Universität Würzburg

Institut für Theoretische Physik

Am Hubland, D–97074 Würzburg, Germany

Specialization processes in on–line unsupervised learning

Michael Biehl, Ansgar Freking, Georg Reents, Enno Schlösser

Ref.: WUE-ITP-97-003
e-mail : biehl@physik.uni-wuerzburg.de
ftp : ftp.physik.uni-wuerzburg.de

Specialization processes in on–line unsupervised learning

M. Biehl, A. Freking, G. Reents, and E. Schlösser
Institut für Theoretische Physik
Universität Würzburg, Am Hubland
D-97074 Würzburg, Germany

Abstract

From the recent analysis of supervised learning by on–line gradient descent in multilayered neural networks it is known that the necessary process of student specialization can be delayed significantly. We demonstrate that this phenomenon also occurs in various models of unsupervised learning. A solvable model of competitive learning is presented, which identifies prototype vectors suitable for the representation of high–dimensional data. The specific case of two overlapping clusters of data and a matching number of prototype vectors exhibits non–trivial behavior like almost stationary plateau configurations. As a second example scenario we investigate the application of Sanger's algorithm for principal component analysis in the presence of two relevant directions in input space. Here, the fast learning of the first principal component may lead to an almost complete loss of initial knowledge about the second one.

1 Introduction

The theory of learning has benefitted from the application of methods developed in statistical physics, which complement the tools from computational learning theory. In the physics framework, averaging over the disorder introduced by the randomness in the training data allows to investigate the typical behavior of very large systems. Most of this work has focused on models for the supervised learning of a rule from examples by adaptation of parameters in a feed–forward neural network (Hertz et al., 1991; Bishop, 1995). Reviews of the statistical physics approach to these processes are for example (Opper and Kinzel, 1996) and (Watkin et al., 1993). Recently, also unsupervised learning (Hertz et al., 1991; Bishop, 1995) has been studied in a statistical mechanics context (see for instance Biehl and Mietzner, 1994; Barkai and Sompolinsky, 1994; Lootens and van den Broeck, 1995; and references therein).

It is the nature of many learning problems that they require the specialization of disjoint subsets of adaptive parameters with respect to specific tasks. In a multilayered neural network, for instance, these subsets are the weight vectors connecting the input with the different hidden units. From the recent analysis of learning by on–line gradient descent in such systems, it is known that the necessary student specialization may dominate by far the computation time consumed for successful training (see e.g. Biehl and Schwarze, 1995; Saad and Solla, 1995; Biehl et al., 1996). We show in the following that delayed specialization is also of great importance in the context of unsupervised learning. First we discuss the identification of prototype vectors representing a large number of correlated data by means of competitive learning (Hertz et al., 1991; Bishop, 1995). Then, we briefly report first results concerning a standard on–line algorithm for principal component analysis.

2 The dynamics of competitive learning

One possible objective of unsupervised learning is the identification of prototype vectors for a given (often very large) set of data $\left\{\boldsymbol{\xi}^\mu \in \mathrm{R}^N\right\}_{\mu=1,\ldots,P}$. The aim is to represent this set as faithful as possible by use of only a few typical vectors $\left\{\mathbf{J}_k \in \mathrm{R}^N\right\}_{k=1,\ldots,K}$ $(K \ll P)$ which capture the relevant features of the data. This is closely related (however, not identical) with clustering problems, where the goal is to group the vectors $\boldsymbol{\xi}^\mu$ into several sets of similar inputs (Bishop, 1995; Hertz, 1991).
Here, we will consider an on–line learning prescription (see for instance Amari, 1993; Saad and Solla, 1995; Biehl et al., 1996, and references therein). In this framework, a sequence of example vectors $\boldsymbol{\xi}^\mu$ is presented by the environment and only the latest from this sequence is used for the actual update of the prototypes.
We restrict the analysis to the case of only two prototype vectors ($K = 2$), which are updated according to a training scheme of the form

$$\mathbf{J}_k(\mu) = \mathbf{J}_k(\mu-1) + \frac{\eta}{N}\left(\boldsymbol{\xi}^\mu - \mathbf{J}_k(\mu-1)\right)\; p_k\left[\mathbf{J}_1(\mu-1), \mathbf{J}_2(\mu-1), \boldsymbol{\xi}^\mu\right]. \qquad (1)$$

