A dilution algorithm for neural networks

1992 J. Phys. A: Math. Gen. 25 L593

(http://iopscience.iop.org/0305-4470/25/9/014)

J. Phys. A: Math. Gen. 25 (1992) L593-L598. Printed in the UK

LETTER TO THE EDITOR

A dilution algorithm for neural networks

P Kuhlmann, R Garcés and H Eissfeller

Institut für Theoretische Physik III, Justus-Liebig-Universität Giessen, Heinrich-Buff-Ring 16, D-W6300 Giessen, Federal Republic of Germany

Received 9 December 1991, in final form 13 February 1992

Abstract. A dilution algorithm to enlarge the storage capacity per synapse α_{eff} of neural networks is proposed. The algorithm is a hybrid method, where Hebb's rule is used to select a fraction of couplings to be removed. Afterwards the perceptron of optimal stability for the remaining couplings is learned. We present an analytical calculation and the results of the numerical simulations. In comparison with the fully connected or the randomly diluted perceptron, the effective storage capacity α_{eff} is remarkably enlarged.

In recent years attractor and feedforward neural networks have gained a great amount of interest (for an introduction see e.g. Hertz *et al* [1]). Feedforward networks can be used to classify a set of given patterns. One of the interesting parameters of such a network is the critical storage capacity α_c, which is the ratio of the maximum number p of patterns that can be learned perfectly and the number N of input neurons. The perceptron [2] is the simplest feedforward network and in the thermodynamic limit ($N \to \infty$) its critical storage capacity has been shown to be $\alpha_c = 2$, if the p given patterns are in general position [3]. A recent result by Bouten *et al* [4] seems to contradict, at first sight, the above limitation for α_c. Bouten *et al* performed a Gardner calculation [5] of the phase space volume in order to determine α_c in the case of an optimally diluted perceptron. The ratio between the maximum number of patterns p, that can be learned perfectly and the number of the remaining neurons in the network Nf,

$$\alpha_{\text{eff}} = \frac{p}{Nf}$$

can be much greater than 2 and even diverges logarithmically in the limit $f \to 0$.

Although analytical calculations [4, 6, 7] and extensive numerical simulations [8] have been done in the field of the dilution of the perceptron, there has not yet been found any algorithm that yields $\alpha_{\text{eff}} > 2$. In this letter we present both an analytical calculation and numerical simulations for a dilution algorithm that results in $\alpha_{\text{eff}} > 2$.

To introduce our algorithm we consider a simple feedforward perceptron that consists of N neurons S_j and a single binary output. The network is required to store $p = \alpha N$ patterns ξ_j^μ, $\mu \in \{1, \ldots, p\}$, $j \in \{1, \ldots, N\}$. The ξ_j^μ are chosen independently according to the Gaussian distribution

$$p(\xi_j^\mu) = \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}(\xi_j^\mu)^2). \tag{1}$$

Each pattern has a binary output $S^\nu \in \{-1, 1\}$ that has to be retrieved by the network. The outputs are chosen independently with $p(S^\nu) = \frac{1}{2}$.

0305-4470/92/090593+06$04.50

The perceptron problem is to find a vector $J=(J_1,\ldots,J_N)^T$ that maps all the $p=\alpha N$ aptterns to the right outputs:

$$S^\nu=\operatorname{sign}\left(\sum_{j=1}^{N} J_j\xi_j^\nu\right) \qquad \text{for all } \nu=1,\ldots,p. \tag{2}$$

This is equivalent to the condition

$$E_\nu=\frac{1}{\sqrt{N}}\sum_{j=1}^{N} J_j\sigma_j^\nu>0 \qquad \text{for all } \nu=1,\ldots,p \tag{3}$$

for the local fields E^ν of the p modified patterns $\sigma_j^\nu=S^\nu\xi_j^\nu$. The *stability* κ of the perceptron is defined as

$$\kappa=\min_\nu\{E^\nu\}/\sqrt{Q} \tag{4}$$

where $Q=1/N\,\Sigma_{j=1}^{N} J_j^2$ is the square of the norm of the couplings. The *perceptron of maximal stability* fulfils

$$E^\nu\geqslant 1 \qquad \text{for all } \nu=1,\ldots,p \tag{5}$$

with minimal Q. The critical capacity $\alpha_c(\kappa)$ of this perceptron of maximal stability has been calculated by Gardner [5].

