Statistical physics of learning: Phase transitions in multilayered neural networks

Michael Biehl, Martin Ahr, Enno Schlösser

Institut für Theoretische Physik, Julius-Maximilians-Universität Würzburg,
Am Hubland, D-97074 Würzburg

Summary: The statistical physics of disordered systems provides tools for the investigation of learning processes in adaptive information processing. The methods and objectives of this approach are exemplified in terms of a specific model scenario: the supervised learning of a rule with a multilayered neural network. The model exhibits a discontinuous dependence of the student performance on the number of example data. This phenomenon can be interpreted as a symmetry breaking phase transition, which results from the competition of (formal) energy and entropy.

1 Introduction

For more than a decade, statistical physics methods have been used to investigate adaptive, information processing systems [1, 2]. This approach has introduced two essential ingredients to the theory of learning: the consideration of very large systems, i.e. the thermodynamic limit, and the concept of performing averages over the randomness contained in the training data. Statistical physics aims at describing the typical behavior of the systems. It thus complements results from other disciplines such as Computational Learning Theory. For more complete reviews of this active field see, for instance, [3, 4, 5].

A superb example for the success of the statistical physics approach to learning is the discussion of phenomena which are the counterparts to phase transitions in statistical mechanics. A specific model is discussed in the following which displays such discontinuous behavior. Here it serves mainly as an example to illustrate the basic concepts of the theory and its outcome. A discussion of the role such phase transitions can play in various other learning systems is given, for example, in [6].

In the following section the model scenario is introduced, an outline of the statistical mechanics analysis is given and its results are presented. The last section summarizes and mentions some of the many other problems that have been treated within the statistical physics of learning.

2 An example: phase transitions in soft–committee machines

2.1 The student teacher scenario

Feedforward neural networks are relevant for applications as they can implement non–trivial regression and classification schemes. These systems have attracted considerable interest due to their ability to learn from example data. Hence they provide an alternative to the rule–based implementation of expert systems, for instance.

For networks with differentiable output successful, practical training algorithms are available [1, 2, 7]. Here, the process of learning in two–layered architectures with differentiable transfer functions is investigated. A specific framework will be considered in the following: the learning of a rule in a fully connected two–layered neural network with output

$$\sigma(\boldsymbol{\xi}) \;=\; \frac{1}{\sqrt{K}} \sum_{j=1}^{K} g(x_j) \quad \text{where} \quad x_j = \frac{1}{\sqrt{N}}\, \mathbf{J}^{(j)} \cdot \boldsymbol{\xi} \tag{2.1}$$

upon presentation of an N–dimensional input vector $\boldsymbol{\xi}$. Given the so–called hidden unit activation $g(x)$, the adaptive weights $\mathbf{J}^{(i)} \in I\!R^N$ parametrize the input–output relation. The term *soft–committee machine* has been coined for this type of network [8, 9], as it can be interpreted as a continuous version of the thoroughly studied (*hard*) committee of binary hidden units (see [10, 11, 12, 13, 14, 15] and references therein). Here, the weights of the linear hidden–to–output relation are fixed to the particular value $1/\sqrt{K}$.

Frequently $g(x)$ is taken to be a sigmoidal function of its argument, e.g. the hyperbolic tangent. The similiar but more convenient function $g(x) = \mathrm{erf}(x/\sqrt{2})$ simplifies the mathematical treatment to a great extent but should not alter the basic features of the model otherwise [8, 9].

In the following, the unknown rule $\tau(\boldsymbol{\xi})$ is taken to 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 orthogonal 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. (2.1), depends explicitely on the length of the weight vectors, this latter constraint corresponds to significant *a priori* knowledge of the rule's structure. The assumption of perfectly matching student complexity constitutes, of course, an important restriction. However, essential aspects of the results derived in the following carry over to the more complex scenarios with unlearnable rules (e.g. [9]) or *over–sophisticated* students [16].

The training process is based on a given set of reliable example data $I\!D = \{\boldsymbol{\xi}^\mu, \tau(\boldsymbol{\xi}^\mu)\}$ $(\mu = 1,2,\ldots,P)$, which contains all the information available about the rule. Training can be, for instance, guided by the minimization of the cost function or training error

$$\varepsilon_t \;=\; \frac{1}{P} \sum_{\mu=1}^{P} \frac{1}{2} \left[\sigma(\boldsymbol{\xi}^\mu) - \tau(\boldsymbol{\xi}^\mu)\right]^2 . \tag{2.2}$$

Thus, learning is formulated as an optimization problem. Here, the corresponding objective function is based on a quadratic error measure, which compares student and teacher output for any given input. 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, 2, 7].

