J Comput Neurosci (2011) 30:699–720
DOI 10.1007/s10827-010-0287-7

Capacity analysis in multi-state synaptic models: a retrieval probability perspective

Yibi Huang · Yali Amit

Received: 13 April 2010 / Revised: 20 September 2010 / Accepted: 5 October 2010 / Published online: 27 October 2010

Abstract We define the memory capacity of networks of binary neurons with finite-state synapses in terms of *retrieval probabilities* of learned patterns under standard asynchronous dynamics with a predetermined threshold. The threshold is set to control the proportion of non-selective neurons that fire. An optimal inhibition level is chosen to stabilize network behavior. For any local learning rule we provide a computationally efficient and highly accurate approximation to the retrieval probability of a pattern as a function of its age. The method is applied to the sequential models (Fusi and Abbott, Nat Neurosci 10:485–493, 2007) and metaplasticity models (Fusi et al., Neuron 45(4):599–611, 2005; Leibold and Kempter, Cereb Cortex 18:67–77, 2008). We show that as the number of synaptic states increases, the capacity, as defined here, either plateaus or decreases. In the few cases where multi-state models exceed the capacity of binary synapse models the improvement is small.

Action Editor: Mark van Rossum

Electronic supplementary material The online version of this article (doi:10.1007/s10827-010-0287-7) contains supplementary material, which is available to authorized users.

Supported in part by NSF ITR DMS-0706816.

Y. Huang (✉)
Department of Statistics, University of Chicago,
5734 S University Ave, Chicago, IL 60637, USA
e-mail: yibih@uchicago.edu

Y. Amit
Departments of Statistics and Computer Science,
University of Chicago, Chicago, IL, USA
e-mail: amit@marx.uchicago.edu

Keywords Hebbian learning · Network dynamics · Inhibition · Sparse coding

1 Introduction

Inspired by the *delay activity* observed in numerous *delayed match to sample* experiments (see Fuster 1995; Miyashita and Hayashi 2000; Wang 2001 for reviews), the stable states of neural network dynamics are considered to be a promising candidate mechanism underlying working memory (see Del Giudice et al. 2003 for a review on this conjecture). The simplest model for synaptic modification leading to memory retrieval can be traced to Willshaw et al. (1969) where synapses between two 'on' binary neurons are simply set to 1 and the others to 0. This network is not dynamic in that learning is performed once for all the patterns. A stochastic modification which allows for dynamic learning and gradual erasure of older patterns from the memory trace was proposed in Amit and Fusi (1994). The network is trained with an uninterrupted sequence of uncorrelated binary stimuli and the synapses are binary variables that transition between the potentiated and depressed states based on the activity of the pre and post-synaptic neurons of a presented stimulus. This model represents the simplest form of dynamic learning with purely local Hebbian modification of the synapses. Stability of the learned patterns is determined in terms of the signal-to-noise ratio (SNR) of the fields of selective/non-selective neurons for a pattern learned in the past. No explicit readout mechanism for accessing the information stored in the synapses is addressed. The learning process is analyzed using Markov chain theory showing that the memory trace of old patterns decays

exponentially fast with the number of learned patterns. For any fixed setting of the parameters (coding level and learning rates) the network capacity was shown to grow at most *logarithmically* in the number of neurons in the network. Higher capacity, up to nearly *quadratic* in the number of neurons, is achieved by adjusting the coding level and the transition rates to the size of the network.

Fusi et al. (2005) and Fusi and Abbott (2007) extended the analysis to multi-state synapses and hidden synaptic states. Again capacity is studied in terms of the SNR, which can be viewed as the ideal observer point of view. In this context two different but related quantities are studied, the storage capacity on one hand and the information content per synapse on the other. In other words how much information is stored in the synapses assuming it can somehow be fully recovered.

One possible readout mechanism involves pattern retrieval in some form of dynamics. In a recurrent network pattern retrieval means the stored pattern is a stable state of the dynamics. In a two-layer feedforward network pattern retrieval means the activity of a pattern in the cue layer is sufficient to evoke activity of another pattern in the target layer, but there is no self-sustained activity. In either case, a threshold must be chosen. Requiring the SNR above a certain level is necessary but not sufficient for retrieval in both cases. Leibold and Kempter (2008) analyzed a number of learning models in the feedforward setting in terms of retrieval probabilities with a preset threshold. In Barret and van Rossum (2008) there is an attempt to identify the optimal synaptic modification rule. The analysis is limited to a feedforward network with one binary output neuron and the goal is to distinguish between two categories, learned and unlearned patterns. Capacity is related to the error rate of this two-class problem.

Retrieval is more difficult in the recurrent setting than in the feedforward setting. The threshold must on the one hand control the number of non-selective neurons that fire and on the other hand guarantee that a large percent of the selective neurons persist in firing. In recent work, Amit and Huang (2010) and Romani et al. (2008), a mechanism to determine this threshold has been proposed, in terms of the network parameters, that enables stable retrieval of patterns in asynchronous dynamics. Stability of the recurrent network especially in the presence of varied coding levels requires the introduction of inhibition. The retrieval probability computation as introduced in Amit and Huang (2010) requires a more careful analysis of the noise, i.e. the variance of the fields, which involves covariances of the synapses. This is then used in a normal approximation to the fields (similar to Leibold and Kempter 2008 and Barret and van Rossum 2008). In addition, the tail of a certain binomial distribution is computed, arising from the approximating assumption that the fields of selective neurons can be considered independent. As shown in Amit and Huang (2010) and in simulations below, these approximations are remarkably accurate when compared to simulation and allow us to obtain precise retrieval probabilities as a function of pattern age. This yields a precise capacity prediction for any size network.

We note that once memory readout is defined as a particular outcome of neural dynamics, whether a particular outcome in a feedforward network, or stability of a pattern in a recurrent network, the notion of information storage in the synapses is not relevant. The synaptic states are simply means to enable the correct readout and are not addressed directly in any way.

In this paper, the retrieval probability framework is used to analyze multi-state synaptic models with binary neurons. In the two-state case, the eigenvalues of the Markov transition matrix can be written explicitly. Except for a few special cases, it is difficult to solve for the eigenvalues of the Markov chain in the multi-state case, not to mention expressing them in terms of the learning parameters. However, apart from explicitly expressing the powers of the transition matrix, much of the analysis in the two-state model can be extended to the multi-state case. A general method to predict the capacity of any local finite multi-state synapse models is given, including the hidden-state, or the so called meta-plasticity models. We apply this method to the sequential models in Fusi and Abbott (2007), and the cascade models and the serial-state models in Ben Dayan Rubin and Fusi (2007), Fusi et al. (2005) and Leibold and Kempter (2008).

As mentioned above, Barret and van Rossum (2008) studied the problem of optimizing capacity, in the feedforward setting, over an entire family of potentiation rules. Here we primarily focus on the more conservative model of Hebbian learning, i.e. if pre and post-synaptic neurons are on potentiation occurs with some probability. If the pre-synaptic neuron is on and the post-synaptic neuron is off depression occurs with some probability. In this framework we show the importance of optimizing the ratio of potentiation and depression probabilities. This one parameter optimization yields, in the recurrent setting, results that are qualitatively similar to those of Barret and van Rossum (2008).

We note that a detailed analysis of retrieval in recurrent networks with binary synapses and the more complex and realistic *integrate-and-fire neurons* can be found in Amit and Brunel (1997a, b) and Curti et al. (2004). For reviews see Amit and Mongillo (2003) and

Brunel (2003). The emphasis there is the behavior of the time continuous stochastic dynamical system and its stable states, mainly using mean field approximations. Not much can be found on maximal retrieval capacity for such networks.

Our main results are summarized as follows.

1. We show that standard asynchronous dynamics with inhibition and a properly determined threshold yields stable retrieval in networks of binary neurons with discrete synapses. The threshold is set to control the proportion of non-selective neurons that fire. An optimal inhibition level is chosen to stabilize network behavior.
2. For any local learning rule we provide a computationally efficient and highly accurate approximation to the retrieval probability of a pattern as a function of its age. Capacity is defined as the expected number of retrievable patterns over all ages and is equivalent to the sum of the retrieval probabilities over all ages.
3. Many prior studies have pointed out that capacity increases with the sparseness of patterns, but sparseness must be restricted (Amit and Fusi 1994; Amit and Huang 2010), especially for multi-state synapse models (Leibold and Kempter 2008). Indeed, increasing sparseness reduces the initial SNR, and below a certain level no memory trace is left. However, this can be remedied by increasing the ratio between the rate of depression and potentiation.
4. For the seven families of models we consider, as the number of synaptic states increases, the retrieval capacities either drop to zero or plateau at a certain level.
5. With a proper choice of parameters, some of the multi-state models slightly outperform the two-state synapse models, however in the low coding regimes where capacity is large the difference is negligible. Dramatic improvement in retrieval capacity beyond two-state models is not observed. In each model retrieval capacity is optimized by decreasing the coding level down to some critical level, while optimizing the ratio of potentiation and depression.

The article is organized as follows. Section 2 provides a general framework for all multi-state synapse models, including the mean, variance, distribution of the field (Section 2.2), threshold selection and inhibition (Section 2.3), and an approximate formula for retrieval probability as a function of pattern age (Section 2.4). The accuracy of the predicted probabilities is demonstrated on a large network of 80,000 neurons. The retrieval probability analysis is then applied to two sequential models in Section 3.2. We also show, numerically, the asymptotic quadratic behavior of capacity as the network size increases for one of these models. In Section 3.3 we treat cascade models and compare them to the two-state model. The probability approximations are further validated with simulation in Section 4. For simplicity, we demonstrate the method for fully connected networks with stimuli of a single coding level. With a slight modification, the method can be applied to randomly connected networks (Section 5.1), multiple-level coding stimuli (Section 5.2), and feed-forward networks (Section 5.3). In Section 6 we discuss the question of observed higher coding levels in the brain, which in Fusi and Abbott (2007) was one of the motivations for introducing multi-state models. We also recap the main conclusions of the analysis.

2 Multi-state and hidden-state synapse models

2.1 Framework

Synaptic states In a multi-state synapse model, synapses assume a finite number of *states* $\{\alpha_1, \alpha_2, \ldots, \alpha_M\}$. Let $\mathbf{w} = (w(\alpha_1), \ldots, w(\alpha_M))$ be the vector of efficacies corresponding to the M states. When more than one state corresponds to the same efficacy level, the states are called *hidden states* or *meta states* as they are not directly observable. Synapses moving between hidden states do not necessarily change efficacy but they may modify the probability of changing the efficacy. Such variation in synaptic plasticity is called *meta-plasticity* (Abraham and Bear 1996).

We represent the state of the synapse with presynaptic neuron j and post synaptic neuron i using an indicator vector

$$J_{ij} = (J_{ij1}, \ldots, J_{ijM}),$$

where $J_{ijm} = 0/1$ according to whether the synapse is at state α_m. The efficacy of the synapse W_{ij} is thus the vector product $W_{ij} = J_{ij}\mathbf{w}^T$.

Stimulus/Pattern For a fully-connected network of N neurons, the input stimulus is coded as $\xi = (\xi_1, \ldots, \xi_N)$ where $\xi_i = 0/1$ represents the firing status of neuron i when stimuli ξ is present. For a given stimulus ξ the *selective* neurons are those with $\xi_i = 1$, the others are called *non-selective*. The sequence of the training stimuli is denoted $\xi^{(1)}, \xi^{(2)}, \ldots$. In this paper, the two terms "stimulus" and "pattern" are used interchangeably.

Learning rule Learning is assumed *local*. A synapse transitions from state α_m to $\alpha_{m'}$ with probability $q_{mm'}^{kk'}$, which only depends on the firing status k', $k = 0/1$ of the pre- and post-synaptic neuron. Let Q^{11} be the $M \times M$ matrix

$$Q^{11} = (q_{mm'}^{11})_{m,m'=1,\ldots,M} \tag{1}$$

and Q^{01}, Q^{10}, Q^{00} are defined accordingly.

Network dynamics We use simple asynchronous dynamics, updating one randomly selected neuron at a time. If the current state of the network is denoted as $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_N^{(t)})$ and if neuron i is being updated we have

$$x_i^{(t+1)} = \begin{cases} 1 & \text{if } \sum_{j:j\neq i} W_{ij} x_j^{(t)} - \eta \sum_j x_j^{(t)} > \theta \\ 0 & \text{Otherwise,} \end{cases} \tag{2}$$

and $x_j^{(t+1)} = x_j^{(t)}$ for all other neurons. We assume that learning does not occur during the dynamics, i.e., $\{W_{ij}\}$ is fixed. The choices of the threshold θ and the inhibition factor η are detailed in Section 2.3. The quantity

$$h_i(\mathbf{x}; W) = \sum\nolimits_{j:j\neq i} W_{ij} x_j - \eta \sum\nolimits_j x_j$$

is called the *field* of neuron i, in which $W = (W_{ij})$ is the synaptic efficacy matrix, and $\sum_{j:j\neq i} W_{ij} x_j$ is the synaptic input to neuron i. Note an additional inhibitory input $-\eta \sum_j x_j$ is added to the conventional form of the field. The reason is detailed in Section 2.3.

Let $J_{ij}^{(p)}$ be the state of the synapse from neuron j to i after learning $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)}$ and let $W^{(p)} = (W_{ij}^{(p)}) = (J_{ij}^{(p)} \mathbf{w}^T)$. Whether the first pattern $\xi^{(1)}$ can be retrieved depends on whether it is a stable fixed point of the dynamics with $W = W^{(p)}$, i.e, that $h_i(\xi^{(1)}; W^{(p)}) > \theta$ if $\xi_i^{(1)} = 1$, and $< \theta$ if $\xi_i^{(1)} = 0$. Practically, the retrieved pattern does not have to be exactly the same as the stored pattern. If initialized with $\xi^{(1)}$, the network stabilizes at a pattern sufficiently close to $\xi^{(1)}$, then the retrieval is considered successful. Table 1 summarizes the notation used in the paper.