Motivated by Sompolinsky [9] and Domany *et al* [10], we introduce the following algorithm: In order to dilute we calculate the N Hebb couplings [11]

$$B_j=\frac{1}{\sqrt{N}}\sum_{\nu=1}^{p}\sigma_j^\nu \tag{6}$$

and use them for the removal of $N(1-f)$ many sites: All the sites j whose absolute values $|B_j|$ are lower than a threshold value s will be removed. On the remaining Nf sites the problem of finding the perceptron of optimal stability has to be solved.

The dilution f can be computed as a function of the threshold value s as follows:

$$f=\frac{1}{N}\sum_{j=1}^{N}\Theta(|B_j|-s) \qquad \text{where } \Theta(x)=\begin{cases}1 & \text{for } x\geqslant 0\\ 0 & \text{for } x<0.\end{cases} \tag{7}$$

Since the $\boldsymbol{\sigma}^\nu=(\sigma_1^\nu,\ldots,\sigma_N^\nu)^T$ are Gaussian, the vector $\boldsymbol{B}=(B_1,\ldots,B_N)^T$ has an uncorrelated Gaussian distribution.

As f is self-averaging for $N\to\infty$, we obtain

$$f=2\Phi\left(-\frac{s}{\sqrt{\alpha}}\right) \tag{8}$$

where $\Phi(x)=\int_{-\infty}^{x}\mathrm{d}\lambda/\sqrt{2\pi}\,\mathrm{e}^{-\frac{1}{2}\lambda^2}$.

So we know how to choose the threshold for the Hebb couplings in order to get a dilution f. We label the remaining sites by k and start the Gardner calculation with the canonical partition function

$$\begin{aligned}Z&=\left(\prod_{k=1}^{Nf}\int_{-\infty}^{+\infty}\mathrm{d}J_k\right)\left(\prod_{\nu=1}^{p}\Theta\left(\frac{1}{\sqrt{Nf}}\sum_{k=1}^{Nf}J_k\sigma_k^\nu-1\right)\right)\exp\left(-\beta\sum_{k=1}^{Nf}J_k^2\right)\\&=\left(\frac{\beta}{\pi}\right)^{N[(1-f)/2]}\left(\prod_{j=1}^{N}\int_{-\infty}^{+\infty}\mathrm{d}T_j\right)\\&\quad\times\left(\prod_{\nu=1}^{p}\Theta\left(\frac{1}{\sqrt{Nf}}\sum_{j=1}^{N}c_jT_j\sigma_j^\nu-1\right)\right)\exp\left(-\beta\sum_{j=1}^{N}T_j^2\right)\end{aligned} \tag{9}$$

where the T_j are our new integration variables. c_j denotes whether a site is removed:

$$c_j = \Theta(|B_j| - s) = \Theta\left(\frac{1}{\sqrt{N}}\left|\sum_{\nu=1}^{p} \sigma_j^\nu\right| - s\right). \tag{10}$$

Note that in equation (9) the temperature parameter β controls the square of the norm of the couplings. β must not be confused with the temperature parameter in [12], which controls the number of errors. We checked that the ansatz in equation (9) is equivalent to the one in [5].

We assume that the free energy $g = -(1/\beta N) \ln Z$ is self-averaging:

$$\lim_{N\to\infty} g = \lim_{N\to\infty} -\frac{1}{\beta N} \langle\langle \ln Z \rangle\rangle \tag{11}$$

where $\langle\langle \ldots \rangle\rangle$ denotes the average over the modified patterns $\{\sigma_j^\nu\}$.

