Frequency-based error back-propagation in a cortical network

Rafal Bogacz
Dept. of Computer Science
University of Bristol
Bristol BS8 1UB, U.K.
R.Bogacz@bristol.ac.uk

Malcolm W. Brown
Dept. of Anatomy
University of Bristol
Bristol BS8 1TD, U.K.
M.W.Brown@bristol.ac.uk

Christophe Giraud-Carrier
Dept. of Computer Science
University of Bristol
Bristol BS8 1UB, U.K.
cgc@cs.bris.ac.uk

This paper presents a biologically plausible mechanism of back-propagating network output error to previous layers of processing in a particular multi-layer neural network. This mechanism is used in a network that is designed to mimic familiarity discrimination as performed by the perirhinal cortex of the temporal lobe. In the algorithm, the error of the network during an initial classification period regulates the frequency of neuronal activity in a succeeding memorising period via an inhibitory circuit, such that the frequency in this memorising period is proportional to the error. Synaptic weight modifications are made according to activity-dependent Hebbian rules, such as may be used in the brain. The magnitude of the modification depends on the frequency of the activity. Hence, the magnitude of weight modification is proportional to the network error.

1 Introduction

In most training algorithms for multi-layer neural networks, weights are optimised according to the gradient of error, where the errors obtained at the output layer influence the modification of all the weights in the network. Although many algorithms for multi-layer network training have been proposed (e.g. [1, 13, 7]), there has been no biologically plausible explanation offered for how the error may be propagated back to previous layers of processing.

Much evidence indicates that discrimination of the relative familiarity of stimuli is dependent on the part of the temporal lobe – the perirhinal cortex [6, 10, 11]. In this area of the brain, a proportion of neurons respond strongly to the sight of novel objects but respond only weakly or briefly when these objects are seen again [15] (see Figure 1). This paper describes a biologically plausible method of error back-propagation, that has been implemented in a network performing familiarity discrimination designed to mimic processing in the perirhinal cortex. This method, although specific to this network, maybe generalisable to other models.

In [5], we present the model of familiarity discrimination in the perirhinal cortex, which is consistent with many experimental observations. Here, an extension of this model is described, which allows more precise weight modifications and results in improved network performance. The proposed algorithm utilises the property of synaptic plasticity in the brain, i.e., that a higher intensity of activity is more likely to produce a higher magnitude of synaptic weight modification [3]. The error of the network during the initial classification period regulates the frequency of neuronal activity in the succeeding memorising period via an inhibitory circuit. This mechanism causes the magnitude of change of all synaptic weights in the network to be proportional to the network error. As the information about the error is carried to synapses by the frequency of neuronal activity, the learning algorithm is referred to as frequency-based error back-propagation, or FreqProp for short.

The paper starts with an overview of our model of the perirhinal cortex in Section 2. Then FreqProp is presented in Section 3. Section 4 shows the results of simulations and Section 5 discusses the relationships between FreqProp and other learning algorithms. A proof of convergence of the presented algorithm is given in the Appendix.

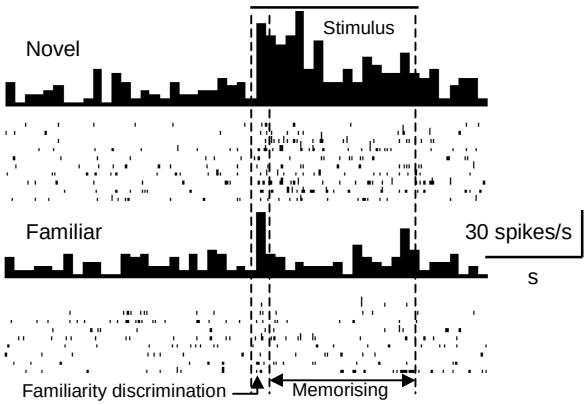

Figure 1. Response of perirhinal neurons to first and repeat presentations of ten pictures recorded from a monkey's brain [15]

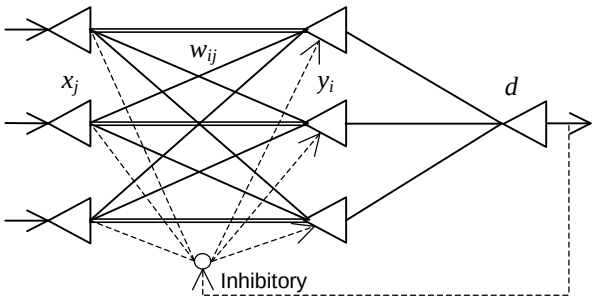

Figure 2. Network architecture in the model of perirhinal cortex

2 Perirhinal network model

Our model of the perirhinal network is briefly described here. In this description, a simple binary model of neurons is used. Full details are in [5].

