Sign-constrained linear learning and diluting in neural networks

1991 J. Phys. A: Math. Gen. 24 L495

(http://iopscience.iop.org/0305-4470/24/9/008)

J. Phys. A: Math. Gen. 24 (1991) L495–L502. Printed in the UK

LETTER TO THE EDITOR

Sign-constrained linear learning and diluting in neural networks

H M Köhler and D Widmaier

Institut für Theoretische Physik der Georg-August-Universität, D-3400 Göttingen, Federal Republic of Germany

Received 12 November 1990, in final form 25 March 1991

Abstract. For neural networks with predefined effects of the synapses, excitatory or inhibitory, the simplex algorithm is applied as a learning rule. It is assumed that the given signs of the synapses can never be changed during learning. The maximum possible dilution of synapses, as a result of the learning, is found at the maximum storage capacity of a model with only positive or randomly distributed signs. For the case of infinitely many neurons, a replica symmetric calculation of the free energy and the distribution of coupling strengths is presented. The linear algorithm is also applied to networks with a more suitable choice of sign constraints, with the result of a higher storage capacity.

Neural networks of formal neurons have been shown to work well as associative memories or as trainable input–output machines. The reason for studying those systems is twofold. They are interesting technical devices on the one hand and a promising concept for understanding the nervous systems of organisms on the other hand. In order to get a better insight into the cooperative behaviour of neurons, biologically motivated requirements have to be taken into account when simple mathematical models are proposed. One of those requirements is the fact that excitatory and inhibitory synapses are different and consequently can never be transformed into one another. A somewhat stronger, but very similar, requirement was formulated as Dale's law [1] in 1935. Usually all synapses of a neuron only release one kind of neurotransmitter and consequently all synapses connecting the axon of this neuron with others are of one and the same sign. Amit and co-workers [2] proposed an algorithm for the sign restricted learning problem, which is similar to the Rosenblatt perceptron [3]. They also give a proof for convergence. Apparently their algorithm is neither very fast in convergence, nor does it converge to any optimal solution.

In this letter we will formulate sign-constrained learning as a linear optimization problem, which uses standard linear programming, and therefore is easy to handle and fast converging. We will show its relation to the dilution of synapses, which is another biological requirement; neurons in the central nervous system are not fully interconnected. We will then give a replica symmetric calculation of the proposed model for the case of positive-only synapses.

Learning rules for attractor neural networks or perceptrons usually produce synaptic couplings J_{ij} $(i, j = 1, \ldots, N)$, which stabilize a set of patterns $\xi_i^\mu \in \{-1, 1\}$

0305-4470/91/090495+08$03.50

($\mu = 1, \ldots, p$), so that their internal fields $h_i^\mu = N^{-1/2} \sum_{j(\neq i)} J_{ij} \xi_j^\mu$ and the respective patterns have the same sign

$$E_i^\mu = h_i^\mu \xi_i^\mu \geqslant c > 0 \,. \tag{1}$$

For simplicity only the case of randomly distributed patterns $p(\xi_i^\mu = \pm 1) = \frac{1}{2}$ will be considered. We conjecture that the statistics of learning for a model with prescibed random signs of synapses can simply be gauge transformed to a model with only positive synapses. For neuron i we therefore get N conditions (we may also presume non-negative self-couplings):

$$J_{ij} \geqslant 0 \qquad \forall j = 1, \ldots, N \,. \tag{2}$$

For a fixed c, say $c = 1$, and free normalization the condition for a great basin of attraction is to have J_{ij} of some minimal norm. Kinzel and Opper [4] define as minimal stability for neuron i

$$\kappa = \min_\nu \frac{\xi_i^\nu \sum_j J_{ij} \xi_j^\nu}{\sqrt{\sum_j J_{ij}^2}} = \frac{c}{\sqrt{\sum_j J_{ij}^2}} \tag{3}$$

which gives an approximate measure for the size of the basin of attraction (Kepler and Abbot [5]). This requirement leads to a quadratic optimization problem, where the quadratic norm of J must be minimized. This is used in Adaline learning, formulated by Widrow and Hoff [6], or in the AdaTron learning algorithm of Anlauf and Biehl [7]. Linear learning has been proposed for the general, not sign restricted, case with normalization $|J| \leqslant 1/\sqrt{N}$ [8]. Dual formulation yields a minimization of the maximum norm of J. In our case the sign of the synapses is positive so we can simply minimize the linear norm for neuron i:

