To appear in Neural Compuation

Inhomogeneities in heteroassociative memories with linear learning rules

David C. Sterratt and David Willshaw

20th December 2006

Abstract

We investigate how various inhomogeneities present in synapses and neurons affect the performance of feedforward associative memories with linear learning, a high level network model of hippocampal circuitry and plasticity. The inhomogeneities incorporated into the model are: differential input attenuation, stochastic synaptic transmission and memories learnt with varying intensity. For a class of local learning rules, we determine the memory capacity of the model by extending previous analysis. We find that the signal to noise ratio (SNR), a measure of fidelity of recall, depends on the coefficients of variation (CVs) of the attenuation factors, the transmission variables, and the intensity of the memories, as well as the parameters of the learning rule, pattern sparsity and the number of memories stored. To predict the effects of attenuation due to extended dendritic trees, we use distributions of attenuations appropriate to unbranched and branched dendritic trees. Biological parameters for stochastic transmission are used to determine the CV of the transmission factors. The reduction in SNR due to differential attenuation is surprisingly low compared to the reduction due to stochastic transmission. Training a network by storing memories at different intensities is equivalent to using a learning rule incorporating weight decay. In this type of network, new memories can be stored continuously at the expense of older ones being forgotten (a 'palimpsest'). We show that there is an optimal rate of weight decay that maximises the capacity of the network, which is a factor of e lower than its non-palimpsest equivalent.

1 Introduction

Associative memory network models resemble the circuitry and presumed function of the CA3 and CA1 areas of the hippocampus (McNaughton and Morris, 1987; Treves and Rolls, 1994; Levy, 1989), the mushroom bodies of the insect olfactory system (Laurent and Naraghi, 1994; Huerta et al., 2004) and the mammalian olfactory cortex (Haberly and Bower, 1989). These networks can store memory patterns for later recall until the memory capacity of the network is reached. The dependence of the capacity on the number of units in the network, and other parameters such as the sparseness of memory patterns or connectivity, has been calculated for many variants of the associative memory model (Willshaw et al., 1969; Willshaw, 1971; Anderson, 1972; Kohonen, 1972; Hopfield,

1982; Amit et al., 1985; Palm, 1988; Dayan and Willshaw, 1991; Palm and Sommer, 1996; Graham and Willshaw, 1997).

The neurons in high-level associative memory models are 'point' neurons. In reality, biological neurons have electrically extended dendritic trees which attenuate distal inputs more than proximal ones *en route* to the soma, a phenomenon we call *differential attenuation*. The synapses in high-level models are deterministic. By contrast, biological synapses exhibit *stochastic transmission* in both the occurrence and magnitude of postsynaptic synaptic currents.

The first aim of this paper is to incorporate these inhomogeneities into a high-level associative memory model to determine how much they affect the capacity of the memory. Differential attenuation is of particular interest in the context of experimental data showing that mean synaptic conductances increase with distance from the soma (Magee and Cook, 2000), leading to somatic excitatory postsynaptic potential (EPSP) amplitudes that are independent of distance, when the neuron is quiescent.

The second aim of this paper is to determine the capacity of the network when the different memories are stored with differing intensities in the network. This is partly motivated by evidence that behavioural stress at the time of learning leads to greater synaptic synaptic potentiation or depression (Xu et al., 1997), suggesting that memories learnt in particularly significant contexts may have more intense traces. Our study of networks with variable storage intensities is also motivated by 'forgetful' learning rules, where the intensity of the traces of memories decays with time. Various types of associative memory with weight decay can be used to eliminate old memories as new ones are learnt (Willshaw, 1971; Nadal et al., 1986). Networks with this property are called *palimpsests* by analogy with the ancient practice of cleaning old texts from papyrus to make way for new ones, leaving a faint impression of the original text (Nadal et al., 1986).

The associative memory model studied is the heteroassociative memory network with linear learning (Willshaw, 1971; Palm, 1988; Willshaw and Dayan, 1990; Dayan and Willshaw, 1991; Chechik et al., 2001). This network allows us to specify arbitrary local learning rules such as heterosynaptic long term depression (Lynch et al., 1977) or the covariance rule (Sejnowski, 1977b). The network comprises an input layer of binary-valued neurons connected by real-valued feedforward synaptic weights to an output layer of binary-valued neurons. During the training phase, the network learns to associate activity patterns on the output layer with input activity patterns. Each pair of patterns is stored by changing each synaptic weight by an amount defined by the learning rule, which is a function only of the activity in the two neurons the synapse connects. This dependence only on activity local to the synapse, but not on the activity of other neurons in the network, means the learning rule is classified as a local learning rule. Since there are four possible combinations of pre- and postsynaptic activity at a synapse, four parameters define the learning rule. A previously stored output pattern is recalled by the network by each output neuron computing the weighted sum of the input pattern vector and thresholding this quantity appropriately. The network is linear in the sense that the sum of the synaptic changes over all patterns determines the synaptic strength, in contrast to associative memory models where the weights are clipped at an upper value (Willshaw et al., 1969).

The performance of the network depends strongly on how the threshold is set. Clearly if it is set very low, all output units will be active for any input pattern presented, or conversely, will be always off if the threshold is set too high. This suggests that there is an optimum threshold. Signal to noise ratio analysis can be used to show what the

optimal performance is (Palm, 1988; Palm and Sommer, 1996; Dayan and Willshaw, 1991; Chechik et al., 2001).

A critical assumption about setting the threshold is whether all output units have the same threshold or whether each output unit can have its own threshold. Palm and coworkers (Palm, 1988; Palm and Sommer, 1996) made the assumption that all output units have the same threshold, which can be adjusted to optimise performance. Their signal to noise ratio (SNR) analysis (Palm and Sommer, 1996) shows that in general there is a finite limit on the capacity of the network, regardless of the number of input units. The only exception to this are networks in which the covariance learning rule is operating, where the SNR depends linearly on the number of input units. Chechik et al. (2001) considered how to rescue the 'ineffective' learning rules by a homoeostatic neuronal regulation mechanism similar to the activity-dependent scaling of synaptic weights observed in biology (Turrigiano et al., 1998). This has the effect of normalising the weights onto each postsynaptic neuron, and leads to the capacity of the network scaling linearly with the number of input units. This is mathematically equivalent to a restriction of the class of possible learning rules, and there is a mapping from any ineffective learning rule

In contrast Dayan and Willshaw (Dayan and Willshaw, 1991; Willshaw and Dayan, 1990) allowed each output unit to have its own threshold, which can be adjusted to optimise performance. Interestingly, subsequent experimental work has shown that neurons can adjust their level of excitability homoeostatically, so as to maintain a constant average level of output activity (Desai et al., 1999). Their SNR analysis showed that with optimal thresholds, there are two classes of learning rules. In *balanced* learning rules (Sejnowski, 1977a) the mean change in synaptic weights is zero and the capacity increases linearly with the number of input units. In *unbalanced* rules the mean change in synaptic weights is nonzero and the capacity increases with the square root of the number of input units. The covariance learning rule (Sejnowski, 1977b) is a balanced learning rule, and is in fact optimal for randomly generated memory patterns; the standard Hebbian rule is an example of an inferior unbalanced learning rule.

The strategies of optimising performance by synaptic neuronal regulation or by individual optimal thresholds are compatible. The set of learning rules produced by the neuronal regulation mechanism of Chechik et al. (2001) are all *balanced*, so neuronal regulation operating in a network with individual optimal thresholds and with an unbalanced learning rule will improve the scaling of the capacity with the size of the network.

1.1 Biological background

The different types of inhomogeneity we study are: differential attenuation of inputs; stochastic synaptic transmission and different numbers of repetitions of each pattern during the training phase.

Differential input attenuation: Excitatory postsynaptic potentials (EPSPs) tend to attenuate *en route* from synapse to soma because of the cable properties of passive dendrites, the amount of attenuation varying with the path distance of the synapse from the soma (Rall, 1964). Magee and Cook (2000) found that the mean EPSP amplitude of Schaffer collateral synapses measured at the soma of a hippocampal CA1 cell does not depend on distance *in vitro*. This was due to the synaptic conductances being scaled according to distance so that distal synapses had higher conductance synapses than more proximal ones (Andrásfalvy and Magee, 2001). Whether this result extends to *in vivo*

conditions is a subject to debate. London and Segev (2001) used a passive model of a dendritic tree to suggest that *in vivo* synaptic scaling would be 'self-defeating', since larger distal synaptic conductances imply a reduction in membrane resistance and consequently reduce the electrotonic length, leading to smaller EPSPs from more distal synapses. However, this model left out a number of features that might rescue synaptic scaling (Magee and Cook, 2001) such as active, amplifying conductances (Magee and Johnston, 1995; Lipowsky et al., 1996; Gillessen and Alzheimer, 1997) and proximal shunting inhibition. While it is possible that active conductances reduce location-dependence of synaptic efficacy and time course (Rudolph and Destexhe, 2003), it is unlikely that all such differences can be eliminated; the attenuation suffered by inputs from different parts of the tree may fluctuate with the level of background activity.

Stochastic synaptic transmission: The release of synaptic vesicles in response to action potentials at CA3 boutons is stochastic, with a transmission probability ranging between 0.06 and 0.63 (Hessler et al., 1993; Stricker et al., 1996) though perhaps as high as 0.8 in potentiated states (Stevens and Wang, 1994; Bolshakov et al., 1997). Measurements of the quantal variability (QV) of excitatory postsynaptic currents (EPSCs) at CA3–CA1 synapses vary from under 0.1 (Stricker et al., 1996) to around 0.3 or 0.45 at potentiated synapses (Bolshakov et al., 1997; Forti et al., 1997).

Inhomogeneous learning intensities: Some memories may be learnt more robustly than others. This could be because they appear relatively frequently or because one of a host of molecules linked to behaviour changes the intensity of Long Term Potentiation (LTP) and/or Long Term Depression (LTD) during their storage (Sanes and Lichtman, 1999). For example stressed animals have reduced LTP and increased LTD (Shors et al., 1989; Xu et al., 1997). The intensity of a memory may decay through time. Chronic recordings *in vivo* suggested that LTP in various forebrain areas has dual exponential decay with a fast time constant of around 1.5 hours and a slow time constant of around five days (Racine et al., 1983). More recent recordings suggest that the persistence of LTP depends on the intensity of the induction protocol and the richness of the environment in which the animals are kept after induction (Abraham et al., 2002). With a weak protocol the synaptic strength falls back to baseline exponentially with a decay time constant of around a day, regardless of the environment. LTP resulting from more intense stimulation protocols can be stable for up to a year when the animals are kept in an unstimulating environment after LTP induction, but this gives way to exponential decay with a timescale of days when the animals are kept in a more stimulating environment (Abraham et al., 2002). These results suggest that learning new memories causes synaptic weights to decay. The dependence of the decay time constant on the strength of induction hints at synapses with states with different persistences, as modelled in a recent paper by Fusi et al. (2005). An alternative hypothesis is that stronger memories are rehearsed more often during sleep, leading to their greater persistence (Geszti and Pázmándi, 1987).

1.2 Theoretical background

Previous theoretical work has analysed the effects of certain inhomogeneities in associative memory networks. Graham (2001) studied differential attenuation using an associative network with a clipped Hebbian learning rule (Willshaw et al., 1969) embedded in a compartmental model of a hippocampal CA1 cell with stochastic synapses. He showed

that this reduced the SNR found in an abstract network by about 40%, for a particular loading level of the network. Scaling synapses to compensate for distance increased the SNR by about 5%, and various other strategies such as amplifying active conductances also improved performance. Whether the inputs arrived synchronously or asynchronously affected the SNR, depending on the type of compensation used. Stochastic transmission has been analysed in autoassociative networks with inhibitory neurons (Bennett et al., 1994) where it facilitates the recall of memories from partial cues, though it also degrades the retrieval state slightly.

Inhomogeneities in memory intensity have been studied extensively. Willshaw (1971) investigated probabilistic weight decay in associative networks with binary weights. He showed that in an associative net with binary-valued synapses, randomly switching off previously activated synapses enabled the memory to act as a palimpsest, but at the expense of forcing the memory to function under non-optimal conditions. Probabilistic weight decay where the probability of decay depends on time has also been studied (Henson and Willshaw, 1995). Hopfield (1982) suggested both weight decay and keeping the weights between prescribed maxima and minima as methods for allowing networks to continue learning new memories whilst forgetting old ones. Nadal et al. (1986) studied a network where each new memory is stored more intensely than the previous one. This guarantees perfect recall of the last stored pattern and, according to simulations, partial recall for around half of the number of memories that could be stored by a standard network. Analytical mean field studies of networks with weight decay followed (Mézard et al., 1986; van Hemmen and Zagrebnov, 1987). The capacity of a palimpsest Hopfield network was found to be about $1/e$ of a standard Hopfield network of the same size (Mézard et al., 1986). Networks incorporating the 'learning within bounds' feature have the palimpsest property, as shown numerically (Parisi, 1986; Nadal et al., 1986), analytically with a combined signal to noise and random walk analysis (Gordon, 1987) and by a sophisticated analysis including a Markov chain representation of the iterative learning procedure (van Hemmen et al., 1988). If the bounds exceed a certain threshold level they have little effect and the performance of the network deteriorates catastrophically, whereas if they are very small, only the most recently stored memory pattern is retrieved accurately. Again, capacity is about $1/e$ that of a standard Hopfield network, for the optimal bound.

