Biological Cybernetics manuscript No.
(will be inserted by the editor)

Andrea Greve, David C. Sterratt, David I. Donaldson, David J. Willshaw, Mark C. W. van Rossum

Optimal learning rules for familiarity detection

May 31, 2008

Abstract It has been suggested that the mammalian memory system has both familiarity and recollection components. Recently, a high-capacity network to store familiarity has been proposed. Here we derive analytically the optimal learning rule for such a familiarity memory using a signal-to-noise ratio analysis. We find that in the limit of large networks the covariance rule, known to be the optimal local, linear learning rule for pattern association, is also the optimal learning rule for familiarity discrimination. The capacity is independent of the sparseness of the patterns, as long as the patterns have a fixed number of bits set. The corresponding information capacity is 0.057 bits per synapse, less than typically found for associative networks.

1 Introduction

There is substantial psychological and neuro-physiological evidence that suggests that mammalian memory system has at least two components during recall: One that relies on recollection of the event (the typical what, where, when memories) and one that relies on familiarity (Yonelinas (2001); Fortin et al (2004)). A typical example of familiarity memory is our extensive familiarity memory for faces which, embarrassingly, often goes unaccompanied by an episodic memory of when or where we met the person. Our daily experience with this specific example indicates already that the capacity for familiarity memory is high.

Recently, a model of familiarity memory was proposed, and the number of patterns it can store the familiarity of was shown to be on the order of the number of synapses, which equals the number of units squared (Bogacz et al (2001); Bogacz and Brown (2003)) (see Yakovlev et al (2008) for another recent familiarity network). When quantified this way, the capacity is much higher than for hetero- and auto-associative networks. For instance, the number of patterns that the Hopfield net can store is on the order of the number of units. The high capacity for the familiarity network can be understood intuitively: the familiarity network needs to store just a single bit (familiar/non-familiar) for each pattern. In contrast, a Hopfield network needs to store

Andrea Greve
Doctoral Training Centre for Neuroinformatics, School of Informatics, University of Edinburgh, 5 Forrest Hill, Edinburgh, EH1 2QL, UK

David C. Sterratt, David J. Willshaw, Mark C. W. van Rossum
Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh, 5 Forrest Hill, Edinburgh, EH1 2QL, UK

David I. Donaldson
School of Psychology, University of Stirling, Stirling, FK9 4LA, UK

a vector with a length equal to the number of units for each pattern stored in order for it to be able to do pattern completion.

Storage capacity can be affected by the sparsity of the coding. In the brain codes are generally sparse, which in models is often modeled by a small fraction of 'on' bits in binary patterns. It is well known that in most associative and auto-associative networks the number of patterns that can be stored increases as the patterns are more sparse, provided the learning rule is adjusted correctly (Tsodyks and Feigelman (1988); Meunier and Nadal (1995)). This is not surprising as the information contained in a sparse pattern is less than in a dense pattern. Indeed, the information stored per synapse remains on the order of one bit, relatively independent of sparseness (Willshaw et al (1969); Nadal and Toulouse (1990)). It is unclear whether such arguments hold for a familiarity memory as the network needs to store one bit per pattern independent of sparseness. This raises the question how sparseness, known to be a realistic feature of neural codes, affects the capacity of familiarity memory and secondly, which learning rule should be used to reach optimal performance.

In this paper we calculate explicitly the signal-to-noise ratio for a class of learning rules. We find that in the limit of large networks that store a large number of patterns, the covariance rule yields optimal storage capacity. The capacity is critically dependent on the details of the sparseness implementation. The Shannon capacity as measured in bits per synapse is somewhat less than typically found in other networks. The calculations are supported by simulations.

1.1 Definitions

The network is an all-to-all connected network with binary units. There are m units in the network ($i = 1 \dots m$), which are connected with m^2 synapses. The patterns to be stored are denoted as vectors $\mathbf{x}^\mu$ with elements x_i^μ. The index $\mu = 1 \dots \Omega$ labels the patterns, Ω is the number of patterns to be stored in the network. The pattern elements are binary. Binary patterns can be chosen either to be $x_i^\mu \in \{0, 1\}$, or $x_i^\mu \in \{-1, 1\}$. For ease we assume the ± 1 case, unless indicated otherwise.

The patterns are random, but the probability for $+1$ ('on') and -1 ('off') need not be equal. With $\overline{x}$ we indicate the mean value of x_i^μ (averaged over many patterns). In the limits $\overline{x} \to \pm 1$ the patterns are sparse, whereas the patterns are dense when $\overline{x} = 0$. When the on-probability of each bit in a pattern is given, the actual number of on-bits in a pattern randomly varies around the mean sparsity. We will refer to this choice of pattern statistics as *average sparseness*. An alternative choice is that each pattern has always a given number of on-bits (but in varying positions). In other words, all patterns, familiar and novel, are permutations of each other. We will refer to this as *fixed sparseness*.

We will measure the familiarity by the so-called energy that a certain pattern μ yields, which is given by (Bogacz et al (2001))

$$E^\mu = \sum_{i,j=1}^{m} x_i^\mu w_{ij} x_j^\mu$$

The network's task is to distinguish a novel and known pattern based on this energy. As we concentrate on the learning rule, we do not consider how the energy is actually read out.

The network's architecture is much like the Hopfield network (Hopfield (1982)). However, unlike the Hopfield network, there is no dynamics in the network. Indeed, the transfer function of the units does not enter. Relaxing the network into an attractor state would seriously reduce the capacity of the network to be of order $O(m)$ - the classic capacity result for Hopfield networks (Hertz et al (1991)).

