Universität Würzburg

Institut für Theoretische Physik

Am Hubland, D–97074 Würzburg, Germany

Statistical physics and practical training of soft-committee machines

Martin Ahr, Michael Biehl, Robert Urbanczik

Ref.: WUE-ITP-98-057
e-mail : biehl@physik.uni-wuerzburg.de
ftp : ftp.physik.uni-wuerzburg.de, directory pub/preprint/1998/

www : http://theorie.physik.uni-wuerzburg.de/publications.shtml

Statistical Physics and Practical Training of Soft–Committee Machines

M. Ahr[1], M. Biehl[1], and R. Urbanczik[2]
(1) Institut für Theoretische Physik
Julius-Maximilians-Universität Würzburg
Am Hubland, D-97074 Würzburg
(2) Neural Computing Research Group
Aston University
Aston Triangle, Birmingham B4 7ET, U.K.

Abstract

Equilibrium states of large layered neural networks with differentiable activation function and a single, linear output unit are investigated using the replica formalism. The quenched free energy of a student network with a very large number of hidden units learning a rule of perfectly matching complexity is calculated analytically. The system undergoes a first order phase transition from unspecialized to specialized student configurations at a critical size of the training set. Computer simulations of learning by stochastic gradient descent from a fixed training set demonstrate that the equilibrium results describe quantitatively the plateau states which occur in practical training procedures at sufficiently small but finite learning rates.

Methods from statistical physics have been applied with great success within the theory of learning in adaptive systems. One prominent example is the investigation of feedforward neural networks which are capable of learning an unknown rule from example data [1, 2]. Frequently, the training procedure, i.e. the choice of network parameters (weights), is based on an energy function which measures the agreement of the student network with the rule in terms of the given example outputs. Statistical mechanics techniques can be applied if training is interpreted as a stochastic process which leads to a properly defined thermal equilibrium [3-5]. A particularly interesting topic is that of phase trasitions in this context, see [6] for a recent review. In multilayered neural networks, for example, underlying symmetries can cause a discontinuous dependence of the success of learning on the size of the training set, see e.g. [14-19].

In this paper we present the first treatment of learning in fully connected soft–committee machines by means of the replica method. This type of network consists of a layer with K hidden units, all of which are connected with the entire input, and the total output of the net is proportional to the sum of their states. Previous studies have addressed large soft committees ($K \to \infty$) with binary weights within the so–called Annealed Approximation [14] or networks with finite K in the limit of high training temperature [20].

Here, analytical results for the learning of a perfectly realizable rule at arbitrary training temperatures are derived (for very large K) within a replica symmetric ansatz. With an increasing size of the training set, the model exhibits a first order transition from unspecialized student configurations to specialized states with better performance. This transition is due to the invariance of the soft–committee output under permutation of hidden units. The same symmetry is known to result in quasi–stationary plateaus of the learning dynamics in on–line learning from a sequence of independent training examples [7-11], see [12] for a recent overview of this framework. Here, on the contrary, we will consider off–line learning from a fixed, limited set of examples. Furthermore we demonstrate that the statistical physics results, if interpreted correctly, describe the behavior of practical learning prescriptions. To this end we compare our results with the outcome of a stochastic variant of the well–known *backpropagation of error* algorithm [1, 2, 21].

We investigate a student–teacher scenario where the rule is parametrized as

$$\tau(\underline{\xi}) = \frac{1}{\sqrt{K}} \sum_{j=1}^{K} g(y_j) \quad \text{with} \quad y_j = \frac{1}{\sqrt{N}} \underline{B}_j \cdot \underline{\xi} \tag{1}$$

We assume an isotropic teacher with orthonormal weight vectors: $\underline{B}_j \cdot \underline{B}_k = N\,\delta_{jk}$ for all j,k. The training of a perfectly matching student with outputs $\sigma(\underline{\xi}) = \sum_{j=1}^{K} g(x_j)/\sqrt{K}$ is considered, where the arguments $x_j = \underline{J}_j \cdot \underline{\xi}/\sqrt{N}$ are defined through adaptive weights $\underline{J}_j$ with $\underline{J}_j^2 = N$. The particular choice of the hidden unit activation function, $g(x) = \mathrm{erf}(x/\sqrt{2})$, simplifies the mathematical treatment to a large extent [7, 8, 20]. We expect, however, that our results apply qualitatively to a large class of sigmoidal functions including the very similar and frequently used hyperbolic tangent.