2 The model and key results

2.1 The model

Our model is a generalisation of the mathematical framework introduced by Palm (1988) and developed by Willshaw and Dayan (1990). We choose Willshaw and Dayan's development of the theory over that of Palm and Sommer (1996) and Chechik et al. (2001) because each unit is assumed to be able to optimise its own threshold to improve performance, as appears to be the case in nature (Desai et al., 1999). As noted in section 1, the idea of 'neuronal regulation' (Chechik et al., 2001) is compatible with individual optimal thresholds.

The network comprises N associative inputs indexed by i and an unspecified number of output neurons, indexed by j. Ω memories have been stored; the ωth memory is a pair of strings $(a^{(\omega)}, b^{(\omega)})$ with binary-valued components $a_i^{(\omega)}$ and $b_j^{(\omega)}$. The typical element

$b_j^{(\omega)}$ of output pattern $b^{(\omega)}$ is assigned the 'high' value h with probability r and the 'low' value l with probability $1-r$. The typical element $a_i^{(\omega)}$ of the input pattern $a^{(\omega)}$ is assigned the 'high' value 1 with a probability p and the 'low' value c with probability $1-p$. c can take on any value apart from 1. In a standard Hopfield network it would be set at -1, but a biologically-realistic value is 0. When each output unit can have its threshold set independently, c is a scaling parameter (Dayan and Willshaw, 1991) and the results of signal to noise calculations are independent of c; we use this fact to check our calculations in this paper.

The synaptic strength from input i to output neuron j is

$$w_{ij} = \sum_{\omega=1}^{\Omega} \kappa^{(\omega)} \Delta_{ij}^{(\omega)} \quad , \tag{1}$$

where $\kappa^{(\omega)}$ is the *intensity* of the ωth memory and where the *weight contribution* $\Delta_{ij}^{(\omega)}$ depends on the input and output patterns presented during the training phase and the four parameters of the generalised local learning rule, α, β, γ and δ. These are allocated as shown in Table 1, which also gives the special cases of the unbalanced Hebbian and balanced covariance learning rules.

general

$\Delta_{ij}^{(\omega)}$		$b_j^{(\omega)}$	
		l	h
$a_i^{(\omega)}$	c	α	β
	1	γ	δ

Hebbian

$\Delta_{ij}^{(\omega)}$		$b_j^{(\omega)}$	
		l	h
$a_i^{(\omega)}$	c	0	0
	1	0	1

covariance

$\Delta_{ij}^{(\omega)}$		$b_j^{(\omega)}$	
		l	h
$a_i^{(\omega)}$	c	pr	$-p(1-r)$
	1	$-(1-p)r$	$(1-p)(1-r)$

Table 1: The general local learning rule and its Hebbian and covariance instantiations.

During recall of the output pattern associated with the ωth input pattern, the dendritic sum is calculated as

$$d_j^{(\omega)} = \sum_{i=1}^{N} w_{ij} f_i g_{ij}^{(\omega)} a_i^{(\omega)} \quad , \tag{2}$$

where f_i is the *attenuation factor* of the ith input and $g_{ij}^{(\omega)}$ is the *transmission factor* of the ijth synapse during presentation of the ωth input pattern. Inclusion of f_i and $g_{ij}^{(\omega)}$ allows the attenuation due to the geometry and electrical properties of real neurons to be incorporated in the model. We view the transmission factors as random variables which model quantal failure and variance in quantal amplitude.

Each unit has a threshold θ_j so that its output o_j takes the value h ('high') when $d_j^{(\omega)} > \theta_j$ and l ('low') otherwise. We assume that it is possible to set an optimal threshold for each output separately (Willshaw and Dayan, 1990).

As an aid to comprehension of the necessarily long calculations in section 3, we present an overview of our analysis and the key result of the paper at the beginning of the next section. In the rest of the paper, this result is applied to differential attenuation (section 5), stochastic transmission (section 6) and memories stored with different intensities (section 7).

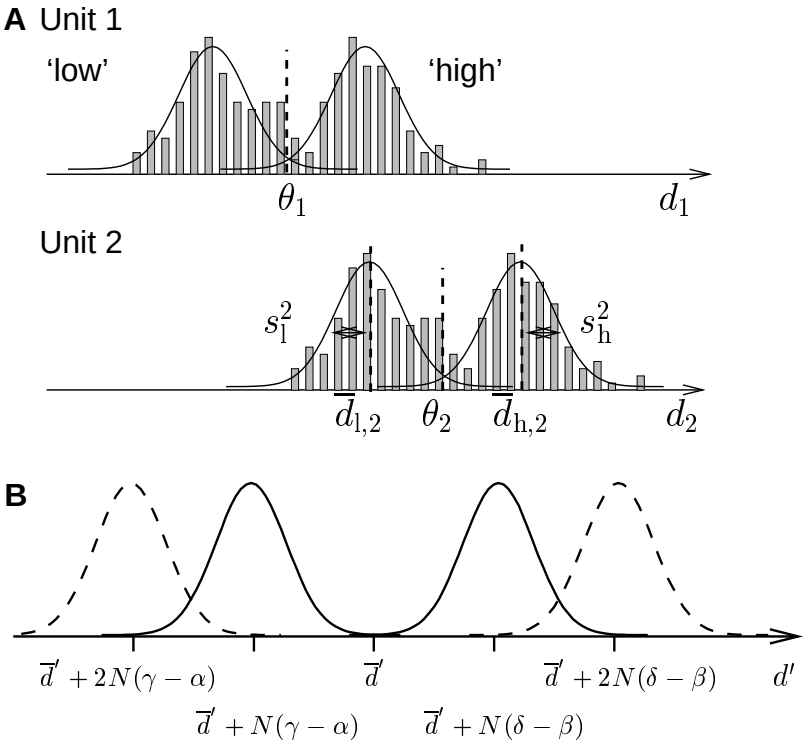

Figure 1: **A**, Schematic diagram of the 'high' and 'low' distributions of dendritic sums d_1 and d_2 of units 1 and 2 in a network. Comparing the two units, the means of the distributions are shifted with respect to each other but the separation between the high and low distributions remains the same. **B**, The distribution of dendritic sums when half the memories have an intensity κ of 1 (solid lines) and half have an intensity κ of 2 (dashed lines). In order simplify the plot, we show a normalised version of the dendritic sum d', where $d' = d/(p(1-p))$. The centres are at $\kappa N(\delta - \beta)$ and $\kappa N(\gamma - \alpha)$ relative to the mean of all dendritic sums. The distributions for $\kappa = 2$ are further apart than for $\kappa = 1$.

2.2 Overview of analysis and key results

We take the signal to noise ratio (SNR) approach to analysing associative memories (Willshaw, 1971; Anderson, 1972; Palm, 1988; Palm and Sommer, 1996; Dayan and Willshaw, 1991; Chechik et al., 2001). For autoassociative memories, mean field approaches are necessary to determine the capacity at which they break down due to catastrophic interference, but the SNR analysis usually gives the same scaling of the capacity with the number of units in the network (Hertz et al., 1991).

The SNR analysis (see Figure 1A) is based on the expected distribution of the dendritic sum for memories whose output should be 'high' ($b_j^{(\omega)} = h$) and the expected distribution of the dendritic sum for memories whose output should be 'low' ($b_j^{(\omega)} = l$). The area of the 'high distribution' to the left of the threshold gives the number of 'high' memories where the retrieved output will be 'low', and the area of the 'low distribution' to the right of the threshold gives the number of 'low' memories where the retrieved output will be 'high'. Thus the total number of erroneously recalled memories (the *bit error*) depends on the threshold, which is set so as to minimise the bit error.

The SNR is a measure of the discriminability of the distributions. It is calculated as the square of the expected difference between the means $\overline{d}_{\mathrm{h}}$ and $\overline{d}_{\mathrm{l}}$ for the high and low

patterns (the 'signal') divided by the sum of the variances $s_{\rm h}^2$ and $s_{\rm l}^2$ of the high and low dendritic sum distributions (the 'noise'):

$$\rho = \frac{(\langle \overline{d}_{\rm h} - \overline{d}_{\rm l} \rangle)^2}{\frac{1}{2}(s_{\rm h}^2 + s_{\rm l}^2)} \quad . \tag{3}$$

Willshaw and Dayan (1990) observed that whereas the difference between means for the high and low distributions is the same from unit to unit, the value of the two means themselves are shifted according to the fraction of memories with high and low outputs stored in a particular unit (Figure 1A). Therefore different units have different optimal thresholds. To ensure that the SNR is a measure of the discriminability when optimal thresholds are set, the variances of the high and low distributions in the SNR have to be computed as the sum of squared deviations from the unit high and low means of each unit individually. Using this definition of the variance, Dayan and Willshaw (1991) obtained a general expression for the SNR which scales with the number of input neurons N. If the variance is computed with respect to the means of the distributions averaged over *all* units (Palm and Sommer, 1996), the SNR tends to a limiting value for large N, except when c can be set to a biologically-dubious non-zero value.

Inserting the parameters of the covariance and Hebbian learning rules into Dayan and Willshaw's general expression (see equation 12) leads to the following expressions for the SNRs due to the covariance and Hebbian rules:

$$\rho^{\rm cov} = \frac{N}{\Omega r(1-r)} \text{ and } \rho^{\rm Hebb} = \frac{N(1-p)}{\Omega^2 p r^2} \quad . \tag{4}$$

At the number of stored memories Ω increases, the SNR decreases. The capacity $\Omega_{\rm max}$is defined as the maximum number of patterns that can be stored before the SNR falls below a desired minimum level $\rho_{\rm min}$. Setting $\rho = \rho_{\rm min}$ and $\Omega = \Omega_{\rm max}$ in these formulae shows that for the covariance rule the capacity is proportional to $N/\rho_{\rm min}$ and for the Hebbian rule the capacity is proportional to $\sqrt{N/\rho}$. In general, for balanced learning rules ($pr\delta + p(1-r)\gamma + (1-p)r\beta + (1-p)(1-r)\alpha$=0) the capacity is proportional to N and for unbalanced learning rules, it is proportional to $\sqrt{N}$ in the limit of large N (Dayan and Willshaw, 1991).

We have used the same method as Dayan and Willshaw (1991) to compute the SNR for the network incorporating differential attenuation, stochastic transmission and variable storage intensities. The attenuation and transmission factors do not change the qualitative form of the high and low distributions and the SNR. In contrast, with inhomogeneous memory intensities, the total distribution of the high or low dendritic sums is a superposition of distributions due to memories stored with different intensities.

To make this clear, we consider an example in which one half of the stored memories have been chosen at random to be twice as intense ($\kappa = 2$) as the other half ($\kappa = 1$). We now need to look at the four distributions for high and low units with $\kappa = 1$ and $\kappa = 2$ as shown in Figure 1B. The means of the dendritic sum distributions of the stronger memories are further apart than the means of the weaker ones, meaning stronger memories have a greater 'signal'. It turns out (see appendix A) that the variances are independent of κ (for unbalanced rules this is only true in a network with a large number of input units), so the 'noise' component is the same regardless of the intensity of the memory. Thus the SNR depends on the strength of the memory.

We have calculated a full expression for $\rho(\kappa)$, the SNR of a memory stored at intensity κ (equation (10) in section 3. This expression becomes much simpler when expressed in

terms of the SNR $\hat{\rho}$ of a homogeneous network with the same input and output pattern sparsities and learning rule. Assuming that c takes the biologically-plausible value of 0, when υ_f, υ_g and υ_κ are the coefficients of variation (CVs) of the attenuation factors, the transmission factors and the intensities respectively, the SNR of a pattern stored with intensity $\rho(\kappa)$ is

$$\rho(\kappa) = \frac{\kappa^2}{(1+\upsilon_f^2)(1+\upsilon_g^2/(1-p))(1+k\upsilon_\kappa^2)}\hat{\rho} \quad , \tag{5}$$

where the constant k is an expression whose form depends on whether the learning rule is balanced or unbalanced. $\rho(\kappa)$ also depends on p, r, the parameters of the learning rule, and, in the case of unbalanced learning rules, Ω.

Since the dendritic sum distributions depend on the memory intensity, at first sight it may appear that the optimum threshold should depend on the intensity of the memory being recalled, which might be difficult to arrange in biology. However, this turns out not to be a problem. As more and more memories are stored, the least intense memories in the network will fall below the criterion threshold $\rho_{\min}$ first. For any number of patterns stored Ω, there will be a memory intensity κ for which the SNR is at the criterion. If the threshold is set to minimise the bit error for memories with this intensity, the more intense memories will be retrieved with less than this bit error (even if they are not retrieved as well as they could if their own optimal threshold had been used).

This means the network capacity is still well-defined. Thus the scaling factor in equation (5) also scales the capacity of the network with balanced learning rules and its square root scales the capacity in the case of unbalanced learning rules .

A key insight from the compact expression (5) is that the reduction in performance of the network due to the inhomogeneity factors depends only on their CVs. Furthermore, the effects of differential input attenuation, stochastic transmission and inhomogeneous learning intensities are independent of each other.

