x^\mu > 0)}} (B^{0\mu})^2 = -\alpha_{\text{eff}} = -\alpha P(x^\mu > 0) \quad . \tag{29}$$

Identifying again the distribution of the embedding strengths with the probability distribution for $x^0$ and normalizing the couplings to 1 we get

$$g \;=\; -\alpha_+ \int\limits_{-\infty}^{\kappa} \mathcal{D}t \;-\; \alpha_- \int\limits_{-\infty}^{\kappa} \mathcal{D}t\, \Phi(t) \quad . \tag{30}$$

Here the first and second parts describe the infuence of $(+)$- and $(-)$-patterns, whereby compared with eqn. (28) a factor $1/2$ was taken into account, since only one of the two subnets ist needed for $(-)$-patterns. The capacity for finite stabilities is again calculated from the KT conditions (27) through

$$L_i^2 \;=\; \frac{1}{N}\sum_{\mu\nu} x_i^\mu B_i^{\mu\nu} x_i^\nu \;=\; \frac{1}{N}\sum_{\mu} x_i^\mu E_i^\mu = \frac{1}{N}\sum_{\mu} x_i^\mu \tag{31}$$

$$\;=\; \alpha \int \mathrm{d}x_i\, x_i w(x_i) = \frac{L_i}{1+g} \int\limits_{-\infty}^{\kappa} \mathcal{D}t\, (\alpha_+ + \alpha_- \Phi(t)) \cdot (\kappa - t)\, . \tag{32}$$

Again, with $\kappa = 1/L$, and multiplying by $(1+g)$, we obtain finally from our cavity method

$$1 \;=\; \alpha \int\limits_{-\infty}^{\kappa} \mathcal{D}t\, \left( \tfrac{1+b}{2} \;+\; \tfrac{1-b}{2}\, \Phi(t) \right) \cdot (1 + \kappa(\kappa - t)) \quad . \tag{33}$$

Obviously, every other method to store a pattern would lead to a smaller learning capacity with our method. The maximal capacity for $\kappa = 0$ is then with our method (i.e. with $\alpha = \alpha_{\mathrm{Cav}}$)

$$\begin{aligned} (\alpha_c^{-1})_{\mathrm{Cav}} \;&=\; \frac{1+b}{2} \int\limits_{-\infty}^{0} \mathcal{D}t \;+\; \frac{1-b}{2} \int\limits_{-\infty}^{0} \mathcal{D}t\, \Phi(t) \\ &=\; \frac{1+b}{4} + \frac{1-b}{16} \;=\; \frac{5-3b}{16} \quad . \end{aligned} \tag{34}$$

This corresponds again to the limit $g = -1$ of eqn. (30) and is compatible with the exact bounds (21), although it is somewhat lower than the condition (22). This, again, is no surprise, since we always expect a somewhat too low estimate from our method in case of RSB situations.

Now, in contrast, we perform the Griniasty step, see [9], to neglect the non-diagonal terms of $\overleftrightarrow{B}$ and assuming at the same time $g = 0$. This would lead to "RS" results, namely

$$
\begin{aligned}
1 &= \frac{1}{N}\,\vec{x}^{\,T}_{\mathrm{RS,\,i}}\,\overleftrightarrow{B}_{\mathrm{RS,\,i}}\,\vec{x}_{\mathrm{RS,\,i}} = \frac{1}{N}\sum_{\mu}(x^{\mu}_{\mathrm{RS,\,i}})^2 \\
&= \alpha_{\mathrm{RS}}\int_{-\infty}^{\kappa}\mathcal{D}t\,\left(\frac{1+b}{2}+\frac{1-b}{2}\,\Phi(t)\right)\cdot(\kappa-t)^2 \quad , \qquad (35)
\end{aligned}
$$

which differs from eqn. (32). For $\kappa = 0$ this yields the result of [12], namely

$$
(\alpha_c^{-1})_{\mathrm{RS}} = \frac{1+b}{4} + \frac{1-b}{2}\,0.045422528\ldots \quad . \qquad (36)
$$

These results are for non-overlapping receptive fields (NRF). The maximal capacity $\alpha_c$ is compared for both approximations in fig. 7. Additionally, also the result for overlapping receptive fields (ORF, see below) is presented as the dashed curve, which is slightly lower for the present case, but not always. This will be discussed later in more detail.

The fact that the RS result differs from our cavity result is, according to our experience, a strong hint for the necessity of RSB. The extent of replica symmetry breaking seems to increase with higher loading and larger percentage of $(-)$-patterns.

For $\alpha_+ = \alpha_-$ the maximal capacity according to our theory is $\alpha_c = 3.2$, whereas the RS approximation leads to $\alpha_c = 3.667$ and eqn. (22) to $\alpha_c = 3.31$. Probably the true result is again smaller than the RS value, but not too far from it. Thus, for $\alpha_+ = \alpha_-$, the RS appoach yields again a good estimate although replica symmetry is probably broken.

It would be interesting to check the local stability of the RS *ansatz*, since the RS result and our cavity result deviate already for small values of $\alpha$. Interestingly, there is no *gap* in the "RS" distribution of oriented fields, so Bouten's sufficient criterion for RSB in the case of simple perceptrons above $\alpha_c$, namely the existence of such a gap, see [4], is not given in the present case, i.e. it is not possible to guess the answer without performing the complicated calculation. Moreover, the limit $g = -1$, above which the system is obviously over-determined concerning the number of couplings, is already reached below $\alpha_c$. On the other hand, if in the perceptron case not the Gardner-Derrida cost function (2), but the Perceptron function (4) would be minimized above $\alpha_c$, then the limit $g \to (-1)$ would yield the correct result, as mentioned in section 2.2.