The weights between the units is given by the matrix w_{ij}. The question we address is which learning rule gives the best performance. We restrict ourselves to learning rules of the form: $w_{ij} = \sum_{\nu=1}^{\Omega} \Delta w_{ij}^\nu$ where $\Delta w_{ij} = \alpha$, if $x_i = x_j = -1$, $\Delta w_{ij} = \beta$, if $x_i = -x_j$, and $\Delta w_{ij} = \gamma$, if $x_i = x_j = 1$. This is shown in Table 1. We will vary the parameters α, β, and γ to optimize the learning rule. This will yield the optimal linear, additive, and local learning rule. However, it is

Δw_{ij}^{ν}	$x_i^{\nu}=-1$	$x_i^{\nu}=1$
$x_j^{\nu}=-1$	α	β
$x_j^{\nu}=1$	β	γ

Table 1 Diagram of the learning rule. The total weight between unit i and unit j is given by the sum over the contributions of the different patterns $w_{ij}=\sum_{\mu}\Delta w_{ij}^{\nu}$

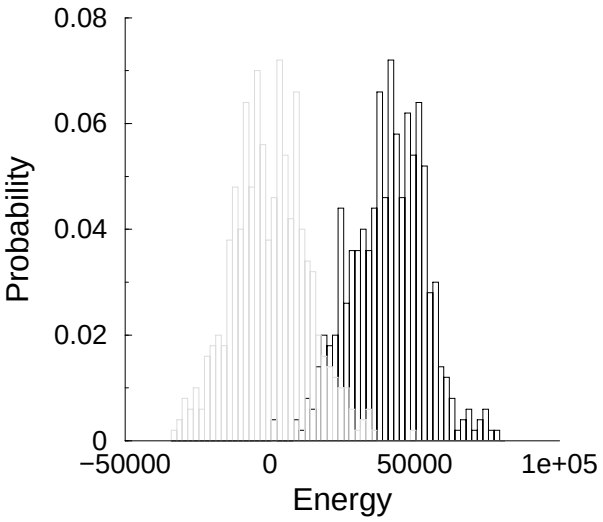

Fig. 1 Example distributions of the energies when the network is tested with familiar (black) and novel (grey) patterns. The signal-to-noise ratio measures the performance of the familiarity discriminator. Simulated network with $m=100$ units, in which $\Omega=500$ patterns were stored. The signal-to-noise ratio was about 10 (±1 patterns, sparsity 0.5, optimal learning rule).

good to remember that the class of possible learning rules is larger as the linearity, locality, and additivity assumptions could be dropped (see Discussion). After learning the Ω patterns we have

$$w_{ij}=\sum_{\nu=1}^{\Omega}[\alpha+2\beta+\gamma]+[\gamma-\alpha]x_i^{\nu}+[\gamma-\alpha]x_j^{\nu}+[\alpha-2\beta+\gamma]x_i^{\nu}x_j^{\nu}$$

In addition, we impose that $w_{ii}=0$. Setting these diagonal terms zero simplifies the calculation, while in the simulations the results are virtually identical with or without this assumption. As a matter of convenience we introduce the variables $p_0=(\alpha+2\beta+\gamma)/4$, $p_1=(\gamma-\alpha)/4$ and $p_2=(\alpha-2\beta+\gamma)/4$, so that

$$w_{ij}=4\sum_{\nu=1}^{\Omega}p_0+p_1x_i^{\nu}+p_1x_j^{\nu}+p_2x_i^{\nu}x_j^{\nu}.$$

After this substitution the task is to find the optimal values for p_0, p_1, and p_2. General principles can guide us to find the optimal p variables. We would expect identical network performance when we scale the all the weights $w_{ij}\rightarrow fw_{ij}$, where f is an arbitrary constant. This indicates that rather than having a unique optimal learning rule, we expect a family of optimal learning rules (line in α,β,γ-space). Similarly, adding a constant to all the weights $w_{ij}\rightarrow w_{ij}+w_0$, should not change the signal to noise rate. However, as we imposed $w_{ii}=0$, this invariance is broken.

1.1.1 Optimal signal to noise ratio

To find the optimal learning we calculate the signal-to-noise ratio (Amit (1989); Dayan and Willshaw (1991)). The idea is that we present a set containing both familiar and novel patterns to the

network, and test whether the network correctly identifies them. The familiar patterns will give a higher energy than the novel ones. However, each familiar pattern will yield a slightly different energy, so that we have a distribution of energies for the familiar patterns, Fig. 1. Similarly, we have a distribution of energies for the novel patterns. Overlap in the distributions will lead to errors; the better the two distributions are separated, the better the performance will be. The separation is measured through the signal to noise ratio (SNR).

To calculate the signal to noise ratio we need the average response to a familiar (high) and novel (low) pattern and the corresponding variances. The mean response to a familiar pattern is defined as

$$\langle E_F \rangle = \langle \frac{1}{\Omega} \sum_{\mu=1}^{\Omega} E^\mu \rangle$$

where the subscript 'F' indicates familiar. Without loss of generality we assume that we have an equal number of random novel patterns (with the same sparseness) with which we test the network. These novel patterns are labeled $\mu > \Omega$, so that μ is distinct from all learned, familiar patterns. The mean response to a novel pattern is therefor

$$\langle E_N \rangle = \langle \frac{1}{\Omega} \sum_{\mu=\Omega+1}^{2\Omega} E^\mu \rangle$$

where the subscript 'N' indicates novel. The signal to noise ratio is defined as

$$SNR = \frac{[\langle E_F \rangle - \langle E_N \rangle]^2}{\frac{1}{2}\langle E_F^2 \rangle_c + \frac{1}{2}\langle E_N^2 \rangle_c} \tag{1}$$

in which we denote the variance (or second cumulant) with $\langle E^2 \rangle_c$. In the Appendix we calculate the SNR as a function of the learning rule constants, the number of the units in the network, the number of patterns stored, and the pattern sparseness. Once the expression for the SNR is known, it can be optimized w.r.t. the p_0, p_1, and p_2 parameters to find the optimal learning rule and the corresponding SNR.