3 General theory

3.1 Distribution of dendritic sums

For a particular output unit j, the sample mean of the high distribution for memories with intensity κ is

$$\overline{d}_{\mathrm{h}j}(\kappa) = \frac{1}{\Omega_{\mathrm{h}j}(\kappa)} \sum_{\{\omega : b_j^{(\omega)} = h, \kappa^{(\omega)} = \kappa\}} d_j^{(\omega)} \quad , \tag{6}$$

where $\Omega_{\mathrm{h}j}(\kappa)$ is the number of high patterns stored by unit j with an intensity of κ. In appendix A.1 we show that the expected value of $\overline{d}_{\mathrm{h}j}(\kappa)$ is

$$\left\langle \overline{d}_{\mathrm{h}j}(\kappa) \right\rangle = N \left\langle f_i \right\rangle \left\langle g_{ij}^{(\omega)} \right\rangle \left(\kappa p(1-p)(1-c)(\delta - \beta) + \left\langle \kappa \right\rangle \sigma(\Omega_{\mathrm{h}j}\phi + \Omega_{\mathrm{l}j}\psi) \right) \quad , \tag{7}$$

where $\sigma = p + (1-p)c$ is the expected value of an input, and $\Omega_{\mathrm{h}j}$ and $\Omega_{\mathrm{l}j}$ are respectively the total number of high and low memories stored in the weights of unit j. The expected sample variance or *dispersion* of the high patterns with intensity κ is

$$s_{\mathrm{h}}^2(\kappa) = \left\langle \frac{1}{\Omega_{\mathrm{h}}(\kappa)} \sum_{\{\omega : b_j^{(\omega)} = h, \kappa^{(\omega)} = \kappa\}} \left(d_j^{(\omega)} \right)^2 - \overline{d}_{\mathrm{h}j}^{\,2}(\kappa) \right\rangle \quad . \tag{8}$$

$$
\begin{array}{ll}
R_1 = p(1-p)(r(\delta-\beta)^2 + (1-r)(\gamma-\alpha)^2) & S_1 = \frac{p+(1-p)c^2}{p(1-p)(1-c)^2} R_1 \\
\quad + r(1-r)(\phi-\psi)^2 & \\
R_2 = (1-2p)(\delta-\beta+\gamma-\alpha)(r\phi+(1-r)\psi) & S_2 = \frac{c+1}{c-1} R_2 \\
R_3 = (r\phi+(1-r)\psi)^2 & S_3 = \frac{p+(1-p)c^2}{p(1-p)(1-c)^2} R_3 \\
R_4 = r(p\delta^2+(1-p)\beta^2) + (1-r)(p\gamma^2+(1-p)\alpha^2) & S_4 = \frac{p+(1-p)c^2}{p(1-p)(1-c)^2} R_4
\end{array}
\tag{11}
$$

Table 2: Expressions for the components of equation (10). $\phi = p\delta + (1-p)\beta$ is the expected weight contribution of a high pattern and $\psi = p\gamma + (1-p)\alpha$ is the expected weight contribution of a low pattern.

Analogous equations apply for $\overline{d}_{\mathrm{l}j}$, and $s^2_{\mathrm{l}}(\kappa)$.

The strict definition of the *signal to noise ratio* of memories with intensity κ is:

$$
\rho(\kappa) = \frac{\left(\left\langle \overline{d}_{\mathrm{h}j}(\kappa) - \overline{d}_{\mathrm{l}j}(\kappa) \right\rangle\right)^2}{\frac{1}{2}(s^2_{\mathrm{h}}(\kappa) + s^2_{\mathrm{l}}(\kappa))} \quad . \tag{9}
$$

We calculate it to be

$$
\rho(\kappa) = \frac{Np(1-p)(\delta-\gamma-\beta+\alpha)^2\tilde{\kappa}^2}{\Omega(1+\upsilon_f^2)\left(R_1 + R_2\tilde{\kappa} + R_3\Omega + R_4\upsilon_\kappa^2 + \upsilon_g^2(S_1 + S_2\tilde{\kappa} + S_3\Omega + S_4\upsilon_\kappa^2)\right)} \tag{10}
$$

where $\tilde{\kappa} = \kappa / \langle\kappa\rangle$ is the normalised memory intensity, υ_f^2 is the squared coefficient of variation (CV) of the attenuation factors, υ_g^2 is the squared CV of the transmission factors, υ_κ^2 is the CV of the memory intensities and the other factors are functions of the parameters α, β, γ, δ, p, r and c, as given in Table 2.

Any threshold is optimal for only one memory intensity κ. As discussed in section 2.2, we assume the threshold is optimal for memories with intensity $\kappa_{\min}$ for which the SNR is at the performance criterion $\rho_{\min}$. In the next section we show there is no dependence of the noise on κ for balanced rules, and the dependence vanishes for large N with unbalanced rules. From (7), we can see that as long as $\delta \geq \beta$, memories which are more intense than $\kappa_{\min}$ will have the mean of their high distribution further from the threshold. Since they have the same dispersion as weaker memories, they will have fewer bits omitted erroneously. A similar argument applies for the low distribution, as long as $\alpha \leq \gamma$. These conditions hold in all the learning rules we consider.

3.2 Comparison with previous analysis

By setting all the CVs to zero and $\kappa = 1$, equation (10) reduces to the expression derived by Dayan and Willshaw (1991):

$$
\hat{\rho} = \frac{Np(1-p)(\delta-\gamma-\beta+\alpha)^2}{\Omega(R_1 + R_2 + R_3\Omega)} \quad . \tag{12}
$$

In the case of balanced learning rules since R_2 and R_3 are zero,

$$
\rho(\kappa) = \frac{Np(1-p)(\delta-\gamma-\beta+\alpha)^2\tilde{\kappa}^2}{\Omega(1+\upsilon_f^2)\left(1 + \upsilon_g^2\frac{p+c^2(1-p)}{p(1-p)(1-c)^2}\right)(R_1 + R_4\upsilon_\kappa^2)} \quad . \tag{13}
$$

From this, the value of k in (5) is calculated to be R_4/R_1.

In the case of unbalanced learning rules at large Ω, the R_1 terms are negligible. As long as $\kappa^{(\omega)} \ll \langle\kappa\rangle\,\Omega$, the R_2 terms can be neglected too. Thus the terms in Ω and in υ_κ^2 dominate the denominator of (10) to give

$$\rho(\kappa) \approx \frac{Np(1-p)(\delta-\gamma-\beta+\alpha)^2\tilde{\kappa}^2}{\Omega^2(1+\upsilon_f^2)\left(1+\upsilon_g^2\frac{p+c^2(1-p)}{p(1-p)(1-c)^2}\right)(R_3+R_4\upsilon_\kappa^2/\Omega)} \quad . \tag{14}$$

From this, the value of k in (5) is calculated to be $R_4/(R_3\Omega)$.

This shows that the reduction in SNR due to (i) differential attenuation, (ii) stochastic transmission and (iii) differential memory intensity combine multiplicatively.

4 Simulations

In order to confirm our theory, simulation results are presented alongside the theoretical SNR curves in some of the figures in the rest of this paper. In the simulations, a network with $N = 1000$ input units and 100 output units learns randomly-generated patterns with $p = 0.2$ and $r = 0.2$ using the covariance rule. The mean and dispersion of the 'high' and 'low' dendritic sums are computed for each of the output units.

The sample mean of the difference in dendritic sums is an estimator for the components $\left\langle \overline{d}_{\mathrm{h}} - \overline{d}_{\mathrm{l}} \right\rangle$ of the SNR, and the sample error in the mean is the sample standard deviation of the differences in dendritic sums, divided by the square root of the number of output units. Likewise the sample mean and error in the mean of the denominator $s_{\mathrm{h}}^2 + s_{\mathrm{l}}^2$ of the SNR can be calculated. We used these values to compute the sample SNR and combined the errors to obtain the error in the SNR.

The simulation code was written in the R language (http://www.r-project.org) and is available from http://www.anc.ed.ac.uk/∼dcs/pubs/inhomog-assoc-net .

5 Differential input attenuation

We now study the effects of differential input attenuation on memory performance. To do this we assume that transmission and memory intensity are homogeneous ($\upsilon_g = \upsilon_\kappa = 0$) by setting $g_i^{(\omega)} = \kappa^{(\omega)} = 1$. The only inhomogeneity remaining is in the attenuation factors f_i. Using (5), for an arbitrary distribution of attenuation factors, the general expression for the SNR (10) reduces to:

$$\rho = \frac{1}{1+\upsilon_f^2}\hat{\rho} \quad . \tag{15}$$

This formula shows that differential attenuation simply reduces the SNR of the homogeneous network by a factor only involving the coefficient of variation of the attenuation factors. The capacity is reduced at most by the same factor (for balanced learning rules) or by the square root of the factor (for unbalanced learning rules in the limit of large N). Since the reduction in SNR is independent of all of the parameters of the network, the differential attenuation has no effect on the optimality of the learning rule or on the dependence of the capacity on network size.

In the remainder of this section, we apply equation (15) to various distributions of f_i that might arise out of the spatial distribution of inputs on a dendritic tree and the geometry of the tree itself.

5.1 Unbranched dendrite, linear attenuation

Our first application of equation (15) is to an unbranched dendrite of uniform thickness, with a uniform distribution of inputs per unit area and a linear dependence of attenuation on distance. Linear dependence has the virtue of simplicity and is a good approximation to the more biologically realistic exponential case when the dendrite is shorter than its electrotonic length. In this case the factors f_i will be uniformly distributed. For factors distributed uniformly between 1 and F, and approximating sums by integrals, we obtain:

$$\rho^{\text{linatt}}(F) = \frac{3}{4}\left(1 + \frac{1}{F + 1 + 1/F}\right)\hat{\rho}$$

Figure 2A shows the theoretical curve, which was confirmed by simulations.

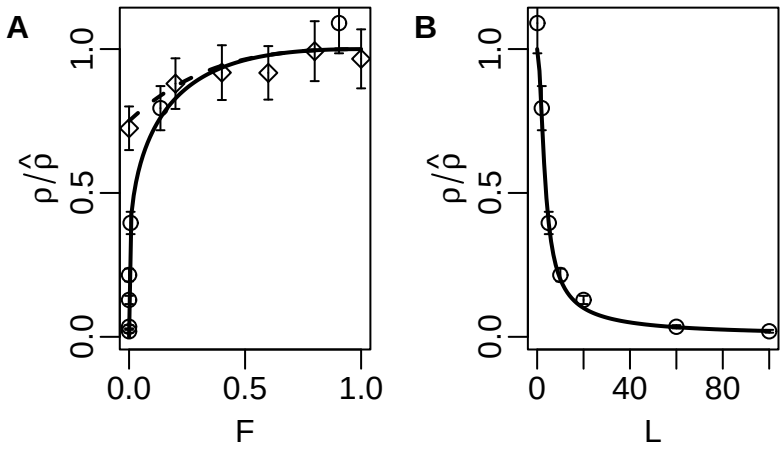

Figure 2: Reduction in SNR for unbranched dendrites with linear and decaying exponential attenuation. Lines show theoretical predictions and points indicate simulation results. **A**, Reduction in SNR ($\rho/\hat{\rho}$) versus attenuation at distal end of dendrite F for linear attenuation (dashed line, diamonds) and exponential attenuation (solid line, circles; $F = e^{-L}$). **B**, Reduction in SNR with exponential attenuation versus length of dendrite L. In the simulations, $\Omega = 500$ patterns have been stored in the network.

In data recorded in CA1 neurons *in vitro*, the attenuation of the distal inputs is roughly double that of the more proximal ones (Magee and Cook, 2000), implying F=1/2. For this value of F, $\rho^{\text{linatt}} \approx 0.96\hat{\rho}$. This reduction in SNR equates to a 4% reduction in capacity (with balanced learning rules) or 2% (with unbalanced learning rules).

Under *in vivo* conditions we might expect the range of attenuation to be greater because of background inputs to the neuron putting it into a high conductance state (London and Segev, 2001; Destexhe et al., 2003). Assuming that in this case there is a tenfold difference between the attenuation of the proximal and distal inputs, $\rho^{\text{linatt}} \approx 0.82\hat{\rho}$, reducing the capacity by 18% at most. In the limiting case, where $F \to \infty$, $\rho^{\text{linatt}} \to 3/4\rho$, giving a maximum reduction in capacity of 25%.

5.2 Unbranched dendrite, exponential attenuation

The next step up in biological plausibility is a decaying exponential dependence of attenuation on distance, as predicted by cable theory applied to passive dendrites. If X is the

electrotonic distance along an unbranched dendrite, then

$$f = \xi(X) = e^{-X} \ ,$$

where the dendrite extends from $X = 0$ to $X = L$. The corresponding distribution of f is proportional to the inverse of the gradient of the attenuation:

$$p(f) \propto 1/|\xi'(X)| = 1/|\xi'(\xi^{-1}(f))| = 1/f \ . \tag{16}$$

Calculating the squared CV of this distribution and substituting the result substituted in equation (15), we obtain

$$\rho^{\text{expatt}} = \frac{2(1 - e^{-L})}{L(1 + e^{-L})}\hat{\rho} \ .$$

The SNR is plotted versus electrotonic length in Figure 2B. For electrotonically long dendrites (large L), the SNR of the network decays to zero. To compare exponential attenuation with linear attenuation, Figure 2A shows the SNR as a function of the amount of attenuation at the end (position $X = L$) of the dendrite $F = e^{-L}$. For $F = 1/2$, the performance is virtually unchanged from the linear attenuation case. For $F = 1/10$ (the estimate under high-conductance conditions) the ratio of attenuated to non-attenuated SNR is 0.71, as compared to 0.82 in the linear attenuation case.

