

High Performance Associative Memory and Sign Constraints

NEIL DAVEY, ROD ADAMS
Department of Computer Science
University of Hertfordshire
College Lane, Hatfield, AL10 9AB
UK

N.Davey@herts.ac.uk <http://www.homepages.feis.herts.ac.uk/~neil>
R.G.Adams@herts.ac.uk <http://www.homepages.feis.herts.ac.uk/~comqrga>

Abstract: - The consequences of imposing a sign constraint on the standard Hopfield architecture associative memory model, trained using perceptron like learning rules, is examined. Such learning rules have been shown to have capacity of at most half of their unconstrained versions. This paper reports experimental investigations into the consequences of constraining the sign of the network weights in terms of: capacity, training times and size of basins of attraction. It is concluded that the capacity is roughly half the theoretical maximum, the training times are much increased and that the attractor basins are significantly reduced in size.

Key-Words: - Associative Memory, Hopfield Networks, Sign Constraints, Capacity, Basins of Attraction.

1 Introduction

Neural networks designed to function as associative memories are usually based around the standard Hopfield architecture. It has been known for some time [1] that a variety of local learning rules can produce models with much better performance than the original Hebbian learning, proposed by Hopfield. These learning rules either act as approximators to the projection weight matrix or use perceptron style learning. It is also thought that networks that purport to biological plausibility should adhere to Dale's law [5], which suggests that neurons either make exclusively excitatory, or inhibitory, connections. In all three types of learning rules, mentioned above, the weights in the resulting network have no such restriction. However, it has been demonstrated that for the perceptron type learning rules, it is possible to constrain the signs of the weights, so that they adhere to Dale's law, whilst still producing convergence on suitable training data. In this paper we examine the performance of such learning rules in terms of: their capacity, learning time and capabilities as effective associative memories.

2 Models Examined

In this section we take a set, Π , of N -ary, bipolar (+1/-1) training vectors, $\{\xi^p\}$. The N by N weight matrix is denoted by \mathbf{W} , and the state of the i 'th unit is denoted by S_i . The loading of the network is defined as:

$$\alpha = \frac{|\Pi|}{N}.$$

All the high capacity models studied here are modifications to the standard Hopfield network. The net input, or *local field*, of a unit, is given by:

$$h_i = \sum_{j \neq i} w_{ij} S_j$$

The next state of the unit is then given by:

$$S_i(t+1) = \begin{cases} 1 & \text{if } h_i > \theta_i \\ -1 & \text{if } h_i < \theta_i \\ S_i(t) & \text{if } h_i = \theta_i \end{cases}$$

where the threshold, θ_i , is normally taken as zero. The update can be synchronous or asynchronous. Here we use asynchronous, random order updates. These network dynamics and a symmetric weight matrix guarantee simple point attractors in the phase space.

If the *aligned local fields*, $h_i \xi_i$, are all non-negative then the pattern ξ , will be stable in the network.

2.1 Perceptron Style Learning

In the late 1980s it was demonstrated that perceptron like learning could be applied to associative memory networks to produce much higher capacity than the basic model. In fact as Gardner [6] showed a Hopfield type network of N units could store up to $2N$ uncorrelated patterns,

with this optimal capacity increasing for correlated patterns. Learning rules of this type are designed to drive the aligned local fields of patterns in the training set over a threshold value, T . As shown above, a necessary and sufficient condition for the training patterns to be learnt is that T is non-negative, and often, for ease of training, a value of 1 (or even 0) is taken. Nevertheless increasing T may improve the attractor performance of the network [1]. Some care must be taken though, since if we consider a network in which all the training patterns are stable, that is $h_i \xi_i > T$ for all patterns and units, i , then any uniform, upward scaling of the weight matrix will increase the aligned local fields, but will obviously not increase the attractor performance. In fact the optimal attractor performance is achieved when the threshold is maximised with respect to the size of the weights. For this reason the relevant characterization is the *normalised stability measure*, defined as: $\gamma_i = \frac{h_i \xi_i}{|W_i|}$ where W_i is the incoming weight vector to unit i . The minimum of all the γ_i therefore gives a measure of the likely attractor performance and we take $\kappa = \min_{p,i}(\gamma_i^p)$.