1.2 The optimal learning rule

As the calculation, which is the core of this work, is rather involved, the theoretical value for the SNR and the optimal learning rules are given in the Appendix. Calculation of the SNR and subsequent optimization of the learning rule leads to the optimal SNR depicted in Fig. 2. The wiggly lines are simulation results where we trained the network according to the optimal learning rule and subsequently measured the SNR. On the left is the result for patterns consisting of ± 1 entries. The solid line gives the result for 'average sparseness' (see above), while the dashed line is for 'fixed sparseness'.

In the limit of large m and Ω, the optimal learning rules reduce to $p_0 = \overline{x}^2$, and $p_1 = -\overline{x}$, $p_2 = 1$. This is the covariance rule, as can be seen after expression in the variables of Table 1. This gives $\alpha = (1 + \bar{x})^2$, $\beta = -1 + \bar{x}^2$, $\gamma = (1 - \bar{x})^2$. It is easily checked that this corresponds to $\Delta w_{ij} = (x_i - \bar{x})(x_j - \bar{x})$, i.e. the covariance rule. The corresponding signal-to-noise ratio for the optimal learning rule in the limit of a large network storing many patterns ($\Omega \gg m \gg 1$) is

$$SNR = \frac{m^2(1 - \overline{x}^2)}{2\Omega(1 + \overline{x}^2)} \tag{2}$$

For the parameters in Fig. 2 this is almost indistinguishable from the general result plotted there.

For average sparseness the capacity of the network deteriorates when the sparseness is small, as observed earlier (Bogacz and Brown (2002)). The reason is that for average sparseness patterns a lot of the variation amongst the pattern energies are due to fluctuation in the number of on-bits.

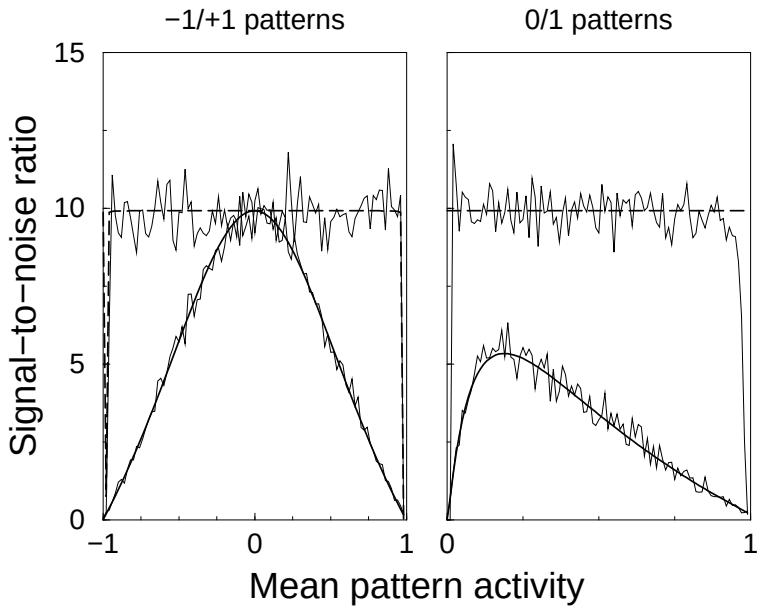

Fig. 2 The Signal-to-noise ratio for a network with optimal learning rules. Left: patterns have elements +1 and -1; right: patterns with elements 0 and 1. The dashed line is the result for fixed sparsity, the solid line for average sparsity. The thin lines denote the results of simulations, the thick lines analytical results, according to the equations in the Appendix. Parameters: $m = 100$ units, $\Omega = 500$ patterns.

These will deteriorate the signal to noise ratio. In contrast, with fixed sparseness the SNR remains at a constant level (Bogacz and Brown (2002)). The SNR for the fixed sparseness approaches $m^2/2\Omega$ (see appendix), which is 10 for the parameters in the figure. It is not hard to see that when the number of on-bits is fixed, the p_0 term in the energy, $p_0 \sum'_{i,j} x_i x_j$ is constant across patterns. Because this term is the same for familiar and novel patterns, both the mean and the variance term in the SNR are independent of p_0. The optimal learning rule therefor has the additional freedom that p_0 can take any value.

