Universität Würzburg

Institut für Theoretische Physik

Am Hubland, D–97074 Würzburg, Germany

Dynamics of On–line Competitive Learning

Michael Biehl, Ansgar Freking, Georg Reents

Ref.: WUE-ITP-96-025
ftp : ftp.physik.uni-wuerzburg.de

e-mail : biehl@physik.uni-wuerzburg.de

Dynamics of On–line Competitive Learning

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

Abstract

We present a solvable model of unsupervised competitive learning, which determines prototype vectors suitable for the representation of high-dimensional data. In the thermodynamic limit, the dynamics of on–line training is described exactly by a system of coupled first order differential equations for a set of order parameters. As an example application of the formalism we discuss the identification of two prototypes in the case of two overlapping clusters of data. This specific model exhibits non–trivial behavior like almost stationary plateau configurations which correspond to weakly repulsive fixed points of the dynamics. The ability of the system to escape from this fixed point as well as its asymptotic behavior depend critically on the learning rate used in the algorithm.

Methods from statistical physics have been applied successfully in the theory of learning. In order to investigate the typical behavior of large systems an average is performed over the disorder introduced by the randomness in the training data. Most of this theory has been concerned with the supervised learning of a rule. An overview of statistical mechanics contributions to this field can be found in *e.g.* [1, 2]. Recently, also unsupervised learning [3, 4, 5] has been studied in the statistical physics framework, see [6–12] and references therein.

One important possible objective of unsupervised learning is the determination of prototypes from a given (often very large) set of data $\left\{\boldsymbol{\xi}^\mu \in \mathbb{R}^N\right\}_{\mu=1,\ldots,P}$. The aim is to represent this set as faithful as possible by use of only a few typical *prototype* vectors $\{\mathbf{J}_k\}_{k=1,\ldots,K}$ ($K \ll P$) which capture the relevant features of the data. This is closely related, yet not identical, with the problem of clustering where the $\boldsymbol{\xi}^\mu$ are to be grouped into sets of similar vectors [3–9].

The learning process is usually guided by the minimization of an appropriate cost or energy function which incorporates the assignment of vectors $\boldsymbol{\xi}^\mu$ to the N–dimensional prototypes $\mathbf{J}_k$ and at the same time evaluates the quality of the representation. Very often, both is based on the Euclidean distances $d_k^\mu = (\boldsymbol{\xi}^\mu - \mathbf{J}_k)^2$ of the data from the respective prototype vectors.

We will consider in the following an on–line learning scenario, see for instance [13–16] and references therein. In this framework, only the latest from a sequence of example vectors $\boldsymbol{\xi}^\mu$ is used for updating the prototypes and the change is according to the negative gradient of an instantaneous energy of the form

$$H^\mu = \frac{1}{2} \sum_{k=1}^{K} d_k^\mu \;\; p_k \left(d_1^\mu, d_2^\mu, \ldots, d_K^\mu\right) \quad - \frac{1}{2} \left(\boldsymbol{\xi}^\mu\right)^2 . \tag{1}$$

Here, the labelling functions p_k obey the normalization constraint $\sum_{k=1}^{K} p_k = 1$ and therefore could be interpreted as a probabilistic assignment of the data to the prototypes. The constant $\left(\boldsymbol{\xi}^\mu\right)^2/2$ cancels exactly the first term from expanding out d_k^μ and is otherwise irrelevant.

In the update rule

$$\mathbf{J}_k(\mu) = \mathbf{J}_k(\mu-1) - \frac{\eta}{N} \left(\nabla_{\mathbf{J}_k} H^\mu\right)\Big|_{\left\{\mathbf{J}_k(\mu-1)\right\}} \qquad k = 1, \ldots, K, \tag{2}$$

the quantity η controls the step size of this stochastic gradient descent procedure and is scaled with the dimension N of the data. Usually, given a specific $\boldsymbol{\xi}^\mu$, the prescription (2) will affect one or several prototypes more efficiently than the others. This "competition" of the prototypes for updates is controlled by the labelling functions p_k. The term "competitive learning" has been coined for a variety of algorithms of this or a similar form [5].

