Universität Würzburg

Institut für Theoretische Physik

Am Hubland, D–97074 Würzburg, Germany

The dynamics of on–line principal component analysis

Michael Biehl and Enno Schlösser

Ref.: WUE-ITP-97-047
ftp.physik.uni-wuerzburg.de

e-mail : biehl@physik.uni-wuerzburg.de
http://www.physik.uni-wuerzburg.de

The dynamics of on–line principal component analysis

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

Abstract

The learning dynamics of an on–line algorithm for principal component analysis is described exactly in the thermodynamic limit by means of coupled ordinary differential equations for a set of order parameters. It is demonstrated that learning is delayed significantly because existing symmetries among student vectors have to be broken. A closely related effect is the perfect or partial loss of initial knowledge in the course of learning. The analysis shows that different learning rates for the student vectors improve the performance of the algorithm drastically.

Various supervised and unsupervised learning techniques [1, 2] have been studied successfully by means of statistical mechanics methods in recent years [3, 4]. Although the focus of these efforts has been on supervised learning in feed-forward neural networks, interesting results have also been obtained for models of unsupervised learning [5, 6, 7, 8, 9].

A problem of particular importance in data analysis is the reduction of high-dimensional data to lower-dimensional representations which contain as much information about the original data as possible. One of the standard methods for this task is principal component analysis (PCA) which determines for a given set of observed data the eigenvectors corresponding to the largest eigenvalues of the empirical covariance matrix. Projections on these characteristic vectors can serve as a faithful linear representation of the data, see e.g. [1, 2] for detailed discussions. So far, the statistical mechanics analysis of PCA has been restricted to the learning of a single characteristic direction [5, 6].

In memory based off–line or batch learning prescriptions all the example data are stored and define a cost function. The learning process is then guided by the minimization of this function. In this Letter, however, we analyse the simultaneous learning of a number of principal components by use of on–line algorithms which are based on the presentation of single example data vectors [1, 10]. In practical situations data often are provided sequentially with no possibility to store them. In addition, on–line learning is the natural tool when the unknown data distribution is assumed to change during the learning process (on the same time scale) [11].

A deeper theoretical understanding of unsupervised on–line learning processes should be helpful in improving standard techniques or constructing new efficient algorithms. As a first step towards this goal we investigate *Sanger's Rule*, an on–line learning scheme which is well–known and widely used for PCA [1, 12].

The theoretical investigation of the learning dynamics is based on two important concepts: the consideration of high–dimensional inputs, and the performing of averages over the disorder introduced by the randomness in the data. A few order parameters are sufficient to describe the typical behavior of the system for an arbitrary number M of student vectors (with M small compared to the input dimension N).

We will demonstrate in this Letter that the necessary breaking of symmetries among the student vectors can dominate the time needed for the successful identification of the principal components. Furthermore, the analysis shows how this learning time can be reduced drastically by using learning rates which differ from student to student.

PCA takes into account second order statistics of the observed data only. Therefore, we consider a particularly simple input distribution: N-dimensional vectors $\underline{\xi}$ are taken to consist of components independently drawn from Gaussian distributions with zero mean. We assume the existence of M relevant directions $\{\underline{B}_i\}_{i=1,\ldots,M}$ in $I\!R^N$ (with $M \ll N$). The correlation matrix $C = \left\langle \underline{\xi}\,\underline{\xi}^\top \right\rangle$ is taken to be of the form

$$\underline{\underline{C}} = \underline{\underline{I}} + \sum_{i=1}^{M} (b_i^2 + 2b_i) \underline{B}_i \underline{B}_i^\top \tag{1}$$

with orthonormal vectors $\underline{B}_i^\top \underline{B}_j = \delta_{ij}$, the identity matrix $\underline{\underline{I}}$ and positive parameters $\{b_i\}_{i=1}^M$. The distribution can be interpreted as the result of deforming an isotropic N–dim. Gaussian cluster with data points $\underline{\tilde{\xi}}$ according to

$$\underline{\xi} = \underline{\tilde{\xi}} + \sum_{i=1}^{M} b_i (\underline{B}_i^\top \underline{\tilde{\xi}}) \underline{B}_i. \tag{2}$$

Note that the $\underline{B}_i$ are the eigenvectors of the correlation matrix $\underline{\underline{C}}$ corresponding to the eigenvalues $(1+b_i)^2$. Hence, assuming $b_1 \geq b_2 \geq \ldots \geq b_M > 0$, the set of vectors $\underline{B}_i$ corresponds by definition to the ordered principal components of the data distribution.

