In search of an optimal dilution algorithm for feedforward networks

1992 J. Phys. A: Math. Gen. 25 L1335

(http://iopscience.iop.org/0305-4470/25/23/012)

J. Phys. A: Math. Gen. 25 (1992) L1335-L1342. Printed in the UK

LETTER TO THE EDITOR

In search of an optimal dilution algorithm for feedforward networks

R Garcés, P Kuhlmann and H Eissfeller

Institut für theoretische Physik III, Julius-Maximilians-Universität Würzburg, Am Hubland, D-W8700 Würzburg, Federal Republic of Germany

Received 6 July 1992

Abstract. In a recent paper we presented a dilution algorithm that yields a storage capacity per synapse α_{eff} larger than 2. In this letter, two new algorithms with even higher α_{eff} values are introduced. We study a one-step dilution algorithm, where the perceptron is used to select a fraction of the couplings to be removed. For the remaining bonds the perceptron of optimal stability is relearned. We further compare simulation data from an iterative version of the one-step dilution algorithm with phase-space volume calculations.

Several dilution models for feedforward as well as for attractor neural networks have been proposed in the past. For diluted attractor neural networks calculations of thermodynamical properties have been performed and the sizes of the basins of attraction have been determined [1-4]. Feedforward neural networks have been used to learn a set of given patterns perfectly. Their dilution has been treated analytically for different models [5-7]. Our motivation is to find an algorithm for feedforward neural networks which fulfils a given task with a minimum amount of synapses, since this is desirable for hardware realizations.

In a recent paper [8] we have shown that by using the Hebb couplings for the selection of the bonds to be removed, and learning the perceptron of optimal stability afterwards, a high storage capacity per synapsis can be achieved.

In this letter we address the question of the minimum number of synapses that a feedforward perceptron needs to map p given patterns to the corresponding outputs. Two algorithms will be presented that result in an α_{eff} greater than the one reached by the hybrid method [8]. As a first step we construct the preceptron of optimal stability instead of using the Hebbian couplings. We remove all the bonds with absolute value lower than a threshold and learn the perceptron of optimal stability for the remaining bonds again. For this one-step algorithm we are able to calculate the critical storage capacity analytically. The second algorithm is an iterative version, where in each step only the weakest bond is removed and the perceptron of optimal stability is learned again on the remaining sites.

We consider a simple perceptron consisting of an input layer of N neurons $S_j, j \in \{1, \ldots, N\}$ feeding directly into the output S. The network is required to store $p = \alpha N$ patterns ξ_j^ν, $\nu \in \{1, \ldots, p\}, j \in \{1, \ldots, N\}$. The ξ_j^ν are chosen randomly according to the Gaussian distribution function

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

0305-4470/92/231335+08$07.50

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 the probability $p(S^\nu = \pm 1) = \frac{1}{2}$.

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

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

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 $Q = (1/N) \sum_{j=1}^{N} J_j^2$. For a given α the perceptron of optimal stability is unique. In the limit $N \to \infty$ its critical storage capacity has been calculated by Gardner [9] ($\alpha_c = 2$).

Instead of selecting the bonds that have to be removed according to the Hebbian learning rule [8], we calculate the perceptron of optimal stability. In order to obtain a desired dilution value f in one step, we have to cut the couplings at a threshold w, which is given by the restriction

$$f = \frac{1}{N} \sum_{j=1}^{N} \Theta(|J_j| - w) \qquad \text{where} \qquad \Theta(x) = \begin{cases} 1 & \text{for } x > 0 \\ 0 & \text{for } x \leqslant 0. \end{cases} \tag{5}$$

This results in the following algorithm:

1. Learn the perceptron of optimal stability with a coupling vector $\boldsymbol{J} = (J_1, \ldots, J_N)^T$.
2. Remove all the sites j whose corresponding couplings have absolute values $|J_j|$ smaller than a threshold w. Therefore $N(1-f)$ sites are removed, where f is the fraction of the remaining sites.
3. Relearn the perceptron of optimal stability on the remaining sites.

