On–line Learning of Prototypes and Principal Components

M. Biehl, A. Freking, M. Hölzer, G. Reents, and E. Schlösser

Institut für Theoretische Physik
Universität Würzburg
Am Hubland
D–97074 Würzburg, Germany

Ref: WUE-ITP-98-007

Abstract

We review our recent investigation of on–line unsupervised learning from high–dimensional structured data. First, on–line competitive learning is studied as a method for the identification of prototype vectors from overlapping clusters of examples. Specifically, we analyse the dynamics of the well–known winner–takes–all or K–means algorithm. As a second standard learning technique, the application of Sanger's rule for principal component analysis is investigated. In both scenarios the necessary process of student specialization may be delayed significantly due to underlying symmetries.

1 Introduction

Methods from statistical physics have been applied to the theory of adaptive systems with great success in recent years. Perhaps the most prominent example is the analysis of feedforward neural networks which can learn from example data. The statistical mechanics approach allows to investigate typical properties of very large systems on average over the randomness contained in the data. It complements results from computational learning theory and other disciplines.

Most of the investigations concern the supervised learning of a rule. For reviews of the field see for instance (Watkin *et al.* 1993; Opper and Kinzel, 1996). A particularly successful line of research was initiated in (Kinzel and Rujan, 1990; Kinouchi and Caticha, 1992) and aims at the analysis of the physics of on–line learning schemes (Amari, 1967 and 1993; Hertz *et al.*, 1991). On–line learning is attractive from a practical point of view because it uses only the latest in the sequence of examples. Obviously storage needs and computational effort are reduced in comparison with batch– or off–line learning (Hertz *et al.*, 1991).

On the other hand, the simplicity of on–line algorithms has allowed to study a variety of learning scenarios and network architectures, including the simple perceptron (e.g. Kinzel and Rujan, 1990; Kinouchi and Caticha, 1992) and multilayered networks with threshold units (e.g. Sompolinsky *et al.*, 1995, Copelli and Caticha, 1995) or sigmoidal activation functions respectively (e.g. Saad and Solla, 1995; Biehl *et al.*, 1996). Despite their simplicity, on–line algorithms compete well with costly off–line prescriptions (e.g. Kinouchi and Caticha, 1992; Opper, 1996; van den Broeck and Reimann, 1996; Kim and Sompolinsky, 1996; Copelli *et al.*, 1997).

Models of unsupervised learning have also been studied in the statistical mechanics framework, see e.g. (Biehl and Mietzner, 1994; Watkin and Nadal, 1994; Barkai and Sompolinsky, 1994; Lootens and van den Broeck, 1995). The investigation of on–line unsupervised learning schemes (Biehl, 1994; van den Broeck and Reimann, 1996) has provided new insights in this context as well.

In the following we review our recent investigation of on–line unsupervised learning from high–dimensional structured data (Biehl *et al.*,1997; Biehl and Schlösser, 1998). In the next section, on–line competitive learning is discussed as a method for the identification of prototype vectors from overlapping clusters of examples. Here, the focus will be on the well–known *winner–takes–all* or *K–means* algorithm (Duda and Hart, 1973; Hertz *et al.*, 1991; Bishop, 1995). Section 3 revisits a second standard unsupervised learning technique: the identification of principal components by means of Sanger's rule (Sanger, 1989; Hertz *et al.*, 1991; Bishop, 1995).

In all these scenarios, the necessary process of student specialization can be delayed significantly. This effect is due to underlying symmetries which result in quasi–stationary plateau configurations in the learning dynamics.

We conclude with a summary of the main results and a discussion of possible extensions in section 4.

2 Competitive Learning

2.1 The Model

One of the possible objectives of unsupervised learning is the identification of prototype vectors from a given set of data $\left\{\boldsymbol{\xi}^\mu \in \mathbb{R}^N\right\}$ $(\mu = 1, \ldots, P)$. The aim is to find a faithful representation of this set by use of only a few typical vectors $\left\{\boldsymbol{J}_k \in \mathbb{R}^N\right\}$ $(k = 1, \ldots, K \ll P)$ which capture the relevant features of the data. This is closely related to (yet not identical with) *clustering problems*, where the goal is to group the vectors $\boldsymbol{\xi}^\mu$ into several sets of similar inputs (Hertz *et al.*, 1991; Bishop, 1995).

Frequently, the identification of prototypes is guided by the Euclidean distances

$$d_k(\boldsymbol{\xi}) \;=\; (\boldsymbol{\xi} - \boldsymbol{J}_k)^2. \tag{2.1}$$

In particular, the family of so–called competitive learning algorithms updates the students $\boldsymbol{J}_k$ according to a prescription of the generic form

$$\boldsymbol{J}_k^\mu = \boldsymbol{J}_k^{\mu-1} + \frac{\eta}{N} \left(\boldsymbol{\xi}^\mu - \boldsymbol{J}_k^{\mu-1} \right) \; p_k\left(\{d_k^\mu\}\right) \tag{2.2}$$

The change of the weight vectors is always along $(\boldsymbol{\xi}^\mu - \boldsymbol{J}_k^{\mu-1})$, i.e. the prototypes are moved in the direction of the presented example. The step size of this change is determined by the learning rate $\eta > 0$ which is scaled with the dimension N of the data.

The factors p_k define the actual algorithm. Here, they are taken to be non–negative functions of the set of distances $d_k^\mu = (\boldsymbol{\xi}^\mu - \boldsymbol{J}_k^{\mu-1})^2$ and obey the normalization constraint

$$\sum_{k=1}^{K} p_k\left(\{d_k^\mu\}\right) \; = \; 1 \tag{2.3}$$

which fixes the total contribution of a single example to the learning process. Typically, a specific example will affect the prototypes which are closest in distance more efficiently than others. This *competition* of the students for updates is controlled by the assignment or labeling functions p_k.

Note that no normalization constraint is imposed on the student vectors. This is in contrast to models of directional clustering, where only characteristic directions are searched in input space (Biehl and Mietzner, 1994; Watkin and Nadal, 1994; Lootens and van den Broeck, 1995). In the following, the magnitude of the student vectors is $\boldsymbol{J}_k^2 = \mathcal{O}(1)$, whereas $\boldsymbol{\xi}^2 = \mathcal{O}(N)$ holds true for the example data.

