Supervised learning from clustered input examples

Carmela Marangi
Dipartimento di Fisica dell'Universita' di Bari
and I.N.F.N., Sez. di Bari
Via Orabona 4, 70126 Bari, Italy

Michael Biehl[1] and Sara A. Solla[2]
CONNECT, The Niels Bohr Institute
Blegdamsvej 17, 2100 Copenhagen Ø, Denmark

Abstract

In this paper we analyse the effect of introducing a structure in the input distribution on the generalization ability of a simple perceptron. The simple case of two clusters of input data and a linearly separable rule is considered. We find that the generalization ability improves with the separation ρ between the clusters, and is bounded from below by the result for the $\rho = 0$ unstructured case. The asymptotic behaviour for large training sets, however, is the same for structured and unstructured input distributions. For small training sets, the dependence of the generalization error on the number of examples is observed to be nonmonotonic for certain values of the model parameters.

The problem of learning [1, 2, 3, 4] in layered neural networks is usually addressed in the context of *supervised learning*. A set of examples for the unknown *target rule*, given in the form of input-output pairs, is used to adjust the internal parameters or *weights* of a *student network*. The aim is to obtain a specific network configuration which approximates the unknown input-output relation as well as possible. The process of learning through the adjustment of the weights is usually guided by the minimization of the *training error*, which measures the degree of

[1]permanent address: Inst. für theor. Physik der Universität, Am Hubland, D–97074 Würzburg
[2]on leave from AT&T Bell Laboratories, Holmdel, NJ 07733, USA

disagreement between student and target output for the training set. The success of learning is measured through the *generalization error*: the probability of disagreement between student and rule for an arbitrary input.

Within the framework of statistical physics, the problem of supervised learning is usually formulated as that of a specific architecture (the student) being trained through examples in order to reproduce an input-output relation which is defined through a specific configuration of the same architecture (the *teacher*). In most of the cases analysed to date, the distribution of data in input space is unstructured, i.e. the training inputs are vectors of binary or continous components drawn from a uniform or isotropic distribution.

In many practical situations, however, the input distribution will be structured. For instance, different *clusters* of data points might exist, reflecting certain features of the inputs. The detection of such a structure in input space is usually addressed in the context of *unsupervised learning*, with no reference to a functional relation to an output space [5, 6, 7].

In this work we study a question of crucial importance, namely the problem of incorporating nontrivial input distributions within the framework of supervised learning [8, 9]. Such a structure is expected to enhance the generalization ability whenever the rule output is to a large extent consistent with the clustering in input space. On the other hand, no improvement should arise when the classification to be learned is based on features completely different from the ones relevant to the clustering in input space.

In order to investigate this problem we consider a simple classification task, in which points $\boldsymbol{\xi}$ in N–space are assigned to two categories according to the state of a single output unit $\xi_0 = \pm 1$. The dichotomy corresponds to placing a hyperplane through the origin to separate the two classes, and it is implemented through a single layer perceptron [5]. Such linearly separable target rule is defined by a vector $\mathbf{B} \in I\!\!R^N$ which is perpendicular to the separating hyperplane. $\mathbf{B}$ can be interpreted as the weight vector of a *teacher perceptron* with output

$$\xi_0 = \mathrm{sign}\left(\frac{1}{\sqrt{N}}\mathbf{B}\cdot\boldsymbol{\xi}\right) \quad . \tag{1}$$

The student network is also a single layer perceptron, with couplings $\mathbf{J} \in I\!\!R^N$ chosen through the learning process. Both weight vectors are normalized: $\mid \mathbf{B} \mid^2 = \mid \mathbf{J} \mid^2 = N$.

A structure is imposed on the discrete N-dimensional input space through the choice of a specific vector $\mathbf{C} \in \{+1, -1\}^N$ and a separation ρ along this direction so that the inputs are distributed according to the discrete equivalent of two Gaussian clouds centered at $\pm\rho\mathbf{C}/\sqrt{N}$. The inputs are generated according to the following distribution [6]

$$P(\xi_i^\nu \mid \sigma^\nu) = \frac{1}{2}\left[(1+\rho/\sqrt{N})\delta(\xi_i^\nu - \sigma^\nu C_i) + (1-\rho/\sqrt{N})\delta(\xi_i^\nu + \sigma^\nu C_i)\right] \quad , \qquad (2)$$

where the dummy variable σ^ν determines the cluster, to which $\boldsymbol{\xi}^\nu$ belongs. We assume

$$P(\sigma^\nu) = \frac{1}{2}\left[\delta(\sigma^\nu - 1) + \delta(\sigma^\nu + 1)\right] \quad . \qquad (3)$$

According to the central limit theorem the resulting distribution of overlaps $\mathbf{C}\cdot\boldsymbol{\xi}^\nu/\sqrt{N}$ is a superposition of two Gaussians with mean values $\pm\rho$ and unit width. Note that in an arbitrary direction perpendicular to $\mathbf{C}$ the data appears structureless, as the corresponding distribution of overlaps is a single Gaussian with zero mean and unit variance. Note that the results reported here would also apply to a continous version [7] of distribution (2).