Table 1 Notation

f	Coding level
m	Number of synaptic states
η	Inhibition factor
θ	Threshold
$C_{\delta,f}$	$(1 - \frac{\delta f}{1-f})$-quantile of the standard normal
J_{ij}	Indicator vector for *state* of synapse from j to i
$W_{ij}^{(p)}$	Synaptic weights from j to i at step p
q_+, q_-	Potentiation and depression probability factors
τ	Potentiation to depression ratio
$\xi^{(p)}$	p'th learned pattern
$h_i^{(p)}$	The field on neuron i at step p produced by $\xi^{(1)}$
$\mu_x^{(p)}, \sigma_x^{(p)}$	Mean and SD of $W_{ij}^{(p)}$ when $\xi_i^{(1)} = x = 0/1, \xi_j^{(1)} = 1$
$\rho_x^{(p)}$	Cov. of $W_{ij}^{(p)}, W_{ik}^{(p)}$ when $\xi_i^{(1)} = x, \xi_j^{(1)} = \xi_k^{(1)} = 1$
π	Stationary distribution of synaptic state J
$M_x^{(p)}$	Mean of field at step p when $\xi_i^{(1)} = x = 0/1$
$R_x^{(p)}$	SD of field at step p when $\xi_i^{(1)} = x = 0/1$
δ	Fraction of non-selective neurons allowed to fire
ε	Allowed error rate in retrieval
$\mathsf{P}_\varepsilon^{(p)}$	Retrieval probability with error ε at step p

For all variables with p superscript, when superscript is removed we refer to asymptotic value $p = \infty$

2.2 Mean, variance, and distribution of the field

For notational simplicity, we write $h_i(\xi^{(1)}; W^{(p)})$ as $h_i^{(p)}$, i.e.,

$$h_i^{(p)} = \sum\nolimits_{j:j\neq i} W_{ij}^{(p)} \xi_j^{(1)} - \eta \sum\nolimits_j \xi_j^{(1)}.$$

Let

$$M_x^{(p)} := \mathsf{E}[h_i^{(p)} | \xi_i^{(1)} = x]$$
$$(R_x^{(p)})^2 := \mathsf{Var}(h_i^{(p)} | \xi_i^{(1)} = x)$$

be the mean and variance for the field of selective ($x = 0$) and non-selective neurons ($x = 1$).

For simplicity, assume the training stimuli are random patterns of a single coding level f, that is, neurons will be selective independently with probability f (See Section 5.2 for the multiple-level coding case). *Conditional* on the number of selective neurons $\sum_j \xi_j^{(1)} = n$ of the first learned pattern, $M_x^{(p)}$ and $(R_x^{(p)})^2$ can be evaluated through the following four steps.

1. In terms of the four Q-matrices (Eq. (1)), define

$$P_1 = fQ^{11} + (1-f)Q^{10}, \quad P_0 = fQ^{01} + (1-f)Q^{00} \tag{3}$$

and the two transition matrices

$$P = fP_1 + (1-f)P_0, \tag{4}$$
$$S = fP_1 \otimes P_1 + (1-f)P_0 \otimes P_0. \tag{5}$$

Here $\otimes$ denotes the Kronecker product (see Appendix A.1). Find the stationary distributions π and γ of the transition matrices P and S

$$\pi P = \pi, \ \gamma S = \gamma.$$

2. For $x = 0, 1$, calculate the p-step distribution

$$\pi_x^{(p)} := \mathsf{E}[J_{ij}^{(p)} | \xi_i^{(1)} = x, \xi_j^{(1)} = 1],$$
$$\gamma_x^{(p)} := \mathsf{E}[J_{ij}^{(p)} \otimes J_{ik}^{(p)} | \xi_i^{(1)} = x, \xi_j^{(1)} = \xi_k^{(1)} = 1]$$

iteratively as follows:

$$\pi_x^{(1)} = \pi Q^{x1}, \quad \gamma_x^{(1)} = \gamma(Q^{x1} \otimes Q^{x1}), \tag{6}$$

$$\pi_x^{(p)} = \pi_x^{(p-1)} P, \quad \gamma_x^{(p)} = \gamma_x^{(p-1)} S \text{ for } p = 2, 3, \ldots \tag{7}$$

3. Calculate the p-step mean, variance, and covariance of the synaptic efficacy as

$$\mu_x^{(p)} := \mathsf{E}[W_{ij}^{(p)} | \xi_i^{(1)} = x, \xi_j^{(1)} = 1] = \pi_x^{(p)} \mathbf{w}^T \tag{8}$$

$$(\sigma_x^{(p)})^2 := \mathsf{Var}(W_{ij}^{(p)} | \xi_i^{(1)} = x, \xi_j^{(1)} = 1) = \mathbf{w}\mathrm{Diag}(\pi_x^{(p)})\mathbf{w}^T - (\mu_x^{(p)})^2 \tag{9}$$

$$\rho_x^{(p)} := \mathsf{Cov}(W_{ij}^{(p)}, W_{ik}^{(p)} | \xi_i^{(1)} = x, \xi_j^{(1)} = \xi_k^{(1)} = 1) = \gamma_x^{(p)} (\mathbf{w} \otimes \mathbf{w})^T - (\mu_x^{(p)})^2. \tag{10}$$

Here $\mathrm{Diag}(\pi_x^{(p)})$ is the diagonal matrix with $\pi_x^{(p)}$ on the diagonal.

4. Finally,

$$\begin{aligned} M_x^{(p)} &= n(\mu_x^{(p)} - \eta), \\ (R_x^{(p)})^2 &= n(\sigma_x^{(p)})^2 + n(n-1)\rho_x^{(p)} \end{aligned} \tag{11}$$

See Appendix A.2 for justification.

Since $\pi_x^{(p)} \to \pi$, $\gamma_x^{(p)} \to \gamma$, the stationary mean, variance, and covariance of the synaptic efficacy are

$$\mu := \mu_x^{(\infty)} = \pi \mathbf{w}^T, \tag{12}$$

$$\sigma^2 := (\sigma_x^{(\infty)})^2 = \mathbf{w}\mathrm{Diag}(\pi)\mathbf{w}^T - \mu^2 \tag{13}$$

$$\rho := \rho_x^{(\infty)} = \gamma(\mathbf{w} \otimes \mathbf{w})^T - \mu^2 \tag{14}$$

and the stationary mean and variance of the fields are

$$\begin{aligned} M_1^{(\infty)} &= M_0^{(\infty)} = n(\mu - \eta), \\ R^2 &:= (R_x^{(p)})^2 = n\sigma^2 + n(n-1)\rho. \end{aligned}$$

Though our capacity prediction is not based on the signal-to-noise ratio (SNR), it does provides insight into the capacity of models. The signal is the difference between the mean fields of selective neurons and the stationary mean field $M_1^{(p)} - M_x^{(\infty)} = n(\mu_1^{(p)} - \mu)$. The SNR is the ratio of the squared signal to the stationary variance of the field R^2.

$$\mathrm{SNR} = \frac{n^2(\mu_1^{(p)} - \mu)^2}{n\sigma^2 + n(n-1)\rho} \approx \frac{n(\mu_1^{(p)} - \mu)^2}{\sigma^2} \tag{15}$$

The magnitude of the covariance ρ and $\rho_x^{(p)}$ can be shown to be on the order of the coding level of the stimuli (see Supplementary Material), and hence they are often ignored in the SNR analysis (e.g. Amit and Fusi 1994; Fusi and Abbott 2007; Fusi et al. 2005; Leibold and Kempter 2008). In the retrieval probability framework this leads to capacity overestimation, especially when the coding level is large.

The exact p-step distribution of the field is derived in the Supplementary Material, but involves $(M-1)$-dimensional integration, where M is the number of synaptic states. This is computationally cumbersome especially for large M. Instead we approximate the distribution of the field $h_i^{(p)}$ by a normal distribution with mean $M_x^{(p)}$ and variance $(R_x^{(p)})^2$, conditional on the selectivity x of neuron i and the number of selective neurons $\sum_j \xi_j^{(1)} = n$ of the first pattern. Simulations show this approximation does not lead to loss in accuracy in predicting the retrieval probabilities. Similar approximations were used in Leibold and Kempter (2008) but ρ and $\rho_x^{(p)}$ were not included.

2.3 Inhibition and threshold selection

Successful retrieval of a pattern in network dynamics depends on the threshold. If the threshold perfectly separates the fields of the selective and non-selective neurons, the firing pattern will clearly sustain itself. Without inhibition (i.e., $\eta = 0$), from Eq. (11) one can see that the fields of both selective and non-selective neurons grow with initial pattern size $n = \sum_j \xi_j^{(1)}$. The separating threshold must grow with n as well, or memory retrieval will fail. Clearly the firing threshold cannot be expected to depend on the size of the pattern being retrieved. Consequently, only patterns with size in a very limited range can be retrieved. Inhibition is thus introduced. We assume the inhibitory neurons reduce the field of the excitatory neurons by an amount proportional to the number of active excitatory neurons. For simplicity, the learning of excitatory-inhibitory and inhibitory-inhibitory synapses is not considered.

The primary consideration in setting the threshold is minimizing *false positives*, namely the probability of a non-selective firing. The pool of non-selective neurons is much larger and if even a small proportion fires the original pattern will be swamped. Consequently the threshold must be several standard deviations (SD) above the mean field of non-selective neurons. The asymptotic SD of the mean field conditional on the pattern size is $R = \sqrt{n\sigma^2 + n(n-1)\rho}$. Thus the threshold will be of the form

$$\theta = n(\mu - \eta) + CR = n\left[\mu - \eta + C\sqrt{\rho + \frac{\sigma^2 - \rho}{n}}\right]. \tag{16}$$

This however grows linearly with n. As in Amit and Huang (2010), by choosing

$$\eta = \mu + C\sqrt{\rho}, \tag{17}$$

the threshold

$$\theta_n = nC\left(\sqrt{\rho + \frac{\sigma^2-\rho}{n}} - \sqrt{\rho}\right) = \frac{C(\sigma^2-\rho)}{\sqrt{\rho + \frac{\sigma^2-\rho}{n}} + \sqrt{\rho}}$$

increases as $\sqrt{n}$ and is bounded above by $\frac{C(\sigma^2-\rho)}{2\sqrt{\rho}}$. Although the threshold still increases with n, it can be shown that the number of false positives is insensitive to the choice of n and hence is still under control (see Supplementary Material). Since the average pattern size is $n = Nf$, we set the threshold to be

$$\theta = \theta_{Nf} = \frac{C(\sigma^2-\rho)}{\sqrt{\rho + \frac{\sigma^2-\rho}{Nf}} + \sqrt{\rho}} \tag{18}$$

It remains to determine the constant C. Assume the fields of the non-selective neurons are normal with mean $\mu_0^{(p)} \approx \mu$ and variance $(R_0^{(p)})^2 \approx R^2$, the expected number of non-selective neurons above the threshold is approximately

$$N(1-f)\left(1-\Phi\left(\frac{\theta - n(\mu-\eta)}{R}\right)\right) = N(1-f)(1-\Phi(C))$$

where Φ is the distribution function of the standard normal distribution. Since we want to keep the expected number of non-selective neurons at a fraction δ of the number of selective neurons, and the average number of selective neurons is Nf, we require $N(1-f)(1-\Phi(C)) \le \delta Nf$. Thus C can be chosen to be the $C_{\delta,f} = (1-\frac{\delta f}{1-f})$-quantile of the standard normal. Though the actual distribution of the field is not normal, this approximation works well as long as Nf is not too small. A rule of thumb is $Nf > 30$. Below are values of $C_{\delta,f}$ for some f and δ

f	0.005	0.01	0.02	0.05	0.1
$\delta = 0.005$	4.05	3.89	3.71	3.47	3.26
$\delta = 0.01$	3.89	3.72	3.53	3.28	3.06

In Amit and Huang (2010) we used $C_{\delta,N} = (1-\delta)^{1/N}$-quantile of the standard normal, which is too conservative as it is unnecessary to keep the fields of *all* non-selective neurons below the threshold. The modified value leads to higher capacity levels.

2.4 Retrieval probability and network capacity

Before calculating the retrieval probabilities, a quantitative description of retrieval is required. Suppose ξ is a pattern with n selective neurons. When the network is initialized with ξ, we say ξ is successfully retrieved, if after the dynamics has stabilized, the number of selective neurons that are on stays larger than $(1-\varepsilon)n$.[1]

Since it is difficult to analyze the behavior of the dynamics, we approximate the retrieval by an event that depends only on properties of the field of the initial input. Given $\sum_j \xi_j^{(1)} = n$, the field induced by $(1-\varepsilon)n$ of the selective neurons needs to keep the field of at least $(1-\varepsilon)n$ selective neurons above threshold. If this is true, then the number of active selective neurons is likely to stay equal or above $(1-\varepsilon)n$ throughout the dynamics.

The field induced by $n' = (1-\varepsilon)n$ selective neurons has mean $n'(\mu_1^{(p)} - \eta)$ and variance $n'(\sigma_1^{(p)})^2 + n'(n'-1)\rho_1^{(p)}$. By the normal approximation, the probability that this field is above θ is,

$$\Psi_{n,\varepsilon}^{(p)} = \Phi\left(\frac{\theta - n'(\mu_1^{(p)} - \eta)}{\sqrt{n'(\sigma_1^{(p)})^2 + n'(n'-1)\rho_1^{(p)}}}\right) \tag{19}$$

This approximation works well when $n' = n(1-\varepsilon)$ is sufficiently large.

The probability that the fields of at least $(1-\varepsilon)n$ of the selective neurons are above the threshold is approximated as

$$\mathsf{P}_{n,\varepsilon}^{(p)} \approx \sum_{k \ge n(1-\varepsilon)} \binom{n}{k} (\Psi_{n,\varepsilon}^{(p)})^k (1-\Psi_{n,\varepsilon}^{(p)})^{n-k}. \tag{20}$$

Here we have also assumed independence of the fields of different neurons. Strictly speaking these fields are not independent but their dependence is weak when f is small. As the patterns are of coding level f, the probability that $\xi^{(1)}$ is of size n is $\binom{N}{n} f^n (1-f)^{N-n}$. The probability to retrieve $\xi^{(1)}$ after learning $\xi^{(1)}, \ldots,$ and $\xi^{(p)}$ is

$$\mathsf{P}_{\varepsilon}^{(p)} = \sum_{n=0}^{N} \mathsf{P}_{n,\varepsilon}^{(p)} \binom{N}{n} f^n (1-f)^{N-n} \tag{21}$$

If we use the *expected number of retrievable patterns* as the quantitative measure of *network capacity*, the capacity is simply the sum of the retrieval probabilities over all ages p

$$\text{Capacity} = \sum_{p=1}^{\infty} \mathsf{P}_{\varepsilon}^{(p)} \tag{22}$$

[1] The fraction of non-selective neurons above the threshold is not specified in the criterion since it has been controlled in the threshold selection. Moreover, because of the strong inhibition, there cannot be many non-selective neurons above threshold throughout the dynamics.