We restrict the analysis to the case of only two prototype vectors $\boldsymbol{J}_1$ and $\boldsymbol{J}_2$ in an environment which provides a sequence of independent data drawn from a stochastic source. We assume a bimodal input distribution of the specific form

$$P(\boldsymbol{\xi}) = \frac{1}{2} \sum_{m=1}^{2} P(\boldsymbol{\xi}|m) \quad \text{where} \quad P(\boldsymbol{\xi}|m) = \frac{1}{(2\pi)^{N/2}} \, \exp\left[-\frac{1}{2} \left(\boldsymbol{\xi} - b\,\boldsymbol{B}_m \right)^2 \right]. \tag{2.4}$$

which corresponds to a mixture of two overlapping Gaussians centered at $b\,\boldsymbol{B}_m$. We take the characteristic vectors $\boldsymbol{B}_m$ to be orthogonal and normalized ($\boldsymbol{B}_k \cdot \boldsymbol{B}_m = \delta_{km}$). Hence, the quantity $b > 0$ specifies the offset of the respective cluster centers from the origin. The dummy variable m indicates from which of the clusters $\boldsymbol{\xi}$ is drawn, both centers contribute with the same probability 1/2.

The assumed data distribution is only weakly anisotropic: Given the cluster label m, the distance of the conditional mean $b\,\boldsymbol{B}_m$ from the origin is $b = \mathcal{O}(1)$, whereas the average length of input vectors is $\mathcal{O}(\sqrt{N})$. Similar clustered input distributions have been studied in various models of unsupervised learning as well as in supervised scenarios, see e.g. (Biehl and Mietzner, 1994; Watkin and Nadal, 1994; Barkai and Sompolinsky, 1994; Meir, 1995; Marangi *et al.*, 1995). Note that here, of course, the labels $m = 1, 2$ are not provided with the example data.

We proceed by investigating the model in the thermodynamic limit $N \to \infty$. The random quantities $x_m = \boldsymbol{J}_m \cdot \boldsymbol{\xi}$ and $y_m = \boldsymbol{B}_m \cdot \boldsymbol{\xi}$ are distributed according to a mixture of Gaussians as well. By means of the central limit theorem this holds true for more general $P(\boldsymbol{\xi}|m)$ in the limit $N \to \infty$, provided the first and second moments are the same as in (2.4). The joint density of the overlaps is uniquely determined by their conditional averages and covariances:

$$\begin{aligned}
\langle x_j x_k \rangle_n - \langle x_j \rangle_n \langle x_k \rangle_n &= Q_{jk} = \boldsymbol{J}_j \cdot \boldsymbol{J}_k \\
\langle x_k y_m \rangle_n - \langle x_k \rangle_n \langle y_m \rangle_n &= R_{km} = \boldsymbol{J}_k \cdot \boldsymbol{B}_m \\
\langle y_l y_m \rangle_n - \langle y_l \rangle_n \langle y_m \rangle_n &= \delta_{lm} = \boldsymbol{B}_l \cdot \boldsymbol{B}_m \\
\langle x_k \rangle_n = b\, R_{kn} \quad \text{and} \quad \langle y_m \rangle_n &= b\, \delta_{mn}
\end{aligned} \tag{2.5}$$

Here and in the following $\langle \cdots \rangle_n$ denotes the average over the contribution $P(\boldsymbol{\xi}|n)$ to the mixture density (2.4). Averages over the full joint density of $\{x_1, x_2, y_1, y_2\}$ are to be calculated as a sum of conditional means, one obtains for example

$$\langle x_1 \rangle = \frac{1}{2} \left(\langle x_1 \rangle_1 + \langle x_1 \rangle_2 \right) = \frac{1}{2} b \, (R_{11} + R_{12}) \, .$$

2.2 The Dynamics of Learning

In the thermodynamic limit $N \to \infty$ the set of self–averaging order parameters Q_{jk} and R_{km} is sufficient for a macroscopical description of the system. Further, the dynamics of the learning process can be analysed exactly in terms of these quantities. To this end we derive from (2.2) recursion relations for the evolution of the Q_{jk} and R_{km}, which are then averaged over the latest random input vector. This is possible because the randomness of the data enters only through the projections $\{x_k, y_k\}$. Thus, all averages are over the four–dimensional Gaussian density which is specified by its moments (2.5).

In terms of the continuous time $\alpha = \mu/N$ the dynamics is now described by a set of coupled first order differential equations, see e.g. (Biehl and Schwarze, 1995) and (Saad and Solla, 1995) for a more detailed description of the formalism in the context of supervised learning. A discussion of self–averaging in on–line learning can be found in (Reents and Urbanczik, 1998).

Here, the obtained system of differential equations reads

$$\begin{aligned}
\frac{dR_{km}}{d\alpha} &= \eta \left\langle (y_m - R_{km})\, p_k \right\rangle \\
\frac{dQ_{lm}}{d\alpha} &= \eta \left\langle (x_l - Q_{lm})\, p_m + (x_m - Q_{lm})\, p_l \right\rangle + \eta^2 \left\langle\, p_l\, p_m \,\right\rangle
\end{aligned} \tag{2.6}$$

where the arguments of the assignment functions have been omitted for simplicity. For a discussion of the essential properties of our model we restrict the analysis to a partially symmetric subspace where

$$Q_{11} = Q_{22} = Q, \;\; Q_{12} = Q_{21} = C, \;\; R_{11} = R_{22} = R, \;\; R_{12} = R_{21} = S \tag{2.7}$$

In all cases investigated here, the dynamics (2.6) preserves this symmetry. We observe furthermore that, even if the initial conditions violate (2.7), the system approaches the restricted subspace (or an equivalent one with relabeled students) after a short transient. This is analogous to the findings of e.g. (Saad and Solla 1995) and (Biehl *et al.*, 1996) with respect to supervised learning schemes.

In systems with two students only, a substantial simplification is achieved by introducing the following linear combinations of order parameters:

$$R_{\pm} = R \pm S \;\; \text{and} \;\; Q_{\pm} = Q \pm C. \tag{2.8}$$

In terms of these quantities and under the symmetry assumption (2.7) the equations of motion (2.6) are of the following form:

$$\frac{dR_+}{d\alpha} = \frac{\eta}{2}\,(b - R_+) \qquad \frac{dQ_+}{d\alpha} = \frac{\eta}{2}\,(2bR_+ - 2Q_+ + \eta) \tag{2.9}$$

$$\begin{aligned} \frac{dR_-}{d\alpha} &= \frac{\eta}{2}\left(\langle (y_1 - y_2)\,(p_1 - p_2)\rangle - \langle p_1 - p_2\rangle \; R_-\right) \\ \frac{dQ_-}{d\alpha} &= \eta\left(\langle (x_1 - x_2)\,(p_1 - p_2)\rangle - \langle p_1 - p_2\rangle \; Q_-\right) + \frac{\eta^2}{2}\left\langle (p_1 - p_2)^2\right\rangle . \end{aligned} \tag{2.10}$$

Here we have used the properties

$$p_1 + p_2 = 1 \;\; \text{and} \;\; \langle p_{1,2}\rangle = 1/2 \tag{2.11}$$

of the input distribution and assignment functions.