2.1.1 Local Learning (LL)

Diederich and Oppen's [4] local learning rule is an iterative learning rule in which the local fields for each training pattern are driven to the correct side of $+T$ or $-T$ as appropriate. This is equivalent to the condition that:

$$\forall i, p \bullet h_i^p \xi_i^p \geq T$$

So the learning rule is given by:

Begin with a zero weight matrix

Repeat until all local fields are correct

Set the state of network to one of the ξ^p

For each unit, i , in turn

Calculate $h_i^p \xi_i^p$. If this is less than T then change the weights to unit i according to:

$$\Delta w_{ij} = \frac{\xi_i^p \xi_j^p}{N}$$

This is the perceptron learning rule with a fixed margin of T and a learning rate of $\frac{1}{N}$. The process will converge on a suitable weight matrix if one exists [4], at which point the trained patterns are guaranteed to be stable.

As shown by Abbott [1], this leads to a network in which $\kappa \geq \frac{T}{2T+1} \kappa_{\max}$, where κ_{\max} is the optimal value of κ . From this it is apparent that increasing T will in turn increase the lower bound of κ , and this may give better attractor performance.

2.1.2 Krauth and Mezard Local Learning (KM)

A modification to this learning rule proposed by Krauth and Mezard [8] can be shown to produce a κ value that tends towards κ_{\max} as T increases. In this version the patterns are not presented to the network in an arbitrary order. Instead the pattern that has the smallest aligned local field is chosen as the one for next presentation.

Begin with a zero weight matrix

Repeat until all local fields are correct

For each unit, i , in turn

Select the pattern, ξ^p with lowest aligned local field at this unit and update the incoming weights according to:

$$\Delta w_{ij} = \frac{\xi_i^p \xi_j^p}{N}$$

2.2 Sign Constraints

A possible difficulty with the normal perceptron learning rule is that weights can (and do) change sign during the learning process. The biological equivalent of this would be for a synapse to change from excitatory to inhibitory or visa versa. This is not thought to happen, and indeed Dale's rule [5] states that all synapses at a given neuron are all either excitatory or inhibitory. For a neural network this is equivalent to requiring that all incoming weights to a given unit have the same sign, and this cannot change over time.

The effect of imposing such a constraint on a Hopfield network was first investigated in 1986 [9] where it was shown that the capacity only falls from $\alpha = 0.14$ to $\alpha = 0.09$, for uncorrelated patterns. Later Amit et al. [2] showed that the perceptron learning rule could also be effective under such a constraint. They showed that the theoretical maximum capacity of a sign constrained network was exactly half that of the unconstrained version. This is a surprising result as the volume of weight space that the network may use is reduced by a much higher proportion. They also showed that this capacity is independent of the particular sign constraint used. In particular, a network of units using only

excitatory (or inhibitory) connections could store up to N uncorrelated patterns.

They also suggest how a learning rule based on standard perceptron learning can be modified to comply with a particular sign constraint. The idea is straightforward: whenever a weight change is proposed that will result in a violation of the sign constraint, the change is not made. A variant of this is to zero such a violating weight.

In the experimental work reported here we use, without loss of generality, networks, with only excitatory weights, and investigate the efficacy of the learning rule and the resulting networks. Specifically, we start with weight matrices with weights randomly initialised to lie between 0 and 1 and use learning rules, that in the case of Signed-LL, can be formally stated as:

Repeat until all local fields are correct

Set the state of network to one of the ξ^p

For each unit, i , in turn

Calculate $h_i^p \xi_i^p$.

If this is less than T then change the weights to unit i according to:

$$w'_{ij} = w_{ij} + \frac{\xi_i^p \xi_j^p}{N}$$

whenever the resulting weight is positive

The variant of this, mentioned above, is to use

$$w'_{ij} = \max\left(w_{ij} + \frac{\xi_i^p \xi_j^p}{N}, 0\right)$$

and we will denote this variant as Signed-LL-Zero. Note that this form of learning can be used in any variant of perceptron learning, so that signed KM is straightforwardly derived from the KM algorithm.

As is well known, normal perceptron learning will converge on a solution, if one exists, since the weight changes always move the weight vectors towards ones that embed the training vectors [1]. However, with sign constrained learning, a subtle difference emerges: when a weight change takes place it is constructive, but a weight change will not take place when it results in a violation of the sign constraint. Therefore, learning can reach a local impasse, in which a suitable weight matrix exists but the learning rule is unable to find it, without passing through a point in the weight space that is prohibited. This may result in a reduction in capacity with respect to the theoretical maximum

and this aspect of the constrained learning is investigated in the next sections.

3 Analysing Performance

An effective associative memory model is expected, not only to have the training patterns as fixed points of the network dynamics, but also that these fixed points should act as attractors in the state space.

