Memory capacity of networks with stochastic binary synapses

Alexis M. Dubreuil[1,2], Yali Amit[3] and Nicolas Brunel[1,2]

[1] UMR 8118, CNRS, Université Paris Descartes, Paris, France
[2] Departments of Statistics and Neurobiology, University of Chicago, Chicago, IL, USA
[3] Departments of Statistics and Computer Science, University of Chicago, Chicago, IL, USA

Abstract

In standard attractor neural network models, specific patterns of activity are stored in the synaptic matrix, so that they become fixed point attractors of the network dynamics. The storage capacity of such networks has been quantified in two ways: the maximal number of patterns that can be stored, and the stored information measured in bits per synapse. In this paper we compute both quantities in fully connected networks of N binary neurons with binary synapses, storing patterns with coding level f, in the large N and sparse coding limits ($N \to \infty$, $f \to 0$). We also derive finite-size corrections that accurately reproduce the results of simulations in networks of tens of thousands of neurons. These methods are applied to three different scenarios: (1) the classic Willshaw model, (2) networks with stochastic learning in which patterns are shown only once (one shot learning), (3) networks with stochastic learning in which patterns are shown multiple times. The storage capacities are optimized over network parameters, which allows us to compare the performance of the different models. We show that finite-size effects strongly reduce the capacity, even for networks of realistic sizes. We discuss the implications of these results for memory storage in hippocampus and cerebral cortex.

Author Summary

Two central hypotheses in neuroscience is that long-term memory is sustained by modifications of the connectivity of neural circuits, while short-term memory is sustained by persistent neuronal activity following the presentation of a stimulus. These two hypotheses have been substantiated by several decades of electrophysiological experiments, reporting activity-dependent changes in synaptic connectivity *in vitro*, and stimulus-selective persistent neuronal activity in delayed response tasks in behaving monkeys. They have been implemented in attractor network models, that store specific patterns of activity using Hebbian plasticity rules, which then allow to retrieve these patterns as attractors of the network dynamics. A long-standing question in the field is, how many patterns (or equivalently, how much information) can be stored in such networks? Here, we compute the storage capacity of networks of binary neurons and binary synapses. Synapses store information according to a simple stochastic learning process, that consists in transitions between synaptic states conditioned on the states of pre and post-synaptic neurons. We consider this learning process in two limits: a one shot learning scenario, where each pattern is presented only once;

and a slow learning scenario, where noisy versions of a set of patterns are presented multiple times, but transition probabilities are small. The two limits are assumed to represent in a simplified way learning in hippocampus and neocortex, respectively. We show that in both cases, the information stored per synapse remains finite in the large N limit, when the coding is sparse. Furthermore, we characterize the strong finite size effects that exist in such networks.

Introduction

Attractor neural networks have been proposed as long-term memory storage devices [1, 2, 3]. In such networks, a pattern of activity (the set of firing rates of all neurons in the network) is said to be memorized if it is one of the stable states of the network dynamics. Specific patterns of activity become stable states thanks to synaptic plasticity mechanisms, including both long term potentiation and depression of synapses, that create positive feed-back loops through the network connectivity. Attractor states are consistent with the phenomenon of selective persistent activity during delay periods of delayed response tasks, which has been documented in numerous cortical areas in behaving monkeys [4, 5, 6, 7]. A long standing question in the field has been the question of the storage capacity of such networks. Much effort has been devoted to compute the number of attractor states that can be imprinted in the synaptic matrix, in networks of binary neurons [8, 9, 10, 11]. Models storing patterns with a covariance rule [12, 1, 8, 11] were shown to be able to store a number of patterns that scale linearly with the number of synapses per neuron. In the sparse coding limit (in which the average fraction of selective neurons per pattern f goes to zero in the large N limit), the capacity was shown to diverge as $1/(f|\log(f)|)$. These scalings lead to a network storing on the order of 1 bit per synapse, in the large N limit, for any value of the coding level. Elizabeth Gardner [10] computed the maximal capacity, in the space of all possible coupling matrices, and demonstrated a similar scaling for capacity and information stored per synapse.