In our model learning scenario, a sequence of example vectors $\underline{\xi}^\mu$ is presented, which are generated independently according to the above specified distribution. A set of student vectors $\underline{J}_l \in I\!R^N \quad (l = 1, 2, \ldots, M)$ is updated according to Sanger's rule [12] upon representation of a single example:

$$\underline{J}_l(\mu) = \underline{J}_l(\mu-1) + \frac{\eta_l}{N} x_l^\mu \left(\underline{\xi}^\mu - \sum_{k=1}^{l} x_k^\mu \underline{J}_k(\mu-1) \right) \tag{3}$$

with the student projections $x_l^\mu = \underline{J}_l^\top \underline{\xi}^\mu$. The learning rates η_l control the magnitude of the updates of different students. Note that the dynamics of $\underline{J}_l$ depends only on the vectors $\underline{J}_k$ with $k \leq l$. The algorithm (3) can be shown to converge and yield normalized vectors ($\underline{J}_l^2 = 1$) in the limit $\eta_l \to 0$ [1]. Throughout this paper we assume an explicit normalization at each time step μ which involves additional terms of order η_l^2/N [6].

Sanger's rule (3) enforces an ordering of the student vectors which in general results in an identification of the vectors $\underline{B}_i$ upon presentation of infinitely many examples [1]. In contrast, a similar algorithm due to Oja [1, 13] is known to provide some basis of the corresponding subspace with the actual result depending on the initial configuration.

The following analysis exploits the fact that the quantities $x_k = \underline{J}_k^\top \underline{\xi}$ and $y_j = \underline{B}_j^\top \underline{\xi}$ (indices μ omitted) are zero mean Gaussian variables with covariances

$$\begin{aligned} \langle\, x_k y_j \,\rangle = (1+b_j)^2 R_{kj}, \quad \langle\, y_i y_j \,\rangle = (1+b_j)^2 \delta_{ij}, \quad \text{and} \\ \langle\, x_k x_l \,\rangle = Q_{kl} + \sum_{i}^{M} (b_i^2 + 2b_i) R_{ki} R_{li} \end{aligned} \tag{4}$$

Here the quantities $R_{kl}(\mu) = \underline{J}_k^\top(\mu)\underline{B}_l$ measure the overlaps of the student vectors with the unknown principal components, whereas the $Q_{kl}(\mu) = \underline{J}_k^\top(\mu)\underline{J}_l(\mu)$ correspond to their mutual overlaps ($Q_{kk} = 1$ due to normalization).

The system can be described exactly in terms of these order parameters in the limit $N \to \infty$ as they become self–averaging quantities [14]. On the other hand, it is straightforward to derive recursion relations from (3) which involve the random $\underline{\xi}$ only in terms of the projections $\{x_k, y_k\}$. Hence, the disorder average can be performed time step by time step. Furthermore, the averaged recursions can be interpreted as differential equations in continuous time $\alpha = \mu/N$ for $N \to \infty$, see for instance [15, 16, 17] for a more detailed discussion of the formalism. The dynamics is then described exactly by a system of $(3M^2 - M)/2$ coupled first order differential equations of the following form:

$$\begin{aligned} \frac{dR_{lj}}{d\alpha} &= \eta_l \langle\, x_l y_j \,\rangle - (\eta_l + \eta_l^2/2) \langle\, x_l^2 \,\rangle R_{lj}(t) \\ &\quad - \eta_l \sum_{k=1}^{l-1} \langle\, x_l x_k \,\rangle (R_{kj} - Q_{lk} R_{lj}) (k,l = 1,2,\ldots,M) \\ \frac{dQ_{lm}}{d\alpha} &= (\eta_l + \eta_m) \langle\, x_l x_m \,\rangle - ((\eta_l + \eta_l^2/2) \langle\, x_l^2 \,\rangle + (\eta_m + \eta_m^2/2) \langle\, x_m^2 \,\rangle) Q_{lm} \end{aligned} \tag{5}$$

$$-\eta_l \sum_{k=1}^{l-1} \langle\, x_l x_k \,\rangle (Q_{km} - Q_{kl} Q_{lm}) - \eta_m \sum_{k=1}^{m-1} \langle\, x_m x_k \,\rangle (Q_{lk} - Q_{km} Q_{lm}) (l \neq m)$$

All averages on the r.h.s. can be performed (4), yielding a closed set of equations. For specific initial conditions and learning rates, numerical integration yields the values of the order parameters for arbitrary α. In addition, an analytic treatment of fixed point properties allows to investigate the system in the limit $\alpha \to \infty$ and with respect to intermediate quasi-stationary states.