The network operates in two phases. During a brief initial period (the familiarity discrimination period), the discrimination of familiarity is achieved. This corresponds to the short interval in which the neurons may be active for both novel and familiar stimuli (Figure 1). The subsequent, longer period (the memorising period) of high frequency spike activity for a novel stimulus then produces differential modifications of synaptic weights, thereby storing the occurrence of a novel stimulus. There is no high frequency activity for familiar stimuli (Figure 1), since they are already stored in the network and there is therefore no need for weight modifications. During this memorising period, after presentation of a novel stimulus, neurons respond with high frequency (Figure 1) and alterations of synaptic weights occur according to a Hebbian learning rule. Hence, in the model, only the first spike(s) in an action potential train is used for computation and the remaining spikes, through maintained high frequency activity, produce weight modification.

The model has three layers (Figure 2). The first layer consists of *N representation neurons*, which provide inputs to the familiarity discrimination network. The pattern of activity of the representation neurons is an internal representation of a stimulus encoded as a sequence of *N* bits. The second layer consists of *familiarity detection neurons* (FDNs): FDNs make individual, independent decisions about the familiarity of the stimulus. The values of the FDNs' weights result in more FDNs being active for familiar patterns than for novel ones during the initial period of processing. Since any individual FDN may make a mistake, e.g., be active for a novel stimulus during this initial period, a third layer of *decision neurons* is required to sample the activity of the population of FDNs. Decision neurons receive inputs from the FDNs and are activated only when a majority of their inputs are active. Hence, during the initial period they are active for familiar patterns and inactive for novel ones. These decision neurons govern the subsequent activity of the network. After presentation of a familiar stimulus, the activated decision neurons trigger the inhibitory neurons which prevent activity in the FDNs during the memorisation period. For novel patterns, the decision neurons are inactive, hence the inhibition is not increased and thus the FDNs are active during the memorisation period (see [5] for details). The network computes the following function:

$$d(\bar{x}) = \mathrm{sgn}\left(\sum_{\substack{i=1 \\ x_i=1}}^{N} y_i \right), \text{ where } y_i = \mathrm{sgn}\left(\sum_{j=1}^{N} x_j w_{ij} - \tfrac{1}{2} \right) \tag{1}$$

where y_i denotes the activation of FDN_i. In equation 1, the summing is done only over the activation of those FDN_i for which $x_i = 1$. In the network, this is implemented by driving connections between representation neurons and corresponding FDNs (denoted by double lines in Figure 2). The driving connections have high synaptic weights (not changed during learning), which ensures that to activate a FDN, the corresponding representation neuron must also be active [5]. The weights of the FDNs (denoted by w_{ij}) are initialised to 0 and then modified after each stimulus presentation according to the Hebb rule of Equation 2. In Equation 2, δ is equal to 1 if the pattern was classified as novel (FDNs are active during the memorising period) and to 0 if the pattern is familiar (FDNs are inactive).

$$\Delta w_{ij} = \frac{1}{N} \delta \, y_i x_j \tag{2}$$

3 Frequency-based error back-propagation

The output of the network described in Section 2 returns only a binary decision (novel or familiar). However, familiarity is a fuzzy concept, i.e., some objects seem more familiar than others [2]. This representation of the level of confidence about familiarity may be implemented by introducing a number of decision neurons. The population activity of these decision neurons should be proportional to the population activity of the FDNs in the familiarity discrimination period. Such an effect may be achieved in many ways. For example, the decision neurons may be stochastic, or they may have different thresholds or the connections between FDNs and decision neurons may be sparse. The differences in population activity of decision neurons for different levels of confidence is illustrated by the results of simulation in Figure 3, performed by a more realistic spike-response version of the network (see [5]).

