Solving the problem of overfitting of the pseudo-inverse solution for classification learning.

F. VALLET, J.-G. CAILTON, Ph. REFREGIER
Laboratoire Central de Recherches
Thomson-CSF 91401 ORSAY (cedex) France

Abstract

We investigate the pseudo-inverse solution for the learning of a binary classification. We address the problem of overfitting of this solution, i.e. the fact that the generalization rate can be relatively low although the learning rate is very high. We interpret this phenomenon with respect to the standard deviation of the eigenvalues of the covariance matrix of the learned patterns. We propose two ways for solving this problem: the first one is linear, the second one is a two-layer perceptron. Numerical simulations are given to illustrate these approaches.

1 Introduction

The general question addressed here is the learning of a classification $\mathcal{C}$ defined on the space of dimension N: $\Re^N$.

$$\begin{aligned} \mathcal{C} &: \Re^N \longrightarrow \{-1,+1\} \\ \mathcal{C} &: |x> \longrightarrow \mathcal{C}|x> = d = \pm 1, \end{aligned} \quad (1)$$

($|x>$ is a vector of $\Re^N$, $< x|$ its transposed vector). We try to approximate the real classification $\mathcal{C}$ by an approximated version $\hat{\mathcal{C}}$ which belongs to a given class of functions, and is derived from a set of patterns $|x^\mu>$ (the learning set L, composed of P $= \alpha$N patterns) with their corresponding output bits d^μ. Hereafter we define two recognition rates:

- the learning rate : the probability for a pattern $|x^\mu>$ which belongs to L to be classified well, i.e. to verify $\mathcal{C}|x^\mu> = \hat{\mathcal{C}}|x^\mu>$;
- the generalization rate : the probability for an arbitrary pattern taken anywhere in the whole available space, to be classified well.

The general problem is to find a classification $\hat{\mathcal{C}}$ which is a best approximation of $\mathcal{C}$, that is to say for which the generalization rate is the highest. In fact, what we can do easily is define a cost function on the learning set, the minimization of which gives a high learning rate. The overfitting problem is then the consequence of the non-equivalence between the maximization of two criteria (within the frame of a given model of $\hat{\mathcal{C}}$): the learning rate on the one hand, and the generalization rate on the other hand. It can be interesting to have a lower learning rate to generalize better.

Many classes of classifiers exist, as for example the supervised multi-layer perceptron [5,15,12,13], self-organizing networks [10] and clustering algorithms... We are mainly interested here in the perceptron-type class of classifiers $\hat{\mathcal{C}}$. For this class, the output bit is defined by:

$$\hat{P}_W|X>= sgn(<W|X>), \qquad (2)$$

where $|W\ >$ is the weight vector of the perceptron. General capacity performances can also be estimated, independently of the question of how to reach effectively the optimum form [6,7,8]. Perceptron solutions for classifiers have been widely exhibited in the literature. Iterative algorithms can provide a solution [4,11], as for example the famous perceptron algorithm when there exist a solution [2,14]. Direct solutions can also be exhibited, as for example the pseudo-inverse solution [10] (denoted hereafter PIS), or the Hebb solution, which is widely used within the field of associative memories [9,1,17].

We will recall the expression of the Hebb $|h>$ and the pseudo-inverse $|PI>$ solutions, and decompose this last one into a form which will allow us to interpret the overfitting on the basis of statistical criteria. We will examine the case of small values of α (of order 1), that is to say that the number of learned pattern is of the same order than the dimension of the space. If for example the patterns to be classified are noise spectra in the range 1Hz - 10kHz, with a precision of 0.1 %, the dimension of the space is of order Log(10 000)/Log(1.001) ~ 10 000; it is possible not to have this many patterns in the learning set.

The illustrative simulations refer to the majority detection problem (MDP): we choose for the boolean function $\mathcal{C}$ the linearly separable classification defined by the weight vector $|C>=(1,1,....1)$, and the input space is restricted to the hypercube of dimension N : $\{-1,1\}^N$, where all of 2^N patterns have the same probability to occur (both for learning and generalization).