Thus there is again the question, wether also here local and global stability of the RS solution are identical, or whether perhaps in our case the RS solution becomes *globally* unstable already at a smaller value (here $\alpha = 0$), whereas the *local* instability happens only at the AT-line given by $g = -1$. This would correspond to the scenario suggested for the simple perceptron in the original paper of Gardner and Derrida [3], which was only recently corrected by Bouten, [4].

Our cavity method is not only applicable to the present case of an NRF-AND machine, but also for NRF-machines with *arbitrary Boolean output* functions: If one has found the optimal strategy similar as above, the steps leading to eqn. (33) are identical, and one only has to substitute the embedding strengths, which compensate the normally-distributed random field, in this equation.

Formally this leads to the following formula for the storage capacity $\alpha = \min_i\{\alpha_i\}$: If $\boldsymbol{t}$ is the random vector describing the oriented fields of the subnets and $x_{i,\kappa}(\boldsymbol{t})$ the embedding strength following from the optimal learning strategy for the subnet $i$ (without *response* $g$), then

$$1 \;=\; \alpha_i \int \mathcal{D}\boldsymbol{t}\, [\theta(x_\kappa(\boldsymbol{t})) + \kappa x_{i,\kappa}(\boldsymbol{t})] \quad . \tag{37}$$

Here the different desired outputs, analogous to the $(+)$ and $(-)$-patterns in case of the AND-Machine, have to be taken into account according to their respective probabilities. The optimal storage capacity is obtained, if one of the subnets has reached its capacity limit. If the output value follows from a Boolean function, for which all the neurons in the intermediate layer are equivalent, then all the $\alpha_i$ are identical. The formula analogous to (35), giving an upper estimate for the storage capacity, was for $\kappa = 0$ already determined by Engel *et al.*, [16] and was described in a more abstract way in [12, 9]. In the formulation of (37) it is

$$1 \;=\; \alpha_i \int \mathcal{D}\boldsymbol{t}\; x_{i,\kappa}^2(\boldsymbol{t}) \quad . \tag{38}$$

In contrast, the results for $\alpha_c$ obtained from eqn. (37) with the cavity method are lower estimates. For the *committee-machine*, because of $\alpha_{\text{Cav}} < 2$, they violate the *lower bound* of Mitchison and Durbin [14], $\alpha > 2$, and are therefore without interest.

More interesting would be a RSB calculation as suggested at the end of the section on the simple perceptron for $\alpha > \alpha_c$, since then relevant estimates from below for the true capacity for this model could be derived.

### 3.3 The AND-machine with overlapping receptive fields

Also the fully connected AND machine can be treated with our cavity method. This correponds to a machine as in fig. 5, but with identical patterns presented to both subnets. Thus the index $i$ of the description of the patterns $\{\xi_k^\mu\}$ is now dummy, cf. (23), and in particular it is $\overleftrightarrow{B}_1 = \overleftrightarrow{B}_2 =: \overleftrightarrow{B}$. A very important parameter for the fully connected AND machine is the overlap $R := \sum_k J_{1k} J_{2k}$ of the two subnets.

Since also for this case one expects RSB, we can no longer expect that $R$ agrees for different approaches, as it happened with the generalization problem treated in [18]. Therefore we can no longer simply compare with the RS results of [12].

For the local fields $t_1$ and $t_2$ one gets for given $R$

$$P_R(t_1, t_2) \;=\; \frac{1}{2\pi\sqrt{1-R^2}} \exp\left(-\frac{t_1^2 - 2Rt_1t_2 + t_2^2}{2(1-R^2)}\right) \quad . \tag{39}$$

However, the probability density of the field $t_1$ of a $(+)$-pattern needing explicit embedding is not influenced, since again

$$\int_{-\infty}^{\infty} \mathrm{d}t_2 P_R(t_1, t_2) \;=\; \frac{1}{\sqrt{2\pi}} e^{-t_1^2/2} \quad . \tag{40}$$

For the $(-)$-patterns the situation is more complicated. Negative correlations of the subnets simplify the storing of these patterns, whereas positive correlations enhance the probability that a $(-)$-pattern, i.e. one which is already wrongly classified by one of the subnets, cannot be stored automatically by the other one, i.e. that it must be embedded explicitly. In the limit $b \to -1$, i.e. when exclusively $(-)$-patterns must be embedded, $\underline{J}_1 = -\underline{J}_2$ (with $R = -1$) for arbitrary couplings $\underline{J}_1$, is a general solution of the problem in the limit $\alpha \to \infty$.

For arbitrary $b$, the field of a $(-)$-pattern before learning is

$$\begin{aligned} P_R(t) &= 2P_R(t_1 = t, t_2 < t) = \int_{-\infty}^{t} \mathrm{d}t_2\, P_R(t, t_2) \\ &= \frac{2}{\sqrt{2\pi}}\, e^{-t^2/2}\, \Phi\left(\frac{t\,(1-R)}{\sqrt{1-R^2}}\right) \quad , \end{aligned} \tag{41}$$

and one obtains (28) as special case for $R = 0$. Again, a single groundstate is assumed, and again the arguments follow from eqn. (29) ff.