This population coding of familiarity confidence level is also very useful from the point of view of network performance. The less familiar a stimulus seems to be to the network, the stronger the modification of weights should be, so that the stimulus will be classified as familiar during the next presentation. This principle is implemented in the frequency-based error-back-propagation algorithm, or FreqProp.

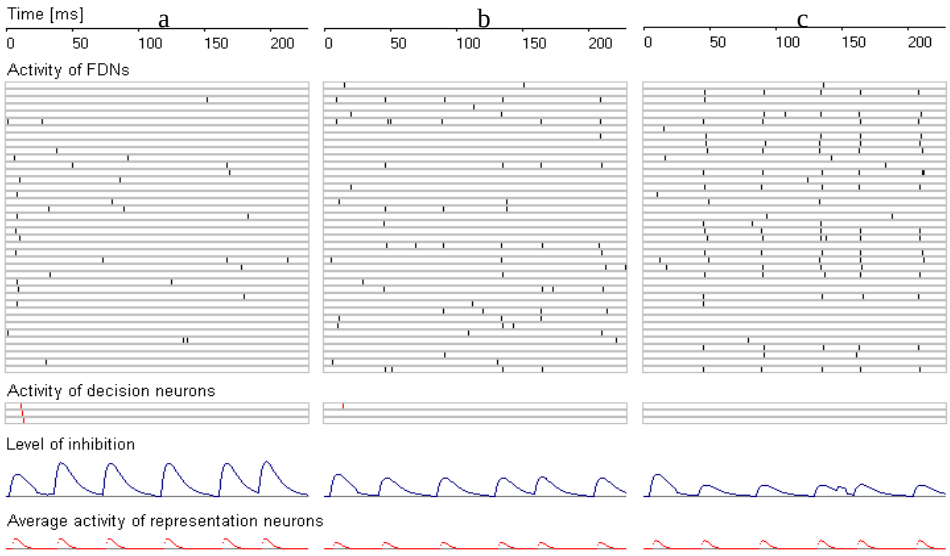

Figure 3. Confidence level encoding in a population activity of decision neurons and its influence on the network activity during the subsequent memorising period. The simulation was performed using a network consisting of 40 representation neurons, 40 FDNs and 3 decision neurons with different thresholds. a) Presentation of a highly familiar pattern causes large activity of FDNs in the initial period, which result in the activation of all the decision neurons (representing classification of the pattern as familiar with high confidence). The large activity of decision neurons causes high level of inhibition in the network, which results in low activity during the following memorising period. b) A stimulus is classified as novel with low confidence, which causes small activity during the memorising period. c) A stimulus is classified as novel with high confidence, which results in strong activity during the memorising period.

In FreqProp, the error is taken as the difference between the confidence level that is required be achieved during the next stimulus presentation, and the current confidence level. In particular, during the next presentation the network should classify the stimulus as familiar, i.e., there should be maximum activity of the decision neurons. Hence, the error is equal to the difference between the maximum and the actual population activity of the decision neurons.

In the perirhinal network, the decision neurons give inputs to inhibitory neurons, which in turn inhibit FDNs (see Figure 2) and thus regulate the neuronal activity in the following memorising period. The lower the number of active decision neurons (i.e., the higher the error), the lower the inhibition and hence the higher the neuronal activity (see Figure 3 for sample simulation results). Since the magnitude of weight modification depends on the frequency of activation (see [3]), higher error causes larger modification of FDNs' weights.

In the network the firing frequency of the FDNs carries information about the value of the error. However, here for simplicity this effect is not simulated explicitly. Instead, we use the term δ in Equation 2, which regulates the magnitude of weight modification, as it can be made to be proportional to the error. To do this, we define the error in terms of the Hopfield energy function, in the Hopfield network having the same weights as the FDNs. This is adequate as familiarity discrimination in the Hopfield network [4] has properties similar to those of the perirhinal network [5], and its mathematical analysis is much simpler. In the Hopfield net, familiarity discrimination is performed not by the network itself but by an external observer calculating the Hopfield energy. The average value of the energy for stored patterns is $-N/2$ (where N = number of neurons in the network), while for novel patterns it is 0 [4]. Therefore, by taking as a threshold the middle value $-N/4$, we can define a familiarity discrimination criterion, namely, if $E < -N/4$, then the pattern is considered familiar, otherwise it is novel.