Learning is guided by the minimization of the training error

$$\epsilon_t = \frac{1}{P} H\left(\left\{\underline{J}_j\right\}\right) = \frac{1}{P} \sum_{\mu=1}^{P} \epsilon\left(\{\underline{J}_i\}, \underline{\xi}^\mu\right) = \frac{1}{P} \sum_{\mu=1}^{P} \frac{1}{2}\left(\sigma(\underline{\xi}^\mu) - \tau(\underline{\xi}^\mu)\right)^2 \tag{2}$$

where P is the number of training examples, which we assume to scale like $P = \alpha N K$ with $\alpha = \mathcal{O}(1)$. The extensive quantity $H = P\epsilon_t$ plays the role of an energy. The replica formalism for the calculation of the corresponding quenched free energy exploits the identity $\langle \ln Z \rangle = \partial \langle Z^n \rangle / \partial n|_{n=0}$ where $\langle \ldots \rangle$ denotes an average over the set of random training examples. Z^n is equivalent to the partition function of n

non-interacting copies (labled $a = 1, 2, \ldots, n$) of the investigated system and reads:

$$Z^n = \int d\mu(\{\underline{J}_i^a\}) \prod_{\mu=1}^{P} \exp\left[-\frac{\beta}{2}\sum_{a=1}^{n}\left(\sigma^a\left(\underline{\xi}^\mu\right) - \tau\left(\underline{\xi}^\mu\right)\right)^2\right]. \tag{3}$$

Here, the measure $d\mu$ is meant to incorporate the normalization $\underline{J}_j^{a\,2} = N$ of the student vectors. We perform the quenched average over all possible sets of independent training inputs $\underline{\xi}^\mu$, the components of which are assumed to be i.i.d. Gaussian random numbers with mean zero and unit variance. One obtains the following form:

$$\langle Z^n\rangle = \int d\mu(\{\underline{J}_i^a\})\, e^{-PG_r} \quad \text{where} \quad G_r = -\ln\left\langle \exp\left[-\frac{\beta}{2}\sum_{a=1}^{n}\left(\sigma^a(\underline{\xi}) - \tau(\underline{\xi})\right)^2\right]\right\rangle_{\xi}. \tag{4}$$

Here and in the following $\langle \ldots \rangle_\xi$ denotes an average over the randomness contained in a single input vector. As the examples are independent, the quenched average over the training set factorizes.

The sample average G_r will only depend on the order parameters $R_{ij}^a = \underline{J}_i^a \cdot \underline{B}_j/N$ and $Q_{ij}^{ab} = \underline{J}_i^a \cdot \underline{J}_j^b/N$. Similarly the generalization error $\epsilon_g = \frac{1}{2}\langle(\sigma - \tau)^2\rangle_\xi$, which measures the success of learning by averaging over arbitrary inputs is given by [8]

$$\epsilon_g = \frac{1}{6} + \frac{1}{K\pi}\sum_{i,j=1}^{K}\left[\arcsin\left(\frac{Q_{ij}^{aa}}{2}\right) - 2\arcsin\left(\frac{R_{ij}^a}{2}\right)\right]. \tag{5}$$

In this paper we restrict ourselves to networks with a very large number K of hidden units. Non-trivial results can be obtained in the limit $K \to \infty$ but with $K \ll N$ by assuming that the relevant student configurations will be site symmetric:

$$R_{ij}^a = \begin{cases} R^a & \text{if } i = j \\ S^a & \text{if } i \neq j \end{cases}, \quad Q_{ij}^{aa}\begin{cases} 1 & \text{if } i = j \\ C^a & \text{if } i \neq j \end{cases}, \text{ and } Q_{ij}^{ab} = \begin{cases} q^{ab} & \text{if } i = j \\ p^{ab} & \text{if } i \neq j \end{cases} \quad \text{for } a \neq b. \tag{6}$$

Here and elsewhere in the paper superscripts a, b label replicas, whereas i and j are hidden unit indices. The restriction (6) allows the system to assume unspecialized ($R^a = S^a$) or specialized states ($R^a > S^a$). Note that the output of a student will be $\mathcal{O}(\sqrt{K})$ and thus on a different scale than the output of the teacher if C^a is on the order of 1. So that the magnitudes of the outputs match, we assume that the hidden unit cross overlaps (C^a, p^{ab} and S^a) are on the order of $1/K$. As a consequence of this scaling one may show [19] that the joint distribution of τ and the σ^a becomes Gaussian in the large K limit.

