ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 125-130

Supervised Learning in Committee Machines by PCA

C. Bunzmann, M. Biehl and R. Urbanczik
Universität Würzburg, Institut für Theoretische Physik

Am Hubland, 97074 Würzburg, Germany

Abstract. A learning algorithm for multilayer perceptrons is suggested which relates to the technique of principal component analysis. The latter is performed with respect to a correlation matrix computed from the example inputs and their target outputs. For large networks it is demonstrated that the procedure requires by far fewer examples for good generalization than traditional on–line training prescriptions.

1 Introduction

Artificial neural networks with complicated input-output relation are often assembled using simple structures which repeat over the network. Thus, the occurrence of symmetries is practically inevitable and — as demonstrated within a variety of learning paradigms — leads to badly generalizing plateau states [8, 9, 1]. These quasi–stationary configurations can be associated with the invariance of the network output under permutation of hidden units.

For large networks trained from randomized i.i.d. example inputs, a statistical physics analysis of on–line gradient descent has been done for various scenarios. These investigations show that good generalization ability is not achieved as long as the number of examples is linear in the number of adjustable parameters. The network is stuck in a plateau state, unless prior knowledge about the target rule is already built into the initial conditions. Note that also optimized training schedules or modified on–line algorithms as suggested in [11, 6, 10, 7] require an unrealistic initial non–trivial initialization.

The purpose of the work is mainly to demonstrate that these findings do not reflect a genuine difficulty in training multilayer networks, but just result from the use of inappropriate training schemes.

To this end we propose a novel approach to the supervised training of multilayer networks which relates to the well–known technique of principal component analysis. In contrast to the conventional use of PCA in the preprocessing of the data, the target outputs enter into the construction of the correlation matrix in our approach. Then the PCA allows us to reduce the effective dimensionality of the learning problem in a first stage of training. The necessary specialization can then be done in terms of a few parameters, the number of which does not increases with the dimensionality of the inputs. As a consequence, good generalization is achievable on the basis of a number of examples which is only linear in the number of free parameters.

In order to demonstrate the usefulness of the suggested training scheme, we investigate its typical properties in a statistical physics framework. Note that the failure of the stochastic gradient descent has been shown within the same idealized type of scenario.

2 The regression problem

In the following we consider a regression problem, in which an unknown rule has to be inferred from a set $\mathcal{P}$ of P example data (ξ^μ, τ^μ) where ξ^μ denote N dimensional input vectors. We assume that the corresponding rule output $\tau^\mu = \tau(B^T\xi)$ can be parameterized in terms of a Soft Committee machine, the *teacher*, as $\tau(B^T\xi) = \frac{1}{\sqrt{K}} \sum_{i=1}^{K} \mathrm{erf}\left(B_i^T\xi\right)$ with orthonormal weight vectors B_i.

The use of the error function as a transfer function is convenient for an analytical treatment of the system. However, this particular choice of the sigmoidal function should not be crucial for the results obtained in the following.

A more substantial restriction is that of assuming that all input components $\xi_j^\mu, j = 1, 2, \ldots N$ are independently drawn from a Gaussian distribution of zero mean and unit variance. We will limit the analysis to this particular case but discuss its significance and ways to relax the restriction in the conclusion.

A first and popular means of extracting information from the example set is Hebbian learning and a simple calculation yields that the expectation of the random variable $\tau(B^T\xi)\xi$ is parallel to $B_{\mathrm{av}} = \frac{1}{\sqrt{K}} \sum_{i=1}^{K} B_i$. Thus the Hebbian vector

$$v_{\mathrm{Hebb}}^{\mathcal{P}} = \frac{1}{P} \sum_{\mu=1}^{P} \tau^\mu \xi^\mu \tag{1}$$

can serve as an estimate of the direction in which the *average teacher vector* B_{av} points. The information obtained by the Hebb rule is obviously not sufficient to determine K different weight vectors $J = [J_1, J_2, \ldots, J_K]$ in a student network of matching complexity.

But one can try to obtain additional information from the higher order statistics of $\tau(B^T\xi)\xi$. A natural extension to Hebbian learning thus is to consider the correlation matrix $\frac{1}{P} \sum_{\mu=1}^{P} (\tau^\mu)^2 \xi^\mu \xi^{\mu\,T}$. More generally, we will use $C^{\mathcal{P}} = \frac{1}{P} \sum_{\mu=1}^{P} F(\tau^\mu) \xi^\mu \xi^{\mu\,T}$, where F is some function of the target output. Whenever numbers are given in the following, we do refer to the special case $F(\tau) = \tau^2$.

