Statistical Mechanics of Unsupervised Structure Recognition

Michael Biehl and Andreas Mietzner
Physikalisches Institut, Julius–Maximilians–Universität
Am Hubland, D–97074 Würzburg, Germany

Abstract

A model of unsupervised learning is studied, where the environment provides N–dimensional input examples that are drawn from two overlapping Gaussian clouds. We consider the optimization of two different objective functions: the search for the direction of the largest variance in the data and the largest separating gap (stability) between clusters of examples respectively.

By means of a statistical mechanics analysis, we investigate how well the underlying structure is inferred from a set of examples. The performances of the learning algorithms depend crucially on the actual shape of the input distribution. A generic result is the existence of a critical number of examples needed for successful learning. The learning strategies are compared with methods different in spirit, such as the estimation of parameters in a model distribution and an information theoretical approach.

PACS. 87.10 02.50 05.90

1 Introduction

One of the major features of Neural Networks is their ability to learn from examples [1]. In supervised learning the environment provides a set of training inputs together with the correct outputs, e.g. the labelling according to a binary classification, available from a teacher. From this information the student–network might infer the unknown rule, which defines the output for any possible input. Various models of supervised learning have been studied by means of statistical mechanics, for reviews see e.g. $[2-4]$.

In unsupervised learning [1, 5, 6] there is no teacher and only unlabelled inputs are available. The task is to infer an underlying structure in the data, i.e. to recognize the relevant features that allow for a clustering or classification. No obvious simple quality measure exists, such as the agreement with the teacher in supervised learning. Therefore it is necessary to constitute criteria that enable us to formulate unsupervised learning as an optimization process. Only recently this type of learning has been considered in a statistical mechanics context $[7-11]$.

A simple model of an unsupervised learning task is the detection of a single direction along which possible inputs form two overlapping "clouds". In the following this structure is imposed by generating data according to a specific probability distribution (section 2). Its parameters allow to shape the clusters and thus model different situations that might occur in real world problems in a similar manner.

The learning strategies are based on the optimization of intuitive objective functions (section 3): the maximization of the output variance and the search for separating gaps between the clusters. These intuitive approaches make no use of any a priori knowledge about the inputs. The specifically chosen distribution is only an example for which the typical properties of the learning prescriptions can be studied. This is done by use of the replica–method [12], in analogy with the theory of supervised learning.

The results (section 4), although obtained for a specific model, demonstrate general difficulties that might occur in any unsupervised learning problem. For instance, the chosen objective function might be inappropriate for the unknown structure to be detected. Furthermore, even though a suitable prescription is used, successful learning can require a minimal number of example inputs.

The discussion in section 5 compares also with approaches different in spirit, such as the estimation of parameters in a model distribution or the choice of a more sophisticated objective function based on information theory.

2 The model data

We consider N–dimensional random inputs $\underline{\xi}^\nu \in I\!R^N$, distributed according to two Gaussian clouds centered around $\pm\frac{\rho}{\sqrt{N}}\underline{B}$, where $\underline{B}$ is an N–dimensional vector with $\underline{B}^2 = N$. The distribution of the overlaps

$$h^\nu_{\underline{B}} = \frac{1}{\sqrt{N}}\underline{B}\cdot\underline{\xi}^\nu \tag{1}$$

is taken to have a double peak structure:

$$P\left(h^\nu_{\underline{B}}\right) = \frac{1}{2\sqrt{2\pi}}\left\{\exp\left[-\frac{1}{2}\left(h^\nu_{\underline{B}}-\rho\right)^2\right] + \exp\left[-\frac{1}{2}\left(h^\nu_{\underline{B}}+\rho\right)^2\right]\right\} \tag{2}$$

where ρ is called the separation of the clouds and the width of a single peak is set to 1.

As an illustrative example one might consider binary $B_j = \pm 1$ and generate independent inputs according to

$$\begin{aligned} P(\widehat{\xi^\nu_j} \mid \sigma^\nu) &= \frac{1+\rho/\sqrt{N}}{2}\delta\left(\widehat{\xi^\nu_j} - B_j\sigma^\nu\right) + \frac{1-\rho/\sqrt{N}}{2}\delta\left(\widehat{\xi^\nu_j} + B_j\sigma^\nu\right) \quad \text{and} \\ P(\sigma^\nu) &= \frac{1}{2}\{\delta(\sigma^\nu - 1) + \delta(\sigma^\nu + 1)\}, \end{aligned} \tag{3}$$

where the dummy variables σ^ν indicate which cloud $\underline{\xi}^\nu$ belongs to. Note that the introduced correlations among the patterns are a factor of $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ weaker than in the case of biased patterns [13], where $\sum_j \xi_j^\nu = \mathcal{O}(N)$. The distribution $P\left(h_{\underline{B}}^\nu = \frac{1}{\sqrt{N}}\underline{B}\cdot\widehat{\underline{\xi}^\nu}\right)$ as given in eq. (2) follows by use of the central limit theorem for $N \to \infty$.

Accordingly, the distribution of the projections on any direction perpendicular to $\underline{B}$ would be a single Gaussian with zero mean and unit variance. Thus, (3) results in clouds which are spherically symmetric about their respective center.

The shape of the clouds can be modified by adding arbitrary contributions from the subspace orthogonal to $\underline{B}$ without changing the $h_{\underline{B}}^\nu$. Therefore we consider in the following inputs with random components given by

$$\xi_j^\nu = \gamma \cdot \widehat{\xi_j^\nu} + (1-\gamma)\left(\frac{1}{N}\underline{B}\cdot\widehat{\underline{\xi}^\nu}\right)B_j\,, \tag{4}$$

where $\gamma > 0$.