In the following we use the notation $\underline{\sigma} = (\sigma^1, \sigma^2, \ldots, \sigma^n, \tau)^\top$, and define a matrix B such that $\underline{\sigma}^\top \mathsf{B}\underline{\sigma} = \sum_{a=1}^{n}(\sigma^a - \tau)^2$. For large K the Gaussian joint distribution of $\underline{\sigma}$ is completely specified through the covariance matrix $\mathsf{M} = \left\langle \underline{\sigma}\,\underline{\sigma}^\top\right\rangle$, the elements of which can be expressed in terms of order parameters. Hence one obtains the effective Hamiltonian G_r, equation (4),

$$G_r = -\ln\left\{\frac{(2\pi)^{\frac{-n-1}{2}}}{\sqrt{\det \mathsf{M}}}\int d^{n+1}\sigma\ \exp\left[-\frac{1}{2}\underline{\sigma}^\top\left(\beta\mathsf{B}+\mathsf{M}^{-1}\right)\underline{\sigma}\right]\right\} = \frac{1}{2}\ln\left\{\det\left[\beta\mathsf{MB}+\mathsf{1}\right]\right\} \tag{7}$$

where the r.h.s. is a function of the site symmetric order parameters (6). A saddle point integration gives $1/N\ \ln\langle Z^n\rangle$ as the extremum (w.r.t. $\left\{R^a_{kl}, Q^{ab}_{kl}\right\}$) of $\exp[-P\,G_r\ + N\,s]$ where

$$s = \frac{1}{N}\ln\int d\mu(\{\underline{J}^a_i\})\prod_{k,l=1}^{K}\prod_{a,b=1}^{n}\delta(Q^{ab}_{kl}-N\underline{J}^a_k\cdot\underline{J}^b_l)\delta(R^a_{kl}-N\underline{J}^a_k\cdot\underline{B}_l). \tag{8}$$

The entropy term can be calculated by means of a saddle point integration itself after writing the δ-functions in their integral representation. One obtains

$$s = 1/2\ln(\det\mathsf{C}) + \text{const.}, \tag{9}$$

where C is the $[(n+1)K]$-dimensional square matrix of all cross- and self-overlaps of (replicated) student and teacher vectors [16, 19]. In the Appendix we sketch a simpler derivation of this result which avoids the saddle point method.

In order to proceed with the analysis, we make a replica symmetric ansatz, in addition to site symmetry (6): $R^a = R, S^a = S, C^a = C$ and $q^{ab} = q, p^{ab} = p$ for $a \neq b$. This assumption simplifies the evaluation of the determinants and allows for a straightforward treatment of the limit $n \to 0$. In agreement with the scaling of the hidden unit cross overlaps, we reparametrize: $S = \widehat{S}/K$, $C = \widehat{C}/(K-1)$, and $p = \widehat{p}/K$. The parameters $\Delta = R - S$ and $\delta = q - p$ now measure the degree of specialization in the network. Inserting these in the saddle point equations we find that the condition $\partial f/\partial\widehat{S} = 0$ can only be satisfied, if $\widetilde{C} = K(1+\widehat{C}-\delta-\widehat{p}) = \mathcal{O}(1)$. After eliminating $\widehat{C}$ accordingly, we obtain the free energy as a function of variables of order one:

$$\frac{2\beta F}{NK} = \alpha\left[\frac{\beta(v-2w+1/3)}{1+\beta(u-v)} + \ln\left[1+\beta(u-v)\right]\right] + \frac{\delta-\Delta^2}{\delta-1} - \ln(1-\delta) - \frac{\delta+\widehat{p}-(\Delta+\widehat{S})^2}{\widetilde{C}}, \tag{10}$$

with $u = 1/3+\widehat{C}/\pi$, $v = [2\arcsin(\delta/2)+\widehat{p}]/\pi$, and $w = [2\arcsin(\Delta/2)+\widehat{S}]/\pi$. Terms of order $(1/K)$ have been neglected on the r.h.s. of equation (10).

