Universität Würzburg

Institut für Theoretische Physik

Am Hubland, D–97074 Würzburg, Germany

Phase transitions in soft–committee machines

Michael Biehl, Enno Schlösser, Martin Ahr

Ref.: WUE-ITP-98-014
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

Phase transitions in soft–committee machines

M.Biehl, E. Schlösser, and M. Ahr
Institut für Theoretische Physik, Julius–Maximilians–Universität Würzburg
Am Hubland, D–97074 Würzburg, Germany

Abstract

Equilibrium statistical physics is applied to the off–line training of layered neural networks with differentiable activation functions. A first analysis of soft–committee machines with an arbitrary number (K) of hidden units and continuous weights learning a perfectly matching rule is performed. Our results are exact in the limit of high training temperatures ($\beta \to 0$). For $K = 2$ we find a second order phase transition from unspecialized to specialized student configurations at a critical size P of the training set, whereas for $K \geq 3$ the transition is first order. The limit $K \to \infty$ can be performed analytically, the transition occurs after presenting on the order of NK/β examples. However, an unspecialized metastable state persists up to $P \propto NK^2/\beta$.

Statistical physics provides tools for the investigation of learning processes in adaptive systems such as feedforward neural networks [1]. The by now standard analysis of off–line or batch learning from a fixed set of example data is based on the interpretation of training as a stochastic process which leads to a properly defined thermal equilibrium. It has been applied with great success to simple networks like single layer preceptrons or specific multilayer architectures with binary threshold units. See e.g. [2-4] for reviews and [5] for a discussion focussed on phase transitions in neural networks.

In a somewhat different framework, the theory of on–line learning, it is assumed that a temporal sequence of independent examples is provided by the environment. For large networks, the inherently stochastic learning dynamics is described exactly by a set of deterministic differential equations, see e.g. [6] for an up to date overview. This approach has made possible the recent progress with respect to layered networks with continuous node activations, see e.g. [7-11] and references therein. Such systems are relevant for applications as they can implement non–trivial regression schemes and practical training algorithms are available [1, 12].

In this Letter we present an analysis of off–line learning in two–layered architectures with differentiable transfer functions by means of equilibrium statistical physics. The considered model exhibits phase transitions in the learning process, i.e. a discontinuous dependence of the student performance on the number of examples. These transitions are the counterparts of quasi–stationary plateau states observed in on–line dynamics [8, 9] and are due to the same inherent symmetries.

The hidden unit specialization studied here is different from the sudden achievement of perfect generalization described in [13] for single, continuous nodes. Previous studies of phase transitions in multilayered neural networks were mainly concerned with networks of binary threshold units and/or non–overlapping architectures, see for instance [14-20]. Off–line learning in soft–committee machines was first studied by Kang et al. in [14], where the analysis was restricted to networks with binary weights and the limit $K \to \infty$ of infinitely many hidden units.

In the following, we will investigate the learning of a rule in a fully connected two–layered neural network with total output

$$\sigma(\boldsymbol{\xi}) \;=\; \tfrac{1}{\sqrt{K}} \textstyle\sum_{j=1}^{K} g(x_j) \quad \text{where} \quad x_j = \tfrac{1}{\sqrt{N}} \mathbf{J}^{(j)} \cdot \boldsymbol{\xi} \tag{1}$$

upon presentation of an N–dimensional input vector $\boldsymbol{\xi}$. Given the hidden unit activation $g(x)$, the adaptive weights $\mathbf{J}^{(i)} \in \mathbb{R}^N$ define the input–output relation. The term *soft–committee machine* has been coined for this type of network [7, 8], as it can be interpreted as a continuous version of the thoroughly studied (*hard*) committee of threshold hidden units (see [14-16] and references therein). Here, the weights of the linear hidden–to–output relation are fixed to the particular value $1/\sqrt{K}$ which differs from the usual scaling considered in the analysis of on–line learning in soft–committees [8, 9].

Off–line learning in networks of type (1) with binary weights $\mathbf{J}^{(i)} \in \{+1, -1\}^N$ has been studied before in the limit $K \to \infty$ by Kang *et al.* [14]. Here we focus on continuous weights and finite K, but include a discussion of large machines for completeness. Note that the specific form of Eq. (1) yields outputs of finite magnitude in the limit $K \to \infty$.