For the *response* $g$ and the capacity $\alpha$ we get

$$g = -\alpha_+ \int_{-\infty}^{\kappa} \mathcal{D}t \;-\; \alpha_- \int_{-\infty}^{\kappa} \mathcal{D}t\, \Phi\left(\frac{t\,(1-R)}{\sqrt{1-R^2}}\right) \quad , \tag{42}$$

$$1 = \alpha \int_{-\infty}^{\kappa} \mathcal{D}t \left(\frac{1+b}{2} + \frac{1-b}{2}\, \Phi\left(\frac{t\,(1-R)}{\sqrt{1-R^2}}\right)\right) \quad . \tag{43}$$

Similar as for the generalization problem in section 2.3 of [18] only that solution $\alpha(R)$ is relevant, for which $R$ is reproduced selfconsistently. For normalized couplings one obtains

$$\begin{aligned} R &= \sum_k J_{1k} J_{2k} = \sum_k \sum_{\mu\nu} x_1^\mu \zeta^\mu \xi_k^\mu x_2^\nu \zeta^\nu \xi_k^\nu = \sum_{\mu\nu} x_1^\mu B^{\mu\nu} x_2^\nu = \sum_\mu x_1^\mu E_2^\mu \\ &= \alpha_- \int_{-\infty}^{\kappa} \mathrm{d}t_1 \int_{-\infty}^{t_1} \mathrm{d}t_2\, P_R(t_1, t_2)\, \frac{\kappa - t_1}{1+g}\, t_2 \\ &\quad + \alpha_+ \int_{-\infty}^{\kappa} \mathrm{d}t_1 \int_{-\infty}^{\infty} \mathrm{d}t_2\, P_R(t_1, t_2)\, \frac{\kappa - t_1}{1+g}\, \max\{\kappa, t_2\} \quad . \end{aligned} \tag{44}$$

In eqn (44) the above-mentioned strategy for the storing of patterns was used: For the $(-)$-patterns the subnet with the smaller oriented field, here $t_2$, is unchanged, whereas the $(+)$-patterns are explicitly embedded in both subnets $i$, if the field $t_i$ is $< \kappa$. Again one multiplies with $(1+g)$ to simplify (44).

For $\kappa = 0$ it is again $g = -1$, and therefore one obtains $R$ from

$$0 \;=\; \frac{1-b}{2} \int\limits_{-\infty}^{0} \mathrm{d}t_1 \int\limits_{-\infty}^{t_1} \mathrm{d}t_2\, P_R(t_1, t_2)\,(-t_1)\, t_2 \tag{45}$$

$$+\; \frac{1+b}{2} \int\limits_{-\infty}^{0} \mathrm{d}t_1 \int\limits_{0}^{\infty} \mathrm{d}t_2\, P_R(t_1, t_2)\,(-t_1)\, t_2 \quad . \tag{46}$$

The result for $R(b)$ is presented in Fig. 7. One should note that it deviates from the result obtained by Griniasty and Grossman in [12]. From a comparison with Fig. 3a in [12] one finds that our value for $R$ is systematically larger, which - as mentioned above - has a negative effect on the storage capacity.

In Fig. 8 the capacity of the fully connected AND-machine is presented for the case that the capacity in (43) is calculated for $\kappa = 0$ with the $R$ determined from (46). For $b = 0$ we get $R = 0.217$ and $\alpha = 3.113$, as opposed to $R = 0$ and $\alpha = 3.512$ obtained by Griniasty und Grossman, [12]. Again we expect that the RS-approximation gives the better estimate. However, our estimate for $b \lesssim -0.65$ is above the *lower bound* (22), as it should, and in the limit $b \to -1$ one gets $\alpha_{\mathrm{Cfull}} \to \infty$.

Apart from the fact that different capacities $\alpha_c$ are obtained with the replica and the cavity approach, there is a second hint on RSB: The replica calculation with the RS ansatz yields an overlap $R$ between the subnets, which is not reproduced by our "optimal" learning algorithm. For the generalization problem, see [18], the perfect agreement of the results obtained with the two approaches was a clear hint on the correctness of the solution, and in particular on the correctness of RS. In the present case, however, one finds that the RS solution apparently yields the optimal capacity only, when an additional component to the coupling vector $\underline{J}$ is introduced, by which the overlap $R$ between the subnets is reduced, but the embedding of the patterns is actually weakened.

In fact, in [9], Griniasty suggests a non-trivial training process for the fully connected AND-machine, by which patterns in the non-affected subnet are unlearned, influencing in this way the correlation $R$. According to our considerations, however, this is not optimal, since after deleting the additional component one can enhance the stability or store additional patterns, using the same distribution of tasks with respect to the different patterns. Thus it is not astonishing that Griniasty [9] with his training process obtains a smaller capacity ($\alpha_c = 3.0$) as with a stochastic algorithm, which leads to $\alpha_c = 3.3$, [12]. This discrepancy, which demands an additional component to the coupling vector, by which the subnets are decorrelated, should become smaller by a RSB calculation.

## 4 Discussion

As we have stated already in [1], our cavity approach is usually technically simpler than a replica calculation. Moreover, it is exact as long as calculations within the replica approach under the assumption of Replica Symmetry (RS) would be correct. Here we have additionally shown that for models with Replica-Symmetry Breaking (RSB), e.g. the simple perceptron above $\alpha_c$, one gets different results, as one should, with our "Kuhn-Tucker cavity method" and the RS approximation: This would not be the case e.g. with the different cavity approximation of Griniasty [9], since Griniasty's theory is always equivalent to the replica calculation in RS approximation. However, although our cavity theory "indicates the necessity of RSB", if RS does not suffice, it is still far from being exact, since the combinatorial explosion of distributing the set of patterns into "good" patterns, which are stored, and "bad ones", which are not, is not considered.

