An exactly solvable model of unsupervised learning

Michael Biehl *
CONNECT, The Niels Bohr Institute
Blegdamsvej 17, DK-2100 Copenhagen Ø, Denmark

Abstract

A model for unsupervised learning from N–dimensional data is studied. Random training examples are drawn such that the distribution of their overlaps with a vector $\underline{B} \in \mathbb{R}^N$ is a mixture of two Gaussians of unit width and a separation ρ. A student vector is generated by an online algorithm, using each example only once. The evolution of its overlap with $\underline{B}$ can be calculated exactly in the thermodynamic limit $N \rightarrow \infty$. As a specific example, a learning algorithm closely related to Oja's rule is investigated. Its dynamics and approach to the stationary solution are solved for both a constant and an optimally chosen time–dependent learning rate. For the latter, the limits of small and infinitely large separation ρ of the peaks are considered. In both limits the analysis suggests the use of an asymptotic $(1/p)$–decay for the learning rate, where p is the number of training examples. In the large–separation limit, the typical number of examples needed for successful learning is found to be $(p/N) \propto \rho^{-2}$, which coincides with a recent result for supervised learning from Gaussian mixtures.

PACS. 87.10, 02.50, 05.90

The ability of Neural Networks to learn from examples has attracted a lot of attention [1]. In the *supervised learning* of a classification scheme a "teacher" provides the learning network with correctly labelled training inputs and the aim is to extract the underlying rule from these examples. Statistical Mechanics methods have been applied successfully to the learning of a rule (e.g. [2, 3]).

Unsupervised learning is the extraction of information from unlabelled data. For example, the training inputs might belong to different clusters in input space

*permanent address: Physikalisches Institut, Julius–Maximilians–Universität, D–8700 Würzburg, Germany

and the network is supposed to detect this structure and implement a classification accordingly. Such clustering or feature detection can be helpful in the construction of multilayer networks, as well as in data compression and encoding [1].

In the following a simple model is studied, in which the input data has a double– peak structure along a single direction $\underline{B}$ in N–space. The environment provides random examples $\underline{\xi}^{\nu} \in I\!R^N$ with a distribution of the overlaps $h^{\nu}_{\underline{B}} = \underline{B} \cdot \underline{\xi}^{\nu}$ given by

$$P\Big(h^{\nu}_{\underline{B}}\Big) = \frac{1}{2\sqrt{2\pi}} \Big(\exp\Big[-\frac{1}{2} \Big(h^{\nu}_{\underline{B}} - \rho \Big)^2 \Big] + \exp\Big[-\frac{1}{2} \Big(h^{\nu}_{\underline{B}} + \rho \Big)^2 \Big] \Big). \tag{1}$$

The parameter ρ measures the separation of the "clouds of data", the width of the single peaks is set to 1.

The input vectors are taken to consist of independent random components with zero mean and unit variance, such that $\underline{\xi}^{\nu} \cdot \underline{\xi}^{\nu} = N$ is exactly satisfied in the thermodynamic limit $N \to \infty$. The effect of this restriction is nontrivial and will be discussed elsewhere [4].

In the case of binary $B_j = \pm 1/\sqrt{N}$, for instance, these properties can be realized by first drawing dummy variables, $\sigma^{\nu} = +1$ or -1 with equal probability, and then the independent, binary components of the examples according to

$$P(\xi^{\nu}_j \mid \sigma^{\nu}) = \frac{1 + \rho/\sqrt{N}}{2} \delta\Big(\xi^{\nu}_j - \sqrt{N} B_j \sigma^{\nu} \Big) + \frac{1 - \rho/\sqrt{N}}{2} \delta\Big(\xi^{\nu}_j + \sqrt{N} B_j \sigma^{\nu} \Big). \tag{2}$$

The distribution (1) then follows for large N, according to the central limit theorem. Note, however, that the explicit form of $P(\xi^{\nu}_j)$ is irrelevant for the following.