The object of our analysis is to investigate the generalization ability of the student as a function of the alignment $\eta = (\mathbf{B}\cdot\mathbf{C})/N$ between the teacher $\mathbf{B}$ and the vector $\mathbf{C}$, and the separation ρ between the centers of the input clusters.

By definition, the rule considered here is learnable for a student perceptron. The situation is different in a similar model recently studied by Meir [9], where the target outputs are defined by the labels σ^ν of the overlapping clusters (2,3). As the clusters overlap, the concept is unrealizable, and strong overtraining effects can occur. In this scenario, simple Hebbian learning outperforms strategies based on the minimization of the training error [9].

For the realizable rule considered here, however, the minimization of the training error is a natural training strategy. As an example of this approach we study the perceptron of optimal stability [10], i.e we analyse the properties of a student which maximizes the quantity

$$\kappa = \min_\nu \left\{ \frac{\xi_0^\nu}{\sqrt{N}} \sum_{i=1}^N J_i \xi_i^\nu \right\}_{\nu=1,\dots P} \qquad (4)$$

over the P examples in the training set. This strategy is known to yield almost optimal generalization in the case of isotropic, unstructured input distributions [11]. Furthermore, deterministic algorithms are available to obtain the weight vector of optimal stability for any linearly separable training set [13, 14, 15], allowing for comparisons between our analytic results and numerical simulations.

Following [10], we interpret the student weights J_j as the N interacting degrees of freedom in a system with energy

$$E = \sum_{\mu=1}^P \Theta\left(\kappa - \frac{\xi_0^\mu}{\sqrt{N}} \sum_{i=1}^N J_i \xi_i^\mu\right), \quad \text{with} \quad \Theta(x) = \begin{cases} 1 & \text{for } x > 0 \\ 0 & \text{else} \end{cases} \quad . \qquad (5)$$

The number of examples is an extensive quantity $P = \alpha N$. In the case of $\kappa = 0$ the *training energy* (5) reduces to the number of misclassified training inputs.

The calculation [10, 11] uses the replica trick [12] within the symmetric ansatz in order to perform the average over the quenched disorder (the set of input examples) and evaluate the thermodynamic quantities of interest. Apart from the order parameter q, representing the typical overlap of two error–free student vectors [10], we introduce two additional order parameters $R = \mathbf{J} \cdot \mathbf{B}/N$ and $D = \mathbf{J} \cdot \mathbf{C}/N$. The limit $\beta \to \infty$ forces the system into its groundstate, and the optimal stability is the maximal value of κ for which this groundstate still has zero training energy for a given α. The corresponding $\mathbf{J}$ is unique and is selected in the $q \to 1$ limit [10]. Together with R and D, the optimal κ is given by the following system of saddle point equations

$$1 - \frac{R^2 + D^2 - 2\eta R D}{1-\eta^2} = \alpha \iint_\Gamma Dt \frac{dh}{\sqrt{2\pi}} e^{-\frac{1}{2}(h-\rho\eta)^2} F^2(h,t) \tag{6}$$

$$2\frac{\eta D - R}{1-\eta^2} = \alpha \frac{\partial}{\partial R} \iint_\Gamma Dt \frac{dh}{\sqrt{2\pi}} e^{-\frac{1}{2}(h-\rho\eta)^2} F^2(h,t) \tag{7}$$

$$2\frac{\eta R - D}{1-\eta^2} = \alpha \frac{\partial}{\partial D} \iint_\Gamma Dt \frac{dh}{\sqrt{2\pi}} e^{-\frac{1}{2}(h-\rho\eta)^2} F^2(h,t) \quad , \tag{8}$$

where

$$F(h,t) = \kappa - t\sqrt{1-R^2} + \rho(\eta R - D)\mathrm{sign} h - |\, h \,|\, R \quad . \tag{9}$$

The integrals are over the two–dimensional domain Γ given by $F(h,t) > 0$, and we use the shorthand $Dt = dt\ \exp(-t^2/2)/\sqrt{2\pi}$.