Most frequently $g(x)$ is taken to be a sigmoidal function of its argument, e.g. the hyperbolic tangent. We choose the similiar but more convenient function $g(x) = \mathrm{erf}(x/\sqrt{2})$ which simplifies the mathematical treatment to a great extent yet should not alter the basic features of the model otherwise [7, 8].

In this Letter, we restrict our analysis to scenarios in which the unknown rule $\tau(\boldsymbol{\xi})$ can be parametrized through a teacher network of perfectly matching architecture and size (K) with weight vectors $\mathbf{B}^{(j)}$. Further, we assume that the $\mathbf{B}^{(j)}$ are normalized vectors of length $\sqrt{N}$ with i.i.d. random components and accordingly impose a normalization $(\mathbf{J}^{(j)})^2 = N$ on the student vectors. Since the continuous student output, Eq. (1), depends explicitely on the length of the weight vectors, this latter constraint corresponds to significant *a priori* knowledge of the rule's structure. We will discuss later how this restriction could be relaxed.

The off-line training process is based on a fixed set of examples $\{\boldsymbol{\xi}^\mu, \tau(\boldsymbol{\xi}^\mu)\}$ $(\mu =$

$1, 2, \ldots, P$) and is guided by the minimization of the cost function or training error

$$\varepsilon_t \,=\, \tfrac{1}{P} \sum_{\mu=1}^{P} \, \tfrac{1}{2} \left[\sigma(\boldsymbol{\xi}^\mu) - \tau(\boldsymbol{\xi}^\mu) \right]^2 . \tag{2}$$

Thus, learning is formulated as an optimization problem. In networks with differentiable activation functions the (approximate) solution could be found by use of gradient descent or similiar methods. The prominent *backpropagation of error*, for example, is widely used in practice [1, 12]. Often, the actual global minimum of (2) cannot be identified by such algorithms, even if the rule is perfectly learnable. Practical learning prescriptions can be trapped in local minima of the training error or stop as soon as a satisfactory performance is achieved.

After training, the quadratic error measure can also be utilized to quantify the success of learning in terms of the so–called generalization error

$$\varepsilon_g \,=\, \tfrac{1}{2} \left\langle \left[\sigma(\boldsymbol{\xi}) - \tau(\boldsymbol{\xi}) \right]^2 \right\rangle . \tag{3}$$

Here $\langle \ldots \rangle$ denotes an empirical average over a test set of (new) examples or over the distribution of random inputs, which is assumed to be known in the model. Throughout the following we take the components of all inputs $\boldsymbol{\xi}$ to be i.i.d. random variables with zero mean and unit variance. In the thermodynamic limit $N \to \infty$ the quantities $x_j = \mathbf{J}^{(j)} \cdot \boldsymbol{\xi} / \sqrt{N}$ and $y_j = \mathbf{B}^{(j)} \cdot \boldsymbol{\xi} / \sqrt{N}$ become zero mean correlated Gaussian variables by means of the central limit theorem. Their joint density is fully characterized by the covariances

$$\langle x_i x_j \rangle = \tfrac{1}{N} \mathbf{J}^{(i)} \!\cdot\! \mathbf{J}^{(j)} = Q_{ij}, \;\; \langle x_i y_j \rangle = \tfrac{1}{N} \mathbf{J}^{(i)} \!\cdot\! \mathbf{B}^{(j)} = R_{ij}, \;\; \text{and} \;\; \langle y_i y_j \rangle = \tfrac{1}{N} \mathbf{B}^{(i)} \!\cdot\! \mathbf{B}^{(j)} = \delta_{ij} \tag{4}$$

where all diagonal $Q_{ii} = 1$. Hence, the average in Eq. (3) reduces to a $2K$–dimensional Gaussian integral which can be performed analytically for the specific activation function $g(x) = \mathrm{erf}(x/\sqrt{2})$. The result depends only on the *order parameters* $\{R_{ij}, Q_{ij}\}$:

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

and was derived in [8] for arbitrary network sizes and more general scenarios.