In order to measure the success of the learning process we consider the linear reconstruction $\underline{\xi}_{est} = \sum_{i=1}^{M} x_i \, \underline{J}_i$ of the original data $\underline{\xi}$ from a given set of projections $\{x_i\}$. The expectation value of the corresponding quadratic error is minimized for $\{\underline{J}_i = \underline{B}_i\}_{i=1,\ldots,M}$ or whenever the two sets of vectors span the same subspace, see [2].

Here, the average estimation error is

$$\varepsilon_{est} = \frac{1}{2} \left\langle \left(\underline{\xi}_{est} - \underline{\xi} \right)^2 \right\rangle - \frac{1}{2} \left\langle \underline{\xi}^2 \right\rangle = -\frac{1}{2} \sum_{k=1}^{M} \left\langle x_k^2 \right\rangle + \sum_{k=1}^{M} \sum_{l=1}^{k-1} \left\langle x_k x_l \right\rangle Q_{kl} \qquad (6)$$

and can be expressed in terms of the order parameters via (4). Note that the irrelevant constant $\left\langle \underline{\xi}^2 \right\rangle /2$ has been subtracted in the definition of ε_{est}. The evolution of the estimation error in the course of learning can be obtained via (6) from integrating (5), yielding the so–called learning curve $\varepsilon_{est}(\alpha)$.

Figure 1 shows a typical example of the evolvement of the cost function with α, the inset displays the corresponding diagonal overlaps R_{ll}. For small enough learning rates η_l the only attractive fixed point of the system is characterized by the asymptotic values

$$\begin{aligned} R_{ll}(\alpha \to \infty) &= \pm \sqrt{\frac{b_l^2 + 2b_l - \eta_l/2}{(b_l^2 + 2b_l)(1 + \eta_l/2)}} \\ R_{lj}(\alpha \to \infty) &= Q_{lj}(\alpha \to \infty) = 0 \qquad \text{for} \qquad l \neq j. \end{aligned} \qquad (7)$$

This configuration reflects the identification of a 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. 1). Very small values of η_l yield good learning success, but many examples are needed. With larger η_l learning becomes fast, but the asymptotic overlaps remain fairly small. Better results can be obtained with decaying, time dependent $\eta_l(\alpha)$, see the discussion below.

Note that for $\eta_l > \eta_l^{crit} = 2b_l(b_l + 2)$ a configuration with the corresponding overlap $R_{ll} = 0$ becomes stable. This means that no learning of the principal components is possible, which is analogous to the findings of [6]. Throughout the following, however, we will assume that all learning rates are smaller than their respective critical value. Although therefore (7) is the only purely attractive state of the system, additional

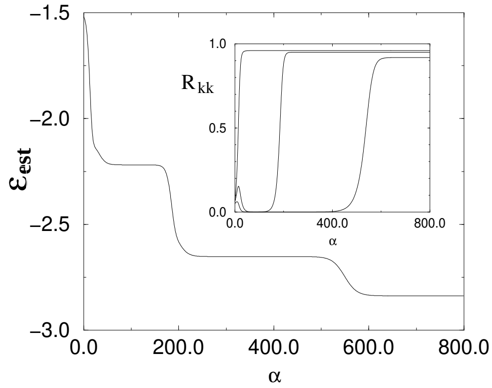

Figure 1: A typical learning scenario with $M = 3$: the cascade–like decrease of ε_{est} and the evolution of the corresponding diagonal overlaps R_{kk}. The learning rate is $\eta_l = 0.1$ for all students, all $R_{jk}(0) = 0.06$ apart from random deviation of the order $\mathcal{O}$ (10^{-10}).