Note that if the synaptic efficacies of all states are transformed as $\mathbf{w} \to a\mathbf{w} + b$ the probabilities $\mathsf{P}_{\varepsilon}^{(p)}$ and the capacity remains the same. As the synaptic efficacy changes scale, the inhibition factor η and firing threshold θ change accordingly, and the retrieval probability is unaffected. This is not surprising as the capacity of a model should not change with the units used to measure the synaptic efficacy.

In Section 4 we present retrieval probability predictions for a number of different synaptic modification models and illustrate the accuracy of these predictions with respect to simulations.

As an initial example, in Fig. 1 we show the accuracy of the retrieval probabilities for a network of 80,000 neurons with binary synapses, where the parameters have been set in such a way that retrieval capacity is around 94,000. In red are the predicted probabilities, in black is the optimal monotone regression on probabilities estimated from ten simulations of this network.

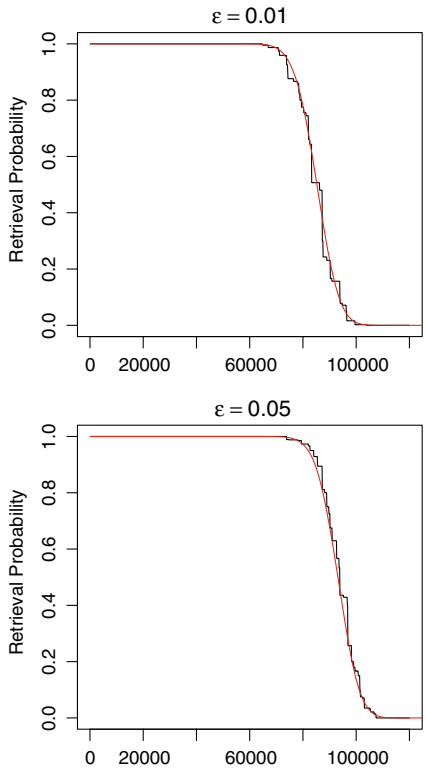

Fig. 1 Predicted retrieval probability for a large network $N =$ 80,000 as a function of pattern age based on Eq. (21) (*red*) vs. retrieval probabilities estimated from ten simulated networks. *x*-axis: age of patterns (p), *y*-axis: retrieval probability $\mathsf{P}_{\varepsilon}^{(p)}$. *Top*: error $\varepsilon = .01$, *Bottom*: error $\varepsilon = .05$. *Black* shows best monotone fit to simulation probabilities. In each run, 120,000 patters are learned and retrieval is assessed using asynchronous dynamics. $f = .002$, $\tau = 1.141$, $\eta = .514$, $\theta = .00024$. Estimated and predicted capacity is 94,000

Note the very close agreement between prediction and simulation for two values of ε—the allowable retrieval error. In the binary synapse case, using bitwise coding of synaptic and neural states it is possible to simulate such a large network rather efficiently. Learning 120,000 patterns and testing their retrieval takes a few hours on a 12 core PC.

2.5 Comparison of the SNR analysis and the retrieval probability approach

The behavior of multi-state synaptic models is much richer than the two-state models, but they are often difficult to analyze algebraically. The retrieval probability approach described above provides a computationally efficient method to numerically predict the capacity with high precision. Given the parameters, the capacity can be computed in seconds. This is particularly helpful for networks that are too large to simulate in computers.

Retrieval probability and SNR analyses do not always agree, however, large SNR is a necessary condition for retrieval. From Eqs. (15) and (16), observe that SNR $< C_{\delta,f}^2$ implies

$$n(\mu^{(p)} - \eta) < \theta_n \leq \theta_{Nf} = \theta$$

if the pattern size $n \leq Nf$ (note ρ is always non-negative). Hence from Eq. (19), $\Psi_{n,\varepsilon}^{(p)} < 0.5$. Note that $\mathsf{P}_{n,\varepsilon}^{(p)} \approx 0$ as long as $\Psi_{n,\varepsilon}^{(p)} < 0.9$. Since the coding level is fixed, the pattern size will not be far from Nf. For n slightly larger than Nf, $\Psi_{n,\varepsilon}^{(p)}$ is still less than 0.9 and hence $\mathsf{P}_{n,\varepsilon}^{(p)} \approx 0$. Thus small SNR ($< C_{\delta,f}^2$) will imply little or no retrieval capacity.

Although the p-step SNR is often intractable, the algebraic form of the initial SNR is sometimes available. From the argument above, initial SNR $< C_{\delta,f}^2$ implies little or no capacity. In some examples in Section 3, we will use the initial SNR to motivate some parameter choices and explain some results obtained via the retrieval probability approach.

3 Examples

In this section we analyze in detail a number of synaptic modification models. All capacity results reported are given in terms of the expected number of retrievable patterns (Eq. (22)) with $\delta = 0.01$, $\varepsilon = 0.05$. Also the value of τ is optimized with respect to this measure of capacity. In some of the models a signal-to-noise analysis helps motivate some of the parameter choices.

3.1 Two-state synapses

Two-state synapse models have been analyzed in detail in Amit and Fusi (1994), Amit and Huang (2010) and Romani et al. (2008). We briefly summarize the results to compare with other models. In this model, a synapse is either potentiated or depressed, with efficacies W_- or W_+, i.e. $\mathbf{w} = (W_-, W_+)$. As the memory capacity is scale and shift invariant, we simply assume $W_- = 0$, $W_+ = 1$. A depressed synapse has probability q_+ to be potentiated when both the pre- and postsynaptic neurons are active, and a potentiated synapse has probability q_- to be depressed when the presynaptic neuron is active and the postsynaptic neuron is silent. Otherwise the efficacy is unchanged. The four Q-matrices are thus

$$Q^{11} = \begin{bmatrix} 1-q_+ & q_+ \\ 0 & 1 \end{bmatrix}, \quad Q^{01} = \begin{bmatrix} 1 & 0 \\ q_- & 1-q_- \end{bmatrix},$$

$$Q^{10} = Q^{00} = I_2,$$

where I_n is the $n \times n$ identity matrix. Suppose $q_- = \tau f q_+/(1-f)$. Then $\pi = (\pi_0, \pi_1) = (\frac{\tau}{1+\tau}, \frac{1}{1+\tau})$, $\mu = \pi_1$, $\sigma^2 = \pi_0\pi_1$,

$$\mu_1^{(p)} = \pi_1 + \lambda^{p-1}\pi_0 q_+, \quad \mu_0^{(p)} = \pi_1 - \lambda^{p-1}\pi_1 q_-$$

where $\lambda = 1 - f^2 q_+ - f(1-f)q_- = 1-(1+\tau)f^2 q_+$. In this model, the algebraic form of SNR is available,

$$\text{SNR} \approx \frac{n(\mu_1^{(p)} - \mu)^2}{\sigma^2} = \frac{n\lambda^{2p-2}\pi_0^2 q_+^2}{\pi_1\pi_0} = n\tau q_+^2 \lambda^{2p-2}.$$

As the stimuli coding level is f, the average pattern size is Nf, and hence $\text{SNR} \approx Nf\lambda^{2p-2}\tau q_+^2$. Requiring $\text{SNR} \geq C$, for some constant C, gives a rough relationship between the capacity and model parameters,

$$p < \frac{\ln(Nf\tau q_+^2/C)}{-2\ln\lambda} \approx \frac{\ln(Nf\tau q_+^2/C)}{2f^2 q_+(1+\tau)}, \tag{23}$$

since $-\ln\lambda = -\ln(1-f^2 q_+(1+\tau)) \approx f^2 q_+(1+\tau)$ when f is small.

A few facts can be observed from Eq. (23). First, for fixed q_+ and τ, the capacity is $O(f^{-2})$. Second, the argument of the logarithm in Eq. (23) must be greater than 1, i.e., the SNR at the first step given by $Nf\tau q_+^2$ must be greater than C, otherwise the capacity is 0. Usually this is viewed as a constraint on the sparseness of the stimuli, especially for slow learning (q_+ small). However, this constraint can be removed by properly adjusting τ to keep the initial SNR above C. The problem is that it is unclear what the constant C should be. With retrieval capacity (Eq. (22)) all constants are determined by retrieval criteria and it is possible to optimize τ for fixed values of f and q_+. For fixed τ, the capacity initially increases as f decreases, peaks at a certain f and then drops to 0 (Fig. 2(a)). When τ is optimized sparser stimuli can be used, and the capacity can be further increased. In Fig. 2(b) τ is optimized for retrieval capacity, which is considerably improved.

Another limit on sparseness that arises when using the retrieval probability analysis is the normal approximation (Eq. (19)). This works well only when $n \approx Nf$ is not too small. Thus we restrict $Nf \geq 30$. Moreover, τ cannot be raised arbitrarily since for two-state synapses the distribution of the mean field is a mixture of Binomial distributions (Amit and Huang 2010). For the normal approximation to work, $n\pi_1 \approx Nf/(1+\tau)$ cannot

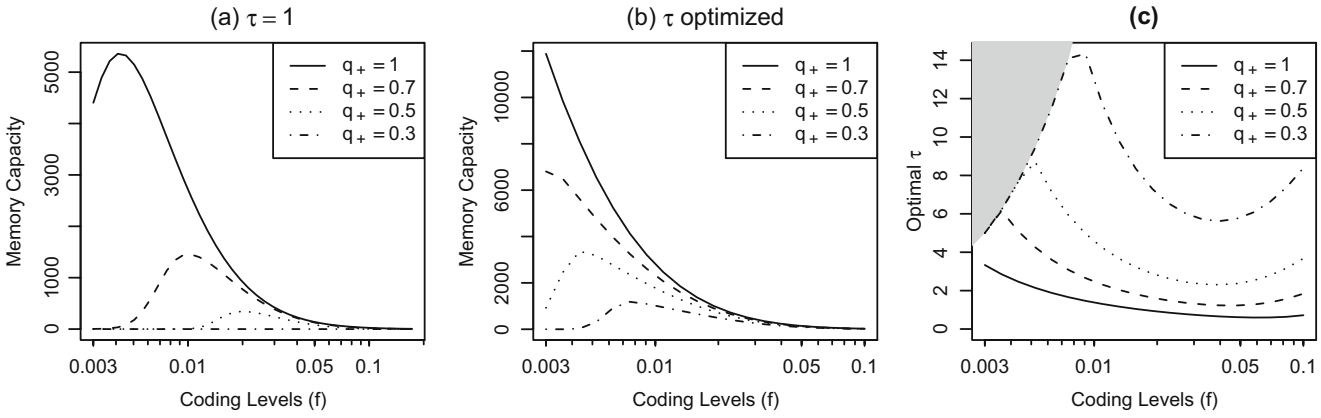

Fig. 2 Capacity of the two-state synapse model for $N = 10{,}000$. (**a**) When τ is fixed at 1, the capacity initially increases as f decreases, peaks at a certain f and then drops to 0. The peak capacity decreases rapidly with q_+, and the optimal f increases as q_+ decreases. For $q_+ \leq 0.3$, the capacity is 0 for all f. (**b**) As τ is optimized for each f and q_+, the capacity for each f and the peak capacity are increased considerably. Even for $q_+ = 0.3$, the capacity can reach 1,000. (**c**) The corresponding optimal τ of (**b**). Note we restrict τ outside the region $n\pi_1 = Nf/(1+\tau) < 5$ (*gray area*)

be too small. We restrict τ in the region $Nf/(1+\tau) \geq 5$ (Fig. 2(c)).

Intuitively smaller τ implies lower probability of depression (q_-), or slow forgetting of learned patterns, and hence the capacity increases. However, decreasing τ will raise the field of background neurons (recall $\pi_1 = \frac{1}{1+\tau}$), reduce the contrast between the selective and background neurons, and hence hinder memory retrieval. Properly obliterating old patterns and keeping the portion of potentiated synapses at a reasonable level is necessary for the network to distinguishing the selective neurons from the non-selective neurons.

3.2 Sequential models

In a sequential model, the synaptic efficacies—w—take $m+1$ discrete values $0, \frac{1}{m}, \frac{2}{m}, \ldots, \frac{m}{m}$, equispaced between 0 and 1. Whenever a synapse with efficacy w is potentiated or depressed, its synaptic efficacy increases or decreases by $1/m$, with probability $q_+\kappa_+(w)$, $q_-\kappa_-(w)$ respectively. Here q_+ and q_- are two scalar parameters between 0 and 1. We assume a synapse is potentiated only when both pre- and postsynaptic neurons are active, and depressed only when the presynaptic neuron is active and postsynaptic is silent. In the notation in Section 2.2, the state space is $\{0, 1, \ldots, m\}$ with efficacy vector $\mathbf{w} = (0, \frac{1}{m}, \frac{2}{m}, \ldots, \frac{m}{m})$. The Q-matrices are $Q^{00} = Q^{10} = I_{m+1}$, $Q^{11} = I_{m+1} + q_+ D_+$, and $Q^{01} = I_{m+1} + q_- D_-$, in which, I_k denotes a $k \times k$ identity matrix, and D_+, D_- are bidiagonal matrices given by

$$D_+ = \begin{bmatrix} -\kappa_+(0) & \kappa_+(0) & & & \\ & -\kappa_+\left(\frac{1}{m}\right) & \kappa_+\left(\frac{1}{m}\right) & & \\ & & \ddots & \ddots & \\ & & & -\kappa_+\left(\frac{m-1}{m}\right) & \kappa_+\left(\frac{m-1}{m}\right) \\ & & & & 0 \end{bmatrix},$$

$$D_- = \begin{bmatrix} 0 & & & & \\ \kappa_-\left(\frac{1}{m}\right) & -\kappa_-\left(\frac{1}{m}\right) & & & \\ & \ddots & \ddots & & \\ & & \kappa_-\left(\frac{m-1}{m}\right) & -\kappa_-\left(\frac{m-1}{m}\right) & \\ & & & \kappa_-(1) & -\kappa_-(1) \end{bmatrix}.$$

We now explore how different plasticity modifications $\kappa_+(w)$ and $\kappa_-(w)$ affect the memory performance.