The generalization error $\epsilon_g(\alpha)$ can be obtained as the average over the input distribution of the error function

$$\epsilon = \Theta\left[-(\mathbf{J} \cdot \boldsymbol{\xi}\,)(\mathbf{B} \cdot \boldsymbol{\xi}\,)\right] \quad . \tag{10}$$

The resulting ϵ_g depends only on the equilibrium values of the selfaveraging order parameters R and D as given for fixed α by the saddle point equations (6–8):

$$\epsilon_g(\alpha) = 1 - \sum_{\sigma=\pm 1} \frac{1}{2\pi\sqrt{1-R^2}} \int\limits_{-\sigma\rho D}^{\infty} \int\limits_{-\sigma\rho\eta}^{\infty} dx dy \exp\left[-\frac{1}{2(1-R^2)}(x^2 + y^2 - 2Rxy)\right] . \tag{11}$$

The generalization error as a function of α is shown for several values of ρ at fixed η in figure 1. The generalization error is observed to decrease monotonically with increasing ρ at fixed α. The $\rho = 0$ curve, corresponding to an unstructured input distribution [11], provides an upper bound to the generalization error computed here.

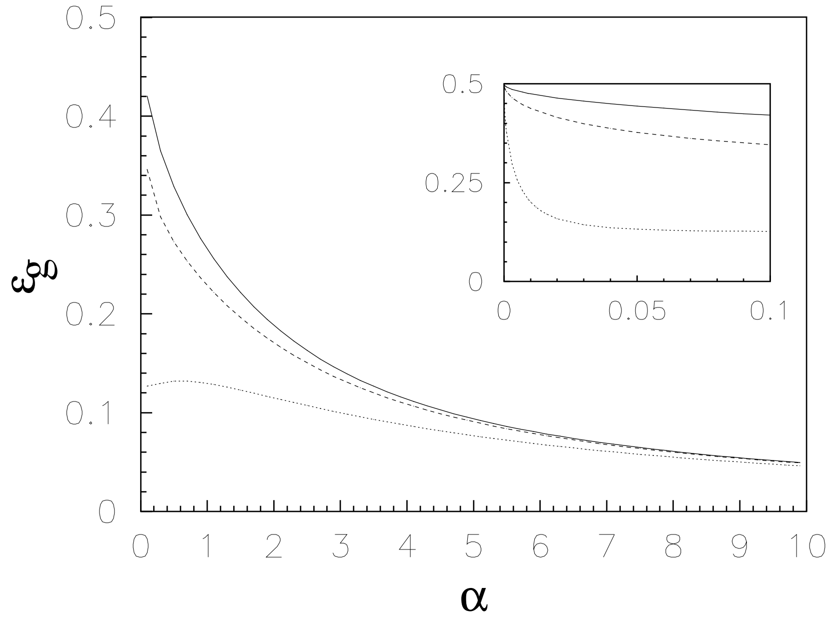

Figure 1: The generalization error versus α for $\eta = 0.3$ and $\rho = 0$ (solid), 2 (dashed), and 4 (dotted). Inset: ϵ_g for small values of α.

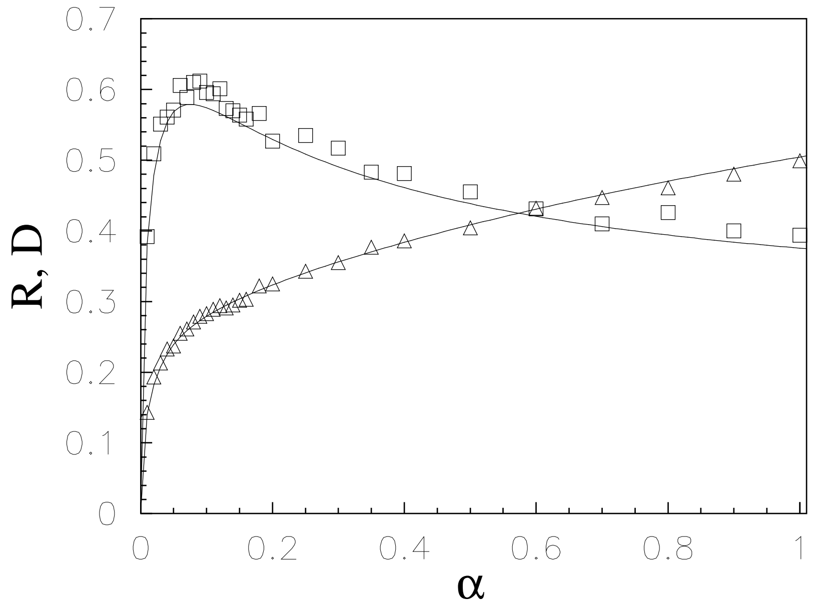

Figure 2: Order parameters D ($\square$) and R ($\triangle$) *vs.* α for $\rho = 5$ and $\eta = 0.3$. The solid lines correspond to the numerical solution of the saddle point equations. The data points represent the results of simulations for a system with $N = 500$ input units, averaged over 100 independent training sets. Standard error bars would be approximately the size of the symbols.

