ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 151-156

Numerical simulations of an optimal algorithm for supervised learning

A. Buhot, J.-M. Torres Moreno and M. B. Gordon *

Département de Recherche Fondamentale sur la Matière Condensée,
CEA/Grenoble, 17 rue des Martyrs, 38054 Grenoble Cedex 9,
France

Abstract. We study numerically an optimal learning algorithm for offline supervised noiseless training of a perceptron, that reaches bayesian generalization. We obtain the finite size corrections of the generalization error and its variance. The latter vanishes like 1/N, thus justifying the assumption of self-averaging done in analytical calculations. We also determined the stabilities distribution of the optimal student.

1. Introduction

Recently, the generalization properties of the optimal (bayesian) perceptron learning a linearly separable task have been predicted [1]. The optimal cost function, whose minimimum corresponds to the bayesian performance, has been also determined [2, 3] through a functional minimization of the generalization error within a statistical mechanics approach. In this paper, we present a numerical study of the offline (batch) learning of the bayesian perceptron using this optimal cost function. The extrapolation of our results to the thermodynamical limit verify the theoretical predictions and allowed us to determine the finite size scaling of the generalization error. A presentation of the cost function is given in section 2. A description of our implementation is described in section 3. Our simulation results are discussed in section 4.

2. Bayesian learning

In this paper we address the problem of learning pattern classification with a perceptron of N inputs, hereafter called student. The N dimensional student weight vector $\mathbf{J}$ has to be found knowing a training set $\mathcal{L}_\alpha$ of $P = \alpha N$ training patterns $\mathbf{S}^\mu$ ($\mu = 1, \ldots, P$) with their corresponding classes $\tau^\mu = \pm 1$. To ensure that the task is learnable *i.e.* linearly separable, the classes are given by

*and Centre National de la Recherche Scientifique.

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 151-156

a *teacher* perceptron of weight vector $\mathbf{B}$ through $\tau^\mu = \text{sign}(\mathbf{B} \cdot \mathbf{S}^\mu)$. As usual, the student's learning performance is measured through the generalization error

$$\epsilon_g = \arccos(R) \tag{1}$$

noindent where

$$R = \frac{\mathbf{J}}{\|\mathbf{J}\|} \cdot \frac{\mathbf{B}}{\|\mathbf{B}\|} \tag{2}$$

with $\|\mathbf{X}\| = \sqrt{\mathbf{X} \cdot \mathbf{X}}$. As τ^μ and the student's outputs $\sigma^\mu = \text{sign}(\mathbf{J} \cdot \mathbf{S}^\mu)$ are both invariant under the transformations $\mathbf{J} \to a\mathbf{J}$, $\mathbf{B} \to a'\mathbf{B}$ with $a, a' > 0$, the student's and the teacher's weight space may be restricted to the hyperspheres $\|\mathbf{J}\| = \|\mathbf{B}\| = \sqrt{N}$ without any loss of generality.

The student's weights are obtained by minimization of a cost that is an additive function of the patterns of the form

$$E(\mathbf{J}, \mathcal{L}_\alpha) = \sum_{\mu=1}^{P} V(\gamma^\mu) \tag{3}$$

where the potential $V(\gamma^\mu)$ depends on the training patterns through their stability: $\gamma^\mu = \tau^\mu \mathbf{J} \cdot \mathbf{S}^\mu / \sqrt{N}$ which is positive if the perceptron with weight vector $\mathbf{J}$ classifies correctly pattern μ. Students trained with different potentials have different properties that can be theoretically studied within the statistical mechanics approach, in the thermodynamic limit, *i.e.* $N \to +\infty$ with α constant. Opper and Haussler [1] determined the smallest generalization error $\epsilon_g^B(\alpha) = \arccos(R_B)$, *i.e.* the bayesian error, that may be reached by offline learning under the assumption that the components of the training patterns, S_i^μ $(i = 1, \cdots, N)$, are independent identically distributed random variables of zero mean $\langle S_i^\mu \rangle = 0$ and variance $\langle S_i^\mu S_j^\nu \rangle = \delta_{ij}\delta_{\mu\nu}$. However, their implementation of the bayesian generalizer needs that an infinite number of perceptrons be trained and compete through a vote. Several authors [4, 5, 6] proposed parametrized cost functions whose minimum endows the trained perceptron with generalization error very close to the bayesian one, provided that the corresponding parameter is adequately tuned. Watkin [8] showed that it exists a weight vector corresponding to bayesian performance within the version space, *i.e.* in the sub-space containing the weights that classify correctly the training set, but none of the proposed cost functions was able to find it.