Their overlaps $h_{\underline{B}}^\nu$ and $h_{\underline{J}}^\nu = \frac{1}{\sqrt{N}}\underline{J}\cdot\underline{\xi}^\nu$ for a vector $\underline{J}$ with $\underline{J}^2 = N$ and $\frac{1}{N}\underline{J}\cdot\underline{B} = R$ are distributed according to the joint density

$$P_R(h_{\underline{J}}^\nu, h_{\underline{B}}^\nu) = \frac{1}{2}\sum_{S=\pm1}\frac{1}{2\pi\gamma\sqrt{1-R^2}}\exp\left[-\frac{1}{2}\frac{\left(h_{\underline{J}}^\nu - Rh_{\underline{B}}^\nu\right)^2}{\gamma^2(1-R^2)} - \frac{1}{2}\left(h_{\underline{B}}^\nu - \rho S\right)^2\right] \tag{5}$$

For $R = 0$, i.e. $\underline{J}\perp\underline{B}$, this density factorizes and we get the independent $P(h_{\underline{B}}^\nu)$ of (2) and

$$P_{R=0}(h_{\underline{J}}^\nu) = \frac{1}{2}\sum_{S=\pm1}\frac{1}{2\pi\gamma}\exp\left[-\frac{1}{2\gamma^2}\left(h_{\underline{J}}^\nu\right)^2\right] \tag{6}$$

Thus the parameter γ controls the width in the orthogonal subspace: $\left\langle (h_{\underline{J}}^\nu)^2 \right\rangle_{R=0} = \gamma^2$. Here and in the following $\langle \ldots \rangle_R$ denotes an average over (5). A value of $\gamma < 1$ yields a higher density close to the symmetry–axis $\underline{B}$, the clouds form prolate "cigars", whereas $\gamma > 1$ results in an oblate shape (N-dimensional "pancakes"). Together with the separation ρ, the parameter γ will determine how well the relevant direction $\underline{B}$ can be detected.

It is important to note that our results would also hold for the equivalent continuous version of (3) [10], yielding the same $P_R(h_{\underline{J}}^\nu, h_{\underline{B}}^\nu)$. The discreteness of $\underline{B}$ and the inputs is never explicitely used in the learning process, there is no such a priori knowledge assumed, see the discussion in section 5.

3 Learning Strategies

In our model learning is the choice of a direction $\underline{J}$ according to a specific criterion. In the following we introduce and discuss two such criteria. For a given set of $p = \alpha N$ example inputs, $\underline{J}$ is taken to minimize a corresponding objective function. The success of learning, however, is measured by the resulting overlap$R = N^{-1}\underline{J}\cdot\underline{B}$.

3.1 Maximal Variance

In many cases one may expect that directions in which the data varies a lot contain much information about an underlying structure (see also the discussion in the last section). The search for maximal variance is a common strategy for obtaining meaningful directions, especially for high dimensional data sets. It is most often referred to as principal component analysis [1, 14].

The corresponding objective function is

$$H = -\sum_{\nu=1}^{p} \left(h_{\underline{J}}^{\nu}\right)^2. \tag{7}$$

In a geometric interpretation this corresponds to maximizing the mean square distance of the input patterns from the hyperplane perpendicular to the vector $\underline{J}$, which separates two classes of patterns. There exist iterative learning algorithms, e.g. Oja's rule or its modifications [1, 14], which indeed minimize H. In our case of zero mean data the minimum of H is given by the normalized eigenvector $\underline{J}$ corresponding to the largest eigenvalue of the correlation matrix $\mathbf{C} = \sum_{\nu=1}^{p} \underline{\xi}^{\nu} \cdot \underline{\xi}^{\nu T}$. In the context of a linear perceptron unit this is equivalent to a maximization of its output variance [15].

Of course we expect this strategy to work well whenever $\underline{B}$ is indeed the direction of maximal variance in the underlying distribution eq. (4).

3.2 Maximal Stability

Maximal stability aims at finding the direction in which the largest gap between two classes of examples can be found. The class membership is not predetermined and can be adjusted to yield a bigger gap.

Maximal stability was originally used to achieve noise tolerant classifications in supervised learning with a threshold perceptron [13, 16]. An objective function associated with this criterion is

$$H = \sum_{\nu=1}^{p} \Theta\left(\kappa - |\, h_{\underline{J}}^{\nu} |\right) \tag{8}$$

and the maximal stability is the largest value of κ for which H can be made to zero.

The geometric interpretation of this strategy is to maximize the distance of the pattern closest to a separating plane through the origin.

The application of maximal stability learning seems to be reasonable in our case, because the probability density eq. (5) in the vincinity of the origin is minimal along the direction $\underline{B}$. Therefore we expect $\underline{B}$ to be the distinguished direction with respect to maximal stability.

Moreover the maximum stability strategy has proven to infer an unknown input–output relation very well from labelled examples [17]. This generalization ability in supervised learning need not translate to unsupervised learning, since the information a labelled example provides about the teacher is different from the information an unlabelled example contains. However, the effect of this difference is not trivial as we will show in the course of this paper.

The minimization of H eq. (8) requires an optimization with respect to the discrete class memberships. Therefore one would have to use a very time consuming simulated annealing procedure or – if satisfied with suboptimal stabilities – faster methods described in [16].