Following the standard statistical physics approach to off–line learning we consider a Gibbs ensemble of networks which is characterized by the partition function

$$Z \,=\, \textstyle\int d\mu(\{\mathbf{J}^{(i)}\}) \, \exp\left[-\beta \, P \, \varepsilon_t \right] . \tag{6}$$

The extensive energy $P\,\varepsilon_t$ is defined in Eq. (2) and the measure $d\mu$ limits the integration to the region in weight space where all $(\mathbf{J}^{(i)})^2 = N$. The formal temperature $1/\beta$ controls the thermal average of the energy in equilibrium. Equivalently, it fixes the amount of noise which is present in a corresponding stochastic training process.

Typical properties of the equilibrium state can be calculated from the associated free energy on average over the quenched randomness contained in the training data $I\!D = \{\boldsymbol{\xi}^\mu, \tau(\boldsymbol{\xi}^\mu)\}$ (denoted as $\langle \ldots \rangle_{I\!D}$). The evaluation of the quenched free energy $-\langle \ln Z \rangle_{I\!D} / \beta$ requires, in general, the application of the replica method. It becomes rather involved in models of the complexity considered here, see e.g. [14, 15, 19] for a treatment of networks with threshold units. In order to obtain first results for the off–line training of soft–committee machines we resort to the simplifying limit of high temperatures. This strategy has proven useful for gaining first insights in a variety of models, see [3, 4, 20] for details and example applications.

In the limit $\beta \to 0$, we can replace $\langle \ln Z \rangle_{I\!D}$ with $\ln \langle Z \rangle_{I\!D}$, i.e. the annealed approximation which circumvents the replica formalism becomes exact. Further, the average over the training data factorizes with respect to the example inputs and one obtains

$$- \langle \ln Z \rangle_{I\!D} / N = \beta f(\{Q_{ij}, R_{ij}\}) = (\beta P/N) \;\; \varepsilon_g(\{Q_{ij}, R_{ij}\}) - s(\{Q_{ij}, R_{ij}\}) \qquad (7)$$

where the r.h.s. is to be minimized with respect to the order parameters. Non–trivial results can only be expected if the effective temperature $(\beta P/N)$ is of order 1, i.e. the high training temperature has to be compensated for by a large number of examples $P \propto N/\beta$. Consequently, training energy ε_t and generalization error ε_g coincide in this limit. Eq. (7) contains the entropy term

$$s(\{Q_{ij}, R_{ij}\}) = (1/N) \ln \int \textstyle\prod_j d^N \mathbf{J}^{(j)} \prod_{i,j} \left(\delta(\mathbf{J}^{(i)} \cdot \mathbf{B}^{(j)} - N R_{ij}) \delta(\mathbf{J}^{(i)} \cdot \mathbf{J}^{(j)} - N Q_{ij}) \right) \qquad (8)$$

which can be evaluated by means of a saddle point integration after rewriting the δ–functions in their integral representation. One obtains $s(\{Q_{ij}, R_{ij}\}) = \ln[\det \mathcal{C}]/2 +$ const, where $\mathcal{C}$ is the $(2K \times 2K)$–matrix of all cross– and self–overlaps of the vectors $\{\mathbf{J}^{(j)}, \mathbf{B}^{(j)}\}$. The constant term is independent of the order parameters and therefore irrelevant.

In order to proceed with the analysis, we assume that the equilibrium student configuration is symmetric with respect to the hidden–units:

$$R_{ij} = R\,\delta_{ij} + S\,(1 - \delta_{ij}) \quad \text{and} \quad Q_{ij} = \delta_{ij} + C\,(1 - \delta_{ij}). \qquad (9)$$

This assumption reflects the symmetry of the rule, yet allows for the *specialization* of student nodes: for $R > S$ each of them has achieved a larger overlap with exactly one of the teacher vectors. In the limiting case $R = 1, S = C = 0$ the student is identical with the teacher and generalizes perfectly ($\varepsilon_g = 0$). Now entropy (apart from irrelevant constants) and generalization error read

$$s = \frac{1}{2} \ln \left[1 + (K-1)C - ((R-S) + KS)^2 \right] + \frac{K-1}{2} \ln \left[1 - C - (R-S)^2 \right] \qquad (10)$$

$$\varepsilon_g = \frac{1}{3} + \frac{K-1}{\pi} \left[\arcsin\left(\frac{C}{2}\right) - 2 \arcsin\left(\frac{S}{2}\right) \right] - \frac{2}{\pi} \arcsin\left(\frac{R}{2}\right). \qquad (11)$$

