Weight decay induced phase transitions in multilayer neural networks

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

Abstract

We investigate layered neural networks with differentiable activation function and student vectors without normalization constraint by means of equilibrium statistical physics. We consider the learning of perfectly realizable rules and find that the length of student vectors becomes infinite, unless a proper weight decay term is added to the energy. Then, the system undergoes a first order phase transition between states with very long student vectors and states where the lengths are comparable to those of the teacher vectors. Additionally in both configurations there is a phase transition between a specialized and an unspecialized phase. An anti-specialized phase with long student vectors exists in networks with a small number of hidden units.

Statistical physics has been applied successfully to the investigation of equilibrium states of neural networks. [1, 2] The by now standard analysis of off-line training from a fixed training set is based on the interpretation of training as a stochastic process which leads to a well-defined thermal equilibrium. Investigations of perceptrons [3, 4, 5] or committee machines [6, 7, 8, 9, 10] have widely improved understanding of learning in neural networks. Meanwhile these studies are being extended to the more application relevant scenario of networks with continuous activation function and output. [11, 12, 13]

The soft-committee machine is a two-layered neural network which consists of a layer of K hidden units, all of which are connected with the entire N-dimensional input $\underline{\xi}$. The total output σ ist proportional to the sum of outputs of all hidden units:

$$\sigma(\underline{\xi}) = \frac{1}{\sqrt{K}} \sum_{j=1}^{K} g(x_j) \;\text{ where }\; x_j = \frac{1}{\sqrt{N}} \underline{J}_j \cdot \underline{\xi} \tag{1}$$

where the weights of the j-th hidden unit are represented by the N-dimensional vector $\underline{J}_j$. We investigate learning of a perfectly matching rule parametrized by a teacher network of the same architecture with output τ and orthogonal vectors $\underline{B}_j$, which we assume to have the length $\sqrt{N}$. The transfer function $g(x)$ is taken to be a sigmoidal function, e.g. the error function. Networks of this type have been studied in the limit of high temperature [11], the annealed approximation [13], and by means of the replica formalism [12]. All these studies imposed the simplifying condition that the order parameters $Q_{ij} = \underline{J}_i \cdot \underline{J}_j/N$ are restricted to the value 1 for $i = j$, so the length of the student vectors is fixed to that of the teacher vectors. This system shows a phase transition between an unspecialized configuration, where the student-teacher overlaps $R_{ij} = \underline{J}_i \cdot \underline{B}_j/N$ are identical for all i, j and a specialized configuration where $R_{ii} \neq R_{ij}$ for $i \neq j$. However, constraining the student lengths implies significant *a priori* knowledge of the rule which is not available in practical applications. So, in this paper we want to obtain first results for soft committee machines which determine student lengths in the course of learning.

Learning is guided by the minimization of the training error

$$\epsilon_t = \frac{1}{P} \sum_{\mu=1}^{P} \frac{1}{2} \left(\sigma(\underline{\xi}_\mu) - \tau(\underline{\xi}_\mu) \right)^2 \tag{2}$$

where P is the number of examples used for training. After training, the success of learning can be quantified by an average of the quadratic error measure over the distribution of possible inputs, the so-called generalization error:

$$\epsilon_g = \frac{1}{2} \left\langle \left(\sigma(\underline{\xi}) - \tau(\underline{\xi}) \right)^2 \right\rangle_{\underline{\xi}} \tag{3}$$

Following the standard statistical physics approach, we consider a Gibbs ensemble, which is characterized by the partition function $Z = \int d\mu(\{J_i\}) \exp(-\beta H(\{\underline{J}_i\}))$ with a formal temperature $1/\beta$ which controls the thermal average of energy in the equilibrium. The extensive energy H is a function of the training error, the standard choice being $H = P\epsilon_t$. Typical equilibrium properties are calculated from the associated quenched free energy $-(1/\beta)\langle \ln Z \rangle =: fN$ where the average is performed over the random set of training examples. The evaluation of $\langle \ln Z \rangle$ in general requires the rather involved replica formalism. To obtain first results we consider the simplifying *high-temperature limit* $\beta \to 0$ [3, 4]. The calculation of equilibrium states is guided by minimization of $\beta f = \tilde{\alpha} K \epsilon_g - s$. Here $\tilde{\alpha} = \beta P/(NK)$ ist the rescaled number of examples, which we assume to be $\mathcal{O}(1)$ and s the entropy per degree of freedom with order parameters held fixed. The latter is given by

$$s = 1/2 \ln \det \underline{\underline{C}} + \text{irrelevant const.} \tag{4}$$

where $\underline{\underline{C}}$ is the $2K \times 2K$-matrix of all cross- and self-overlaps of student and teacher vectors. Equation 4 is of quite general validity and can be derived by means of a saddle point integration from the definition of the entropy. In [12] a simpler derivation is presented.