For $\alpha = \mathcal{O}(1)$, the saddle point equations yield two different types of solution: an unspecialized, committee symmetric branch with $\Delta = \delta = 0$ and specialized solutions with $\Delta, \delta > 0$. In the first case we find $\widehat{p} = \widehat{S} = 1$ and $\widehat{C} = 0$, with the generalization error $\epsilon_g = 1/3 - 1/\pi$ independent of both α and β. In the specialized case we get $\widehat{C} = 0$, $\widehat{p} = 1 - \delta$ and $\widehat{S} = 1 - \Delta$, while δ and Δ as functions of α and β can be determined only numerically.

Figure 1 (left) shows the generalization error as a function of α for three different values of β. The system undergoes a first order phase transition from a committee symmetric state ($R = S$) to a specialized solution with $R > S$. At constant training temperature, a locally stable, specialized configuration appears at a (β–dependent) value α_{min}. For $\alpha > \alpha_{glob}(\beta)$, the specialized solution becomes globally stable. Asymptotically, the corresponding generalization error ϵ_g and the training error ϵ_t decay like $1/(\alpha\beta)$ for large α. In contrast to the unspecialized phase, at a given α the generalization error always decreases with increasing β in the specialized phase.

It is important to note that an unspecialized configuration with constant ϵ_g remains locally stable for all α. For a given β the corresponding training error is constant with respect to the size of the training set, initially. At an additional critical value of α, the order parameter $\delta = q - p$ which measure correlations between students in different replicas assumes a non-zero value, whereas in this phase $\Delta = R - S$ remains zero for all α. This transition does not affect the generalization error but it does cause a first order transition to a slightly higher value of the training error ϵ_t. The training error continues to increase and approaches its asymptotic value $1/3 - 1/\pi$ while $\delta \to 1$ for $\alpha \to \infty$. The latter indicates that, asymptotically, a unique set of unspecialized student weights is chosen in all replicas. Due to the transition , the training and generalization error of the unspecialized configuration coincide in the limit $\alpha \to \infty$.

Figure 1 (right) displays $\epsilon_t(\alpha)$ for $\beta = 100$, where the above mentioned phase transition is located at $\alpha \approx 6.18$ where the training error jumps to a slightly larger value. The transition within the unspecialized phase occurs at values of α which increase rapidly with the training temperature, for instance at $\alpha \approx 139$ for $\beta = 10$ and $\alpha \approx 489$ for $\beta = 5$.

Our results parallel the findings of [15, 16] and [19] for large multilayer networks with threshold activation functions. We have found essentially the same qualitative behavior in the limit of infinite training temperature [20] and by applying the Annealed Approximation. However, the transition within the unspecialized phase cannot be identified in these simpler frameworks. It is further quite possible that even the replica symmmetric decription of this transition is incomplete. For threshold activation functions it was observed in [19] that this transition is affected by replica symmetry breaking, resulting in a lower critical value for α than predicted in replica symmetry and changing the nature of the transition from first to second order. A more detailed discussion of this transition for the present case will be given elsewhere [22].

The limit $\beta \to \infty$ is of particular interest and corresponds to potentially error free training with $\epsilon_t = 0$ for all α. Within our replica symmetric ansatz we find for $\beta \to \infty$ that the system switches from poor to perfect generalization ($\epsilon_g = 0$) at $\alpha_{min} = \alpha_{glob} = 1$, where the number of examples coincides with the number of adjustable weights in the network. This is a consequence of the smooth, differentiable nature of the input–output relation in this type of network. Such a transition to