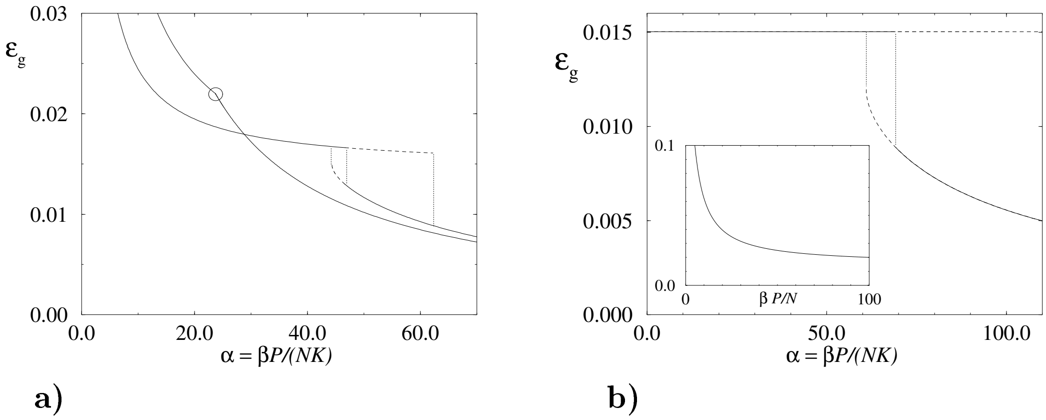

Figure 1: a) Learning curves for small K. The system with $K = 2$ hidden units undergoes a second order transtion at $\alpha_c^{(2)} \approx 23.7$ leading to the kink in $\varepsilon_g(\alpha)$ (marked by the circle). For $K \geq 3$ the transition is first order. Here we display the case $K = 5$, solid lines correspond to the globally stable solution, dashed lines denote local minima, ($\alpha_s^{(5)} \approx 44.3$, $\alpha_c^{(5)} \approx 46.6$ and $\alpha_d^{(5)} \approx 62.8$).
b) The learning curve in the limit $K \to \infty$. The specialized state is a local minimum (dashed) for $\alpha \geq \alpha_s^{\infty} \approx 61.0$ and becomes globally stable (solid) at $\alpha_c^{\infty} \approx 69.1$. The unspecialized configuration remains locally stable up to $\alpha_d^K = 4\pi K$ for large K. The inset displays the initial (unspecialized) decrease of ε_g with $\beta P/N$.

Defining $\alpha = \beta P/(NK)$, the rescaled number of examples per student weight, we have minimized $f = \alpha(K\varepsilon_g) - s$ with respect to the three order parameters R, C, S numerically for different network sizes.

For $K = 2$ we find that the equilibrium configuration is characterized by $R = S$ for $\alpha < \alpha_c^{(2)} \approx 23.7$, whereas above this critical value the only solution obeys $\Delta = |R - S| \neq 0$. The system undergoes a second order phase transition and the specialization Δ increases close to the critical point like $(\alpha - \alpha_c^{(2)})^x$ where $x = 0.5$. Figure 1 shows that this spontaneous symmetry breaking translates into a kink in the learning curve $\varepsilon_g(\alpha)$.

The picture is qualitatively different for all $K \geq 3$ where we observe a first order transition. Again, for small α the equilibrium solution is unspecialized ($R = S$). Then, for $\alpha \geq \alpha_s^{(K)}$ a locally stable solution with $\Delta > 0$ and significantly lower generalization error appears. It becomes the global minimum at a critical value $\alpha_c^{(K)}$ where the free energies of the two solutions coincide. Finally, for $\alpha \geq \alpha_d^{(K)}$, the local minimum with zero specialization disappears. Figure 1 (a) shows the learning curve for $K = 5$ as an example. A more detailed description of the dependence of order parameters on α will be given in a forthcoming publication.

The difference in behavior between networks with $K \geq 3$ and $K = 2$ is due to the higher degree of symmetry in the latter case. The permutation symmetry of hidden units results in a free energy, Eqs. (10,11), which is invariant under exchange of R and S only for $K = 2$.