3.2.1 Hard-bound models

In a hard-bound model, $\kappa_+(w)$ and $\kappa_-(w)$ are independent of the efficacy w,

$$\kappa_+(w) = \kappa_-(w) = 1$$

for $0 < w < 1$, and at the boundary $w = 0$ and 1, any transitions that would move the synaptic efficacy outside the range $[0, 1]$ are truncated, i.e. $\kappa_-(0) = 0$, $\kappa_+(1) = 0$. This model is equivalent to the one-dimensional random walk on integers with two barriers at 0 and m. Suppose $q_- = \tau f q_+/(1-f)$. The case $\tau = 1$ corresponds to the symmetric random walk, since the marginal probability of potentiation $f^2 q_+$ and depression $f(1-f)q_-$ are equal. The stationary distribution of the symmetric random walk is uniform. For $\tau \neq 1$, the stationary distribution is truncated geometric. The mean, variance, and initial SNR can be easily derived and are listed in Table 2 in the Appendix.

For $\tau = 1$, the initial SNR $\approx \frac{12Nfq_+^2}{(m+1)^2}$ decreases quadratically with m. As m increases, the SNR soon falls below C. Hence the capacity is zero unless m is small. These observations are confirmed in terms of retrieval capacity in Fig. 3(a). For $\tau < 1$, the initial SNR is also too small to maintain any memory.

From Fig. 3(b), for $\tau > 1$, one can see that the capacity is nearly constant in m as m gets large. Actually as m gets large, the model approaches the asymmetric simple random walk on non-negative integers 0, 1, 2, ..., with a single barrier at 0. The stationary state of the model is the geometric distribution

$$\pi = (1-\tau^{-1})(1, \tau^{-1}, \tau^{-2}, \ldots).$$

Most synapses stay at the lowest few levels. Higher levels are rarely visited and have little effect on learning. Further increasing the number of synaptic levels does not change the model very much.

Like the two-state model, we find the optimal τ that maximize the capacity for each m and f given. Figure 3(c and d) shows the result for $N = 10{,}000$, $m = 1, 2, 3, 10$, and $.003 \leq f \leq .1$ in which Fig. 3(c) is the optimal capacity and Fig. 3(d) is the corresponding optimal τ. The case $m = 1$ reduces to the two-state model. For low coding levels, the 3-level model ($m = 2$) is the best, at approximately 3-25% higher capacity than the optimal two-state model. More synaptic levels do not further increase capacity. For larger coding levels, the optimal hard-bounds model is no better than the optimal two-state model.

We can also use the capacity formula to numerically analyze the asymptotic capacity as N increases. In Fig. 3(e) we show the square root of the predicted

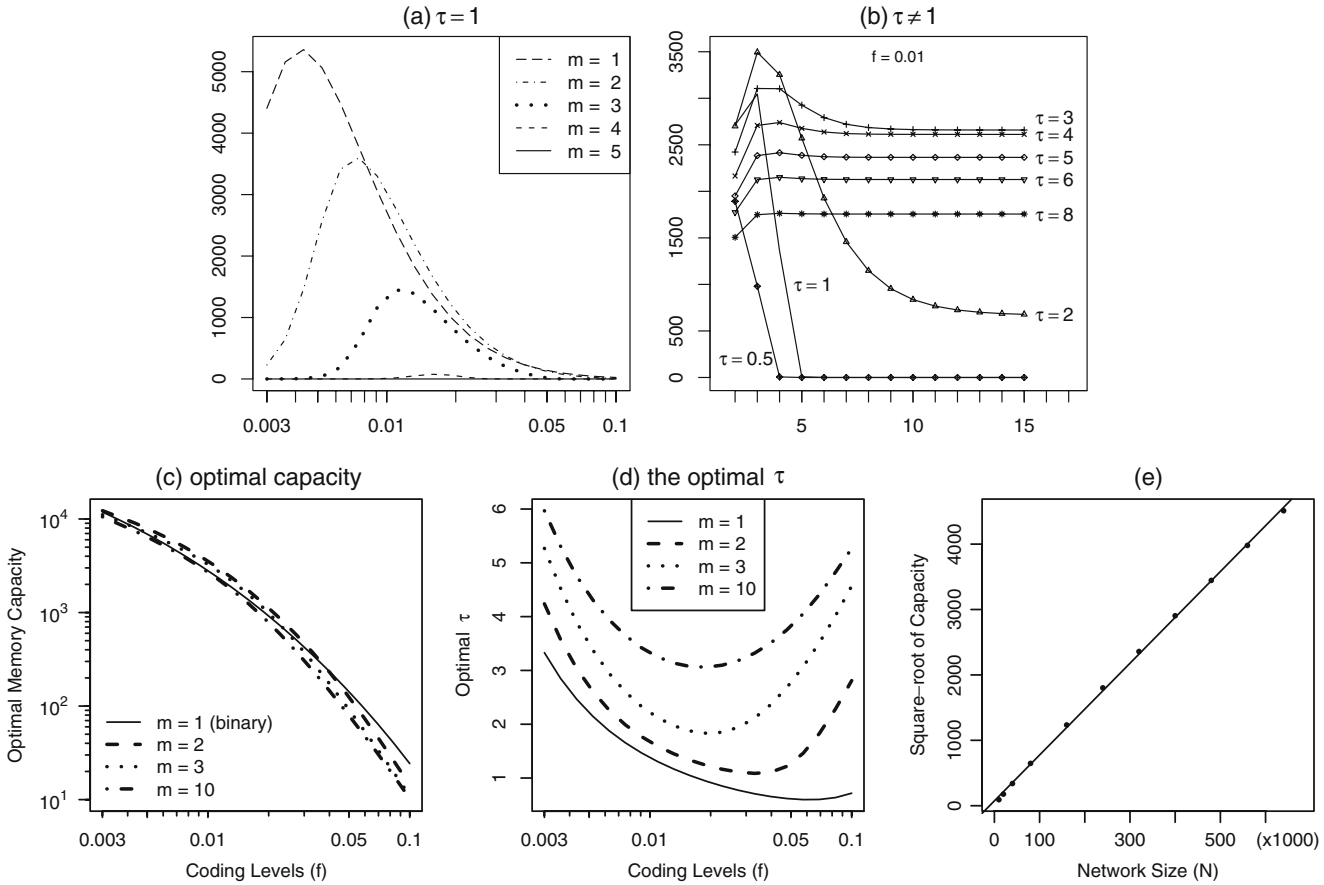

Fig. 3 Hard-bound model, $N = 10{,}000$: (**a**) For $\tau = 1$, the capacity decreases rapidly with m and is nonzero only when m is small. (**b**) When $\tau < 1$, the capacity quickly drops to 0 as m increases. When $\tau > 1$, the capacity is nearly constant in m when m is large. (**c**) The optimal capacity of the hard-bound model for $N = 10{,}000$, $m = 1$, 2, 3, 10, and $.003 \leq f \leq .1$, optimized over τ. (**d**) The corresponding optimal τ. (**e**) Square root of capacity as a function of N for the hard bound model. $m = 5$, $f = 4N/\log N$ and $\tau = 5$, for $N = 10, 20, 40, 80, \ldots, 640 \times 10^3$. The slope is .008

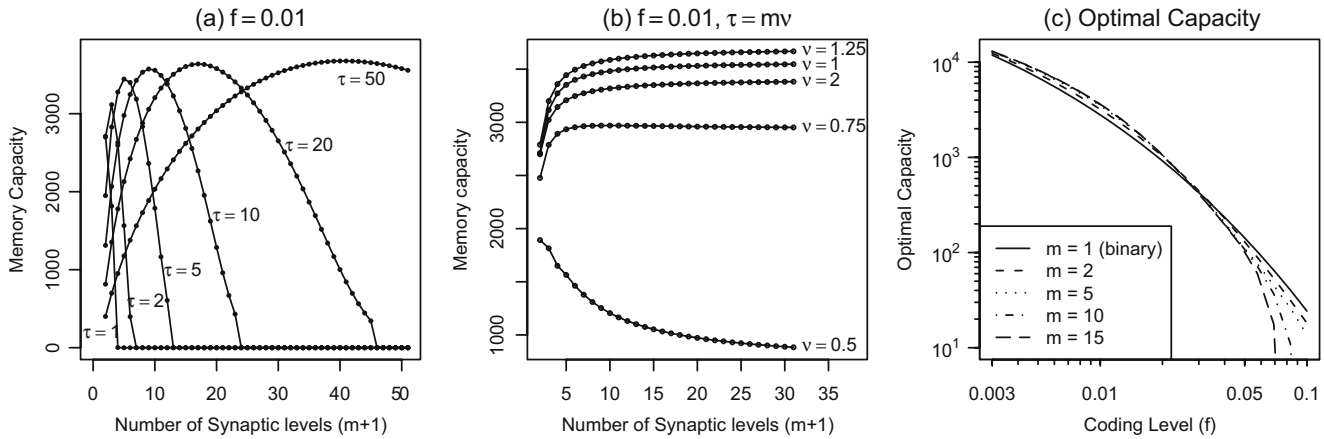

Fig. 4 (**a**) The capacity of the soft-bound model as a function of m, for a number of values of τ with $N = 10{,}000$ and $f = 0.01$. (**b**) The capacity of a soft-bound model for $m = 1$ to 30 and $\tau = \nu m$ is kept proportional to m, when $N = 10{,}000$, $f = 0.01$. The capacity hardly changes with m after $m > 10$, except for $\nu = 0.5$. (**c**) The optimal capacity of the soft-bound model for $N = 10{,}000$, $m = 1, 2, 5, 10, 15$, and $.003 \leq f \leq .1$, optimized over $\nu = m\tau$

capacity as a function of N for the hard-bound model with $m = 5$, where $f = 4N/\log N$ and $\tau = 5$. We see the nearly perfect linear fit with a slope of .008.

3.2.2 Soft-bound model

Instead of truncating the efficacy at the boundaries, a soft-bound model allows $\kappa_+(w)$ and $\kappa_-(w)$ to vary with w, gradually vanishing at the boundaries,

$$\kappa_+(w) = 1 - w,\ \kappa_-(w) = w.$$

Assuming $q_- = \tau f q_+/(1-f)$, the stationary distribution is Binomial$\left(m, \frac{1}{1+\tau}\right)$. The mean, variance, and initial SNR are listed in Table 2 in the Appendix. Since the initial SNR $= \frac{n\tau q_+^2}{m}$ *decreases* with m, for fixed τ the memory capacity must decrease to 0 for large m. Figure 4(a) confirms this observation. Moreover, this plot also shows that for fixed τ, the model capacity first increases and then decrease with m, and the optimal m seems to be proportional to τ.

Indeed, for fixed f, the model capacity is mostly affected by the ratio $\nu = \tau/m$, at least when m is large. In Fig. 4(b), the capacity plateaus as m increases while keeping the ratio ν constant. This is not surprising since the binomial distribution is well-approximated by the Poisson distribution with mean $1/\nu$ when $\tau = \nu m$ for m large,

$$\pi_i \approx \frac{e^{1/\nu}}{\nu^i\, i!}.$$

Moreover, the probability of depression from state i to $i-1$ depends on ν only

$$f(1-f)q_- \times (i/m) = i f^2 q_+ \nu.$$

When m is large relative to i, the probability of potentiation from state i to $i+1$ is scarcely affected by m,

$$f^2 q_+ \times (1 - i/m) \approx f^2 q_+.$$

For $\nu > 1$, most of the mass of Poisson$(1/\nu)$ concentrates on the lowest few synaptic levels. Higher levels are rarely visited and hence not influential.

Figure 4(c) shows the optimal capacity for each coding level f and $m = 1, 2, 5, 10$, and 15 where ν is chosen to maximize the capacity. Note the optimal ν might be slightly different for each m. The case $m = 1$ reduces to the two-state model. For small f, the capacity of the optimal soft-bound model with ten levels is 10–25% better than the optimal two-state model; for larger f, the optimal soft-bound are no better than the optimal two-state model.

In this model, the lowest level is most efficient in learning. Since $\kappa_+(0) = 1$, all synapses in this level will be potentiated once the conditions for potentiation are met. After the lowest level, the second lowest level is the level where depression is least likely to happen, and hence the trace of learned patterns could be preserved longer. In an optimal soft-bound model, most synapses stay at these two levels for effective learning and slow forgetting. This explains the resemblance of the optimal soft-bound models and the two-state models.

3.3 Hidden-state models

In a hidden state model, there may be multiple states corresponding to each level of efficacy. We consider the case of two levels of efficacy W_- and W_+ only. Recall the capacity is scale invariant (Section 2.4). We can assume $W_- = 0$, $W_+ = 1$. Figure 5 depicts the

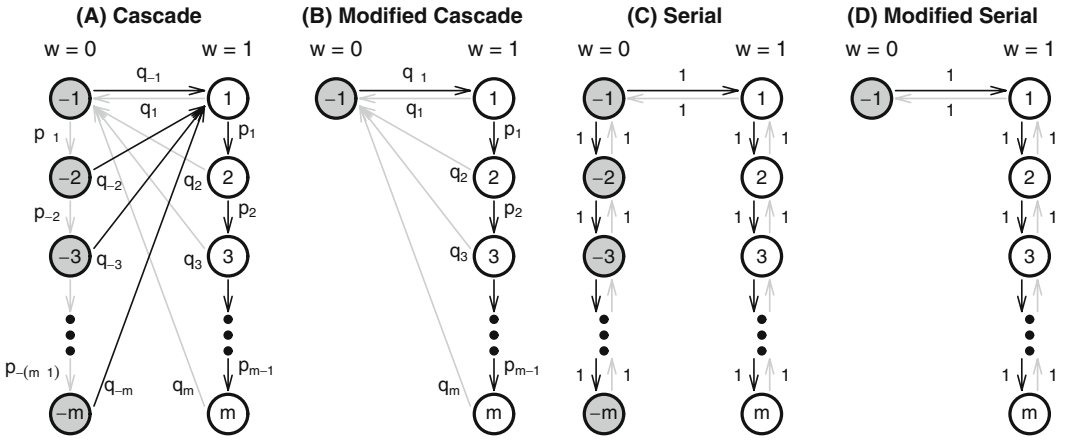

Fig. 5 Structure of the four hidden-state models. In all four models there are only two efficacy levels, $w = 0$ (depressed, *gray circles*) and $w = 1$ (potentiated, *white circles*), each with a number of hidden states. The *black/gray arrows* indicate the allowable transitions for LTP/LTD, and the numbers next to arrows indicate transition probabilities

structure of four hidden-state models we consider. The black/gray arrows indicate the allowable transitions when LTP/LTD happens, and the numbers next to the arrows indicate the transition probabilities.

