Perceptron learning with sign-constrained weights

J. Phys. A: Math. Gen. **22** (1989) 2039–2045. Printed in the UK

Perceptron learning with sign-constrained weights

Daniel J Amit†§, K Y M Wong† and C Campbell‡

† Department of Physics, Imperial College, London SW7 2BZ, UK

‡ Department of Applied Physics, Kingston Polytechnic, Kingston-upon-Thames KT1 2EE, UK

Received 20 April 1989

Abstract. We study neural network models in which the synaptic efficacies are restricted to have a prescribed set of signs. It is proved that such neural networks can learn a set of random patterns by a perceptron-like algorithm which respects the synaptic restrictions at every step. In particular, it shows that learning can take place iteratively in a network which obeys Dale's rule, i.e. in which neurons are exclusively excitatory or inhibitory. The learning algorithm as well as its convergence theorem are stated in perceptron language and it is proved that the algorithm converges under the same conditions as required for an unconstrained perceptron. Numerical experiments show that these necessary conditions can actually be met for relatively large sets of patterns to be learned. We then argue that the results are invariant under the distribution of the signs, due to gauge invariance for random patterns. As a consequence the same sets of random patterns can be learned by networks which have any fixed distribution of synaptic signs, ranging from fully inhibitory to fully excitatory.

1. Introduction

The Edinburgh group has brought perceptron theory [1, 2] into a very fruitful relationship with attractor neural networks [3, 4]. It started with the adaptation of the perceptron learning algorithm [1, 2] as a technique for storing a set of patterns as attractors in a neural network [5]. It culminated with the seminal work of Elizabeth Gardner [6, 7], which established:

(i) the number of random patterns that can be stored as a function of the correlations between the patterns, as well as of the stability parameter of each of the attractors;

(ii) that the perceptron learning convergence theorem can be extended to cases in which stability parameters are imposed to enlarge the basins of attraction of the embedded attractors.

Perceptron theory also stood to gain from these studies. Gardner's shift of attention to typical random patterns permitted specific statements as to the conditions for which the perceptron can find a solution. Such knowledge about existence underlies the learning convergence theorem, as well as the rate at which learning can converge. The rate of convergence depends just on the largest stability parameter for which a solution exists. As far as random patterns are concerned, existence of a solution depends essentially on the number of patterns to be taught. The knowledge of the maximal

§ On leave from the Racah Institute of Physics, Hebrew University, Jerusalem.

0305-4470/89/122039+07$02.50

number of patterns that can be learned, as a function of the stability parameter, gives at one and the same time both the knowledge of the convergence process and an upper bound on the number of correction steps.

If learning algorithms and storage estimates of this type are to be candidates for modelling properties of biological networks, then they will have to conform with at least one constraint, namely that neurons obey in general Dale's rule [8]. This rule stipulates that the synapses emanating from a given neuron are all of the same type, either all excitatory or all inhibitory. In other words, a neural network that is to learn by a perceptron algorithm must be able to do so by modifying synaptic efficacies but keeping fixed the signs of all those which come out of a given neuron. In the perceptron context this biological constraint implies that each of the N weights, leading from the N input elements to the linear threshold output element, will have a pre-assigned sign.

To be more specific, suppose that a neural network is specified by giving the neural activity states $S_i = (\pm 1)$ of N neurons and the synaptic efficacies J_{ij} for $i \neq j$ connecting each pair of neurons. The efficacy J_{ij} measures the amount of post-synaptic potential (local field) induced in neuron i by neuron j. The dynamics, in the absence of noise is

$$S_i(t + \delta t) = \operatorname{sgn}\left(\sum_{j=1}^{N} J_{ij} S_j(t)\right)$$

where we have taken for simplicity zero thresholds. Extensions to finite threshold have been discussed by Gardner [6]. A fixed-point attractor state of this dynamical process has to satisfy

$$\sum_{j=1}^{N} J_{ij} S_j S_i > 0. \qquad \text{S\_i and S\_j have same sign for cumulative majority of weights} \quad (1)$$

For such a network, to 'learn' a set of patterns is to modify the synaptic weights J_{ij} in such a way as to reach a set which will have all the patterns in the set as fixed points. Specifically, the patterns to be learned are a set of p N-bit words, $\xi_i^\mu = \pm 1$, $i = 1, \ldots, N; \mu = 1, \ldots, p$. The network starts with a set J_{ij}^0 and then one pattern at a time is presented to the network, by substituting $S_i = \xi_i^\nu$ on the left-hand side of equation (1). If the inequality is not satisfied the J_{ij} are modified by writing

$$J_{ij}^{(k+1)} = J_{ij}^{(k)} + \Delta_{ij}^{(k)} \qquad (2)$$

where $J_{ij}^{(k)}$ is the set of couplings following k learning passes. The specification of the form of $\Delta_{ij}^{(k)}$ completes the definition of the learning algorithm.