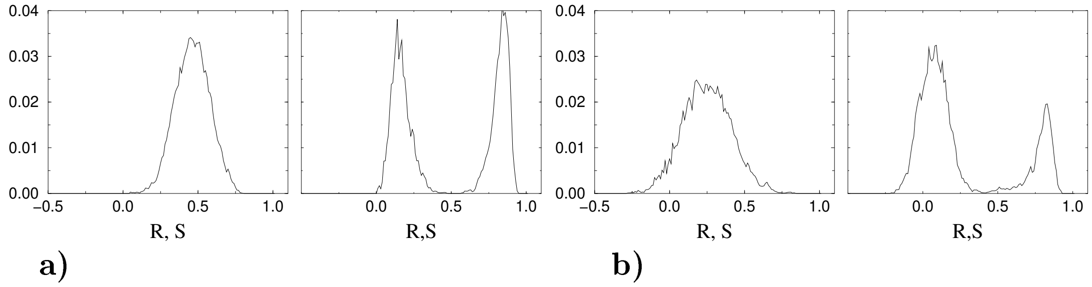

Figure 2: Empirical densities of observed student teacher overlaps in Monte Carlo simulations of the learning process at inverse temperature $\beta = 0.8$. The system size is $N = 50$, after 50000 Monte Carlo steps allowed for equilibration, another 50000 were performed while sampling the densities. Panel a: $K = 2$, $\alpha = 15$ (left) and $\alpha = 40$ (right), panel b: $K = 4$, $\alpha = 30$ (left) and $\alpha = 60$ (right).

The apparently same qualitative difference between learning curves for $K = 2$ and $K \geq 3$ has been found in non–overlapping parity and committee machines [17, 18]. We would like to emphasize, however, that such tree–like architectures are not invariant under permutation of hidden units. There, different types of symmetry are responsible for the occurance and nature of the observed transitions.

We have performed continuous Monte Carlo simulations of the learning process at relatively low temperatures. Figure 2 (a) displays the density of observed student teacher overlaps for $K = 2$ close to equilibrium in the specialized and unspecialized phase ($\alpha = 15$ and $\alpha = 40$ respectively). In panel (b) the corresponding histograms are plotted for a network with $K = 4$ at $\alpha = 30$ and $\alpha = 60$. Note that the total weight of overlaps S should be a factor $(K - 1)$ larger than the contribution of type R when hidden unit symmetry holds. This is confirmed very well and shows that the simplifying assumption (9) is justified also for moderately low temperatures.

Our preliminary simulations do not allow to demonstrate clearly that the transition is first or second order respectively. This could be done, for instance, by an investigation of hysteresis effects close to the critical number of examples [16], which is beyond the scope of this Letter. Note that the equilibration of the continuous degrees of freedom in our model at high temperature is significantly more difficult than for binary weights, where simple Metropolis spin flip techniques can be applied [14, 16].

The behavior of very large networks with binary weights $\mathbf{J}^{(i)} \in \{+1, -1\}^N$ in the limit $K \to \infty$ (but $K \ll N$) has been studied in [14] for the transfer function $g(x) = \tanh(x)$ within the annealed approximation. We proceed with a discussion of this limit in our model with continuous weights. The analysis simplifies significantly due to the choice $g(x) = \mathrm{erf}(x/\sqrt{2})$. Different regimes have to be distinguished, in analogy to previous studies of large multilayered networks [14-16,19]. First, we assume that $\beta P/N = \alpha K = \mathcal{O}(1)$ and find that only an unspecialized solution exists in this initial phase of the learning process. The ansatz $R = S = \widehat{S}/K$ and

$C = \widehat{C}/(K-1)$ yields a simplified free energy of the form $f = (\alpha K)(\widehat{C}/2 - \widehat{S}) - \ln[1 + \widehat{C} - \widehat{S}^2]/2 + \widehat{C}/2 + \mathcal{O}(1/K^2)$ which is minimized for

$$\widehat{S} = \alpha K \,/(\alpha K + \pi) \quad \text{and} \quad \widehat{C} = -\pi \alpha K \,/(\alpha K + \pi)^2 \, . \tag{12}$$

The corresponding generalization error is plotted in the inset of Figure 1 (b) *vs.* αK and approaches the stationary non–zero value $\varepsilon_g = 1/3 - 1/\pi$ for large αK.