3.3.1 Cascade model

The cascade model is proposed in Ben Dayan Rubin and Fusi (2007). There are m hidden states for both levels of efficacy, denoted as $-1, -2, \ldots, -m$ and $1, 2, \ldots, m$. If a synapse in the high/low level of efficacy is further potentiated/depressed, it does not change efficacy but becomes more resistant to efficacy change. The transition probabilities $q_1 > q_2 > \ldots > q_m$, $q_{-1} > q_{-2} > \ldots > q_{-m}$ are usually an exponentially decreasing sequence, reflecting biochemical processes operating on multiple timescales.

In our notation, the efficacy vector is $\mathbf{w} = (0, 0, \ldots, 0, 1, 1, \ldots, 1)$. The transition matrices for potentiation and depression can respectively be written as $I_{2m} + D_+$ and $I_{2m} + D_-$ where

$$D_+ = \begin{pmatrix} B_m^- & C_m^- \\ 0 & A_m^+ \end{pmatrix}, \quad D_- = \begin{pmatrix} A_m^- & 0 \\ C_m^+ & B_m^+ \end{pmatrix}$$

Here $A_m^\pm$ is the bi-diagonal matrix

$$A_m^\pm = \begin{pmatrix} -p_{\pm 1} & p_{\pm 1} & & & \\ & -p_{\pm 2} & p_{\pm 2} & & \\ & & \ddots & \ddots & \\ & & & -p_{\pm(m-1)} & p_{\pm(m-1)} \\ & & & & 0 \end{pmatrix}$$

and $C_m^\pm$ is the matrix with $(q_{\pm 1}, \ldots, q_{\pm m})^T$ in the first column and 0 elsewhere. $B_m^\pm$ is the diagonal matrix with $(-q_{\pm 1}, \ldots, -q_{\pm m})$ in the diagonal.

Correlations among synapses, though negligible for sparsely coded patterns, could considerably inflate the noise level in retrieving densely coded patterns. A special learning rule to remove synaptic correlation is proposed in Ben Dayan Rubin and Fusi (2007): a synapse is potentiated when both the pre- and postsynaptic neurons are active or silent, with probability q_+ and $q_0 = \frac{f^2 q_+}{(1-f)^2}$ respectively; otherwise, it is depressed with probability $q_- = \frac{\tau f q_+}{1-f}$. In our notation, $Q^{11} = I_{2m} + q_+ D_+$, $Q^{01} = Q^{10} = I_{2m} + q_- D_-$, $Q^{00} = I_{2m} + q_0 D_+$. Using this learning rule, the stationary synaptic covariance ρ defined in Eq. (14) will be 0 (see Appendix A.3 for a proof). Note that the non-stationary covariance $\rho^{(p)}$ is still nonzero, but the size is reduced considerably.

Figure 6 gives the capacity of the cascade model for $q_{\pm i} = p_{\pm i} = 2^{-i}$, for $i = 1, \ldots, m-1$, and $q_{\pm m} = 2^{-m+1}$ as in Ben Dayan Rubin and Fusi (2007). The capacity decreases rapidly as the number of hidden levels increases, and complex models have non-zero capacity only when the coding level is large (Fig. 6(a)). This phenomenon is still present even if the model is optimized over τ (Fig. 6(b)). As pointed out in Leibold and Kempter (2008), this is because the initial SNR of the cascade model decreases with the complexity of the synapses. However, as the low initial signal

$$\mu^{(1)} - \mu = \sum_{i=1}^{m} \pi_{-i} q_{-i}$$

is only due to the low potentiation probability q_{-i} of some depressed levels, this problem can be resolved by reducing the number of depressed states. The structure of the modified cascade model is depicted in Fig. 5(b), there is only one depressed state but m potentiated states. This modification preserves the multiple timescale characteristics of the cascade model. In this

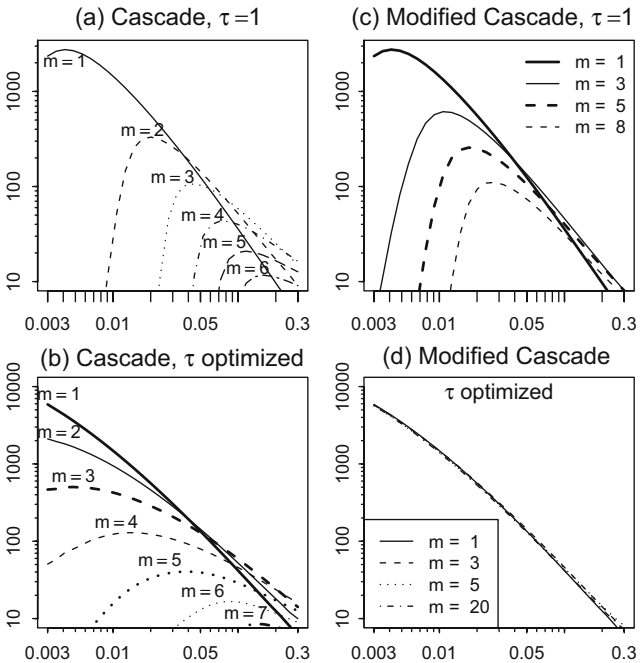

Fig. 6 Memory capacity of the cascade and modified cascade model for $N = 10{,}000$, $q_{\pm i} = p_{\pm i} = 2^{-i}$, for $i = 1, \ldots, m-1$, and $q_{\pm m} = 2^{-m+1}$. x-axis: coding level f; y-axis: memory capacity. (**a**) The original cascade model: when $\tau = 1$, the capacity drops rapidly as the number of hidden states m increases. (**b**) Even when the model is optimized over τ, the capacity still decreases as m increases. (**c**) When the number of depressed state is reduced to 1, the capacity is greatly improved, but still decreases with the number of hidden states. (**d**) When optimized over τ, the optimal modified cascade model is as good as the optimal two-state model

case, $\mathbf{w} = (0, 1, 1, \ldots, 1)$, $Q^{11} = I_{m+1} + q_+ D_+$, $Q^{01} = Q^{10} = I_{m+1} + q_- D_-$, $Q^{00} = I_{m+1} + q_0 D_+$, where

$$D_+ = \begin{pmatrix} -1 & \mathbf{e}_1 \\ \mathbf{0}^T & A_m^+ \end{pmatrix}, \quad D_- = \begin{pmatrix} 0 & \mathbf{0} \\ c_m & B_m^+ \end{pmatrix}.$$

Here $\mathbf{e}_1 = (1, 0, \ldots, 0)$, $c_m = (q_1, \ldots, q_m)^T$, and $\mathbf{0} = (0, \ldots, 0)$. The capacity is greatly improved, but still decreases with the number of hidden levels (Fig. 6(c)). This is because when $\tau = 1$, the stationary distribution is uniform over all states (Table 2), $\mu = \frac{m}{m+1}$, which is high when m is large. As in the two-state model, when the fraction of potentiated synapse is high, the contrast between the selected and the background neurons must be low and hence retrieval is difficult. This problem is resolved when τ is optimized (Fig. 6(d)). However, synapses of all complexity are as good as the two-state synapses.

3.3.2 Serial-state model

Another hidden-state model studied in Leibold and Kempter (2008) has a simpler structure where all synaptic states are connected serially and all transition probabilities equal one (Fig. 5(c)). As in the cascade we also study a modified serial-state model with only one depressed hidden state (Fig. 5(d)). Using the decorrelating learning rule, the Q matrices are of the same form as in the cascade model, but the D_+ and D_- are replaced by the D_+ and D_- of the hard-bound model. The summary of the two models are given in Table 2 in the Appendix.

For small f, the serial-state model also suffers from the problem of small initial SNR as m gets large (Table 2). The retrieval capacity soon drops to zero as m increases, even when optimized over τ (Fig. 7(ab)). On the other hand, serial-state model does improve

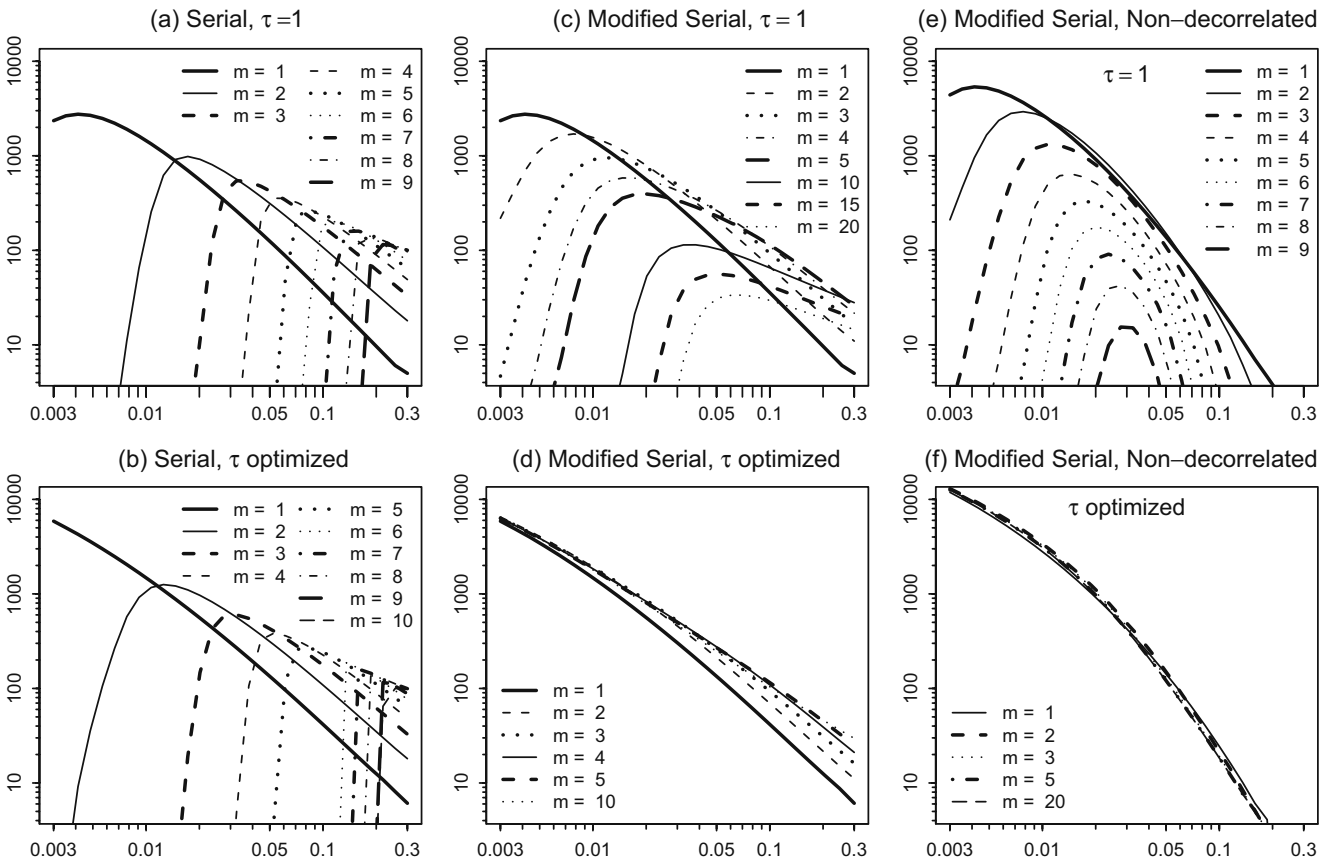

Fig. 7 Memory capacity of the serial-state, modified serial-state, and non-decorrelated modified serial-state model for $N = 10{,}000$, x-axis: coding level f; y-axis: memory capacity. (**ab**) Serial-state model: No matter for $\tau = 1$, or optimized over τ, the minimum coding level f that the model has nonzero capacity increases with m, and complex model does improve the capacity when f is large. (**cd**) Modified serial-state model: when the number of depressed state is reduced to 1 and optimized over τ, the limitation on sparseness due to small initial SNR is removed. The capacity of $m > 1$ is uniformly better than the two-state model over all coding levels. (**ef**) When a non-decorrelating learning rule is used, the capacity decreases as m increases for $\tau = 1$. The improvement in capacity for large coding level disappears

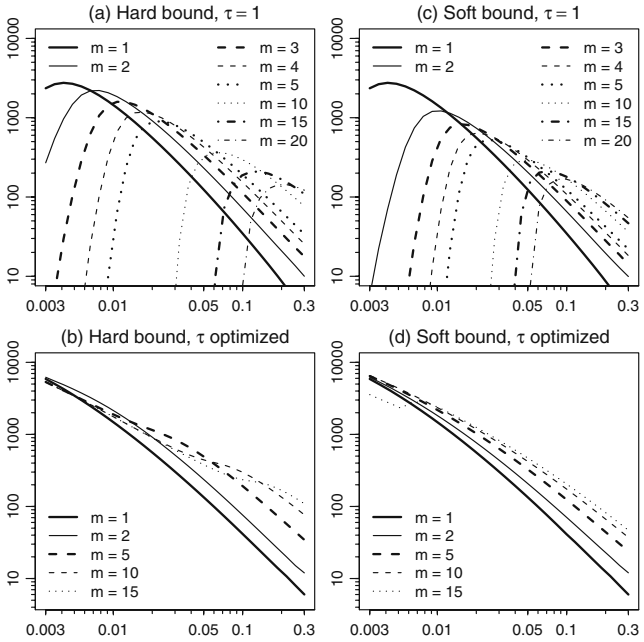

Fig. 8 Memory capacity of the hard-bound and soft bound model for $N = 10{,}000$, using the synapse decorrelation learning rule. *x*-axis: coding level f; *y*-axis: memory capacity

capacity beyond the two-state model for large f (Fig. 7(ab)).

For the modified serial-state model, when optimized over τ (Fig. 7(cd)), the limitation on sparseness due to small initial SNR is completely removed. Moreover, the modified serial-state model with $m > 1$ is uniformly better than the two-state model over all coding levels (Fig. 7(d)), especially for large f, though not as much as the original serial-state model, because the decay rate of the modified serial-state model is larger than that of the original one. On the other hand, the capacity saturates at a certain level as m increases. The optimal capacity for $m = 10$ is nearly the same as that for $m = 5$ (Fig. 7(d)).