$$d_i = \sum_{j(\neq i)} J_{ij} \,. \tag{4}$$

Now the whole problem is linear and can be solved by standard linear programming. Geometrically this results in finding a solution of the minimization problem (3), which lies on the vertices of the polytype given by (1) and (2), whereas minimization of a quadratic norm would generally find a solution on an edge of the polytype. We find solutions up to $\alpha = p/N = 1$. Above this value conditions (1) and (2) do not allow a feasible solution. With the result of Cover [9], which states that the total number of random uncorrelated conditions that cut off planes in the N-dimensional space cannot be higher than $2N$ in the limit $N \to \infty$, we can say that the probability for finding a solution above $\alpha = 1$ is vanishing for big systems. This is in accordance with a result obtained by Amit and co-workers [10] with the replica method as used by Gardner and Derrida [11], and also with a result of Nadal [12], obtained with a geometrical approach simmilar to [9].

Numerical evaluation of the problem shows that the solution is achieved by setting a considerable number of synapses to zero, while the rest of them have positive values. Figure 1 shows the dilution rate of the synapses and the relative frequency or probability of finding a solution as a function of the capacity α. Asterisks show results for positive resricted synapses; connecting lines are spline interpolations and meant

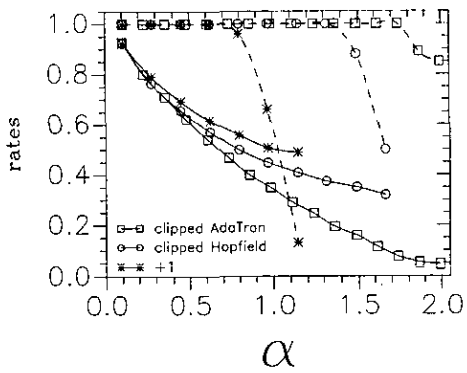

Figure 1. Dilution rate η (full curves) and success rate (broken curves) over capacity α for three different restrictions of the signs of the couplings; $N = 127$.

as a guide to the eye. One can see that at the maximum storage capacity, which is seen at $\alpha = 1$, the dilution rate nearly reaches the value 0.5. It is plausible that this is also the maximum possible value, because we have $2N$ conditions (1) and (2); N of those conditions will be fulfilled automatically; they are linear combinations of each other, because the space of $\{J_j\}$ is N-dimensional. Since patterns and dilution are uncorrelated, we conclude that $N/2$ of embedding strengths are $h_i^\mu \xi_i^\mu = c$, and $N/2$ of the conditions (2) are fulfilled with equality, which means that the dilution rate must be 0.5. Surprisingly this is the same value as the critical dilution rate, given by Bouten *et al* [13] in a recent paper for the case of quenched dilution and non-restricted J_{ij} (note the difference between quenched and annealed).

Leaving aside the easy motivation of biology, any single synapse of one neuron could be restricted to something that depends on the patterns in some way. So we also considered correlations between the sign of the couplings J_{ij} and the stored patterns ξ_j^μ. Now we minimize $\sum_{j(\neq i)} J_{ij} g_{ij}$, where $J_{ij} g_{ij} \geqslant 0$ and $g_{ij} = \pm 1$ for all i, j. Figure 1 gives a comparison between three choices of the sign restrictions. Hopfield sign restrictions with

$$g_{ij} = \operatorname{sign} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \tag{5}$$

already increase the capacity; results are plotted as circles in figure 1. The shape of the curve remains the same, but for equal α dilution rates are lower. Amazingly the maximum capacity considerably exceeds that of randomly or positive-sign-restricted J_{ij}, which means that although the Hopfield matrix does not allow for retrieval for $\alpha > 0.14$, at least an extensive number of couplings have the right sign for exact retrieval of all patterns for α between 0 and ≈ 1.65, where the success rate goes down dramatically. Geometrically we interpret the result as follows. By choosing a part of the couplings with the right sign, the probability of having a set of linear independent conditions (1) and (2) decreases. Thus the effective number of conditions decreases, which allows a higher storage capacity. The maximum possible storage capacity, which is $\alpha_c = 2$, was found for sign restrictions of the optimal perceptron (squares), which was calculated with the AdaTron algorithm. For $\alpha \to 2$ the dilution rate goes to approximately zero, because there is no more freedom for choice of the couplings. For smaller α dilution rates are again lower than those of the last curve, but again its shape remains similar.