The assignment functions p_k are taken to be

$$p_1\left[\mathbf{J}_1, \mathbf{J}_2, \boldsymbol{\xi}\right] = \Theta\left[(\mathbf{J}_2 - \boldsymbol{\xi})^2 - (\mathbf{J}_1 - \boldsymbol{\xi})^2\right] \quad \text{and} \quad p_2 = 1 - p_1.$$

This algorithm (1) updates only the prototype vector with the smallest distance to the currently presented input vector $\boldsymbol{\xi}^\mu$. In this "competition for updates" among the prototypes the "winner takes it all", hence the term "competitive learning" is used for this class of algorithms (Hertz et al., 1991).
The step size of the procedure is determined by the learning rate η in (1) which is scaled with the dimension N of the data. The change is always in the direction $(\boldsymbol{\xi}^\mu - \mathbf{J}_k)$, therefore the updated prototype will be the winner for similar examples later in the sequence with even higher probability. Such a strategy should lead to a

specialization of prototypes, each representing regions of $\mathbf{R}^N$ where examples have been observed frequently. The algorithm is therefore expected to be a suitable tool for the identification of "clusters" of similar data. A corresponding model distribution is specified in the following.

Note that (1) can be interpreted as a stochastic gradient descent for a proper energy function, which incorporates the assignment of the data to the prototypes and at the same time measures the quality of the representation; see (Biehl et al., 1997) for a more detailed discussion. Note furthermore that no normalization constraint is imposed on the prototype vectors in (1). This is in contrast to models of directional clustering or feature detection (Biehl and Mietzner, 1994; Watkin and Nadal, 1994; Lootens and van den Broeck, 1995), where only characteristic directions in the distribution of the data are of interest. Here, the magnitude of the $\mathbf{J}_k$ will typically be of order $\mathbf{J}_k^2 = \mathcal{O}(1)$, whereas $\boldsymbol{\xi}^2 = \mathcal{O}(N)$ holds for the example data.

We study competitive learning from data as provided by a stochastic source according to a density of the specific form

$$P(\boldsymbol{\xi}) = \frac{1}{2} \sum_{m=1}^{2} P(\boldsymbol{\xi}|m) \quad \text{where} \quad P(\boldsymbol{\xi}|m) = \frac{1}{(2\pi)^{N/2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{\xi} - b\mathbf{B}_m\right)^2\right]. \qquad (2)$$

This corresponds to a mixture density of 2 overlapping Gaussians centered at $b\mathbf{B}_1$ and $b\mathbf{B}_2$ respectively. We assume the $\mathbf{B}_m$ to be orthogonal ($\mathbf{B}_1 \cdot \mathbf{B}_2 = 0$) and normalized ($\mathbf{B}_m^2 = 1$). Thus, the non–negative quantity b defines the distance of the cluster centers from the origin. The dummy variable m indicates from which of the clusters $\boldsymbol{\xi}$ was drawn, both centers are taken to contribute with the same probability $1/2$. Note, however, that the labels $m = 1, 2$ are not provided with the data in the course of learning.

The anisotropy of this data distribution is only very weak: the mean of all $\boldsymbol{\xi}$ belonging to cluster m has an offset $b\mathbf{B}_m$ from the origin. This is a vector of length $b = \mathcal{O}(1)$, whereas the average length of the inputs $\boldsymbol{\xi}$ is $\mathcal{O}(\sqrt{N})$. Distributions of this or a similar form have been considered previously in the context of supervised and unsupervised learning (e.g. Biehl and Mietzner, 1994; Watkin and Nadal, 1994; Barkai and Sompolinsky, 1994; Meir, 1995; Marangi et al., 1995).