Nevertheless the results of both learning criteria can be studied analytically by interpreting each objective function as the energy $H(\underline{J})$ of an interacting system of N degrees of freedom. The corresponding partition function

$$Z = \left(\prod_{j=1}^{N} \int dJ_j\right) \quad \delta\left(\sum_{j=1}^{N}(J_j)^2 - N\right) \exp\left[-\beta H(\underline{J})\right] \tag{9}$$

will be studied in the thermodynamic limit $N \to \infty$, $p = \alpha N$. The limit $\beta \to \infty$ yields the groundstate and thus the minimum of the objective function [13]. Assuming that the free energy $F = -\frac{1}{\beta} \ln Z$ is self–averaging with respect to the distribution of inputs (4), the order parameter R can be obtained by means of a saddle–point integration using the replica–method [12]. We assumed replica–symmetry at the saddle–point. The calculation is outlined in the appendix.

4 Results

4.1 Maximal Variance

The saddle–point equations can be solved algebraically to yield the following dependence of R on the number of patterns presented, given the separation ρ and the orthogonal width γ of the peaks:

$$|\,R\,|\,(\alpha) = \begin{cases} \sqrt{\frac{\alpha\theta^2 - 1}{\alpha\theta^2 + \theta}} & \text{for} \quad \alpha \geq \frac{1}{\theta^2} \quad \text{and} \quad \theta > 0, \\ 0 & \text{for} \quad \alpha < \frac{1}{\theta^2} \quad \text{or} \quad \theta \leq 0, \end{cases} \tag{10}$$

where $\theta := \frac{1+\rho^2}{\gamma^2} - 1$ depends only on the relation between the variances $(1+\rho^2)$ along $\underline{B}$ and (γ^2) perpendicular to $\underline{B}$. Therefore θ measures how distinguished $\underline{B}$ is with respect to maximal variance.

Note, that the two solutions $\pm R$ are equivalent, since there is no reason to distinguish between the recognition of $\underline{B}$ and $-\underline{B}$.

For a specific value of ρ, fig. 1 shows $|\,R\,|$ vs. α for various γ according to eq. (10) together with the results of simulations. Finite–size–effects are rather drastic due to the "weak bias" $\rho/\sqrt{N}$ in the pattern distribution. A careful finite–size scaling confirms our result very well, assuming corrections of order $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$.

The vector $\underline{B}$ can only be recognized, if it coincides indeed with the direction of largest variance within the distribution $(\theta > 0)$. Even in this case, it remains unrevealed $(R = 0)$ unless a certain critical number of examples $\alpha_c = 1/\theta^2$ has been presented. The dependence of the critical number of examples on the parameters of the distribution is plotted in fig. 2. Above α_c, R will increase monotonously with α as the underlying structure becomes increasingly evident. In the limit $\alpha \to \infty$ perfect recognition is achieved as $|R| \sim 1 - \frac{1}{2\rho^2\alpha}$ approaches 1.

Note that if $\theta < 0$, the largest eigenvalue of $\mathbf{C}$ (sect. 3.1) for $\alpha \to \infty$ is $(N-1)$ – degenerate and corresponds to the entire subspace orthogonal to $\underline{B}$. Therefore the search for subsequent principal components would not be helpful unless all N eigenvectors were determined.

For fixed γ the typical number needed for successful learning scales like $1/\rho^2$, which coincides with a recent result [18] for supervised learning from similar data in the *large separation limit* $\rho \to \infty$. In this limit the additional information a teacher provides is redundant because the structure in the data is self–evident.

4.2 Maximal Stability

We obtain the replica–symmetric saddle–point equations

$$\begin{aligned} 1 - R^2 &= \alpha \int_{-\tilde{\kappa}}^{\tilde{\kappa}} \frac{dt}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(t - \tilde{\rho}R)^2}{2\sigma^2} \right] (|t| - \tilde{\kappa})^2 \qquad (11) \\ -2R &= \alpha \frac{\partial}{\partial R} \int_{-\tilde{\kappa}}^{\tilde{\kappa}} \frac{dt}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(t - \tilde{\rho}R)^2}{2\sigma^2} \right] (|t| - \tilde{\kappa})^2 \\ \text{where} &\qquad \tilde{\rho} := \rho/\gamma, \quad \tilde{\kappa} := \kappa/\gamma \quad \text{and} \quad \sigma^2 := 1 - \left(1 - \gamma^{-2}\right) R^2 \end{aligned}$$

which have to be solved numerically and yield the optimal stability and the overlap R for given α. The results for κ will not be discussed here, see [16].

Again $R = 0$ solves the saddle–point equations and $\pm R$ are equivalent. Depending on the parameters α, ρ and γ one or even two additional solutions $|R| \neq 0$ exist. The solution with the largest κ has to be interpreted as the actual result of the maximal stability criterion.

In fig. 3 the results for $|R|$ are plotted versus α for various widths γ and separations ρ. Again there exists a critical number of examples α_c, where monotonous improvement of recognition begins. Quite contrary to the preceding criterion the detection of $\underline{B}$ is not perfect, even in the limit $\alpha \to \infty$, where the set of examples is most likely to reflect the features of the underlying distribution. We obtain

$$|R|^{\alpha\to\infty} = \begin{cases} \sqrt{1 - \frac{1}{6(\gamma^2-1)}\left[z - 5 + \operatorname{sign}(1-\gamma)\sqrt{(z-5)^2 - 24} \right]} & \text{for} \quad \gamma \neq 1, \\ \sqrt{1 - \frac{2}{\rho^2}} & \text{for} \quad \gamma = 1, \end{cases} \qquad (12)$$

where $z := \rho^2/(1-\gamma^{-2})$. Obviously $\underline{B}$ itself is not the "best" vector with respect to stability. Only for an infinite separation $\rho \to \infty$ or extreme cigars ($\gamma = 0$) recognition is perfect.