5.3 Branched dendrite, exponential attenuation

We now consider branched dendrites, such as those found in CA1 and CA3 cells. Assuming that the number of inputs per unit length is constant, there will be a greater fraction of inputs further away from the soma. This is in broad agreement with anatomical work suggesting most of the input to CA1 cells is on the oblique branches (Megías et al., 2001). We characterise the density of inputs as a function of electrotonic length by

$$\zeta(X) = e^{X/D} \ ,$$

where D is the characteristic branching distance. This approximates to a situation where the distance (in units of electrotonic length) between successive bifurcations is $D \ln 2$. Incorporating the input density in (16) leads to the distribution of attenuations as a function of the electrotonic length L and the branching distance D:

$$p(f) \propto \zeta(X)/|\xi'(X)| = \zeta(\xi^{-1}(f))/|\xi'(\xi^{-1}(f))| = (1/f)^{1/D+1} \ .$$

Calculating the squared CV of this distribution and substituting it into equation (15), the SNR is

$$\rho^{\text{expatt,branched}} = \frac{(1 - 2D)\left(e^{L(1/D-1)} - 1\right)^2}{(D-1)^2(e^{L/D} - 1)(e^{L(1/D-2)} - 1)}\hat{\rho} \ .$$

The reduction in SNR for branched dendrites is shown in Figure 3. As the bifurcation length becomes large relative to the length of the dendrites, the ratio approaches the value for the unbranched dendrites. For small D compared to L, the performance approaches the unattenuated case. This seemingly paradoxical result is because with profusely branching dendrites, most of the area of the dendrites is concentrated near the tips, so that most of the inputs are attenuated equally. The ratio has a minimum at $D = 1$ of $\frac{L^2}{(1-e^{-L})(e^L-1)}$.

To estimate how much performance might be reduced in hippocampal CA1 cells, we took L to be 2, following Stricker et al. (1996) who found that Schaffer collateral synapses on CA1 cells of synapses were located between 0.3 and 2.1 length constants from the soma. For $L = 2$, the value of the minimum is 0.72, a reduction by a factor of 1.4.

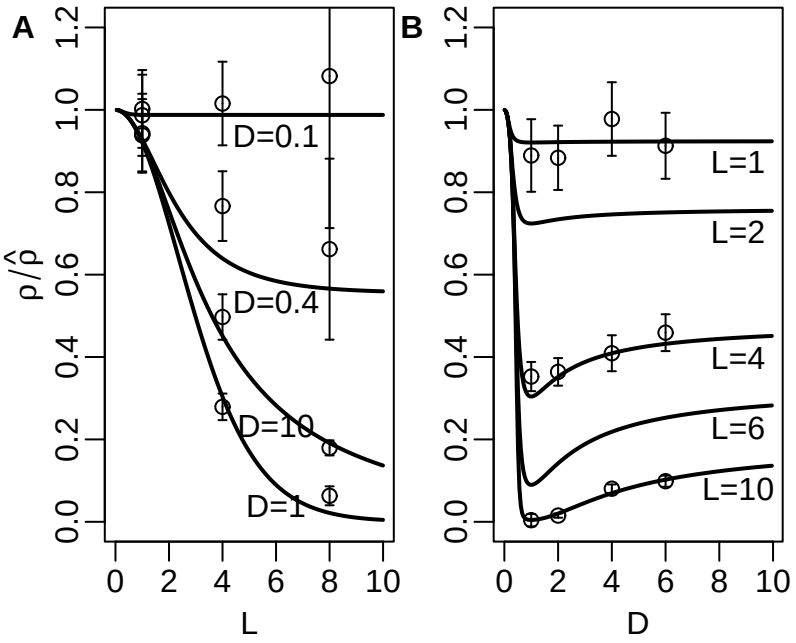

Figure 3: Fractional reduction of SNR $\rho/\hat{\rho}$ in branched dendrites where the attenuation decays exponentially as a function of the branching length constant D. **A**, Reduction as a function of electrotonic length L for various values of D. Performance decreases monotonically with L for all values of D. **B**, Reduction as a function of D for various values of L. Note that there is no decrease in performance for very rapidly branching cells ($D \to 0$), and that the worst performance always occurs when $D = 1$.

6 Stochastic synaptic transmission

We now examine the case with homogeneous attenuation factors and memory intensity and inhomogeneous, stochastic transmission (non-zero υ_g^2 and $\upsilon_f^2 = \upsilon_\kappa^2 = 0$; $f_i = \kappa^{(\omega)} = 1$ for all i and ω).

Rather than using equation (5), where c was set to zero, the theory presented in section 3 is required to derive the most general expression for the reduction in SNR due to stochastic transmission. From the general formula (10) and Table 2, the stochastic transmission reduces the SNR by a factor of

$$\left(1 + \upsilon_g^2 \frac{p + c^2(1-p)}{p(1-p)(1-c)^2}\right)^{-1} \quad . \tag{17}$$

In contrast to the cases studied so far, the value of the low state c appears in the formula. This appears to contradict Dayan and Willshaw (1991), who showed that the SNR is independent of the value of c for all $c \neq 1$. We investigate this by defining canonical inputs $\hat{a}_i^{(\omega)} \in \{0, 1\}$ and deriving the actual inputs from them: $a_i^{(\omega)} = (1-c)\hat{a}_i^{(\omega)} + c$. Then the postsynaptic sum is given by

$$d_j^{(\omega)} = (1-c)\sum_{i=1}^{N} w_{ij}\hat{a}_i^{(\omega)} g_{ij}^{(\omega)} + c\sum_{i=1}^{N} w_{ij} g_{ij}^{(\omega)} \quad .$$

With homogeneous transmission ($g_{ij}^{(\omega)} = 1$) the second term is the same for all ω and so represents a translation. Stochastic transmission causes this term to vary with ω and this

adds to the dispersion of the dendritic sum distribution. Similarly, whereas the first term represents a stretch in the homogeneous case, it adds to the dispersion in the stochastic case.

To predict the effect of stochastic transmission in the hippocampus, we assume synapses release single quanta with a transmission probability t and amplitude q. The squared CV of g is

$$v_g^2 = \frac{v_q^2 + 1 - t}{t} \ , \tag{18}$$

where v_q is the CV of q. Measurements from the experimental literature suggest a maximum value of v_g^2 of about 10 ($t = 0.1$, $v_q = 0.1$; Stricker et al., 1996) and a minimum of 0.5 in a potentiated state ($t = 0.8$, $v_q = 0.45$; Bolshakov et al., 1997). A CV of about 2 may be more realistic for unpotentiated synapses ($t = 0.4$, $v_q = 0.3$; Bolshakov et al., 1997). We estimate the range of p to be 0.2–0.3 on the basis of *in vivo* recordings from CA3 (Barnes et al., 1990; Leutgeb et al., 2004). Substituting these values into equation (17) with $c = 0$, we derive an upper limit on the reduction in SNR of $1 + 10/0.7 \approx 15$ and a lower limit of $1 + 0.5/0.8 \approx 2$. For a CV of 2 we estimate the SNR reduction factor to be $1 + 2/0.8 = 3.5$.

7 Inhomogeneous memory intensities

We now consider networks with homogeneous attenuation factors and perfectly reliable transmission but with inhomogeneous memory intensities. This means that $\kappa^{(\omega)}$ is not uniform and $v_f = v_g = 0$; $f_i = g_i^{(\omega)} = 1$ for all i and ω.

There are a number of ways of imagining how such inhomogeneous memory intensities might be established in a network. For example, particularly significant memories might be learnt with greater weights, or might be rehearsed more, leading to greater weights over a period of time. These scenarios are probably best studied in the context of continuous learning and forgetting of memories. Accordingly, here we apply the general result to the more familiar case of a network where the intensities are graded so that effectively there is weight decay.

7.1 Weight decay

We suppose that the network is learning continuously ($\Omega \to \infty$) with effectively weight decay occurring between each presentation of pattern pairs. Thus each memory's intensity is $\kappa^{(\omega)} = e^{-\frac{\omega}{\tau}}$ where τ is the weight decay (or 'forgetting') time constant and where the index ω now measures the 'age' of a memory; it is 1 for the most recent. This distribution of κ has $\langle\kappa\rangle \approx \tau/\Omega$, $\langle\kappa^2\rangle \approx \tau/(2\Omega)$, $\sigma_\kappa^2 \approx \tau/(2\Omega)$ and $v_\kappa^2 = \Omega/(2\tau)$. Substitution of these quantities into equation (5) leads to a general expression for the SNR of a memory with age ω:

$$\rho(\omega) = \frac{Np(1-p)(\delta - \gamma - \beta + \alpha)^2 e^{-\frac{2\omega}{\tau}}}{\tau\left(R_2 e^{-\frac{\omega}{\tau}} + R_3\tau + R_4/2\right)} \ . \tag{19}$$

For balanced learning rules the R_2 and R_3 terms in the denominator vanish, so the SNR of an ageing memory decays exponentially with a time constant $\tau/2$. For unbalanced learning rules this relation holds approximately, especially for older patterns.

7.2 Palimpsests with balanced learning rules

We now analyse the covariance learning rule ($\alpha = (1-p)(1-r), \beta = -p(1-r), \gamma = -(1-p)r, \delta = (1-p)(1-r)$) as an example of a balanced learning rule. The analysis applies equally to the homosynaptic ($\alpha = 0, \beta = 0, \gamma = -r, \delta = 1-r$) and heterosynaptic ($\alpha = 0, \beta = -p, \gamma = 0, \delta = 1-p$) instantiations of balanced learning rules. The formula for the SNR depends on the precise learning rule, but our basic finding – the dependence of the capacity of the network on the decay time constant – remains unchanged.

Substituting the covariance learning rule into (19) leads to the following expression for the SNR:

$$\rho^{\mathrm{cov}}(\omega) = \frac{2Ne^{-\frac{2\omega}{\tau}}}{r(1-r)\tau} \quad . \tag{20}$$

Figure 4 shows the expected SNR and error levels as a function of the age of the patterns for three different forgetting time constants τ. For small τ, the network forgets quickly. For large τ the network's memory is longer but the interference from older patterns degrades the performance of the more recent patterns. In the figure we have chosen an SNR of 10 to define successful retrieval; this is shown by the horizontal dashed line. The intermediate value of τ shown in the plot provides the largest capacity at this level of performance. Simulations confirmed the results shown in Figure 4.

Rearranging (20) leads to an expression for the capacity in terms of τ and the minimum SNR of $\rho_{\min}$:

$$\Omega_{\max} = \frac{\tau}{2}(\ln 2\hat{\Omega}_{\max} - \ln \tau) \tag{21}$$

where

$$\hat{\Omega}_{\max} = \frac{N}{r(1-r)\rho_{\min}} \tag{22}$$

is the capacity of a homogeneous network with the minimum SNR $\rho_{\min}$. The dependence of $\Omega_{\max}$ on τ is shown in Figure 5. For $\tau > 2\hat{\Omega}_{\max}$ it is not possible for the network to perform at the specified SNR of $\hat{\rho}$ because of interference from older memories. The optimal value of $\Omega_{\max}$ is $\hat{\Omega}_{\max}/e$ which occurs at $\tau = 2\hat{\Omega}_{\max}/e$. Thus the palimpsest property reduces the capacity of the network by a factor of at least e.

The scaling of the capacity of the network with number of synapses N in an output unit depends on whether the time constant is scaled with N or not. If $\tau = kN/(r(1-r)\rho_{\min})$ for $0 < k < 2$, then $\Omega_{\max}$ will scale with N. In this case the initial SNR is independent of N: $\rho(1) = 2\rho_{\min}e^{-2/\tau}/k$. In contrast, if τ is fixed, the capacity only grows as $\ln N$, but the initial SNR is proportional to N. In the Discussion (section 8), we consider whether this is reasonable.

7.3 Palimpsests with a Hebbian rule

Applying equation (19) to the Hebbian learning rule, we obtain an expression for the SNR:

$$\rho^{\mathrm{Hebb}}(\omega) = \frac{N(1-p)e^{-\frac{2\omega}{\tau}}}{\tau r((1-2p)e^{-\frac{\omega}{\tau}} + pr\tau + 1/2)} \quad . \tag{23}$$

In the limit $N \to \infty$ the denominator is dominated by the terms quadratic in τ and we can compare the palimpsest capacity with the standard capacity in the large N limit. This gives a reduction in capacity by a factor of e, as in the balanced case, though the optimal time constant is a factor of 2 smaller at $\hat{\Omega}_{\max}/e$.

With the decay time constant matched to the number of inputs, the memory lifetime is proportional to $\sqrt{N}$ and the initial performance is independent of N. When the decay time constant is fixed, the initial SNR is proportional to N and the lifetime is proportional to $\ln(\sqrt{N})$.

8 Discussion

In this paper we have derived a general expression for the capacity of a heteroassociative memory with continuous weights trained with a general local learning rule with differential input attenuation, stochastic synaptic transmission and inhomogeneous memory intensities. This work extends that of Dayan and Willshaw (1991), which considered completely homogeneous networks.

As far as we are aware, ours is the first analysis of differential attenuation in a mathematically-tractable associative network, though an associative network embedded in a multicompartmental CA1 model was studied numerically by Graham (2001). His approach is more biologically-grounded than ours and can be used to predict the effect of active conductances or the precision of timing of the inputs. The geometry of the cell is an integral part of Graham's compartmental model, rather than being imposed on the model as we have done. Any interactions between inputs are ignored in our model. For example, a distal input might be boosted by activation of proximal NMDA receptors. Nevertheless, our model suggests that it is the spread of effective attenuations (as measured by the CV) that is important rather than the precise dynamics of attenuation.