We proceed by investigating the model in the thermodynamic limit $N \to \infty$. In this limit the random quantities $x_k = \mathbf{J}_k \cdot \boldsymbol{\xi}$ and $y_m = \mathbf{B}_m \cdot \boldsymbol{\xi}$ are distributed according to a mixture of Gaussians as well. Note that this holds true for more general $P(\boldsymbol{\xi}|n)$ with the same first and second moments as in (2) according to the central limit theorem. The joint density of the $\{x_k, y_m\}$ is completely determined by their conditional averages and covariances ($\langle \cdots \rangle_n$ denotes the average over $P(\boldsymbol{\xi}|n)$):

$$\langle x_j x_k \rangle_n - \langle x_j \rangle_n \langle x_k \rangle_n = Q_{jk} = \mathbf{J}_j \cdot \mathbf{J}_k, \quad \langle x_k y_m \rangle_n - \langle x_k \rangle_n \langle y_m \rangle_n = R_{km} = \mathbf{J}_k \cdot \mathbf{B}_m$$

$$\langle y_l y_m \rangle_n - \langle y_l \rangle_n \langle y_m \rangle_n = \delta_{lm} = \mathbf{B}_l \cdot \mathbf{B}_m, \quad \langle x_k \rangle_n = b\, R_{kn}, \quad \langle y_m \rangle_n = b\, \delta_{mn} \qquad (3)$$

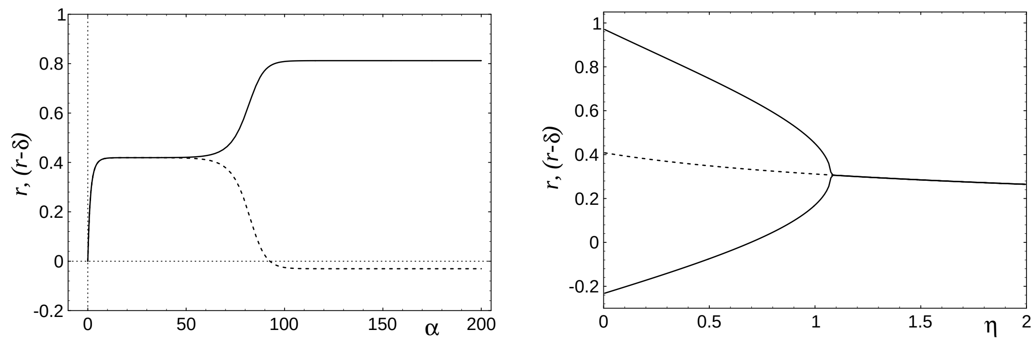
The self–averaging order parameters Q_{jk} and R_{km} suffice in the thermodynamic limit $N \to \infty$ for a macroscopical description of the prototypes and the dynamics of the learning process can be analysed exactly in terms of these quantities.
It is straightforward to derive from (1) recursion relations for the evolution of the order parameters. These equations are then averaged over the latest example, assuming that at each time step a new, uncorrelated vector $\boldsymbol{\xi}$ is drawn according to (2). The randomness enters only through the projections $\{x_k, y_m\}$. In terms of the continuous time $\alpha = \mu/N$ the dynamics is then described by a set of coupled first order differential equations (see e.g. Biehl and Schwarze, 1995; Saad and Solla, 1995) for a more detailed description of the formalism. As the order parameters are self–averaging quantities, their evolution describes the learning process exactly for any random realization of the data sequence in the thermodynamic limit.
The corresponding system of differential equations reads

$$\begin{aligned} \frac{dR_{km}}{d\alpha} &= \eta \left\langle \left(y_m - R_{km}\right) p_k \right\rangle \\ \frac{dQ_{lm}}{d\alpha} &= \eta \left\langle \left(x_l - Q_{lm}\right) p_m + \left(x_m - Q_{lm}\right) p_l \right\rangle + \eta^2 \left\langle p_l \, p_m \right\rangle . \end{aligned} \tag{4}$$