Assuming the same pattern distribution as Opper and Haussler [1], it is possible to derive such optimal potential through a functional minimization of the generalization error. This potential [2, 3] depends implicitly on α through the parameter R (2), that must satisfy:

$$\frac{R^2}{\sqrt{1-R^2}} = \frac{\alpha}{\pi} \int_{-\infty}^{\infty} Dt \frac{\exp(-t^2 R^2/2)}{H(-Rt)} \tag{4}$$

where $Dt = dt \exp(-t^2/2)/\sqrt{2\pi}$ and $H(x) = \int_x^\infty Dt$. This is the same equation that was found by Opper and Haussler [1] for R_B, proving that the optimal

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 151-156

potential will endow the perceptron with bayesian performance. The equations that determine the optimal potential are:

$$\frac{dV}{d\gamma}(\gamma(t)) = -g(t), \tag{5}$$

$$g(t) \equiv \gamma(t) - t = T^2 \frac{d}{dt} \ln H(-\frac{t}{T}) \tag{6}$$

where $T = \sqrt{(1-R^2)/R^2}$. Equation (6) has to be inverted to find the expression of $V(\gamma)$ by integration of (5). It may be shown [2, 3] that $V(\gamma)$ is infinite for $\gamma < 0$ and a monotonic decreasing function of γ for $\gamma > 0$ that vanishes for $\gamma \to \infty$. The convexity of the potential ensures that the cost function has a unique minimum.

The distribution of stabilities $\rho(\gamma)$ of the training patterns may also be deduced within the statistical mechanics approach [7]. In the present case, where all the stabilities are positives, this distribution is nothing else than the distribution of distances of the training patterns to the separating hyperplane (orthogonal to $\mathbf{J}$). in the case of the optimal potential, we find the following analytical expression:

$$\rho(\gamma) = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{t^2(\gamma)}{2}\right) H\left(-\frac{t(\gamma)}{T}\right) \frac{dt}{d\gamma}(\gamma) \tag{7}$$

where $t(\gamma)$ is given by the inversion of (6), that depends implicitly on α through T. The details of the calculations will be published elsewere [3].

3. The algorithm

In practical implementations, as the optimal potential is infinite for negative stabilities, the initial weight vector for the minimization of the cost function (3) must lay inside the version space. As the training set is linearly separable, several algorithms, like the Perceptron learning algorithm, are available to find the initial weights. We used the algorithm Minimerror [9], whose potential $V(\gamma) = 1 - \tanh(\beta\gamma/2)$ was proposed by Gordon and Grempel [4]. The value of β that gives optimal generalization depends on α, but instead of using the theoretical value, Minimerror increases β upon learning, until a convergence criterium is meet [9]. Using the weights determined with Minimerror as starting point, the bayesian student is found by a simple gradient descent on cost function (3) with the optimal potential. At each step of the gradient descent, the new normalized weights $\mathbf{J}(k+1)$ are given by:

$$\mathbf{J}(k+1) = \sqrt{N} \frac{\mathbf{J}'}{\|\mathbf{J}'\|} \tag{8}$$

where

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 151-156

$$\mathbf{J}' = \mathbf{J}(k) - \epsilon(k)\,\delta\mathbf{J}(k) \quad (9)$$

$$\delta\mathbf{J}(k) = \sum_{\mu} V'(\gamma^{\mu}(k))\,\mathbf{S}^{\mu}\tau^{\mu} \quad (10)$$

