An anti-Hebbian model of familiarity discrimination in the perirhinal cortex

Rafal Bogacz [1], Malcolm W. Brown [2]

[1] *Department of Applied and Computational Mathematics, Princeton University, Princeton NJ 08544, USA*

[2] *MRC Centre for Synaptic Plasticity, Department of Anatomy, University of Bristol, Bristol BS8 1TD, UK*

Correspondence to:

Rafal Bogacz

Department of Applied and Computational Mathematics

Princeton University

Princeton NJ 08544, USA

Tel: +1-609-258-6155

Fax: +1-609-258-1735

E-mail: rbogacz@princeton.edu

Key words:

recognition memory, novelty detection, perirhinal cortex, familiarity, hippocampal region,

Abstract

Much evidence indicates that the perirhinal cortex of the temporal lobe is involved in the familiarity discrimination aspect of recognition memory. All previously published models of familiarity discrimination in the perirhinal cortex are based on Hebbian learning. Here we present a biologically plausible model based on anti-Hebbian learning. When the responses of neurons providing input to the familiarity discrimination network are correlated (as is indicated by experimental data), then the anti-Hebbian model achieves a much higher capacity (up to thousands of times) and hence a crucially higher efficiency than models based on Hebbian learning.

1. Introduction

Work in monkeys has established that discrimination of the relative familiarity or novelty of visual stimuli is dependent on the perirhinal cortex, and this finding is consistent with studies of amnesic patients [5,6]. Within the monkey's perirhinal cortex, ~25% of neurons respond strongly to the sight of novel objects but respond only weakly or briefly when these objects are seen again [6,13].

All previously published models of familiarity discrimination in the perirhinal cortex [1,10,11] are based on Hebbian learning. When it is assumed that responses of neurons providing input to the network are uncorrelated, these models achieve very high storage capacity, sufficient potentially to explain human familiarity discrimination capabilities. Under this assumption, if the perirhinal cortex worked akin to these models, it alone could discriminate the familiarity of many more stimuli than current neural network models indicate could be recalled (recollected) by all the remaining areas of the cerebral cortex. [1]. However, experimental evidence indicates that the responses of neurons in the perirhinal cortex are correlated [8]. If so, then the achievable capacities of the published networks based on Hebbian learning are dramatically reduced [2].

In this abstract we present a biologically plausible familiarity discrimination network based on anti-Hebbian learning, that achieves much higher capacity than networks based on Hebbian learning when the inputs are correlated. Due to space limitation this abstract includes only a description of the model and results of simulations. The derivation of the capacity and discussion of the consistency of the model with experimental observations will appear elsewhere [3].

We focus on modelling computations performed by 'novelty' neurons, the ~10% of perirhinal neurons that respond strongly to the first presentations of novel stimuli but only briefly or weakly to presentations of previously seen stimuli [13].

2. Description of the model

For ease of explanation and mathematical analysis, the network is described using a simple model of neurons which does not consider changes of neurons' membrane potentials in time. We assume that each visual stimulus is represented by a specific pattern of activity of the neurons providing input to the familiarity discrimination network and each of these input neurons may be in one of two states: active or inactive.

Figure 1 shows the synaptic plastic changes for the anti-Hebbian model. After presentation of a novel stimulus the synaptic weights of connections from active input neurons are decreased due to homo-synaptic long-term depression, which is known to exist in the perirhinal cortex [7]. This synaptic modification decreases the sum of the synaptic strengths (weights) of the novelty neuron. Hence to maintain the neuron's overall excitability, the synaptic weights of connections from inactive input neurons must be increased (see Figure 1). This increase may be mediated by homeostatic mechanisms that act to maintain average neuronal activity and thus promote network stability (they have been reported in cultures and slices of cortical neurons; for review see [12]). When the same stimulus is presented again, the membrane potential of the novelty neuron will be lower (because the weights of synapses of inputs that were active for this stimulus have been reduced) and the novelty neuron will be inactive (or, more generally, less active). Thus the neuron responds more strongly to novel than familiar stimuli.