The right-hand side of equation (11) is calculated by means of the replica method (see van Hemmen and Palmer [13]). The matrix

$$Q_{\rho\sigma} = \frac{1}{Nf} \sum_{j=1}^{N} T_j^\rho T_j^\sigma \Theta(|B_j'| - s)$$

appears as a saddle point variable in the replica calculation, where B_j' is defined by $f(B_j) = \int_{\infty}^{+\infty} \mathrm{d}B_j' f(B_j') \delta(B_j' - B_j)$.

If we assume replica symmetry, $Q_{\rho\rho} = Q \; \forall \rho$, $Q_{\rho\sigma} = q$, $\rho \neq \sigma$, then Q yields the square of the norm of the couplings and q describes the overlap between two different solutions of the perceptron problem.

The stability $\kappa = 1/\sqrt{Q}$ is given by equation (4). Thus, if we let β tend to infinity, our theory describes the perceptron of optimal stability. After a transformation of variables we obtain the following result in the limit $\beta \to \infty$.

Given a dilution f and a stability κ the critical capacity $\alpha_c(f, \kappa)$ is obtained from

$$\alpha_c = \frac{f}{(f/2C)(\kappa - a)^2 + (1 + a^2)\Phi(a) + (a/\sqrt{\pi}) \exp(-\frac{1}{2}a^2)} \tag{12}$$

where C is given by

$$C(w) = \frac{w}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}w^2) \tag{13}$$

and w is equal to $s/\sqrt{\alpha}$ in equation (8). The saddle point variable a is the solution of

$$f(\kappa - a) = 2C\left(\frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}a^2) + a\Phi(a)\right). \tag{14}$$

The function α_{eff} for our dilution rule is shown in figure 2 below.

Our numerical simulation is a hybrid method, which corresponds exactly to the analytical calculation described above. Hebb's rule is used to learn the given patterns. Then all the synaptical strengths $|B_j| \leqslant s$ are set to zero. Afterwards the couplings are learned on the remaining sites using the AdaTron rule [14].

Since numerical simulations are restricted to a finite number N of neurons, learning binary patterns with Hebb's rule has the unpleasant disadvantage of highly degenerate B_j values. Hence the synapses that are to be cut in order to get a specified dilution f cannot be determined uniquely. Therefore we use Gaussian distributed patterns in order to reduce these finite size effects.

The AdaTron algorithm uses the embedding strengths x^ν as its dynamical variables instead of the couplings $J=(J_1,\ldots,J_N)^T$. The couplings and the embedding strengths are related through the following equation:

$$J_j=\frac{1}{Nf}\sum_{\nu=1}^{p} c_j\sigma_j^\nu x^\nu. \tag{15}$$

The dynamical equation is

$$x^\nu(t+1)=x^\nu(t)+\delta x^\nu(t) \qquad \text{for all } \nu=1,\ldots,p \tag{16}$$

where

$$\delta x^\nu(t)=\max\{\gamma(1-E^\nu),-x^\nu\}. \tag{17}$$

Note that $x^\nu\geqslant 0$ is always fulfilled. We define the positive semidefinite correlation matrix **C** with elements

$$C^{\nu\mu}=\frac{1}{Nf}\sum_{j=1}^{N} c_j\sigma_j^\nu\sigma_j^\mu \tag{18}$$

f being the fraction of all connections that still remain in the system and c_j as in equation (10). The field E^ν of the pattern ν can be written as

$$\boldsymbol{E}=\mathbf{C}\boldsymbol{x} \tag{19}$$

where the p-vectors are given by $\boldsymbol{x}=(x^1,\ldots,x^p)^T$ and $\boldsymbol{E}=(E^1,\ldots,E^p)^T$.

The dynamical updating rule for the x^ν described above will be repeated until the following condition is met for all ν:

$$\text{either } (x^\nu=0 \text{ and } E^\nu\geqslant 1) \text{ or } (x^\nu>0 \text{ and } E^\nu=1). \tag{20}$$

Our diluted network relaxes exponentially towards the perceptron of optimal stability for the Nf remaining neurons. After the AdaTron algorithm has converged, we calculate the stability of the network by means of equation (4).