with $\gamma^{\mu}(k) = \tau^{\mu}\mathbf{J}(k)\cdot\mathbf{S}^{\mu}/\sqrt{N}$. We control the convergence of the algorithm through the norm $||\delta\mathbf{J}_{\perp}(k)|| \equiv ||\delta\mathbf{J}(k) - \mathbf{J}(k)(\delta\mathbf{J}(k)\cdot\mathbf{J}(k))/N||$, which vanishes at the minimum of the cost function. Therefore, we do not need to integrate (5), as only the derivative $V'(\gamma) \equiv dV/d\gamma$ is needed in (10). This only requires the numerical inversion of (6), that has to be done for each value of α. The stopping condition in all our simulations was $||\delta\mathbf{J}_{\perp}(k)|| < 10^{-7}$. The variable learning rate $\epsilon(k)$ was determined at each step $k+1$ as follows: three different students $\mathbf{J}'$ were calculated, with learning rates $\epsilon(k)/2$, $\epsilon(k)$ and $5\epsilon(k)$. We kept for $\epsilon(k+1)$ the value corresponding to the smallest $||\delta\mathbf{J}_{\perp}(k)||$. This procedure helps to prevent the oscillations that may occur with too large learning rates and allows large steps in the regions where the potential is flat. We verified that the time needed for the three weights determinations is compensated because larger steps are allowed when possible. Perhaps more important, the choice of the initial value for the learning rate becomes non crucial, as the algorithm adjusts it automatically. Thus, we do not need to make any test before the different runs, a fact that results in a substantial gain of computer time.

4. Simulation results

We made numerical simulations for several values of α and N. For each N, one teacher vector $\mathbf{B}$ was picked at random from a uniform distribution within the N-dimensional centered hypercube of side 2, and $P = \alpha N$ training patterns, of components $S_i^{\mu} = \pm 1$, were randomly selected. The number of tests for each pair (P, N) ranged between 1000 to 40000, and was chosen large enough to determine the generalization error ϵ_g within $\sim 0.1\%$. We performed most of our tests on a parallel machine that allowed for 64 training sets to be processed simultaneously.

The generalization error is represented as a function of $1/N$ on figure 1 (a). For each value of α we considered values of N large enough, that second order corrections in $1/N$ are negligeable. The lines correspond to linear fits, whose extrapolations to the origin are in agreement with the theoretical prediction for ϵ_g, valid in the limit $N \to \infty$. The finite size corrections are negative and proportionnal to $1/N$ for large N. This behaviour is not very surprising as the information carried by the training set at given α is larger the smaller N.

The variance $\sigma_g^2 = \langle(\epsilon_g - \langle\epsilon_g\rangle)^2\rangle$, plotted as a function of $1/N$ on figure 1 (b), vanishes $1/N \to 0$. This result confirms that the hypothesis of self-averaging, underlying all the theoretical calculations, is correct: the variance of the generalization error tends to zero in the thermodynamical limit for all the α studied.