The improvement in capacity for large f is primary due the the synapse decorrelation learning rule of Ben Dayan Rubin and Fusi (2007). If synapses are potentiated only when both pre- and postsynaptic neurons are active, and depressed only when the presynaptic neuron is active and postsynaptic is not, as in the sequential model, the capacity of the modified serial-state model is still worse than two-state model (Fig. 7(ef)).

When the synaptic decorrelation learning rule is applied to the hard bound and soft bound model, both models exhibit improvement beyond two-state models for large coding levels (Fig. 8), including the serial-state, and modified serial-state models. This is because the small initial SNR due to large m is compensated by a large f. The slow decay rate when m is large starts taking effect and improves the capacity. Without decorrelation, enlarging f does not enlarge SNR because the synaptic correlation increases linearly with f and inflates the noise level. On the other hand, though multi-state models and synaptic decorrelation do increase capacity for large coding levels, the capacity still decreases as f increases even when optimized over both τ and m. The capacity of a network of 10,000 neurons is less than 300 for $f = 0.1$, which is disappointingly low. Moreover, synaptic decorrelation depends precisely on the values of q_+, q_- and q_0. Any slight perturbation will ruin the result. Furthermore, synaptic decorrelation does not work for multi-level coding.

4 Simulations

To show the accuracy of the approximate retrieval probability in Section 2.4, we performed four simula-

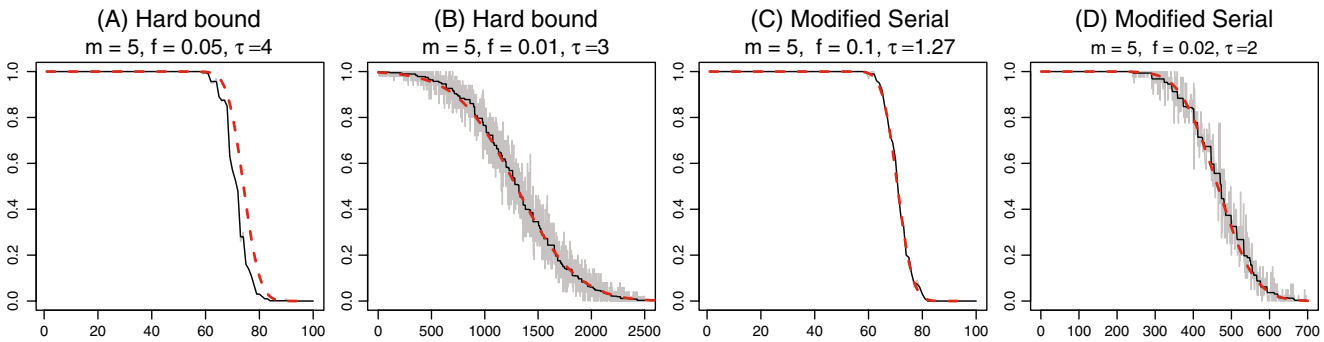

Fig. 9 *x*-axis: age of patterns ; *y*-axis: probability of retrieval. Four simulations with networks of 5,000 neurons for the hard bounds model (**a**, **b**) and the modified serial model (**c**, **d**) with six synaptic states ($m = 5$). The f and τ are marked on the figures. *Jagged grey line*: average number of successful retrievals in 50 runs for each age. *Solid black line*: the non-increasing line that best fits the *jagged grey line* (Robertson et al. 1988). *Red dashed line*: the approximate retrieval probability

tions with fully connected networks of 5,000 neurons for the hard bounds model and the modified serial model with 6 synaptic states ($m = 5$) at two different coding levels (Fig. 9). The coding levels and τ are marked in Fig. 9. In the simulation, the network is first trained with a sequence of random stimuli of the specified coding level. The dynamics is as described in Eq. (2). The criterion for retrieval is described in Section 2.4 with $\varepsilon = 0.05$.

The jagged gray lines in Fig. 9 are the average number of successful retrievals in 50 runs for each age of the patterns (age = 1 is the most recent). The dashed lines are the approximate retrieval probabilities in Eq. (21), which resemble the simulation quite well except for (A), where the approximation slightly overestimates the true retrieval probabilities. Recall that the approximation assumes the independence of the fields. When the coding level gets larger, the correlation between neurons increases and the approximation is less accurate.

In the simulation, the threshold is set with $\delta = 0.01$. That is, we expect the number of non-selective neurons above the threshold to be about 1% of the number of *selective* neurons. In the simulation, the number of false positives relative to the total number of selective neurons are

Simulation	A	B	C	D
False positives	1–5%	3–9%	<0.2%	0.2–1.2%
Skewness	0.151	0.290	−0.044	0.001

higher then the preset value $\delta = 1\%$. Note that the normal approximation to the distribution of the field is accurate when n is large. For the hard bounds model (simulation A and B), the distribution of the field is right-skewed. The right tail is thus fatter than the normal distribution and the simulated false positive rate is higher then the preset value. For simulation C, the field is left-skewed, thus the simulated false positive rate is less than the preset value. Here the skewness is calculated as follows, ignoring the synaptic correlation,

$$\frac{(\sum_i \pi_i w_i^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3\sqrt{Nf}}.$$

5 Extensions

5.1 Partially connected network

In reality networks are never fully connected, rather it is estimated that neurons receive on the order of several thousands of synaptic connections onto their dendrites. For a partially connected network, the capacity depends heavily on the synaptic connectivity configuration, which is beyond the scope of this paper. Here we only study a special case, the *randomly connected network*, where any pair of neurons is *randomly* connected with probability c. The field of a randomly connected network is

$$h_i^{(p)} = \sum\nolimits_{j\neq i} W_{ij}^{(p)} C_{ij} \xi_j^{(1)} - \eta_c \sum\nolimits_j \xi_j^{(1)}.$$

Here C_{ij} are independent Bernoulli(c) and denote the presence of the synapse from neuron j to neuron i. Assume the C_{ij}'s remain fixed throughout learning and pattern retrieval. Given $\sum_j \xi_j^{(1)} = n$, and $\xi_i^{(1)} = x$, the conditional mean and variance of the field (Eq. (11)) become

$$\begin{aligned} M_x^{(p)} &= n(c\mu_x^{(p)} - \eta_c) \\ (R_x^{(p)})^2 &= n[c(\sigma_x^{(p)})^2 + c(1-c)(\mu_x^{(p)})^2] \\ &\quad + n(n-1)c^2\rho_x^{(p)} \end{aligned}$$

The inhibition factor η in Eq. (17) is scaled to give $\eta_c = c\eta$ and the new threshold becomes

$$\theta_c = \frac{C(\sigma^2 + (1-c)\mu^2 - c\rho)}{\sqrt{\rho + \frac{\sigma^2+(1-c)\mu^2-c\rho}{Nfc}} + \sqrt{\rho}} \tag{24}$$

Other parts of the analysis remain unchanged.

To show the accuracy of the approximation above, a simulation is done in a network of $N = 6{,}000$ neurons with connectivity $c = .4$ (Fig. 10(f)). The simulation resembles the approximation quite well.

A randomly connected network of size N with connectivity c is *not* equivalent to a fully connected network of size Nc. Compare the SNR of the two cases,

$$\begin{aligned} \text{former}: &\ \frac{(\mu_1^{(p)} - \mu)^2}{\rho + [\sigma^2 - \rho + (1-c)(\mu^2+\rho)]/(nc)}, \\ \text{latter}: &\ \frac{(\mu_1^{(p)} - \mu)^2}{\rho + (\sigma^2 - \rho)/(nc)}. \end{aligned}$$

The difference could be considerable when c is small. The additional variability in the field is induced by the random connectivity. The number of synapses feeding into a neuron is not always precisely Nc. When c decreases, the capacity will also decrease even when Nc is kept constant. Figure 10 shows the comparison of the optimal capacity of a fully-connected network of size 10,000 and a randomly-connected network of size 100,000 and connectivity $c = 0.1$, in the five models in Section 3. The optimal capacity of a randomly-connected network is about 40% less than that of a fully connected work. Note the optimal τ for a randomly-connected network is greater than that of a fully-connected network.

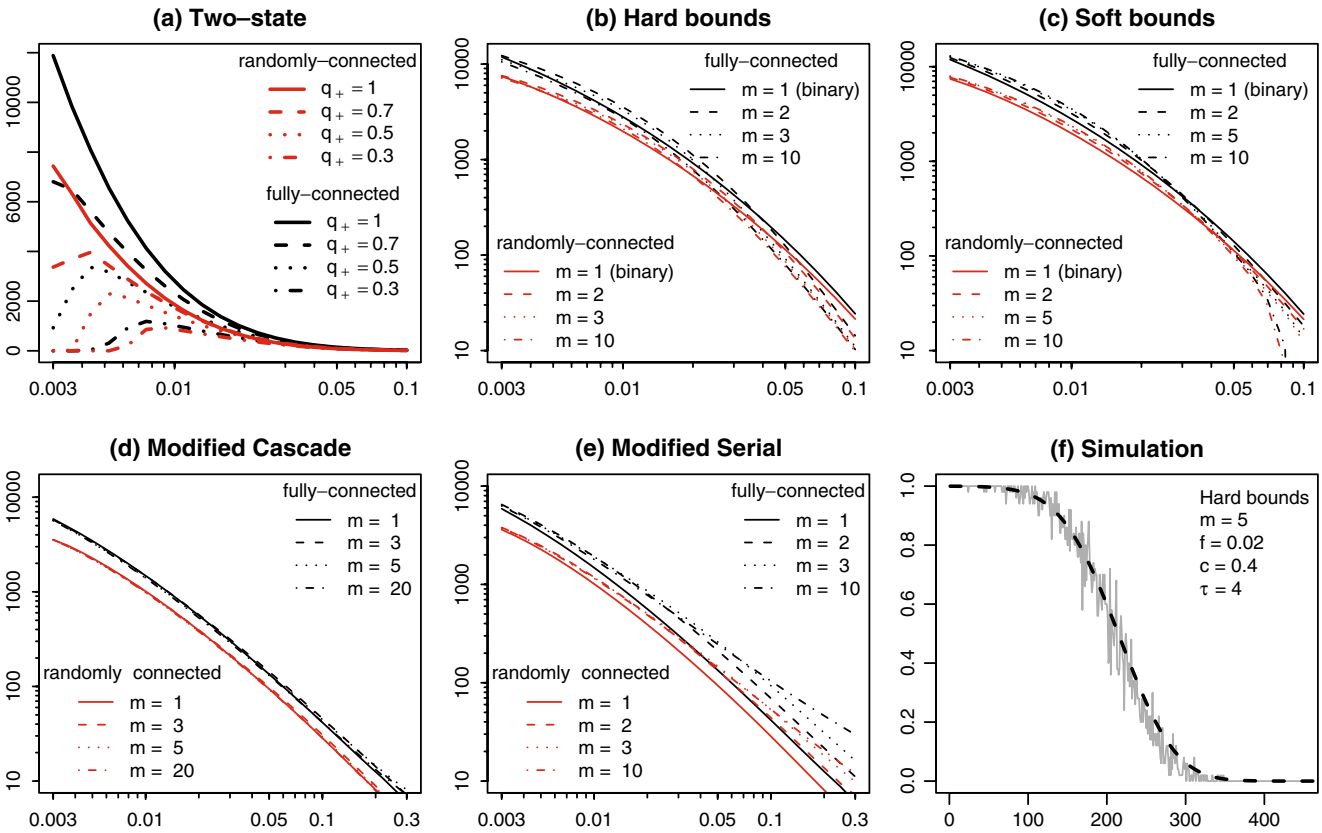

Fig. 10 (**a**–**e**) Comparison of the optimal capacity of a fully-connected network of size 10,000 and a randomly-connected network of size 100,000 and connectivity $c = 0.1$ for each f and m, for the five models considered in Section 3, where the capacity is optimized over τ. x-axis: coding level f; y-axis: memory capacity. The optimal capacity of a randomly-connected network is about 40% less than that of a fully connected work. Note the optimal τ for a randomly-connected network is greater than that of a fully-connected network. Other behaviors of the two networks are similar. (**f**) Randomly connected network of size $N = 6{,}000$ neurons for the hard bound model, $m = 5$, $f = .02$, $c = .4$, $\tau = 4$. *Jagged grey line*—average number of successful retrievals in 50 runs for each age. *Dashed lines*—approximate retrieval probabilities. x-axis: age of patterns ; y-axis: probability of retrieval

Although a reduction in capacity is observed for randomly connected networks, our conclusion about the 5 models remain valid: the limitation on sparseness of multi-state models can be removed by adjusting τ, only a few synaptic states are relevant to improving capacity, and dramatic improvement beyond two-state models is not observed.

5.2 Multiple-level coding stimuli

The analysis above assumes all stimuli come with the same coding level f. This assumption does not appear realistic as a model for real world inputs. It is reasonable to assume that different objects have different numbers of features, i.e. neurons, that are activated. Amit and Huang (2010) extended the analysis of the one-level coding setting to multi-level coding in the case of two-state synaptic models. Here we extend the result to multi-state models.

Since there is no specific preference to any of the inputs we assume that the distribution of ξ_j, $j = 1, \ldots N$ is *exchangeable* in j. The joint distribution then only depends on the number of selective neurons but not on the specific set of active neurons. It can be summarized in terms of marginal probabilities

$$p_{m,n} = \mathsf{P}(\text{the first } m \text{ neurons are } 1, \text{ the next } n \text{ neurons are } 0) \tag{25}$$

where m, n are nonnegative integers with $m + n \leq N$. The most general form we will employ for the joint distribution of the features is

$$p_{m,n} = \int_0^1 f^m (1-f)^n \nu(df), \tag{26}$$

where ν is some distribution on the unit interval. For example, if the stimuli come with k different coding

levels $f_1, \ldots, f_k$, with weights $r_1, \ldots, r_k$, $\sum_{i=1}^{k} r_i = 1$, then

$$p_{m,n} = \sum_{i=1}^{k} r_i f_i^m (1 - f_i)^n.$$

The two transition matrices change to

$$P = p_{2,0} Q^{11} + p_{1,1}(Q^{10} + Q^{01}) + p_{0,2} Q^{00}, \tag{27}$$

$$\begin{aligned} S = {} & p_{3,0} Q^{11} \otimes Q^{11} + p_{1,2} Q^{10} \otimes Q^{10} \\ & + p_{2,1}(Q^{11} \otimes Q^{10} + Q^{10} \otimes Q^{11}) \\ & + p_{2,1} Q^{01} \otimes Q^{01} + p_{0,3} Q^{00} \otimes Q^{00} \\ & + p_{1,2}(Q^{01} \otimes Q^{00} + Q^{00} \otimes Q^{01}). \end{aligned} \tag{28}$$

and the threshold becomes

$$\theta = \frac{C(\sigma^2 - \rho)}{\sqrt{\rho + \frac{\sigma^2 - \rho}{N p_{1,0}}} + \sqrt{\rho}} \tag{29}$$

where $C = C_{\delta, p_{1,0}} = (1 - \delta \frac{p_{1,0}}{p_{0,1}})$-quantile of the standard normal. Other parts of the analysis remain unchanged.