Stochastically-firing units are often considered in capacity calculations of Hopfield networks (Hertz et al., 1991) and stochastic transmission has been incorporated in associative network models (Bennett et al., 1994; Graham, 2001). Numerical analysis of an autoassociative network model of CA3 shows that stochastic firing reduces the capacity of the network, but enhances its ability to recall a pattern from a partial cue (Bennett et al., 1994). Our results also indicate that stochastic firing decreases capacity, though we cannot make the comparison with autoassociative recall dynamics as they are absent from our model.

Weight decay has been studied in binary-weighted associative networks (Willshaw, 1971; Henson and Willshaw, 1995) and Hopfield networks (Mézard et al., 1986; Nadal et al., 1986; van Hemmen and Zagrebnov, 1987). We have incorporated weight decay into an associative model with arbitrary pattern sparsity and local, linear learning rules. In common with Mézard et al. (1986), our approach also covers arbitrary distributions of memory intensities, as well as those arising from weight decay.

8.1 The effects of differential attenuation and stochastic transmission

We have considered how much differential attenuation and stochastic transmission are likely to affect the network performance of the CA3–CA1 network.

In section 5.3 we found the worst-case reduction in capacity with uniform inputs over a branching dendritic tree was a factor of 1.4, assuming the dendritic tree is 2 electrotonic lengths long (Stricker et al., 1996). Mechanisms such as synaptic scaling and active conductances should lead to a tree that is electronically more compact, but the high membrane conductance might lead to a greater electrotonic length.

Our assumptions about the branching structure of dendrites for CA1 neurons are only approximate, though they do suggest that the effect of branching dendrites will not be very great. A more precise estimation of the affect on the SNR could be made by using the statistics of synapse placement on CA1 such as those obtained by Megías et al. (2001).

In section 6 we used values of the sparseness of presynaptic activity, transmission probabilities and CVs of successful transmission taken from the literature to estimate that the stochastic transmission reduces the SNR by a factor of 3.5 (with a possible range of 2–15). It would appear then that transmission noise should lead to a greater decrease in SNR than differential attenuation.

Bursts of presynaptic neuronal activity can lead to reliable synaptic transmission from unreliable synapses, when the burst (rather than individual spikes) is considered as the unit of presynaptic activity (Lisman, 1997). If we assume a squared CV of around 0.3 for the postsynaptic response to a burst and a presynaptic activity $p = 0.2$, this leads to an estimate of about 1.4 for the factorial reduction in SNR, similar to the reduction due to differential attenuation.

Associative memories with linear learning therefore seem to be quite robust to differential input attenuation. This could be important in the biological neural networks such as CA3–CA1 network which might have attenuation profiles varying with the level of background activity. It also raises the question of whether synapses are scaled with distance at all (Magee and Cook, 2000), especially given the potential for this mechanism to defeat itself (London and Segev, 2001). Nevertheless, increasing the homogeneity of the input attenuations does lead to improved performance, so it is perhaps not so surprising that there should be synaptic scaling.

Graham (2001) found that the SNR was reduced by a factor of 2.5 (40%) in an associative network embedded in a compartmental model of a CA1 cell with a synaptic transmission probability of 1 and a quantal amplitude CV of 0.3. We estimate that the stochastic transmission should reduce the SNR by approximately 10%. Combined with our estimate of a reduction of 1.4 due to attenuation differences, this leads to a reduction of 1.5 in the SNR, considerably less severe than the reduction in the multicompartmental model. This discrepancy could arise from differences in the network models used or from our underestimating the effective attenuations. The capacity of the binary-weighted network used in Graham's model depends logarithmically on the number of synapses (Willshaw et al., 1969), as opposed to the linear or square-root dependence in our model. In the binary-weighted network there is no variance in the 'high' distribution but the variable attenuations will smear this out, perhaps increasing the apparent reduction in SNR. A simple test of whether the differences are due to the underlying network model or the neuron model would be to repeat Graham's simulations using a heteroassociative network with linear learning, though negative weights in this model would have to be prevented by some means.

8.2 Optimal forgetting in palimpsests

Our results suggest that for optimal capacity, the decay constant of the memories should be tuned to the number of neurons, consistent with the scaling in Hopfield networks (Nadal et al., 1986; Mézard et al., 1986). The optimal value we find for the forgetting time constant also agrees. For balanced networks, it is a factor $2/e$ times the capacity of the equivalent standard network. The tuning of the time constant need not be very precise, but does have to be less than a critical value as otherwise recall breaks down.

We have shown that if the forgetting rate is fixed, the network capacity scales only with the logarithm of the number of inputs (for balanced rules) or the logarithm of the square root of the number of inputs (for unbalanced rules).

In a model with binary synapses with states with varying levels of persistence, Fusi et al. (2005) showed that memory lifetime can scale with the number of synapses raised to a power less than one, without having to tune the forgetting time constant. This scaling is better than the logarithmic scaling we find for fixed forgetting time constant, but worse than the scaling if the time constant is scaled with the number of synapses.

The question arises of how reasonable is it to tune τ. This will not be a problem that has to be dealt with within an animal's lifetime, as we expect the number of inputs and the sparsity of the memory coding to be fairly constant. It seems feasible that τ could be tuned through evolutionary mechanisms. Our results suggest that the different forgetting time constants should appear in different associative memory systems according to the sparsity of the input and output patterns and the numbers of inputs.

Whether forgetting obeys a power law or an exponential function is a matter of some controversy in the psychophysical literature (Wixted and Ebbesen, 1997; Anderson and Tweney, 1997). In the physiological literature, long term studies suggest LTP decays exponentially (Racine et al., 1983; Abraham et al., 2002). However, LTP results from an artificial protocol, and is probably not subject to mechanisms such as rehearsal or modulation due to behavioural state (Xu et al., 1997). The general results presented in this paper could provide a framework for predicting the memory time courses arising from physiological processes.

Acknowledgements This work is carried out with the financial support of the UK Medical Research Council (Grant P9119632). Our thanks go to Kit Longden, Guy Billings, Fiona Jamieson, Jesus Cortes and other members of the Institute for Adaptive and Neural Computation for their helpful comments during preparation of this paper, and to the referees for their constructive reviews.

A Derivation of SNR

We now devise the expression for the SNR given in equation (10) and the associated relations in Table 2, in section 3.

A.1 Expected difference of high and low dendritic sums

To avoid notational clutter, we drop the j suffix of the postsynaptic neuron throughout this appendix. The expected dendritic sum for a high pattern ω_h with intensity $\kappa_\mathrm{h} = \kappa^{(\omega_\mathrm{h})}$ can be written as

$$\left\langle d^{(\omega_\mathrm{h})} \right\rangle = \sum_{i=1}^{N} \left\langle f_i \left(g_i^{(\omega_\mathrm{h})} \kappa_\mathrm{h} a_i^{(\omega_\mathrm{h})} \Delta_i^{(\omega_\mathrm{h})} + \sum_{\omega \in \mathcal{H}, \omega \neq \omega_\mathrm{h}} g_i^{(\omega)} \kappa^{(\omega)} a_i^{(\omega_\mathrm{h})} \Delta_i^{(\omega)} + \sum_{\omega \in \mathcal{L}} g_i^{(\omega)} \kappa^{(\omega)} a_i^{(\omega_\mathrm{h})} \Delta_i^{(\omega)} \right) \right\rangle$$

where $\mathcal{H} = \{\omega : b^{(\omega)} = h\}$ and $\mathcal{L} = \{\omega : b^{(\omega)} = l\}$. The attenuation factors f_i and the transmission factors $g_i^{(\omega)}$ are independent of each other and all the other variables, so their expectations can be factored out. The weight contributions are independent of

the values of c, h and l. For convenience, and without loss of generality, we define them in terms of canonical input patterns $\hat{a}_i^{(\omega)} \in \{0,1\}$ and output patterns $\hat{b}^{(\omega)} \in \{0,1\}$: $\Delta_i^{(\omega)} = \alpha(1-\hat{a}_i^{(\omega)})(1-\hat{b}^{(\omega)}) + \beta(1-\hat{a}_i^{(\omega)})\hat{b}^{(\omega)} + \gamma\hat{a}_i^{(\omega)}(1-\hat{b}^{(\omega)}) + \delta\hat{a}_i^{(\omega)}\hat{b}^{(\omega)}$. Using the fact that the $\kappa^{(\omega)}$ factors are independent from the inputs and weight increments, we can substitute in the expected values of the products of $a_i^{(\omega_\mathrm{h})}\Delta_i^{(\omega_\mathrm{h})}$ and $a_i^{(\omega)}\Delta_i^{(\omega)}$ for high and low patterns to obtain:

$$\left\langle d^{(\omega_\mathrm{h})} \right\rangle = N \left\langle f_i \right\rangle \left\langle g_i^{(\omega)} \right\rangle \left\langle \kappa_\mathrm{h}(p\delta + (1-p)c\beta) + \sum_{\omega\in\mathcal{H},\omega\neq\omega_\mathrm{h}} \kappa^{(\omega)}\sigma\phi + \sum_{\omega\in\mathcal{L}} \kappa^{(\omega)}\sigma\psi \right\rangle \quad ,$$

where $\sigma = p + c(1-p)$ is the expected activity of an input unit.

We now define Ω_h to be the number of high patterns, Ω_l the number of low patterns, $\overline{\kappa_\mathrm{h}}$ the mean of the high patterns and $\overline{\kappa_\mathrm{l}}$ the mean of the low patterns. These quantities are random, varying between output units. By adding $\kappa_\mathrm{h}\sigma\psi$ to the first sum of the above formula and taking it away from the first term and simplifying we can write this formula as:

$$\left\langle d^{(\omega_\mathrm{h})} \right\rangle = N \left\langle f_i \right\rangle \left\langle g_i^{(\omega)} \right\rangle \kappa_\mathrm{h} \left(p(1-p)(1-c)(\delta-\beta) + \left\langle \Omega_\mathrm{h}\overline{\kappa_\mathrm{h}}\sigma\phi + \Omega_\mathrm{l}\overline{\kappa_\mathrm{l}}\sigma\psi \right\rangle\right) \quad . \tag{24}$$

The equivalent formula for low patterns is

$$\left\langle d^{(\omega_\mathrm{l})} \right\rangle = N \left\langle f_i \right\rangle \left\langle g_i^{(\omega)} \right\rangle \kappa_\mathrm{l} \left(p(1-p)(1-c)(\gamma-\alpha) + \left\langle \Omega_\mathrm{h}\overline{\kappa_\mathrm{h}}\sigma\phi + \Omega_\mathrm{l}\overline{\kappa_\mathrm{l}}\sigma\psi \right\rangle\right) \quad . \tag{25}$$

Hence

$$\left\langle d^{(\omega_\mathrm{h})} - d^{(\omega_\mathrm{l})} \right\rangle = N \left\langle f_i \right\rangle \left\langle g_i^{(\omega)} \right\rangle p(1-p)(1-c)\left(\kappa_\mathrm{h}(\delta-\beta) - \kappa_\mathrm{l}(\alpha-\gamma)\right) \quad . \tag{26}$$

A.2 Dispersion of dendritic sums

The dispersion of the high patterns as defined in equation (8), can be rearranged (Dayan and Willshaw, 1991) into the form

$$\left\langle \frac{\Omega_\mathrm{h}-1}{\Omega_\mathrm{h}} \left(\left(d^{(\omega_\mathrm{h1})}\right)^2 - d^{(\omega_\mathrm{h1})}d^{(\omega_\mathrm{h2})} \right) \right\rangle \quad ,$$

where ω_{h1} and ω_{h2} index two different patterns with high outputs. An approximation to this quantity, which is tractable to compute is:

$$\left\langle \left(d^{(\omega_\mathrm{h1})}\right)^2 \right\rangle - \left\langle d^{(\omega_\mathrm{h1})}d^{(\omega_\mathrm{h2})} \right\rangle \quad .$$