A learning algorithm chooses, after the presentation of, say, $\nu{-}1$ such training inputs, a normalized vector $\underline{J}(\nu{-}1) \in I\!R^N$. Then the overlap $R(\nu{-}1) = \underline{J}(\nu{-}1) \cdot \underline{B}$ indicates how well the underlying structure has been inferred. Given a value of $R(\nu - 1) = R$, the joint probability distribution for the overlaps $h^{\nu}_{\underline{B}}$ and $h^{\nu}_{\underline{J}} = \underline{J}(\nu - 1) \cdot \underline{\xi}^{\nu}$ of a new and uncorrelated input example $\underline{\xi}^{\nu}$ is

$$P\Big(h_{\underline{B}}, h_{\underline{J}}\Big) = \frac{1}{2\pi\sqrt{1-R^2}} \frac{1}{2} \sum_{S=\pm 1} \exp\Big[-\frac{h^2_{\underline{J}} + h^2_{\underline{B}} - 2 h_{\underline{J}} h_{\underline{B}} R}{2(1-R^2)} - \frac{\rho^2}{2} + h_{\underline{B}} \rho S \Big] \tag{3}$$

where the indices ν have been omitted for simplicity.

Note that for $R = 0$, the distribution factorizes and $P(h^{\nu}_{\underline{J}})$ becomes a single Gaussian peak with zero mean, reflecting the fact that the data appears structureless in the subspace perpendicular to $\underline{B}$.

In the following, learning proceeds "online" [1], that is each example is presented only once, the vector $\underline{J}$ is updated "instantaneously" and then the next input is drawn. Examples are not stored and presented iteratively as in memory based or "offline"–schemes. Memory based unsupervised learning from data of the type considered here has been studied recently [5].

The analysis of the learning process follows closely the methods used in e.g. [6, 7, 8] for supervised online learning.

The generic form of the unsupervised online algorithm studied here is

$$\underline{J}(\nu) \propto \underline{J}(\nu-1) + \frac{\gamma}{N} f(h^\nu_{\underline{J}})\, \underline{\xi}^\nu \tag{4}$$

when example ν is being presented. The parameter γ is called the learning rate and the weight function $f(h^\nu_{\underline{J}})$ defines the actual algorithm. It depends only on the overlap of the current example with the student vector, which is clearly available in unsupervised learning. Normalization of $\underline{J}(\nu)$, assuming $|\,\underline{J}(\nu-1)\,|^2 = 1$, yields to first order in $1/N$ the prescription

$$\begin{aligned}
\underline{J}(\nu) &= \left(\underline{J}(\nu-1) + \frac{\gamma}{N} f(h^\nu_{\underline{J}})\underline{\xi}^\nu\right) \Big/ \sqrt{1 + \frac{2\gamma}{N} f(h^\nu_{\underline{J}})h^\nu_{\underline{J}} + \frac{\gamma^2}{N} f^2(h^\nu_{\underline{J}})} \\
&= \underline{J}(\nu-1) + \frac{1}{N}\left(\gamma f(h^\nu_{\underline{J}})\underline{\xi}^\nu - \left[\gamma f(h^\nu_{\underline{J}})h^\nu_{\underline{J}} + \frac{\gamma^2}{2} f^2(h^\nu_{\underline{J}})\right]\underline{J}(\nu-1)\right).
\end{aligned} \tag{5}$$

Thus the normalization of $\underline{J}$ is imposed through a proper weight decay term [1].

From eq.(5) one can immediately derive a recursion relation for the overlap R:

$$R(\nu) = R(\nu-1) + \frac{1}{N}\left(\gamma f(h^\nu_{\underline{J}})h^\nu_{\underline{B}} - \left[\gamma f(h^\nu_{\underline{J}})h^\nu_{\underline{J}} + \frac{\gamma^2}{2} f^2(h^\nu_{\underline{J}})\right] R(\nu-1)\right). \tag{6}$$

The average over the latest example can now be performed by averaging over the joint density (3) with $R = R(\nu-1)$. Assuming selfaveraging properties in the limit $N \to \infty$ the resulting recursion relation can be interpreted as a differential equation for the actual value $R(\alpha)$ in "continuous time" $\alpha = \nu/N$:

$$\frac{dR}{d\alpha} = \gamma\Big\langle f(h_{\underline{J}})h_{\underline{B}}\Big\rangle - \left(\gamma\Big\langle f(h_{\underline{J}})h_{\underline{J}}\Big\rangle + \frac{\gamma^2}{2}\Big\langle f^2(h_{\underline{J}})\Big\rangle\right) R(\alpha) \tag{7}$$

where $\langle \ldots \rangle$ denotes an average over (3).