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

$$\varepsilon_g \;=\; \frac{1}{2}\left\langle [\sigma(\boldsymbol{\xi}) - \tau(\boldsymbol{\xi})]^2 \right\rangle . \tag{2.3}$$

$\langle \ldots \rangle$ denotes an empirical average over a test set of (new) inputs $\boldsymbol{\xi} \notin I\!D$ 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. This corresponds to an extreme situation in which the input data itself is completely uncorrelated with the rule. In practical situations we would expect the vectors $\boldsymbol{\xi}$ to carry features which already hint at the rule output without knowing τ explicitly.

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. Hence, the following is independent of further details of the input distribution and the joint density of the $\{x_j,y_j\}$ is fully characterized by their covariances

$$\langle x_i x_j \rangle = \frac{1}{N}\mathbf{J}^{(i)}\cdot\mathbf{J}^{(j)} \;=\; Q_{ij},\;\; \langle x_i y_j \rangle = \frac{1}{N}\mathbf{J}^{(i)}\cdot\mathbf{B}^{(j)} \;=\; R_{ij}, \langle y_i y_j \rangle = \frac{1}{N}\mathbf{B}^{(i)}\cdot\mathbf{B}^{(j)} \;=\; \delta_{ij} \tag{2.4}$$

where all diagonal $Q_{ii} = 1$. The average in Eq. (2.3) 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* $\{R_{ij},Q_{ij}\}$:

$$\varepsilon_g\,(\{Q_{ij},R_{ij}\}) \;=\; \frac{1}{6} + \frac{1}{\pi}\sum_{i,j=1}^{K}\left[\arcsin\left(\frac{Q_{ij}}{2}\right) - 2\arcsin\left(\frac{R_{ij}}{2}\right)\right] \tag{2.5}$$

and was derived in [9] for arbitrary network sizes and more general scenarios.

2.2 Statistical physics of the learning process

In a by now standard approach to learning one considers formally a Gibbs ensemble of networks which is characterized by the partition function

$$Z \;=\; \int d\mu(\{\mathbf{J}^{(i)}\})\, \exp\left[-\beta\, P\, \varepsilon_t\right] . \tag{2.6}$$

The cost function $P\,\varepsilon_t$, as defined in Eq. (2.2), is interpreted as the extensive energy of the system 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. The Gibbs density naturally corresponds to the assumption that the state of the student is characterized by its energy only and that no further constraints apply.

Typical properties of the thermal equilibrium state can be calculated from the associated free energy on average over the randomness contained in the training data $\mathbb{D} = \{\boldsymbol{\xi}^\mu, \tau^\mu\}$ (denoted as $\langle \ldots \rangle_{\mathbb{D}}$). The evaluation of the quenched free energy $-\langle \ln Z \rangle_{\mathbb{D}} / \beta$ requires, in general, the application of the replica trick [1, 4, 5]. It exploits the identity

$$\ln Z \;=\; \lim_{n \to 0} \frac{Z^n - 1}{n}. \tag{2.7}$$

For integer n, the term Z^n is easily identified as the partition function of n non–interacting copies of the system. However, performing the average over the data introduces effective interactions among the replicated systems. $\langle Z^n \rangle_{\mathbb{D}}$ can be evaluated by means of a saddle point equation; the calculation becomes rather involved in models of the complexity considered here, see e.g. [10, 11, 12, 13, 14, 15, 17]. The subtleties of the method are encountered when taking the limit $n \to 0$, for which one usually applies approximation schemes, e.g. *replica symmetry* [1, 4, 5].

2.3 A simplifying limit

In order to obtain first, meaningful results for soft–committee machines we resort here to the simplifying limit of high training temperatures. This strategy has proven useful for gaining insights in a variety of models, [4, 5, 16] present details and further applications.

In the limit $\beta \to 0$, one can show that it is correct to replace $\langle \ln Z \rangle_{\mathbb{D}}$ with $\ln \langle Z \rangle_{\mathbb{D}}$. Further, the average over the training data factorizes with respect to the example inputs and one obtains

$$-\langle \ln Z \rangle_{\mathbb{D}} / N \;=\; \beta f(\{Q_{ij}, R_{ij}\}) \;=\; (\beta P/N)\;\; \varepsilon_g(\{Q_{ij}, R_{ij}\}) - s\,(\{Q_{ij}, R_{ij}\}) \tag{2.8}$$

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. (2.8) contains the entropy term

$$s(\{Q_{ij}, R_{ij}\}) \;= \frac{1}{N} \ln \int \prod_j d^N \mathbf{J}^{(j)} \prod_{i,j} \left(\delta(\mathbf{J}^{(i)} \cdot \mathbf{B}^{(j)} - N R_{ij}) \delta(\mathbf{J}^{(i)} \cdot \mathbf{J}^{(j)} - N Q_{ij}) \right) \tag{2.9}$$

which is evaluated by means of a saddle point integration. One obtains

$$s(\{Q_{ij}, R_{ij}\}) \;=\; \frac{1}{2} \ln[\det \mathcal{C}] + \text{const}, \tag{2.10}$$

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 can be omitted.