For the sake of simplicity the arguments of the labelling functions have been omitted. All averages $\langle \cdots \rangle$ are over the full distribution of the $\{x_k, y_m\}$ and can be performed analytically (see a forthcoming publication for details). Given a set of initial conditions, this system of seven coupled differential equations can be integrated numerically and thus yields the evolution of the order parameters in the course of learning.
For a first discussion of the essential features of the learning curves we restrict our considerations to a partially symmetric subspace of the system configurations. We assume that the initial conditions obey the relations

$$Q_{11} = Q_{22} = Q, \quad Q_{12} = Q_{21} = Q - \varepsilon, \quad R_{11} = R_{22} = r\sqrt{Q}, \quad R_{12} = R_{21} = (r - \delta)\sqrt{Q}.$$

This symmetry is preserved in the dynamics. The first condition indicates that the lengths of the prototype vectors are equal. It can be satisfied easily in realistic situations, whereas the specific symmetry of the overlaps $\mathbf{J}_k \cdot \mathbf{B}_m$ constitutes a restriction of the dynamics. We have observed, however, that the system generically approaches this subspace quickly from rather general initial conditions. A similar effect occurs in supervised learning by gradient descent and was exploited in e.g. (Saad and Solla, 1995; Biehl et al., 1996). We limit the analysis to the reduced system for the 4 quantities r, δ, Q, and ε, which describe the dynamics exactly within the symmetric subspace.
Figure 1 a) shows the evolution of the order parameters r and δ for $\eta = 1$ and specific initial conditions corresponding to normalized orthogonal vectors $\mathbf{J}_1(0)$ and $\mathbf{J}_2(0)$ with small initial values of r and δ.
In a first phase of the training process the symmetry parameter δ remains very small. This corresponds to a situation where the prototypes have not yet specialized and



a) b)

Figure 1: a) Learning curves for the model explained in the text with $\eta = 1$, $b = 1.2$ and for initial conditions $r(0) = \delta(0) = 10^{-6}$, $Q(0) = \varepsilon(0) = 1$. The evolution of the order parameters r (solid line) and $r - \delta$ (dashed) is shown.
b) Fixed point values of r and δ vs. the learning rate η for a separation $b = 0.8$. For $\eta > \eta_S \approx 1.08$ the system remains in a plateau with an η–dependent value of $r = r - \delta$. Below the critical learning rate this configuration is unstable (dashed line) and a fixed point with non–zero δ becomes attractive.

their overlaps with $\mathbf{B}_1$ and $\mathbf{B}_2$ are almost equal. The fast initial increase of $r \approx r - \delta$ indicates that both prototypes gain information about the relevant directions. Then, the system is trapped in a plateau–state in which the order parameters are approximately constant. Eventually, with increasing α, δ becomes significantly different from zero which indicates the specialization of the prototypes: each $\mathbf{J}_k$ represents mainly data from one of the clusters.
Note that, in contrast to the case of supervised gradient descent in multilayered nets (Saad and Solla, 1995; Biehl et al., 1996) the dominant plateau in our model is not characterized by almost identical $\mathbf{J}_k$. Due to the pronounced competitive nature of the "winner takes all" prescription, a "mutual repulsion" is imposed on the prototype vectors: at each time step only one of the vectors is updated even if $\mathbf{J}_1 \approx \mathbf{J}_2$. This effective repulsion separates the prototypes very fast; however, their projections into the $\mathbf{B}_1$–$\mathbf{B}_2$–plane are almost equal in the plateau state.
The occurence of the plateau is due to the existence of a fixed point of the differential equations (4) with $\delta = 0$. The properties of the fixed point depend critically on the choice of η. Provided the learning rate is smaller than a b–dependent value η_S, the plateau configuration is unstable. A non–zero value of δ is necessary in order to break the symmetry and enables the prototypes to specialize with increasing α. The initial value $r(0)$ is to a large extent irrelevant for the *occurence* of the plateau state. However, the length of the plateau is determined by the initial specialization $\delta(0)$, a more detailed analysis will be published elsewhere. For $\eta > \eta_S$ the unspecialized plateau state is attractive which is again analogous to results from supervised learning (Biehl and Schwarze, 1995). In this case, small deviations from $\delta = 0$ are not sufficient for the system to leave the plateau and a spontaneous specialization of the