For any choice of algorithm, learning rate, and initial condition this equation can be solved (at least numerically) and yields the evolution of R for the learning from the considered type of data.

As a specific example, $f(h^\nu_{\underline{J}}) = h^\nu_{\underline{J}}$ is studied in the following. This choice corresponds to the so–called Hebbian unsupervised learning [1]: The change of the weight vector might be interpreted as "input$\times$output" of a linear unit [1]. The normalized algorithm (5) reads now

$$\underline{J}(\nu) = \underline{J}(\nu-1) + \frac{1}{N}\left(\gamma h^\nu_{\underline{J}}\underline{\xi}^\nu - \left(\gamma + \frac{\gamma^2}{2}\right)(h^\nu_{\underline{J}})^2 \underline{J}(\nu-1)\right). \tag{8}$$

To first order in γ this coincides with Oja's rule [9], which is used in the context of Principal Component Analysis [1] for detecting the direction of largest variance in the input data.

The resulting differential equation for R is, after performing the averages $\left\langle h_{\underline{J}} h_{\underline{B}} \right\rangle = R(1+\rho^2)$ and $\left\langle h_{\underline{J}}^2 \right\rangle = 1 + R^2\rho^2$,

$$\frac{dR}{d\alpha} = -\frac{\gamma^2}{2} R + \rho^2 \left(\gamma R - R^3 \left(\gamma + \frac{\gamma^2}{2} \right) \right). \tag{9}$$

This equation has always a fixpoint $\frac{dR}{d\alpha} = 0$ for $R = 0$. However, this fixpoint is repulsive for

$$\gamma < 2\rho^2. \tag{10}$$

The learning rate has to be sufficiently small in order to enable the system to detect the structure in the data. A constant value of γ which satisfies condition (10) results in a a nontrivial attractive fixpoint

$$R_\infty = \pm \sqrt{\left(\frac{2}{2+\gamma} \right) \left(1 - \frac{\gamma}{2\rho^2} \right)} \tag{11}$$

where the actual sign depends on on the initial condition $R(0) = R_o$. The problem is symmetric in the sense that $+\underline{J}$ and $-\underline{J}$ represent the same classification, with only the orientation changed.

It is important to note that only in the limit $\gamma \to 0$ the student coincides asymptotically with $\pm\underline{B}$ and thus infers the structure of the data perfectly:

$$\mid R_\infty \mid = 1 - \frac{\gamma}{2} \left(1 + \frac{1}{\rho^2} \right) \quad \text{as} \quad \gamma \to 0. \tag{12}$$

For a constant γ the solution of eqn. (9) is given by

$$R(\alpha) = R_\infty \left[1 + \left(\frac{R_\infty^2}{R_o^2} - 1 \right) \exp\left[-\gamma \left(2\rho^2 - \gamma \right) \alpha \right] \right]^{-1/2}. \tag{13}$$

The approach to R_∞ is exponentially fast, but with a relaxation time which diverges as $\gamma \to 0$ and $R_\infty \to 1$. Fig. 1 shows the evolution of R for given values of γ and R_o. Simulations are in good agreement with the analytical result.

The above analysis suggests decreasing the learning rate with the number of examples presented, a strategy familar from other learning schemes or optimization problems. Introducing an α–dependent rate into eqn. (9) formally allows for choosing the optimal γ at each time–step by maximizing $| dR/d\alpha |$, yielding

$$\gamma(\alpha) = \frac{\rho^2(1 - R^2(\alpha))}{1 + \rho^2 R^2(\alpha)}. \tag{14}$$

Unfortunately neither the separation ρ of the data, nor the actual value of R is available in any practical situation. But the following analysis provides an upper bound on the value $R(\alpha)$ that can be achieved by an algorithm of the form (8). Fig. 2 shows, for three different values of ρ, the numerical solution of (9) with $\gamma = \gamma(\alpha)$ from equ. (14), which is also plotted. Note, that $R_\infty = \pm 1$ is guaranteed for any nonzero ρ and R_o by optimally choosing the learning rate.