In FreqProp, we assume that the target value of the energy function corresponding to classification of a pattern as familiar is $-N/2$ (which is the average value of the energy function for stored patterns [4]). The error term from Equation 2 may be defined as:

$$\delta = \begin{cases} \dfrac{N/2 + E(x)}{N/2} & \text{for } E(x) > -N/2 \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

According to Equation 3, if a pattern was classified as familiar with high confidence ($E(x) < -N/2$), then δ is equal to 0 (no weight modification). The higher the level of confidence about the novelty of a pattern (i.e., the higher the energy), the higher the value of δ. The error term δ is defined in such a way, that if the same pattern is presented just after weight modification, the value, E', of the energy function is given by:

$$E'(x) = E(x) - \tfrac{1}{2}\sum_{i=1,\,j=1}^{N} x_i x_j \Delta w_{ij} = E(x) - \tfrac{1}{2}\sum_{i=1,\,j=1}^{N} x_i x_j \,{}^{1}\!/_{N} x_i x_j \frac{N/2 + E(x)}{N/2} = E(x) - {}^{N}\!/_{2}\frac{N/2 + E(x)}{N/2} = -{}^{N}\!/_{2} \quad (4)$$

Since E' is exactly $-N/2$, the pattern is correctly classified as familiar with high confidence. In this sense, the FreqProp demonstrates single-trial learning. We refer to the algorithm described in Section 2 as the binary algorithm to distinguish it from FreqProp. In the binary algorithm, δ is equal to 1 if $E(x)>-N/2$ and 0 otherwise. The binary algorithm does not always have the single-trial learning property. For example, if the energy for a novel pattern is equal to $N/2$ then after the weight modification with the binary algorithm, the energy for this pattern will be equal to 0 during its next presentation and the pattern will be classified as novel again.

Repeated presentations of training patterns usually increase the storage capacity of neural networks [12, 7]. Both FreqProp and binary algorithms have analogous property: repeated presentations increase the number of patterns, which may be classified as familiar. Furthermore, the algorithms are guaranteed to find the values of the weights, resulting in the classification of all the patterns in the training set as familiar with high confidence, if such values of weights exist (see Appendix).

There exists a small probability of classifying a novel pattern as familiar by the network. The patterns for which this false recognition error may occur correspond to the spurious attractors in the Hopfield network (i.e., the patterns having lower energy than the stored patterns), which exist even for a very small number of stored patterns [7]. Therefore FreqProp does not guarantee that novel patterns are classified correctly. It is not possible to guarantee this, because false recognition patterns (corresponding to spurious attractors in the Hopfield net) always exist.

FreqProp converges much faster than the binary algorithm, because in FreqProp the absolute values of weights ($|w|$ in Appendix) are much smaller than in the case of the binary algorithm, where the absolute values of weights grow rapidly. This results in lower probability of false recognition errors in the case of the FreqProp algorithm. The experimental results described in Section 4 confirm these findings.

4 Results of simulation

The storage capacity for familiarity discrimination was compared for FreqProp and the binary algorithm. After training, the classification of a pattern by the network was considered as correct when the corresponding energy function was lower than $-N/4$ in case of a familiar pattern and higher than $-N/4$ in case of a novel pattern.

The storage capacity of the network was calculated after patterns had been presented different numbers of times, i.e. after repeated presentations (iterations). For each number of iterations we calculated the storage capacity of the network, i.e. the maximum number of patterns for which the familiarity classification error is lower than 1%. The tests were performed for two types of data sets, namely, one consisting of stored patterns only and one consisting of equal numbers of stored and novel patterns. The storage capacity is the average over many tests performed on 5000 patterns for each number of iterations. The results of simulations are shown in Figure 4.