prototypes does not occur.
In figure 1 b) the stationary values of r and $r-\delta$ for $\alpha \to \infty$ are plotted versus the learning rate for a particular separation b. The plateau values of r, Q, and ε vary weakly with the actual learning rate η. For $\eta < \eta_S$ the unspecialized state is repulsive and the stable asymptotic configuration is characterized by $\delta \neq 0$. In the limit $\eta \to 0$ the prototypes become for $\alpha \to \infty$ linear combinations of the $\mathbf{B}_k$ with $(\mathbf{J}_1 - \mathbf{J}_2) \propto (\mathbf{B}_1 - \mathbf{B}_2)$; this corresponds to the global minimum of the associated cost function (Biehl et al., 1997).

3 The dynamics of principal component analysis

Principal component analysis is a standard tool for the detection of relevant features of correlated data (Hertz et al., 1991, Bishop, 1995, and references therein). The purpose is to identify a number of relevant directions $\{\mathbf{J}_k\}_{k=1,\ldots,K}$ in input space, so that the projections $x_k = \mathbf{J}_k \cdot \boldsymbol{\xi}$ contain a large amount of information about the original data. One possible application is the K–dimensional encoding of N–dimensional data with $K \ll N$ (Hertz et al., 1991).
Consider a given set of zero mean data $\{\boldsymbol{\xi}^\mu\}_{\mu=1,\ldots,P}$ with the empirical correlation matrix $\mathsf{C} = \sum_\mu \boldsymbol{\xi}^\mu \boldsymbol{\xi}^{\mu\top}/P$. It can be shown that if the $\{\mathbf{J}_k\}_{k=1,\ldots,K}$ are chosen to be the eigenvectors of C corresponding to its K largest eigenvalues [1] the linear reconstruction $\boldsymbol{\xi}^\mu_{rec} = \sum_k x_k^\mu \mathbf{J}_k$ is optimal with respect to the averaged quadratic error $\sum_\mu (\boldsymbol{\xi}^\mu_{rec} - \boldsymbol{\xi}^\mu)^2/P$. This optimality criterion coincides for a purely Gaussian data distribution with the maximization of the information contained in the x_k about the original input. The formal argument and a more detailed discussion can be found in e.g. (Bishop, 1995). Note that the eigenvalues λ_k defined through $\mathsf{C}\,\mathbf{J}_k = \lambda_k\,\mathbf{J}_k$ are given by the variance of the corresponding x_k with respect to the distribution of data.
A number of algorithms have been introduced for the identification of the principal components; for examples see (Hertz et al., 1991). We will focus in the following on a procedure proposed by Sanger (Sanger, 1989; Hertz et al., 1991). The identification of the first principal component only was studied in a similar model previously (Biehl and Mietzner, 1994; Biehl, 1994).
In the on–line training scenario, independent vectors $\boldsymbol{\xi}^\mu$ are assumed to be generated according to an N–dimensional, zero mean Gaussian distribution which is isotropic apart from two distinct, orthogonal directions $\mathbf{B}_1$ and $\mathbf{B}_2$ with $\mathbf{B}_k^2 = 1$ and $\mathbf{B}_1 \cdot \mathbf{B}_2 = 0$. The corresponding correlation matrix is taken to be $\langle \boldsymbol{\xi}\,\boldsymbol{\xi}^\top \rangle = \mathrm{I}_N + (b_1^2 + 2\,b_1)\,\mathbf{B}_1\,\mathbf{B}_1^\top + (b_2^2 + 2\,b_2)\,\mathbf{B}_2\,\mathbf{B}_2^\top$ with the N–dimensional identity matrix I_N.
For arbitrary, normalized vectors $\mathbf{J}_k$, this yields for $N \to \infty$ zero mean Gaussian random variables (index μ omitted) $x_k = \mathbf{J}_k \cdot \boldsymbol{\xi}$ and $y_k = \mathbf{B} \cdot \boldsymbol{\xi}$ with covariances

$$\langle x_k^2 \rangle = 1 + (b_1^2 + 2\,b_1)\,R_{k1} + (b_2^2 + 2\,b_2)\,R_{k2}, \quad \langle y_m^2 \rangle = (1 + b_m)^2,$$