A.2.1 The expectation of $\left(d^{(\omega_\mathrm{h1})}\right)^2$

This can be partitioned into a sum with N terms where the activity is from the same input units and a double sum with $N(N-1)$ terms where the activity is from different units:

$$\begin{aligned}\left\langle \left(d^{(\omega_\mathrm{h1})}\right)^2 \right\rangle = &\left\langle \sum_{i=1}^{N} f_i^2 \left(g_i^{(\omega_\mathrm{h1})}\right)^2 \sum_{\omega=1}^{\Omega}\sum_{\omega'=1}^{\Omega} \kappa^{(\omega)}\kappa^{(\omega')}\Delta_i^{(\omega)}\Delta_i^{(\omega')} \left(a_i^{(\omega_\mathrm{h1})}\right)^2 \right\rangle \\ &+ \left\langle \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} f_i f_j g_i^{(\omega_\mathrm{h1})} g_j^{(\omega_\mathrm{h1})} \sum_{\omega=1}^{\Omega}\sum_{\omega'=1}^{\Omega} \kappa^{(\omega)}\kappa^{(\omega')}\Delta_i^{(\omega)}\Delta_j^{(\omega')} a_i^{(\omega_\mathrm{h1})} a_j^{(\omega_\mathrm{h1})} \right\rangle \end{aligned} \tag{27}$$

k		T_k	V_k
1	${\kappa_\mathrm{h}}^2$	$p\delta^2 + c^2(1-p)\beta^2$	$(p\delta + c(1-p)\beta)^2$
2	$\Omega_\mathrm{h}\overline{{\kappa_\mathrm{h}}^2} - {\kappa_\mathrm{h}}^2$	$\pi(p\delta^2 + (1-p)\beta^2)$	$\sigma^2\phi^2$
3	$\Omega_\mathrm{l}\overline{{\kappa_\mathrm{l}}^2}$	$\pi(p\gamma^2 + (1-p)\alpha^2)$	$\sigma^2\psi^2$
4	$2\Omega_\mathrm{h}\kappa_\mathrm{h}\overline{\kappa_\mathrm{h}} - 2{\kappa_\mathrm{h}}^2$	$(p\delta + c^2(1-p)\beta)\phi$	$(p\delta + c(1-p)\beta)\sigma\phi$
5	$2\Omega_\mathrm{l}\kappa_\mathrm{h}\overline{\kappa_\mathrm{l}}$	$(p\delta + c^2(1-p)\beta)\psi$	$(p\delta + c(1-p)\beta)\sigma\psi$
6	${\Omega_\mathrm{h}}^2\overline{{\kappa_\mathrm{h}}^2} - (2\Omega_\mathrm{h}\kappa_\mathrm{h}\overline{\kappa_\mathrm{h}} - 2{\kappa_\mathrm{h}}^2)$ $-(\Omega_\mathrm{h}\overline{\kappa_\mathrm{h}} - {\kappa_\mathrm{h}}^2) - {\kappa_\mathrm{h}}^2$	$\pi\phi^2$	$\sigma^2\phi^2$
7	$2\Omega_\mathrm{l}\Omega_\mathrm{h}\overline{\kappa_\mathrm{h}\kappa_\mathrm{l}} - 2\Omega_\mathrm{l}\kappa_\mathrm{h}\overline{\kappa_\mathrm{l}}$	$\pi\phi\psi$	$\sigma^2\phi\psi$
8	${\Omega_\mathrm{l}}^2\overline{{\kappa_\mathrm{l}}^2} - \Omega_\mathrm{l}\overline{{\kappa_\mathrm{l}}^2}$	$\pi\psi^2$	$\sigma^2\psi^2$

Table 3: Components of $\left\langle \left(d^{(\omega_{\mathrm{h}1})}\right)^2 \right\rangle$.

Under the assumption that the attenuation and transmission factors are independent from each other and from the inputs, we can apply the expectations to each factor in the sums:

$$\begin{aligned}\left\langle \left(d^{(\omega_{\mathrm{h}1})}\right)^2 \right\rangle = N\left\langle f^2\right\rangle\left\langle g^2\right\rangle \underbrace{\left\langle \sum_{\omega=1}^{\Omega}\sum_{\omega'=1}^{\Omega} \kappa^{(\omega)}\kappa^{(\omega')}\Delta_i^{(\omega)}\Delta_i^{(\omega')}\left(a_i^{(\omega_{\mathrm{h}1})}\right)^2 \right\rangle}_{=:\,T_\mathrm{h}} \\ +N(N-1)(\langle f\rangle)^2(\langle g\rangle)^2 \underbrace{\left\langle \sum_{\omega=1}^{\Omega}\sum_{\omega'=1}^{\Omega} \kappa^{(\omega)}\kappa^{(\omega')}\Delta_i^{(\omega)}\Delta_i^{(\omega')}a_i^{(\omega_{\mathrm{h}1})}a_j^{(\omega_{\mathrm{h}1})} \right\rangle}_{=:\,V_\mathrm{h}}\end{aligned} \tag{28}$$

We define T_h to be the inner sums for the same-unit terms and V_h the inner sums for the cross-unit terms. The expectation of each of the Ω^2 terms of T_h and V_h depends on whether ω or ω' are equal to each or other or $\omega_{\mathrm{h}1}$. There are eight different types of combinations of ω, ω' and $\omega_{\mathrm{h}1}$. We index the combinations with k and denote the expectation of a combination $\left\langle \Delta_i^{(\omega)}\Delta_i^{(\omega')}\left(a_i^{(\omega_{\mathrm{h}1})}\right)^2 \right\rangle$ by T_k and $\left\langle \Delta_i^{(\omega)}\Delta_i^{(\omega')}a_i^{(\omega_{\mathrm{h}1})}a_j^{(\omega_{\mathrm{h}1})} \right\rangle$ by V_k. The expectation of the whole of the inner sum is then the sum of the products of the expectations with the sums of the intensities $\kappa^{(\omega)}\kappa^{(\omega')}$, similar to equation (A.1). We then rearrange the sums (in a similar way to equation (24)) so that we have expressions in terms of the κ_h, $\overline{\kappa_\mathrm{h}}$ and $\overline{\kappa_\mathrm{l}}$ etc. Table 3 gives the values of each of these 8 terms together with the appropriate prefactors.

From the table, we can write down an expression for T_h:

$$\begin{aligned}T_\mathrm{h} &= {\kappa_\mathrm{h}}^2(T_1 - T_2 - 2T_4 + 2T_6) + \kappa_\mathrm{h}\overline{\kappa_\mathrm{h}}\Omega_\mathrm{h}(2T_4 - 2T_6) + \kappa_\mathrm{h}\overline{\kappa_\mathrm{l}}\Omega_\mathrm{l}(2T_5 - 2T_7) \\ &\quad + \overline{{\kappa_\mathrm{h}}^2}\Omega_\mathrm{h}(T_2 - T_6) + \overline{{\kappa_\mathrm{l}}^2}\Omega_\mathrm{l}(T_3 - T_8) + (\Omega_\mathrm{h}\overline{\kappa_\mathrm{h}})^2T_6 + 2\Omega_\mathrm{h}\overline{\kappa_\mathrm{h}}\Omega_\mathrm{l}\overline{\kappa_\mathrm{l}}T_7 + (\Omega_\mathrm{l}\overline{\kappa_\mathrm{l}})^2T_8 \\ &= {\kappa_\mathrm{h}}^2p(1-p)(1-c^2)(1-2p)(\delta-\beta)^2 \\ &\quad + 2p(1-p)(1-c^2)(\delta-\beta)(\kappa_\mathrm{h}\overline{\kappa_\mathrm{h}}\Omega_\mathrm{h}\phi + \kappa_\mathrm{h}\overline{\kappa_\mathrm{l}}\Omega_\mathrm{l}\psi) \\ &\quad + \pi p(1-p)(\overline{{\kappa_\mathrm{h}}^2}\Omega_\mathrm{h}(\delta-\beta)^2 + \overline{{\kappa_\mathrm{l}}^2}\Omega_\mathrm{l}(\gamma-\alpha)^2) + \pi(\Omega_\mathrm{h}\overline{\kappa_\mathrm{h}}\phi + \Omega_\mathrm{l}\overline{\kappa_\mathrm{l}}\psi)^2\end{aligned} \tag{29}$$

We can write down a similar equation for V_h and there are analogous expressions for T_l and V_l which are obtained by interchanging Ω_h and Ω_l, δ and γ, β and α, and ϕ and ψ.

k		U_k	W_k
1	$2{\kappa_\mathrm{h}}^2$	$(p\delta^2+c(1-p)\beta^2)\sigma$	$(p\delta+c(1-p)\beta)\sigma\phi$
2	${\kappa_\mathrm{h}}^2$	$(p\delta+c(1-p)\beta)^2$	$(p\delta+c(1-p)\beta)^2$
3	${\kappa_\mathrm{h}}^2$	$(p\delta+c(1-p)\beta)^2$	$\sigma^2\phi^2$
4	$2(\Omega_\mathrm{h}\kappa_\mathrm{h}\overline{\kappa_\mathrm{h}}-2{\kappa_\mathrm{h}}^2)$	$(p\delta+c(1-p)\beta)\sigma\phi$	$(p\delta+c(1-p)\beta)\sigma\phi$
5	$2\Omega_\mathrm{l}\kappa_\mathrm{h}\overline{\kappa_\mathrm{l}}$	$(p\delta+c(1-p)\beta)\sigma\psi$	$(p\delta+c(1-p)\beta)\sigma\psi$
6	$2(\Omega_\mathrm{h}\kappa_\mathrm{h}\overline{\kappa_\mathrm{h}}-2{\kappa_\mathrm{h}}^2)$	$(p\delta+c(1-p)\beta)\sigma\phi$	$\sigma^2\phi^2$
7	$2\Omega_\mathrm{l}\kappa_\mathrm{h}\overline{\kappa_\mathrm{l}}$	$(p\delta+c(1-p)\beta)\sigma\psi$	$\sigma^2\phi\psi$
8	$\Omega_\mathrm{h}\overline{{\kappa_\mathrm{h}}^2}-2{\kappa_\mathrm{h}}^2$	$\sigma^2(p\delta^2+(1-p)\beta^2)$	$\sigma^2\phi^2$
9	$\Omega_\mathrm{l}\overline{{\kappa_\mathrm{l}}^2}$	$\sigma^2(p\gamma^2+(1-p)\alpha^2)$	$\sigma^2\psi^2$
10	${\Omega_\mathrm{h}}^2\overline{\kappa_\mathrm{h}}^2-4(\Omega_\mathrm{h}\kappa_\mathrm{h}\overline{\kappa_\mathrm{h}}-2{\kappa_\mathrm{h}}^2)$ $-(\Omega_\mathrm{h}\overline{{\kappa_\mathrm{h}}^2}-2{\kappa_\mathrm{h}}^2)-4{\kappa_\mathrm{h}}^2$	$\sigma^2\phi^2$	$\sigma^2\phi^2$
11	$4\Omega_\mathrm{l}\Omega_\mathrm{h}\overline{\kappa_\mathrm{h}\kappa_\mathrm{l}}-4\Omega_\mathrm{l}\kappa_\mathrm{h}\overline{\kappa_\mathrm{l}}$	$\sigma^2\phi\psi$	$\sigma^2\phi\psi$
12	${\Omega_\mathrm{l}}^2\overline{\kappa_\mathrm{l}}^2-\Omega_\mathrm{l}\overline{{\kappa_\mathrm{l}}^2}$	$\sigma^2\psi^2$	$\sigma^2\psi^2$

Table 4: Components of $\left\langle d^{(\omega_\mathrm{h1})}d^{(\omega_\mathrm{h2})}\right\rangle$

A.2.2 The expectation of $d^{(\omega_\mathrm{h1})}d^{(\omega_\mathrm{h2})}$

This can be partitioned similarly into N terms from the same input unit and $N(N-1)$ terms where the activity is from different units:

$$\begin{aligned}\left\langle d^{(\omega_\mathrm{h1})}d^{(\omega_\mathrm{h2})}\right\rangle = &\left\langle \sum_{i=1}^{N} f_i^2 g_i^{(\omega_\mathrm{h1})} g_i^{(\omega_\mathrm{h2})} \sum_{\omega=1}^{\Omega}\sum_{\omega'=1}^{\Omega} \kappa^{(\omega)}\kappa^{(\omega')}\Delta_i^{(\omega)}\Delta_i^{(\omega')}a_i^{(\omega_\mathrm{h1})}a_i^{(\omega_\mathrm{h2})}\right\rangle \\ &+\left\langle \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} f_i f_j g_i^{(\omega_\mathrm{h1})} g_j^{(\omega_\mathrm{h1})} \sum_{\omega=1}^{\Omega}\sum_{\omega'=1}^{\Omega} \kappa^{(\omega)}\kappa^{(\omega')}\Delta_i^{(\omega)}\Delta_i^{(\omega')}a_i^{(\omega_\mathrm{h1})}a_j^{(\omega_\mathrm{h2})}\right\rangle \end{aligned} \quad (30)$$

Again, we factor out the expectations of the attenuation and transmission factors:

$$\begin{aligned}\left\langle d^{(\omega_\mathrm{h1})}d^{(\omega_\mathrm{h2})}\right\rangle = &N\left\langle f^2\right\rangle(\langle g\rangle)^2 \underbrace{\left\langle \sum_{\omega=1}^{\Omega}\sum_{\omega'=1}^{\Omega} \kappa^{(\omega)}\kappa^{(\omega')}\Delta_i^{(\omega)}\Delta_i^{(\omega')}a_i^{(\omega_\mathrm{h1})}a_i^{(\omega_\mathrm{h2})}\right\rangle}_{=:U_\mathrm{h}} \\ &+N(N-1)(\langle f\rangle)^2(\langle g\rangle)^2 \underbrace{\left\langle \sum_{\omega=1}^{\Omega}\sum_{\omega'=1}^{\Omega} \kappa^{(\omega)}\kappa^{(\omega')}\Delta_i^{(\omega)}\Delta_i^{(\omega')}a_i^{(\omega_\mathrm{h1})}a_j^{(\omega_\mathrm{h2})}\right\rangle}_{=:W_\mathrm{h}} \end{aligned} \quad . \quad (31)$$