From the results of the present paper it can be seen in detail that for cases with RSB, Griniasty's "RS-cavity theory" approach usually yields too optimistic estimates, whereas in the same cases, our Kuhn-Tucker cavity method is apparently "too pessimistic" in the error rates. However, the limiting negative field value, below which patterns are no longer learnable, (e.g. $t \approx -0.5$ in Fig. 3), is well approximated by our theory, and the theory also gives good results, if one starts with the field distribution by a 1-step-RSB calculation (see below).

Moreover, the RS approximation follows for our models, except of the last-mentioned case of the fully connected AND machine, always from the formally "crude approximation" of vanishing response factor $g$ and vanishing off-diagonal elements $B^{\mu\nu}$.

Another important point of our cavity method is the demand that the constraints, which the solutions impose on the couplings, are actually realizable, which means that no more degrees of freedom are fixed than are available with the given couplings. For the fully connected AND-Machine an additional postulate is that the correlation of the subnets is self-consistently reproduced by the embedding strengths. Thus we can interprete our result as an estimate adapted to our training algorithm.

Although replica calculations in RS approximation do not at all take care of such details, we have found that they yield good estimates for the storage capacity of the simple perceptron above $\alpha_c$, and probably also for the AND machine. In this respect, the virtue of our approach is based on two facts:

At first, it yields an independent estimate, which seems to be a lower bound for $\alpha_c$ and an upper bound for the error fraction $f$.

Second, our cavity theory visualizes the "internal stresses" inherent in the RS approach and shows, which quantities and order parameters depend most sensitively on the assumptions made.

Finally we repeat that our Kuhn-Tucker cavity approach, starting from field distributions, which Majer and Engel, see [6], obtained in 1-step RSB for the perceptron above $\alpha_c(\kappa)$, shows that the 1-step RSB results must already be very near to the truth.

## References


[1] Gerl, F., Krey, U.: J. Phys. **A 27**, 7353 (1994).

[2] Gerl, F., Bauer, K., Krey, U.: Z. Physik **B 88**, 339 (1992)

[3] Gardner, E., Derrida, B.: J. Phys. **A 21**, 271 (1988)

[4] Bouten, M.: J. Phys. **A 27**, 66021 (1994)

[5] Griniasty, M., Gutfreund, H.: J. Phys. **24**, 715 (1990)

[6] Majer, P., Engel, E., Zippelius, A.: J. Phys. **A 26**, 7405 (1993)

[7] Anlauf, J.K., Biehl, M.: Europhys. Lett. **10**, 687 (1989)

[8] Opper, M.: Phys. Rev. **A38**, 3824 (1988)

[9] Griniasty, M.: Phys. Rev. **E 47**, 4496 (1993) J. Phys. I (France) **2** 167 (1992)

[10] Fontanari, J.F., Theumann, W.K., J. Phys. bf A 26, L1233 (1993)

[11] Mézard, M., Parisi, G., Virasoro, M.A.: *Spin glass theory and beyond*, World Scientific, Singapore 1987

[12] Griniasty,M., Grossman, T., Phys. Rev. **A 45** 8924 (1992)

[13] Minsky, M.L., Papert, S.A., *Perceptrons*, MIT Press, Cambridge MA., 1969.

[14] Mitchison, G.J, Durbin, R.M., Biol. Cybernetics **60** 345 (1989)

[15] Barkai, E., Hansel, D., Sompolinsky, H., Phys. Rev. **A 45** 4146 (1992)

[16] Engel, A., Köhler, H.M., Tschepke, F., Vollmayr H., Zippelius, A., Phys. Rev. **A 45** 7590 (1992)

[17] Wendemuth, a:, to appear in *Proceedings of the 1994 Congress on fuzzy and intelligent techniques (EUFIT 94)*, Aachen, 20-23 September 1994

[18] Gerl, F., Krey, U.: accepted for publication by J. Phys. **A**.

## Figure Captions

Fig. 1: For two different error rates $f$ the storage capacity $\alpha$ is presented as a function of the stability $\kappa$ for both estimates (14) and (16).

Fig. 2: The stability $\kappa$ is presented as a function of the error rate $f$ for $\alpha = 2.0$, 0.8 und 0.4 . Results are shown for the cavity-method and the replica calculation in RS and 1-step-RSB approximation.

Fig. 3: The probability density $P(t)$ of the local fields $t$ before learning is presented as a function of $t$ at $\alpha = 1$ und $\kappa = 1$. Again results are shown for the cavity-method and the replica calculation in RS and 1-step-RSB approximation. The distribution after training is given by pushing the middle segment to a $\delta$-Peak at $\kappa = 1$ in every case. The cavity-method yields an error rate $f = 0.29092$, the replica method gives $f = 0.13073$ in RS and $f = 0.13576$ in 1-step-RSB approximation. The fraction of explicitly learned patterns is 0.55042 for the cavity approximation and 0.71062 and 0.66745 respectively for the replica calculations.

Fig. 4: The error rate $f$ is presented as a function of $\alpha$ for $\kappa = 1.25$, 0.5 und 0. The results are given for the cavity method, the RS and 1-step-RSB approximation as well as the result $f_{\text{KT}}$, which is calculated by putting together the Kuhn-Tucker conditions and the local fields of the RSB-solution. The best estimate is very likely between these RSB-graphs, i.e. $f_{\text{RSB}}$ and $f_{\text{KT}}$, and probably closer to $f_{\text{RSB}}$.