Note that the dynamics of R_+ and Q_+ decouples from the remaining equations. Remarkably, Eqs. (2.9) are independent of the specific learning algorithm provided it satisfies the conditions (2.11). Hence, for two prototypes in the presence of the bimodal input distribution (2.4), the temporal evolution of R_+ and Q_+ is the same for all competitive learning schemes and can be studied beforehand. We obtain the analytic solution

$$\begin{aligned} R_+(\alpha) &= b + Ae^{-\eta\alpha/2} \\ Q_+(\alpha) &= \frac{\eta}{2} + b^2 + 2bAe^{-\eta\alpha/2} + Be^{-\eta\alpha} \end{aligned} \tag{2.12}$$

where the constants $A = R_+(0) – b$ and $B = -\eta/2 + b^2 + Q_+(0) – 2bR_+(0)$ depend on the initial conditions. The asymptotic ($\alpha \to \infty$) configuration $R_+ = b$ and

$Q_+ = b^2 + \eta/2$ represents the only fixed point of the subsystem (2.9) and is approached exponentially fast with increasing α from all initial settings and for all values of η.

The quantity R_+ measures the overlaps of the prototype vectors with the sum $(\boldsymbol{B}_1 + \boldsymbol{B}_2)$. It therefore marks the unspecialized identification of the direction in which the center of mass of the input distribution is found.

On the contrary, the overlap $R_- = \boldsymbol{J}_1 \cdot (\boldsymbol{B}_1 - \boldsymbol{B}_2) = \boldsymbol{J}_2 \cdot (\boldsymbol{B}_2 - \boldsymbol{B}_1)$ quantifies the *specialization* of the students which is, in a way, the genuine aim of competitive learning.

2.3 The Specialization Process

We will investigate the dynamics of the specialization process for a particularly simple and efficient choice for the labeling functions:

$$p_k = \prod_{j \neq k}^{K} \Theta(d_j^\mu - d_k^\mu) = \begin{cases} 1 & \text{if } d_k^\mu < d_j^\mu \quad \text{for all } j \neq k \\ 0 & \text{else.} \end{cases} \tag{2.13}$$

The corresponding training scheme updates at each time step only the student with the minimal distance to the current input. All other $K-1$ prototypes remain unchanged, hence the term *winner–takes–all*–algorithm has been coined for this prescription (Hertz *et al.*, 1991). It is identical with an on–line realization of the well–known *K–means*–algorithm (Duda and Hart, 1973; Bishop, 1995).

The resulting prescription can be interpreted as the stochastic gradient descent minimization of an instantaneous energy

$$\varepsilon\,(\boldsymbol{\xi}^\mu) = \frac{1}{2} \sum_{k=1}^{K} d_k^\mu \, p_k - \frac{1}{2} (\boldsymbol{\xi}^\mu)^2 \tag{2.14}$$

which measures the *representation error* of input vector $\boldsymbol{\xi}^\mu$ in terms of Euclidean distances, see (Hertz *et al.*, 1991; Biehl *et al.*, 1997) for details.

As the change of weights is always in the direction $(\boldsymbol{\xi}^\mu - \boldsymbol{J}^{\mu-1})$, the updated prototype will be the *winner* for similar examples later in the sequence with even higher probability. Therefore, the strategy should yield specialized prototypes, each of which represents a region in input space where many examples have been observed in the course of learning.

For two competing students Eq. (2.13) reduces to

$$p_1 = \Theta(d_2^\mu - d_1^\mu) \qquad p_2 = \Theta(d_1^\mu - d_2^\mu) = 1 - p_1. \tag{2.15}$$

By use of the property $p_k p_m = p_k \delta_{km}$ of the Heaviside–function all averages in the Eqs. (2.10) can be performed analytically and one obtains

$$\frac{dR_-}{d\alpha} = \frac{\eta}{2} \left(-b - R_- + 2b\Phi \left[\frac{bR_-}{\sqrt{2Q_-}} \right] + \frac{2R_-}{\sqrt{\pi Q_-}} \exp \left[-\frac{b^2 R_-^2}{4Q_-} \right] \right)$$

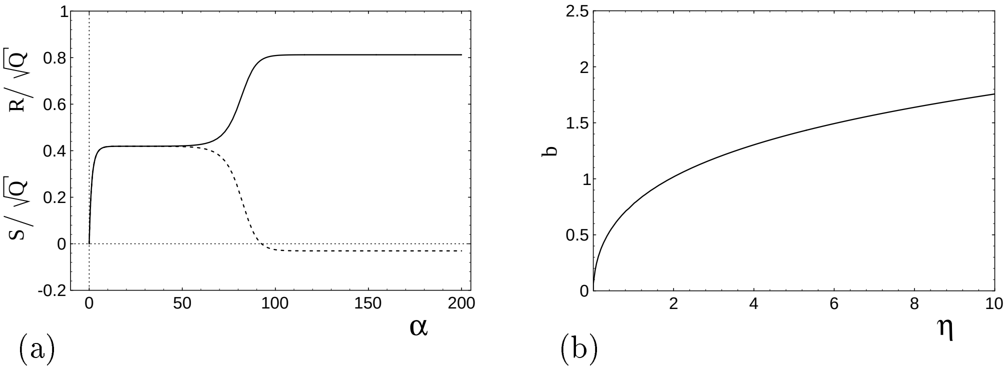


Figure 1 (winner–takes–all algorithm)
(a) Normalized order parameters $R/\sqrt{Q}$ (solid) and $S/\sqrt{Q}$ (dashed) vs. α for $\eta = 1$, $b = 1.2$. Here, initial conditions were $R(0) = 10^{-6}$, $S(0) = C(0) = 0$, and $Q(0) = 1$.
(b) The critical learning rate η_c (2.18) separates values of η and b for which the plateau state is stable from those which allow the specialization of prototypes.

$$\frac{dQ_-}{d\alpha} = \frac{\eta}{2}\left(\eta - 2Q_- - 2bR_- + 4bR_-\Phi\left[\frac{bR_-}{\sqrt{2Q_-}}\right] + \frac{4\sqrt{Q_-}}{\sqrt{\pi}}\exp\left[-\frac{b^2R_-^2}{4Q_-}\right]\right) \tag{2.16}$$

where $\quad \Phi[x] = \int\limits_{-\infty}^{x} \frac{dy}{\sqrt{2\pi}} e^{-y^2/2} = \frac{1}{2}\left(1 + \mathrm{erf}\left[x/\sqrt{2}\right]\right).$

The two subsystems for R_+, Q_+ and R_-, Q_- decouple completely in this model: (2.16) is independent of (2.9) and vice versa. However, in order to evaluate the original order parameters $\{R, S, Q, C\}$ as functions of α one has to combine the analytic result (2.12) with the numerical integration of the above set of equations.