Eq. (12) is only valid for parameters ρ and γ that allow recognition at all. Analoguous to $\theta > 0$ for the maximal variance criterion we find a corresponding minimal $\rho_c^{\alpha\to\infty}(\gamma)$ for the maximal stability strategy. For various α $\rho_c(\gamma)$ is shown for both criteria in fig. 4.

An interesting effect in $|R|$ (α) can be observed above $\gamma_s \simeq 1.35$, where the examples are so widely spread, that two locally maximal gaps can occur. Then the saddle–point equations have three solutions, corresponding to three extrema of $\kappa(|R|)$. One maximum at $R_0 = 0$, a minimum at $R_1 \neq 0$, and another maximum at $R_2 > R_1$. So the two maxima R_2 and $R_0 = 0$ compete for the highest stability. In that case recognition starts out with jumping from $R = 0$ to a finite R, when the critical number of examples is presented. This critical number where $|R| \neq 0$ becomes the global maximum of stability, i. e. the "physical" solution, will be denoted by α_c. We will not discuss the values of α for the appearance of a local maximum at $|R| \neq 0$ and for the $|R| \neq 0$ solution becoming the only maximum, see fig. 3.

Fig. 4 depicts the region in parameter space of ρ and γ, where a first order transition in understanding can be found. For given α it is the region between the closed and dashed line, where κ still has a local maximum at $R = 0$. Though this is not the physical solution, an algorithm like simulated annealing might be trapped in this local maximum.

Unfortunately no simple algorithm exists to achieve maximal stability for a given set of examples in the unsupervised case. Thus the above results cannot be explicitely confirmed by simulations. However simulations were performed using the socalled HopfTron–algorithm [16], which yields suboptimal yet high stabilities. This algorithm can be subdivided into two stages. The first one obtains a classification $\sigma^\nu = \text{sign}\left(\underline{J}_H \cdot \underline{\xi}^\nu\right)$ of fairly high stability, that can be defined through a vector of Hebbian form $\underline{J}_H = N^{-1} \sum_{\nu=1}^{p} \underline{\xi}^\nu \sigma^\nu$. In a second step this labelling is held fixed and the corresponding direction $\underline{J}_P$ of optimal stability by means of supervised learning [19 – 21]. This second step significantly increases the stability.

Depending on the distribution's parameters, the values for $|R_H|$ and $|R_P|$ respectively differ more or less from the calculated resluts of the maximal stability criterion, fig. 3. Nevertheless the increase of stability as achieved by the second stage of the algorithm, always moves $|R_P|$ in the direction of our maximal stability result. We suspect, however, that our analytic result concerning optimal stability is not exact and that replica–symmetry–breaking [12] must be considered.

4.3 Comparison

We found two important basic differences between the criteria of maximal variance and maximal stability.

Firstly, in the limit $\alpha \to \infty$, where the input examples represent the true structure of the data, maximal variance perfectly recognizes the vector $\underline{B}$, whenever it is indeed the direction of maximal width in the underlying distribution. Maximal stability on the other hand will – apart from two unnatural extreme shapes – not perfectly detect $\underline{B}$, no matter how many examples are presented. This is possible, because the underlying distribution exhibits no actual gap along $\underline{B}$, neither does any other direction: $\kappa(\alpha \to \infty) = 0$. Thus R is the result of two competing influences: the probability for having patterns close to the origin is low for large values of R, but on the other hand a smaller R corresponds to a larger subspace of directions that can be searched for the biggest gap.

Secondly, the two strategies require very different critical numbers of examples to initiate recognition, which certainly is most important in situations where only a limited number of examples is available. Fig. 2 gives a comparison between $\alpha_c(\rho, \gamma)$ for both criteria. For some regions in the parameter space of ρ, γ the underlying structure becomes noticable to the maximal stability criterion before it does for the maximal variance criterion and vice versa. The distribution being fairly flat ($\gamma \geq 1.35$) is a precondition for the maximal stability to have the possibility of being superior. However, we found maximal variance to be the better strategy in most cases.

For practical applications this indicates no necessity for the development of an algorithm that indeed achieves maximal stability (if not specifically interested in the stability itself). This was confirmed by all the simulations we performed with the HopfTron algorithm. HopfTron inferred the underlying structure better than (fig. 3a) or as good as (fig. 3b) our calculation predicted for the physical solution of the maximal stability criterion. Note that in fig. 3a even below the critical number of examples $\alpha_c(\rho, \gamma)$ the relevant direction was almost perfectly detected. For increasing flatness γ the HopfTron even outperformed the maximal variance strategy.

For a completely unknown structure it seems promising to use various strategies, e.g. maximal variance and the HopfTron algorithm, and compare the results.

5 Discussion

We have studied unsupervised learning based on the use of ad–hoc objective functions. As a simple specific example we have considered data drawn from two overlapping Gaussian clouds in N–space. Our results reflect some generic problems that can occur in unsupervised learning or clustering [5].

It is intuitively clear that the more examples are provided, the better the structure in the data can be detected. Our analysis shows that even a critical number can exist, below which successful learning is impossible. This behavior depends crucially on the chosen learning prescription and, of course, on how pronounced the structure of the data is.

As a rather drastic example we have shown that a fairly reasonable strategy, i.e. looking for the largest gap separating clusters of examples, might be of little use if the true input distribution reveals no such gaps. This doesn't indicate, however, that the stability–criterion will be inappropriate in general, it rather depends on the structure of the input distribution. For example the HopfTron algorithm [16], originating from the idea of gap search, proved to be very successful.

The maximal variance strategy seems to be a natural choice for the considered data. Yet it is not guaranteed that it will extract the interesting information in any case, as we have demonstrated for large γ.

A problem common to all types of learning, supervised as well as unsupervised, is that of a priori knowledge or assumptions on the complexity of the task. If we knew about the type of the underlying structure we could choose our learning strategy accordingly.