For $f=1$ one knows that it is difficult to obtain small values of κ by numerical simulations [15], because near the critical point α_{eff} for $\kappa=0$ the learning time diverges for any known perceptron learning rule [16]. In order to obtain α_{eff} from the simulation data one has to extrapolate to $\kappa=0$.

However, in our case we can calculate analytically $\kappa(\alpha)$ for any given value of f. So we are able to compare these curves directly to the results of our numerical simulations. This can be seen in figure 1 for some dilution values f. We find that the data from the simulations, performed for $N=200$ neurons, and averaged over 50 samples, is in very good agreement with the theoretical curves.

In figure 2 we plotted the α_{eff} values, i.e. the fraction p/Nf where $\kappa=0$, versus the dilution f. For comparison two curves determined by Bouten *et al* are also given. The straight line corresponds to the 'quenched dilution' case, in which a certain fraction f of couplings is removed at random without retraining. As one could expect this procedure does not change the effective storage capacity, $\alpha_{\text{eff}}(f)=2$ for all f. The upper curve corresponds to the (as named by Bouten *et al*) 'annealed dilution' case. This is the result of their replica symmetric calculation, yielding the curve for the 'optimal' α_{eff}. But we have to keep in mind that up until now there has been no algorithm that corresponds to these theoretical calculations. We also want to remark, at this point, that it has recently been found that the replica symmetric solution of Bouten *et al* is unstable [17]. In our case replica symmetry breaking effects, if they

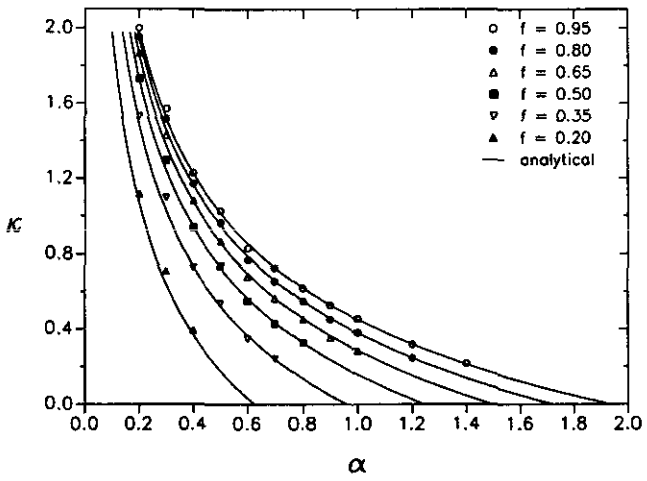

Figure 1. Shown is the stability $\kappa(\alpha)$ for the given values of f. Our numerical simulations (symbols) are compared with the analytical results (solid curves). The statistical errors are of the same order as the symbol sizes.

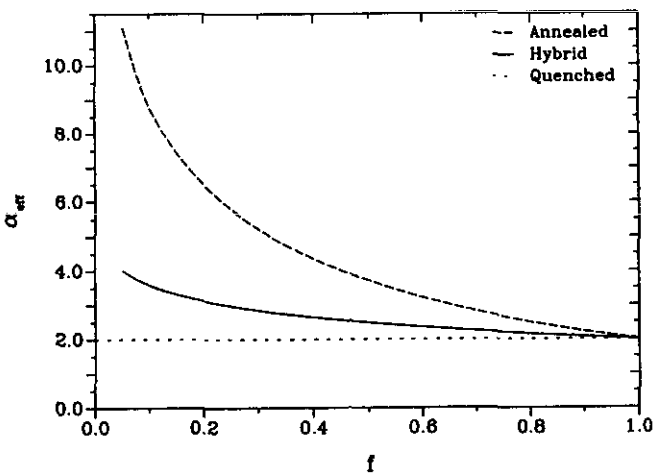

Figure 2. The effective storage capacity $\alpha_{\rm eff} = \alpha/f$ is given as a function of the dilution f for $\kappa = 0$ for the annealed dilution (upper curve) case for the hybrid method (middle curve), and the quenched dilution (lower curve) case from Bouten *et al.*

exist, are expected to be less important with respect to $\alpha_{\rm eff}$, since the results of the simulations coincide with the analytical results.