Figure 1 (a) displays, as an example, the resulting learning curves (normalized overlaps $R/\sqrt{Q}$ and $S/\sqrt{Q}$ vs. α) for $\eta = 1$, $b = 1.2$ and initial conditions $R(0) = 10^{-6}$, $S(0) = C(0) = 0$, $Q(0) = 1$. These correspond to normalized, orthogonal student vectors with very small initial overlaps R and specialization $R_- = R - S$. Note that for randomly drawn N–dimensional vectors one would expect random values $R(0)$ and $S(0)$ of the order $\mathcal{O}(1/\sqrt{N})$ in realistic learning scenarios with no *a priori* knowledge.

As shown in the Fig. 1 (a), the overlaps R and S increase rapidly for small α but without achieving considerable specialization $R_- = R - S$. Only after an extended *plateau*–like phase of the dynamics with almost constant order parameters, the specialization increases drastically and the system approaches an apparently stable configuration with $R_- = \mathcal{O}(1)$ for large α.

In order to gain a theoretical understanding of the observed behavior we study the fixed point structure of (2.16). For all fixed points, the values of R_+ and Q_+ are given by the asymptotic form ($\alpha \to \infty$) form of Eqs. (2.12).

We observe that configurations with

$$R_- = 0, \qquad Q_- = Q_-^{(\pm)} = \frac{4 + \eta\,\pi \pm 2\,\sqrt{4 + 2\eta\pi}}{2\,\pi} \tag{2.17}$$

are stationary under the dynamics (2.16). A linearization of the system shows that the fixed point $(0, Q_-^{(-)})$ is always repulsive, whereas $(0, Q_-^{(+)})$ becomes attractive for

$$\eta > \eta_c = \frac{2}{\pi}\left(b^4 + 2b^2\right) \tag{2.18}$$

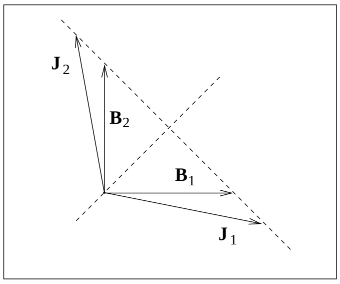
which defines a critical learning rate of the process. In Fig. 1 (b) it is shown as a line in the (η, b)–plane which separates the region in which no specialization occurs from the one with a non–zero asymptotic value of R_-.

For small enough learning rates $\eta \leq \eta_c$ all fixed points with $R_- = 0$ are repulsive and the students will eventually specialize upon presentation of an increasing number of examples. However, generic initial conditions with $R \approx S \approx 0$ will cause the system to approach a state in the vicinity of (2.17). A linearization around the fixed point $(0, Q_-^{(+)})$ shows that the specialization increases exponentially with α: $R_- \propto R_-(0)\, e^{\lambda\alpha}$ where λ is the relevant eigenvalue of the linearized system. The characteristic time needed to achieve significant specialization $R_- = \mathcal{O}(1)$ is therefore proportional to $-\ln[R_-(0)]/\lambda$.

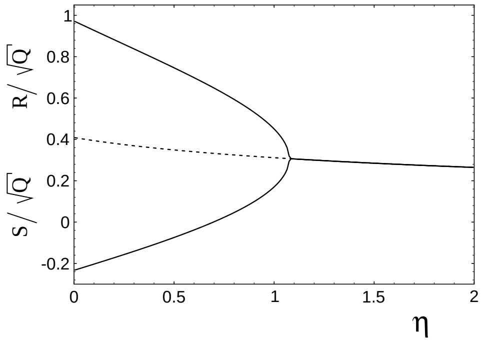
This behavior is strongly reminiscent of the plateau states which have been found to delay supervised learning by gradient descent in multilayered networks (Biehl and Schwarze, 1995; Saad and Solla, 1995; Biehl *et al.*, 1996), see other contributions to this volume. Note, however, that the dominant plateau in the winner–takes–all scenario is *not* characterized by almost identical student vectors. An effective mutual repulsion is imposed on the prototype vectors due to the pronounced competitive nature of the learning algorithm. At each time step only one of the students can be updated even if $\boldsymbol{J}_1 \approx \boldsymbol{J}_2$. The prototypes separate very fast, however their projections in the $(\boldsymbol{B}_1, \boldsymbol{B}_2)$–plane are almost equal in the plateau state.

Eventually the system approaches a stable configuration where R_- and Q_- satisfy the conditions

$$\begin{aligned} Q_- &= \frac{\eta^2}{8}\frac{r_-^2}{\left(\Phi\left[r_-\right] - \frac{1}{2}\right)^2 \left(b^2 - 2r_-^2\right)^2} \\ \frac{e^{-r_-^2/2}}{\sqrt{\pi}} &= \frac{\sqrt{Q_-}}{2} - \frac{b^2\left(\Phi\left[r_-\right] - \frac{1}{2}\right)}{\sqrt{2}\, r_-} \end{aligned} \tag{2.19}$$



(a)



(b)

Figure 2 (winner–takes–all algorithm)
(a) The asymptotic configuration for $\eta \to 0$ (sketch), see the discussion in the text.
(b) Asymptotic values of the normalized $R/\sqrt{Q}$ and $S/\sqrt{Q}$ for a cluster offset $b = 0.8$. For learning rates $\eta > \eta_c \approx 1.08$ the prototypes do not specialize, below η_c the plateau state is unstable (dashed line) and a fixed point with non–zero R_- becomes attractive.

with the abbreviation $r_- = bR_-/\sqrt{2Q_-}$. In general, these fixed point values have to be obtained numerically. The limit of small learning rates $\eta \to 0$ yields

$$R_- = \sqrt{Q_-} = -b + 2\,b\,\Phi\left[\frac{b}{\sqrt{2}}\right] + \frac{2}{\sqrt{\pi}}\exp\left[-\frac{b^2}{4}\right], \quad \text{i.e.} \quad r_- = \frac{b}{\sqrt{2}}. \tag{2.20}$$

This asymptotic configuration of the system is characterized by student vectors which are linear combinations of the $\boldsymbol{B}_k$ and obey

$$(\boldsymbol{J}_1 - \boldsymbol{J}_2) \propto (\boldsymbol{B}_1 - \boldsymbol{B}_2).$$

Note, however, that this does not imply a perfect alignment or even identity of the $\boldsymbol{J}_k$ which one of the cluster centers. The above mentioned energy function (2.14) favors indeed well separated prototypes which are, in a sense, more typical for the data than the actual centers of the overlapping clusters.