These initial studies, performed on the simplest possible networks (binary neurons, full connectivity, unrestricted synaptic weights) were followed by a second wave of studies that examined the effect of adding more neurobiological realism: random diluted connectivity [9], neurons characterized by analog firing rates [13], learning rules in which new patterns progressively erase the old ones [14, 15]. The above mentioned modifications were shown not to affect the scaling laws described above. One particular modification however was shown to have a drastic effect on capacity. A network with binary synapses and stochastic on-line learning was shown to have a drastically impaired performance, compared to networks with continuous synapses [16, 17]. For finite coding levels, the storage capacity was shown to be on the order of $\sqrt{N}$, not N stored patterns, while the information stored per synapse goes to zero in the large N limit. In the sparse coding limit however ($f \sim \log(N)/N$), the capacity was shown to scale as $1/f^2$, and therefore a similar scaling as the Gardner bound, while the information stored per synapse remains finite in this limit. These scaling laws are similar to the Willshaw model [18], which can be seen as a particular case of the Amit-Fusi [17] rule. The model was then subsequently studied in greater detail by Huang and Amit [19, 20] who computed the storage capacity for finite values of N, using numerical simulations and several approximations for the distributions of the 'local fields'

of the neurons. However, computing the precise storage capacity of this model in the large N limit remains an open problem.

In this article we focus on a model of binary neurons where binary synapses are potentiated or depressed stochastically depending on the states of pre and post synaptic neurons [17]. We first introduce analytical methods that allow us to compute the storage capacity in the large N limit, based on a binomial approximation for the synaptic inputs to the neurons. We first illustrate it on the Willshaw model and to recover the well-known result on the capacity of this model [18, 21, 22]. We then move to a stochastic learning rule, in which we study two different scenarios: (i) in which patterns are presented only once - we will refer to this model as the SP (Single Presentation) model [17]; (ii) in which noisy versions of the patterns are presented multiple-times - the MP (Multiple presentations) model [23]. For both models we compute the storage capacity and the information stored per synapse in the large N limit, and investigate how they depend on the various parameters of the model. We then study finite size effects, and show that they have a huge effect even in networks of tens of thousands of neurons. Finally we show how capacity in finite size networks can be enhanced by introducing inhibition, as proposed in [19, 20]. In the discussion we summarize our results and discuss the relevance of the SP and MP networks to memory maintenance in the hippocampus and cortex.

Results

1 Storage capacity in the $N \to \infty$ limit

1.1 The network

We consider a network of N binary (0,1) neurons, fully connected through a binary (0,1) synaptic connectivity matrix. The activity of neuron i ($i = 1...N$) is described by a binary variable, $\sigma_i = 0, 1$. Each neuron can potentially be connected to every other neurons, through a binary connectivity matrix $\mathbf{W}$. This connectivity matrix depends on P random uncorrelated patterns ('memories') $\vec{\xi}^{\mu}, \mu = 1, \ldots, P$ that are presented during the learning phase. The state of neuron $i = 1, \ldots, N$ in pattern $\mu = 1, \ldots, P$ is

$$\xi_i^{\mu} = \begin{cases} 1 & \text{with probability } f \\ 0 & \text{with probability } 1 - f \end{cases} \tag{1}$$

where f is the coding level of the memories. We study this model in the limit of low coding level, $f \to 0$ when $N \to \infty$. In all the models considered here, P scales as $1/f^2$ in the sparse coding limit. Thus, we introduce a parameter $\alpha = Pf^2$ which stays of order 1 in the sparse coding limit.