The usefulness of computing this correlation matrix is easily seen in the limit $P \to \infty$, which yields that the expectation of $C^{\mathcal{P}}$ has three different eigenvalues: One eigenvalue, $\lambda_0 = \frac{2}{\pi}\arcsin(2/3)$, has an $N-K$ dimensional eigenspace orthogonal to all teacher weight vectors. The eigenvector of the largest eigenvalue $\lambda_{\text{unspec}} = \lambda_0 + 8\left(K-1+1/\sqrt{5}\right)/(3K\pi)$ is the averaged teacher weight vector B_{av}. The eigenspace of the smallest eigenvalue $\lambda_{\text{spec}} = \lambda_0 - 8(1-1/\sqrt{5})/(3K\pi)$ is $(K-1)$ dimensional and spanned by the vectors $B_i - B_K$ $(i = 1, 2, \ldots, K-1)$. So the eigenspace of the smallest eigenvalue yields additional information about the teacher vectors.

For finite P we thus determine the eigenvectors $\Delta_1^{\mathcal{P}}, \ldots, \Delta_{K-1}^{\mathcal{P}}$ of $C^{\mathcal{P}}$ which have the smallest eigenvalues. Then, if P is sufficiently large, the linear space $V^{\mathcal{P}}$ spanned by $v_{\text{Hebb}}^{\mathcal{P}}$ and $\Delta_1^{\mathcal{P}}, \ldots, \Delta_{K-1}^{\mathcal{P}}$ will approximate the space V^∞ spanned by the teacher vectors $B_1, \ldots, B_K$.

In a second stage, we use an on-line training algorithm to identify the best approximation of the teacher weight vectors in the space $V^{\mathcal{P}}$. In the case we focus on, that K is much smaller than N, this two stage procedure achieves a drastic reduction in the dimensionality of the learning problem from KN to K^2 dimensions. It thus reduces, and in the limit $N \to \infty$ eliminates, the plateau problem in traditional on-line training.

2.1 Approximation of the teacher space by PCA

The first stage of our procedure obtains an approximation of V^∞ by applying the eigensystem analysis to the estimator $C^{\mathcal{P}}$. To investigate the quality of the approximation, we calculate the overlap r_{spec} of its $(K-1)$ dimensional part $\Delta^{\mathcal{P}} = [\Delta_1^{\mathcal{P}}, \ldots, \Delta_{K-1}^{\mathcal{P}}]$ with V^∞,

$$r_{\text{spec}} = \sqrt{\frac{\mathrm{Tr}\Delta^{\mathcal{P}^T} B B^T \Delta^{\mathcal{P}}}{K-1}}, \tag{2}$$

where B is the matrix with columns $B_1, \ldots, B_K$. Note that r_{spec} takes its maximal value 1 iff the vectors $\Delta_1^{\mathcal{P}}, \ldots, \Delta_{K-1}^{\mathcal{P}}$ lie in the space V^∞.

To analyze the procedure theoretically, we introduce the Gibbs distribution

$$P(J) = \frac{\exp\left(-\beta\, J^T C^{\mathcal{P}} J\right)}{Z(\mathcal{P})}, \qquad Z(\mathcal{P}) = \int \mathrm{d}J\, \exp(-\beta\, J^T C^{\mathcal{P}} J), \tag{3}$$

where the integral is over the unit sphere. In the limit $\beta \to \infty$, the distribution is dominated by the eigenvector $J^\star$ of $C^{\mathcal{P}}$ with the smallest eigenvalue. We denote the overlaps of $J^\star$ with the teacher vectors by $R^\star = B^T J^\star$.

In the limit $N \to \infty$, $P = \alpha K N$, using the replica trick a statistical physics calculation [5] yields the typical value of $R^\star$ as the solution of the following variational problem

$$R^\star(\alpha, K) = \arg\max_R \left(\min_\gamma \; R^T A(\alpha,\gamma) R + a(\alpha,\gamma) \right) \tag{4}$$

using $G(\tau(y)) = \frac{F(\tau(y))}{1+2\gamma F(\tau(y))}$, $A_{jk}(\alpha,\gamma) = -\alpha K \left\langle G(\tau(y))\,(y_j y_k - \delta_{jk})\right\rangle_y - \frac{\delta_{jk}}{2\gamma}$ and $a(\alpha,\gamma) = -\alpha K \left\langle G(\tau(y))\right\rangle_y + \frac{1}{2\gamma}$, where the angle brackets denote an average