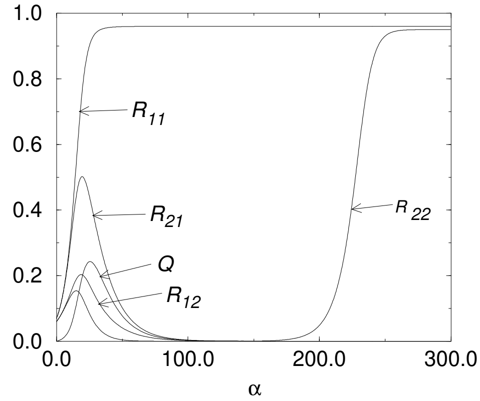

Figure 2: A learning scenario with M=2, described by the evolution of all overlaps of the students and the principal components. Initial conditions are: $R_{lk} = \mathcal{O}$ $(10^{-2}), Q_{12} = Q = \mathcal{O}$ (10^{-4}) and $X = \mathcal{O}$ (10^{-10}). $\eta_l = 0.1$ for both students.

repulsive states exist due to the underlying symmetries of the learning problem. After a proper relabeling of the students all these states could be characterized by (7) with some or all overlaps R_{ll} set to zero. Before approaching its asymptotic values, the system of Fig. 1 is trapped subsequently in the vicinity of such repulsive fixed points. There, the configuration is almost stationary and only the presentation of a large number of further examples enables another student to approach one of the principal components. This behavior is similar to the occurence of plateau states in supervised learning (see e.g. [16, 17]).

For simplicity we discuss the structure of these quasi–stationary states and their relevance for the learning dynamics in terms of the model with $M = 2$ and identical learning rates $\eta_1 = \eta_2 = \eta$ to begin with.

It is straightforward to show 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 [6] since the update of $\underline{J}_1$ is independent of all other vectors $\underline{J}_k (k \geq 2)$. A non–zero initial value of R_{11} will eventually yield the value given in (7). Here we focus on the dynamics of the subsequent students and assume $R_{11}(0) > 0$.

Due to the hierarchical structure of Sanger's algorithm the above consideration does not carry over directly to the evolution of R_{22} and other overlaps, i.e. $dR_{22}/d\alpha \neq 0$ can hold even if $R_{22} = 0$. Instead, it can be shown that the quantity

$$X = \begin{vmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{vmatrix} = R_{11}R_{22} - R_{12}R_{21} \quad \text{satisfies} \quad \frac{dX}{d\alpha} = 0 \quad \text{for} \quad X = 0. \tag{8}$$

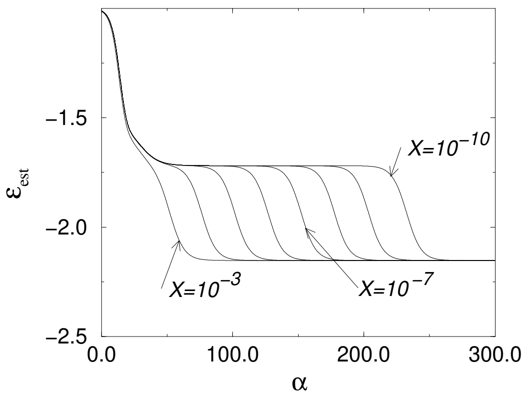

Figure 3: The plateau in the learning process shows a logarithmic dependence in X. Apart from X, initial conditions are the same as in fig. 2.

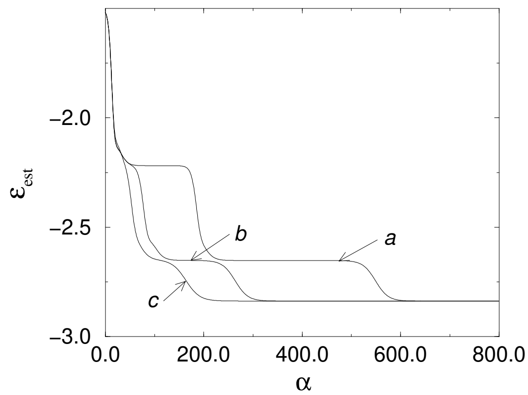

Figure 4: Estimation error for equal learning rates as shown in fig. 1 (a), learning rates $\eta_2 = 1.009\eta_1$ and $\eta_3 = 0.99\eta_1$ (b) and learning rates $\eta_2 = 1.09\eta_1$ and $\eta_3 = 0.9\eta_1$ (c).

A value of $X = 0$ corresponds to vectors $\underline{J}_i$ with linear dependent projections in the space spanned by the principal components. As a consequence of (8), a set of students with $X(0) = 0$ is not able to identify both principal components even with considerable initial knowledge ($R_{jk} > 0$ for all j, k). The conserved symmetry (8), together with the asymptotic orthogonalization of student vectors, enforces $R_{22} \to 0$ (and $R_{21} \to 0$) when R_{11} increases. This effective loss of prior information is illustrated in Fig. 2.