After the learning phase, we choose one of the P presented patterns $\vec{\xi}^{\mu_0}$, and check whether it is a fixed point of the dynamics:

$$\sigma_i(t+1) = \Theta[h_i(t) - fN\theta], \tag{2}$$

where

$$h_i(t) = \sum_{j=1}^{N} W_{ij}\sigma_j(t) \tag{3}$$

is the total synaptic input ("field") of neuron i, θ is a scaled activation threshold (constant independent of N), and Θ is the Heaviside function.

1.2 Field averages

When testing the stability of pattern $\vec{\xi}^{\mu_0}$ after learning P patterns, we need to compute the distribution of the fields on selective neurons (sites i such that $\xi_i^{\mu_0} = 1$), and of the fields on non-selective neurons (sites i such that $\xi_i^{\mu_0} = 0$). The averages of those fields are fNg_+ and fNg respectively, where

$$g_+ = \mathbb{P}(W_{ij} = 1 | \xi_i^{\mu_0} = \xi_j^{\mu_0} = 1) \tag{4}$$

and

$$g = \mathbb{P}(W_{ij} = 1 | \, (\xi_i^{\mu_0}, \xi_j^{\mu_0}) \neq (1,1) \,). \tag{5}$$

Pattern $\vec{\xi}^{\mu_0}$ is perfectly imprinted in the synaptic matrix if $g_+ = 1$ and $g = 0$. However, because of the storage of other patterns, g_+ and g take intermediate values between 0 and 1. Note that here we implicitly assume that the probability of finding an potentiated synapse between two neurons i, j such that $\xi_i^{\mu_0} = \xi_j^{\mu_0} = 0$ or $\xi_i^{\mu_0} \neq \xi_j^{\mu_0}$ is the same. This is true for the models we consider below. g_+ and g are function of α, f, and other parameters characterizing learning.

1.3 Information stored per synapse

One measure of the storage capability of the network is the information stored per synapse :

$$i = \frac{P_{max} N(-f \log_2 f - (1-f)\log_2(1-f))}{N^2} \tag{6}$$

$$\underset{f \to 0}{\simeq} \alpha \frac{|\log_2 f|}{fN} \tag{7}$$

where P_{max} is the size of a set of patterns in which each pattern is a fixed point of the dynamics with probability one. When α is of order one, for the information per synapse to be of order one in the large N limit, we need to take f as

$$f = \beta \frac{\ln N}{N}. \tag{8}$$

In this case the information stored per synapse has the simple expression:

$$i = \frac{\alpha}{\beta \ln 2} \tag{9}$$

1.4 Computing the storage capacity

Our goal here is to compute the size $P_{max} = \alpha/f^2$ of the largest set of patterns that can be stored in the connectivity matrix. The criterion for storage that we adopt is that if one picks a pattern in this set, then this pattern is a fixed point of the dynamics with probability 1. We thus need to compute the probability $\mathbb{P}_{ne}$ of no error in retrieving a particular pattern μ_0. To compute this probability, we first need to estimate the probabilities that a single selective/non-selective neuron is in its right state when the network is initialized in a state corresponding to pattern μ_0. For a pattern with M selective neurons, and neglecting correlations between neurons (which is legitimate if $f \ll 1/\sqrt{N}$ [17]), we have

$$\mathbb{P}_{ne} = (1 - \mathbb{P}(h_i \leq fN\theta | \xi_i^{\mu_0} = 1))^M \, (1 - \mathbb{P}(h_i \geq fN\theta | \xi_i^{\mu_0} = 0))^{N-M} \tag{10}$$