Recently, Watkin and Nadal [10] studied unsupervised learning based on the estimation of parameters in a model–distribution [5]. If it is assumed that the inputs are drawn from a distribution of the form (3) or its continuous equivalent, the model parameters $\underline{J}$ can be chosen in an optimal way, so as to maximize the expected value of R.

Watkin and Nadal point out, that knowledge about the discreteness of a vector $\underline{B} \in \{+1,-1\}^N$ leads to a much faster learning. For continuous $\underline{B}$ and spherical clusters ($\gamma = 1$), the optimal procedure turns out to be only slightly better than the search for maximal variance in continuous $\underline{J}$–space. In particular, the same minimal number of examples $\alpha_c N$ is needed for successful learning.

Of course, if a priori knowledge is available, fitting an appropriate model should be superior to any ad–hoc principle. On the other hand, problems will arise if the assumptions made on the inputs were wrong [10].

Strategies like the ones discussed in this paper can be useful if the knowledge is very poor. If, for example, the true distribution could be a mixture of three or more Gaussians as well, it is still likely that the maximal variance strategy would extract some relevant information from the examples.

As a last point we shortly discuss an information theoretic approach. Linskers Infomax–principle [22] suggests, in our language, to choose $\underline{J}$ such that the value of $h_{\underline{J}}$ contains on average as much information about the input as possible. This mutual information $I(\underline{J})$ is for such a deterministic input/output–relation simply the entropy of the output, not knowing the input [22]. Would such a more sophisticated criterion give "better" results than the simple maximum–variance principle?

Consider a vector $\underline{J}$ with a given value of R. The output entropy averaged over the distribution of inputs depends only on R:

$$I(R) = -\int_{-\infty}^{\infty} dh_{\underline{J}} P_R(h_{\underline{J}}) \ln \left[P_R(h_{\underline{J}}) \right] \tag{13}$$

A maximization of this quantity gives the overlap R_I for the direction of largest mutual information. This is to be compared with the value R corresponding to the largest $\left\langle h_{\underline{J}}^2 \right\rangle_R$ as it is achieved by the maximal variance strategy for $\alpha \to \infty$. Fig. 5 shows the numerical results for two values of ρ in dependence of γ.

Note that $R_I = 0$ above the same critical value $\gamma_c = \sqrt{1+\rho^2}$ as found in section 3 for maximal variance. This is due to the fact that for $R \approx 0$ the distribution $P_R(h_{\underline{J}})$ is approximately Gaussian and its entropy is given by

$$I(R) \approx \frac{1}{2}\left(1 + \ln \left[2\pi \left\langle h_{\underline{J}}^2 \right\rangle_R \right] \right) \tag{14}$$

which is a monotonous function of the variance.

Only in a certain range of subcritical γ, the resulting overlaps differ, indicating that some intermediate direction preserves more information on the inputs than $\underline{B}$ itself. This difference is particularly pronounced for large ρ where the distribution for nonzero R differs significantly from a Gaussian. For small γ maximal mutual information coincides with maximal width again. Details will be discussed elsewhere.

Forthcoming studies should incorporate more realistic situations, e.g. more complex structures with a higher number of clusters to detect [10]. In this context it might be interesting to study strategies similar to what is known as competetive learning [1]. The maximization of the mutual information [22] might as well be very useful in a more general situation and deserves further investigation.

Acknowledgement

We thank W. Kinzel for useful discussions and T.L.H. Watkin and J.-P. Nadal for communicating the results of [10] prior to publication. This work is part of the Ph.D. thesis of A. Mietzner. It is supported by the Volkswagenstiftung and the Deutsche Forschungsgemeinschaft. Simulations were performed on the Cray Y–MP–EL of the Rechenzentrum der Universität Würzburg.

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, Rev.Mod.Phys. **65**, 499 (1993).

[4] W. Kinzel and M. Opper, *Statistical Mechanics of Generalization*, preprint 1993.

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

[6] S. Becker, Int. J. of Neural Systems **2**, 17 (1991).

[7] M. Benaim, Europhys.Lett. **19**, 241 (1992).

[8] A. Prügel–Bennett and J.L. Shapiro, J.Phys. **A26**, 2343 (1993).

[9] J.–P. Nadal and N. Parga, *Dual Learning Machines: a bridge between Supervised and Unsupervised Learning*, preprint 1992

[10] T.L.H. Watkin and J.–P. Nadal, *Optimal Unsupervised Learning*, preprint 1993.

[11] M. Biehl and A. Mietzner, Europhys.Lett., in press

[12] M. Mezard, G. Parisi, and M.A. Virasoro, *Spin Glass Theory and beyond*, World Scientific (Singapore), 1987.

[13] E. Gardner, J.Phys. **A21**, 257 (1988).

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

[15] A. Krogh and J.A. Hertz, in *Parallel Processing in Neural Systems and Computers*, eds. R. Eckmiller, G. Hartmann, and G. Hauske, Elsevier (Amsterdam), pg. 183.

[16] A. Mietzner, W. Kinzel and M. Opper, *Optimal Stability in Unsupervised Learning*, preprint 1993.

[17] M. Opper, W. Kinzel, J. Kleinz, and R. Nehl, J.Phys. **A23**, L581 (1990).

[18] N. Barkai,H.S. Seung, and H. Sompolinsky, Phys.Rev.Lett. **70**, 3167 (1993).

[19] W. Krauth, M. Mezard, J.Phys. **A20**, L745 (1987).

[20] J.K. Anlauf and M. Biehl, Europhys.Lett. **10**, 687 (1990).

[21] P. Rujan, J. Phys. I (Paris) **3**, 277 (1993).