For very small α all three curves meet at almost total dilution. In the limit of $N \to \infty$, only a finite number O(1) of synapses is needed to satisfy the constraints of a finite number of patterns. Simulations were done for systems of size $N = 127$ and all data points were averaged over 100 independent samples. A verification of many data points with systems of $N = 255$ produced good agreement.

We also checked anti-Hopfield sign restrictions $g_{ij} = -\text{sign} \sum_\mu \xi_i^\mu \xi_j^\mu$, but could not find a single solution; apparently all conditions are contradictory. Perhaps a solution could be found if a part of the couplings was chosen with the sign of Hopfield, while the rest would have anti-Hopfield sign. Similarly for the signs of the optimal perceptron, higher dilution rates could possibly be achieved for $\alpha < 2$ by choosing the right amount of synapses with the opposite sign.

We remark that the computer time required to find the solution of a linear problem with the standard simplex algorithm depends on system size as well as on the specific problem (see e.g. Klee and Minty [14]). For the given problem the solution can be found very quickly at data points with high rates of success; concerning the necessary computer time the linear learning algorithm is comparable to or better than other fast converging algorithms such as the AdaTron. Only at data points with low rates of success does the computer time increase considerably.

We now check the asymptotic behaviour of our learning rule for the case of positive-only couplings. We calculate the characteristic function $g(k)$ of the probability distribution $P(J)$ for the resulting couplings in the limit $N \to \infty$. Therefore $g(k)$ is written as formal thermodynamic limit of an average over the quenched variables ξ:

$$g(k) = \lim_{\beta \to \infty} \left\langle Z^{-1} \int_0^\infty \prod_j \mathrm{d}J_j \mathrm{e}^{\mathrm{i}kJ_l} \mathrm{e}^{-\beta \sum_j J_j} \prod_\mu \Theta(N^{-1/2} \sum_j J_j \xi_j^\mu - \kappa) \right\rangle \tag{6}$$

where κ stands for for the smallest linear stability and $\Theta(x)$ for the unit step function. The index $l \in \{1, \ldots, N\}$ is arbitrary and can later be omitted. Z is given by

$$Z = \int_0^\infty \prod_j \mathrm{d}J_j \prod_\mu \Theta(N^{-1/2} \sum_j J_j \xi_j^\mu - \kappa) \mathrm{e}^{-\beta \sum_j J_j}. \tag{7}$$

This calculation is technically similar to the Oppers calculation of learning times of the MinOver algorithm for the unrestricted perceptron [15]. Here the Hamiltonian H of the system is simply the linear norm (4). The helping parameter β ensures in the limit $\beta \to \infty$ that H takes its minimum. The average over the ξ can be performed by using the replica approach. Introducing replicas J_j^a, $a = 1 \ldots n$, brings us to

$$g(k) = \lim_{\substack{\beta \to \infty \\ n \to 0}} \left\langle \int_0^\infty \prod_{ja} \mathrm{d}J_j^a \mathrm{e}^{-\beta \sum_{ja} J_j^a} \prod_{\mu a} \Theta(N^{-1/2} \sum_j J_j^a \xi_j^\mu - \kappa) \mathrm{e}^{\mathrm{i}kJ_l^1} \right\rangle. \tag{8}$$

The integrals are evaluated with standard techniques of replica symmetric treatment. As the unrestricted perceptron does not show any effects of replica symmetry breaking, we do not expect them to occur here, because the sign-restricted linear problem remains convex. The testing field $\mathrm{e}^{\mathrm{i}kJ_l^1}$ has no effect on the saddlepoint, and the saddlepoint equations determine the value of the order parameters $q^{ab} = N^{-1} \sum_j J_j^a J_j^b$ and its conjugates r^{ab}. Our replica symmetric ansatz needs four parameters $q^{ab} = \delta_{a,b} q^* + q$ and $r^{ab} = \delta_{a,b} r^* + r$. To perform the limit $\beta \to \infty$ we use the scaling

$$\beta q^* = v \qquad -r^*/\beta = s \qquad \frac{r^* + r}{\beta^2} = t \tag{9}$$

where v, s and t are finite. We find the following saddlepoint equations:

$$s = \frac{\alpha}{2v} \int_{-\kappa/\sqrt{q}}^{\infty} \mathrm{D}y\,(\kappa/\sqrt{q} + y)y \tag{10}$$

$$t = \frac{\alpha}{2v^2} \int_{-\kappa/\sqrt{q}}^{\infty} \mathrm{D}y\,(\kappa + y\sqrt{q})^2 \tag{11}$$