Here we assume the components of all examples to be independent random numbers with mean zero and unit variance. Then, in the thermodynamic limit $N \to \infty$ the generalization error can be calculated analytically, if we choose the activation function $g(x) = \mathrm{erf}(x/\sqrt{2})$ [14, 15] which is very similar to the more popular hyperbolic tangent, so the basic features of the model should not be altered:

$$\epsilon_g = \frac{1}{6} + \frac{1}{K\pi} \sum_{i,k=1}^{K} \left[\sin^{-1}\left(\frac{Q_{ik}}{\sqrt{(1+Q_{ii})(1+Q_{kk})}} \right) - 2\sin^{-1}\left(\frac{R_{ik}}{\sqrt{2(1+Q_{ii})}} \right) \right] \tag{5}$$

In the following, we will first investigate the simplest case $K = 1$, i.e. a network consisting of one single unit to show the basic principles. Then we will study networks with arbitrary K and finally investigate the limit $K \to \infty$ of very large networks.

In the $K = 1$ case equations 4 and 5 read:

$$\epsilon_g = \frac{1}{6} + \frac{1}{\pi}\sin^{-1}\left(\frac{Q}{1+Q}\right) - \frac{2}{\pi}\sin^{-1}\left(\frac{R}{\sqrt{2(1+Q)}}\right) \tag{6}$$

$$s = \frac{1}{2}\ln\left(Q - R^2\right) \tag{7}$$

Trying to minimize $\tilde{\alpha}\epsilon_g - s$, we find that ϵ_g remains of order 1 for arbitrary R, Q while s becomes infinite for $Q \to \infty$, yielding $f \to -\infty$. This means that in thermal equilibrium the length of the student vector increases to infinity, while its overlap with the teacher becomes irrelevant. Of course, this is not the desired result of training. The method of choice to avoid this behavior, is to "punish" configurations with large Q with an additional energy called "weight decay". This is a method of *regularization* which is widely used in practice in order to improve the generalization ability of feedforward neural networks [1]. So we introduce $H = P\epsilon_t + \lambda N Q$ [16, 17, 19, 20, 21, 22] and obtain $\beta f = \tilde{\alpha}\epsilon_g + \tilde{\lambda} Q - s$ with $\tilde{\lambda} = \beta\lambda$ which has to be minimized w.r.t. R and Q. In Figure 1 we show ϵ_g as a function of the rescaled number of examples, $\tilde{\alpha}$ for $\tilde{\lambda} = 0.001$. For small $\tilde{\alpha}$ the network is in a state with large Q (and ϵ_g). For $\tilde{\alpha} \geq 12$ a second state with small Q and small ϵ_g exists, which becomes globally stable at $\tilde{\alpha} \approx 15$. At $\tilde{\alpha} \approx 21.6$ the state with large Q becomes even locally unstable. We remark that this phase transition is solely due to the differentiable nature of the activation function, which causes the energy to depend on the length of the student vector, and does not occur in the simple perceptron. It was also not found for the simpler linear unit with $g(x) = x$ [19], where the training error is more sensitive to a mismatched Q than in the case of a bounded, saturating transfer function. The phase transition disappears for $\tilde{\lambda} \geq 0.006$.

We have performed continuous Monte-Carlo simulations of a Metropolis-like learning process of the single unit. The results shown in Figure 1 confirm our theoretical results.

In order to extend our analysis to networks with $K \geq 2$ we assume the network configuration to be *site-symmetric* with respect to the hidden units so the order