[1] (or alternatively, any orthogonal basis of this subspace)

$$\langle x_k\, y_m \rangle = (1+b_m)^2\, R_{km}, \quad \langle y_1\, y_2 \rangle = 0, \quad \text{and} \quad \langle x_1\, x_2 \rangle = Q_{12}.$$

All order parameters are defined as in the previous section; note, however, that we consider normalized vectors $\mathbf{J}_k$ in the following. We furthermore assume, without loss of generality, that $b_1 > b_2$ in the following. Thus, $\mathbf{B}_1$ corresponds to the largest eigenvalue of the correlation matrix and $\mathbf{B}_2$ to the second principal component of the data distribution.

We study the learning of two feature vectors by use of an on–line version of Sanger's rule:

$$\begin{aligned} \mathbf{J}_1(\mu) &= \mathbf{J}_1(\mu-1) + \frac{\eta}{N}\, x_1^\mu \left[\boldsymbol{\xi}^\mu - x_1^\mu\, \mathbf{J}_1(\mu-1)\right] \\ \mathbf{J}_2(\mu) &= \mathbf{J}_2(\mu-1) + \frac{\eta}{N}\, x_2^\mu \left[\boldsymbol{\xi}^\mu - x_1^\mu\, \mathbf{J}_1(\mu-1) - x_2^\mu\, \mathbf{J}_2(\mu-1)\right] \end{aligned} \tag{5}$$

In the limit $\eta \to 0$ this update rule preserves an initial normalization $\mathbf{J}_k^2 = 1$. For larger learning rates, this constraint can be imposed explicitely and introduces additional weight decay terms of order $\mathcal{O}(\eta^2/N)$, see forthcoming publications. It can be shown that, asymptotically, the prescription (5) yields the normalized eigenvectors of the correlation matrix corresponding to its largest and second largest eigenvalue. In contrast, a similar learning scheme proposed by Oja (see Hertz et al., 1991) is known to provide some basis of the same subspace, but with the asymptotic result depending on the initial configuration.

The analysis of the dynamics is to a large extent analogous to the considerations in the previous section. Here we only present first results as obtained by numerical integration of the corresponding system of differential equations for the quantities R_{km} and Q_{12}.

Figure 2 shows the evolution of the order parameters in continuous time $\alpha = \mu/N$; initial conditions are given in the caption. Generically, we observe a fast increase of both overlaps R_{k1} with the most relevant direction $\mathbf{B}_1$. After this initial phase, the effective orthogonalization implemented in (5) enforces the specialization of the $\mathbf{J}_k$ and only one of the vectors keeps approaching $\mathbf{B}_1$. As an interesting side effect, the overlap of the remaining student vector with the second relevant direction (R_{22} in Fig. 2) also decreases rapidly. Prior knowledge, as expressed in the non–zero initial value of this order parameter, deteriorates almost completely during this phase of the learning process. Only after a considerable delay, $\mathbf{J}_2$ acquires a significant overlap with $\mathbf{B}_2$ and R_{22} approaches its final (η–dependent) value. For larger values of the learning rate similar effects are observed. However, additional plateaus in the cross–overlaps may occur and the asymptotic configuration is characterized by smaller values of R_{11} and R_{22}. Preliminary investigations of learning according to Oja's subspace algorithm (Hertz et al., 1991) reveal the same effects qualitatively. Both schemes will be discussed in greater detail in a forthcoming publication.