Fig. 5: The AND-Machine with tree structure (NRF). The weights between the intermediate layer and the output neuron are fixed A threshold between 0 and 1 at the output-neuron takes care that there only is a positive output if both intermediate neurons are positive.

Fig. 6: The storage capacity of the AND-machine is presented as a function of $\kappa$ for $b = 0$. The results for the cavity method and for the replica calculation are given. The storage capacity $\alpha_{\textbf{full}}$ of the fully connected AND-Machine, see 3.3, and $\alpha_{\text{Perc}}$ of the simple perceptron are presented for comparison

Fig. 7: The overlap $R$ of both subnetworks of the fully connected AND-machine is presented as a function of the bias parameters $b$ from (20). The result for the cavity method gives a higher correlation of the subnetworks compared to *Fig.* 3.a in [12]. We have $R = 0$ at $b = -1/3$,

which means $\alpha_- = 2\alpha_+$, in contrast to $R = 0$ at $b = 0$ in [12].

Fig. 8: The maximal storage capacity of the AND-machine is presented as a function of the bias parameter $b$ from (20). The result of the cavity method for the tree structure is compared to the RS-result from [12] and the cavity result for the fully connected AND-machine. The largest deviations occur for small values of $b$. For the fully connected AND-machine the storage capacity is again smaller than with the replica calculation (see *Fig.* 3.b in [12]). We have $\alpha_{\text{C-full}} = \alpha_{\text{C-tree}}$ at $b = -1/3$, when $R = 0$. For $b \to -1$ we have $\alpha_{\text{C-full}} \to \infty$ as in [12].

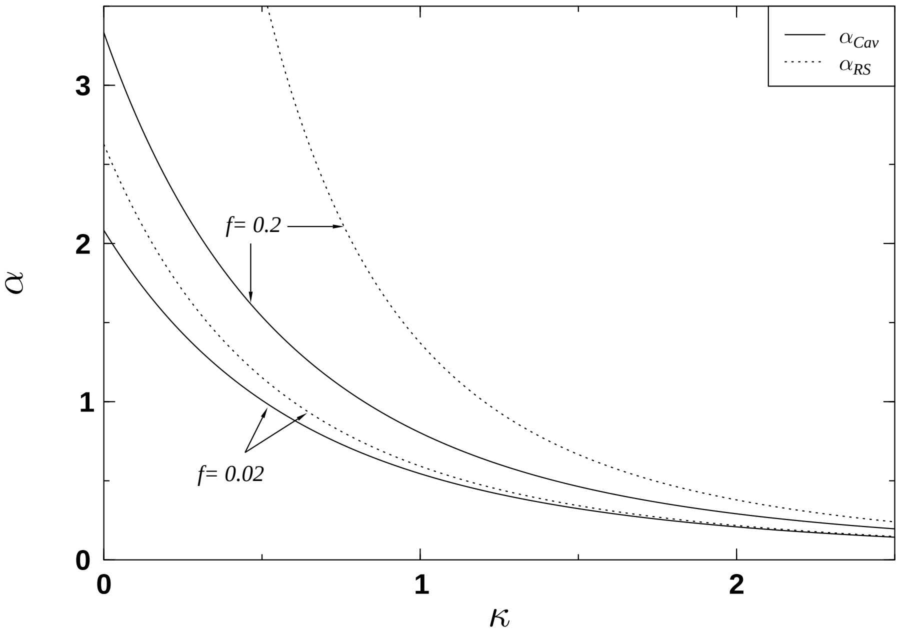


Figure 1: For two different error rates $f$ the storage capacity $\alpha$ is presented as a function of the stability $\kappa$ for both estimates (14) and (16).

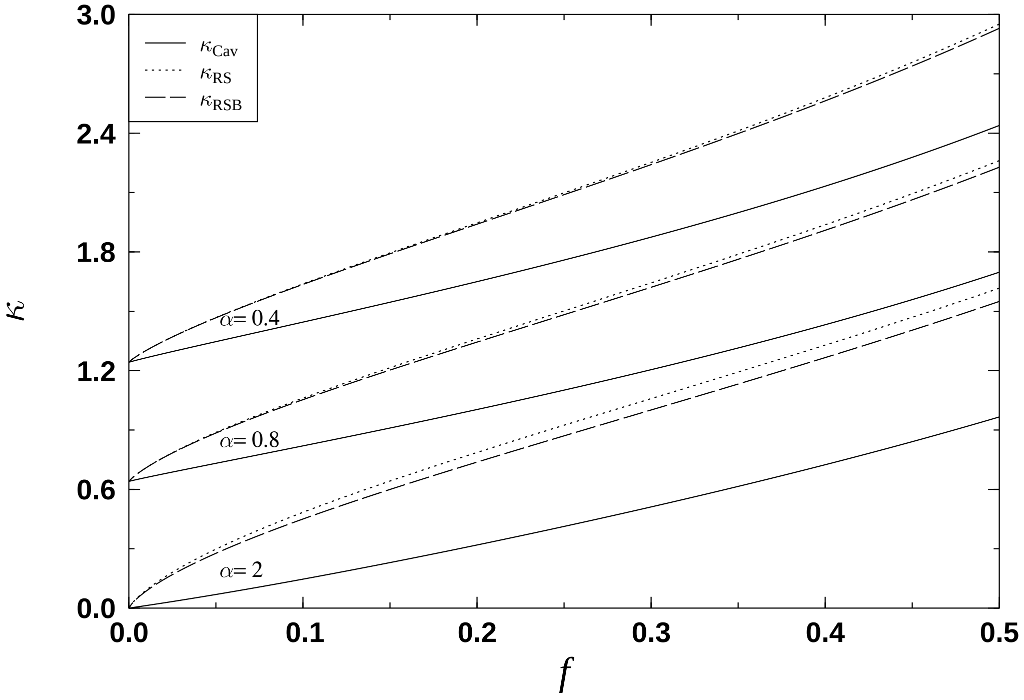


Figure 2: The stability $\kappa$ is presented as a function of the error rate $f$ for $\alpha = 2.0$, $0.8$ und $0.4$ . Results are shown for the cavity-method and the replica calculation in RS and 1-step-RSB approximation.