1.3 The case of 0/1 patterns

Next we analyze the case for which the units takes values 0 and 1, rather than ± 1. Perhaps unexpectedly, the results depend on this choice. We can make a transformation of variables defining $y_i = \frac{1}{2} + \frac{1}{2}x_i$. Although the weight matrix has the same form after this transformation, the energy has not. The energy for pattern μ

$$\begin{aligned} E_y^\mu &= \sum_{i,j} y_i^\mu w_{ij} y_j^\mu \\ &= \frac{1}{4}\sum_{i,j} x_i^\mu w_{ij} x_j^\mu + \frac{1}{2}\sum_{i,j} w_{ij} x_i^\mu + \frac{1}{4}\sum_{i,j} w_{ij} \\ &= \frac{1}{4}E_x^\mu + \frac{1}{2}\sum_{i,j} w_{ij} x_i^\mu + \text{const.} \end{aligned}$$

Here the first term is the scaled energy term in terms of the ± 1 patterns, the third term is just a constant that does not enter in the SNR, but the second term is new. In other words, the energy measure used to discriminate familiar from novel patterns is not invariant under the transformation from the 0/1 to the ± 1 case. Of course, one could add compensation terms to make the energy for

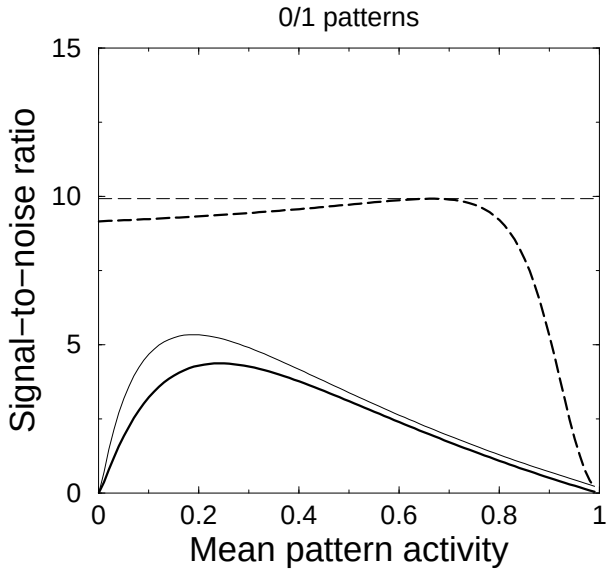

Fig. 3 Even for moderate sized networks ($m = 100$ units, $\Omega = 500$ patterns), the covariance rule yields sub-optimal SNR. The thin lines denote the SNR resulting from using optimal learning rule (which includes finite size effects), the thick line show the SNR using the covariance rule. Dashes line are for fixed sparseness, solid lines are for average sparseness. In larger networks this effect disappears.

± 1 and 0/1 patterns identical, but here we choose to explore the differences. A similar issue occurs for the Hopfield auto-associator network where it has led to considerable discussion (Tsodyks and Feigelman (1988); Horner (1989); Dayan and Willshaw (1991)).

On the right in Fig. 2 the result is shown for patterns consisting of 0's and 1's. In the case of fixed sparseness the SNR is again independent of $\bar{x}$ and high. But here the difference between fixed and average sparseness is even more clear. Here the network with average sparseness does not reach a SNR of 10. In the limit of large m and Ω, the learning rule again reduces to the covariance rule $p_0 = \overline{x}^2$, and $p_1 = -\overline{x}$, $p_2 = 1$. Or using the variables of Table 1, $\alpha = \bar{x}^2$, $\beta = \bar{x}(1-\bar{x})$, $\gamma = (1-\bar{x})^2$. In the limit of large networks the covariance rule yields an SNR of

$$SNR = \frac{1}{2}\frac{m^2\bar{x}(1-\bar{x})}{m(1-\bar{x})^2 + \Omega x(1+\bar{x})}$$

An interesting feature of the average sparseness 0/1 patterns, is that the optimal sparsity depends on the number of patterns to be stored. In the limit of large networks the maximal SNR is obtained when $\bar{x} = [1 - \sqrt{2\Omega/m}]^{-1}$, i.e. when many patterns are stored they should be sparse. This contrasts to the ± 1 case where the optimal sparseness is always $\bar{x} = 0$, independent of the number of patterns.

Although the theory says that the covariance rule to be optimal for large networks, the approximation to that solution is quite slow. The correction for small networks storing few patterns is shown in Fig. 3. The SNR is plotted for the optimal learning rule (thin lines), and the covariance rule (thick lines). Although the results are similar, it is obvious that the covariance is sub-optimal despite the decent size of the network.

1.4 Information capacity per synapse

So far we have calculated the signal to noise ratio of the familiarity detector. Next, one can set a threshold on the SNR which signifies good storage and calculate how many patterns can be stored.

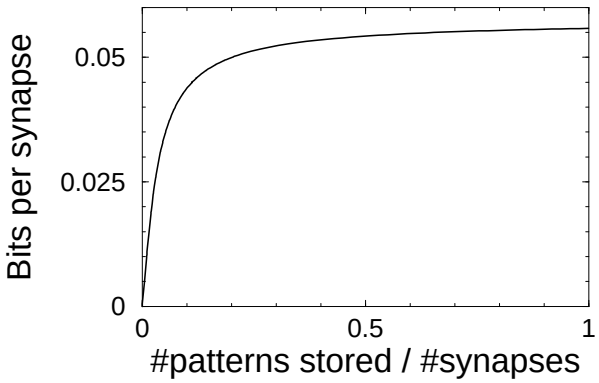

Fig. 4 The capacity of the network in bits per synapse, versus the number of patterns stored divided by the number of synapses in the network.

For instance, from Eq.(2), if the threshold on the SNR is set to 1, one obtains that one can store $\Omega_{max} = \frac{1}{2}m^2$ patterns. The exact value of the threshold is however a bit arbitrary.