In order to proceed with the analysis, we make the simplifying assumption that, in equilibrium, the student configuration is symmetric with respect to the hidden units:

$$R_{ij} = \begin{cases} R & \text{for } i=j \\ S & \text{for } i \neq j \end{cases} \qquad \text{and} \qquad Q_{ij} = \begin{cases} 1 & \text{for } i=j \\ C & \text{for } i \neq j \end{cases} . \tag{2.11}$$

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 the entropy and generalization error read

$$\begin{aligned} s &= \frac{1}{2}\ln\left[1+(K-1)C-((R-S)+KS)^2\right] \\ &\quad +\frac{K-1}{2}\ln\left[1-C-(R-S)^2\right], \end{aligned} \tag{2.12}$$

$$\varepsilon_g = \frac{1}{3}+\frac{K-1}{\pi}\left[\arcsin\left(\frac{C}{2}\right)-2\arcsin\left(\frac{S}{2}\right)\right]-\frac{2}{\pi}\arcsin\left(\frac{R}{2}\right) \tag{2.13}$$

Defining $\alpha = \beta P/(NK)$, the rescaled number of examples per student weight, we have to minimize the free energy

$$f(R,S,C) = \alpha K \varepsilon_g - s \tag{2.14}$$

w.r.t. the remaining order parameters R,S,C. Given a value α, one then obtains the macroscopic properties of a student network, and hence the typical *learning curve* $\varepsilon_g(\alpha)$.

For the model with $K = 2$ hidden units in student and teacher we find that the equilibrium state satisfies $R = S$ for all $\alpha < \alpha_c^{(2)} \approx 23.7$, whereas above this critical value the only solution obeys $\Delta = |R - S| \neq 0$. The system undergoes a second order phase transition and the quantity Δ, which measures the student specialization, increases close to the critical point like $\Delta \propto (\alpha - \alpha_c^{(2)})^{1/2}$. 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 shows the learning curve for $K = 5$ as an example.

The qualitative difference 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, Eq. (2.14), which is invariant under exchange of R and S only for $K = 2$.

In the specialized phase, the generalization error decreases asymptotically like $\varepsilon_g \propto 1/\alpha$ for $\alpha \to \infty$. The metastable state can be shown to disappear at $\alpha_d^{(K)} = 4\pi K$ when K is large [18]. The latter result is an extremely good approximation for K as small as 4 or 5 already.

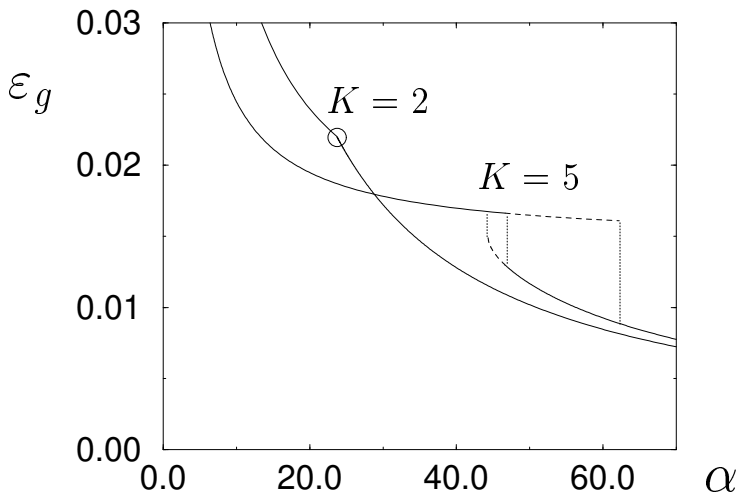

Figure 1 Learning curves for small K. The system with $K = 2$ hidden units undergoes a second order transtion at $\alpha_c^{(2)} \approx 23.7$ leading to the kink in $\varepsilon_g(\alpha)$ (marked by the circle). For $K \geq 3$ the transition is first order. Here we display the case $K = 5$, solid lines correspond to the globally stable solution, dashed lines denote local minima, ($\alpha_s^{(5)} \approx 44.3$, $\alpha_c^{(5)} \approx 46.6$ and $\alpha_d^{(5)} \approx 62.8$).

2.4 Summary

In the above section we have applied equilibrium statistical physics to analyse learning in two–layered soft–committee machines. The results are obtained in the simplifying limit of high training temperature, which allows to calculate the quenched free energy analytically.