2 Hebb solution and pseudo-inverse solution (PIS).

The Hebb solution is given by:

$$|h>\equiv \frac{1}{P}\sum_{\mu=1}^{P} d^\mu|x^\mu> \qquad (3)$$

In the case of MDP (majority detection problem), if the components of the patterns are randomly and independently taken in {+1,-1}, with equal probability, the learning and generalization rates can be calculated exactly versus α, in the $N\ \rightarrow\ \infty$ limit [16]:

$$L(\alpha)=\frac{1}{\sqrt{\pi}}\int_0^\infty e^{-u^2}erfc(-u\sqrt{\frac{2\alpha}{\pi}}-\frac{1}{\sqrt{2\alpha}})du$$

$$G(\alpha)=1-\frac{1}{\pi}Arctg(\sqrt{\frac{\pi}{2\alpha}}) \qquad (5)$$

($erfc(x)=\frac{2}{\sqrt{\pi}}\int_x^\infty e^{-u^2}du$ is the complementary error function).

The Hebb solution tends toward the good one when $\alpha\rightarrow\infty$. We have plotted these rates on figure 1.

The pseudo-inverse solution $|PI\ >$ is the vector $|W\ \ >$ which minimizes a quadratic cost:

$$C=\sum_{\mu=1}^{P}(<W|x^\mu>-d^\mu)^2. \qquad (6)$$

To find the solution, we differentiate equation (5) and write dC = 0, that is to say:

$$(d<W|).\sum_{\mu=1}^{P}|x^\mu>(<x^\mu|W>-d^\mu)=0, \qquad (7)$$

i.e.

$$H|W>=|h>, \qquad (8)$$

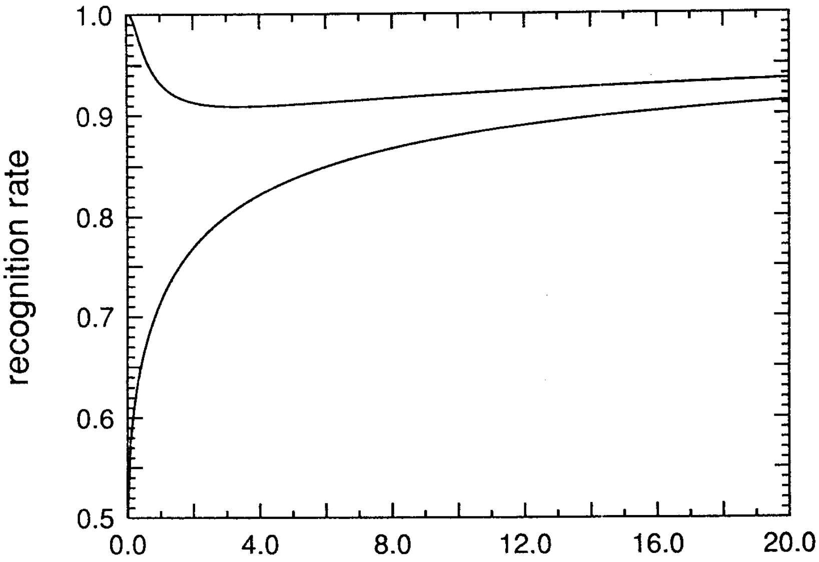

figure 1: Learning rate (upper curve) and generalization rate (lower curve), versus $\alpha = P/N$, with the Hebb rule, for the majority detection problem (MDP): the classification defined on the hypercube is linearly separable, and the corresponding weight vector is $|C>=(1,1,....1)$. (This is an analytical result).

where H is the Hebb matrix (the covariance matrix when the $|x^{\mu}>$'s are centered):

$$H = \frac{1}{P}\sum_{\mu} |x^{\mu}><x^{\mu}|. \qquad (9)$$

As H is a real symmetric matrix, it is diagonalizable (the $|\alpha>$'s are the orthonormalized eigen-vectors of H):

$$H = \sum_{\alpha=1}^{M} \lambda_{\alpha}|\alpha><\alpha|, \qquad (10)$$

where we keep only the non-zero eigenvalues, so that M is the dimension of the space spanned by the patterns of the learning set (we have in general $M = inf(P,N)$). As $|h>$ belongs by definition to this space, equation (7) is invertible, and we obtain:

$$|PI>= H^{-1}|h>, \qquad (11)$$

where we define $H^{-1} = \sum_{\alpha=1}^{M} \lambda_{\alpha}^{-1}|\alpha><\alpha|$ (so that $H^{-1}H = HH^{-1}$ is the orthogonal projection on the learning space).