[22] R. Linsker, Computer (IEEE) **21**, 105 (1988).

[23] E. Gardner and B. Derrida, J.Phys. **A21**, 271 (1988).

[24] M. Griniasty and H. Gutfreund, J.Phys. **A24**, 715 (1991).

Appendix

We shall calculate $\langle \ln Z \rangle$ from (9) using the replica trick, where $\langle \ldots \rangle$ denotes the average over the inputs. The calculation resembles the ones of [13, 17, 23, 24] and is only sketched in the following.

For simplicity we consider the case $\gamma = 1$, i.e. $\xi_j^\nu = \widehat{\xi_j^\nu}$ from eq. (3). The replicated partition function reads

$$\begin{aligned} Z^n \;&=\; \left(\prod_{j,a}\int dJ_j^a\right)\left(\prod_{\nu,a}\int\frac{dh_a^\nu dx_a^\nu}{2\pi}\right)\left[\prod_{a=1}^{n}\delta\left(\sum_{j=1}^{N}J_j^{a\,2}-N\right)\right] \\ &\times\; \prod_{\nu,a}\exp\left[ix_a^\nu\left(h_a^\nu-\frac{1}{\sqrt{N}}\sum_j J_j^a\widehat{\xi_j^\nu}\right)\right]\prod_a\exp\left[-\beta\sum_\nu g(h_a^\nu)\right] \end{aligned} \tag{15}$$

Here a denotes the replica index. The energy–contribution of a single input example is in replicon a

$$g(h_a^\nu) = \begin{cases} -(h_a^\nu)^2 & \text{for the maximal variance strategy} \\ \Theta(\kappa - \mid h_a^\nu \mid) & \text{for optimal stability.} \end{cases} \tag{16}$$

We can now average over the inputs $\widehat{\xi_j^\nu}$ according to (3) and obtain

$$\begin{aligned}\langle Z^n\rangle &= \left(\prod_{a<b}\int dq_{ab}\right)\left(\prod_a\int dR^a\right)\left(\prod_a\int dJ^a\right)\exp\left[N\alpha\Phi\right]\\ &\times \prod_a\left\{\delta\left(NR^a-\sum_j J_j^a B_j\right)\delta\left(N-\sum_j(J_j^a)^2\right)\right\}\prod_{a<b}\delta\left(Nq_{ab}-\sum_j J_j^a J_j^b\right) \quad (17)\end{aligned}$$

where

$$\begin{aligned}\exp\left[\Phi\right] &= \left(\prod_a\int\frac{dx_a dh_a}{2\pi}\right)\exp\left[-\beta\sum_a g(h_a)\right]\\ &\times \frac{1}{2}\sum_{\sigma=\pm1}\exp\left[i\sum_a x_a h_a - i\rho\sigma\sum_a x_a R^a-\sum_{a<b}q_{ab}x_a x_b-\frac{1}{2}\sum_a x_a^2\right].\end{aligned}$$

In the replica–symmetric Ansatz $q_{ab}=q$, $R^a=R$ we get after performing the integrals over the J_j^a and x_a

$$\frac{\partial}{\partial n}\bigg|_{n=0}\frac{\langle Z^n\rangle}{N}=\beta\frac{1-R^2}{2v}+\alpha\sum_{\sigma=\pm1}\int\frac{dt\,e^{-t^2/2}}{\sqrt{2\pi}}\ln\left\{\int\frac{dh}{\sqrt{2\pi}}\exp\left[-\beta\left(g(h)+\frac{(h-t-\sigma\rho R)^2}{2v}\right)\right]\right\}$$

where $v=\beta(1-q)$ is assumed to be of order $\mathcal{O}(1)$ in the simultanuous limit $q\to1,\beta\to\infty$ [23]. This limit corresponds to forcing the system in a unique groundstate.

The integral over h is for $\beta\to\infty$ dominated by the maximal integrand and can be performed according to the choice of $g(h)$ [24], yielding the respective groundstate (free) energies F.

For $\gamma\neq1$ the calculation is done in complete analogy, averaging over the $\widehat{\xi_j^\nu}$, but with a modified δ–function defining

$$h_a^\nu=\gamma\left(\frac{1}{\sqrt{N}}\sum_j J_j^a\widehat{\xi_j^\nu}\right)+R(\gamma-1)\left(\frac{1}{\sqrt{N}}\sum_j B_j\widehat{\xi_j^\nu}\right) \qquad (18)$$

in equation (15). We get for the maximal variance strategy

$$\frac{-F}{N}=\frac{1-R^2}{2v}+\alpha\,\frac{\gamma^2+(1-\gamma^2+\rho^2)R^2}{1-2\gamma^2 v}, \qquad (19)$$

and for the learning with maximal stability we find

$$\frac{-F}{N}=\frac{1}{2v}\left\{(1-R^2)-\alpha\int_{-\tilde\kappa}^{\tilde\kappa}\frac{dt}{\sqrt{2\pi}S}\exp\left[-\frac{(t-\tilde\rho)^2}{2S^2}\right](|t|-\tilde\kappa)^2\right\}, \qquad (20)$$

with the abreviations

$$\tilde\rho=\frac{\rho}{\gamma},\quad \tilde\kappa=\frac{\kappa}{\gamma},\quad \text{and}\quad S^2=1-\left(1-\frac{1}{\gamma^2}\right)R^2. \qquad (21)$$

The analytic solution (10) and the saddle–point equations (11) follow from the condition $\frac{\partial F}{\partial v}=\frac{\partial F}{\partial R}=0$.