The anti-Hebbian model includes a single layer of novelty neurons receiving projections from the input neurons. If each novelty neuron makes its own decision about stimulus familiarity, the overall response ("answer") of the network is encoded in the population activity of the novelty neurons. It is necessary to ensure that individual novelty neurons remain independent assessors of familiarity if the information storage capacity of the network is to be maximised [1].

Otherwise, should all the novelty neurons be active after the presentation of each of a series of novel stimuli, then the synaptic weights of each of the novelty neurons would be modified in the same way, and hence all the novelty neurons would come to have highly correlated weights. Thus, eventually, they would all be active or inactive together and the whole network would have the same capacity as a single novelty neuron. To avoid this problem, the number of novelty neurons active for any one stimulus must be limited, i.e. only a subset of novelty neurons must respond to any given stimulus. This limitation of the number of active novelty neurons is achieved in the model by inhibitory competition: only the fraction of neurons with the highest membrane potentials are selected to be active, the activity of the remainder being suppressed by inhibition, and only these most active neurons have their weights modified [10]. For mathematical details of the simulated version of the model see the Appendix.

3. Storage capacity for correlated responses of input neurons

Storage capacity is defined as the number of presented stimuli for which a network can discriminate familiarity with an accuracy of 99% [1]. To ease explanation, the capacity was established using simple binary patterns. Further, sparse coding was not assumed (the probability of each input neuron being active was 50%). However, the sparseness of coding does not have a great influence on the capacity of familiarity discrimination networks [4].

The simple binary patterns were generated so that (the modulus of) the correlation between each pair of input neurons was constant. Thus at the beginning of a simulation a binary template pattern ξ^{temp} was generated randomly. All the patterns ξ^{μ} were biased towards ξ^{temp}, such that the probability of $\xi^{\mu} = \xi^{\text{temp}}$ equalled ½ + ½b, where b is the parameter controlling bias. Additionally, to keep the level of activity constant across the neurons, at random moments in

time the template was inverted, i.e. each bit in template was switched ($\xi^{temp} \leftarrow -\xi^{temp}$). For patterns generated in this way, the correlation r_{ij} between a pair of inputs was equal to b^2 or $-b^2$.

Figure 2a shows that the capacity of the model [1] based on Hebbian learning (these simulations are described in [2]) decreases very markedly even when the correlation is very small. Figure 2b shows that correlation reduces the capacity of the anti-Hebbian model far less than the network based on Hebbian learning.

Furthermore, for familiarity discrimination networks based on Hebbian learning the influence on capacity of correlation between responses of input neurons increases when the size of the network grows. By contrast, for the anti-Hebbian model the effect of correlation on capacity decreases with increasing network size. Hence for large networks, the anti-Hebbian model achieves a capacity much greater than any of the networks based on Hebbian learning when there are even very small correlations between the responses of the input neurons.

4. Discussion

The difference in capacities of the models based on Hebbian and anti-Hebbian learning may be explained intuitively by the fact that the Hebbian models have a natural tendency to extract features; hence they focus on elements common to all the input patterns (i.e. features). By contrast, the anti-Hebbian model focuses on elements characteristic to individual patterns rather than their common features (for more details and formal explanation see [3]).

The consistency of the anti-Hebbian and other models with the results of experimental observations is compared in [3]. In [3] we also estimate the capacity of putative networks of novelty neurons in the human perirhinal cortex. These estimations show that if perirhinal cortex worked akin to the anti-Hebbian model, it could discriminate familiarity for up to thousands of times more stimuli that if it worked according to the models based on Hebbian learning.

References

[1] R. Bogacz, M.W. Brown and C. Giraud-Carrier, Model of familiarity discrimination in the perirhinal cortex, J. Comp. Neurosci. 10 (2001) 5-23.

[2] R. Bogacz, Computational models of familiarity discrimination in the perirhinal cortex, Ph.D. thesis, University of Bristol, 2001. (also available at: http://www.math.princeton.edu/~rbogacz).