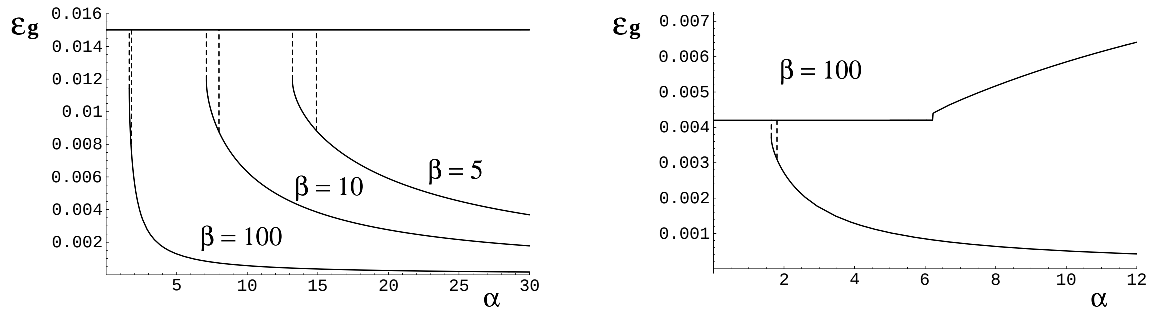

Figure 1: Generalization error and training error as functions of $\alpha = P/(KN)$.
Left panel: ϵ_g vs. α for three different training temperatures in the Gibbs ensemble. For each temperature, the leftmost dashed line indicates the occurance of a locally stable specialized state; the second vertical line marks α_{glob} where it becomes globally stable.
Right panel: ϵ_t vs. α for $\beta = 100$. An additional (first order) transition occurs at which ϵ_t begins to increase in the unspecialized solution while ϵ_g remains constant. For $\alpha \to \infty$ the training error approaches the value $\epsilon_t = \epsilon_g = 1/3 - 1/\pi$.

$\epsilon_g = 0$ is not observed in networks with threshold activation functions and continuous weights. The achievement of perfect generalization observed in networks with binary weights is due to a completely different mechanism, i.e. a freezing transition in the discrete configuration space, see e.g. [4, 5, 15].

It is of course a crucial question, whether our statistical mechanics treatment can give relevant results for practical applications. We have followed the standard approach and analysed a heat bath ensemble, i.e. a Gibbs distribution of network configurations. One might reproduce the Gibbs density in simulations of the learning process by use of an appropriate Langevin or Monte Carlo dynamics. However, these prescriptions are out of the question for practical applications in the case of continuous weights and differentiable outputs. Much faster and more effective methods exist, the most prominent one is certainly the so–called *backpropagation of error* [1, 2, 21].

When can we expect the statistical physics results to be relevant for such a practical prescription? Under certain restricting assumptions one can show, for instance, that stochastic gradient descent produces a stationary distribution which approximates a Gibbs density in the limit of infinitesimally small learning rates. This has been investigated in detail for simple systems in the vicinity of local energy minima [23, 24, 25]. But heat bath results can be interpreted in a broader context. Whenever an algorithm yields network configurations with a probability which depends exclusively on the training energy, one could in principle analyse an appropriate ensemble. All such ensembles, including the heat bath, refer to the same microcanonical density. Hence, for fixed energy, the system chooses among the same set of possible states with equal probablity and the same macroscopic features emerge. Stability properties, however, will depend strongly on the considered ensemble which has to be specified in order to locate a phase transition, for instance.

In Figure 2 (left panel) we have plotted the generalization error *vs.* the corresponding training error at $\alpha = 5$ by eliminating β in all saddle point solutions (regardless their local or global stability). Clearly, this dependence could be derived from the microcanonical density as well. According to the above reasoning, the same graph is valid for all procedures which produce configurations with a purely energy dependent probability.

In the following we demonstrate that the backpropagation algorithm appears to fulfill this requirement very well for a range of learning rates. To this end we have performed simulations of a stochastic version [2] with updates

$$\underline{J}_i^{t+1} = \sqrt{N} \;\; \left(\underline{J}_i^t - \eta \nabla_{\underline{J}_i} \epsilon(\{\underline{J}_i^t\}, \underline{\xi}^{\mu(t)})\right) \;/\; \left|\underline{J}_i^t - \eta \nabla_{\underline{J}_i} \epsilon(\{\underline{J}_i^t\}, \underline{\xi}^{\mu(t)})\right|. \qquad (11)$$

The current training example $\left\{\underline{\xi}^{\mu}(t), \tau^{\mu}\right\}$ is drawn randomly from the pool of $P = \alpha K N$ independent input-output pairs with probability $1/P$ at each time step. The learning rate η controls the step size of this stochastic gradient descent and the weights are normalized explicitly. The number of hidden units was $K = 10$ in all simulations shown in Figure 2.