Figure 2 (a) shows the asymptotic configuration for $\eta \to 0$ schematically. For non–zero $\eta < \eta_c$ the picture remains qualitatively the same, however, contributions from the space orthogonal to $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ persist even in the limit $\alpha \to \infty$. For practical purposes, an appropriate time dependent, decaying learning rate should be used with $\eta \propto 1/\alpha$ for large α (e.g. Bishop, 1995). In Fig. 2 (b) the asymptotic values of R and S are displayed as a function of the learning rate for a specific cluster offset b.

3 Principal Component Analysis

3.1 The Model

Another important problem in data analysis is the faithful representation of high-dimensional data by low–dimensional feature vectors which contain as much information about the original inputs as possible.

One standard method for this task is principal component analysis (PCA). It determines, for a given set of observed data, the eigenvectors of the empirical covariance matrix which correspond to its largest eigenvalues. Projections on these characteristic vectors serve as a useful linear representation of the data, see e.g. (Hertz *et al.*, 1991; Bishop, 1995; Deco and Obradovic, 1996) for the theoretical background.

The purpose of competitive learning is to provide (few) typical prototype vectors of the same dimensionality as the (many) original input vectors. On the contrary, PCA aims at the low–dimensional representation of each of the examples by detecting the most relevant features of the data.

Principal component analysis takes into account only first and second moments of the observed data (i.e. their covariance matrix). Hence, we can consider a particularly simple model distribution in the following. Input vectors $\boldsymbol{\xi}$ are taken to consist of random components independently drawn from zero mean Gaussian distributions. We assume that M relevant directions $\{\boldsymbol{B}_i\}_{i=1,\ldots,M}$ in $I\!\!R^N$ (with $M \ll N$) exist which determine the correlation matrix $C = \left\langle \boldsymbol{\xi}\boldsymbol{\xi}^\top \right\rangle$:

$$C \;=\; I_N + \sum_{i=1}^{M} (b_i^2 + 2b_i) \boldsymbol{B}_i \boldsymbol{B}_i^\top \tag{3.1}$$

with the N–dimensional identity matrix I_N. The vectors $\{\boldsymbol{B}_i\}$ are taken to be orthogonal and normalized: $\boldsymbol{B}_i \cdot \boldsymbol{B}_j \;=\delta_{ij}$. This weakly anisotropic distribution can be interpreted as the result of deforming a single, isotropic Gaussian cluster with data points $\tilde{\boldsymbol{\xi}} \in I\!\!R^N$:

$$\boldsymbol{\xi} = \tilde{\boldsymbol{\xi}} + \sum_{i=1}^{M} b_i (\boldsymbol{B}_i \cdot \tilde{\boldsymbol{\xi}}) \boldsymbol{B}_i . \tag{3.2}$$

Distributions of this type have been previously considered in e.g. (Lootens and van den Broeck, 1995).

The directions B_i are, by construction, the eigenvectors of C corresponding to eigenvalues $(1 + b_i)^2$. Assuming $b_1 \geq b_2 \geq \ldots \geq b_M > 0$ without loss of generality, the set of vectors $\boldsymbol{B}_i$ coincides therefore with the ordered principal components of the data distribution.

As before, we assume that the environment generates a sequence of independent example vectors $\boldsymbol{\xi}^\mu$ according to the above input distribution. A matching number of student vectors $\boldsymbol{J}_l \in I\!R^N \quad (l = 1, 2, \ldots, M)$ is updated following Sanger's rule (Sanger, 1989; Hertz *et al.*, 1991) when a new example is presented:

$$\boldsymbol{J}_l^\mu = \boldsymbol{J}_l^{\mu-1} + \frac{\eta_l}{N} x_l^\mu \left(\boldsymbol{\xi}^\mu - \sum_{k=1}^{l} x_k^\mu \boldsymbol{J}_k^{\mu-1} \right), \tag{3.3}$$

with the student projections $x_l^\mu = \boldsymbol{J}_l^{\mu-1} \cdot \boldsymbol{\xi}^\mu$.

The learning rates η_l control the step size of the updates for the students. In contrast to the previous section we include the possibility of using different rates for different $\boldsymbol{J}_k$.

For small learning rates $\eta_l \to 0$ one can show that Sanger's rule yields normalized vectors (Hertz *et al.*, 1991). Throughout this paper, however, we assume that an explicit normalization at each time step μ guarantees $\boldsymbol{J}_l^2 = 1$ for all μ. This introduces additional terms of order η_l^2/N in the full form of the algorithm, see (Biehl, 1994) for the case $M = 1$.

The prescription (3.3) leads to an ordering of the student vectors which (in general) corresponds to the identification of one $\boldsymbol{B}_i$ by each student when a large number of examples is presented (Hertz *et al.*, 1991). An alternative scheme was suggested by Oja (Oja, 1982, Oja and Karhunen, 1985, Hertz *et al.*, 1991, Oja, 1996) and has been proven to provide an arbitrary basis of the subspace spanned by the $\{\boldsymbol{B}_i\}$, the actual result depends on the initial choice of the $\boldsymbol{J}_l(0)$.

Again, the analysis is based on the fact that the quantities x_k and $y_j = \boldsymbol{B}_j \cdot \boldsymbol{\xi}$ (indices μ omitted) are Gaussian variables. In the thermodynamic limit this holds true for more general input vectors consisting of non–Gaussian random components ξ_i with the same second order statistics. Here, the relevant moments are

$$\langle x_k y_l \rangle = (1 + b_l)^2 R_{kl}, \qquad \langle y_k y_l \rangle = \delta_{kl} \, (1 + b_k)^2, \tag{3.4}$$

$$\langle x_k x_l \rangle = Q_{kl} + \sum_{i=1}^{M} (b_i^2 + 2b_i) R_{li} R_{ki}, \tag{3.5}$$

and $\langle x_k \rangle = \langle y_k \rangle = 0$ for all k. The order parameters $R_{kl} = \boldsymbol{J}_k \cdot \boldsymbol{B}_l$ and $Q_{kl} = \boldsymbol{J}_k \cdot \boldsymbol{J}_l$ are defined as in the previous section but with all $Q_{kk} = 1$ due to the above mentioned normalization.