[3] R. Bogacz and M.W. Brown, Comparison of computational models of familiarity discrimination in the perirhinal cortex, Hippocampus (in press).

[4] R. Bogacz and M.W. Brown, The restricted influence of the sparseness of coding on the capacity of the familiarity discrimination networks, Network (in press).

[5] M.W. Brown and J.P. Aggleton, Recognition memory: what are the roles of the perirhinal cortex and hippocampus? Nat. Rev. Neurosci. 2 (2001) 51-62.

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

[7] K. Cho, N. Kemp, J. Noel, J.P. Aggleton, M.W. Brown, Z.I. Bashir, A new form of long-term depression in the perirhinal cortex, Nat. Neurosci. 3 (2000) 150-156.

[8] C.A. Erickson, B. Jagadeesh, R. Desimone, Clustering of perirhinal neurons with similar properties following visual experience in adult monkey, Nat. Neurosci. 3 (2000) 1143-1148.

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

[10] K.A. Norman and R.C. O'Reilly, Modelling hippocampal and neocortical contributions to recognition memory: a complementary learning systems approach, Technical Report 01-02, University of Colorado, Boulder, 2001.

[11] V.S. Sohal and M.E. Hasselmo, A model for experience-dependent changes in the responses of infero-temporal neurons, Network 11 (2000) 169-190.

[12] G.G. Turrigiano and S.B. Nelson, Hebb and homeostasis in neuronal plasticity, Cur. Opin. Neurobiol. 10 (2000) 358-364.

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

Figure 1. Synaptic plasticity in the anti-Hebbian model. In each panel, the triangle represents an excitatory novelty neuron [13] and lines on the left side of each panel denote inputs to the network, which are axons of neurons whose activity encodes visual stimuli. “Spikes” over the lines indicate that the corresponding neuron is active, a lack of spikes that it is inactive. The thickness of the lines indicates the strength of the synaptic connections. The left panel illustrates synaptic weights and neuronal responses for a novel stimulus; the right panel when this stimulus is presented again (i.e. is familiar).

Figure 2. The capacities of familiarity discrimination networks for correlated patterns of a fully connected a) Hebbian [1] and b) anti-Hebbian model. X-axes: square root of $|r_{ij}|$, i.e. b; y-axes: capacity P. Methods of calculating capacity as in [1]. For each network and for each number of neurons N, the familiarity discrimination error was estimated for different numbers of stored patterns P, and the capacity P_{max} was taken as the maximum number of stored patterns P, for which the error rate was $\leq 1\%$. For given N and P, the discrimination error was estimated as follows. During each test, P patterns were presented to the network, and then accuracy tested on all the presented (stored) patterns and equal number of novel patterns (generated in the same way as the stored patterns). These tests were repeated until the network had been tested with 5000 stored patterns and 5000 novel patterns, e.g. for P=100, the tests were repeated 50 times. The average accuracy over the tests is plotted. Grey curves show theoretical predictions of capacity according to the equations below the charts. The equation for capacity of the Hebbian model is derived in [2], the anti-Hebbian in [3].

Appendix. Details of the model

The notation in the Appendix follows that of previous work on auto-associative memories [9]. Denote the active state of a neuron by 1, and the inactive by –1. Denoting the inactive state by –1 rather than 0 simplifies the calculation of capacity for familiarity discrimination networks without changing their capacity (see [1]). Also, it is assumed that the probability of an input neuron being active is 50%.

