Sign Constrained High Capacity Associative Memory Models

N.Davey, R.G.Adams

Department of Computer Science, University of Hertfordshire,
College Lane, Hatfield, AL10 9AB. United Kingdom
N.Davey@herts.ac.uk, R.G.Adams@herts.ac.uk

Abstract Biological neural networks do not allow the synapses to choose their own sign: excitatory or inhibitory. The consequences of imposing such a sign-constraint on the weights of the standard Hopfield associative memory architecture, trained using perceptron like learning, are examined in this paper. The capacity and attractor performance of these networks is empirically investigated, with sign-constraints of varying correlation and training sets of varying correlation. It is found that the specific correlation of the signs affects both the capacity and attractor performance in a significant way.

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 (Abbott, 1990) that a variety of local learning rules can produce models with much better performance than the original rule proposed by Hopfield. It is also thought that networks that purport to biological plausibility should adhere to Dale's law (Dale, 1935), which suggests that neurons either make exclusively excitatory, or inhibitory, connections. In all the learning rules, mentioned above, the weights in the resulting network have no such restriction. However, it has been demonstrated that for 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 and capabilities as effective associative memories, in relation to random data with varying degrees of correlation

2 Network Dynamics

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: $h_i = \sum_{j \neq i} w_{ij} S_j$ where w_{ij} is the weight on the connection from unit *j* to unit *i*. The dynamics of the network is given by: the standard update: $S_i' = \Theta(h_i)$, where Θ is the heavyside function. Unit states may be updated synchronously or asynchronously. Here we use asynchronous, random order updates. A *symmetric* weight matrix and asynchronous updates ensures that the network will evolve to a fixed point. If a training pattern ξ^{μ} is one of these fixed points then it is successfully stored. A network state is stable if, and only if, all the local fields are of the same sign as the corresponding unit, equivalently the *aligned local fields*, $h_i S_i$, should be positive.

3 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 (Dale, 1935) states that all the efferent synapses from a given neuron are all either excitatory or inhibitory. For a neural network this is equivalent to requiring that all outgoing weights from a given unit have the same sign, and this cannot change over time. There are now known to be exceptions to this picture, so that, for example, the sign of the synapse may be determined by properties of the post-synaptic cell (Wong and Campbell, 1992). A general sign constraint mechanism consists of a matrix of signs, $g_{ij} = \pm 1$, corresponding to each weight in the network, together with requirement that: $g_{ij} w_{ij} > 0$. The *sign-bias* of these weights is the ratio of positive to negative weights.

The effect of imposing a sign constraint on every connection in a standard Hopfield network was first investigated in 1986 (Sompolinsky, 1986) where it was shown that the capacity (the ratio of the maximum number of random patterns that the network can store to number of units in the network) only falls from $\alpha = 0.14$ to $\alpha = 0.09$, for uncorrelated patterns. Later Amit et al. (Amit et al., 1989b) showed that the perceptron learning rule could also be effective under such a constraint. They also showed that the theoretical maximum capacity of a sign constrained network was exactly half that of the

unconstrained version, namely α = 1.0 for signed nets and α = 2.0 for unconstrained nets (Amit et al., 1989a). 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 (for unbiased patterns) is independent of the particular sign constraint used. However the presence of correlated training data will make the capacity of the network sensitive to the specific *sign-bias*.

Viswanathan (Viswanathan, 1993) studied the special case of networks which strictly adhered to Dale's rule, so that all the outgoing weights at a given neuron had the same sign, $i, i' \quad g_{ij} = g_{i'j}$. The results showed that the theoretical capacity of such networks was always greatest when the number of excitatory and inhibitory neurons was equal, $\langle g_{ij} \rangle = 0$. Moreover when the training data becomes increasingly correlated the theoretical capacity increases, so that with the optimal sign constraint ($\langle g_{ij} \rangle = 0$) the initial capacity for unbiased data of α = 1.0 would increase as the data correlation increased.

The dynamics of the network is also affected by the sign bias. Wong and Campbell (Wong and Campbell, 1992) showed that in a diluted network, with any sign constraint that had a non-zero bias of positive or negative weights, developed a new form of attractor: the uniform state (all +1/-1). As the sign-bias increases then the uniform state becomes progressively more likely to attract other states. It is likely that this behaviour would extend to fully connected networks, since for example, in a network with positive weights only, the energy function $E\{\boldsymbol{S}\} = -\frac{1}{2}\sum_{i,j} w_{ij} S_i S_j$, will have a *global* minima at the uniform, +1, state. A consequence of the increasing influence of the uniform attractor could be to decrease the attractor basin size of the stored patterns.

1.1 Learning Rules

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 (Gardner, 1988) 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 in Section 2 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 (Abbott, 1990). 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 ξ, 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 (Kepler and Abbott, 1988) and we take $\kappa = \min_{p,i}(\gamma_i^p)$.