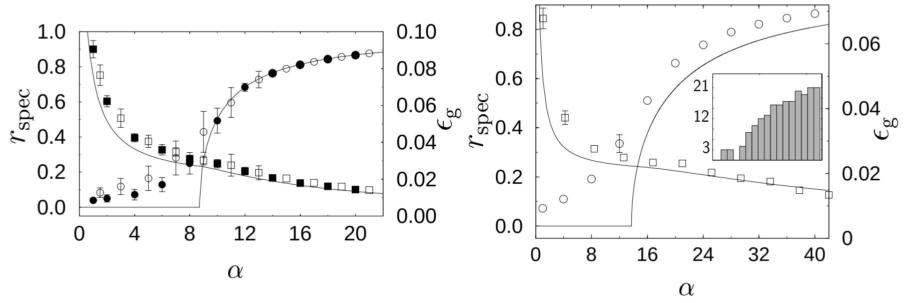

Figure 1: **left, K = 3**: Left axis: $r_{\rm spec}$ for $N = 400$ (○) and $N = 1600$ (●). Right axis: $\epsilon_{\rm g}$ found when the two stages of our procedure are combined, $N = 400$ (□) and $N = 1600$ (■), α refers to the number of examples in both training sets, $\alpha = (P + \hat{P})/KN$. Full lines: theoretical prediction for the thermodynamic limit $N \to \infty$. Where not shown, error bars are smaller than the symbol size. **right, K = 7**, using the optimized $C^{\mathcal{P}}$: $r_{\rm spec}$ (○) and $\epsilon_{\rm g}$ (□) for $N = 2000$. Full lines: Theoretical curves found in the large K limit. Inset: histogram of the 200 smallest eigenvalues of $C^{\mathcal{P}}$ for a single training set $\mathcal{P}$ with $P = 22KN$. A gap separates the 6 smallest eigenvalues from the rest of the spectrum. The range of the eigenvalues shown is $[-0.1, -0.07]$

over the K-dimensional random variable y with i.i.d. Gaussian components with zero mean and unit variance.

Since (4) is quadratic in R, the extremal problem has a solution with $R^\star \neq 0$ only if A is singular. From the symmetries one easily obtains, that A has just two eigenvalues. The first is $A_{11} - A_{12}$, its degeneracy is $K - 1$ and it is the relevant eigenvalue in our case. The degeneracy shows that the difference between the $K - 1$ smallest eigenvalues of $C^{\mathcal{P}}$ vanishes for large N. So in the thermodynamic limit, the simple procedure of analyzing the properties of the vector $J^\star$ minimizing $J^T C^{\mathcal{P}} J$, in fact, yields the properties of the $K - 1$ eigenvectors to the smallest eigenvalues of $C^{\mathcal{P}}$. Thus, the theoretical prediction of the quantity $r_{\rm spec}$ defined in (2) for an example set of size αKN is just the Euclidean norm of $R^\star(\alpha, K)$.

We find that the resulting $r_{\rm spec}(\alpha, K)$ is nonzero only for $\alpha > \alpha_{\rm c}(K)$. Above this critical value, Monte Carlo simulations show excellent agreement with the theoretical prediction. Fig.1 (left panel) shows as an example the results obtained for $K = 3$ where $\alpha_{\rm c}(3) = 8.70$.

In the limit of large K, but $K \ll N$, the variational problem can be simplified using the central limit theorem, see [3] for details of this calculation. The large K expansion yields that $\alpha_{\rm c}(K)$ diverges as

$$\alpha_{\rm c}(K) = \frac{27 \arcsin^2(2/3)}{16\left(1 - 1/\sqrt{5}\right)^2} K\,. \tag{5}$$

Also for large K, one can show that the optimal choice for F is not $F(\tau) = \tau^2$, but $F(\tau) = \tau^2 - \lambda_0^2$ (Fig.1,right). This optimal choice reduces $\alpha_{\rm c}$ to 2/3 of the value given in (5).

2.2 Choosing weight vectors

The results of the previous section show that for $\alpha > \alpha_\mathrm{c}(K)$ the eigenvectors $\Delta_i^\mathcal{P}$ of the $K-1$ smallest eigenvalues of $C^\mathcal{P}$ supplement the information obtained by the Hebbian vector (1). So in $V^\mathcal{P}$ (spanned by the $\Delta_i^\mathcal{P}$ and $v_\mathrm{Hebb}^\mathcal{P}$) a student network J can be found which generalizes well in terms of the quadratic error $\epsilon_\mathrm{g}(J) = \frac{1}{2}\left\langle \left(\tau(J^T\xi) - \tau(B^T\xi)\right)^2 \right\rangle_\xi$.

To actually find such a student network we set $J = \hat{B}\Gamma$, where the basis is $\hat{B} = [\Delta_1^\mathcal{P}, \ldots, \Delta_{K-1}^\mathcal{P}, v_\mathrm{Hebb}^\mathcal{P}]$ and optimize the $K{\times}K$ parameter matrix Γ by on-line gradient descent. For theoretical convenience, we use an independent training set $\hat{P}$ not contained in $\mathcal{P}$. So after the presentation of the ν-th example in $\hat{P}$ the matrix $\Gamma^{\nu+1}$ is $\Gamma^{\nu+1} = \Gamma^\nu + \eta\nabla_\Gamma \frac{1}{2}\left(\tau(\Gamma^{\nu T}\hat{B}^T\xi^\nu) - \tau(B^T\xi^\nu)\right)^2$. With increasing N, we can now scale the learning rate η and the number of additional examples $\hat{P}$ such that: $\eta \to 0$, $K^2 \ll \eta\hat{P}$ but $\hat{P} \ll P$. Then, in the large N limit, the on-line procedure performs gradient descent in ϵ_g (in the restricted space), reaches a minimum but uses a negligible number of additional examples.