As stated above the perceptron type learning rules will have the training vectors embedded, when the aligned local fields have been driven to be non-negative. Moreover the larger these aligned local fields become, the better the attractor performance should be. Therefore we examine the performance of the signed constrained networks by varying both the loading, α , and the training threshold T .

We also consider the effect of correlations in the training patterns. An uncorrelated training set is one in which the patterns are completely random. Correlation can be increased by varying the probability that a given bit in a training pattern is ± 1 . If the probability of any bit being $+1$ in the training set is the bias, b , $\forall i, p \cdot \text{prob}(\xi_i^p = +1) = b$, then a bias of 0.5 corresponds to an uncorrelated training set and a bias of 1 to a completely correlated one.

To estimate the size of the attractor basins of the learned patterns R , the normalized mean radius of the basins of attraction [3] is measured. It is defined as $R = \left\langle \left\langle \frac{1-m_0}{1-m_1} \right\rangle \right\rangle$ where m_0 is the minimum overlap

an initial state must have with a stored pattern for the network to return the state to the training pattern, and m_1 is the largest overlap of the state with the rest of the training patterns. The angled braces denote an average over all training patterns. Details of the algorithm used can be found in [3].

The training time of the local learning rules is reported as the number of epochs, complete presentations of the training set, needed for convergence.

4 Results

4.1 Capacity

To measure capacity we successively trained the networks on training sets of gradually increasing size. At each loading increment, the learning rule

had to embed five different training sets. The last loading point at which a learning rule succeeded was designated as its maximum loading. Note that this calculation of capacity is different from the normal one applied to the standard Hopfield one-shot learning rule, which is obviously guaranteed to converge, but may do so without embedding the training patterns.

The capacity of the first version of local learning, with 100 units, was found at differing biases, and the results can be seen in Figure 1. The first point that is apparent from this result is that the capacity of unbiased patterns ($b = 0.5$) is 55 patterns, much smaller than the theoretical prediction of 100. Secondly it can be seen that the capacity falls off rapidly as the patterns become correlated. For example with patterns with a bias of $b = 0.7$, the capacity has fallen to 10 patterns and at $b = 0.8$ it is just 2 patterns. Such a decrease is, of course, characteristic of one-shot Hebbian learning, but not of normal perceptron learning.

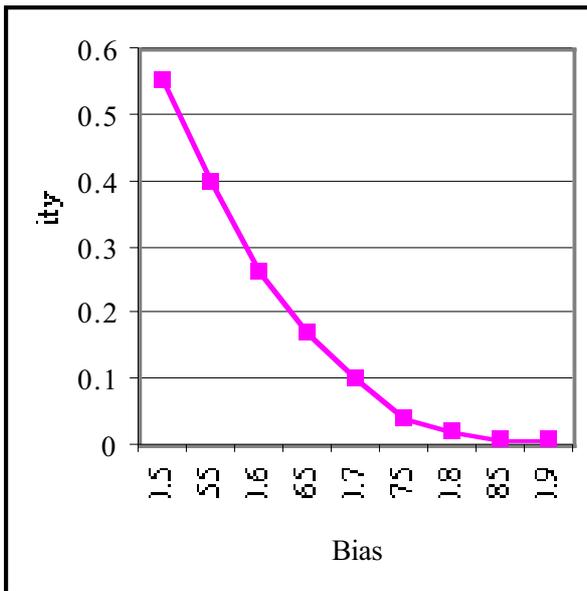


Figure 1: The capacity, the maximum loading, of a one hundred unit network trained with patterns sets of varying bias.

The second experiment repeated this analysis for the other two learning rules, Signed-LL-Zero and Signed-KM. No significant difference in capacity was found. This suggests that these capacity results are not simply an artifact of a particular learning rule but are a reflection of a deeper characteristic of the problem.

4.2 Performance

The next set of results compares the unconstrained versions of LL and KM with the sign constrained modifications. In all cases the loading of the (100 unit) network is $\alpha = 0.3$, the patterns are unbiased and the results are averages over fifty runs. Three different learning thresholds (T) are used.

Table 1 shows the attractor performance of the Signed-LL and Signed-KM. It can be seen that both versions have very similar performance, with the optimal KM rule not showing a noticeably better performance. The increase in T brings about the expected increase in minimum *normalised stability measure*, κ . The R values, though, are very small, suggesting that, on average, no more than between 1 and 5 bits of a 100 bit training pattern will be corrected. The explanation for this may be that, a strong attractor in a signed network is the pattern that conforms to the sign constraint exactly, in the case here the pattern consisting of +1's. This has the potential advantage of being an obvious parasitic attractor, but can also interfere with the basins of attraction of the trained patterns.