Our results are given by the solid line. Like the two other curves it also starts at $f = 1$ and $\alpha_{\rm eff} = 2$. For decreasing f the performance of our hybrid algorithm is much better than random dilution, and $\alpha_{\rm eff}$ diverges in the limit $f \to 0$. So the qualitative behaviour is similar to the replica symmetric solution of Bouten *et al.*

A quantitative comparison suggests that other algorithms exist with a better performance than our hybrid method. For example it is possible to dilute in several steps, each consisting of removing bonds according to some rule and relearning for the remaining ones. We have some preliminary results of such a method using the AdaTron algorithm

after each dilution step [18]. They indicate that this method has a better performance than the hybrid method, but still the curve presented by Bouten *et al* is not reached. Details will be given in a forthcoming paper.

In this sense our algorithm can be regarded as a first step. But such an iterative procedure is not possible with Hebb's rule. After cutting and applying it again, the values of the remaining couplings do not change. An interesting and still open question is whether any algorithm can be found, that meets the 'optimal curve'. For practical applications we think the efficiency of the hybrid method presented here is better than an iterative dilution process. Hebb's rule needs much less numerical operations than an iterative algorithm, where for example the optimal perceptron has to be learned after each dilution step. The benefits of a somewhat greater storage capacity will be compensated by the great amount of computational effort.

To summarize, we presented a first, rather simple algorithm for dilution, which yields $\alpha_{\text{eff}} > 2$. By using Hebbs rule to classify couplings of low importance, removing them and calculating the optimal perceptron matrix, the effective storage capacity is remarkably enlarged.

The authors would like to thank M Opper, W Kinzel and M Biehl for many stimulating discussions. The numerical simulations were carried out on the CRAY Y-MP of the HLRZ Jülich. The work was supported by grants of the Deutsche Forschungsgemeinschaft. It is part of the PhD Thesis of PK and of HE.

References

[1] Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Redwood City: Addison-Wesley)
[2] Rosenblatt F 1958 *Psychoanal. Rev.* **65** 3 86
Minsky M and Papert S 1969 *Perceptrons* (Cambridge, MA: MIT Press)
[3] Cover T M 1965 *IEEE Transactions on Electronic Computers* **14** 326
[4] Bouten M *et al* 1990 *J. Phys. A: Math. Gen.* **23** 4643
[5] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[6] Wong K Y M and Bouten M 1991 *Europhys. Lett.* **16** 525
[7] Bollé D, Dupont P and van Mourik J 1991 *Leuven Preprint*
[8] Kürten K E 1990 *J. Physique* **51** 1585
[9] Sompolinsky H 1987 *Heidelberg Coll. on Glassy Dynamics and Optimization* ed J L van Hemmen and I Morgenstern (Berlin: Springer)
[10] Domany E, Kinzel W and Meir R 1989 *J. Phys. A: Math. Gen.* **22** 2081
[11] Hebb O D 1949 *The Organisation of Behavior* (New York: Wiley)
[12] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[13] van Hemmen J L and Palmer R G 1979 *J. Phys. A: Math. Gen.* **12** 563
[14] Anlauf J K and Biehl M 1989 *Europhys. Lett.* **10** 687
[15] Krauth W and Mezard M 1987 *J. Phys. A: Math. Gen.* **20** L745
[16] Opper M 1988 *Phys. Rev.* A **38** 3824
[17] Bouten M, Engel A and Wong K Y M Private communications
[18] Garcés R 1991 *Diploma thesis* Giessen