Curves describing the capacity when only stored patterns are presented (filled dots in Figure 4) show that FreqProp converges to the optimal solution much faster than the binary algorithm. This is especially clear when there have been just a few iterations, for example after two iterations the capacity for FreqProp is three times higher than that of the binary algorithm. Curves describing capacity when both stored and novel patterns are presented (empty dots in Figure 4) show that the capacity of Freq-Prop does not depend on the number of iterations. In the case of the binary algorithm the capacity decreases after many iterations due to the increase in the absolute values of the weights.

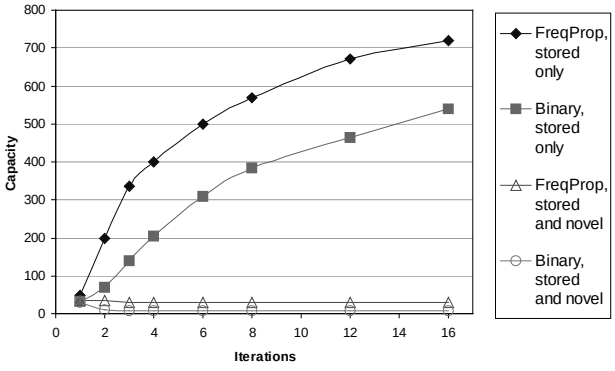

Figure 4. Storage capacity of familiarity discrimination network consisting of $N = 50$ FDNs for different numbers of training iterations.

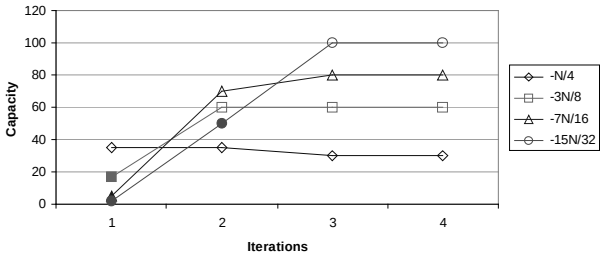

Figure 5. Storage capacity of familiarity discrimination network consisting of $N = 50$ FDNs for different values of discrimination threshold and different number of training iterations. The empty dots represents cases where the majority of errors made by network are false recognitions, while the filled dots represent cases where the non-recognition errors prevail.

Curves for which the proportion of novel to familiar patterns is neither zero nor equal will lie between the curves presented in Figure 4. In everyday life it might be expected that most patterns will be familiar (indeed often experienced), the minority being novel. Accordingly, the analysis establishes that the capacity of the network is likely to be closer to the upper curves in situations that parallel reality.

When there is more than one iteration, the only mistakes made by network are false recognition. Hence, for higher numbers of iterations the capacity may be improved by decreasing, i.e. making more negative, the discrimination threshold (which was equal to $-N/4$ in the above experiment). Figure 5 shows the storage capacity when both stored and novel patterns are presented for different discrimination thresholds and different numbers of training iterations. By decreasing the threshold, the capacity may be increased. The optimal value of the threshold depends on the number of training iterations: the higher the number of iterations, the lower the value of the optimal threshold.

5 Discussion

FreqProp is related to the standard error back-propagation algorithm, where the magnitude of the weight modification of hidden unit j is defined as [13]:

$$\Delta w_{jk} = \eta f'(\phi_j) x_k \sum_i w_{ij} \delta_i \tag{5}$$

Δw_{jk} depends on two groups of parameters. The first group are the parameters local to the neuron (i.e., input to the neuron x_k, membrane potential ϕ_j, derivative of the transition function f' and learning rate η). The second group are the parameters propagated from the next layers (i.e., the error terms in the next layer δ_i and the weights w_{ij} of the forward connections of the neuron). The parameters propagated from the next layer are individual for each hidden unit, because the weights of neurons in the next layer (w_{ij}) have different values.

The specific property of the perirhinal network is that all the weights of output units are equal (see Equation 1) and not modified during learning. Hence, the error propagated back to the hidden unit is the same for each neuron and therefore it can be encoded in one value, namely the frequency of activity in the memorising period. The generalisability of this error correction algorithm to other networks is a subject for future work.