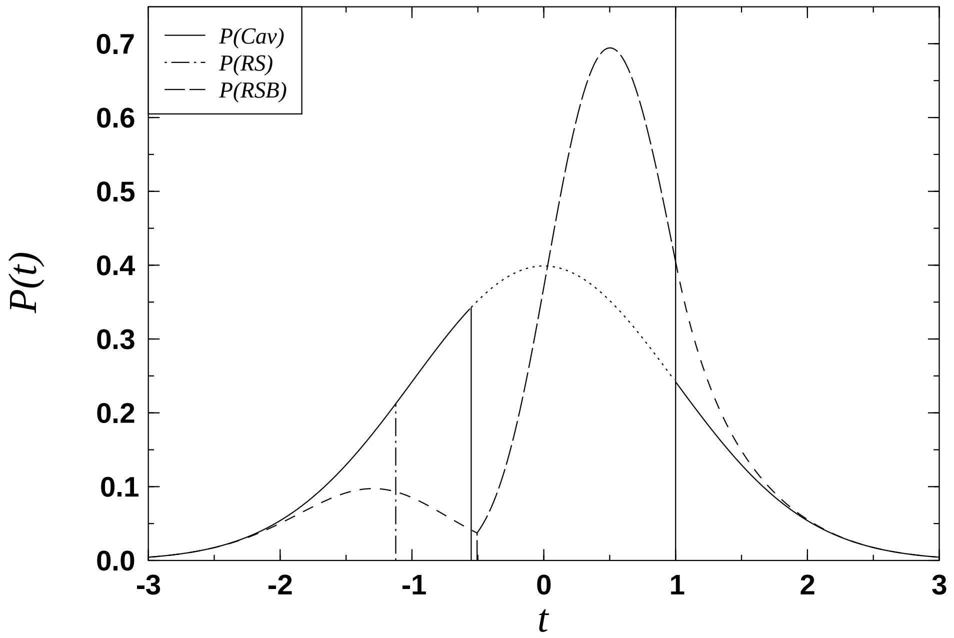


Figure 3: The probability density $P(t)$ of the local fields $t$ before learning is presented as a function of $t$ at $\alpha = 1$ und $\kappa = 1$. Again results are shown for the cavity-method and the replica calculation in RS and 1-step-RSB approximation. The distribution after training is given by pushing the middle segment to a $\delta$-Peak at $\kappa = 1$ in every case. The cavity-method yields an error rate $f = 0.29092$, the replica method gives $f = 0.13073$ in RS and $f = 0.13576$ in 1-step-RSB approximation. The fraction of explicitly learned patterns is 0.55042 for the cavity approximation and 0.71062 and 0.66745 respectively for the replica calculations.

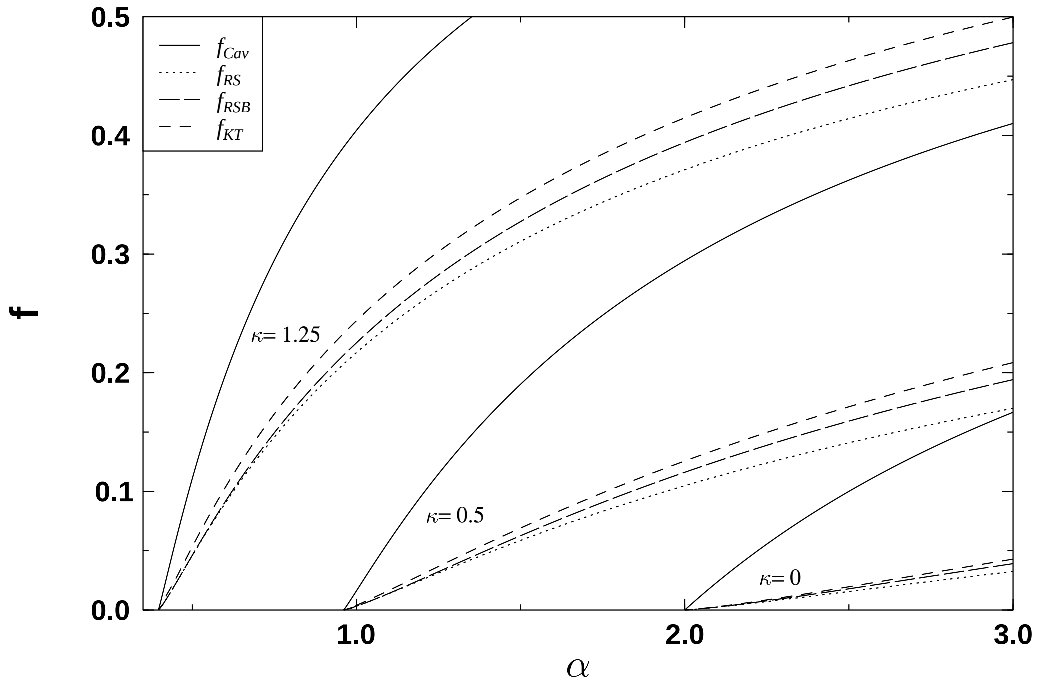


Figure 4: The error rate $f$ is presented as a function of $\alpha$ for $\kappa = 1.25$, 0.5 und 0. The results are given for the cavity method, the RS and 1-step-RSB approximation as well as the result $f_{\mathrm{KT}}$, which is calculated by putting together the Kuhn-Tucker conditions and the local fields of the RSB-solution. The best estimate is very likely between these RSB-graphs, i.e. $f_{\mathrm{RSB}}$ and $f_{\mathrm{KT}}$, and probably closer to $f_{\mathrm{RSB}}$.