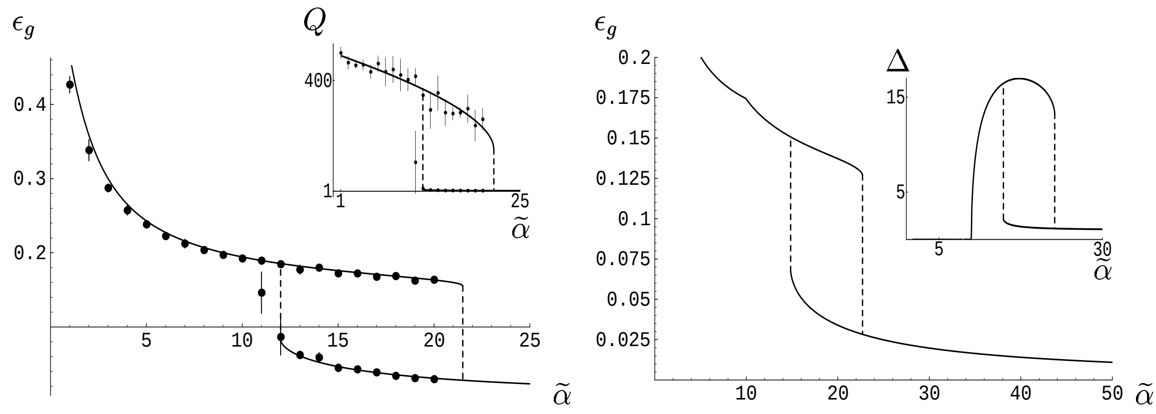

Figure 1: left panel: $\epsilon_g(\tilde{\alpha})$ and $Q(\tilde{\alpha})$ (inset) as obtained analytically (solid line) and results of Monte-Carlo simulations (dots) for $K = 1$, $\tilde{\lambda} = 0.001$ and $\beta = 0.2$. (system size $N = 100$, averages over 5 runs with 10000 M.C. steps each, 5000 of which were used for equilibration, 5000 for sampling measurements) We get two locally stable states with different student lengths for some $\tilde{\alpha}$, depending on the starting value of the student vector. We have used the same strategy as in [18] to obtain the hysteresis behaviour. right panel: $\epsilon_g(\tilde{\alpha})$ and $\Delta(\tilde{\alpha})$ (inset) for $K = 2$ and $\tilde{\lambda} = 0.001$.

parameters fulfill the conditions $R_{ij} = R\delta_{ij} + S(1-\delta_{ij})$ and $Q_{ij} = Q\delta_{ij} + C(1-\delta_{ij})$. This assumption reflects the symmetry of the rule yet allows for specialization of the student, as student overlaps with teacher vectors can yield different values for $i = j$ and $i \neq j$. Now generalization error and entropy read:

$$
\begin{aligned}
\epsilon_g &= \tfrac{1}{6} + \tfrac{1}{\pi}\sin^{-1}\left(\tfrac{Q}{1+Q}\right) + \tfrac{K-1}{\pi}\left[\sin^{-1}\left(\tfrac{C}{1+Q}\right) - 2\sin^{-1}\left(\tfrac{S}{\sqrt{2(1+Q)}}\right)\right] \\
&\quad -\tfrac{2}{\pi}\sin^{-1}\left(\tfrac{R}{\sqrt{2(1+Q)}}\right) \qquad (8) \\
s &= \tfrac{1}{2}\ln\left[(K-1)C + Q - (R + (K-1)S)^2\right] \\
&\quad + \tfrac{K-1}{2}\ln\left[Q - C - (R-S)^2\right] \qquad (9)
\end{aligned}
$$

The weight decay term introduced for the single unit generalizes naturally to $\lambda \sum_{i=1}^{K} Q_{ii}$, so the free energy becomes $\beta f = \tilde{\alpha} K \epsilon_g + \tilde{\lambda} K Q - s$ in a site symmetric state. Numerical minimization leads to the results shown in Figure 1 and 2 for $K = 2$ and $K = 3$. In addition to the first order phase transition already observed at the single unit, which connects states with different lengths of student vectors, we observe transitions between phases which are characterized by the parameter $\Delta := R - S$ indicating specialization features. As both transitions are due to independent mechanisms, namely on the one hand a change of student vector *lengths* and on the other hand an alteration of their *directions*, specialized ($\Delta > 0$) and unspecialized ($\Delta = 0$) phases can exist both in the large-Q configuration and in the small-Q regime. Indeed for $K \geq 3$ first order transitions between specialized

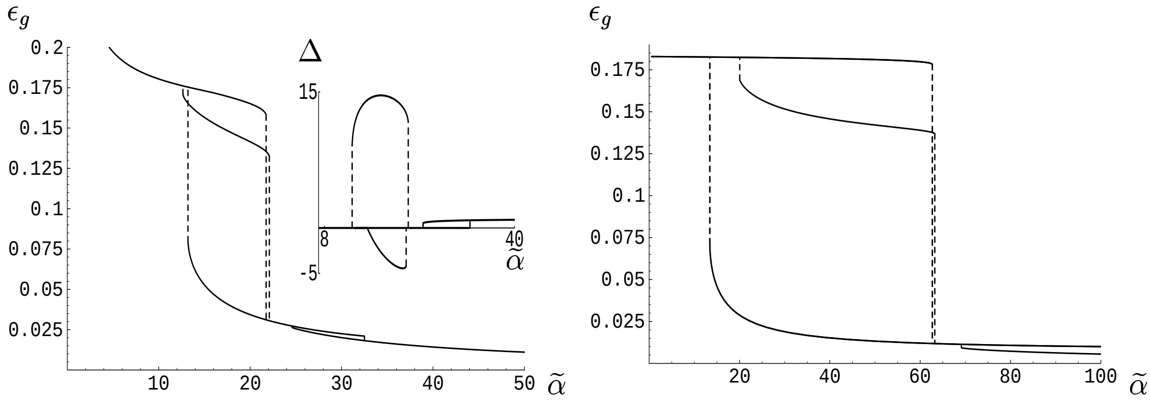

Figure 2: left: $\epsilon_g(\widetilde{\alpha})$ for $K = 3$, $\widetilde{\lambda} = 0.001$ and $\Delta(\widetilde{\alpha})$ (inset). Different starting values were used in numerical minimization to calculate as many local minima of the free energy as possible. right: $K = \infty$, $\widetilde{\lambda} = 0.0001$. All local minima of the free energy have been calculated from the saddle point equations.