Other, non-gradient based algorithms for training multi-layer networks have been proposed, which do not require error back-propagation and which use only locally accessible information for weight modification. For example, the Alopex algorithm modifies the weights based on the correlation between the previous weight modification and the change in error [14]. In regression by independent component analysis, the weights of hidden units are found by unsupervised learning on both inputs and outputs [9]. Algorithms similar to FreqProp could be incorporated in the implementation of the above algorithms in networks designed to model processing in the brain. For example, in the Alopex algorithm the information about the error produced by the network must be delivered to each synapse, which could be implemented in a way similar to FreqProp.

Appendix A Convergence proof

This Appendix contains the proof that the binary algorithm finds values of the weights resulting in the correct classification of patterns from the training set if such values exist. The proof extends naturally to FreqProp. For simplicity the Hopfield energy function is used as the discrimination criterion. The proof is analogous to the convergence proof of the perceptron rule [12].

Let us denote the i-th bit of stored pattern μ by ξ_i^μ and the weight between representation neuron j and FDN_i after n-th weight correction by $w_{ij}^{(n)}$. Let us also define $F = -2E$, where E is value of the Hopfield energy, i.e.

$$F = \sum_{i,j} w_{ij}^{(n)} x_i x_j \tag{A.1}$$

Hence, when $F = N$ a pattern is classified as familiar with high confidence. The weights are modified according to Equation 2, which becomes:

$$\text{if } F < N, \text{ then modify weights: } \Delta w_{ij} = \frac{1}{N} x_i x_j \tag{A.2}$$

Let J denote a set of weights resulting in the correct classification of all the patterns from the training set as familiar with high confidence, i.e.,

$$\forall \mu \quad \sum_{i,j} J_{ij} \xi_i^\mu \xi_i^\mu \geq N \tag{A.3}$$

The convergence theorem essentially states that after a finite number m of weight corrections, the algorithm finds the weights resulting in the correct classification of all the training patterns if such weights exist, i.e.,

$$\text{IF } \exists J \, \forall \mu \; \sum_{i,j} J_{ij} \xi_i^\mu \xi_i^\mu \geq N \quad \text{THEN } \exists m \forall \mu \; \sum_{i,j} w_{ij}^{(m)} \xi_i^\mu \xi_i^\mu \geq N \tag{A.4}$$

Proof. In the proof, the weight matrices are sometimes treated as vectors created by the concatenation of matrices' rows. Hence the "length" and the "scalar product" of matrices are defined here as:

$$|A| = \sqrt{\sum_{i,j} A_{ij}^2}\ ,\ A \circ B = \sum_{i,j} A_{ij} B_{ij} \tag{A.5}$$

Let us define G by:

$$G = \frac{J \circ w}{|J||w|} \tag{A.6}$$

Since $A^{\circ}B = |A||B|\cos\angle(A,B)$, hence:

$$\frac{J \circ w}{|J||w|} = \cos\angle(J, w) \le 1 \tag{A.7}$$

$G = 1$ only when $J = \lambda w$, where λ is a non-zero constant. Let us denote $N^{(n)} = J \circ w^{(n)}$ and $D^{(n)} = |J||w^{(n)}|$. Let us calculate the change in $N^{(n)}$ due to the n-th weight correction.

$$\Delta N^{(n)} = \sum_{i,j} J_{ij} \Delta w_{ij} = \frac{1}{N} \sum_{i,j} J_{ij} \xi_i^{\mu} \xi_i^{\mu} \tag{A.8}$$

Since J are the weights resulting in the correct classification, from Equation A.3, $\Delta N^{(n)} \ge (1/N)N = 1$. Since at the beginning all the weights are initialised to zero (so $N^{(0)} = 0$), then

$$N^{(n)} = \sum_{k=1}^{n} \Delta N^{(k)} \ge n \tag{A.9}$$

To calculate $D^{(n)}$, let us first estimate the change in $|w^{(n)}|^2$ due to the n-th weight correction.