Assume that a network consists of N novelty neurons, receiving information from N input neurons whose activity pattern represents a visual stimulus. For simplicity assume that each novelty neuron is connected to all the input neurons and denote the strength of the synaptic connection between input neuron j and novelty neuron i by w_{ij}. Denote the activity of input neuron j by x_j, and define the membrane potential of novelty neuron i as:

$$h_i = \sum_{j=1}^{N} w_{ij} x_j \tag{1}$$

In the anti-Hebbian model, the number of active novelty neurons is limited by competition. After presentation of a stimulus, the membrane potentials of novelty neurons are calculated according to Equation 1 and a threshold set such that exactly half of the novelty neurons with the highest membrane potentials are allowed to be active. In a real network such selection of a proportion of the most active neurons may be achieved by inhibition and competition (see e.g. [10]). The pattern of activity of the novelty neurons is denoted by y, i.e. y_i=1 if neuron i has membrane potential among N/2 highest membrane potentials in the network, and y_i=–1 otherwise. The weights of the active novelty neurons are updated according to the rule (illustrated in Figure 1):

$$\Delta w_{ij} = -\frac{\eta}{2N}\left(y_i + 1\right) x_j \tag{2}$$

In Equation 2, η denotes the learning rate – a parameter determining the magnitude of weight modification; its optimal value depends on N (see [3]).

The initial network response is equal to the response (proportional to the membrane potential) of neurons selected to be active:

$$d(x) = \sum_{i=1}^{N} \psi_i h_i \tag{3}$$

As the detailed explanation of how such a function may be calculated by a biologically plausible neural network is long, it is not given here, but is in [2]. Due to the anti-Hebbian weight modifications produced by previous occurrences, d is lower for familiar patterns than for novel. The familiarity of stimuli may be discriminated reliably by evaluating d. The familiarity discrimination threshold may be taken as the middle value between the average decision function for novel and for familiar stimuli. For simplicity, during the simulations, the biologically plausible network computing d was not simulated explicitly, but instead the familiarity of a stimulus was evaluated by the simulator program computing function d of Equation 3.

Figure 1

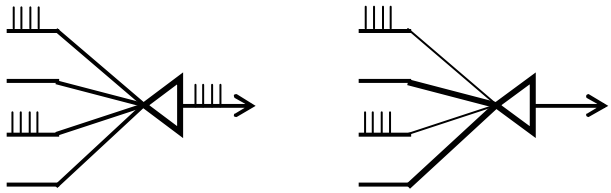
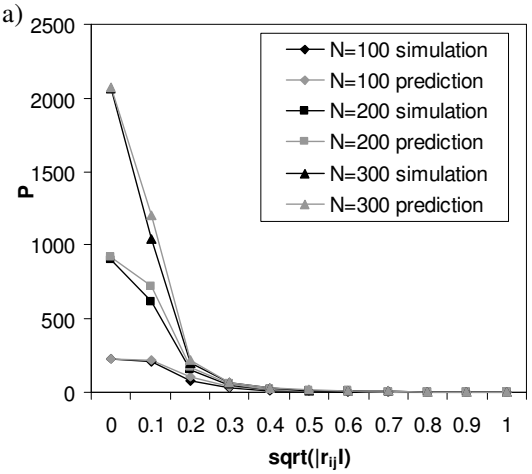


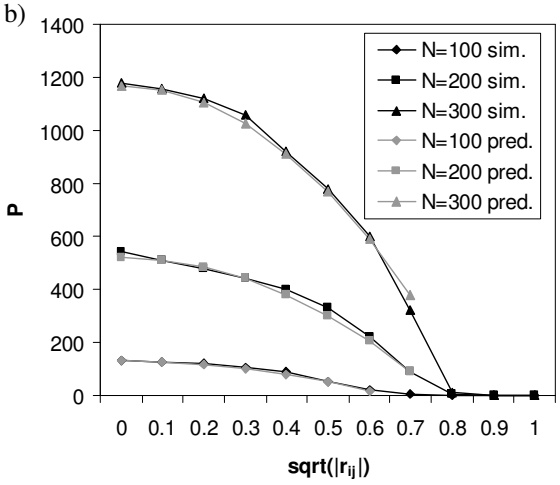
Figure 2

a)



$$P_{\max} \approx \frac{-1+\sqrt{1+0.185N^3r^3}}{4Nr^3}$$

b)



$$P_{\max} \approx 0.013N^2 - 0.31N^{3/2}r$$