For the analytical calculation we assume that f is self-averaging in the limit $N \to \infty$ with respect to taking averages over the first perceptron and the patterns. Since the couplings of the first perceptron are normally distributed, this leads to:

$$f = 2\Phi(-w) \tag{6}$$

where

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

To start the Gardner calculation [9] of the fractional phase-space volume V, we choose the p modified patterns and learn the first perceptron with couplings $\boldsymbol{J}$ and stability κ_1. Let the η_k^ν, $k \in \{1, \ldots, Nf\}$, denote the σ_j^ν with $c_j = \Theta(|J_j| - w) = 1$, i.e. the patterns on the remaining sites k after the dilution procedure. If we require a stability κ_2 for the diluted perceptron the fractional phase-space volume on the remaining

sites is

$$V=\left(\prod_{k=1}^{Nf}\int_{-\infty}^{+\infty}\mathrm{d}L_k\right)\delta\left(\sum_{k=1}^{Nf}L_k^2-Nf\right)\prod_{\nu=1}^{p}\Theta\left(\frac{1}{\sqrt{Nf}}\sum_{k=1}^{Nf}L_k\eta_k^\nu-\kappa_2\right)$$
$$\times\left[\left(\prod_{k=1}^{Nf}\int_{-\infty}^{+\infty}\mathrm{d}L_k\right)\delta\left(\sum_{k=1}^{Nf}L_k^2-Nf\right)\right]^{-1}$$
$$=\frac{1}{C_{\mathrm{norm}}}\left(\prod_{j=1}^{N}\int_{-\infty}^{+\infty}\frac{\mathrm{d}T_j}{\sqrt{2\pi}}\right)\delta\left(\sum_{j=1}^{N}c_jT_j^2-Nf\right)\exp\left(-\frac{1}{2}\sum_{j=1}^{N}(1-c_j)T_j^2\right)$$
$$\times\prod_{\nu=1}^{p}\Theta\left(\frac{1}{\sqrt{Nf}}\sum_{j=1}^{N}c_jT_j\sigma_j^\nu-\kappa_2\right) \tag{7}$$

with

$$C_{\mathrm{norm}}=\left(\prod_{j=1}^{N}\int_{-\infty}^{+\infty}\frac{\mathrm{d}T_j}{\sqrt{2\pi}}\right)\delta\left(\sum_{j=1}^{N}c_jT_j^2-Nf\right)\exp\left(-\frac{1}{2}\sum_{j=1}^{N}(1-c_j)T_j^2\right).$$

We assume that the entropy $s=(1/N)\ln V$ is self-averaging with respect to a replica-average over the first perceptron and an average over the modified patterns:

$$\lim_{N\to\infty}s=\lim_{N\to\infty}\frac{1}{N}\langle\!\langle\langle\ln V(\boldsymbol{J})\rangle_{\{\boldsymbol{J}\}}\rangle\!\rangle_{\{\sigma_j^\nu\}} \tag{8}$$

where $\langle\!\langle\ldots\rangle\!\rangle$ is the average with respect to the probability distribution of the modified patterns. Since the $S^\nu=\pm1$ are chosen with equal probability the distribution of the modified patterns is identical to the one of the patterns themselves (1).

The average $\langle\ldots\rangle$ over the first perceptron has to be calculated by means of a replica average [10, 11]:

$$\langle\ln V(\boldsymbol{J})\rangle_{\{\boldsymbol{J}\}}=\lim_{m\to0}\left(\prod_{\beta=1}^{m}\left(\prod_{j=1}^{N}\int_{-\infty}^{+\infty}\mathrm{d}J_j^\beta\right)\delta\left(\sum_{j=1}^{N}(J_j^\beta)^2-N\right)\right)$$
$$\times\left(\prod_{\beta=1}^{m}\prod_{\nu=1}^{p}\Theta\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N}J_j^\beta\sigma_j^\nu-\kappa_1\right)\right)\ln\bar{V}(\boldsymbol{J}^1). \tag{9}$$