$$\Delta\left|w^{(n)}\right|^2 = \left|w^{(n)} + \Delta w\right|^2 - \left|w^{(n)}\right|^2 = \sum_{i,j}\left(w_{ij}^{(n)} + \Delta w_{ij}\right)^2 - \sum_{i,j}\left(w_{ij}^{(n)}\right)^2 = \frac{2}{N}\sum_{i,j} w_{ij}^{(n)} \xi_i^{\mu} \xi_i^{\mu} + \frac{1}{N^2}\sum_{i,j}\left(\xi_i^{\mu} \xi_i^{\mu}\right)^2 \tag{A.10}$$

Since $\xi_i^{\mu} \in \{-1, 1\}$, the second term in the right-hand side of Equation A.10 is equal to 1. Since the weight corrections are made only when a pattern is incorrectly classified, then from A.1 and A.2 the first tem in Equation A.10 is smaller or equal to $(2/N)N = 2$. Therefore, $\Delta|w^{(n)}|^2 \le 3$, which implies that $|w^{(n)}|^2 \le 3n$, hence:

$$\left|w^{(n)}\right| \le \sqrt{3n}\ , \text{ and hence } D^{(n)} \le \sqrt{3n}|J| \tag{A.11}$$

From Equations A.9 and A.11 we obtain:

$$G^{(n)} = \frac{N^{(n)}}{D^{(n)}} \ge \frac{n}{\sqrt{3n}|J|} = \frac{\sqrt{n}}{\sqrt{3}|J|} \tag{A.12}$$

The sequence $G^{(n)}$ grows with the square root of n, which means that it is unbounded (its limit is infinite). However, from A.7 it is bounded by 1, which means that, after a number of steps m, the process of weight corrections must finish (otherwise G would exceed 1). The process can finish only if there is no more correction to be made, i.e., when all the patterns are classified correctly as familiar with high confidence. This proves, that after a number of steps the learning algorithm finds the weights resulting in the correct classification, if such weights exist.

References

[1] Amari S. (1967). Theory of Adaptive Pattern Classifiers. *IEEE Transactions on Electronic Computers*, 16, 299-307.

[2] Begg, I., & Rowe, E.J. (1972). Continuos judgement of word frequency and familiarity. *J. Exp. Psych.*, 95, 48-54.

[3] Bliss, T.V.P., & Collingridge, G.L. (1993). A synaptic model of memory: long-term potentiation in hippocampus. *Nature*, 361, 31-9.

[4] Bogacz, R., Brown, M.W., & Giraud-Carrier C. (1999). High capacity neural networks for familiarity discrimination. In *Proceedings of ICANN'99*, 773-778.

[5] Bogacz, R., Brown, M.W., & Giraud-Carrier C. (submitted). Model of familiarity discrimination in the perirhinal cortex.

[6] Brown, M.W., & Xiang, J.Z. (1998). Recognition memory: Neuronal substrates of the judgement of prior occurrence. *Prog. Neurobiol.*, 55, 149-189.

[7] Hertz, J., Krogh, A., & Palmer, R.G. (1991). *Introduction to the theory of neural computation.* Addison Wesley.

[8] Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities, *Proc. of Nat. Acad. of Sc.*, 79, 2554-2558.

[9] Hyvarinen, A. (1999). Regression using independent component analysis, and its connection to multi-layer perceptrons. In *Proceedings of ICANN'99*, 491-496.

[10] Murray, E.A. (1996). What have ablation studies told us about the neural substrates of stimulus memory? *Semin. Neuros.*, 8, 13-22.

[11] Murray, E.A., & Bussey, T.J. (1999). Perceptual-mnemonic functions of the perirhinal cortex. *Trends in Cogn. Sci.*, 3, 142-151.

[12] Rosenblatt, F. (1962). *Principles of neurodynamics.* New York: Spartan.

[13] Rumelhart, D.E., Hinton, G.E., & Williams R.J. (1986). Learning representations by back-propagating errors. *Nature*, 323, 533-536.

[14] Unnikrishnan, K.P., & Venugopal, K.P. (1994). Alopex: A correlation-based learning algorithm for feedforward and recurrent neural networks. *Neural Computation*, 6, 469-490.

[15] Xiang, J.Z., & Brown, M.W. (1998). Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. *Neuropharmacol.*, 37, 657-676.

Acknowledgements: This work is supported in part by ORS, Wellcome Trust and MRC grants.