Clearly, for $\mathbb{P}_{ne}$ to go to 1 in the large N limit, the probabilities for the fields of single neurons to be on the wrong side of the threshold have to vanish in that limit. A first condition for this to happen is $g_+ > \theta > g$ - if these inequalities are satisfied, then the average fields of both selective and non-selective neurons are on the right side of the threshold. When g_+ and g are sufficiently far from θ, the tail probabilities of the distribution of the fields are

$$\begin{aligned} \mathbb{P}\left(h_i \leq fN\theta | \xi_i^{\mu_0} = 1\right) &= \exp\left(-M\Phi(g_+, \theta) + o(M)\right) & (11) \\ \mathbb{P}\left(h_i \geq fN\theta | \xi_i^{\mu_0} = 0\right) &= \exp\left(-M\Phi(g, \theta) + o(M)\right) & (12) \end{aligned}$$

where $\Phi(g_+, \theta)$, $\Phi(g, \theta)$ are the rate functions associated with the distributions of the fields (see Methods A). Neglecting again correlations between inputs, the distributions of the fields are binomial distributions, and the rate functions are

$$\Phi(x, \theta) = \theta \ln \frac{\theta}{x} + (1 - \theta) \ln \frac{1 - \theta}{1 - x} \tag{13}$$

Inserting Eqs. (11,12,13,8) in Eq. (10), we find that

$$\mathbb{P}_{ne} = \exp\left[-\exp\left(X_s\right) - \exp\left(X_n\right)\right] \tag{14}$$

where

$$\begin{aligned} X_s &= -\beta\Phi(g_+, \theta) \ln N + \ln\ln N + o(\ln\ln N) \\ X_n &= -\beta\Phi(g, \theta) \ln N + \ln N + o(\ln N). \end{aligned} \tag{15}$$

For $\mathbb{P}_{ne}$ to go to 1 in the large N limit, we need both X_s and X_n to go to $-\infty$ in that limit. This will be satisfied provided

$$\begin{aligned} \Phi(g_+, \theta) &> \frac{\ln\ln N}{\beta \ln N} & (16) \\ \Phi(g, \theta) &> \frac{1}{\beta} & (17) \end{aligned}$$

These inequalities are equivalent in the large N limit to the inequalities

$$g_+ > \theta > g + \zeta \tag{18}$$

where ζ is given by the equation $\Phi(g+\zeta,\theta)=1/\beta$.

The maximal information per synapse is obtained by saturating inequalities (16) and (17), and optimizing over the various parameters of the model. In practice, for given values of α, and parameters of the learning process, we compute g and g_+; we can then obtain the optimal values of the threshold θ and the rescaled coding level β as

$$\theta \underset{N\to+\infty}{\longrightarrow} g_+ \tag{19}$$

$$\beta = \frac{1}{\Phi(g,\theta)}, \tag{20}$$

and compute the information per synapse using Eq. (9). We can then find the optimum of i in the space of all parameters.

Before applying these methods to various models, we would like to emphasize two important features of these calculations:

- In Eq. (16), note that the r.h.s. goes to zero extrememely slowly as N goes to ∞ (as $\ln\ln N/\ln N$) - thus, we expect huge finite size effects. This will be confirmed in Section 5 where these finite size effects are studied in detail.
- In the sparse coding limit, a Gaussian approximation of the fields gives a poor approximation of the storage capacity, since the calculation probes the tail of the distribution.

2 Willshaw model

The capacity of the Willshaw model has already been studied by a number of authors [18, 21, 22]. Here, we present the application of the analysis described in Section 1 to the Willshaw model, for completeness and comparison with the models described in the next Section. In this model, after presenting P patterns to the network, the synaptic matrix is described as follows: $W_{ij}=1$ if at least one of the P presented patterns had neuron i and j co-activated, $W_{ij}=0$ otherwise. Thus, after the learning phase, we have,

$$\begin{aligned} g_+ &= 1 \\ g &= 1-(1-f^2)^P \simeq 1-\exp(-\alpha) \text{ for small } f \end{aligned} \tag{21}$$

Saturating the inequalities (19),(20) with g fixed, one obtains the information stored per synapse,

$$i_{opt} = \ln(1-g)\ln g\frac{1}{\ln 2} \tag{22}$$

The information stored per synapse is shown as a function of g in figure 1a. A maximum is reached for $g=0.5$ at $i_W=\ln 2=0.69$ bits/synapse, but goes to zero in both the $g\to 0$ and $g\to 1$ limits. The model has a storage capacity comparable to its maximal value, $i_{opt}>0.5 i_W$ in a large range of values of g (between 0.1 and 0.9. We can also optimize capacity for a given value of β, as shown in figure 1b. It reaches its maximum at $\beta=1.4$, and goes to zero in the small and large β limits. Again, the model has a large storage capacity for a broad range of β, $i_{opt}>0.5 i_W$ for β between 0.4 and 10.