Note that in contrast to a standard replica calculation of an averaged free energy [12], in this replica average one replicon (for instance the first one, see equation (9)) is specified. The replica-index $\beta\in\{1,\ldots,m\}$ has been introduced to calculate the average over the first perceptron. To average $\ln V(\boldsymbol{J})$ over the modified patterns a second replica index $\rho\in\{1,\ldots,n\}$ is needed. The following overlaps appear as saddle-point variables in the calculation:

$$P_{\beta\gamma}=\frac{1}{N}\sum_{j=1}^{N}J_j^\beta J_j^\gamma \tag{10}$$

$$R_{\beta\rho}=\frac{1}{N}\sum_{j=1}^{N}J_j^\beta c_jT_j^\rho \tag{11}$$

$$Q_{\rho\sigma}=\frac{1}{Nf}\sum_{j=1}^{N}c_jT_j^\rho T_j^\sigma. \tag{12}$$

We assume replica symmetry for $P_{\beta\gamma}$ and $Q_{\rho\sigma}$, $P_{\beta\beta}=1\forall\beta$, $P_{\beta\gamma}=p\forall\beta,\beta\neq\gamma$, and accordingly for $Q_{\rho\sigma}$. Since in equation (9) the replicon $\beta=1$ is specified, $R_{1\rho}$ has to be treated separately. Replica symmetry is assumed for the remaining replica. Therefore we set $r_1=R_{1\rho}\forall\rho$ and $r=R_{\beta\rho}\forall\beta=2,\ldots,m,\forall\rho$.

The calculation of the critical storage capacity for the second perceptron is simplified, if we require the first perceptron to have optimal stability. The limit $p\to 1$ then yields:

$$\alpha(\kappa_1)=\left[(1+\kappa_1^2)\Phi(\kappa_1)+\frac{\kappa_1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\kappa_1^2\right)\right]^{-1} \tag{13}$$

with $\Phi(x)$ from equation (6) as calculated by Gardner [9].

We calculate the function $l=\lim_{q\to 1}(1-q)s$ at the critical point of the second perceptron, where s is the entropy from (8). After a transformation of variables we obtain:

$$\begin{aligned} l(\kappa_1,a,b,\kappa_2,f) &= -\frac{1}{2}\alpha(\kappa_1)\int_{-\kappa_2}^{\infty}\frac{dz}{\sqrt{2\pi}}(\kappa_2+z)^2\left[\exp\left(-\frac{1}{2}z^2\right)\Phi\left(-\frac{\kappa_1+za}{\sqrt{1-a^2}}\right)\right. \\ &\quad +\frac{1}{\sqrt{1-a^2+b^2}}\exp\left(-\frac{1}{2}\frac{(z+\kappa_1(a-b))^2}{1-a^2+b^2}\right) \\ &\quad \left.\times\Phi\left(\sqrt{\frac{1-a^2+b^2}{1-a^2}}\left(\kappa_1+\frac{b}{1-a^2+b^2}(z+\kappa_1(a-b))\right)\right)\right] \\ &\quad -\frac{fa^2}{2E}+\frac{b^2}{2}+\frac{f}{2} \end{aligned} \tag{14}$$

where a and b are the remaining order parameters and $E=f+\sqrt{2/\pi}\,w\exp(-\frac{1}{2}w^2)$, where w is the threshold for the cutting of the couplings.