3.2 The Dynamics of Learning

The analysis of the temporal evolution of order parameters proceeds by deriving a system of $(3M^2 - M)/2$ coupled first order differential equations which reads

$$\frac{dR_{lj}}{d\alpha} = \eta_l \langle x_l y_j \rangle - (\eta_l + \eta_l^2/2) \left\langle x_l^2 \right\rangle R_{lj} - \eta_l \sum_{k=1}^{l-1} \langle x_l x_k \rangle \, (R_{kj} - Q_{lk} R_{lj})$$

$$\frac{dQ_{lj}}{d\alpha} = (\eta_l + \eta_j) \langle x_l x_j \rangle - ((\eta_l + \eta_l^2/2) \left\langle x_l^2 \right\rangle + (\eta_j + \eta_j^2/2) \left\langle x_j^2 \right\rangle) Q_{lj}$$
$$-\eta_l \sum_{k=1}^{l-1} \langle x_l x_k \rangle (Q_{kj} - Q_{kl} Q_{lj}) - \eta_j \sum_{k=1}^{j-1} \langle x_j x_k \rangle (Q_{lk} - Q_{kj} Q_{lj}) \tag{3.6}$$

where $l, j = 1, 2, \ldots, M$ ($l \neq j$ in the second equation). Note that all averages are given in (3.4). Therefore, a closed set of differential equations is obtained for arbitrary $M (\ll N)$. In addition to the numerical integration of (3.6), an analytic treatment of its fixed point properties allows for an investigation of the system in the limit $\alpha \to \infty$. Further, plateau–like states in the transient dynamics can be studied.

As a measure of success we consider the deviation of the linear reconstruction

$$\boldsymbol{\xi}_{est} = \sum_{i=1}^{M} x_i \, \boldsymbol{J}_i \tag{3.7}$$

from the original data $\boldsymbol{\xi}$. The expectation value of the associated quadratic error is minimized for $\{\boldsymbol{J}_i = \boldsymbol{B}_i\}$ $(i = 1, \ldots, M)$ or whenever the two sets of vectors span the same subspace. The simple linear combination of vectors $\boldsymbol{J}_k$ (3.7) coincides with the optimal (with respect to the quadratic error) linear reconstruction only asymptotically, i.e. for $\{\boldsymbol{J}_k \to \boldsymbol{B}_k\}$ (Bishop, 1995; Deco and Obradovic, 1996). Nevertheless, (3.7) can serve as a rough measure of the achieved quality of the representation.

The estimation error on average over the assumed input distribution is given by

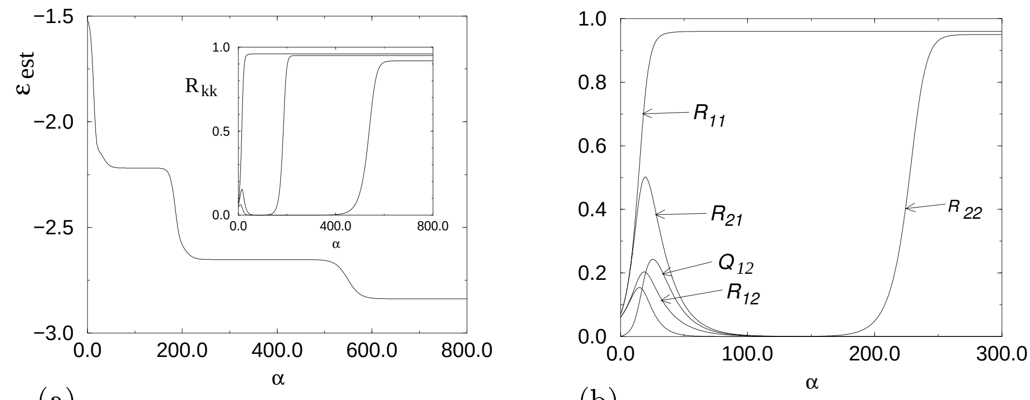
$$\varepsilon_{est} = \frac{1}{2} \left\langle (\boldsymbol{\xi}_{est} - \boldsymbol{\xi})^2 \right\rangle - \frac{1}{2} \left\langle \boldsymbol{\xi}^2 \right\rangle = -\frac{1}{2} \sum_{k=1}^{M} \left\langle x_k^2 \right\rangle + \sum_{k=1}^{M} \sum_{l=1}^{k-1} \langle x_k x_l \rangle \, Q_{kl} \tag{3.8}$$

and can be expressed in terms of the order parameters by means of (3.4). Note that an irrelevant constant $\left\langle \boldsymbol{\xi}^2 \right\rangle / 2$ has been subtracted in the definition of ε_{est}.

In Figure 3 (a) a generic example of the temporal evolvement of diagonal overlaps R_{ll} and of the cost function is shown. For small enough learning rates η_l the following attractive fixed point configuration is approached asymptotically (with $\alpha \to \infty$):

$$R_{ll} = \pm \sqrt{\frac{b_l^2 + 2b_l - \eta_l/2}{(b_l^2 + 2b_l)(1 + \eta_l/2)}}, \qquad R_{lj} = Q_{lj} = 0 \;\; \text{for } l \neq j. \tag{3.9}$$

This configuration reflects the identification of one specific principal component by each student. However, the achievable absolute values $|R_{ll}|$ remain smaller than 1 for non–zero learning rates (all $\eta_l = 0.1$ in Fig. 3 (a)). Very



(a) (b)

Figure 3: (Sanger's rule)
(a) A typical learning curve for $M = 3$: The representation error ε_{est} decreases in a cascade–like manner, the corresponding evolution of the diagonal order parameters R_{kk} is displayed in the inset. Here, $\eta_l = 0.1$ and all $R_{jk}(0) = 0.06 +$ random deviations of the order $\mathcal{O}$ (10^{-10}).
(b) The evolution of all overlaps in the case $M = 2$ with initial conditions $R_{lk}(0) = \mathcal{O}(10^{-2})$, $Q_{12}(0) = \mathcal{O}(10^{-4})$ and $X(0) = \mathcal{O}(10^{-10})$ (cf. Eq. (3.11)), learning rates $\eta_l = 0.1$ for both students.

small values of η_l yield good learning success, but many examples are needed. With larger η_l learning is fast, but the asymptotic error remains large. Better results could be achieved by applying an appropriate annealing schedule for α–dependent learning rates $\eta_l(\alpha)$.