An alternative measure of the capacity of the network, is the Shannon information capacity. It expresses how much information about the novelty pattern is gained by observation of the energy. Consider first that the signal to noise ratio is very large, so that no errors are made during test. For each pattern just one bit of information is stored (familiar/novel). When Ω patterns are stored the capacity is $C = \Omega/m^2$ bits per synapse (the network contains $m(m-1) \approx m^2$ synapses). However, as more and more patterns are stored the SNR reduces, and less information can be used as patterns will be misclassified.

To estimate how many bits are lost by the reduced SNR, we assume an equal number of familiar and novel patterns and a discrimination threshold halfway between the peaks of the two distributions (Fig. 1). The noise entropy is then $H_{noise}(p_e) = -(1-p_e)\log_2(1-p_e) - p_e \log_2 p_e$. The p_e is the probability for mis-classification which can be expressed in the SNR as $p_e = \frac{1}{2}\mathrm{erfc}\left(\frac{\sqrt{SNR}}{2\sqrt{2}}\right)$. Assuming patterns with fixed sparseness, the optimal learning rule yields $SNR = m^2/2\Omega$, and the error rate is $p_e^{opt} = \frac{1}{2}\mathrm{erfc}\left(\frac{m}{4\sqrt{\Omega}}\right)$. The number of bits stored per synapse is

$$C = \frac{\Omega}{m^2}[1 - H_{noise}(p_e^{opt})]$$

As more and more patterns are stored, the higher the capacity, Fig. 4. Although storing more patterns leads to a lower SNR and more errors, this is offset by the increased number of patterns. For $\Omega \to \infty$, the capacity converges to $C = (8\pi \log 2)^{-1} \approx 0.057$.

2 Discussion

Our results show that in the limit of large networks the covariance rule is the optimal learning rule for familiarity discrimination. The use of the covariance rule was proposed earlier (Bogacz and Brown (2003)), but here its optimality is explicitly demonstrated. This opens the possibility to combine both the amiliarity and episodic (Hopfield) components into one network using the same units and learning rules (Greve, Donaldson and van Rossum, submitted).

However, there a few subtleties. The first is that the network's performance is in general better when the number of 'on' bits in the patterns is fixed as was analyzed earlier (Bogacz and Brown (2002)). In the case that the number of on-bits is constant and the same for each pattern, the optimal SNR is independent of the sparseness. Biologically, precisely tuned inhibition could cause that always the same number of neurons is active. Whether this is the case in biology has not been studied to our knowledge.

In the case that the number of on-bits is probabilistic, the optimal SNR decreases with decreasing sparseness. Secondly, the learning rule is quite sensitive to deviations from the optimal learning when the network is not large or only a few patterns are stored.

We also analyzed the storage capacity using Shannon mutual information. We found that each synapse stores maximally 0.057 bits/synapse. This is still substantially less as the capacity reached in hetero and auto-associator networks, which is typically 0.1..1 bits per synapse (Meunier and Nadal (1995); Brunel (1994)). This opens the possibility then that better familiarity learning rule might exist.

The learning rule we used was linear, local, and additive. Each pattern adds a certain amount to the weights, independently of the other patterns (additive). The weight change only depends on the activation of the two units that it connects (local). It is not known whether this is the optimal learning for the familiarity network when this restriction is dropped.

In the case of the Hopfield network, the covariance rule gives 0.14 bits/synapse for dense patterns. For instance, the pseudo-inverse rule is has a higher capacity (1 bit per synapse), but it is neither linear, nor local (Kanter and Sompolinsky (1987); Hertz et al (1991)).

A related issues is that it is unclear whether the energy is the best familiarity detector imaginable. As an alternative, Hopfield originally suggested to use the temporal derivative of the energy (Hopfield (1982)). This indeed can be used as a familiarity detector, but its performance is not superior to the energy (Cortes, Greve, Barrett and van Rossum, submitted).

3 Appendix

We present the details of the SNR calculation, first for the case that the patterns consist of -1 and $+1$ and the case of 'average sparsity' in which is bit is set with a certain probability.

3.0.1 Mean energy

First we calculate the mean response to a familiar pattern. The result is

$$\begin{aligned}\langle E_F\rangle &= \langle \frac{1}{\Omega}\sum_{\mu=1}^{\Omega} E^\mu\rangle \\ &= \langle \frac{1}{\Omega}\sum_{i,j=1}^{m}{}' \sum_{\mu,\nu=1}^{\Omega} x_i^\mu x_j^\mu [p_0 + p_1 x_i^\nu + p_1 x_j^\nu + p_2 x_i^\nu x_j^\nu]\rangle \end{aligned} \tag{3}$$

where the prime in the sum indicates that $i \neq j$, which follows from the condition that $w_{ii} = 0$. Below we will also use this notation to sum over patterns for which $\mu \neq \nu$.

When the network is repeatedly probed with different sets of patterns, the energy fluctuates around its true mean, see e.g. Fig. 2. The angular brackets denote the average over different choices of the set of random patterns to be stored. We assume that the patterns are uncorrelated, so that

$$\begin{aligned}\langle x_i^\mu x_j^\nu\rangle &= (1-\delta_{ij}\delta_{\mu\nu})(\overline{x})^2 + \delta_{ij}\delta_{\mu\nu}\ \overline{x^2} \\ &= (1-\delta_{ij}\delta_{\mu\nu})\bar{x}^2 + \delta_{ij}\delta_{\mu\nu} \end{aligned} \tag{4}$$

This correlation structure is simple: when all indices are the same we use the fact that $(x_i^\mu)^2 = 1$, while otherwise we replace x_i^μ with its mean value. This means that we need to proceed by separating out the terms for the indices are equal. We find that

$$\langle E_F\rangle = m(m-1)\left\{p_0\Omega\bar{x}^2 + 2p_1[\bar{x} + (\Omega-1)\bar{x}^3] + p_2[1 + (\Omega-1)\bar{x}^4]\right\} \tag{5}$$