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 151-156

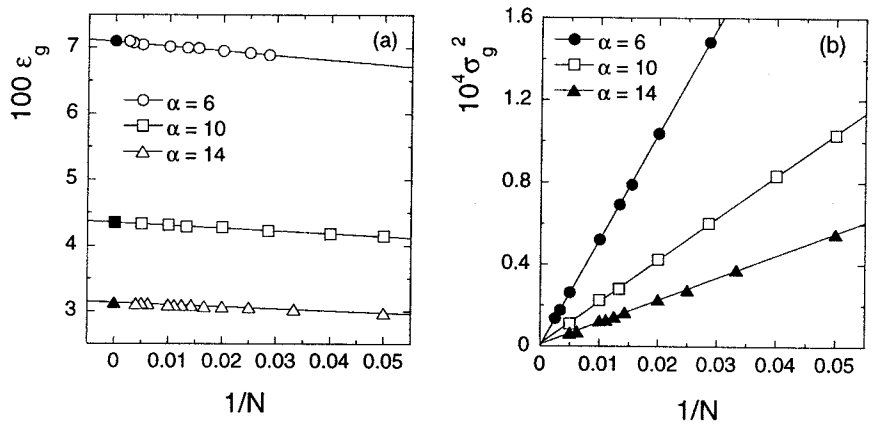

Figure 1: (a) Generalization error vs. $1/N$ for $\alpha = 6, 10$ and 14. Open symbols are numerical simulation results, full symbols are the theoretical predictions. Error bars are not visible at the scale of the figure. (b) Variance of ϵ_g vs. $1/N$ for $\alpha = 6, 10$ and 14.

The distribution of stabilities of the optimal student has been determined from our simulation results. The results corresponding to $\alpha = 6$ are plotted on figure 2 for two values of N, together with the theoretical prediction for $N \to +\infty$ (7). For comparison, we included on the same figure the teacher's distribution $\rho_t(\gamma)$, which in the thermodynamic limit has the following simple analytical expression:

$$\rho_t(\gamma) = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\gamma^2}{2}\right). \tag{11}$$

The teacher's and the student's distributions agree far from the separating hyperplane. The bayesian student has a maximum of patterns at a finite distance to the separating hyperplane, but contrary to the Maximal Stability Perceptron [10], there is a finite fraction of training patterns closer to the hyperplane.

5. Conclusion

In this paper we presented numerical simulations of the optimal (bayesian) perceptron learning a linearly separable task from examples. They confirm the theoretical predictions in the thermodynamical limit $N \to \infty$. We show that for finite N, the mean value of the generalization performance is *better* than the theoretical predictions, which should thus be considered as a theoretical *upper* bound for the average generalization error.

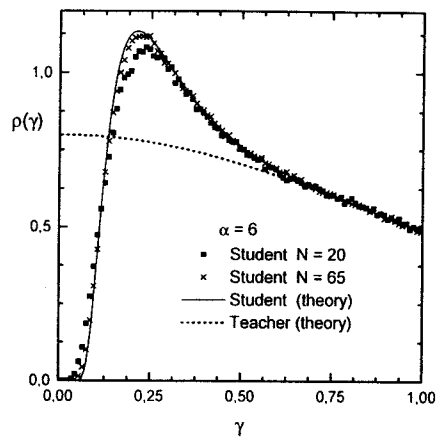

Figure 2: Distributions of stabilities $\rho(\gamma)$ for $\alpha = 6$.

References

[1] M. Opper and D. Haussler, *Phys. Rev. Lett.*, **66**, 2677-2680 (1991).

[2] O. Kinouchi and N. Caticha, *Phys. Rev. E*, **54**, R54-R57 (1996).

[3] A. Buhot and M. B. Gordon (to be published).

[4] M. B. Gordon and D. Grempel, *Europhys. Lett.*, **29**, 257-262 (1995).

[5] R. Meir and J.F. Fontanari, *Phys. Rev. A*, **45**, 8874-8884 (1992).

[6] M. Bouten, J. Schietse and C. Van den Broeck, *Phys. Rev. E*, **52**, 1958-1967 (1995).

[7] M. Griniasty and H. Gutfreund, *J. Phys. A: Math. Gen.*, **24**, 715-734 (1991).

[8] T. L. H. Watkin, *Europhys. Lett.*, **21**, (8), 871-876 (1993).

[9] B. Raffin and M. B. Gordon, *Neural Computation*, **7**, 1206-1224 (1995).

[10] W. Krauth and M. Mézard, *J. Phys. A: Math. Gen.*, **20**, L745-L752 (1987).