Perceptron learning for training these networks was proposed by Diederich and Opper (Diederich and Opper, 1987), who denoted the method *local learning* (LL). As the units in the network are treated independently the resulting weights are not symmetric. However Gardner (Gardner, 1988) showed that symmetry could be maintained if weight changes to w_{ij} were always mirrored by changes to w_{ji}. Surprisingly this does not decrease the capacity of the network (Nardulli and Pasquariello, 1991), and we denote this rule as symmetric local learning (SLL).

Amit et al (Amit et al., 1989b) 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. Specifically, given a particular sign-bias, $g_{ij} = \pm 1$, and an initialisation of zero weights the Signed version of LL, *Signed-LL*, can be formally stated as:

```
Repeat until all local fields are correct
Set the network state to one of the ξ^p
For each unit, i, in turn:
  Calculate h_i^p ξ_i^p. If this is less than
  T then change the weights to unit i
  according to: w'_ij = w_ij + ξ_i^p ξ_j^p / N whenever
  the resulting weight meets the sign
  constraint, g_ij w'_ij > 0, otherwise leave
  the weight unchanged
```

Of course weight symmetry can also be maintained for signed networks, by first requiring that the sign constraints are symmetric, $g_{ij} = g_{ji}$ and

secondly by using SLL modified to adhere to the sign constraint, as above. This learning rule is here denoted as *Signed-SLL*. The only change in the above algorithm is that the weight change is now given by:

$$j \neq i \quad w'_{ij} = w_{ij} + \frac{\xi_i^p \xi_j^p}{N} \qquad w'_{ji} = w_{ji} + \frac{\xi_i^p \xi_j^p}{N}$$

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. With the sign constrained version it is also possible to show (Amit et al., 1989a) a similar result.

4 Analysing Performance

We use, *R*, the normalised mean radius of the basins of attraction (Kanter and Sompolinsky, 1987), as a measure of attractor performance in these networks. 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 fundamental memory for the network to converge (in the sense described above) on that fundamental memory, and m_1 is the largest overlap of the initial state with the rest of the fundamental memories. The angled braces denote a double average over sets of training patterns and initial states.

As one of the learning rules used here can produce non-symmetric weights it is interesting to examine the symmetry of the weight matrix that results.

To this end the symmetry measure of Krauth, Nadal and Mezard (Krauth et al., 1988) was applied to the resulting weight matrices. It is defined as: $\sigma = \frac{\sum_{i,j} w_{ij} w_{ji}}{\sum_{i,j} w_{ij}^2}$. For a symmetric matrix this takes the value +1. For an anti-symmetric matrix it takes the value –1 and for a random set of weights it will be roughly zero.

5 Results

In the following tests randomly created training sets are used. Correlation within the training sets is varied by biasing the patterns: the bias of a training set is the probability that an individual bit in a pattern is +1, so a bias of 0.5 corresponds to uncorrelated data.

5.1 Capacity

The first set of results measures the capacity of signed networks trained using Signed-LL, varying both the bias of the training sets, and the weight sign-bias. The actual capacity can only be estimated; an incremental search was undertaken for the first point at which the network failed to learn five different sets of random patterns. The highest loading for which this was possible was taken as the capacity of the network.

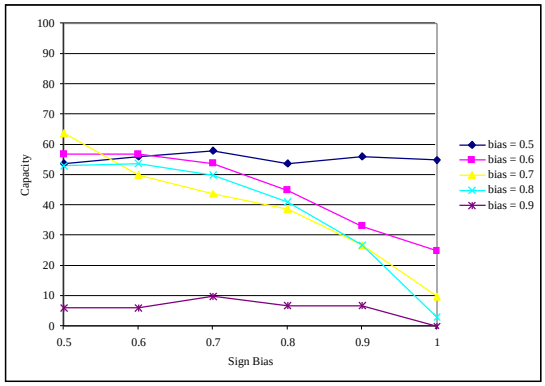

Figure 1: Capacity of 100 unit networks, trained using Signed-LL, with varying degrees of Sign Bias and with different correlations within the training sets (data-bias 0.5 to 0.9).

In Figure 1 it can be seen that when the patterns are not correlated (data-bias = 0.5) the capacity is independent of the specific sign bias, as expected. However this capacity is significantly less than the theoretically predicted one of 100 patterns in a 100 unit network. As the training sets become more correlated, an increasing sign bias causes the capacity to fall considerably. This is in accord with the theoretical prediction of Viswanathan (Viswanathan, 1993) for the special case of networks that adhere to Dale's law. The exception is with highly correlated patterns (data-bias = 0.9) where capacity is very low whatever the sign bias. It is also noteworthy that the networks can withstand some bias in the signs: with these networks capacity was maintained reasonably up to a sign bias of 0.8.