For the derivation we used the following considerations: For the first term in Eq.(5) we directly apply Eq. (4) in Eq. (3). The second term in Eq. (3) can be written as

$$\langle \sum_{i,j=1}^{m}{}' \sum_{\mu,\nu=1}^{\Omega} x_i^\mu x_j^\mu x_i^\nu \rangle = \langle \sum_{i,j=1}^{m}{}' \sum_{\mu,\nu=1}^{\Omega} \delta_{\mu\nu} x_i^\mu x_j^\mu x_i^\nu + \sum_{i,j=1}^{m}{}' \sum_{\mu,\nu=1}^{\Omega}{}' x_i^\mu x_j^\mu x_i^\nu \rangle$$
$$= \Omega m(m-1)\bar{x} + \Omega(\Omega-1)m(m-1)\bar{x}^3$$

This can be seen as a signal and a noise term. The first part is the energy contribution of the test pattern with its stored version (or $\mu = \nu$). The second part is an interference term describing the energy contribution of other stored patterns ($\mu \neq \nu$).

For the third term in Eq. (3) we again split of the terms for which $\mu = \nu$ and $\mu \neq \nu$, and get

$$\langle \sum_{i,j=1}^{m}{}' \sum_{\mu,\nu=1}^{\Omega} x_i^\mu x_j^\mu x_i^\nu x_j^\nu \rangle = \langle \sum_{i,j=1}^{m}{}' \sum_{\mu,\nu=1}^{\Omega} \delta_{\mu\nu} x_i^\mu x_j^\mu x_i^\nu x_j^\nu + \sum_{i,j=1}^{m}{}' \sum_{\mu,\nu=1}^{\Omega} (1-\delta_{\mu\nu}) x_i^\mu x_j^\mu x_i^\nu x_j^\nu \rangle$$
$$= \Omega m(m-1) + \Omega(\Omega-1)m(m-1)\bar{x}^4$$

Adding all contributions yields Eq. (5).

The calculation is easily repeated for the mean energy for novel patterns. The difference is that now all test patterns are distinct from the stored patterns, in other words there is no $\delta_{\mu\nu}$-term. This is reflected having μ run from $\Omega + 1$ to 2Ω.

$$\langle E_N \rangle = \langle \frac{1}{\Omega} \sum_{\mu=\Omega+1}^{2\Omega} E^\mu \rangle$$
$$= m(m-1)\Omega \left\{ p_0 \bar{x}^2 + 2p_1 \bar{x}^3 + p_2 \bar{x}^4] \right\}$$

For the signal to noise ratio we need the difference in energies of familiar and novel patterns, which is

$$\langle E_F \rangle - \langle E_N \rangle = m(m-1) \left\{ 2p_1[\bar{x} - \bar{x}^3] + p_2[1 - \bar{x}^4] \right\}$$
$$= m(m-1)(1-\overline{x}^2) \left\{ 2p_1 x + p_2(1+\bar{x}^2) \right\} \tag{6}$$

Note that the difference in energy increases quadratically in m. This is because energy is proportional to the number of synapses, which scales with the network size squared. The difference in energy is also proportional to $(1 - \bar{x}^2)$, which means it reduces as the patterns become more sparse.

3.0.2 Variance of the familiarity

To calculate the SNR we also need the variance in the energy of both familiar and novel patterns The variance for the energies for novel patterns is defined as $\langle E_N^2 \rangle_c = \frac{\Omega}{\Omega-1} \langle \frac{1}{\Omega} \sum_{\mu=\Omega+1}^{2\Omega} E^\mu E^\mu - \langle E_N \rangle^2 \rangle$. However, it would be incorrect to use the theoretical value for the $\langle E_N \rangle$ derived above. Namely, from trial-to-trial the mean energy will vary. Therefor the spread around the sample mean is lower than around the average mean. This is also known as the quenching or dispersion correction (Dayan and Willshaw (1991)). The size of the correction term disappears for a large number of patterns. However, the quantity that is relevant is the number of on-bits and when the patterns are sparse, the number of on-bits can be quite small even for a large number of patterns. By explicitly writing out the sums these corrections can be dealt with.

We define $f(\mu,\nu) = \sum_{ij}{}' x_i^\mu x_j^\mu (p_0 + p_1 x_i^\nu + p_1 x_j^\nu + p_2 x_i^\nu x_j^\nu)$ so that the energy for a pattern is $E_\mu = \sum_\nu f(\mu,\nu)$. Now

$$\begin{aligned}\langle E_N^2\rangle_c &= \frac{\Omega}{\Omega-1}\langle\frac{1}{\Omega}\sum_{\mu,\nu,\lambda} f(\mu,\nu)f(\mu,\lambda) - \frac{1}{\Omega^2}\sum_{\mu,\kappa} E_\mu E_\kappa\rangle \\ &= \frac{1}{\Omega-1}\sum_{\mu,\nu,\lambda}\langle f(\mu,\nu)f(\mu,\lambda)\rangle - \frac{1}{\Omega(\Omega-1)}\sum_{\mu,\kappa,\nu,\lambda}\langle f(\mu,\nu)f(\kappa,\lambda)\rangle \\ &= \frac{1}{\Omega}\sum_{\mu,\nu,\lambda}\langle f(\mu,\nu)f(\mu,\lambda)\rangle - \frac{1}{\Omega(\Omega-1)}\sum_{\mu,\kappa}{}'\sum_{\nu,\lambda}\langle f(\mu,\nu)f(\kappa,\lambda)\rangle\end{aligned}$$

where the sum over μ and κ is over the novel patterns, i.e. from $\Omega+1,\ldots,2\Omega$, while ν,λ indicate sums over learned patterns $(1\ldots\Omega)$. However, we need one additional step, as ν can be equal to λ.