It is instructive to investigate two special cases:

$\boldsymbol{\rho \rightarrow 0}$**:** In the limit of a very small separation, $\rho \rightarrow 0$, the optimal choice of the rate is $\gamma = \rho^2(1-R^2)$, which is substituted into (9) to obtain

$$\frac{dR}{d\alpha} = \frac{1}{2}\rho^4 R(1-R^2)^2. \tag{15}$$

The solution is asymptotically of the form

$$\begin{aligned} R^2(\alpha) &\approx 1 - \frac{1}{\rho^4 \alpha} \\ \text{and correspondingly} \quad \gamma(\alpha) &\approx \frac{1}{\rho^2 \alpha} \quad \text{as} \quad \alpha \rightarrow \infty. \end{aligned} \tag{16}$$

The typical number of examples needed for succesful learning diverges like $1/\rho^4$ as $\rho \rightarrow 0$.

$\boldsymbol{\rho \rightarrow \infty}$: In the large–separation limit, if $\rho^2 >> 1/R^2(\alpha)$ is fulfilled for all α, the optimal learning rate is to first order in $1/\rho^2$ given by $\gamma(\alpha) = \frac{1}{R^2} - 1$, and (9) reads

$$\frac{dR}{d\alpha} = \frac{1}{2}\rho^2 \frac{1}{R}\left(1-R^2\right)^2 \tag{17}$$

with the solution (for all α)

$$\begin{aligned} R^2(\alpha) &= 1 - \frac{1}{\rho^2 \alpha + c} \\ \text{and} \quad \gamma(\alpha) &= \frac{1}{\rho^2 \alpha + c - 1} \quad \text{where} \quad c = \frac{1}{1-R_o^2}. \end{aligned} \tag{18}$$

Given a large separation, only very few examples are needed to detect the direction of $\underline{B}$, typically $\alpha \propto 1/\rho^2$. It is interesting to note that the same scaling has been found in supervised learning from Gaussian mixtures in the large–separation limit [10]. In this regime the structure of the data is so "obvious" that the additional information given by a teacher is redundant.

Learning rate schedules with $\gamma(\alpha)$ being linear (or roughly constant) for small α and decreasing like $(1/\alpha)$ asymptotically have proven useful in other supervised and unsupervised learning schemes (e.g. [11, 12, 13]). An asymptotic $\gamma \propto (1/\alpha)$ is known to yield optimal convergence for more general stochastic gradient methods [14, 15].

It might be interesting to extend the presented analysis to both more sophisticated learning algorithms and the learning from more complicated data. In particular, it seems promising to apply the techniques used here to the problem of *competetive* unsupervised learning [1].

Acknowledgement
This work has benefitted very much from discussions with W. Kinzel, B. Lautrup, A. Mietzner, H. Schwarze, and S. Solla. I thank L.K. Hansen and J. Hertz for bringing refs. [12–15] to my attention. I am grateful for the hospitality and most inspiring atmosphere at CONNECT and the Niels Bohr Institute. This work was supported by the Deutsche Forschungsgemeinschaft.

References

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

[2] H.S. Seung, H. Sompolinsky, and N. Tishby, Phys.Rev. **A45**, 6056 (1992).

[3] T.L.H. Watkin, A. Rau, and M. Biehl, *The Statistical Mechanics of Learning a Rule*, Rev. Mod. Phys., 1993 , in press

[4] M. Biehl and A. Mietzner, in preparation.

[5] M. Biehl and A. Mietzner, *Statistical Mechanics of Unsupervised learning*, preprint Universität Würzburg, 1993

[6] W. Kinzel and P. Rujan, Europhys.Lett. **13**, 473 (1990).

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

[8] M. Biehl and H. Schwarze, Europhys. Lett. **20**, 733 (1992).

[9] E. Oja, J.Math.Biol. **15**, 267 (1982).