There are twelve different types of combinations of ω, ω', ω_h1 and ω_h2 (see Figure 6). We denote the expectation of a combination $\left\langle \Delta_i^{(\omega)}\Delta_i^{(\omega')}a_i^{(\omega_\mathrm{h1})}a_i^{(\omega_\mathrm{h2})}\right\rangle$ by U_k and $\left\langle \Delta_i^{(\omega)}\Delta_i^{(\omega')}a_i^{(\omega_\mathrm{h1})}a_j^{(\omega_\mathrm{h2})}\right\rangle$ by W_k. These expectations, along with the prefactors, are shown in Table 4. Adding up the terms leads to this expression for U_h:

$$\begin{aligned}U_\mathrm{h} = &{\kappa_\mathrm{h}}^2(2U_1+U_2+U_3-4U_4-4U_6-2U_8+6U_{10}) \\ &+\kappa_\mathrm{h}\overline{\kappa_\mathrm{h}}\Omega_\mathrm{h}(2U_4+2U_6-4U_{10})+\kappa_\mathrm{h}\overline{\kappa_\mathrm{l}}\Omega_\mathrm{l}(2U_5+2U_7-4U_{11}) \\ &+\overline{{\kappa_\mathrm{h}}^2}\Omega_\mathrm{h}(U_8-U_{10})+\overline{{\kappa_\mathrm{l}}^2}\Omega_\mathrm{l}(U_9-U_{12}) \\ &+(\overline{\kappa_\mathrm{h}}\Omega_\mathrm{h}U_{10}+\overline{\kappa_\mathrm{l}}\Omega_\mathrm{l}U_{12})^2 \end{aligned} \quad , \quad (32)$$

and by subtracting U_{h} from T_{h} (equation (29)), we obtain

$$\begin{aligned} T_{\mathrm{h}} - U_{\mathrm{h}} = p(1-p)(1-c)^2\big\langle {\kappa_{\mathrm{h}}}^2(6p(p-1)+1)(\delta-\beta)^2 \\ + 2(1-2p)(\delta-\beta)(\kappa_{\mathrm{h}}\overline{\kappa_{\mathrm{h}}}\Omega_{\mathrm{h}}\phi + \kappa_{\mathrm{h}}\overline{\kappa_{\mathrm{l}}}\Omega_{\mathrm{l}}\psi) \\ + p(1-p)(\overline{{\kappa_{\mathrm{h}}}^2}\Omega_{\mathrm{h}}(\delta-\beta)^2 + \overline{{\kappa_{\mathrm{l}}}^2}\Omega_{\mathrm{l}}(\gamma-\alpha)^2) \\ + (\Omega_{\mathrm{h}}\overline{\kappa_{\mathrm{h}}}\phi + \Omega_{\mathrm{l}}\overline{\kappa_{\mathrm{l}}}\psi)^2\big\rangle \end{aligned} \quad . \tag{33}$$

A similar computation yields $V_{\mathrm{h}} - W_{\mathrm{h}} = 0$, so there is no contribution from the cross-unit terms.

This absence of cross-term contributions means that the dispersion of the high and low patterns depends only on T_{h}, U_{h}, T_{l} and U_{l}. The expression for the high patterns is

$$s_{\mathrm{h}}^2(\kappa_{\mathrm{h}}) = N\left\langle f^2\right\rangle\left((\langle g\rangle)^2(T_{\mathrm{h}} - U_{\mathrm{h}}) + (\left\langle g^2\right\rangle - (\langle g\rangle)^2)T_{\mathrm{h}}\right) \tag{34}$$

and there is an analogous expression for the low patterns. We define

$$R = \frac{T_{\mathrm{h}} + T_{\mathrm{l}} - U_{\mathrm{h}} - U_{\mathrm{l}}}{2(\langle\kappa\rangle)^2\Omega p(1-p)(1-c)^2} \text{ and } T^{\dagger} = \frac{T_{\mathrm{h}} + T_{\mathrm{l}}}{2(\langle\kappa\rangle)^2\Omega p(1-p)(1-c)^2} \tag{35}$$

so that we can write down the SNR as a function of intensity which is of the same form as equation (10) in section 3:

$$\rho(\kappa) = \frac{Np(1-p)(\delta-\gamma-\beta+\alpha)^2(\kappa/\langle\kappa\rangle)^2}{\Omega(1+\upsilon_f^2)\left(R + \upsilon_g^2 T^{\dagger}\right)} \quad . \tag{36}$$

A.2.3 Calculation of expectations involving κ_{h}

The terms which are linear in $\overline{\kappa_{\mathrm{h}}}$, $\overline{{\kappa_{\mathrm{h}}}^2}$ and Ω_{h} are straightforward since $\left\langle\overline{{\kappa_{\mathrm{h}}}^2}\right\rangle = \langle{\kappa_{\mathrm{h}}}^2\rangle$ and $\langle\overline{\kappa_{\mathrm{h}}}\rangle = \langle\kappa_{\mathrm{h}}\rangle$. As these are independent of Ω_{h}, the expectations $\left\langle\overline{{\kappa_{\mathrm{h}}}^2}\Omega_{\mathrm{h}}\right\rangle$ and $\langle\overline{\kappa_{\mathrm{h}}}\Omega_{\mathrm{h}}\rangle$ factorise. In order to evaluate the term $\langle(\Omega_{\mathrm{h}}\overline{\kappa_{\mathrm{h}}}\phi + \Omega_{\mathrm{l}}\overline{\kappa_{\mathrm{l}}}\psi)^2\rangle$, we compute the expectation of $\overline{\kappa_{\mathrm{h}}}^2$ conditional on Ω_{h}

$$\left\langle\overline{\kappa_{\mathrm{h}}}^2|\Omega_{\mathrm{h}}\right\rangle = \frac{1}{\Omega_{\mathrm{h}}}\sigma_{\kappa_{\mathrm{h}}}^2 + (\langle\kappa_{\mathrm{h}}\rangle)^2 \tag{37}$$

This means that

$$\left\langle{\Omega_{\mathrm{h}}}^2\overline{{\kappa_{\mathrm{h}}}^2}\right\rangle = \left\langle\left\langle{\kappa_{\mathrm{h}}}^2|\Omega_{\mathrm{h}}\right\rangle{\Omega_{\mathrm{h}}}^2\right\rangle = \langle\Omega_{\mathrm{h}}\rangle\,\sigma_{\kappa_{\mathrm{h}}}^2 + \left\langle{\Omega_{\mathrm{h}}}^2\right\rangle(\langle\kappa_{\mathrm{h}}\rangle)^2 \tag{38}$$

Hence

$$\begin{aligned} &\left\langle(\Omega_{\mathrm{h}}\overline{\kappa_{\mathrm{h}}}\phi + \Omega_{\mathrm{l}}\overline{\kappa_{\mathrm{l}}}\psi)^2\right\rangle = \\ &\quad \phi^2(\langle\Omega_{\mathrm{h}}\rangle\,\sigma_{\kappa_{\mathrm{h}}}^2 + \left\langle{\Omega_{\mathrm{h}}}^2\right\rangle(\langle\kappa_{\mathrm{h}}\rangle)^2) + \psi^2(\langle\Omega_{\mathrm{l}}\rangle\,\sigma_{\kappa_{\mathrm{l}}}^2 + \left\langle{\Omega_{\mathrm{l}}}^2\right\rangle(\langle\kappa_{\mathrm{l}}\rangle)^2) \\ &\qquad + 2\phi\psi\,\langle\Omega_{\mathrm{h}}(\Omega-\Omega_{\mathrm{h}})\overline{\kappa_{\mathrm{h}}\kappa_{\mathrm{l}}}\rangle \\ &\quad = \Omega(r\phi^2\sigma_{\kappa_{\mathrm{h}}}^2 + (1-r)\psi^2\sigma_{\kappa_{\mathrm{l}}}^2) \\ &\quad +\Omega r(1-r)(\langle\kappa_{\mathrm{h}}\rangle\,\phi - \langle\kappa_{\mathrm{l}}\rangle\,\psi)^2 + \Omega^2(r\,\langle\kappa_{\mathrm{h}}\rangle\,\phi + (1-r)\,\langle\kappa_{\mathrm{l}}\rangle\,\psi)^2 \end{aligned} \tag{39}$$

We can use equation (39) to remove the expectations over Ω_{h} and Ω_{l} from T_{h} and T_{l}. When we substitute the new expressions for T_{h} and T_{l} into the $T^{\dagger}$ (defined in equation (35)),

and ignore terms in $1/\Omega$, we obtain:

$$\begin{aligned} T^{\dagger} \approx & \frac{1+c}{1-c}(\delta-\beta+\gamma-\alpha)(r\phi+(1-r)\psi)\kappa/\langle\kappa\rangle \\ & + \frac{\pi}{(1-c)^2}\left(r(\delta-\beta)^2+(1-r)(\gamma-\alpha)^2\right)\left\langle\kappa^2\right\rangle/(\langle\kappa\rangle)^2 \\ & + \frac{\pi}{p(1-p)(1-c)^2}\left((r\phi^2+(1-r)\psi^2)\sigma_\kappa^2/(\langle\kappa\rangle)^2+r(1-r)(\phi-\psi)^2\right) \\ & + \frac{\pi}{p(1-p)(1-c)^2}(r\phi+(1-r)\psi)^2\Omega \end{aligned} \quad . \tag{40}$$

Similarly, we can remove the expectations from $T_{\mathrm{h}} - U_{\mathrm{h}}$ to give:

$$\begin{aligned} R \approx & (1-2p)(\delta-\beta+\gamma-\alpha)(r\phi+(1-r)\psi)\kappa/\langle\kappa\rangle \\ & + p(1-p)(r(\delta-\beta)^2+(1-r)(\gamma-\alpha)^2)\left\langle\kappa^2\right\rangle/(\langle\kappa\rangle)^2 \\ & + (r\phi^2+(1-r)\psi^2)\sigma_\kappa^2/(\langle\kappa\rangle)^2+r(1-r)(\phi-\psi)^2 \\ & + (r\phi+(1-r)\psi)^2\Omega \ . \end{aligned} \tag{41}$$

By rewriting the $\langle\kappa^2\rangle/(\langle\kappa\rangle)^2$ in terms of the coefficient of variation υ_κ and grouping terms we arrive at the expression for the SNR given by equation (10) and Table 2 in section 3 of the main text.

References

Abraham, W. C., Logan, B., Greenwood, J. M., and Dragunow, M. (2002). Induction and experience-dependent consolidation of stable long-term potentiation lasting months in the hippocampus. *J. Neurosci.*, 22(21):9626–9634.

Amit, D. J., Gutfreund, H., and Sompolinsky, H. (1985). Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, 55:1530–1533.

Anderson, J. A. (1972). A simple neural network generating an interactive memory. *Math. Biosci.*, 14:197–220.

Anderson, R. B. and Tweney, R. D. (1997). Artifactual power curves in forgetting. *Mem. Cognit.*, 25(5):724–730.

Andrásfalvy, B. K. and Magee, J. C. (2001). Distance-dependent increase in AMPA receptor number in the dendrites of adult hippocampal CA1 pyramidal neurons. *J. Neurosci.*, 21(23):9151–9159.

Barnes, C. A., McNaughton, B. L., Mizumori, S. J. Y., Leonard, B. W., and Lin, L.-H. (1990). Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. *Prog. Brain Res.*, 83:287–300.

Bennett, M. R., Gibson, W. G., and Robinson, J. (1994). Dynamics of the CA3 pyramidal neuron autoassociative memory network in the hippocampus. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 343(1304):167–187.

Bolshakov, V. Y., Golan, H., Kandel, E. R., and Siegelbaum, S. A. (1997). Recruitment of new sites of synaptic transmission during the cAMP-dependent late phase of LTP at CA3–CA1 synapses in the hippocampus. *Neuron*, 19:635–651.

Chechik, G., Meilijson, I., and Ruppin, E. (2001). Effective neuronal learning with ineffective Hebbian learning rules. *Neural Comput.*, 13:817–840.

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

Desai, N. S., Rutherford, L. C., and Turrigiano, G. G. (1999). Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nat. Neurosci.*, 2:515–520.

Destexhe, A., Rudolph, M., and Paré, D. (2003). The high-conductance state of neocortical neurons *in vivo*. *Nat. Rev. Neurosci.*, 4:739–751.

Forti, L., Bossi, M., Bergamaschi, A., Villa, A., and Malgaroli, A. (1997). Loose-patch recordings of single quanta at individual hippocampal synapses. *Nature*, 388:874–878.

Fusi, S., Drew, P. J., and Abbott, L. F. (2005). Cascade models of synaptically stored memories. *Neuron*, 45:599–611.

Geszti, T. and Pázmándi, F. (1987). Learning within bounds and dream sleep. *J. Phys. A Math. Gen.*, 20:L1299–L1303.

Gillessen, T. and Alzheimer, C. (1997). Amplification of EPSPs by low Ni^{2+} and amiloride-sensitive Ca^{2+} channels in apical dendrites of rat CA1 pyramidal neurons. *J. Neurophysiol.*, 77(3):1639–1643.

Gordon, M. B. (1987). Memory capacity of neural network learning within bounds. *J. Phys. (Paris)*, 48:2053–2058.

Graham, B. and Willshaw, D. (1997). Capacity and information efficiency of the associative net. *Network Comp. Neural Syst.*, 8:35–54.

Graham, B. P. (2001). Pattern recognition in a compartmental model of a CA1 pyramidal neuron. *Network Comp. Neural Syst.*, 12:473–492.

Haberly, L. B. and Bower, J. M. (1989). Olfactory cortex: model circuit for study of associative memory? *Trends Neurosci.*, 12(7):258–264.

Henson, R. and Willshaw, D. J. (1995). Short-term associative memory. In *Proceedings of the INNS World Congress on Neural Networks, 1995, Washington DC.*

Hertz, J. A., Krogh, A. S., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation.* Addison-Wesley, Reading, Massachusetts.

Hessler, N. A., Shirke, A. M., and Malinow, R. (1993). The probability of transmitter release at a mammalian central synapse. *Nature*, 366:569–572.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.*, 79:2554–2558.