Consequently, the unspecialized configuration is given by $\widehat{S} = 1$ and $\widehat{C} = 0$ to first order in $1/K$ when $\alpha = \beta P/(NK)$ is of order one. However, specialization is possible in this regime and including a nonzero $\Delta = R - S = \mathcal{O}(1)$ in the above ansatz yields a solution with $\widehat{S} = KS = 1 - \Delta$ and $\widehat{C} = KC = 0$ to first order. The corresponding specialization Δ is the largest positive solution of

$$\alpha = \frac{\pi \Delta \sqrt{4 - \Delta^2}}{(1 - \Delta^2)(2 - \sqrt{4 - \Delta^2})} \tag{13}$$

which can only be satisfied for $\alpha > \alpha_s^\infty \approx 60.99$. Comparing the free energies of specialized and unspecialized solution yields the critical value $\alpha_c^\infty \approx 69.09$.

In the specialized phase, the generalization error decreases asymptotically like $\varepsilon_g \propto \alpha^{-1}$ for $\alpha \to \infty$ which holds true also for general (small) K. The local minimum with $\Delta = 0$ remains locally stable even in the limit $\alpha \to \infty$. However, we can show that the metastable state disappears at $\alpha_d^{(K)} = 4\pi K$ when K is large. The latter result is a very good approximation for K as small as 4 or 5 already.

In summary, we have used equilibrium statistical physics to analyse off–line learning in two–layered soft–committee machines. First results for finite K are obtained in the high temperature limit which allows to calculate the quenched free energy analytically. Specifically, we have studied networks with K hidden units learning a perfectly matching rule. For $K = 2$ we find a second order phase transition from unspecialized to specialized student configuration at a critical number of examples. In secenarios with $K \geq 3$ the transition from poor to good generalization is first order.

It is interesting to note that also the investigation of on–line learning in soft–committees revealed qualitatively different features for $K = 2$ and $K \geq 3$, see the discussion of the fixed point structure in [9]. A detailed discussion of this correspondence will be given elsewhere.

The analysis of the limit $K \to \infty$ shows that a critical number $P \propto NK/\beta$ of examples is needed for specialization, in agreement with earlier findings of Kang et al. [14]. As a novel result we observe that the metastable unspecialized state persists up to $P = 4\pi NK^2/\beta$ for large K.

Our results (for $K \geq 3$) seem to correspond to a large extent the findings of [14-16] for hard committee machines. We do not expect this correspondence to extend to very low temperatures, however. In fact, for $T \to 0$, one could speculate that a transition to perfect generalization might occur after presenting on the order of

$\mathcal{O}(NK)$ examples to the continuous and differentiable student. Note in this context, that the computationally cheap on–line gradient descent already realizes an asymptotic decay of ϵ_g which is exponential in $\alpha = P/(NK)$ for all (finite) K [8, 9]. In the case of a single unit, the fast on–line asymptotics is known to parallel a transition to $\epsilon_g = 0$ in off–line learning [13]. It is unclear if this analogy carries over to $K > 1$, and even more so, if any practical algorithm could realize the sudden achievement of generalization.

It will be necessary to complete the picture by extending the analysis into the low temperature regime within the annealed approximation or the replica formalism. The application of the latter to the case of large K should follow the work of [15, 19]. Further investigations will address unlearnable rules as well as over–sophisticated students [19, 20]. It might also be interesting to extend our analysis to networks with multiple outputs [21] and to consider an optimized low temperature training prescription [22]. The introduction of an additional weight decay term [1] to the training energy allows to relax the somewhat unnatural *a priori* normalization of student weights. First results concern a single unit and show a non–trivial dependence of the performance on the weight decay parameter.

We would like to thank W. Kinzel, G. Reents, and R. Urbanczik for stimulating discussions, and J. Hertz for drawing our attention to the analysis of large soft–committees in [14].

References

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

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

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

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

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

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

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

[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] H. Schwarze, M. Opper, and W. Kinzel, Phys. Rev. A **46**, R6185, 1992

[21] K. Kang, J.–Oh, C. Kwon, and Y. Park, Phys. Rev. E **54**, 1811, 1996

[22] O. Winther, B. Lautrup, J.–B. Zhang, Phys. Rev. E **55**, 836, 1997