Note that no normalization constraint is imposed on the prototype vectors. This is in contrast to models of directional clustering or feature detection [8, 11, 12], 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 learning from data which are provided by a stochastic source and are distributed according to a density of the specific form

$$P(\boldsymbol{\xi}) = \frac{1}{M} \sum_{m=1}^{M} P(\boldsymbol{\xi}|m) \quad \text{where} \quad P(\boldsymbol{\xi}|m) = \frac{1}{(2\pi)^{N/2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{\xi} - b_m \mathbf{B}_m\right)^2\right] . \tag{3}$$

This corresponds to a mixture of M overlapping Gaussians which are centered at $b_m \mathbf{B}_m$. We assume the $\mathbf{B}_m$ to be normalized ($\mathbf{B}_m^2 = 1$), so that the non–negative b_m are the distances of the centers from the origin. The dummy variable m indicates from which of the clusters $\boldsymbol{\xi}$ was drawn, all centers are taken to contribute with the same probability $1/M$. Note, however, that the labels m are not provided with the data in the course of learning.

The anisotropy of this data distribution is only very weak: given the cluster label m, the offset of the mean of $\boldsymbol{\xi}$ from the origin is $b_m \mathbf{B}_m$, which is a vector of length $b_m = \mathcal{O}(1)$. In comparison, the average length of the vectors $\boldsymbol{\xi}$ is $\mathcal{O}(\sqrt{N})$. Similar

distributions of data have been considered previously in models of both supervised and unsupervised learning (*e.g.* [8, 9, 11, 12, 17, 18]).

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 (3) 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 $P(\boldsymbol{\xi}|n)$):

$$\begin{aligned} \langle x_j x_k \rangle_n - \langle x_j \rangle_n \langle x_k \rangle_n &= Q_{jk} = Q_{kj} = \mathbf{J}_j \cdot \mathbf{J}_k \\ \langle x_k y_m \rangle_n - \langle x_k \rangle_n \langle y_m \rangle_n &= R_{km} = \mathbf{J}_k \cdot \mathbf{B}_m \\ \langle y_l y_m \rangle_n - \langle y_l \rangle_n \langle y_m \rangle_n &= T_{lm} = T_{ml} = \mathbf{B}_l \cdot \mathbf{B}_m \\ \langle x_k \rangle_n = b_n R_{kn} \quad &\text{and} \quad \langle y_m \rangle_n = b_n T_{mn} \end{aligned} \tag{4}$$

The self–averaging order parameters Q_{jk} and R_{km} together with the model parameters T_{mn} and b_m ($j, k = 1, \ldots, K$ and $m, n = 1, \ldots, M$) suffice in the limit $N \to \infty$ for a macroscopical description of the prototypes' properties. For example, the average $\langle H \rangle$ of the instantaneous energy over $P(\boldsymbol{\xi})$ (3) can be performed as an average over the joint density of the $\{x_k, y_m\}$ defined above. The result depends on the vectors $\mathbf{J}_k$ only implicitly through the order parameters.

The dynamics of the learning process can also be analysed in terms of the order parameters. It is straightforward to derive from (2) recursion relations for the evolution of these quantities. These equations are averaged over the latest example, assuming that at each time step a new, uncorrelated vector $\boldsymbol{\xi}$ is drawn according to (3). Again, 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.* [14, 19] 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.

As a specific example we have studied the model with $K = M = 2$, $T_{nm} = \delta_{nm}$, and $b_1 = b_2 = b$. The labelling functions are taken to be $p_1 = \Theta(d_2 - d_1)$ and $p_2 = \Theta(d_1 - d_2)$ in the following. For this choice a vector $\boldsymbol{\xi}$ is assigned deterministically to the prototype which is currently closest in distance. The corresponding training scheme

$$\mathbf{J}_k(\mu) = \mathbf{J}_k(\mu - 1) + \frac{\eta}{N} \left(\boldsymbol{\xi}^\mu - \mathbf{J}_k(\mu - 1) \right) \; p_k(d_1^\mu, d_2^\mu) \Big|_{\{\mathbf{J}_k(\mu-1)\}} \tag{5}$$

updates only the "winner" of this "competition", hence the name "winner takes all" algorithm for such a procedure [5]. Note that in (5) additional terms which involve derivatives of the Heaviside functions cancel exactly. The gradient is always in the direction $(\boldsymbol{\xi}^\mu - \mathbf{J}_k)$, hence 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 $\mathbb{R}^N$ where examples have been observed frequently. It is therefore a natural tool for the identification of clusters of data as modelled by the distribution (3).

The corresponding system of differential equations reads

$$\begin{aligned} \frac{\mathrm{d}R_{km}}{\mathrm{d}\alpha} &= \eta \left\langle (y_m - R_{km})\; p_k \right\rangle \\ \frac{\mathrm{d}Q_{lm}}{\mathrm{d}\alpha} &= \eta \left\langle (x_l - Q_{lm})\; p_m + (x_m - Q_{lm})\; p_l \right\rangle + \eta^2 \left\langle p_l\, p_m \right\rangle . \end{aligned} \tag{6}$$

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.