3 Overfitting of the PIS.

Concerning the MDP, the simulations for $\alpha = 1$ give for the Hebb solution and PIS, learning rates of 93% and 100% respectively, while the generalization rates are 72% and 56%. We clearly face a problem of overfitting for the PIS. We now interpret this result by considering carefully the eigenvalues of H. To do this, we define new variables x_{α} which are the normalized projections of $|x>$ on the α^{th} principal axes of H:

$$|x> \longrightarrow x_{\alpha} = \frac{1}{\sqrt{\lambda_{\alpha}}} <x|\alpha>, \qquad (12)$$

(we define h_{α} in the same way). With this notation, the PIS gives:

$$<PI|x>= \sum_{\alpha=1}^{M} h_{\alpha}x_{\alpha} \qquad (13)$$

It is straightforward to show that the mean value of x_{α}^2 (the standar deviation of the x_{α} when centered) over the learning

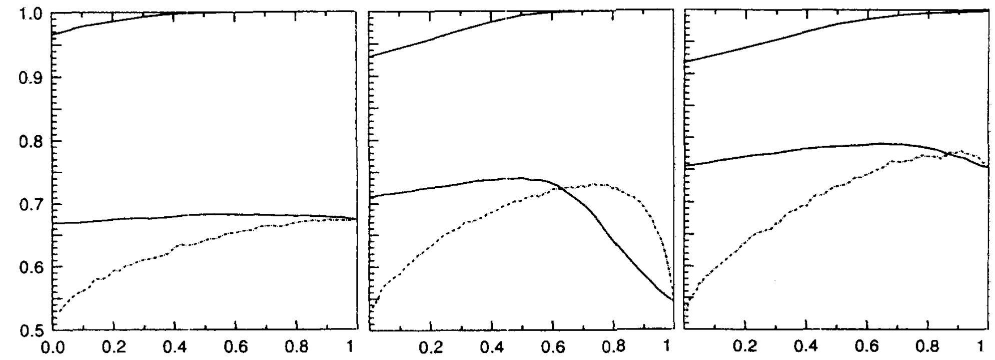

figure 2: *Continued lines*: learning rate (upper curve) and generalization rate (lower curve) for the MDP, versus k, for the linear extension of the PIS : $|k>=H^{-k}|h>$.
Dashed lines: generalization rate versus the proportion of the largest eigenvalues kept in the compression-like form.
Three different values of α have been plotted: a) $\alpha = 0.5$, b) $\alpha = 1$, c) $\alpha = 1.5$. The simulations are done in dimension N=100, and the results are averaged over 100 draws of different learning sets. k=0 is the Hebb solution, k=1 is the PIS.

set is equal to 1 for each value of α. On the other hand, the same mean value calculated over all of the 2^N possible patterns (the ones we take into account for the generalization rate) is equal to $1/\sqrt{\lambda_\alpha}$. We then see clearly that for an arbitrary test pattern $|x>$, the values x_α in (12) corresponding to small eigenvalues can explode and give rise to an aberrant contribution (over the learning set, these components are by definition perfectly suited, but they may not be relevant for the general classification). We think that the overfitting problem comes mainly from this fact. We propose now several ways to avoid such a problem.

4 Linear extension of the PIS.

The basic idea is to lower the importance of the α terms in (12) which correspond to small values of λ_α, by multiplying each term by a scalar which grows when λ_α gets larger:

$$\sum_{\alpha=1}^{M} h_\alpha x_\alpha \longrightarrow \sum_{\alpha=1}^{M} h_\alpha x_\alpha f(\lambda_\alpha), \qquad (14)$$

where f is an increasing function. We could take for f a step function: we eliminate in this case all the terms in (12) which correspond to eigenvalues λ_α which are below a given threshold (fig. 2). This solution is a known compression technique [3]. We can also use a smoother function f:

$$f(\lambda) = \lambda^{1-k}, \qquad (15)$$

where k is a real value. This is equivalent to consider a perceptron-type solution for which the weight vector is:

$$|k>=H^{-k}|h> . \qquad (16)$$

For k=1 we obtain the PIS, for k=0 it is the Hebb solution. We show on figure 2 the results obtained for the learning and generalization rates versus k, for three values of α (still for the MDP). We see clearly the phenomenon of overfitting for the PIS.