For step sizes

$$\eta_l \;>\; \eta_l^c \;=\; 2b_l(b_l+2) \tag{3.10}$$

configurations with the corresponding overlaps $R_{ll} = 0$ become stable which is analogous to the result of (Biehl, 1994) for $M = 1$. Throughout the following we will assume that all learning rates are smaller than their respective critical value. Hence, (3.9) is the only attractive configuration of the system.

3.3 Symmetries and Specialization

Additional repulsive fixed points of the system exist due to underlying symmetries of the learning problem. After relabeling the students all these repulsive states are characterized by (3.9) with some or all of the diagonal $R_{ll} = 0$.

The influence of these repulsive states on the learning dynamics is exemplified in Fig. 3 (a): Before approaching its asymptotic configuration (3.9), the system is trapped close to a number of such fixed points. The intermediate configurations are almost stationary and resemble the plateau states observed in supervised learning and in the context of the competitive algorithm (see previous section).

To begin with, we discuss the relevance and structure of the plateau–like configurations in terms of the model with two student vectors ($M = 2$) and equal learning rates $\eta_1 = \eta_2 = \eta$.

First we note that for any initial configuration with $R_{11}(0) = 0$, this overlap will remain zero in the course of learning. This property of Sanger's rule is already apparent in a system with only one student (Biehl, 1994) since the update of $\boldsymbol{J}_1$ is independent of all other vectors $\boldsymbol{J}_k (k > 1)$. A non–zero $R_{11}(0)$ will enable the system to achieve the asymptotic value given in (3.9). In realistic situations one would expect $R_{11} = \mathcal{O}(1/\sqrt{N})$ indicating that a characteristic time of the order $\ln N$ is needed to produce a significant overlap R_{11}. Here we focus on the dynamics of the subsequent students and assume a fairly large $R_{11}(0)$ throughout the following.

Because of the hierarchical structure of the algorithm (3.3) the above property does not simply carry over to other overlaps. For instance, $dR_{22}/d\alpha \neq 0$ holds true even with $R_{22} = 0$, in general. Instead, we can show that the specific symmetry measure $X = (R_{11}R_{22} - R_{12}R_{21})$ satisfies

$$\frac{dX}{d\alpha} = 0 \quad \text{for} \quad X = 0. \tag{3.11}$$

A value of $X = 0$ corresponds to unspecialized vectors $\boldsymbol{J}_i$ with linear dependent projections in the space spanned by $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$. For a set of students with $X(0) = 0$ it is impossible to identify both principal components because the conservation of the symmetry (3.11), together with the effective orthogonalization for $\alpha \to \infty$, enforces

$$R_{22} \to 0 \quad \text{and} \quad R_{21} \to 0$$

when R_{11} increases in the course of learning. This is possible even with $R_{22}(0), R_{21}(0) > 0$. Such a case is illustrated in Fig. 3 (b) which demonstrates the effective loss of initial overlaps.

A linearization of the equations of motion around the specific configuration $R_{jk} = 0$ for $j, k = 1, 2$ and $Q_{12} = 0$ shows that a small, non–zero $|X|$ increases exponentially with α. Eqs. (3.6) decouple in the vicinity of the configuration and one obtains

$$X(\alpha) = X(0) \cdot e^{\lambda\,\alpha} \qquad \text{with} \quad \lambda = (b_1^2 + 2b_1 + b_2^2 + 2b_2)\eta - \eta^2. \tag{3.12}$$

It is interesting to note that the learning rate at which λ becomes zero is given by $(\eta_1^c + \eta_2^c)/2$ with η_l^c from Eq. (3.10). Thus, the stability of $X(0) = 0$ is directly linked to the critical rates associated with the diagonal overlaps themselves.

The characteristic time which is needed to achieve a nonzero $X = \mathcal{O}(1)$, i.e. to leave the repulsive fixed point will be proportional to $-\ln|X(0)|/\lambda$.

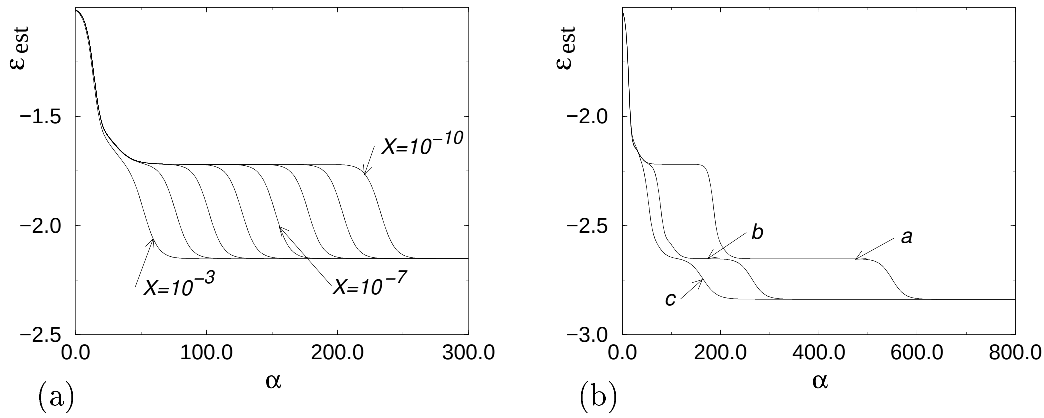


Figure 4: (Sanger's algorithm)
(a) The plateau length of the learning process ($M = 2$) depends logarithmically on X (from left to right: $-\log_{10} X = 3, 4, 5, \ldots 10$). Apart from the deviations X, initial conditions are the same as in Fig. 3 (b)
(b) Average estimation error ($M = 3$) for equal learning rates as in Fig. 3 (a) (line a), learning rates $\eta_1 = 1, \eta_2 = 1.009\eta_1$ and $\eta_3 = 0.99\eta_1$ (line b) and $\eta_1 = 1, \eta_2 = 1.09\eta_1$ and $\eta_3 = 0.9\eta_1$ (line c).

This logarithmic dependence of the *plateau length* is displayed in Fig. 4 (a) for a number of initial values $X(0)$.

We would like to point out that the symmetry $X = 0$ (eq. (3.11)) is preserved in the course of learning only if the learning rates are identical ($\eta_1 = \eta_2$). The relation

$$\frac{dX}{d\alpha} = aX + b(\eta_1 - \eta_2) \tag{3.13}$$

indicates that $|X|$ will increase with α even for $X(0) = 0$ as soon as $\eta_1 \neq \eta_2$. The coefficients a and b are in general non-zero and depend on the R_{kj} and the Q_{kl}. Small differences $|\eta_1 - \eta_2|$ enable the system to escape from the symmetric configuration after a time of order $-\ln|\eta_1 - \eta_2|$. Note that the effect is *not* related to the natural ordering of the principal components. The improvement in comparison with $\eta_1 = \eta_2$ does not depend critically on which of the learning rates is taken to be the larger one.