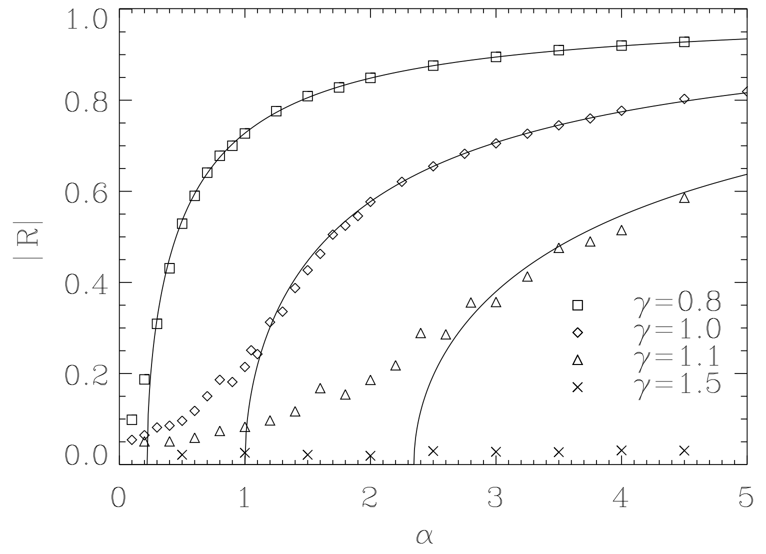

Figure 1: Absolute value of R vs. α for the maximal variance strategy. The curves are for a separation $\rho = 1$ and three different values of the orthogonal width. Simulations were performed with $N = 1000$ or $N = 500$ (for $\alpha > 2$) respectively. The results were averaged over 100 independent runs and standard error bars would be approximately the size of the symbols. A careful finite–size scaling confirms the analytical predictions very well, note that the typical value of $\mid R \mid$ of a randomly chosen $\underline{J}$ is of the order $1/\sqrt{N} \approx 0.03$ for $N = 1000$.

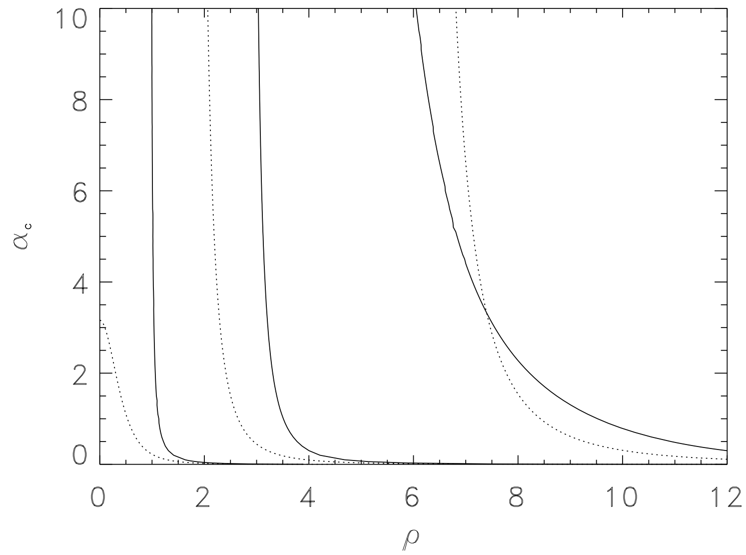

Figure 2: Critical number of examples α_c vs. the separation ρ of the peaks in the data to be learned. The dotted lines correspond to the maximal variance strategy, whereas the closed lines represent the maximal stability result. From left to right the widths γ were taken to be 0.8, 2.0, and 6.0, respectively.

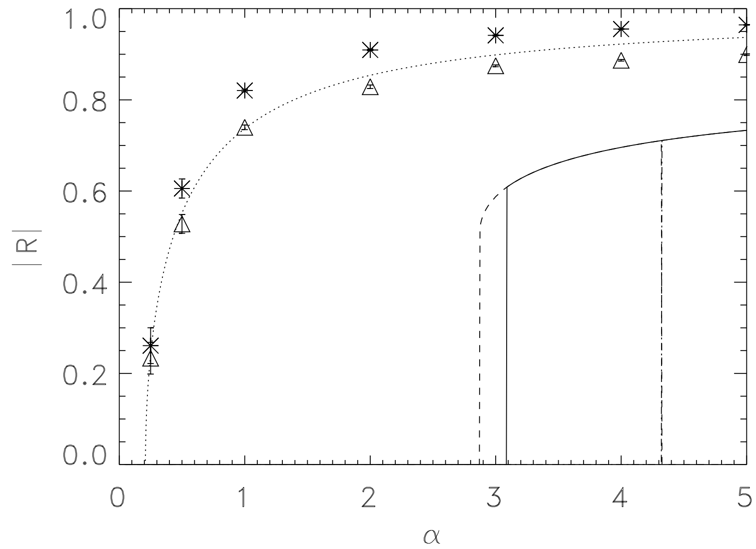

Figure 3a): $|R|$ vs. α for a separation of $\rho = 2.6$ and width $\gamma = 1.56$. The dotted line corresponds to the result of the maximal variance strategy. The closed line is the physical solution of the maximal stability criterion. The dashed line represents a local maximum of $\kappa(|R|)$, whereas the dash–dotted line marks the α, where the $R = 0$ solution becomes a minimum of $\kappa(|R|)$.
The corresponding results of the HopfTron algorithm are represented by stars (first step, low stability) and triangles (second step, high stability). Simulations were done for $N = 400$, the error bars depict the standard error for 20 to 25 runs.