Given κ_2 and f, the order parameters a, b and the critical stability κ_1 are calculated by solving the following system of equations (similar to [14]):

$$\begin{aligned} &l(\kappa_1,a,b,\kappa_2,f)=0 \\ &\frac{\partial l}{\partial a}=0 \qquad \frac{\partial l}{\partial b}=0. \end{aligned} \tag{15}$$

Together with equation (13) the maximum storage capacity $\alpha_c(\kappa_2,f)$ is calculated. This data is given in figure 1, where we have plotted κ_2 as a function of α for some values of f. For comparison the results of the numerical simulations are also given. The simulations were run on a system of $N=200$ neurons, and averaged over 50 samples. The perceptron of optimal stability can be learned using various algorithms [15, 16], we used the AdaTron algorithm [17]. The analytical results and the simulation data are shown in figure 1. They are in very good agreement.

The fraction

$$\alpha_{\text{eff}}=\frac{\alpha(\kappa_2=0)}{f} \tag{16}$$

represents the critical storage capacity per synapse. The corresponding curve for the one-step algorithm is plotted in figure 2.

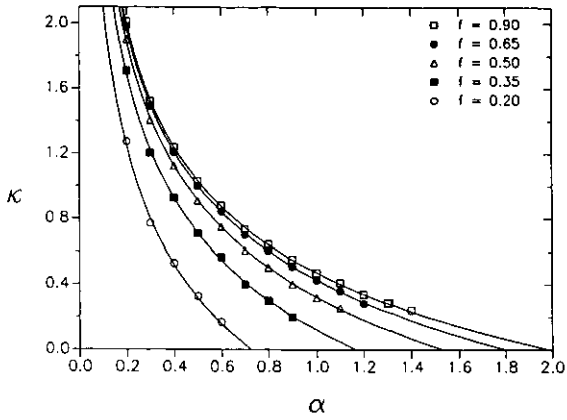

Figure 1. The stability $\kappa(\alpha)$ is given as a function of α for some values of f. The numerical simulations (symbols) are compared with the analytical results (solid curves). The results were averaged over 50 runs with systems of $N = 200$ neurons each. The statistical errors are smaller than the symbol sizes.

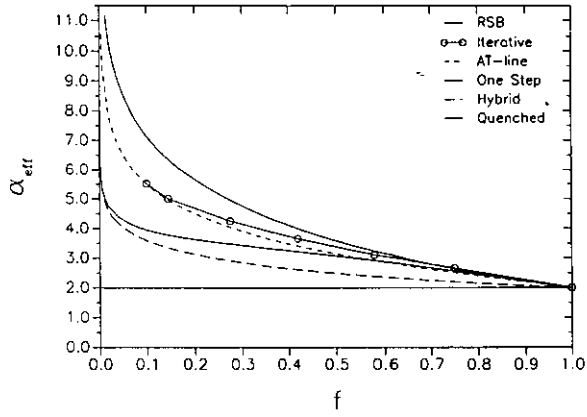

Figure 2. The effective storage capacity $\alpha_{eff} = \alpha_c / f$ is given as a function of the dilution f. Shown from top to bottom are the RSB solution of the annealed dilution case, the multistep method, the AT-line, the perceptron one-step method, the hybrid algorithm as well as the quenched dilution case.

In the system of equations (15) the order parameter a can be interpreted as an overlap between the first and second perceptron, both of them being required to have optimal stability. Therefore a provides a measurement for the importance of the relearning procedure, since decreasing overlaps a indicate an increasing difference between the coupling vectors of the two perceptrons,

$$a = \frac{r}{\sqrt{f}} \tag{17}$$

where r has been defined above as $r = R_{\beta\rho}$, $\beta \geq 2$ (see equation (11)). The analytical and numerical results for a are in very good agreement, both of them are shown in figure 3. As expected for strongly diluted networks (small f values) and large storage