Huerta, R., Nowotny, T., Garca-Sanchez, M., Abarbanel, H. D., and Rabinovich, M. I. (2004). Learning classification in the olfactory system of insects. *Neural Comput.*, 16(8):1601–1640.

Kohonen, T. (1972). Correlation matrix memories. *IEEE Trans. Comput.*, C-21:353–359.

Laurent, G. and Naraghi, M. (1994). Odorant-induced oscillations in the mushroom bodies of the locust. *J. Neurosci.*, 14:2993–3004.

Leutgeb, S., Leutgeb, J. K., Treves, A., Moser, M.-B., and Moser, E. I. (2004). Distinct ensemble codes in hippocampal areas CA3 and CA1. *Science*, 305(5688):1295–1298.

Levy, W. B. (1989). A computational approach to hippocampal function. In Hawkins, R. D. and Bower, G. H., editors, *Computational Models of Learning in Simple Neural Systems*, volume 23 of *The Psychology of Learning and Motivation*, pages 243–305. Academic Press, San Diego.

Lipowsky, R., Gillessen, T., and Alzheimer, C. (1996). Dendritic Na^+ channels amplify EPSPs in hippocampal CA1 pyramidal cells. *J. Neurophysiol.*, 76(4):2181–2191.

Lisman, J. E. (1997). Bursts as a unit of neural information: making unreliable synapses reliable. *Trends Neurosci.*, 20:38–43.

London, M. and Segev, I. (2001). Synaptic scaling *in vitro* and *in vivo*. *Nat. Neurosci.*, 4(9):853–854.

Lynch, G. S., Dunwiddie, T., and Gribkoff, V. (1977). Heterosynaptic depression: a postsynaptic correlate of long-term depression. *Nature*, 266:737–739.

Magee, J. C. and Cook, E. P. (2000). Somatic EPSP amplitude is independent of synapse location in hippocampal pyramidal neurons. *Nat. Neurosci.*, 3:895–903.

Magee, J. C. and Cook, E. P. (2001). Reply to "Synaptic scaling *in vitro* and *in vivo*". *Nat. Neurosci.*, 4(9):854–855.

Magee, J. C. and Johnston, D. (1995). Synaptic activation of voltage-activated channels in the dendrites of hippocampal pyramidal neurons. *Science*, 268:301–304.

McNaughton, B. L. and Morris, R. G. M. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends Neurosci.*, 10(10):408–415.

Megías, M., Emri, Z., Freund, T. F., and Gulyás, A. I. (2001). Total number and distribution of inhibitory and excitatory synapses on hippocampal CA1 pyramidal cells. *Neuroscience*, 102(3):527–540.

Mézard, M., Nadal, J. P., and Toulouse, G. (1986). Solvable models of working memories. *J. Phys. (Paris)*, 47:1457–1462.

Nadal, J. P., Toulouse, G., Changeux, J. P., and Dehaene, S. (1986). Networks of formal neurons and memory palimpsests. *Europhys. Lett.*, 1:535–542.

Palm, G. (1988). On the asymptotic information storage capacity of neural networks. In Eckmiller, R. and von der Malsburg, C., editors, *Neural computers*, volume F41 of *NATO ASI Series*, pages 271–280. Springer-Verlag.

Palm, G. and Sommer, F. T. (1996). Associative data storage and retrieval in neural networks. In Domany, E., van Hemmen, J. L., and Shulten, K., editors, *Models of Neural Networks III: Association, Generalization and Representation*, pages 79–118. Springer-Verlag, New York.

Parisi, G. (1986). A memory which forgets. *J. Phys. A Math. Gen.*, 19:L617–L620.

Racine, R. J., Milgram, N. W., and Hafner, S. (1983). Long-term potentiation phenomena in the rat limbic forebrain. *Brain Res.*, 260:217–231.

Rall, W. (1964). Theoretical significance of dendritic trees for neuronal input-output relations. In Reis, R. F., editor, *Neural Theory and Modeling*. Stanford University Press, Palo Alto. Reprinted in Segev et al. (1995).

Rudolph, M. and Destexhe, A. (2003). A fast-conducting, stochastic integrative mode for neocortical neurons *in vivo*. *J. Neurosci.*, 23(6):2466–2476.

Sanes, J. R. and Lichtman, J. W. (1999). Can molecules explain long-term potentiation? *Nat. Neurosci.*, 2:597–604.

Segev, I., Rinzel, J., and Shepherd, G. M., editors (1995). *The Theoretical Foundation of Dendritic Function: Selected Papers of Wilfrid Rall with Commentaries*. MIT Press, Cambridge, Massachusetts.

Sejnowski, T. J. (1977a). Statistical constraints on synaptic plasticity. *J. Theor. Biol.*, 69:385–389.

Sejnowski, T. J. (1977b). Storing covariance with nonlinearly interacting neurons. *J. Math. Biol.*, 4:303–321.

Shors, T. J., Seib, T. B., Levine, S., and Thompson, R. F. (1989). Inescapable versus escapable shock modulates long-term potentiation in the rat hippocampus. *Science*, 244:224–226.

Stevens, C. F. and Wang, Y. (1994). Changes in reliability of synaptic function as a mechanism for plasticity. *Nature*, 371:704–707.

Stricker, C., Field, A. C., and Redman, S. J. (1996). Statistical analysis of amplitude fluctuations in EPSCs evoked in rat CA1 pyramidal neurons *in vitro*. *J. Physiol.*, 490(2):419–441.

Treves, A. and Rolls, E. T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus*, 4(3):374–391.

Turrigiano, G. G., Leslie, K. R., Desai, N. S., Rutherford, L. C., and Nelson, S. B. (1998). Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature*, 391:892–896.

van Hemmen, J. L., Keller, G., and Kühn, R. (1988). Forgetful memories. *Europhys. Lett.*, 5(7):663–668.

van Hemmen, J. L. and Zagrebnov, V. A. (1987). Storing extensively many weighted patterns in a saturated neural network. *J. Phys. A Math. Gen.*, 20:3989–3999.

Willshaw, D. (1971). *Models of distributed associative memory*. PhD thesis, University of Edinburgh.

Willshaw, D. J., Buneman, O. P., and Longuet-Higgins, H. C. (1969). Non-holographic associative memory. *Nature*, 222:960–962.

Willshaw, D. J. and Dayan, P. (1990). Optimal plasticity in matrix memories: What goes up must come down. *Neural Comput.*, 2:85–93.

Wixted, J. T. and Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Mem. Cognit.*, 25(5):731–739.

Xu, L., Anwyl, R., and Rowan, M. J. (1997). Behavioural stress facilitates the induction of long-term depression in the hippocampus. *Nature*, 387:497–500.

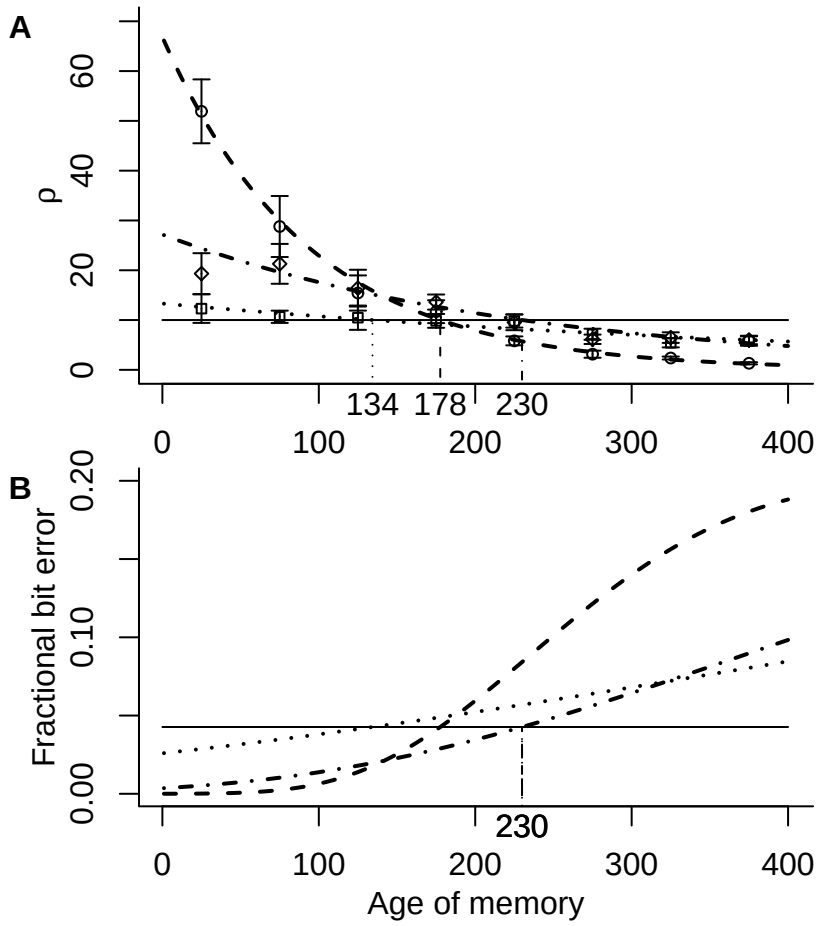

Figure 4: SNR (**A**) and bit error (**B**) as functions of the age of a memory in a network with a covariance learning rule with three different forgetting time constants: fast ($\tau = 187$; dashed line), optimal ($\tau = 460$; dash-dotted line) and slow ($\tau = 939$; dotted line). The solid horizontal lines indicate the SNR defining good retrieval $\rho_{\min}(\mathbf{A})$ or bit error thresholds (**B**). The capacity of the network is defined by the age of memory at the point where the SNR or bit error curve crosses the SNR or bit error criterion. The bit error is derived from the SNR ρ according to the formula of Dayan and Willshaw (1991): $(1-r)\Phi(-\frac{\sqrt{\rho}}{2}+\frac{1}{\sqrt{\rho}}\ln\frac{r}{1-r})+r\Phi(-\frac{\sqrt{\rho}}{2}-\frac{1}{\sqrt{\rho}}\ln\frac{r}{1-r})$ where $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{u^2}{2}}du$. The capacities for the three curves are indicated along the x-axis.

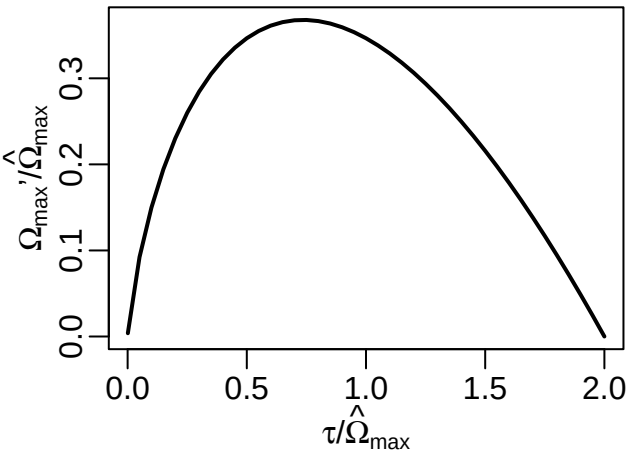

Figure 5: Capacity of a covariance rule palimpsest as a function of the forgetting time constant. The capacity of the network $\Omega_{\max}$ and the forgetting time constant τ are given as fractions of the capacity of a standard network $\hat{\Omega}_{\max}$. The maximum capacity $\hat{\Omega}_{\max}/e$ is attained when $\tau = 2\hat{\Omega}_{\max}/e$.

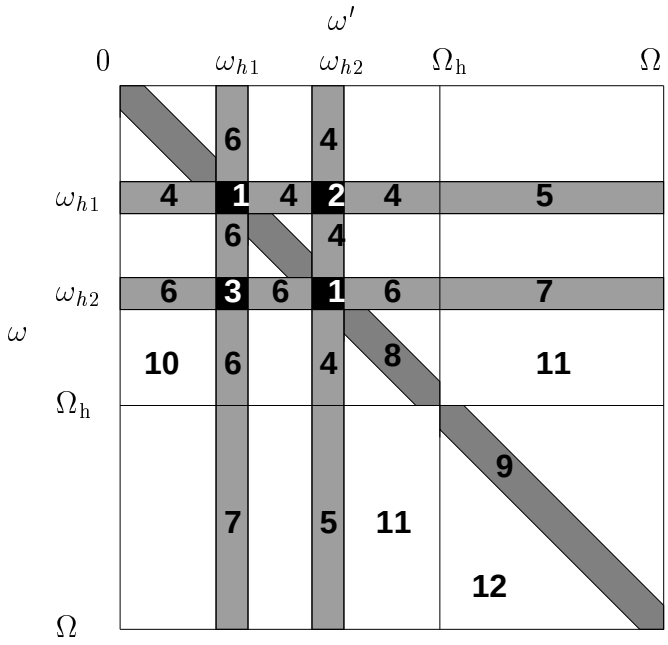

Figure 6: Diagrammatic representation of the twelve different types of combinations of ω,ω', $\omega_{\rm h1}$ and $\omega_{\rm h2}$ present in the cross-pattern terms in equation (30). In the light-shaded regions either ω or ω' is equal to $\omega_{\rm h1}$ or $\omega_{\rm h2}$. In the darker shaded region, $\omega = \omega'$. In the black regions, ω is $\omega_{\rm h1}$ or $\omega_{\rm h2}$, as is ω'. The numbers in the regions refer to the suffix k of U_k and W_k (see Table 4).