5.3 Associative and feed-forward network

So far the networks we consider are auto-associative networks (AAN) that store patterns one-by-one and require the self-sustainability of patterns in the network dynamics. There are also associative networks (AN) that store paired patterns–cue and target–and require activities in the cue pattern to evoke activities in the target pattern. The feed-forward network (FFN) also stores paired patterns, but in two layers of network, with cue in one layer and target in another, whereas in AN, both cue and target are in the same layer. The cue and target pattern might overlap in AN but not in FFN. Though the biological interpretation of AN and FFN are quite different, in capacity analysis they are very similar.

For AN and FFN, the framework and methodology in Section 2.1 can be applied directly with slight revision. For AN, suppose the learning rule of the synapse only depends on whether the presynaptic neuron is selective for the cue pattern ($k' = 0/1$) and whether the postsynaptic neuron is selective for the target pattern ($k = 0/1$. The learning rule can still be described by the four matrices $Q^{kk'}$, $k, k' = 0/1$ in Section 2.1. The overlap of cue and target pattern does not pose any problem. For FFN, the same rule applies. Assume the cue and target patterns are of coding level f_c and f_t respectively. The size of the AN is N. The size of the cue layer and target layer of FFN are N_c and N_t respectively. Simply replace P_0 and P_1 in Eq. (3) by

$$P_1 = f_c Q^{11} + (1 - f_c) Q^{10}, \; P_0 = f_c Q^{01} + (1 - f_c) Q^{00}. \tag{30}$$

and the two transition matrices P and S by

$$P = f_t P_1 + (1 - f_t) P_0, \tag{31}$$

$$S = f_t P_1 \otimes P_1 + (1 - f_t) P_0 \otimes P_0.. \tag{32}$$

The mean, variances, and covariances $\pi, \gamma, \mu, \sigma, \rho, \pi_x^{(p)}$, $\gamma_x^{(p)}$, $\mu_x^{(p)}$, $(\sigma_x^{(p)})^2$, $\rho_x^{(p)}$, and the inhibition factor η can be calculated in exactly the same way as in Sections 2.2 and 2.3, except that n is replaced by $n_c = N_c f_c$, the size of the cue pattern, $C_{\delta, f}$ is replaced by C_{δ, f_t}, and $x = 0/1$ depends on whether the neuron is in the target pattern. The only subtle difference of AN from FFN is the following. If a neuron is in the cue pattern, the n in the formula of $\mu_x^{(p)}$ and $(\sigma_x^{(p)})^2$ is replaced by $n_c - 1$, not n_c,

The threshold θ in Eq. (18) is replaced by

$$\theta = \frac{C(\sigma^2 - \rho)}{\sqrt{\rho + \frac{\sigma^2 - \rho}{N f_c}} + \sqrt{\rho}}.$$

Like the AAN, a quantitative description of the retrieval criterion for AN and FFN is required. Given ε, and the target pattern size n_t, if the field induced by the cue pattern keeps the field of at least $(1 - \varepsilon) n_t$ target neurons, then we consider the retrieval is successful.[2]

This criterion is simpler than the criterion of AAN as the stability of the dynamics is not an issue and the probability of retrieval can be calculated directly. Replace Eq. (19) by

$$\Psi_{n_c, \varepsilon}^{(p)} \approx \Phi\left(\frac{\theta - n_c(\mu_1^{(p)} - \eta)}{\sqrt{n_c(\sigma_1^{(p)})^2 + n_c(n_c - 1)\rho_1^{(p)}}} \right) \tag{33}$$

and Eq. (20) by

$$\mathsf{P}_{n_t, n_c, \varepsilon}^{(p)} \approx \sum_{k \geq n_t(1-\varepsilon)} \binom{n_t}{k} (\Psi_{n_c, \varepsilon}^{(p)})^k (1 - \Psi_{n_c, \varepsilon}^{(p)})^{n_t - k} \tag{34}$$

and Eq. (21) by

$$\begin{aligned} \mathsf{P}_{\varepsilon}^{(p)} = {} & \sum_{n_c=0}^{N_c} \sum_{n_t=0}^{N_t} \mathsf{P}_{n_t, n_c, \varepsilon}^{(p)} \binom{N_t}{n_t} f_t^{n_t} (1 - f_t)^{N_t - n_t} \\ & \times \binom{N_c}{n_c} f_c^{n_c} (1 - f_c)^{N_c - n_c}. \end{aligned} \tag{35}$$

[2] The amount of allowable non-selective neurons above the threshold is not specified in the criterion since it has been controlled in the threshold selection.

If $\varepsilon = 0$, $\mathsf{P}_{\varepsilon}^{(p)}$ can be simplified as

$$\mathsf{P}_0^{(p)} = \sum_{n_c=0}^{N_c} (f_t \Psi_{n_c,0}^{(p)} + 1 - f_t)^{N_t} \binom{N_c}{n_c} f_c^{n_c} (1 - f_c)^{N_c - n_c}.$$

The capacity can still be defined as the expected number of retrievable paired patterns, as in Eq. (22).

6 Discussion

Intuitively the larger the synaptic state space, the more information the network should be able to store. And indeed if the network readout mechanism in the brain can access these synaptic states capacity should increase with the number of states as in Fusi and Abbott (2007). However, in terms of the concrete readout mechanism defined in this paper—the expected number of retrievable patterns—the capacity of the seven families of multi-state models considered above does not seem to be significantly better than that of two state-models. We note that the retrieval probability measure yields qualitatively similar results to the simpler SNR analysis and in the case of feedforward networks this simply follows from the Gaussian distribution of the fields on the output neurons (as in Leibold and Kempter 2008 and Barret and van Rossum 2008). However in the recurrent setting, retrieval is a function of the network dynamics and therefore it is not straightforward to connect the SNR to the retrieval probability. On the other hand our simulations show that our proposed approximation of the retrieval probability is very accurate, and indeed is closely related although not identical to the SNR analysis.

Among the multi-state models, those with optimal capacity are very similar to the two-state models in that the majority of the synapses concentrate at two synaptic states. Other states are less or rarely visited and have little impact on the learning process. Further increasing the state space, the network capacity either plateaus or decreases. The improvement in retrieval capacity beyond the two-state model is small (20–30% at most). The main reason for this phenomenon is that each training pattern is presented *only once*, so that large capacity can be achieved only through fast learning. Slow learning models will have little or no capacity since they cannot form memory from one-shot presentations. As the number of synaptic levels increases, either the change in efficacy is decreased, or fewer synapses are potentiated, learning becomes slower leading to loss in capacity. If the patterns are allowed to be presented repeatedly, slow learning models can have larger capacity. In upcoming work we hope to extend the retrieval probability analysis allowing repetition in the presentation of the training patterns.

One of the motivations for the work in Ben Dayan Rubin and Fusi (2007) was the observation that certain brain regions exhibit relatively high coding levels. The implications for memory capacity with the original two-state models in Amit and Fusi (1994) were not encouraging where capacity is expected to be very low at high coding levels. Moreover, both Ben Dayan Rubin and Fusi (2007) and Leibold and Kempter (2008) remarked the sparseness of stimuli must be restricted as the synapses grow more complex. However, these constraint are not as stringent as they appear. Our results show that when the ratio τ between the rate of depression and potentiation is properly adjusted, complex synapse models can have large capacity even when the stimuli are sparse (<0.01), and in fact capacity grows with sparseness for optimal τ. For a fully-connected network of size 10,000, the capacity can reach 3,000 when $f = 0.01$, 8,000 when $f = 0.005$ and 12,000 when $f = 0.003$. In the seven models we considered, the capacity is around 30 for dense coding ($f \geq 0.1$). For hard-bound and soft-bound models, increasing the number of synaptic states makes things worse when f is large. Only when the synapse decorrelation learning rule is used the capacity is improved to 100, but this is still very low. The conclusion is the same for partially-connected networks though the capacity is reduced.

Regarding the supposed problem of experimentally observed high coding levels we note that in the hippocampus coding levels appear to be sparse ($f = 0.01 - 0.04$) for both granular (Barnes et al. 1990) and pyramidal cells (Jung and McNaughton 1993), and $f = 0.03$ in the medial temporal lobe for visual stimuli (Quiroga et al. 2005). In inferotemporal cortex, larger coding levels are reported (0.2–0.3) in response to visual stimuli (Rolls and Tovee 1995; Sato et al. 2007). However, as only neurons responding twice as much to faces as to non-faces are included in Rolls and Tovee (1995), i.e., only neurons likely to be selective to visual stimuli are included, the true sparseness could be much lower. Moreover, the measure of sparseness used in Rolls and Tovee (1995) and Sato et al. (2007) creates a significant upwards bias especially if underlying coding levels are very low. This is detailed in the Appendix A.4. Finally, from the perspective of energy consumption, Attwell and Laughlin (2001) and Lennie (2003) estimated that at any given moment only 2% of the population of cortical neurons can afford to be significantly active. This imposes another upper bound on the coding level of the mammalian cortex. See Olshausen and Field (2004) for a review on other advantages of sparse coding.

Appendix

A.1 Kronecker product

If A is an $m \times n$ matrix and B is a $p \times q$ matrix, then the Kronecker product $A \otimes B$ is the $mp \times nq$ block matrix

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

The Kronecker product is bilinear and associative

- $A \otimes (B + C) = A \otimes B + A \otimes C$,
- $(A + B) \otimes C = A \otimes C + B \otimes C$,
- $(kA) \otimes B = A \otimes (kB) = k(A \otimes B)$,
- $(A \otimes B) \otimes C = A \otimes (B \otimes C)$,

where A, B and C are matrices and k is a scalar. If A, B, C and D are matrices of such size that one can form the matrix products AC and BD, then

$$(A \otimes B)(C \otimes D) = AC \otimes BD.$$

This is called the *mixed-product property* because it mixes the ordinary matrix product and the Kronecker product.

A.2 Transition matrices and covariances

As the synaptic modification rule is local, the evolution of the state of a synapse is a Markov chain. Summing over the possible firing states of the pre- and postsynaptic neuron, the probability a synapse transits from α_m to $\alpha_{m'}$ is

$$f^2 q^{11}_{mm'} + f(1-f)q^{10}_{mm'} + (1-f)f q^{01}_{mm'} + (1-f)^2 q^{00}_{mm'},$$

so the transition matrix can then be written as

$$P = f^2 Q^{11} + f(1-f)Q^{10} + f(1-f)Q^{01} + (1-f)^2 Q^{00}.$$

Similarly, a pair of synapses with a common postsynaptic neuron $(J^{(p)}_{ij}, J^{(p)}_{ik})$ is also a Markov chain, where the state space is the cross product $\{\alpha_1, \alpha_2, \ldots, \alpha_M\} \times \{\alpha_1, \alpha_2, \ldots, \alpha_M\}$. Again summing over the firing status of the three neurons i, j, k, the transition probability from (α_l, α_m) to $(\alpha_{l'}, \alpha_{m'})$ is

$$\begin{aligned} &f[fq^{11}_{ll'} + (1-f)q^{10}_{ll'}][fq^{11}_{mm'} + (1-f)q^{10}_{mm'}] \\ &+ (1-f)[fq^{01}_{ll'} + (1-f)q^{00}_{ll'}][fq^{01}_{mm'} + (1-f)q^{00}_{mm'}] \end{aligned}$$

so the transition matrix is

$$\begin{aligned} S = {} & f[fQ^{11} + (1-f)Q^{10}] \otimes [fQ^{11} + (1-f)Q^{10}] \\ & + (1-f)[fQ^{01} + (1-f)Q^{00}] \otimes [fQ^{01} + (1-f)Q^{00}] \end{aligned}$$

Suppose the network is initialized at its stationary state. $\pi^{(0)}_x = \pi$ and $\gamma^{(0)}_x = \gamma$. To obtain the mean and variance after the first step of learning (Eq. (6)), given $\xi^{(1)}_i = x$, since the presynaptic neuron is on $\xi^{(1)}_j = \xi^{(1)}_k = 1$, the transition probability for $J^{(p)}_{ij}$ from α_m to $\alpha_{m'}$ is $q^{x1}_{mm'}$, and the transition probability for $(J^{(p)}_{ij}, J^{(p)}_{ik})$ from (α_l, α_m) to $(\alpha_{l'}, \alpha_{m'})$ is $q^{x1}_{ll'} q^{x1}_{mm'}$. This explains Eq. (6).

Since $J^{(p)}_{ij}$ is a indicators variable, $\mathsf{E}[J^{(p)}_{ij} | \xi^{(1)}_i = x, \xi^{(1)}_j = 1]$ and $\mathsf{E}[J^{(p)}_{ij} \otimes J^{(p)}_{ik} | \xi^{(1)}_i = x, \xi^{(1)}_j = \xi^{(1)}_k = 1]$ are exactly the p-step distribution of the Markov chains $\pi^{(p)}_x$ and $\gamma^{(p)}_x$, and Eq. (7) comes from the Kolmogorov equations.

Equation (8) is straightforward since $W^{(p)}_{ij} = J^{(p)}_{ij}\mathbf{w}^T$. One can verify that $\mathsf{Var}(J^{(p)}_{ij} | \xi^{(1)}_i = x, \xi^{(1)}_j = 1) = \mathrm{Diag}(\pi^{(p)}_x) - \pi^{(p)}_x (\pi^{(p)}_x)^T$ and deduce Eq. (9), where $\mathrm{Diag}(\pi^{(p)}_x)$ is the diagonal matrix with $\pi^{(p)}_x$ on the diagonal. To obtain Eq. (10) we use:

$$\begin{aligned} \mathsf{Cov}(J_{ij}\mathbf{w}^T, J_{ik}\mathbf{w}^T) &= \mathsf{E}[(J_{ij}\mathbf{w}^T)^T J_{ik}\mathbf{w}^T] - \mu^2 \\ &= \mathbf{w}\mathsf{E}[J^T_{ij} J_{ik}]\mathbf{w}^T - \mu^2 \\ &= \gamma(\mathbf{w} \otimes \mathbf{w})^T - \mu^2. \end{aligned}$$

where the last equality comes from the *mixed-product property* of the Kronecker product (See Appendix A.1).