and unspecialized phases can be observed in both configurations. Additionally, there is a second order phase transition between the unspecialized large Q phase and an anti-specialized phase ($\Delta < 0$) with large Q at $\widetilde{\alpha} \approx 15$. The $K = 2$ system shows a second order transition in the large-Q regime, while an unspecialized configuration with small Q cannot be observed. This difference in behavior results from the higher degree of symmetry in the $K = 2$ system, where the free energy is invariant under exchange of R and S. Consequently there is no physical difference between specialized and anti-specialized configurations in the $K = 2$ system.

To study the behaviour of very large networks ($K \to \infty$) scaling assumptions of order parameters have to be made. Supposing C to be $\mathcal{O}(1)$, the output of the student will be $\mathcal{O}(\sqrt{K})$ and thus on a different scale as the teacher output. So we assume the hidden unit overlaps to be $\mathcal{O}(1/K)$, writing $C = \hat{C}/(K-1)$ and further introduce $S = \hat{S}/K$, while Δ and Q remain $\mathcal{O}(1)$. Inserting this and performing $\lim_{K\to\infty} \beta f/K$ we find that the condition $\partial f/\partial S = 0$ can be fulfilled only if $Q + \hat{C} - (\Delta + \hat{S})^2$ is assumed to be $\mathcal{O}(1/K)$. So we substitute $\widehat{C} = \widetilde{C}/K + (\Delta + \widehat{S})^2 - Q$ before performing the limit $K \to \infty$. The corresponding generalization error is shown in Figure 2 as a function of $\widetilde{\alpha}$. For small $\widetilde{\alpha}$, the network is in an unspecialized phase with large Q. At $\widetilde{\alpha} \approx 13$ a locally stable, unspecialized configuration with small Q appears, which is globally stable between $\widetilde{\alpha} \approx 22$ and $\widetilde{\alpha} \approx 88$, where the specialized small Q configuration becomes globally stable. However, the unspecialized configuration remains locally stable. Additionally, at $\widetilde{\alpha} \approx 20$ the specialized large Q phase appears, the free energy of which is smaller than that of the unspecialized large Q phase for $\widetilde{\alpha} > 22.5$. Anti-specialized configurations do not exist in the limit $K \to \infty$. We expect them to be a characteristic feature of systems with small $K \geq 3$.

In summary, we have shown by means of statistical physics that learning an unknown rule without a priori knowledge in the form of normalized student vectors leads to a much more complicated behaviour than learning with normalized students. The number of phases in which the system can exist increases. Further, student lengths tend to infinity unless the network weights are regularized by means of a proper weight decay.

Further investigations will extend research to finite temperatures by applying the replica formalism and study the relevance of our results for practical training processes.

Acknowledgement

We thank W Kinzel and A Freking for stimulating discussions and a critical reading of the manuscript.

References

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

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

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

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

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

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

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

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

[11] M Biehl, E Schlösser and M Ahr *Europhys. Lett.* **44** (2) pp. 261-267 (1998)

[12] M Ahr, M Biehl and R Urbanczik *to be published*

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

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

[15] D Saad and S A Solla *Phys. Rev. E* **52**, 4225, 1995

[16] S Bös, *Phys. Rev. E* **58**, 833, 1998 and S Bös and M Opper *J. Phys. A* **31**, 4835, 1998

[17] P Sollich *J. Phys. A* **27**, 7771, 1994

[18] H Schwarze and J Hertz, in: *Advances in Neural Information Processing Systems V*, eds. S J Hanson et. al., Morgan Kaufmann, San Mateo, 1993

[19] A P Dunmur and D J Wallace *J. Phys. A* **26**, 5767, 1993

[20] A Krogh, *J. Phys. A* **25**, 1119, 1992 and A Krogh and J A Hertz J. Phys. A **25**, 1135, 1992

[21] D Saad and M Rattray *Phys. Rev. E* **57**, 2170, 1998

[22] D Barber and P Sollich, in: *On-line Learning in Neural Networks*, ed. D Saad, Cambridge University Press, Cambridge, 1998