The second of Viswanathan's theoretical predictions, that increasing correlation should increase capacity is however, not confirmed for the general sign bias studied here.

5.2 Basins of Attraction and Symmetry of Weights

Here the mean normalised radii of the basins of attraction, R, associated with fundamental memories is estimated. The minimum of the normalised stability factors, κ, and the symmetry of the weights σ, is also reported. All three sets of results are with 15 random patterns in 100 unit networks, with results averaged over 50 runs. This loading is chosen as, in most cases, it is well within the capacity of the networks. Considering the results in Table 1, where

it can be seen that the signed networks show progressively poorer performance (R values) as the sign of the weights becomes more correlated. This confirms the increasing importance of the uniform attractor, as the sign of the weights become similar. However for each sign-bias the κ values of each type of network are very similar so that the normal relation between R and κ is broken; we have never come across this behaviour in this type of network before.

It is also interesting to note that the non-symmetric version of the signed nets, Signed-LL, performs better than the symmetric version, Signed-SLL. Normally the symmetric weight models are preferred, as they have simpler dynamics, and it is particularly unusual that networks with a relatively low degree of symmetry ($\sigma = 0.41$), as in the case of the 0.50 sign-bias version of Signed-LL should perform so well.

As the sign bias increases the weights become progressively more symmetric, so that at a Sign-Bias of 1.00 the weights are very nearly symmetric. $\sigma = 0.95$. This is not unexpected: as the sign bias of the weights increases the more likely it is that two weight pairs, w_{ij} and w_{ji}, will have the same sign and can therefore take similar values. For comparison the unrestricted learning rule SLL is also included and it can be seen that it attains a κ value roughly twice that of the signed networks. This is in accord with the theoretical prediction: for any given kappa the maximum theoretical capacity of a signed net is half that of its unsigned counterpart (and vice-versa)

Network	Sign-Bias	R	κ	σ
Signed-LL	0.50	0.78	0.99	0.41
	0.75	0.52	0.98	0.54
	1.00	0.23	1.00	0.95
Signed-SLL	0.50	0.65	0.95	1.00
	0.75	0.39	0.95	1.00
	1.00	0.20	0.94	1.00
SLL		0.96	1.84	1.00

Table 1: Uncorrelated Data (bias 0.5). Attractor Performance, R, κ and σ for three different types of network. Each result is for 100 unit networks trained with 15 patterns averaged over 50 runs.

6 Conclusions

Complete freedom in assigning weights to connections may not be an adequate model of biological systems, where amongst other constraints, connections may be only excitatory or only inhibitory. The proportion of signed to unsigned weights in a network, its sign-bias, may affect the behaviour of the network. One of the important results here is that the actual capacity of a sign constrained network is a lot less than the theoretical maximum. The presence of correlation in the training data decreased the capacity, contrary to both the behaviour of unsigned nets and the theoretical prediction of Viswanathan (Viswanathan, 1993).

The degree of correlation in the signs of the weights was shown to affect the dynamics of the trained networks, so that the best attractor performance (R values) was attained with neutral sign correlation, where the uniform attractor was not significant. It was also observed that with sign constrained networks, the normal static measure of likely performance, the smallest normalised stability measure, was not a good predictor of performance.

The specific sign bias of these networks is important in attaining good performance and it suggests that in biological systems the ratio of excitatory to inhibitory synapses will not be accidental.

References

Abbott, L. F. (1990) *Network: Computational Neural Systems,* **1,** 105-122.

Amit, D. J., Campbell, C. and Wong, K. Y. M. (1989a) *Journal of Physics A: Mathematical and General,* **22,** 4687.

Amit, D. J., Wong, K. Y. M. and Campbell, C. (1989b) *Journal of Physics A: Mathematical and General,* **22,** 2039.

Dale, H. H. (1935) *Proceedings of the Royal Society of Medecine,* **28,** 319-332.

Diederich, S. and Opper, M. (1987) *Physical Review Letters,* **58,** 949-952.

Gardner, E. (1988) *Journal of Physics A,* **21,** 257-270.

Kanter, I. and Sompolinsky, H. (1987) *Physical Review A,* **35,** 380-392.

Kepler, T. B. and Abbott, L. F. (1988) *Journal Physique de France,* **49,** 1657-1662.

Krauth, W., Nadal, J.-P. and Mézard, M. (1988) *Journal of Physics A,* **21,** 2995-3011.

Nardulli, G. and Pasquariello, G. (1991) *Journal of Physics A: Mathematical and General,* **24,** 1103.

Sompolinsky, H. (1986) *Physics Review A,* **34,** L519-L523.

Viswanathan, R. R. (1993) *Journal of Physics A: Mathematical and General,* **26,** 6195.

Wong, K. Y. M. and Campbell, C. (1992) *Journal of Physics A: Mathematical and General,* **25,** 2227.