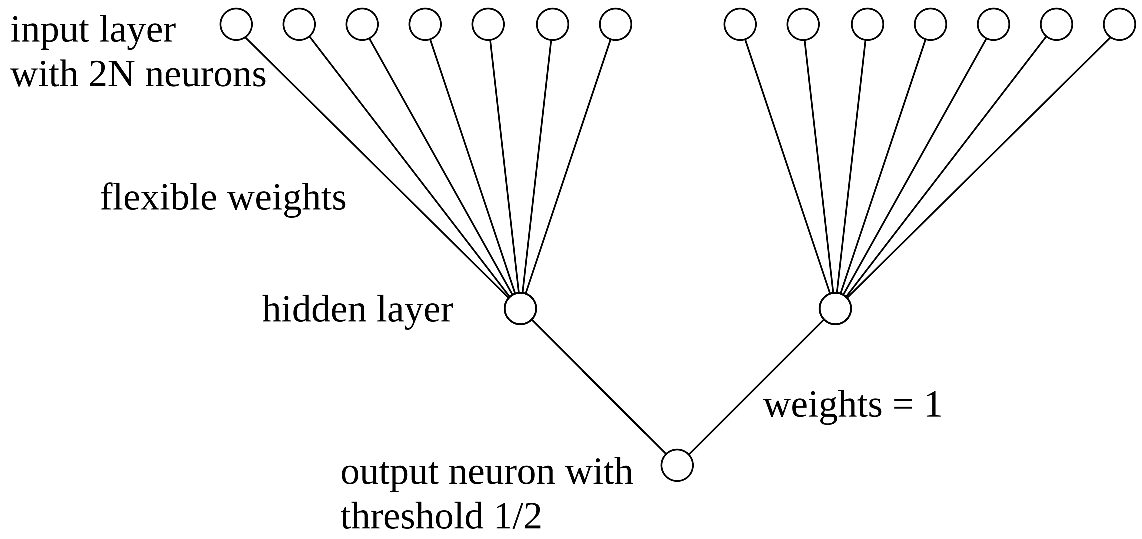


Figure 5: The AND-Machine with tree structure (NRF). The weights between the intermediate layer and the output neuron are fixed A threshold between 0 and 1 at the output-neuron takes care that there only is a positive output if both intermediate neurons are positive.

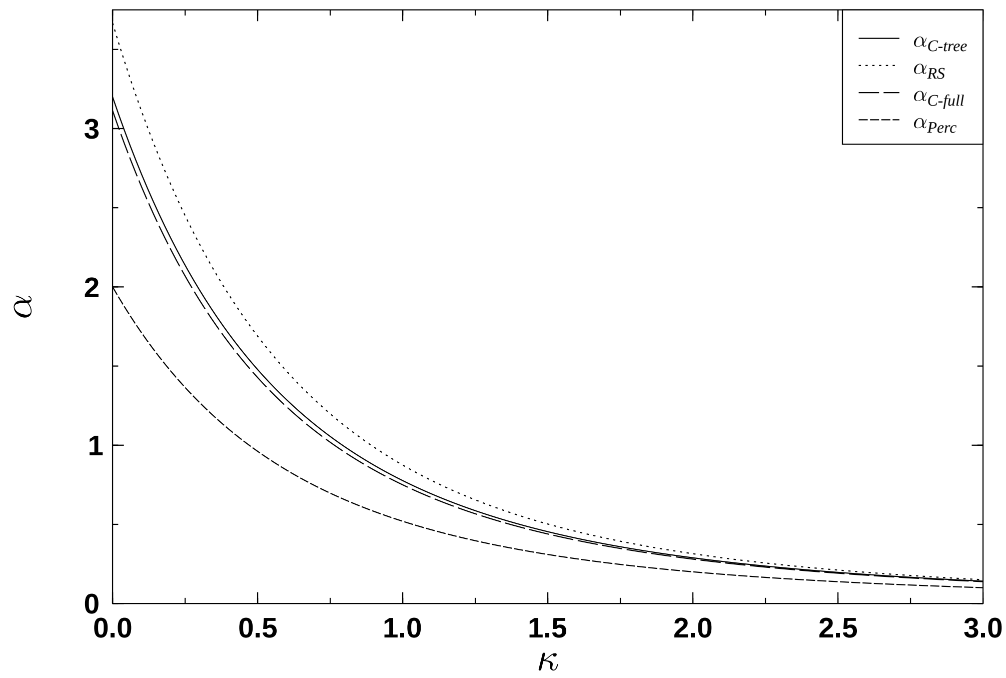


Figure 6: The storage capacity of the AND-machine is presented as a function of $\kappa$ for $b = 0$. The results for the cavity method and for the replica calculation are given. The storage capacity $\alpha_{\text{full}}$ of the fully connected AND-Machine, see 3.3, and $\alpha_{\text{Perc}}$ of the simple perceptron are presented for comparison