$$v = \frac{1}{2s}\left(1 - \int_{-1/\sqrt{2t}}^{\infty} \mathrm{D}y\,(1/\sqrt{2t} + y)y\right) \tag{12}$$

$$q = \frac{1}{4s^2} + \frac{t}{2s^2} - \frac{1}{4s^2} \int_{-1/\sqrt{2t}}^{\infty} \mathrm{D}y\,(1 + y\sqrt{2t})^2 \tag{13}$$

where $\mathrm{D}y = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y^2/2} \mathrm{d}y$. Integrals can be reduced to error functions and κ can be scaled out, so that the four equations can easily be solved numerically. For the free energy we find for $\beta \to \infty$

$$f = -\lim_{\substack{\beta\to\infty \\ n\to 0}} \frac{1}{\beta N n}(\langle Z^n \rangle - 1) = 2tv - 2sq\,. \tag{14}$$

The free energy is plotted in figure 2; it diverges at $\alpha = 1$. Independently the characteristic function $g(k)$ is calculated as

$$g(k) = \int_{-\infty}^{-\Delta} \mathrm{D}y\, \mathrm{e}^{-\mathrm{i}k(1+y/\Delta)/2s} + \int_{-\Delta}^{\infty} \mathrm{D}y \tag{15}$$

where $\Delta = 1/\sqrt{2t}$. The probability distribution is found by back Fourier transformation:

$$P(J) = \delta(J) \int_{-\Delta}^{\infty} \mathrm{D}y \;+\; \Theta(J)\, \mathrm{e}^{-\Delta^2(1+2sJ)^2} \tag{16}$$

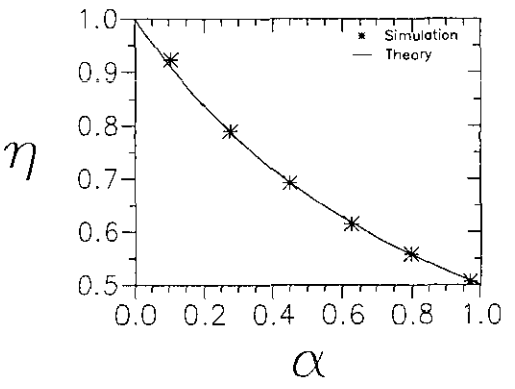

Figure 2. Free energy from the replica symmetric calculation (dimensionless). Note the divergence at $\alpha = 1$.

The first term gives the probability that a synapse takes the value zero. Comparing this probability with data from simulations produces a good agreement as shown in figure 3. The dilution rate as a function of α is plotted for both, theory and simulation. The second term gives the probability distribution of the strength of the (positive) synapses. We do not have maximum possible dilution at $0 < \alpha < 1$, because we would expect a convex shaped function for the dilution rate, rather than a concave one, to have a greater dilution than Bouten's linear function for the quenched case. In figure 4 we plot the distribution for two values of α, where we scaled κ so that $q = 1$. Curves are again given for theory and simulation and once more good agreement is observed.

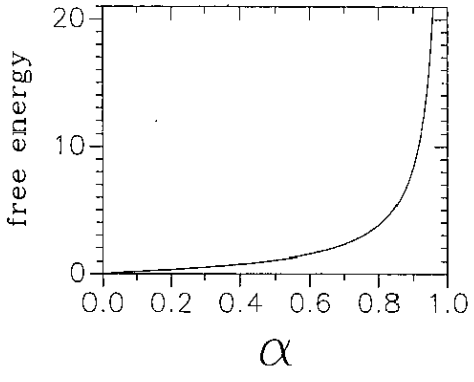

Figure 3. Dilution rate η over capacity α for the exitatory network after linear learning for theory and simulation with $N = 127$.

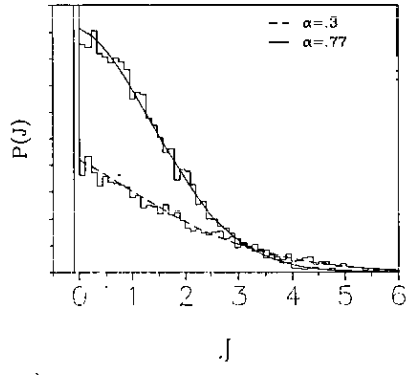

Figure 4. Probability distribution of the coupling strengths after learning, normalized to $q = 1$. Shown are curves for tho values of α for theory and simulation; the simulated data were accumulated in narrow chanels over 100 independent runs.