Since in the above description different output neurons are independent, it is sufficient to consider just one. This reduces the formulation to that of the perceptron. Here the learning can be defined as follows: there is a set of N input variables S_i and a set of weights A_i. To learn a set of p patterns $\Phi_i^\mu = \pm 1$ is to modify the A_i recursively, so as to reach a set of A_i which will satisfy

$$\sum_{j=1}^{N} A_j \Phi_j^\mu > 0$$

for all $\mu = 1,\ldots,p$. The patterns are tested as to whether they satisfy the inequality. If some pattern does not, the weights are modified. Following the kth error in the sequence of presentations of the patterns, the weights have the values $A_i^{(k)}$. Upon the detection of the next error, the ith weight is modified by the addition of $\Delta_i^{(k)}$.

The evaluation of the maximal storage capacity of perceptrons, and hence of neural networks, has been obtained [9, 10, 6] by inspection of the *full* space of weights (or, equivalently, synaptic efficacies or coupling constants). The same can be said about the learning algorithm. In what follows we will constrain every synaptic weight to have a given sign. This requirement can be stated formally by choosing a fixed set $g_i = \pm 1$ $(i = 1,\ldots,N)$, which classifies our neurons as inhibitory or excitatory, and demanding that a solution will satisfy $A_i^{(k)} g_i > 0$, for all i at all stages k.

There are two contexts in which the Dale constraint has been implemented in neural networks. In the Hopfield model [3], which has a random distribution of synaptic efficacies, the synaptic matrix has been diluted by eliminating synapses until it conforms with the rule [11]. The network then continues to function quite effectively as an associative memory. The second context is that of Willshaw's model [12, 13]. In this model a set of sparsely coded patterns can be stored in a purely excitatory network. While this is a very effective model it is different in spirit from perceptron learning in that the sparse coding eliminates interference between patterns and the encoding is, therefore, not a process of error detection and synaptic modification, but rather a direct writing of the synapses according to the patterns presented until errors start to appear.

Here we address the question as to whether there exists a learning algorithm of automatic modification of synapses, which respects Dale's rule, and brings about the embedding of a set of random patterns as attractors. Stated in perceptron language the question will be about the existence of a learning algorithm, respecting a prescribed distribution of signs for the weights, which can classify a set of random patterns into two classes, one giving the target output $+1$ and the other -1. Given such an algorithm, one would like to know the conditions for its convergence and the convergence rate. The other part of Gardner's program, that of the conditions for the existence of a solution and correspondingly of the storage capacity, will be described elsewhere [14].

One may consider different learning algorithms which respect Dale's rule and which converge under the same conditions as those for an unconstrained perceptron, namely:

if a solution to the classification problem exists, then the algorithm will converge to a solution after having made a number of corrections at most polynomial in N.

2. Description of learning algorithms

Consider, for example, the following algorithm which automatically respects the signs of all the weights. Encountering an error upon the presentation of pattern ν, i.e.

$$\sum_{i=1}^{N} A_i^{(k)} \phi_i^{\nu} < 0 \tag{3}$$

the weights are modified according to

$$\Delta_i^{(k)} = \phi_i^{\nu} \theta[A_i^{(k)}(A_i^{(k)} + \phi_i^{\nu})] \tag{4}$$

where $\phi_i = \Phi_i/\sqrt{N} = \pm 1/\sqrt{N}$, and the function θ is 1 for positive argument and zero otherwise. In this process the weight is modified as in the usual perceptron, unless the modification would lead to a change in the sign of that weight. In the latter case the weight remains unmodified. Clearly, if the initial conditions are such that $A_i^{(0)} g_i > 0$, then each weight will preserve the sign of the corresponding g_i throughout the process. Note that the requirement on the initial A is quite plausible, since the excitatory or inhibitory nature of a synapse is not learned.