Specifically, we have studied networks with K hidden units learning a perfectly matching rule. The competition of the formal energy and entropy in the model causes a discontinuous dependence of the student performance on the size of the training set $\mathbb{D}$. For $K = 2$ we find a second order phase transition from unspecialized to specialized student configuration at a critical number of examples. In scenarios with $K \geq 3$ the transition from poor to good generalization is first order.

A more detailed presentation of the above is given in [18]. By use of the replica method, the analysis has been extended to low temperatures [17]. There, it was also demonstrated that the typical unspecialized and specialized student configurations are indeed observed in practical training schemes such as stochastic gradient descent methods. Furthermore, the case of noisy training data and the approximation of discrete classification schemes with continuous networks have been studied in [19].

It should be noted, that the results discussed here (for $K \geq 3$) parallel to a large extent earlier findings of [10, 11, 12, 13, 14, 15] and others for similar network architectures.

3 Concluding remarks

The above discussed example illustrates the basic ideas used in the statistical physics approach to learning. The consideration of very large systems with many degrees of freedom allows to formulate learning as an optimization problem in a high–dimensional space. Typical properties of such systems can be evaluated by considering a (formal) thermal equilibrium with respect to the objective function or training energy. Here, a temperature like parameter controls the average training error. As the energy is defined with respect to one particular set of examples $\mathbb{D}$ only, one performs an additional average over the randomness or *quenched disorder* contained in the data. The structure of the problem is reminiscent of the questions that arise in spin glass physics (see [1, 5] for references) and the same techniques can be applied.

This strategy has been used to analyse a variety of learning scenarios and model situations. We wish to conclude by referring to only a few of the problems that could be treated or are under consideration currently. For a more complete overview and further references we can only refer to the literature, e.g. [3, 4, 5].

In the context of supervised learning, many different network architectures and types of rules have been considered. These range from single layered architectures like the simple perceptron [1, 4, 5] to the more sophisticated multilayered networks considered here or the so–called Support Vector Machines [20]. The successful treatment within statistical physics has included classification and regression as well as mismatched situations, e.g. unlearnable rules and over–sophisticated students [4, 5]. A question of particular interest is that of training data which is corrupted by some stochastic process (noise), see for instance [19]. The effect of correlations among the training data or with the unknown rule can also be studied, see [21] and references therein.

Another interesting set of problems arises in the context of unsupervised learning [1, 2]. Here, information is to be inferred from input data only. It amounts to the extraction of relevant features in the data, e.g. for the purpose of a low dimensional representation of the inputs or the identification of typical prototype vectors. When dealing with high–dimensional data, statistical physics methods can be applied in a fashion very similar to the analysis described here. Examples for the treatment of unsupervised learning can be found, for instance, in [22, 23].

The investigation of the dynamics of learning processes has followed slightly different lines. If one assumes that training data is presented in a temporal sequence of uncorrelated examples, it is possible to describe the systems in terms of ordinary differential equations for the order parameters. The very successful analysis of *On–line Learning* has been applied to almost all the scenarios mentioned above, including multilayered networks [9, 24]. A recent overview and further references can be found in [25] and [26]. One of the advances that emerged from these studies concerns algorithm design. Initially within the framework of on–line learning, the application of variational methods has allowed to construct potentially optimal training procedures, see [27, 28, 29] for examples.

The analysis of learning problems continues to be an active field within statistical physics. It bears the promise to contribute significantly to the advancement of the theory of adaptive information processing.

Bibliography

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

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

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

[10] H. Schwarze and J. Hertz, Europhys. Lett. **21**, 785 (1993).

[11] H. Schwarze, J. Phys. A **26**, 5781 (1993).

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

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

[14] B. Schottky, J. Phys A **28**, 4515 (1995)

[15] B. Schottky and U. Krey, J. Phys A **30**, 8541 (1997).

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

[17] M. Ahr, M. Biehl, and R. Urbanczik, Eur. Phys. J. B **10**, 583 (1999).

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

[19] M. Ahr, M. Biehl, and R. Urbanczik, J. Phys. A **32**, L531 (1999).

[20] R. Dietrich, M. Opper, and H. Sompolinsky, Phys. Rev. Lett. **82**, 2975 (1999).

[21] C. Marangi, M. Biehl, and S.A. Solla, Europhys. Lett. **30**, 117 (1995).

[22] M. Biehl and A. Mietzner, Europhys. Lett. **24**, 421 (1993).

[23] P. Reimann, C. Van den Broeck, and G.J. Bex, J. Phys. A **29**, 3521 (1996).

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

[25] D. Saad (ed.), *On–line learning in neural networks.* (Cambridge University Press, 1998).

[26] M. Biehl and N. Caticha, in *The Handbook of Brain Theory and Neural Networks*, ed. M.A. Arbib, MIT Press (in press).

[27] O. Kinouchi and N. Caticha, J. Phys. A **28**, 6243 (1992).

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

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