[10] N. Barkai, H.S. Seung, and H. Sompolinsky, *Scaling Laws in Learning of Classification Tasks*, preprint Hebrew University Jerusalem, 1993

[11] T.D. Sanger, in *Advances in Neural Information Processing Systems 1*, ed. D.S. Touretzky, Morgan Kaufmann, San Mateo, 1989, pg. 11

[12] C. Darken and J. Moody, in *Advances in Neural Information Processing Systems 3*, eds. R.P. Lippmann, J.E. Moody, and D.S. Touretzky , Morgan Kaufmann, San Mateo, 1991, pg. 832

[13] C. Darken and J. Moody, in *Advances in Neural Information Processing Systems 4*, eds. J.E. Moody, S.J. Hanson, and R. P. Lippmann, Morgan Kaufmann, San Mateo, 1992, pg. 1009

[14] L. Goldstein, "Mean Square optimality in the continous time Robbins Monro procedure", Technical Report DRB-306. Dept. of Math., Univ. of Southern California, Los Angeles, 1987

[15] H. Whites, Neural Computation **1**, 425 (1989).

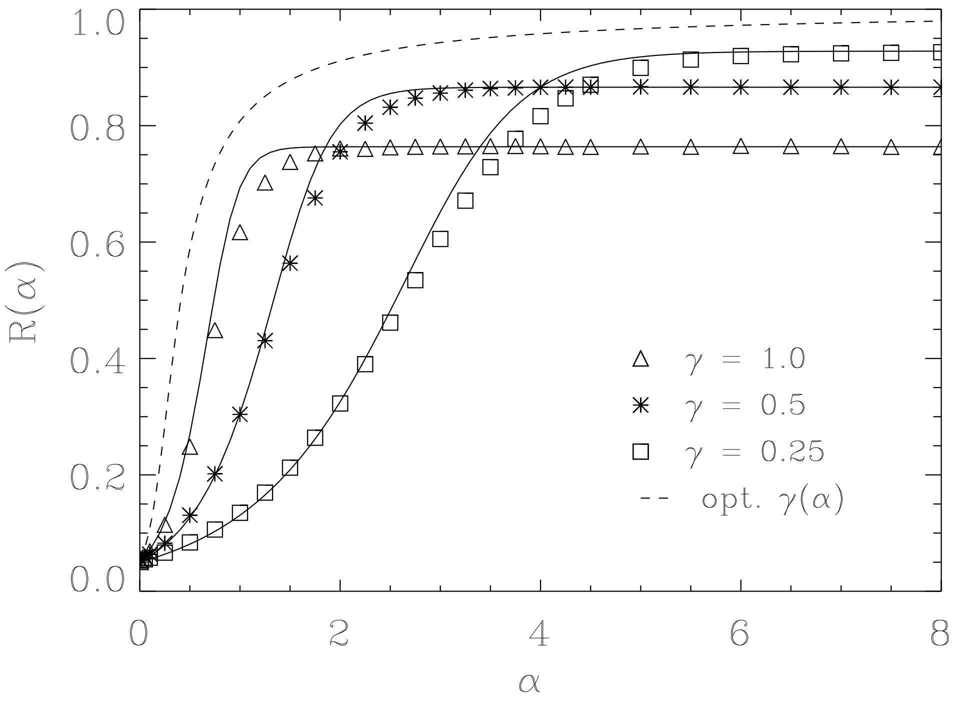
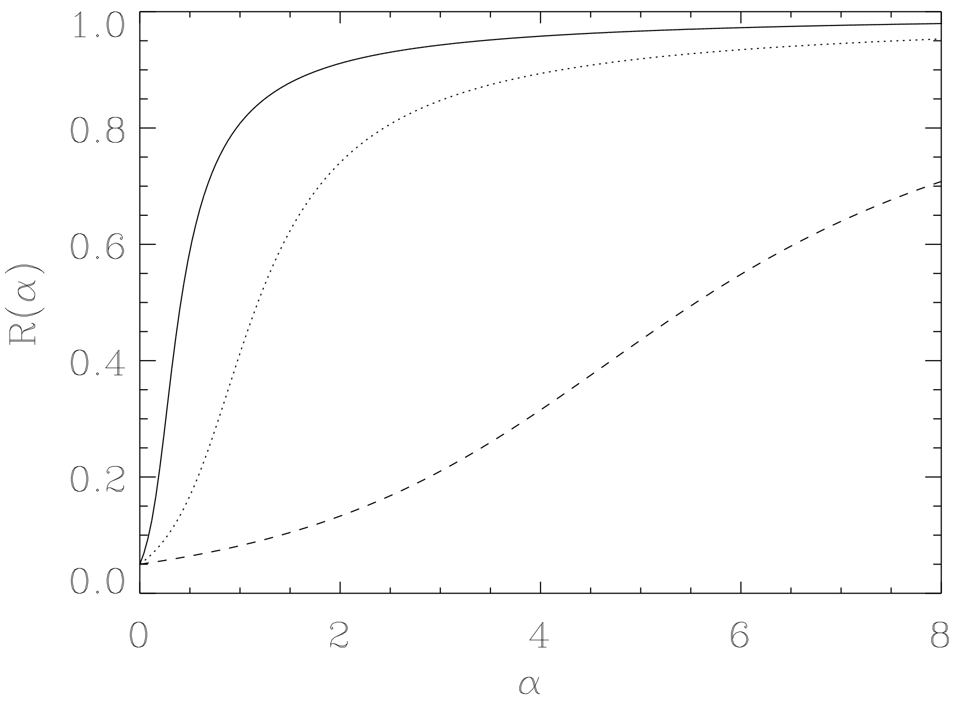
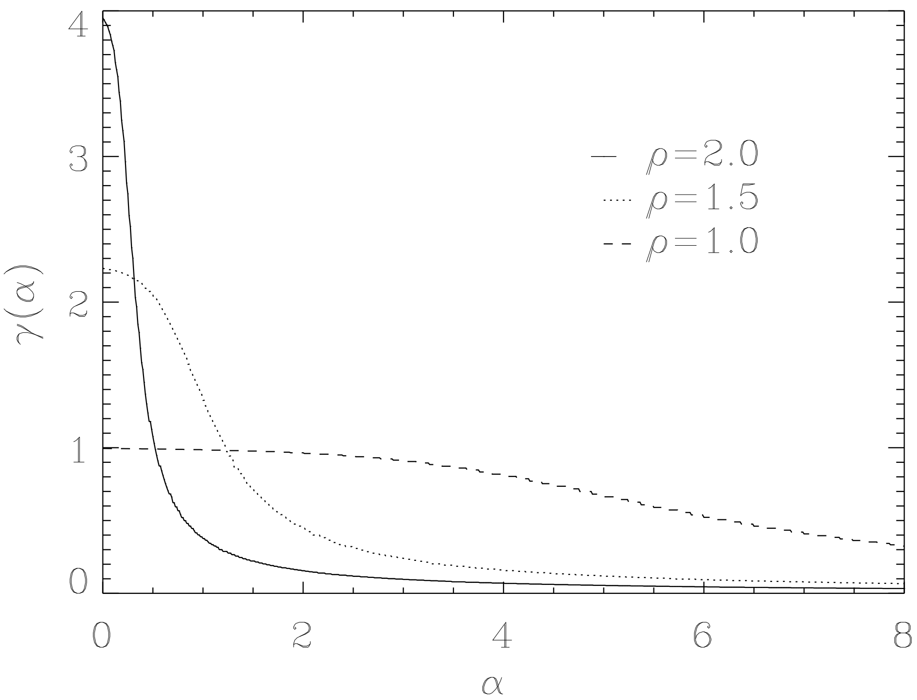


Figure 1: Evolution of the overlap R with α, according to eqn. (13). All curves are for a separation $\rho = 2$ and initial overlap $R_o = 0.05$, but for different learning rate. As γ is decreased, the asymptotic value R_∞ approaches 1, but the relaxation slows down. Simulations were done for $N = 2000$, the averages were performed over 200 independent sets of training examples. Error bars would be smaller than the symbols. Additionally: $R(\alpha)$ for the optimally chosen learning rate $\gamma(\alpha)$, bounding all curves for constant γ.



a)

Figure 2: Learning with an optimally chosen learning rate, for an initial condition $R_o = 0.05$ and different values of ρ. Fig. 2a) shows $R(\alpha)$, 2b) the corresponding rate $\gamma(\alpha)$ according to eqn. (14).



b)