A linearization of the equations around $R_{jk} = 0$ for $j, k = 1, 2$ and $Q_{12} = 0$ shows that a non–zero $|X|$ will increase exponentially with α. Eqs. (5) decouple in the vicinity of this configuration and one obtains

$$X(\alpha) = X(0) \cdot e^{\lambda\,\alpha} \quad \text{with} \quad \lambda = (b_1^2 + 2b_1 + b_2^2 + 2b_2)\eta - \eta^2. \tag{9}$$

Hence, the time needed for the students to achieve a significant $X = \mathcal{O}(1)$ is $\alpha_p \approx -\ln|X(0)|/\lambda$. This is a measure for the time that the system will spent in the vicinity of the above specified repulsive fixed point. The logarithmic dependence of the *plateau length* α_p is demonstrated in Fig. 3 for a range of initial values $X(0)$. Randomly drawn initial student vectors without prior knowledge will result in small overlaps $R_{jk}(0)$ of order $\mathcal{O}(1/\sqrt{N})$ and $X(0) = \mathcal{O}(1/N)$. Thus the plateau length will diverge as $\alpha_p \propto \ln N$ with the system size in realistic scenarios.

It is important to observe that the conservation of the symmetry $X = 0$ (eq. (8)) holds true only for exactly identical learning rates $\eta_1 = \eta_2$. Without this restriction one obtains an equation of the form

$$dX/d\alpha = aX + b(\eta_1 - \eta_2)$$

indicating that $|X|$ will increase with α even for $X(0) = 0$. The quantities a and b are in general non-zero and depend on the R_{kj} and the Q_{kl}. For small differences $|\eta_1 - \eta_2|$ and $X(0) = 0$ the system escapes from the restricted subspace after a time of order $-\ln|\eta_1 - \eta_2|$. Sufficiently different learning rates prevent the system from approaching the vicinity of the intermediate state with $R_{22} = 0$. Note that this effect is not related to the natural ordering of the principal components. The improvement in comparison with $\eta_1 = \eta_2$ does not depend crucially on which learning rate is taken to be the smaller one.

In the model with M students and $\eta_l = \eta$ for all l the determinant $X^{(M)}$ of the matrix of overlaps R_{kj} shows a behavior analogous to eqs. (8) and (9). Due to the hierarchical structure of Sanger's rule the same is true for the sub-determinants $X^{(k)}$ with $k = 1, 2, ..., M$.

In figure 1 one can notice how a system with $M = 3$ units visits different plateaus consecutively, which correspond to $X^{(1)} = R_{11} \approx 0$, $X^{(2)} = X \approx 0$ and $X^{(3)} \approx 0$ respectively. Slightly different learning rates break these symmetries efficiently and enable the system to leave the quasi-stationary states much faster, see fig. 4.

Further improvements could be achieved with time-dependent learning rates $\eta_l(\alpha)$ which are roughly constant for small α and decay like $1/\alpha$ for $\alpha \to \infty$ [2]. Note that this modification affects the asymptotic result rather than the occurrence of the above described plateaus.

The impressive success of avoiding plateaus by simply choosing different learning rates is not limited to on-line PCA. Similar results are obtained for the supervised training of multilayer networks and other learning scenarios. These findings will be presented in forthcoming publications. Current research furthermore concerns the optimization of the learning rates η_l by means of variational methods [18, 19]

References

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

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

[3] T Watkin, A Rau, and M Biehl, Rev.Mod.Phys. **65** (1993) 499

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

[5] M Biehl and A Mietzner, Europhys. Lett. **24** (1993) 421, J.Phys. A **27** (1994) 1885

[6] M Biehl, Europhys. Lett. **25** (1994) 391

[7] E Lootens and C van den Broeck, Europhys. Lett. **30** (1995) 381

[8] C van den Broeck and P Reimann, Phys.Rev.Lett. **76**(1996) 2188

[9] A Freking, G. Reents, and M. Biehl, Europhys.Lett. **38** (1997) 1

[10] S Amari, IEEE Trans. Elect. Comput. **EC-16** (1967) 299, Neurocomp. **5** (1993) 185

[11] T. Heskes and B. Kappen, Phys. Rev. A **44** (1991) 2718

[12] T Sanger, Neural Networks **2** (1989) 549

[13] E Oja and J Karhunen, J.Math.An.Appl. **106** (1985) 69

[14] G Reents and R Urbanczik, *Self-averaging and on-line learning*, in preparation

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

[16] D Saad and S A Solla, Phys. Rev. Lett. **74** (1995) 4337, Phys. Rev. E **52** (1995) 4225

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

[18] O Kinouchi and N Caticha, J. Phys. A **25** (1992) 6243

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