Another possibility would be a variation on the previous process, namely if the modification of a weight would lead to a change of sign, that weight is set to zero. Once set to zero, the weight can be modified only if the modification respects the prescribed sign.

3. Convergence of learning algorithms

The proof follows the canonical perceptron path. One assumes the existence of a solution, somewhat stronger than that being learned, i.e. one assumes that there exists a set A_i^* satisfying

$$|A^*|^2 \equiv \sum_{i=1}^{N} (A_i^*)^2 = 1 \tag{5}$$
$$A_i^* g_i > 0$$

and, for all $\mu = 1, \ldots, p$

$$\sum_{i=1}^{N} A_i^* \phi_i^\mu > \delta > 0. \tag{6}$$

Then, one monitors the angle cosine between the presumed solution, A^*, and the consecutive learned solutions, $A^{(k)}$, i.e.

$$G^k \equiv \frac{A^* \cdot A^{(k)}}{|A^{(k)}|}.$$

Since A^* is normalised $G^k \leq 1$. The proof consists of demonstrating that the numerator in G^k increases faster than the denominator, when corrections are being made.

Consider the first algorithm. If upon the presentation of the pattern ϕ^ν, following k previous correction steps, there is an error, i.e. inequality (3) holds, then a correction has to be made according to equation (4).

The change in the numerator of G will be

$$\delta N^k = \sum_{i=1}^{N} A_i^* \phi_i^\nu \theta[A_i^{(k)}(A_i^{(k)} + \phi_i^\nu)]$$
$$= \sum_{i=1}^{N} A_i^* \phi_i^\nu + \sum_{i=1}^{N} A_i^* \phi_i^\nu \{\theta[A_i^{(k)}(A_i^{(k)} + \phi_i^\nu)] - 1\}.$$

The first term in the last line is bounded from below by δ, as was assumed in (6). The second term in this line is positive. To see this, note that $A_i^{(k)}$ has the same sign as

A_i^*. When the curly brackets in each term in the sum is non-zero it is negative, e.g. -1 and it is non-zero only if $A_i^{(k)}$ and hence also A_i^* are of opposite sign to ϕ_i^ν. Each non-vanishing term in the second sum is, therefore, a product of two negative factors. As a result

$$\delta N^k > \delta$$

and the numerator grows at least linearly with k.

As far as the denominator is concerned, the change in its square is

$$\begin{aligned}\delta(D^k)^2 &= 2\sum_{i=1}^{N} A_i^{(k)}\Delta_j^{(k)} + \sum_{i=1}^{N}(\Delta_j^{(k)})^2 \\ &= 2\sum_{i=1}^{N} A_i^{(k)}\phi_i^\nu\theta[A_i^{(k)}(A_i^{(k)} + \phi_i^\nu)] + \sum_{i=1}^{N}(\Delta_j^{(k)})^2.\end{aligned}$$

Due to the normalisation of the patterns ϕ, the second term is bounded from above by 1, for all k. The first term can be written as

$$2\sum_{i=1}^{N} A_i^{(k)}\phi_i^\nu + 2\sum_{i=1}^{N} A_i^{(k)}\phi_i^\nu\{\theta[A_i^{(k)}(A_i^{(k)} + \phi_i^\nu)] - 1\}.$$

The first term in the above expression is negative, because we have had an error, i.e. because of equation (3). The second term is positive, but has an upper bound independent of k. To see this, note that as before the term in the curly brackets vanishes unless ϕ_i is of opposite sign to $A_i^{(k)}$ *and* is greater in absolute value. Hence, the above expression can be bounded from above by

$$2\sum_{i=1}^{N}(\phi_i^\nu)^2 < 2.$$

All of this leads to the conclusion that the square of the denominator grows at most linearly with k. The denominator itself, therefore, grows at most like $\sqrt{k}$ and the algorithm must converge, much as in the original proof[2]. This proof can be extended to the second algorithm without any special difficulty.

4. Discussion

Clearly, the above theorem displaces the whole question to the availability of the sufficient conditions for the convergence, i.e. whether a solution of the type of inequalities (5) exists for a given set of patterns. This is answered by the other aspect of Gardner's approach and here will be deferred to a subsequent report [14]. Nevertheless, numerical experiments show that such solutions exist for random patterns, up to at least $p = N/2$, and that convergence is slower than in the unconstrained case. The first algorithm is slower than the second.