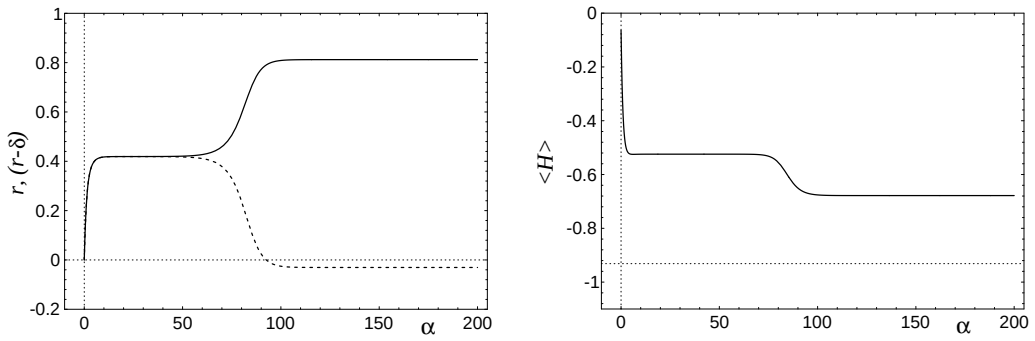
For a first discussion of the essential features of such learning curves we restrict our considerations to a partially symmetric subspace of the system configurations. We assume that the initial conditions obey the relations

$$Q_{11} = Q_{22} = Q, \quad Q_{12} = Q_{21} = Q - \varepsilon, \quad R_{11} = R_{22} = r\sqrt{Q}, \quad R_{12} = R_{21} = (r - \delta)\sqrt{Q}. \tag{7}$$

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 observe, 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.* [14, 15]. 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 $r(0) = \delta(0) = 10^{-6}$ and $Q(0) = \varepsilon(0) = 1$, the latter corresponding to normalized orthogonal vectors $\mathbf{J}_1(0)$ and $\mathbf{J}_2(0)$. In realistic situations with no *a priori* knowledge of the $\mathbf{B}_k$ one would also expect small values of r and δ initially. The averaged instantaneous energy $\langle H \rangle$ is plotted as a function of α in figure 1 b). This quantity is a measure for the success of the learning process: it evaluates the expected distance of a random vector $\boldsymbol{\xi}$ from the corresponding prototype. In a sense it is analogous to the so–called generalization error in supervised learning [2, 5, 20].

We will shortly discuss the main properties of the learning curves in the following. 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 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,



a) b)

Figure 1: Learning curves for the model explained in the text with $\eta = 1$, $b = 1.2$ and for initial conditions $r(0) = \delta(0) = 10^{-6}$, $Q(0) = \varepsilon(0) = 1$. In a) the evolution of the order parameters r (solid line) and $r - \delta$ (dashed) is shown, b) displays the corresponding averaged instantaneous energy as a function of α. The horizontal dotted line marks the minimal value of $\langle H \rangle$.

the system is trapped in a plateau–state in which the order parameters as well as $\langle H \rangle$ are approximately constant. Eventually, with increasing α, δ becomes significantly different from zero which corresponds to a specialization of the prototypes: each $\mathbf{J}_k$ represents mainly data from one of the clusters. This is also reflected in a decreasing value of the averaged instantaneous energy. The asymptotic behavior of the system for $\alpha \to \infty$ will be discussed below.

The phenomenon of plateaus in the learning curve has also been found in the context of supervised learning by gradient descent [14, 15, 19]. Note, however, that the dominant plateau in our model is not characterized by almost identical $\mathbf{J}_k$. A "mutual repulsion" is imposed on the prototype vectors which is due to the pronounced competitive nature of the "winner takes all" algorithm: 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 in the $\mathbf{B}_1, \mathbf{B}_2$–plane are almost equal in the plateau state.

The occurence of such a plateau is due to the existence of a fixed point of the differential equations (6) 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 [19]. 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.