In the course of learning one observes quasi–stationary states in which both ϵ_t and ϵ_g remain almost constant over a large number of updates. These are reminiscent of the *plateaus* found in on–line training of soft–committees [7-9] where each example is presented exactly once. We have identified plateaus according to a heuristic criterion in our simulations and determined the corresponding values of ϵ_g and ϵ_t. Note that several such states can be approached successively while learning with a fixed rate η. Details of the simulations will be explained in a forthcoming publication [22].

Figure 2 (left) shows the observed pairs of values (ϵ_g, ϵ_t) for learning rates $0.1 \leq \eta \leq 4.0$. Simulation results are in good aggreement with the theoretical analysis for a range of finite η. The algorithm favors configurations from either one of the two predicted phases, the occurance of states in between the specialized and the unspecialized branch is presumably due to the finite size of the system. The data with ϵ_g significantly larger than predicted correspond to plateaus found in simulations with relatively large η. Figure 2 (right panel) displays the observed values of ϵ_g *vs.* η. For small enough learning rates the predicted competition of specialized and unspecialized states is confirmed. For $\eta \gtrsim 2$, the value of ϵ_g can deviate significantly from the prediction, its sudden increase at $\eta \approx 5$ is reminiscent of the presence of a critical learning rate in on–line learning from a sequence of uncorrelated examples [7, 8].

As argued above, the location of a sharp transition from poor to good generalization cannot be expected to carry over from the heat bath to backpropagation results. We could not establish a relation between the control parameters β and η since the specific density of plateau states as produced by the training algorithm is unknown. Our simulations support, however, the assumption that it is purely energy dependent for reasonable η. The calculation of student–student overlaps provides further evi-

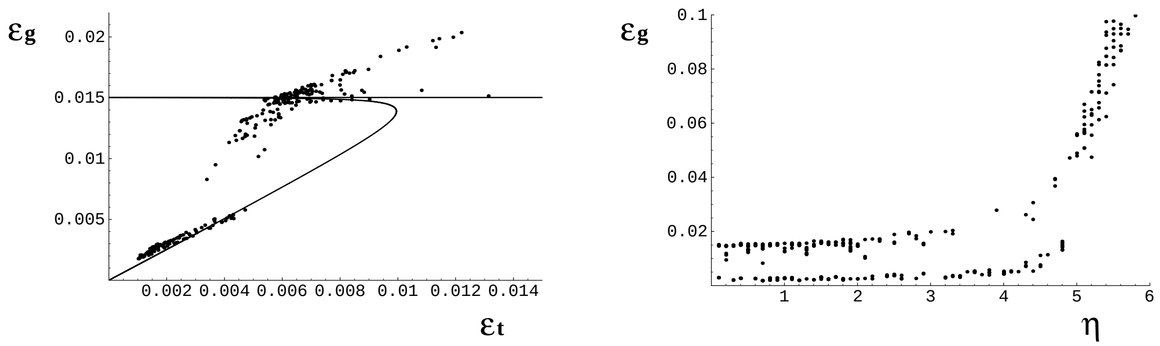

Figure 2: Stochastic backpropagation in a system with $N = 150$ and $K = 10$ at $\alpha = P/(KN) = 5$. Dots represent values found in plateau states of single runs, see the description in the text.
Left panel: Solid lines show ϵ_g *vs.* ϵ_t as obtained from the Gibbs ensemble by eliminating β and disregarding stability criteria. The dots display the data pairs observed in simulations with learning rates between $\eta = 0.1$ and 4.0.
Right panel: ϵ_g as found in plateau states as a function of the learning rate η. Note that for $\eta \tilde{>} 2$, ϵ_g can deviate significantly from the prediction. These results contribute to the set of points clearly above the horizontal line in the left panel. The generalization error increases drastically for $\eta \tilde{>} 5$ (not shown in the left panel).

dence for this hypothesis: we find the predicted scaling $C \propto 1/K^2$ for small learning rates, whereas $C = \mathcal{O}(1)$ independent of K for large η. Apparently, stochastic gradient descent with large learning rates prefers, among the states of a certain energy, those with highly correlated hidden unit vectors.

In summary, we have presented an analytic description of learning in large soft–committee machines by means of a replica symmetric treatment of the corresponding Gibbs ensemble. A characteristic feature of this model is the existence of a first order phase transition from poor to good generalization at a temperature dependent, critical size of the training set. In the limit of error free training ($\beta \to \infty$) the transition is to perfect generalization and occurs at $\alpha = 1$.