As shown in figure 1, improved performance at finite α arises even when there is only weak correlation between the labelling direction $\mathbf{B}$ and the direction $\mathbf{C}$ which characterizes the clustering of the input data. This effect is more pronounced as η increases from $\eta = 0$ ($\mathbf{B} \perp \mathbf{C}$) to $\eta = 1$ ($\mathbf{B} \parallel \mathbf{C}$) : at finite α, the generalization error decreases with increasing η for all ρ , and becomes independent of ρ at $\eta = 0$. The upper bound at $\eta = 0$ is thus again given by the $\rho = 0$ result [11].

The advantage of learning from structured data disappears asymptotically: in the $\alpha \to \infty$ limit the generalization error becomes independent of both ρ and η. In this limit $R \to 1$ and $D \to \eta$ so that $D \to \eta R$, and the generalization error exhibits the same $\epsilon_g(\alpha) \approx 0.50/\alpha$ decay as in the unstructured case [4]. Note the merging of the curves for different values of ρ with increasing α in figure 1.

An interesting behaviour appears at small α. As shown in figure 1, the generalization error can exhibit a nonmonotonic dependence on α. We have observed this effect for intermediate values of η and ρ. To identify the underlying mechanism, consider the α dependence of the order parameters R and D, shown in figure 2, where the numerical solution to the saddle point equations is shown to be in very good agreement with simulation results. While R increases monotonically with α, it does so at a slower rate than for $\rho = 0$. The small α behaviour is dominated by alignment with $\mathbf{C}$ instead of $\mathbf{B}$, for as long as the labelling of the examples by the rule output is consistent with the cluster labels σ. In this regime a separating hyperplane perpendicular to $\mathbf{C}$ tends to be selected by the maximal stability requirement, resulting in a fast increase of the order parameter D with α. As α increases, the sampling associated with a given cluster eventually crosses the hyperplane perpendicular to $\mathbf{B}$, and the cluster label is no longer consistent with the rule output. The data now indicates that $\mathbf{C}$ is not a valid hypothesis, and D decreases monotonically towards its $D = \eta R$ asymptotic value, while the student vector monotonically approaches the teacher $\mathbf{B}$ as $R \to 1$.

A nonmonotonic behaviour of the generalization error in the small α regime has recently been observed [4, 16] for a continuous perceptron. As the input distribution considered by Bös et al. is isotropic, the effect they observe cannot be explained in terms of competition between two order parameters.

The behaviour observed here for small α is unlikely to be an artifact of optimal stability training. We believe that the nonmonotonicity arising from the competition between $\mathbf{C}$ and $\mathbf{B}$ is an intrinsic property of the model. Alternative training strategies such as Hebbian learning [17], zero stability [1], and optimal learning [18, 19] are currently being considered.

Acknowledgements
The authors gratefully aknowledge discussions with G. Jung, B. Lautrup, M. Opper, H. Schwarze, and O. Winther.
Simulations were done on the Cray–YMP–EL of the Rechenzentrum der Universität Würzburg.
M. Biehl has been supported by the Deutsche Forschungsgemeinschaft.
C. Marangi has been partially supported by the Istituto Nazionale di Fisica Nucleare.
All three authors wish to express their appreciation for the hospitality extended by CONNECT and The Niels Bohr Institute while this work was being performed.

References

[1] Györgyi G. and Tishby N., in *Neural Networks and Spin Glasses*, edited by W.K. Theumann and R. Koeberle (World Scientific, Singapore) 1990.

[2] Seung H.S., Sompolinsky H. and Tishby N., *Phys. Rev. A*, **45** (1992) 6056.

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

[4] Opper M. and Kinzel W., *Statistical Mechanics of Generalization*, to appear in *Physics of Neural Networks III*, series eds. E. Domany, J.L. van Hemmen, and K. Schulten, to be published by Springer (Berlin)

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

[6] Biehl M. and Mietzner A., *J. Phys. A: Math. Gen.* , **27** (1994) 1885.

[7] Watkin T.L.H. and Nadal J.–P., *J. Phys. A: Math. Gen.*, **27** (1994) 1899.

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

[9] Meir R., *Empirical risk minimization versus maximal likelihood estimation: a case study*, *Neural Computation*, in press.

[10] Gardner E., *J. Phys. A:Math. Gen.*, **21** (1988) 257.

[11] Opper M., Kinzel W., Kleinz J., and Nehl R., *J. Phys. A:Math. Gen.*, **23** (1990) L581

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

[13] Krauth W. and Mezard M., *J. Phys. A:Math. Gen.*, **20** (1987) L745.

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

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

[16] Bös S., Kinzel W., and Opper M., *Phys. Rev. E*, **47** (1993) 1384.

[17] Vallet F., *Europhys. Lett.*, **9** (1989) 315

[18] Opper M. and Haussler D., *Phys. Rev. E*, **66** (1991) 2677.

[19] Watkin T.L.H., *Europhys. Lett.*, **21** (1993) 871.