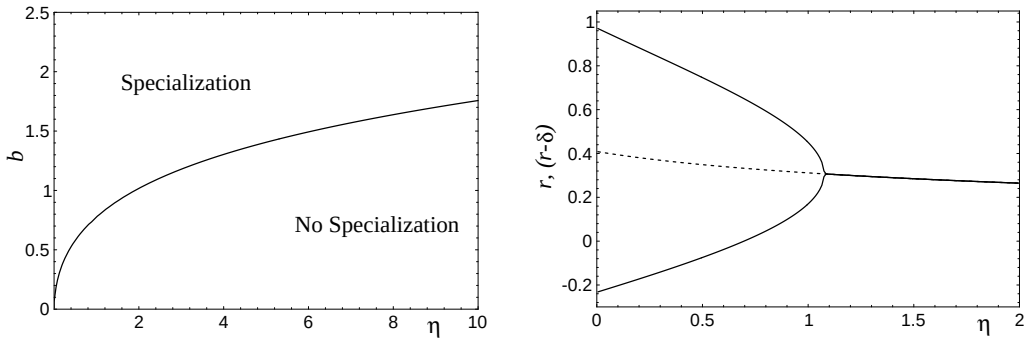


Figure 2: The role of the learning rate: In a) the solid line separates values of η and b for which the plateau state is stable from those which allow a specialization of the prototypes. The possible stationary values of r and $(r-\delta)$ are shown in b) for a separation $b = 0.8$. For $\eta > \eta_S \approx 1.08$ 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. In the limit $\eta \to 0$ the asymptotic values of r and $r - \delta$ correspond to the minimum of $\langle H \rangle$.

Figure 2 a) shows the region of possible specialization in the (η, b)–plane. The critical value of the learning rate increases with b. Thus, a more pronounced clustering of the data results in a larger range of η–values allowing successful learning. In figure 2 b) the stationary values of r and $r - \delta$ for $\alpha \to \infty$ are plotted versus the learning rate for a given 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 becomes repulsive and the stable asymptotic configuration is characterized by $\delta \neq 0$.

The minimum of the averaged energy is achieved by prototypes which are given as linear combinations of the $\mathbf{B}_k$ with $(\mathbf{J}_1 - \mathbf{J}_2) \propto (\mathbf{B}_1 - \mathbf{B}_2)$. Note, however, that this does not correspond to a perfect alignment or identity of each $\mathbf{J}_k$ with one of the cluster centers. This reflects a property of the cost function which is related to the above mentioned effective repulsion: the energy favors a larger relative distance of the prototypes [4]. The competitive training process does not aim explicitly at the identification of the cluster centers but at a good reliable representation of the data. In the large separation limit ($b \to \infty$) these two objectives become truly identical.

Note furthermore that only for $\eta \to 0$ the system is enabled to realize the smallest possible value of $\langle H \rangle$ (as for instance displayed by the dotted line in figure 1 b). Nonzero learning rates allow the prototypes to keep contributions from the space orthogonal to the $\mathbf{B}_k$. We find that the stationary value of $\langle H \rangle$ in the limit $\alpha \to \infty$ grows linearly with η in the vicinity of $\eta = 0$. This result suggests to employ an α–dependent step size as it has proven useful in similar learning scenarios [3, 5, 10, 13].

In summary, we have presented an exactly solvable model of unsupervised competitive learning. For the specific example of a "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 occurence of plateaus in the learning curves and the critical role of the learning rate.

A problem of particular interest is the choice of an appropriate time–dependent learning rate in order to guarantee convergence to the minimum of $\langle H \rangle$. Further investigations will also address more complex model situations, for instance with a mismatched number of prototypes ($K \neq M$) or non–spherical clusters of data. Different cost functions should be considered which go beyond the simple "winner takes all" scheme. The effect of a probabilistic "soft competitive" assignment of the prototypes on the dynamics and the success of learning is an important open question. Such algorithms are closely related to maximum likelihood methods or similar techniques of density estimation [3, 4, 9].

* * *

We would like to thank W. Kinzel, P. Riegler, and E. Schlösser, for useful discussions and a critical reading of the manuscript.

References

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

[2] Watkin, T.L.H., Rau, A., and Biehl, M., *Rev. Mod. Phys.*, **65** (1993)

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

[4] Duda, R.O. and Hart, P.E., *Pattern Classification and Scene Analysis*, (Wiley, New York) (1973) .

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

[6] Blatt, M., Wiseman, S., and Domany, E., *Phys. Rev. Lett.*, **76** (1996) 3251

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

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

[9] Barkai, N. and Sompolinksy, H., *Phys. Rev. E*, **50** (1994) 1766

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

[11] Biehl, M. and Mietzner, A., *J. Phys. A*, **27** (1994) 1885

[12] Watkin, T.L.H. and Nadal, J.–P., *J. Phys. A*, **27** (1994) 1889

[13] Amari, S., *Neurocomp.*, **5** (1993) 185

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

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

[16] Copelli, M., Kinouchi, and O., Caticha, N., *Phys. Rev. E*, **53** (1996) 6341

[17] Meir, R., *Neural Computation*, **7** (1995) 144

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

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

[20] Hansen, L.K. and Larsen, J., Danish Technical University Preprint (1996)