In a system of $M > 2$ students and $\eta_l = \eta$ ($l = 1, 2, \ldots, M$) statements analogous to the Eqs. (3.11,3.12) can be made with respect to

$$X^{(M)} = \begin{vmatrix} R_{11} & R_{12} & \ldots & R_{1M} \\ R_{21} & \ldots & \ldots & R_{2M} \\ \ldots & \ldots & \ldots & \ldots \\ R_{M1} & \ldots & \ldots & R_{MM} \end{vmatrix} . \tag{3.14}$$

The full determinant of overlaps R_{lm} as well as the determinants of all upper–left submatrices $X^{(k)}$ (related to the first k students and eigenvectors) are found to satisfy

$$\frac{dX^{(k)}}{d\alpha} = 0 \quad \text{if} \quad X^{(k)} = 0 \qquad (k = 1, 2, \ldots, M). \tag{3.15}$$

This reflects again the hierarchical structure of Sanger's algorithm. Note that for $k = 1, 2$ the above discussed properties of the overlap R_{11} and the symmetry $X = X^{(2)}$ (3.11) is included here.

In Figure 3 (a) and 4 (b) (line a) it is shown how in a system of $M = 3$ students with $\eta_1 = \eta_2 = \eta_3$ two subsequent plateaus are visited which are characterized by $X^{(2)} = X \approx 0$ and $X^{(3)} \approx 0$ respectively. The use of only slightly different learning rates already breaks these symmetries very efficiently as can be seen in Fig. 4 (b) (lines b and c).

4 Summary and Outlook

In summary we have discussed two solvable models of unsupervised learning from high–dimensional data. In the thermodynamic limit, the dynamics of the considered on–line learning processes is described in terms of differential equations for a small number of order parameters.

Here we have focused on the process of student specialization in the course of learning. In both scenarios, plateau–like intermediate states of the system can dominate the time needed for successful training. These plateaus are due to the existence of weakly repulsive fixed points of the dynamics and reflect characteristic underlying symmetries.

In particular, we have studied the determination of prototype vectors from clustered example data by means of competitive learning. The investigated winner–takes–all procedure is identical with the on–line realization of the prominent K–means algorithm. It assigns each input deterministically to the prototype which is closest in distance. Obviously it is irrelevant precisely which of the prototypes represents which data cluster. This is similar to the permutation symmetry obeyed by the hidden nodes in a fully connected multilayered neural network and has analogous consequences.

The investigation of Sanger's rule for principal component analysis shows that repulsive fixed points and plateaus exist even though the algorithm imposes, by construction, a natural ordering on the student vectors. We have identified the relevant underlying symmetries and studied their effect on the learning dynamics.

Here, the effect of choosing different step sizes for different students has been demonstrated. Preliminary results show a similar, drastic improvement for the supervised training of over–sophisticated students, i.e. multilayer nets with an inappropriately large number of hidden units (Schwarze, 1998).

Further investigations shall address more complex model situations, for instance competitive learning with more than two clusters and prototypes. In particular, situations with a number of students which does not match the structure of the example data should be interesting.

Possible modifications of the competitive learning algorithm replace the *hard* step functions in Eq. (2.13) by a *soft minimum* type of assignment. As an example one could consider a stochastic gradient ascent maximization of the likelihood associated with a model distribution of the form (2.4), see e.g. (Bishop, 1995).

By introducing a topology in the space of prototypes it should be possible to extend the analysis to the dynamics of self–organized feature maps, see e.g. (Hertz *et al.*, 1991) for an introduction and related references.

Non–linear extensions of principal component analysis (Oja *et al.*, 1996) could be studied, which take into account higher moments of the presented data. In a sense such algorithms fill in the gap between the methods of prototype identification and low–dimensional representation.

Finally, for all these learning algorithms the use of a proper annealing schedule for the learning rates seems promising.

References

Amari, S., 1967, IEEE Trans. Elect. Comput. **EC–16**, 299

Amari, S., 1993, Neurocomputing **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. **38**, 1

Biehl, M. and E. Schlösser, 1998, J. Phys. A **31**, L97

Bishop C. M., 1995, 'Neural Networks for Pattern Recognition' (Clarendon Press, Oxford)

Copelli, M. and N. Caticha, 1995, J. Phys. A **28**, 1615

Copelli, M., R. Eichhorn, O. Kinouchi, M. Biehl, R. Simonetti, P. Riegler, and N. Caticha, 1997, Europhys. Lett. **37**, 432

Deco, G. and D. Obradovic, 1996, 'An Information-Theoretic Approach to Neural Computing' (Springer, Berlin)

Duda, R.O. and P.E. Hart, 1973, 'Pattern Classification and Scene Analysis' (Wiley, New York)

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

Kim, J. and H. Sompolinsky, 1996, Phys. Rev. Lett. **76**, 3021

Kinouchi, O. and N. Caticha, 1992, Phys. Rev. E **26**, 6243

Kinzel, W. and P. Rujan, 1990, Europhys. Lett. **13**, 473

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

Oja, E., 1982, J. Math. Biol. **15**, 267

Oja, E. and J. Karhunen, 1985, J. Math. An. Appl. **106**, 69

Oja, E., J. Karhunen, L. Wang, and R. Vigario, 1996, in 'Neural Nets, WIRN Vietri–95', eds. M. Marinaro and R. Tagliaferri (World Scientific, Singapore)

Opper, M., 1996, Phys. Rev. Lett. **77**, 4671

Opper, M. and W. Kinzel, 1996, in: 'Models of Neural Networks', Vol. III, eds. E. Domany, J.L. van Hemmen, and K. Schulten (Springer, Berlin)

Reents, G. and R. Urbanczik, 1998, 'Self–averaging and On–line learning', preprint Universität Würzburg

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

Sanger, T.D., 1989, Neural Networks **2**, 549

Schwarze, S., 1998, diploma thesis Universität Würzburg

Sompolinsky, H., N. Barkai, and H.S. Seung, 1995, in: 'Neural Networks: The Statistical Mechanics Perspective', eds. J.H. Oh, C. Kwon, and S. Cho (World Scientific, Singapore),

van den Broeck, C. and P. Reimann, 1996, Phys. Rev. Lett. **76**, 2188

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