We expect our results to be relevant for a large class of practical algorithms which do not favor particular network configurations among those of equal training error. Simulations of learning by stochastic gradient descent with sufficiently small but finite learning rates show qualitative and quantitative agreement of plateau states with the theoretical predictions. This indicates that the considered training procedure provides network configurations with a purely energy dependent probability. The latter feature is lost if the learning rate is too large.

We will provide a more detailed study of stochastic backpropagation in a forthcoming publication. Future research will furthermore address learning from noisy examples, unrealizable rules, and the training of networks with a finite number of hidden units.

Acknowledgement:
We thank G. Reents and E. Schlösser for stimulating discussions and a critical reading of the manuscript.

References

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

[2] C. Bishop, *Neural Networks for Pattern Recognition*, (Clarendon, Oxford) 1995

[3] M. Opper and W. Kinzel, in *Models of Neural Networks III*, eds. E. Domany, J.L. van Hemmen, and K. Schulten, (Springer, Berlin) 1996

[4] S. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056, 1992

[5] T.L.H. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys. **65**, 499, 1993

[6] W. Kinzel, Phil. Mag. B **77**, 1455, 1998

[7] M. Biehl and H. Schwarze, J. Phys. A **28**, 643, 1995

[8] D. Saad and S.A. Solla, Phys. Rev. Lett **74**, 4337, 1995 and Phys. Rev. E **52**, 4225, 1995

[9] M. Biehl, P. Riegler, and C. Wöhler, J. Phys. A **29**, 4769, 1996

[10] R. Vicente and N. Caticha, J. Phys. A **30**, L559, 1997

[11] D. Saad and M. Rattray, Phys. Rev. Lett. **79**, 2578, 1997

[12] D. Saad (ed.), *On–line learning in neural networks*, Cambridge University Press, in press

[13] S. Bös, W. Kinzel, and M. Opper, Phys. Rev. E **47**, 1384, 1993

[14] K. Kang, J.–H. Oh, C. Kwon, and Y. Park, Phys. Rev. E **48**, 4805, 1993

[15] H. Schwarze and J. Hertz, Europhys. Lett. **21**, 785, 1993

[16] H. Schwarze, J. Phys. A **26**, 5781, 1993

[17] M. Opper, Phys. Rev. Lett. **72**, 2113, 1994

[18] B. Schottky, J. Phys. A **28**, 4515, 1995, and B. Schottky and U. Krey, J. Phys. A **30**, 8541, 1997

[19] R. Urbanczik, J. Phys A **28**, 7097, 1995, and Phys. Rev. E **58**, 2298, 1998

[20] M. Biehl, E. Schlösser, and M. Ahr, Europhys. Lett. **44**, 261, 1998

[21] Y. Chauvin and D.E. Rumelhart (eds.), *Backpropagation: Theory, Architecture, and Applications* (Lawrence Erlbaum, Hillsdale, NJ) 1995

[22] M. Ahr, M. Biehl, E. Schlösser, and R. Urbanczik, in preparation.

[23] L.K. Hansen, R. Pathria, and P. Salomon, J. Phys. A **26**, 63, 1993

[24] G. Radons, J. Phys. A **26**, 3455, 1993

[25] T. Heskes and B. Kappen, in *Mathematical Foundations of Neural Networks*, ed. J.G. Taylor, (Elsevier, Amsterdam) 1993

Appendix

We want to calculate a volume of the form

$$V(\mathsf{Q}) = \int d\mathsf{J}\ \boldsymbol{\delta}(N\mathsf{Q} - \mathsf{J}^\top \mathsf{J}) = \int d\mathsf{J} \prod_{a,b=1(a\le b)}^{n} \delta\left(NQ_{ab} - \underline{J}^a \cdot \underline{J}^b \right) \tag{12}$$

where Q is a symmetric, positive definite (n,n)-matrix of overlaps and J is the (N,n)-matrix which is composed of the n vectors $\underline{J}^a \in \mathbb{R}^N$.