It is not clear which normalization is most appropiate, in terms of $\kappa = c/|\boldsymbol{J}|$, to be a good meassure for the basins of attraction, so one has to be careful when comparing stabilities. However, since quadratic stability (3) is the most common, we give in figure 5 a comparison between this quantity for our linear solution and the optimal quadratic stability for the sign-constrained device, as calculated by Amit and co-workers, which is just the result of Gardner and Derrida, with α replaced by $\alpha/2$ (broken curve). We scaled our linear stability to $q = 1$, so that it is measured in terms of (3), and necessarily less than optimal; smooth curves denote theory and asterisk

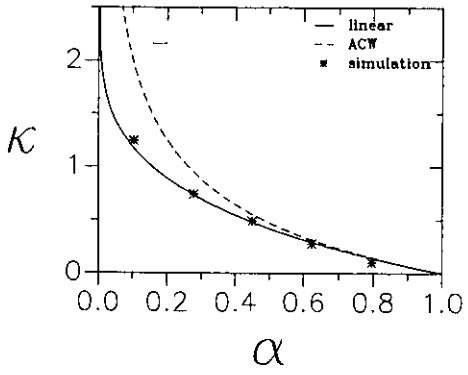

Figure 5. Quadratic stability κ: The smooth curve shows stability of linear theory scaled to $q = 1$ and asterisk symbols show the same for simulation with $N = 127$. The broken curves shows optimal quadratic stability as calculated by [10] (ACW).

symbols denote simulation data, which support the theory. Both theoretical curves are surprisingly close together, at least for α not too small, so that we expect to have comparable basins of attraction with a much smaller number of couplings.

In conclusion, we have shown that on the average the maximum capacity $\alpha_c = 1$ of a sign-constrained device, where the constraints are uncorrelated to the patterns, can be reached with linear programming. Our linear algorithm produced at $\alpha = 1$ the maximum possible dilution of synapses and reasonable dilution for smaller α, with the benefit of great stabilities.

Using methods of replica symmetric theory we calculated the nature of the infinite linear learning system of purely exitatory synapses; we found very good agreement with simulated systems of a size of the order of a hundred. Furthermore we showed that a network or a perceptron with sign constraints of clipped Hopfield can store up to appoximately $1.65N$ patterns.

We would like to thank M Opper, A Zippelius, R Kree and S Mertens for inspiring and useful discussions. This work was supported by the BRAIN initiative of the commission of the European Community. The computer time for the simulations was granted by the Forschungszentrum in Jülich, FRG.

References

[1] Dale H H 1935 *Proc. R. Soc. Med.* **28** 319
[2] Amit D J, Wong K Y M and Campbell C 1989 *J. Phys. A: Math. Gen.* **22** 2039
[3] Rosenblatt F 1958 *Psychoanalytic Rev.* **65** 386
Minsky M and Papert S 1988 *Perceptrons* (Cambridge, MA: MIT Press)
[4] Kinzel W and Opper M 1991 *Physics of Neural Networks* ed J L von Hemmen, E Domany and K Schulten (Berlin: Springer)
[5] Kepler T B and Abbott L F 1988 *J. Physique* **49** 1657
[6] Widrow B and Hoff M 1960 *WESCON Convention, Report IV* p 96
[7] Anlauf J K and Biehl M 1989 *Europhys. Lett.* **10** (7) 687
Anlauf J K and Biehl M 1990 *Europhys. Lett.* **11** (4) 387
[8] Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* **20** L745
[9] Cover T M 1965 Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition *IEEE Trans.* **EC-14** 326
[10] Amit D J, Campbell C and Wong K Y M 1989 *J. Phys. A: Math. Gen.* **22** 4687

[11] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257; 1987 *Europhys. Lett.* **4** 481
Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[12] Nadal J-P 1990 On the storage capacity with sign constrained synaptic couplings *Preprint* Ecole *Normale Superieure*
[13] M Bouten, A Engel, A Komoda and R Serneels 1990 Quenched versus annealed dilution in neural networks 1990 *J. Phys. A: Math. Gen.* **23** 4643–57
[14] Klee V and G J Minty G J 1972 *Inequalities III* ed O Shisha (New York: Academic) pp 159–75
[15] Opper M 1988 Learning times of neural networks: exact solution for a PERCEPTRON algorithm *Phys. Rev.* A **38** 3824