So, for large N, the theoretical prediction $\epsilon_\mathrm{g,opt}(\alpha, K)$ for the generalization error of a student found by running the on-line procedure is given by the minimum of ϵ_g in $V^\mathcal{P}$. To calculate its value we need r_spec and the fact that the normalized overlap of the Hebbian vector $v_\mathrm{Hebb}^\mathcal{P}$ with the teacher average B_av is $r_\mathrm{unspec}(\alpha, K) = \left(1 + \frac{3\arcsin(2/3)}{2\alpha K}\right)^{-1/2}$. Then $\epsilon_\mathrm{g,opt}(\alpha, K)$ may be calculated using the explicit expression for $\epsilon_\mathrm{g}(J)$ given in [8].

In Fig.1, the generalization error obtained in our simulations is compared to $\epsilon_\mathrm{g,opt}$. The bend in the curve of the generalization error approximately at the critical value of α indicates that the on-line procedure leads to a specialization of the weight vectors for $\alpha > \alpha_\mathrm{c}(K)$.

3 Conclusion

We have shown that for fixed K committee machines can be efficiently learned from a number of examples which scales linearly in N if the input distribution is isotropic. This is in contrast to the findings in the case of stochastic gradient descent schemes for this scaling of the training set size. There, only a sub-optimally generalizing plateau state is reached, if the training set is sampled without replacement. (Despite recent theoretical efforts [4], the situation is still unclear for sampling with replacement or even batch learning.) Thus our findings indicate that the difficulties encountered in on-line learning result from the limited power of such algorithms and not from any intrinsic difficulties of the learning problem for these architectures. Note that our algorithm can be extended to classification tasks [3]. Further, since the eigenvalue spectrum of the correlation matrix $C^{\mathcal{P}}$ has a gap separating the $K-1$ lowest eigenvalues (inset of Fig.1), the number of hidden units can in fact be determined from $C^{\mathcal{P}}$. Thus our algorithm also contributes to the problem of model selection.

Of course, from a practical point of view, it is highly unrealistic to assume isotropic inputs. With regard to second order statistics this can be easily fixed by whitening the inputs. Interestingly we have found that whitening can improve the performance of our algorithm even when the inputs sample an isotropic Gaussian since the empirical distribution on a finite sample is by no means isotropic. However, given the NP-completeness of the loading problem already for very simple multilayer networks (e.g. [2]), one should not expect that our algorithm handles any input distribution well. It remains to be seen whether such malicious distributions play a major rôle in practical applications.

References

[1] M. Biehl, P. Riegler, and C. Wöhler. Transient dynamics of on-line learning in two-layered neural networks. *J. Phys. A*, 29:4769, 1996.

[2] A. Blum and R. Rivest. Training a 3-node neural network is NP-complete. In D. Haussler and L. Pitt, editors, *COLT 88*, pages 9–18. Morgan Kaufmann, 1988.

[3] C. Bunzmann, M. Biehl, and R. Urbanczik. Efficiently learning multilayer perceptrons. *Phys. Rev. Lett.*, 86:2166–2169, 2001.

[4] A.C.C. Coolen, D. Saad, and Y. Xiong. On-line learning from restricted trainings sets in multilayer neural networks. *Europhys. Lett.*, 51:691 – 697, 2000.

[5] A. Engel and C. Van den Broeck. *Statistical Mechanics of Learning.* Cambridge University Press, Cambridge, 2001.

[6] M. Rattray and D. Saad. Globally optimal on-line learning rules. In *NIPS 10.* MIT Press, 1997.

[7] M. Rattray, D. Saad, and S. Amari. Natural gradient descent for on-line learning. *Phys. Rev. Lett.*, 81:5461 –5464, 1998.

[8] D. Saad and S. Solla. Exact solution for on-line learning in multilayer neural networks. *Phys. Rev. Lett.*, 74:4337 – 4340, 1995.

[9] H. Schwarze and J. Hertz. Generalization in fully connected committee machines. *Europhys. Lett.*, 21:785 – 790, 1993.

[10] R. Vicente and N. Caticha. Functional optimization of online algorithms in multilayer neural networks. *J. Phys. A*, 30:L599 –L605, 1997.

[11] A.H.L. West and D. Saad. Adaptive back-propagation in on-line learning of multilayer networks. In *NIPS 8.* MIT Press, 1995.