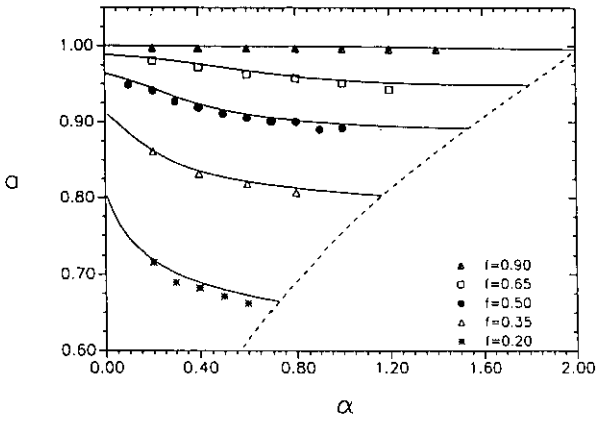

Figure 3. Shown is the overlap a between the first and second perceptron as a function of the storage capacity α for different dilution values f. The numerical simulations (symbols) are compared with the analytical results (solid curves). The broken curve represents the limit $\kappa = 0$. The results were averaged over 50 runs with systems of $N = 200$ neurons each. The statistical errors are smaller than the symbol sizes.

capacities α, i.e. large α_{eff}, the relearning procedure is more effective. The dashed line in figure 3 corresponds to the critical curve $\kappa_2 = 0$. The curve represents the lower bound for the overlaps a.

In order to improve our algorithm the following iterative procedure has been used:

1. Learn the perceptron of optimal stability with a coupling vector $\boldsymbol{J} = (J_1, \ldots, J_N)^{\mathrm{T}}$.
2. Remove only the site j with the coupling that has the smallest absolute value $|J_j|$.
3. Relearn the perceptron of optimal stability on the remaining sites.
4. Iterate steps two and three until the desired dilution is reached.

Since an analytical solution for this iterative algorithm has not yet been found, the simulation data must be extrapolated to $\kappa_2 = 0$ numerically. However, it is difficult to gain small κ values from simulations because near the critical storage capacity $\alpha_c(\kappa = 0)$, the learning time diverges for any known perceptron learning rule [18].

We found that the analytical curves calculated in the one-step dilution case give the best fit for our data, if we use f as a free parameter. The resulting curve $\alpha_{\text{eff}}(f)$ is given in figure 2.

For comparison we have also included the quenched dilution curve, results of the hybrid method [8], the AT-line [19] as well as the curve calculated in first-order replica symmetry breaking [20]. In the quenched dilution case the sites are removed at random and the diluted system is simply a network of Nf neurons with random outputs and uncorrelated patterns. Therefore α can be rescaled and $\alpha_{\text{eff}} = 2$ holds.

For the one-step dilution algorithm the storage capacity per synapsis is remarkably enlarged by using the perceptron of optimal stability instead of Hebb's rule for the removal of the bonds. A further increase is gained using the iterative algorithm. Nevertheless for practical applications we believe that the iterative method is not appropriate, since a great amount of computation is needed. So the one-step dilution algorithms should be preferred.

The results of the iterative algorithm are of greater interest from a theoretical point of view. The most important question in this context is: *What is the maximum α_{eff} that can be reached by any dilution algorithm*? This question has already been addressed by Bouten *et al* in [5]. Nevertheless it has recently been found that their replica

symmetric calculation is only stable below the AT-line in figure 2 [19]. The AT-line represents a lower bound that an optimal dilution algorithm must reach. Above this bound the replica symmetry is broken. A replica symmetry breaking (RSB) correction in first-order has recently been calculated [20] yielding an upper bound for α_{eff}. Since we suspect that for this problem a replica symmetry breaking behaviour analogous to the SK model [21] is present, further orders of RSB might be needed. Since the first-order RSB calculation remarkably decreases the α_{eff}, it is indeed possible that the iterative algorithm yields the optimal curve.

Up until now the highest values for α_{eff} have been achieved using the iterative algorithm. In principle every multistep dilution algorithm can be replaced by a one-step dilution procedure. The problem remains to discover the one-step selection rule realized by the multistep dilution procedure. If one knew this rule explicitly one could learn the optimal perceptron, remove the couplings according to this rule in one single step and relearn the perceptron of optimal stability.