Sign Constrained	T	κ	R
LL	1	0.34	0.01
LL	10	0.47	0.04
LL	100	0.49	0.05
KM	1	0.38	0.03
KM	10	0.49	0.04
KM	100	0.50	0.04

Table 1: The performance of the signed LL and KM learning rules, under a loading of 0.3 (30 patterns in a 100 unit network). The patterns are unbiased and results are averages over 50 runs.

Table 2 shows the equivalent results for the unconstrained versions of these two rules. Once again LL and KM give similar results. In comparison with the sign constrained networks the κ values are much higher here (the theoretical maximum value of κ at this loading is 1.27), and as a result the attractor performance, as measured by R , is far better.

Unconstrained	T	κ	R
LL	1	0.84	0.57
LL	10	1.14	0.64
LL	100	1.18	0.63
KM	1	0.87	0.57
KM	10	1.19	0.66
KM	100	1.23	0.64

Table 2: The performance of the unconstrained LL and KM learning rules, under a loading of 0.3 (30 patterns in a 100 unit network). The patterns are unbiased and results are averages over 50 runs.

The training time of the constrained and free versions of LL are shown in Table 3. As the KM rule does not follow a simple epoch by epoch approach its training epochs are not reported. It can be seen that the signed network requires about seven times as many epochs to obtain a suitable weight matrix, which suggests that the constrained learning task is much harder.

T	Unconstrained	Sign Constrained
1	7.7	53.4
10	54.8	358.7
100	500.6	3425.5

Table 3: The number of epochs required to embed 30 patterns in a 100 unit network by the unconstrained and signed LL learning rule. Averages over 50 runs are reported.

5 Conclusion

The imposition of a sign constraint on a fully connected set of perceptrons is interesting from the perspective of biological plausibility. Theoretically it has been shown that this only halves the maximum capacity of such a network, when trained with uncorrelated patterns. Here we have empirically investigated how perceptron style learning rules perform under the sign constraint. The first important result is that the actual capacity is significantly less than the theoretical maximum, just over $N/2$ rather than N . Moreover this result is robust to modifications in the learning rule, with all three variants of perceptron learning producing the same result. The reason for this could be that the theoretical maximum is only obtainable for very large N , or that, as mentioned earlier, the learning rules reach local impasses, with respect to the sign

constraint, that prevents them from reaching the optimal set of weights.

As the training patterns become more correlated the performance of the constrained learning rules falls off drastically, a phenomenon worthy of further investigation.

The performance of the constrained networks as associative memories, their ability to embed training patterns in wide attractors, is shown to be very poor, particularly in comparison with the unconstrained version. This is perhaps the most significant result in this paper and raises questions about the validity of these networks as biologically plausible models.

References:

- [1] Abbott, L. F. (1990). Learning in Neural Network Memories. *Network: Comp. Neural Sys.* 1:105-122
- [2] Amit, D. J., K. Y. M. Wong and C. Campbell (1989). Perceptron Learning with sign-constrained weights. *Journal of Physics A* **22**, 2039-2045.
- [3] Davey, N, and Hunt S. P. (2000). A Comparative Analysis of High Performance Associative Memory Models. *Proceedings of 2nd International ICSC Symposium on Neural Computation (NC'2000)*
- [4] Diederich, S. and M. Opper (1987). Learning of Correlated Patterns in Spin-Glass Networks by Local Learning Rules. *Physical Review Letters* **58**, 949-952
- [5] Dale, H. H. (1935). Pharmacology and nerve-endings, *Proceedings of R. Soc. Med.* **28**, 319-332
- [6] Gardner, E. (1988). The space of interactions in neural network models, *J. Phys. A* **21**, 257-270;
- [7] Gardner, E., H. Gutfreund and I. Yekutieli (1989). The Phase Space of Interactions in Neural Networks with definite Symmetry, *J. Phys. A* **22**.
- [8] Krauth, W. and Mezard, M. (1987), Learning algorithms with optimal stability for neural networks, *J.Phys. A* **20** L745-L752.
- [9] Sompolinsky, H. (1986), Neural Networks with nonlinear synapses and a static noise, *Physics Review A* **34**, L519-L523.