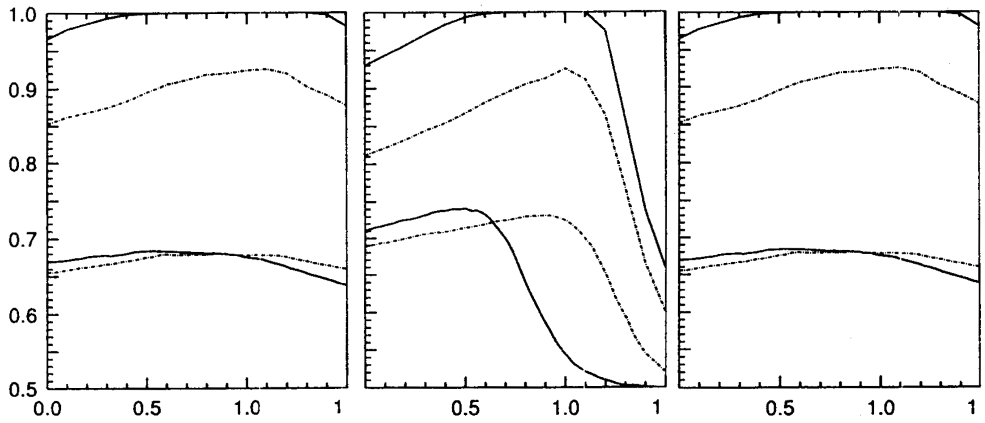

figure 3: Learning rate (upper curve) and generalization rate (lower curve) for the MDP, versus k, of the linear extension of the PIS (continued curves) and its saturated two-layer version (dashed curve): $|k >\underset{T}{=} H^{k-1}|h >$. Three different values of α have been plotted: a) $\alpha = 0.5$, b) $\alpha = 1$, c) $\alpha = 1.5$. (N=100, 100 draws).

5 Non-linear extension of the PIS: a two-layer perceptron.

Another way to avoid the explosion of certain terms in (12) is to saturate the components x_α to a value of the same order as their root mean square defined on the learning set, that is to say 1. This is a simple way to project the test vector $|x >$ into the statistical region defined by the learning set. Doing this is equivalent to introducing a hidden layer composed of the M new parameters x_α, to which is applied a non-linear function σ(x) which saturate when $|x|$ is greater than 1 (hyperbolic tangent for example). This two-stage solution can be written as follows, when combined with the preceding linear solution (parameterized by k):

$$|x > \longrightarrow sgn\left(\sum_{\alpha=1}^{M} \lambda_\alpha^{1-k} h_\alpha \sigma(x_\alpha)\right). \quad (17)$$

We give on figure 3 the curves of the simulation corresponding to this solution: the overfitting is clearly avoided, and the optimum generalization rate appears to come back to k=1, that is to say to the saturated version of the PIS (in these simulations, the σ function is a simple saturation: $\sigma(\mathrm{x}) = \mathrm{x}$ if $|x| < 1$, $\sigma(\mathrm{x}) = \mathrm{sgn(x)}$ otherwise).

6 Overfitting versus α.

A very strange observation is that, in the case of PIS, the generalization rate is not a monotonic increasing function, but shows a maximum, decreases and then increases again (see figure 4). This can be interpreted partly by the fact that the relative number of small eigenvalues responsible for the explosion of terms in (12) first increases with α (when $\alpha < 1$), and then decreases. To illustrate this fact, we have plotted on the same figure (fig 4), the generalization rate, the minimum value λ_{min} and the standard deviation of the eigenvalues of H which reflects of course the relative number of small eigenvalues (still for the majority detection problem). To see clearly the result, the eigenvalues have been normalized so that their mean value

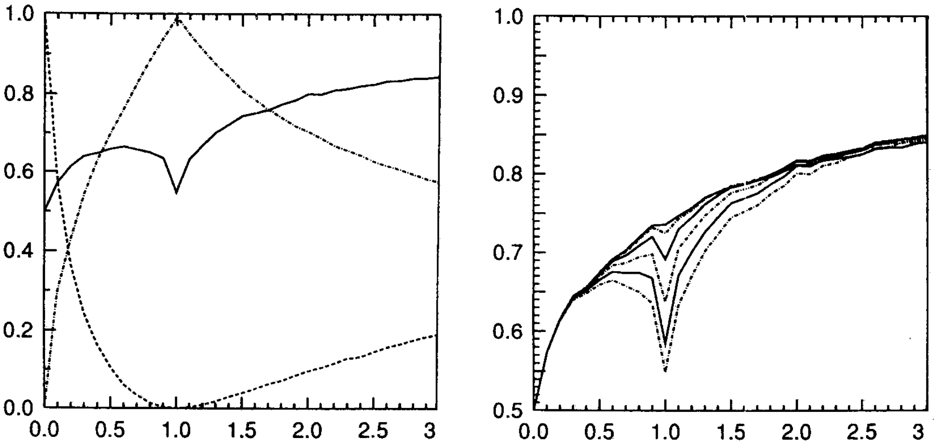

figure 4: a) Generalization rate for k=1 (continued curves), minimum value λ_{min} (dotted curves) and standard deviation (dashed curves) of the normalized eigenvalues of H, versus α. The classification to learn is the MDP, (N=101, 46 draws).

b) Generalization rate for several values of k (.5, .6, .7, .8, .9, 1), versus α, for the MDP (N=101, 46 draws). The curves with deepening valleys correspond to growing values of k.

is 1, and only the non-zero eigenvalues are considered. We then see that when the relative number of small eigenvalues increases, the generalization rate drops dramatically.