$$\begin{aligned}\langle E_N^2\rangle_c &= \frac{1}{\Omega}\sum_{\mu,\nu,\lambda}\langle f(\mu,\nu)f(\mu,\lambda)\rangle - \frac{1}{\Omega(\Omega-1)}\sum_{\mu,\kappa}{}'\sum_{\nu,\lambda}{}'\langle f(\mu,\nu)f(\kappa,\lambda)\rangle \\ &\quad -\frac{1}{\Omega(\Omega-1)}\sum_{\mu,\kappa}{}'\sum_{\nu,\lambda}\delta_{\nu,\lambda}\langle f(\mu,\nu)f(\kappa,\lambda)\rangle \\ &= \frac{1}{\Omega}\sum_{\mu}\sum_{\nu,\lambda}{}'\langle f(\mu,\nu)f(\mu,\lambda)\rangle + \frac{1}{\Omega}\sum_{\mu,\nu}\langle f(\mu,\nu)f(\mu,\nu)\rangle \\ &\quad -\frac{\Omega-1}{\Omega}\langle E_N\rangle^2 - \frac{1}{\Omega(\Omega-1)}\sum_{\nu}\sum_{\mu,\kappa}{}'\langle f(\mu,\nu)f(\kappa,\nu)\rangle \qquad (7)\end{aligned}$$

We expand all terms w.r.t. to the p-variables, this gives terms with products of up to 8 terms in x. As above in the calculation of the mean energy, we seperate the x's in which the indices are equal from those in which they are different. Repeated application of Eq.(4) and careful counting, allows us to calculate these terms as above, yielding

$$\begin{aligned}\langle E_N^2\rangle_c = 2m(m-1)(1-x^2)\Omega\,\{&p_0^2\Omega[1+(2m-3)x^2] + 4p_0p_1\Omega x[1+(2m-3)x^2] \\ &+2p_0p_2\Omega x^2[1+(2m-3)x^2] + 2p_1^2[1+(m+2\Omega-2)x^2+(1-6\Omega-m+4m\Omega)x^4] \\ &+4p_1p_2[x+(m+\Omega-2)x^3+(1-3\Omega-m+2m\Omega)x^5] \\ &+p_2^2[1+x^2+(-5+2m+\Omega)x^4+(\Omega-1)(2m-3)x^6] \qquad (8)\end{aligned}$$

This is a complex expression, but all terms can be checked independently against simulations by varying the parameters p's in the learning rule.

We also need the variance for familiar patterns. For small Ω, the variance of familiar and novel energies are unequal. For the familiar patterns, one has a slightly more complicate expression for Eq. (7),

$$\begin{aligned}\langle E_F^2\rangle_c = 2m(m-1)(1-x^2)\{&p_0^2\Omega^2[1+(2m-3)x^2] + \\ &4p_0p_1\Omega[(m+\Omega-2)x+(\Omega-1)(2m-3)x^3] + 2p_0p_2\Omega(\Omega-1)[x^2+(2m-3)x^4] + \\ &2p_1^2[m+\Omega-2+(\Omega-1)(-8+5m+2\Omega)x^2+2(2m-3)(\Omega-1)^2x^4] \\ &4p_1p_2(\Omega-1)[x+(-4+2m+\Omega)x^3+(\Omega-1)(2m-3)x^5] + \\ &p_2^2(\Omega-1)[1+x^2+(-5+2m+\Omega)x^4+(2m-3)(\Omega-1)x^6]\} \qquad (9)\end{aligned}$$

From the above expressions, Eqs.(6), (8), and (9), the signal-to-noise ratio follows from Eq. (1).

3.1 Optimal learning rule

We differentiate the SNR w.r.t. to p_0 and p_1 and solve the equations to yield their optimal values, labeled p_i^{opt}

$$p_0^{opt} = p_2^{opt}\,\frac{\bar{x}^2}{\Omega Z}\left[-(\Omega-1)(2\Omega+m-2)+[-2+4m-2m^2+2\Omega+m^2\Omega-2m\Omega^2]\overline{x}^2+\right.$$
$$\left.[(8-9m+2m^2-30\Omega+31m\Omega-7m^2\Omega+18\Omega^2-18m\Omega^2+4m^2\Omega^2)]\overline{x}^4\right]$$
$$p_1^{opt} = -p_2^{opt}\,\frac{\overline{x}^3}{Z}\left[(m-2)(2\Omega-3)\{1+(2m-3)\overline{x}^2\}\right] \quad (10)$$

Where

$$Z = 2\Omega+m-2+(6\Omega m+m^2-8\Omega-8m+10)\overline{x}^2+(4m^2\Omega-10m\Omega-7m^2+6\Omega+21m-16)\overline{x}^4 \quad (11)$$

As expected from the scale invariance of the weights, both p_0 and p_1 are linear in p_2. This extra degree of freedom can easily be fixed without loss of generality by setting $p_2 = 1$.