To gain first hints about such a selection rule, we examined the distribution of the initial couplings which were removed by the iterative algorithm. Therefore we stored all the coupling values of the perceptron that were learnt in the first step and kept track of all the sites that were removed during the iterations. The distribution of the coupling values of the first perceptron which correspond to these removed sites is plotted in figure 4. This distribution is Gaussian for all dilution values f, with zero mean and standard deviations

$$\sigma_f = (1-f)\sigma_I \tag{18}$$

where σ_I is the standard deviation of the initial coupling distribution. The last equation arises from the simple fact that all zero couplings of the initial perceptron are removed with probability one.

In a first attempt to convert the iterative algorithm presented above into a one-step algorithm, we tried to cut the couplings independently by using the probability distributions from figure 4. Unfortunately this results in an α_{eff} lower than the one achieved by cutting the weakest couplings according to a bound w. Hence our assumption of

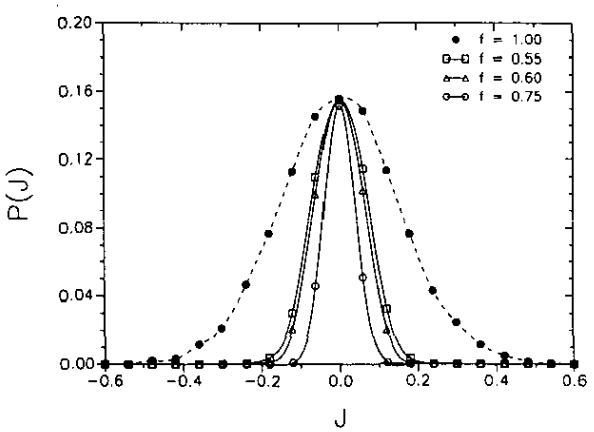

Figure 4. The probability distribution of the coupling coefficients that were removed from the system are shown as a function of different dilution values f. The initial storage capacity is $\alpha = 1.00$. The simulations were averaged over 50 samples of $N = 200$ neurons. The initial probability distribution for the fully connected system is shown by the dashed curve.

an independent probability distribution to dilute the sites does not yield an improvement. Therefore future one-step algorithm versions have to take care of the correlation matrix between the couplings of the first perceptron.

We 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 from the Deutsche Forschungsgemeinschaft. It is part of the PhD Thesis of PK and HE as well as the diploma thesis of RG.

References

[1] Sompolinsky H 1987 *Heidelberg Colloqium on Glassy Dynamics and Optimization* ed J L van Hemmen and I Morgenstern (Berlin: Springer)
[2] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167
[3] Domany E, Kinzel W and Meir R 1989 *J. Phys. A: Math. Gen.* **22** 2081
[4] Opper M, Kleinz J, Köhler H and Kinzel W 1989 *J. Phys. A: Math. Gen.* **22** L407
[5] Bouten M, Engel A, Komoda A and Serneels R 1990 *J. Phys. A: Math. Gen.* **23** 4643
[6] Bollé D, Dupont P and v Mourik J 1991 *Preprint* Leuven
[7] Wong K Y M and Bouten M 1991 *Europhys. Lett.* **16** 525
[8] Kuhlmann P, Garcés R and Eissfeller H 1992 *J. Phys. A: Math. Gen.* **25** L593
[9] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[10] Mézard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
[11] Fischer K H and Hertz J 1991 *Spin Glasses* (Cambridge: Cambridge University Press)
[12] Sherrington D and Kirkpatrick S 1978 *Phys. Rev.* A **35** 1792
[13] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[14] Engel A *et al* 1992 *Phys. Rev.* A **45** 7590
[15] Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* **20** L745
[16] Ruján P 1991 *Preprint*
[17] Anlauf J K and Biehl M 1989 *Europhys. Lett.* **10** 687
[18] Opper M 1988 *Phys. Rev.* **38** 3824
[19] Wong K Y M private communications
[20] Kuhlmann P 1992 in preparation
[21] Parisi G 1980 *J. Phys. A: Math. Gen.* **13** L115