A.3 Decorrelating the synapses

If the four Q matrices are of the form $Q^{11} = I_{2m} + q_+ D_+$, $Q^{01} = Q^{10} = I_{2m} + q_- D_-$, $Q^{00} = I_{2m} + q_0 D_+$, where $q_0 = \frac{f^2 q_+}{(1-f)^2}$, $q_- = \frac{\tau f q_+}{1-f}$, then

$$P_1 = I_{2m} + f q_+ (D_+ + \tau D_-),$$

$$P_0 = I_{2m} + \frac{f^2 q_+}{1-f}(D_+ + \tau D_-),$$

$$P = I + 2f^2 q_+ (D_+ + \tau D_-)$$

If π is the stationary distribution of P, i.e. $\pi P = \pi$, then $\pi(D_+ + \tau D_-) = 0$, which implies $\pi P_1 = \pi P_0 = \pi$. Hence, $\gamma = \pi \otimes \pi$ of is the stationary distribution of S since

$$\begin{aligned} (\pi \otimes \pi)S &= (\pi \otimes \pi)[f P_1 \otimes P_1 + (1-f) P_0 \otimes P_0] \\ &= f(\pi P_1 \otimes \pi P_1) + (1-f)(\pi P_0) \otimes (\pi P_0) \\ &= f(\pi \otimes \pi) + (1-f)(\pi \otimes \pi) = (\pi \otimes \pi). \end{aligned}$$

Thus the stationary synaptic covariance ρ defined in Eq. (14) must be 0.

This learning rule only works for single-level coded stimuli, but not for multi-level coded stimuli.

Table 2 Summary of properties of different multi-state models

Model	States	$\mathbf{w}$	τ	π	μ	σ^2	$\mu_1^{(1)} - \mu$ [a]	Initial SNR [b]
Hard bound	$0, 1, 2, \ldots, m$	$\left(0, \frac{1}{m}, \ldots, \frac{m}{m}\right)$	$= 1$	$\frac{1}{m+1}(1, \ldots, 1)$	0.5	$\frac{1}{12} + \frac{1}{6m}$	$\frac{q_+}{m+1}$	$\frac{12nq_+^2}{(m+1)^2}$
			> 1	$\frac{\tau-1}{\tau-\tau^{-m}}\left(1, \frac{1}{\tau}, \ldots, \frac{1}{\tau^m}\right)$	$\frac{1}{m}\left(\frac{1}{\tau-1} - \frac{m+1}{\tau^{m+1}-1}\right)$	$\approx \frac{\tau}{m^2(\tau-1)^2}$	$\approx \frac{q_+}{m}$	$\approx \frac{nq_+^2(\tau-1)^2}{\tau}$
			< 1				$\approx \frac{q_+\tau}{m}$	$\approx nq_+^2\tau(\tau-1)^2$
Soft bound	$\{0, 1, 2, \ldots, m\}$	$\left(0, \frac{1}{m}, \ldots, \frac{m}{m}\right)$		$\text{Binomial}\left(m, \frac{1}{1+\tau}\right)$ $(\pi_0, \pi_1, \ldots, \pi_m)$, $\pi_i = \binom{m}{i}\left(\frac{1}{1+\tau}\right)^i\left(\frac{\tau}{1+\tau}\right)^{m-i}$, $0 \le i \le m$	$\frac{1}{1+\tau}$	$\frac{\tau}{m(1+\tau)^2}$	$\approx \frac{n\tau q_+}{m(1+\tau)}$	$\approx \frac{n\tau q_+^2}{m}$
Cascade	$\{-1, -2, \ldots, -m,$ $1, 2, \ldots, m\}$	$(0, 0 \ldots, 0,$ $1, 1, \ldots, 1)$	$= 1$	$\frac{1}{2m}(1, \ldots, 1)$	0.5	$\mu(1-\mu)$	$\frac{q_+}{2m}$	nq_+^2/m^2
			> 1	$(\pi_{-1}, \ldots, \pi_{-m},$ $\pi_1, \ldots, \pi_m)^*$	$***$		$\pi_1 q_+ \frac{1+\tau}{2}$	$\approx \frac{n(\tau^2-1)\pi_1}{4}$
			< 1		$\approx \tau$			$\approx \frac{n(1+\tau)\pi_1^2}{\tau(1-\tau)}$
Modified cascade	$\{-1, 1, 2, \ldots, m\}$	$(0, 1, 1, \ldots, 1)$	$= 1$	$\frac{1}{m+1}(1, \ldots, 1)$	$\frac{m}{1+m}$	$\mu(1-\mu)$	$\frac{q_+}{m+1}$	nq_+^2/m
			> 1	$(\pi_0, \pi_1, \ldots, \pi_m)^{**}$	$1 - \pi_0$		$q_+\pi_0$	$\approx nq_+^2\frac{\tau-1}{2}$
			< 1					$\approx nq_+^2\pi_0$
Serial	$\{-1, -2, \ldots, -m,$ $1, 2, \ldots, m\}$	$(0, 0 \ldots, 0,$ $1, 1, \ldots, 1)$	$= 1$	$\frac{1}{2m}(1, \ldots, 1)$	0.5	$\mu(1-\mu)$	$\frac{q_+}{2m}$	nq_+^2/m^2
			$\ne 1$	$\frac{\tau-1}{1-\tau^{-2m}}(\tau^{-m}, \ldots, \tau^{-1},$ $\tau^{-m-1}, \ldots, \tau^{-2m})$	$\frac{1}{(1+\tau^m)}$		$\frac{q_+(\tau-1)\tau^{-m}}{1-\tau^{-2m}}$	$\frac{nq_+^2(\tau-1)^2\tau^{-m}}{(1-\tau^{-m})^2}$
Modified serial	$\{-1, 1, 2, \ldots, m\}$	$(0, 1, 1, \ldots, 1)$	$= 1$	$\frac{1}{m+1}(1, \ldots, 1)$	$\frac{m}{1+m}$	$\mu(1-\mu)$	$\frac{q_+}{m+1}$	nq_+^2/m
			$\ne 1$	$\frac{\tau-1}{\tau-\tau^{-m}}\left(1, \frac{1}{\tau}, \ldots, \frac{1}{\tau^m}\right)$	$\frac{\tau^m-1}{\tau^{m+1}-1}$		$q_+(1-\mu)$	$\frac{nq_+^2(\tau-1)}{1-\tau^{-m}}$

[a] $\mu_1^{(1)} = \pi Q^{11}\mathbf{w}$.

[b] The initial SNR$= n(\mu_1^{(1)} - \mu)^2/\sigma^2$, where the synaptic covariance is ignored.

* For the cascade model: $\pi_1 = \pi_{-1} = \frac{\tau-1}{\tau+1}(\tau r_2^{m-2} - r_1^{m-2}/\tau)^{-1}$, $\pi_i = \pi_1 r_1^{i-1}$, $\pi_{-i} = \pi_1 r_2^{i-1}$ for $1 \le i < m$, $\pi_m = \pi_1 r_1^{m-2}/\tau$, $\pi_{-m} = \pi_1 r_2^{m-2}\tau$, and $\mu = (1 - r_1^{m-2})(\tau r_2^{m-2} - r_1^{m-2}/\tau)^{-1}$. Here $r_1 = 2/(1+\tau)$, $r_2 = 2\tau/(1+\tau)$.

** For the modified cascade model: $\pi_0 = \frac{\tau-1}{\tau+1}(1 - r^{m-1}/\tau)^{-1}$, $\pi_i = \pi_0 r^i$, for $1 \le i < m$, $\pi_m = \pi_0 r^{m-1}/\tau$. Here $r = 2/(1+\tau)$

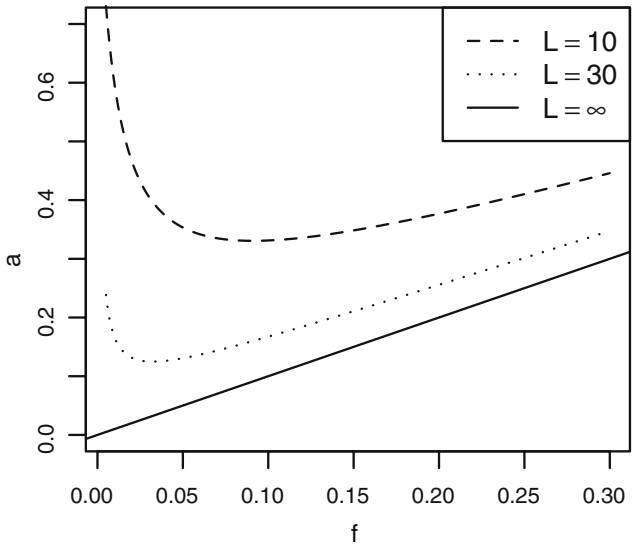

Fig. 11 The sparseness measure a (Eq. (36)) versus the actual coding level f for $L = 10$ and 30

A.4 Bias in sparseness measurement

In Rolls and Tovee (1995) and Sato et al. (2007) coding level is defined as

$$a = \frac{(\sum_i^n r_i/n)^2}{\sum_i^n r_i^2/n} \tag{36}$$

where r_i is the firing rate of the neuron to the ith stimulus in a set of n stimuli. Say the baseline firing rate of the neuron is r, and if the neuron is selective to the stimulus, the firing rate is Lr $(L > 1)$, and assume the true sparseness is f. Then on average $\sum_i r_i/n \approx fLr + (1-f)r$, $\sum_i r_i^2/n \approx fL^2r^2 + (1-f)r^2$.

$$a \approx \frac{(fL + 1 - f)^2}{fL^2 + 1 - f} > f$$

Though $a \to f$ as $L \to \infty$, when $f \to 0$, $a \approx 1 - f$. The relationship between a and f is not monotone and when $L = 30$, a is always above 0.1 whatever f is (Fig. 11). If the baseline firing rate is subtracted from each r_i, then L will be larger and make a closer to f.

References

Abraham, W. C., & Bear, M. F. (1996). Metaplasticity: The plasticity of synaptic plasticity. *Trends in Neurosciences, 19*(4), 126–130. doi:10.1016/S0166-2236(96)80018-X.

Amit, D. J., & Brunel, N. (1997a). Dynamics of recurrent network of spiking neurons before and following learning. *Network, 8*, 373–404.

Amit, D. J., & Brunel, N. (1997b). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral Cortex, 7*, 237–252.

Amit, D. J., & Fusi, S. (1994). Learning in neural networks with material synapses. *Neural Computation, 6*, 957–982.

Amit, D. J., & Mongillo, G. (2003). Selective delay activity in the cortex: Phenomena and interpretation. *Cerebral Cortex, 13*, 1139–1150.

Amit, Y., & Huang, Y. (2010). Precise capacity analysis in binary networks with multiple coding level inputs. *Neural Computation, 22*(3), 660–688. doi:10.1162/neco.2009.02-09-967.

Attwell, D., & Laughlin, S. B. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow and Metabolism, 21*(10), 1133–1145.

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

Barret, A. B., & van Rossum, M. C. (2008). Optimal learning rules for discrete synapses. *PLoS Computational Biology, 4*, 1–7.

Ben Dayan Rubin, D. D., & Fusi, S. (2007). Long memory lifetimes require complex synapses and limited sparseness. *Frontiers in Computational Neuroscience, 1*, 1–14.

Brunel, N. (2003). Dynamics and plasticity of stimulus-selective persistent activity in cortical network models. *Cerebral Cortex, 13*, 1151–1161.

Curti, E., Mongillo, G., La Camera, G., & Amit, D. J. (2004). Mean-field and capacity in realistic networks of spiking neurons storing sparsely coded random memories. *Neural Computation, 16*, 2597–2637.

Del Giudice, P., Fusi, S., & Mattia, M. (2003). Modelling the formation of working memory with networks of integrate-and-fire neurons connected by plastic synapses. *Journal of Physiology Paris, 97*(4–6), 659–681. doi:10.1016/j.jphysparis.2004.01.021.

Fusi, S., & Abbott, L. F. (2007). Limits on the memory storage capacity of bounded synapses. *Nature Neuroscience, 10*, 485–493.

Fusi, S., Drew, P. J., & Abbott, L. (2005). Cascade models of synaptically stored memories. *Neuron, 45*(4), 599–611. doi:10.1016/j.neuron.2005.02.001.

Fuster, J. (1995). *Memory in the cerebral cortex: An empirical approach to neural networks in the human and nonhuman primate*. Cambridge, MA: MIT Press.

Jung, M. W., & McNaughton, B. L. (1993). Spatial selectivity of unit activity in the hippocampal granular layer. *Hippocampus, 3*(0), 165–182.

Leibold, C., & Kempter, R. (2008). Sparseness constrains the prolongation of memory lifetime via synaptic metaplasticity. *Cerebral Cortex 18*, 67–77.

Lennie P (2003) The cost of cortical computation. *Current Biology, 13*(6), 493–497. doi:10.1016/S0960-9822(03)00135-0.

Miyashita, Y., & Hayashi, T. (2000). Neural representation of visual objects: Encoding and top-down activation. *Current Opinion in Neurobiology, 10*(2), 187–194.

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology, 14*(4), 481–487. doi:10.1016/j.conb.2004.07.007.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature, 435*, 1102–1107.

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order restricted statistical inference*. Wiley.

Rolls, E. T., & Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Physiology, 73*(2), 713–726.

Romani, S., Amit, D., & Amit, Y. (2008). Optimizing one-shot learning with binary synapses. *Neural Computation, 20*, 1928–1950.

Sato, T., Uchida, G., & Tanifuji, M. (2007). The nature of neuronal clustering in inferotemporal cortex of macaque monkey revealed by optical imaging and extracellular recording. In *34th Ann. meet. of soc. for neuroscience*. San Diego, USA.

Wang, X. J. (2001). Synaptic reverberation underlying mnemonic persistent activity. *Trends in Neurosciences, 24*(8), 455–463. doi:10.1016/S0166-2236(00)01868-3.

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