In the limit that $\Omega\overline{x}^2 \gg m\overline{x}^2 \gg 1$, only the highest order terms in $\overline{x}$ remain in the nominator, Eqs.(10) and denominator, Eq.(11). We have $p_0 = \overline{x}^2$, and $p_1 = -\overline{x}$, $p_2 = 1$. This is the covariance rule.

The case of patterns with 0/1 elements is similar, but the correlation structure is different, namely $\langle x_i^\mu x_j^\nu\rangle = (1-\delta_{ij}\delta_{\mu\nu})\bar{x}^2 + \delta_{ij}\delta_{\mu\nu}\,\bar{x}$, but the derivation goes similar.

3.2 Fixed sparseness -1/+1

When the number of on-bits in the pattern is fixed, the pattern correlation structure is more complicated. For all patterns it holds that $\sum_i x_i^\mu = \bar{x}$. This yields for instance

$$\sum_{ij}{}' x_i^\mu x_j^\mu = \sum_i x_i^\mu \sum_j x_j^\mu - \sum_i (x_i^\mu)^2$$
$$= m^2\bar{x}^2 - m$$

Higher order terms are also non-trivial, for instance $\sum_{ij}{}' \sum_{k,l}{}' x_i^\mu x_j^\mu x_k^\mu x_l^\mu = m^2(m^2\bar{x}^4 - 2m\bar{x}^2+1)$.

The resulting expression for the SNR is independent of p_0. The reason is that when the number of on-bits is fixed, the p_0 term is constant, independent of the pattern, hence it does not contribute to the SNR. As a result the expression for the SNR is more compact

$$SNR = \frac{m(m-2)(m-3)[2(m-1)p_1x+(m(1+x^2)-2)p_2]^2}{2(\Omega-1)\{2m(m-1)(m-2)(m-3)p_1^2x^2+4m^2(m-1)(m-3)p_1p_2x^3+[-8+12m-6m^2+m^3+4m^2x^2-2m^3x^2-3m^3x^4+2m^4x^4]p_2^2\}}$$

The optimal learning is obtained when, after setting $p_2 = 1$, $p_1 = -(mx^2-m+2)/[mx(m-3)] \approx -x$.

Finally, when the patterns entries are 0 and 1 and pattern sparseness is fixed, we find

$$SNR = \frac{(m-2)(m-3)[2(m-1)2p_1+(m+mx-1)p_2]^2}{2(\Omega-1)\{2(m-1)(m-2)(m-3)p_1^2+4(m-1)(m-3)(-1+mx)p_1p_2+[-3+2m+m^2+6m(1-m)x+m^2(-3+2m)x^2]p_2^2\}}$$

Again for the optimal learning rule, the value of p_0 does not matter. If we set $p_2 = 1$, then $p_1 = -(mx-2)/(m-3) \approx -x$.

The optimal SNR for 'fixed sparseness' for both 0/1 and ±1 patterns is $SNR_{opt} = \frac{m^2-m-2}{2(\Omega-1)}$ (for any m, Ω) , which is independent of the sparseness.

3.3 Acknowledgements

We would like to thank Alessandro Treves for discussion. A. G. was supported by the EPSRC Doctoral Training Centre in Neuroinformatics.

References

Amit D (1989) Modeling brain function: The worldof attractor neural networks. Cambridge University Press

Bogacz R, Brown MW (2002) The restricted influence of sparseness of coding on the capacity of familiarity discrimination networks. Network 13(4):457–485

Bogacz R, Brown MW (2003) Comparison of computational models of familiarity discrimination in the perirhinal cortex. Hippocampus 13:494–524

Bogacz R, Brown MW, Giraud-Carrier C (2001) Model of familiarity discrimination in the perirhinal cortex. J of Comput Neurosc 10:5–23

Brunel N (1994) Storage capacity of neural networks: effect of the fluctuations of the number of active neurons per memory. Phys A 27:4783–4789

Dayan P, Willshaw DJ (1991) Optimising synaptic learning rules in linear associative memories. Biol Cybern 65:253–265

Fortin NJ, Wright SP, Eichenbaum H (2004) Recollection-like memory retrieval in rats is dependent on the hippocampus. Nature 431:188–191

Hertz J, Krogh A, Palmer RG (1991) Introduction to the theory of neural computation. Perseus, Reading, MA

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci 79:2554–2558

Horner H (1989) Neural networks with low levels of activity: Ising vs. McCulloch-Pitts neurons. Zeitschrift für Physik B Condensed Matter 75(1):133–136

Kanter I, Sompolinsky H (1987) Associative recall of memory without errors. Phys Rev A 35:350–392

Meunier C, Nadal JP (1995) Sparsely coded neural networks. In: Arbib MA (ed) The handbook of Brain theory, 1st edition, MIT press, Cambridge, MA

Nadal JP, Toulouse G (1990) Information storage in sparsely coded memory nets. Network 1:61–74

Tsodyks MV, Feigelman MV (1988) The enhanced storage capacity in neural networks with low activity level. Europhys Lett 6:101–105

Willshaw DJ, Buneman OP, Longuet-Higgins HC (1969) Non-holographic associative memory. Nature 222:960–993

Yakovlev V, Amit DJ, Romani S, Hochstein S (2008) Universal memory mechanism for familiarity recognition and identification. J Neurosci 28(1):239–248, DOI 10.1523/JNEUROSCI.4799-07.2008, URL http://dx.doi.org/10.1523/JNEUROSCI.4799-07.2008

Yonelinas AP (2001) Components of episodic memory: the contribution of recollection and familiarity. Proc R Soc B 356:1363–1374