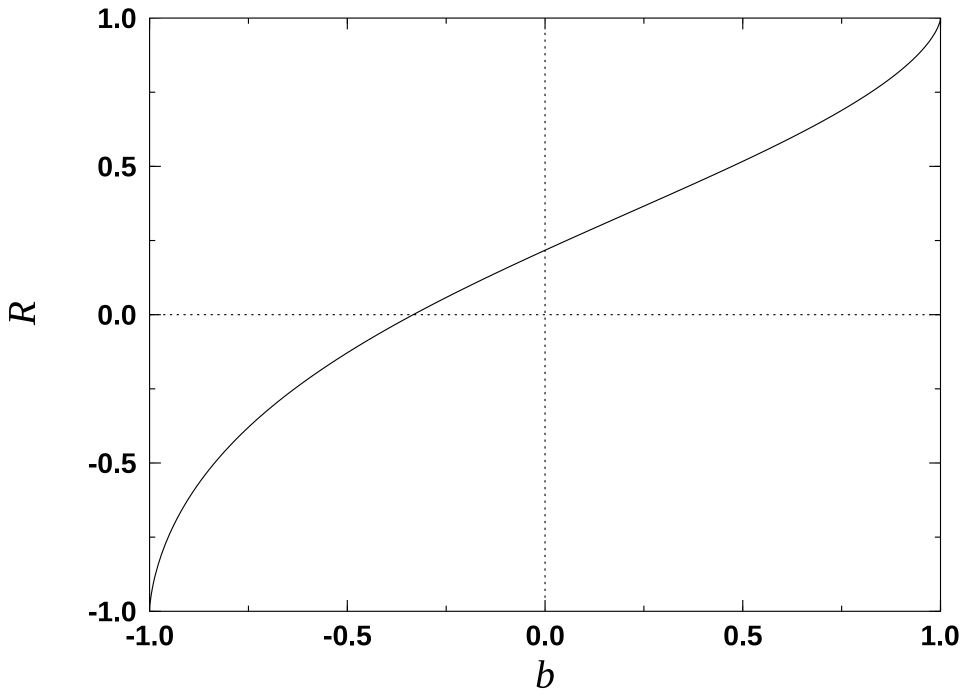


Figure 7: The overlap $R$ of both subnetworks of the fully connected AND-machine is presented as a function of the bias parameters $b$ from (20). The result for the cavity method gives a higher correlation of the subnetworks compared to *Fig.* 3.a in [12]. We have $R = 0$ at $b = -1/3$, which means $\alpha_- = 2\alpha_+$, in contrast to $R = 0$ at $b = 0$ in [12].

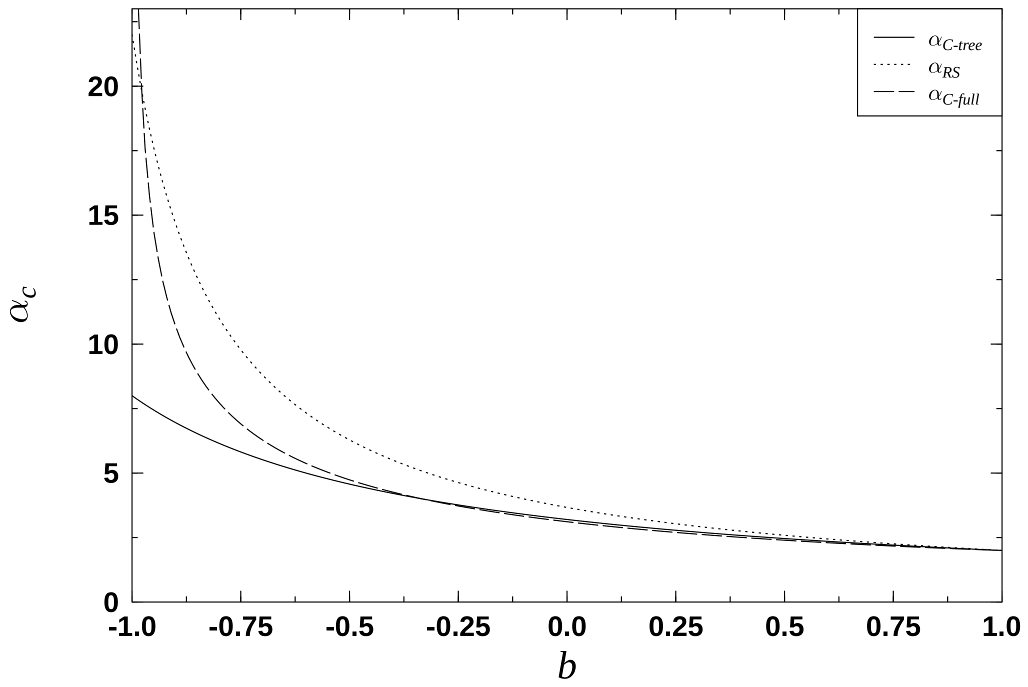


Figure 8: The maximal storage capacity of the AND-machine is presented as a function of the bias parameter $b$ from (20). The result of the cavity method for the tree structure is compared to the RS-result from [12] and the cavity result for the fully connected AND-machine. The largest deviations occur for small values of $b$. For the fully connected AND-machine the storage capacity is again smaller than with the replica calculation (see *Fig.* 3.b in [12]). We have $\alpha_{\text{C-full}} = \alpha_{\text{C-tree}}$ at $b = -1/3$, when $R = 0$. For $b \to -1$ we have $\alpha_{\text{C-full}} \to \infty$ as in [12].