7 Discussion.

The interpretation around the spectrum of the eigenvalues of H comes from the fact that the statistics of the realized learning set are not the same as the general statistics over the total set of all possible input patterns (in the case of small values of α). In the MDP for example, the general statistics are spherical (all the 2^N patterns are equally probable), whereas the statistics of the learning set can be strongly non-uniform (as shown by the possibly extended spectrum of eigenvalues of H). The principal idea developed here is to find a way to project the test patterns in the limit of the region defined by the learning set. This approach is shown to work well on the simple case of the MDP. We have tried it on a real problem of classification of under-sea noise spectra, and the results are quantitatively very similar.

8 Concluding remarks.

In the case of small values of α (number of learned patterns / dimension) the pseudo-inverse solution (PIS) can show strong overfitting behavior, and thus is not the optimal linear solution. We have illustrated this point with simulations about the majority detection problem (MDP). Two solutions to this overfitting problem have been exhibited (a one-layer and a two-layer perceptron), and are based on the interpretation of the spectrum of the eigenvalues of the covariance matrix H. They work quite well on the MDP, due to the fact that the learning set is statistically rather different than the total set of all possible patterns. Another related strange behavior can also be explained with statistical considerations: the PIS can decrease its generalization capacities when more examples are learned. This fact is not conform to the intuition that systems which have learned more patterns have better performance for recognition.

References

[1] D. J. Amit, H. Gutfreund, and H. Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, 55(14):1530–1533, 1985.

[2] H. D. Block. *Rev. Mod. Phys*, 34(123), 1962.

[3] G. W. Cottrell, P. Munro, and D. Zipser. *Image Compression by Back-Propagation: An Example of Extensional Programming.* In *Advances in cognitive science.* Volume 3, Norwood, NJ: Ablex, 1987.

[4] S. Diederich and M. Opper. Learning of correlated patterns in spin-glass networks by local learning rules. *Phys. Rev. Lett.*, 58(9):949–952, 1987.

[5] E. Domany, R. Meir, and W. Kinzel. Storing and retrieving information in a layered spin system. *Europhys. Lett.*, 2(3):175–185, 1986.

[6] E. Gardner. Maximum storage capacity in neural networks. *Europhys. Lett.*, 4(4):481–485, 1987.

[7] E. Gardner and B. Derrida. Optimal storage properties of neural network models. *J. Phys. A: Math Gen*, 21:271–284, 1988.

[8] E. Gardner and B. Derrida. Three unfinished works on the optimal storage capacity of networks. (to appear in *J. Phys*).

[9] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. In *Proc. Natl. Acad. Sci. USA*, pages 2554–2558, 1982.

[10] T. Kohonen. *Self organisation and associative memories.* Springer Verlag, 1981.

[11] W. Krauth and M. Mézard. Learning algorithms with optimal stability in neural networks. *J. Phys. A*, 20:L745–L752, 1987.

[12] E. Meir and E. Domany. Stochastic dynamics of a layered neural network, exact solution. *Europhys. Lett.*, 4(6):645, 1987.

[13] M. Mézard and J. Nadal. Learning in feedforward layered networks: the tiling algorithm. 1989. preprint submitted to *J. Phys. A*.

[14] M. Minsky and S. Papert. *Perceptrons.* MIT press, Cambridge, 1988. (exp. ed.).

[15] T. J. Sejnowski, P. K. Kienker, and G. E. Hinton. Learning symmetry groups with hidden units: beyond the perceptron. *Physica*, 22D:260–275, 1986.

[16] F. Vallet. The Hebb rule for learning linearly separable functions: learning and generalisation. 1989. preprint submitted to *Europhys. Lett.*

[17] G. Weisbuch and F. Fogelman-Soulié. Scaling laws for the attractors of Hopfield networks. *J. Physique Lett.*, 46:L623–L630, 1985.