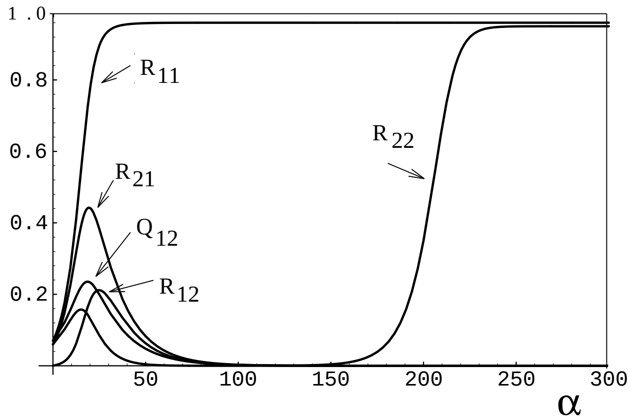


Figure 2: Learning curves for Sanger's algorithm as explained in the text for $b_1 = 0.6, b_2 = 0.4$ and initial conditions $R_{11} = R_{21} = 0.07$, $R_{22} = R_{12} = 0.06$ and $Q_{12} = 0$. The evolution of the order parameters is shown for the learning rate $\eta = 0.1$.

4 Summary and outlook

In summary, we have presented exactly solvable models of unsupervised learning.
For the specific example of a competitive ("winner takes all") prescription we have demonstrated that the dynamics of the learning process can be described by a small number of order parameters which obey a set of first order differential equations in the thermodynamic limit. Here we have focused our analysis on the specialization of prototype vectors and the occurence of plateaus in the learning curves. In this context we have also investigated the critical role of the learning rate.
First results concerning an on–line principal component analysis indicate that also in this scenario successful learning may be considerably delayed by the necessary specialization. As a particularly interesting effect we observe that the fast identification of the first principal component may cause an almost complete loss of prior knowledge about less relevant directions.
The choice of an appropriate time–dependent learning rate in order to yield fast convergence to the optimal solution of the respective problem is of particular interest for both cases. The influence of initial configurations on the occurence and length of plateaus should be studied carefully in analogy to the investigation of supervised gradient descent.
Further investigations will also address more complex model situations, for instance with a mismatched number of prototypes or student vectors ($K \neq M$) or more complicated data distributions. Modifications of the learning algorithms based on different objective functions should be considered which go beyond the standard schemes studied so far. For example, the effect of a probabilistic "soft competitive" assignment of the prototypes on the learning dynamics is an important open question. Such algorithms are closely related to maximum likelihood methods or similar techniques of density estimation (Bishop, 1995). Furthermore, non–linear extensions of principal component analysis, which take into account higher moments of the data distribution, will be investigated in a forthcoming project.

References

Amari, S., 1993, Neurocomp. **5**, 185.
Barkai, N. and H. Sompolinksy, 1994, Phys. Rev. E **50**, 1766.
Biehl, M. and A. Mietzner, 1994, J. Phys. A **27**, 1885.
Biehl, M., 1994, Europhys. Lett. **30**, 391.
Biehl, M. and H. Schwarze, 1995, J. Phys. A **28**, 643.
Biehl, M., P. Riegler, and C. Wöhler, 1996, J. Phys. A **29**, 4769.
Biehl, M., A. Freking, and G. Reents, 1997, Europhys. Lett., in press.
Bishop C. M., 1995, *Neural Networks for Pattern Recognition* (Clarendon Press, Oxford)
Hertz, J.A., A. Krogh, and R.G. Palmer, 1991, *Introduction to the Theory of Neural Computation* (Addison Wesley, Redwood–City, CA)
Lootens, E. and C. van den Broeck, 1995, Europhys. Lett. **30**, 381.
Marangi, C., M. Biehl, S.A. Solla, 1995, Europhys. Lett. **30**, 117.
Meir, R., 1995, Neural Comp. **7**, 144.
Opper, M. and W. Kinzel, 1996, in *Models of Neural Networks III* , eds. E. Domany, J.L. van Hemmen, and K. Schulten (Springer, Berlin)
Saad, D. and S.A. Solla, 1995, Phys. Rev. Lett. **74**, 4337 and Phys. Rev. E **52**, 4225.
Sanger, T.D., 1989, in *Advances in Neural Information Processing Systems 1* , ed. D.S. Touretzky (Morgan Kaufmann, San Mateo), 1009.
Watkin, T.L.H. and J.–P. Nadal, 1994, J. Phys. A **27**, 1889.
Watkin, T.L.H., A. Rau, and M. Biehl, 1993, Rev. Mod. Phys. **65**, 499.