For a suitable orthogonal (n,n)-matrix o and a diagonal (n,n)-matrix D one can write Q as $\mathsf{Q} = \mathsf{o}^\top \mathsf{D}\mathsf{D}\mathsf{o}$. We now apply the linear transformation $\mathsf{J} \to \mathsf{J}\mathsf{D}\mathsf{o}$ to the above integral. Its determinant is $\det \mathsf{D}^{\mathbf{N}}$ and we obtain

$$V(\mathsf{Q}) \;=\; \int d\mathsf{J}\ \boldsymbol{\delta}(\mathsf{o}^\top \mathsf{D}(N\mathsf{1} - \mathsf{J}^\top \mathsf{J})\mathsf{D}\mathsf{o})\ \ \det \mathsf{D}^N\ . \tag{13}$$

The Fourier representation of the δ-function yields

$$\boldsymbol{\delta}(\mathsf{o}^\top \mathsf{D}(N\mathsf{1} - \mathsf{J}^\top \mathsf{J})\mathsf{D}\mathsf{o}) = C_n \int d\hat{\mathsf{Q}}\ \exp\left(i \operatorname{Tr}\left[\hat{\mathsf{Q}} \mathsf{o}^\top \mathsf{D}(N\mathsf{1} - \mathsf{J}^\top \mathsf{J})\mathsf{D}\mathsf{o} \right] \right)\ . \tag{14}$$

The integration runs over symmetric (n,n)-matrices and $C_n = (2\pi)^{-n(n+1)/2} 2^{n(n-1)/2}$, where the second factor arises from the fact that the off-diagonal elements are counted twice in the trace. Using

$$\operatorname{Tr}\left[\hat{\mathsf{Q}} \mathsf{o}^\top \mathsf{D}(N\mathsf{1} - \mathsf{J}^\top \mathsf{J})\mathsf{D}\mathsf{o} \right] \;=\; \operatorname{Tr}\left[\mathsf{D}\mathsf{o}\hat{\mathsf{Q}}\mathsf{o}^\top \mathsf{D}(N\mathsf{1} - \mathsf{J}^\top \mathsf{J}) \right]$$

and transforming $\hat{\mathsf{Q}}$ via $\hat{\mathsf{Q}} \to \mathsf{o}^\top \mathsf{D}^{-1} \hat{\mathsf{Q}} \mathsf{D}^{-1} \mathsf{o}$ yields

$$\begin{aligned} \boldsymbol{\delta}(\mathsf{o}^\top \mathsf{D}(N\mathsf{1} - \mathsf{J}^\top \mathsf{J}\mathsf{D}\mathsf{o}) \;&=\; C_n \det \mathsf{D}^{-n-1} \int d\hat{\mathsf{Q}}\ \exp\left(i \operatorname{Tr}\left[\hat{\mathsf{Q}}(N\mathsf{1} - \mathsf{J}^\top \mathsf{J}) \right]) \right) \\ &=\; C_n \det \mathsf{D}^{-n-1} \boldsymbol{\delta}(N\mathsf{1} - \mathsf{J}^\top \mathsf{J}) \end{aligned} \tag{15}$$

and thus $V(\mathsf{Q}) = \det \mathsf{D}^{N-n-1} V(N\mathsf{1})$. Now $V(N\mathsf{1})$ is just a normalization constant and of course $\det \mathsf{D}^2 = \det \mathsf{Q}$. Hence, in the limit $N \to \infty$ with n of order one, one obtains

$$\frac{1}{N} \ln V(\mathsf{Q}) \;=\; \frac{1}{2} \ln \det \mathsf{Q} + \mathcal{O}(1).$$

The case where one considers an additional (N,m)-Matrix B of m teacher vectors and wants to evaluate $\int d\mathsf{J}\ \boldsymbol{\delta}(N\mathsf{Q} - \mathsf{J}^\top \mathsf{J})\, \boldsymbol{\delta}(N\mathsf{R} - \mathsf{J}^\top \mathsf{B})$ reduces to the above consideration by noting that the integral will not depend on the choice of B, as long as the matrix of teacher overlaps $\mathsf{T} = \mathsf{B}^\top \mathsf{B}/N$ is held fixed. Thus, one may in addition integrate over all B which have correlation matrix T.

For the system of K teacher vectors and nK replicated students we define the $(n+1)K$–dimensional square matrix of overlaps

$$\mathsf{C} = \begin{pmatrix} \mathsf{Q} & \mathsf{R} \\ \mathsf{R}^\top & \mathsf{T}. \end{pmatrix}$$

for which the above result yields equation (9).