The existence of solutions with a prescribed set of synaptic signs has rather surprising implications. It implies that the perceptron, and hence the network, can

learn a set of *random* patterns of ± 1 given any distribution of the signs, g_i, on the connections. The question of the distribution of these (quenched) variables has not entered the discussion of the convergence theorem and is implicit in the assumption of the existence of A^*. But if a solution A^* exists for one set of g_i, then changing the signs of *any* subset of them can be compensated by the same change in the corresponding components of all the patterns. The new patterns will still be words of random ± 1. The corresponding change in the components of A^* produces a solution with the new distribution of signs for the new set of patterns. Thus, the probability of finding a solution A^* for a set of p random patterns is independent of the particular realisation of the set of signs g_i.

In the thermodynamic limit ($N \to \infty$) the implications are even more striking. If there is a non-vanishing probability for a solution of inequalities (5) for a given set of signs, g_i, in the space of all possible sets of p random patterns, then this probability approaches unity as $N \to \infty$. Since this probability is unchanged by the change of signs of a subset of the g_i, one finds that there is also a solution with the new set of signs for almost every set of p patterns. Consequently, given a fixed (quenched) set of p random patterns, if a solution A^* exists for a particular realisation of the g_i, then the probability that a solution B^* exists for *any* other realisation of the signs approaches unity as $N \to \infty$. This is a *local gauge invariance* of the theory. The search for a solution, however, is constrained to a subspace of the total space of couplings, a subspace smaller by a factor of 2^{-N} and the volume of the subspace of solutions is reduced by at least this same factor.

In particular, changing the distribution includes the possibility of changing the relative number of $+1$ and -1, even reaching the extreme of having the signs of all the weights positive. In other words, this innocuous result implies that one can comfortably store an arbitrary set of random patterns, even in a neural network that is purely excitatory, up to the same storage limit as for any other network which has a prescribed distribution of signs (g_i). Of course, one has to carefully study the basins of attraction, and especially how to avoid falling too often into the ferromagnetic states with all neurons active at maximum rate or with all neurons quiescent. The possibility of complying with a pre-assigned distribution of the local fields [15, 16, 17] or of improving the convergence rate of the learning algorithm [18] should be the subject of further study.

Acknowledgments

We would like to acknowledge useful discussions with Professor D Sherrington. DJA is indebted to the SERC for a fellowship which has made his stay at Imperial College possible and to J Taylor for discussions.

References

[1] Rosenblatt F 1962 *Principles of Neurodynamics* (New York: Spartan)
[2] Minsky M L and Papert S 1969 *Perceptrons* (Cambridge, MA: MIT Press)
[3] Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554
[4] Amit D J 1987 *Heidelberg Colloq. on Glassy Dynamics* ed L B Van Hemmen and I Morgestern (Berlin: Springer); *Modeling Brain Function* (Cambridge: Cambridge University Press)

[5] Wallace D J 1985 *Advances in Lattice Gauge Theory* ed D W Duke and J F Owens (Singapore: World Scientific)
[6] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[7] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[8] Eccles J C 1964 *Physiology of Synapses* (Berlin: Springer)
[9] Cover T M 1965 *IEEE Trans. Electron. Comput.* **EC-14** 965
[10] Venkatesh S 1986 *Proc. Conf. on Neural Networks for Computing, Snowbird, UT (AIP Conf. Proc.* **151**) ed J S Denker (New York: American Institute of Physics)
[11] Shinomoto S 1987 *Biol. Cybern.* **57** 197
[12] Willshaw D J, Buneman O P and Longuet-Higgins H C 1969 *Nature* **222** 960
[13] Rubin N and Sompolinsky H 1989 Neural networks with low local firing rates *Preprint* Racah Institute, Jerusalem
[14] Amit D J, Campbell C and Wong K Y M 1989 The interaction space of neural networks with sign-constrained synapses, in preparation
[15] Abbott L F and Kepler T B 1989 *J. Phys. A: Math. Gen.* **22** 2031
[16] Krauth W, Mézard M and Nadal J-P 1989 Basins of attraction in a perceptron-like neural network *Preprint* LPTENS 88/8
[17] Kepler T B and Abbott L F 1988 *J. Physique* **49** 1657
[18] Abbott L F and Kepler T B 1988 Optimal learning in neural network memories *Preprint* Brandeis University BRX-TH-255