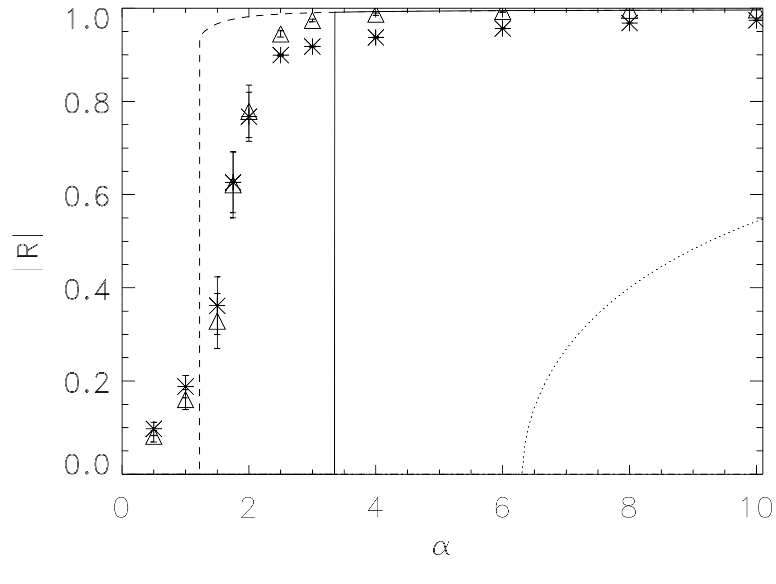

Figure 3b): Just as a), but with $\rho = 10$, $\gamma = 8.5$ and 10 runs for $\alpha \geq 4$.. For both parameter sets the maximal variance strategy approaches $|R| = 1$, whereas the maximal stability criterion tends to a value of $|R| < 1$ for $\alpha \to \infty$ as given by eq. 12. Note, that the performance of the HopfTron algorithm is comparable to whichever strategy is more successful.

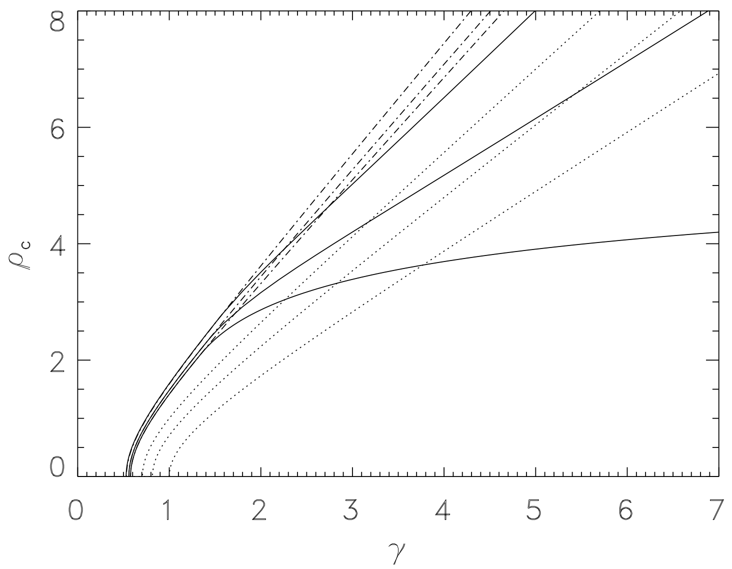

Figure 4: The minimally necessary separation ρ_c vs. width γ for various α. The dotted lines correspond to the maximal variance strategy, the closed lines to the physical solution of the maximal stability criterion and the dashed–dotted ones to the separations, where $R = 0$ is no maximum of $\kappa(|R|)$ anymore. In every family of curves the upper ones correspond to $\alpha = 1.0$, the ones in the middle to $\alpha = 4.0$ and the lowest ones to $\alpha \to \infty$. For values of γ, where the closed and the dashed–dotted curves are already split, recognition by the maximal stability criterion takes place as a first order transition. Only for fairly high γ and α there exist separations ρ, where the maximal stability criterion is able to start detecting the underlying structure, whereas maximal variance does not.

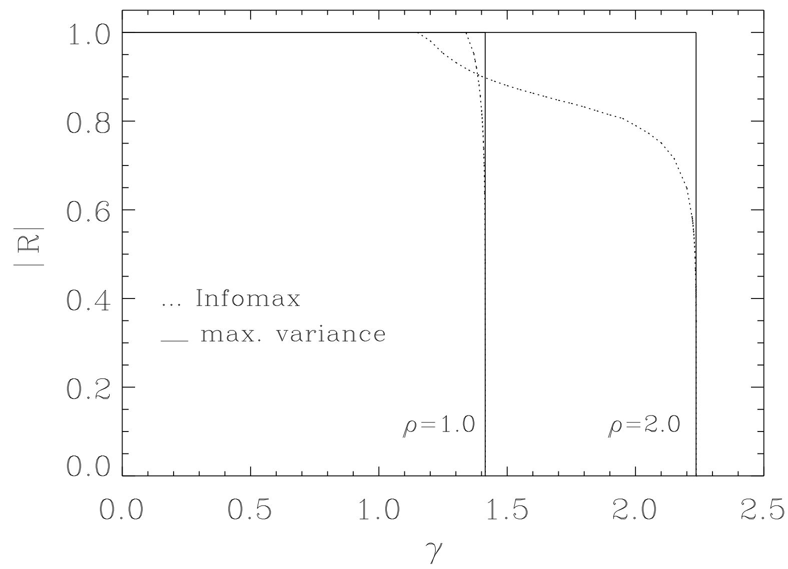

Figure 5: Comparison of the direction of largest information content (see sec. 5) and the direction of largest variance in the model data for two different separations ρ. For most values of γ the corresponding overlaps $|\ R\ |$ coincide, in particular R becomes zero at the same value $\gamma_c = \sqrt{1+\rho^2}$.