Previous studies [18, 21] have found an optimal capacity of 0.69 bits/synapse. Those studies focused on a feed-forward network with a single output neuron, with no fluctuations

in the number of selective neurons per pattern, and required that the number of errors on silent outputs is of the same order as the number of selective outputs in the whole set of patterns. In the calculations presented here, we have used a different criteria, namely that a given pattern (not all patterns) is exactly a fixed point of the dynamics of the network with a probability that goes to one in the large N limit. Another possible definition would be to require that **all** the P patterns are exact fixed points with probability one. In this case, for patterns with fixed numbers of selective neurons, the capacity drops by a factor of 3, $\ln(2)/3 = 0.23$, as already computed by Knoblauch et al [22].

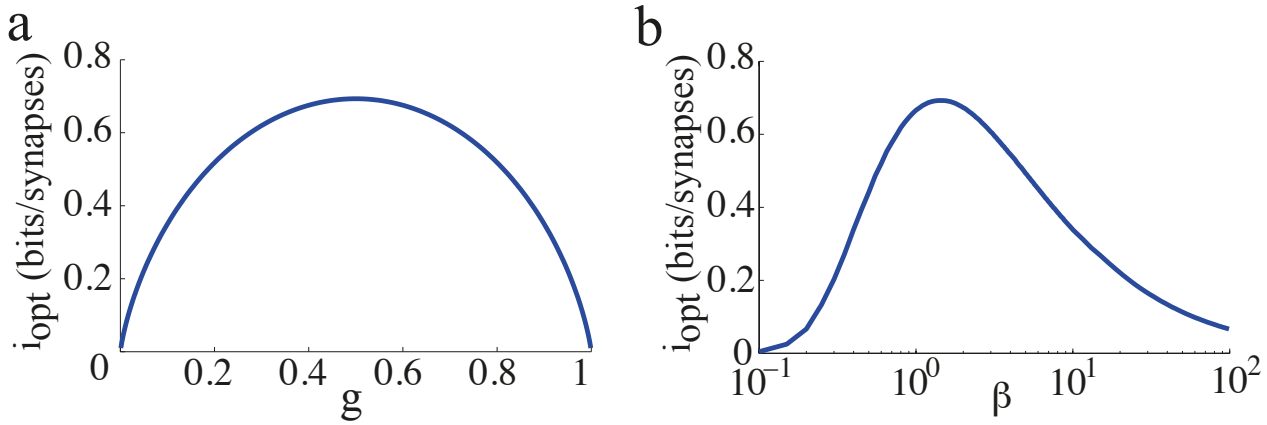

Figure 1: *Optimized information capacity of the Willshaw model in the limit $N \to +\infty$. Information is optimized by saturating (19) ($\theta = 1$) and (20): a. i_{opt} as a function of g, b. i_{opt} as a function of $\beta = fN/\ln N$.*

3 Amit-Fusi model

A drawback of the Willshaw learning rule is that it only allows for synaptic potentiation. Thus, if patterns are continuously presented to the network, all synapses will eventually be potentiated and no memories can be retrieved. In [17] Amit and Fusi introduced a new learning rule that maintains the simplicity of the Willshaw model, but allows for continuous on-line learning. The proposed learning rule includes synaptic depression. At each learning time step μ, a new pattern $\vec{\xi}^{\mu}$ with coding level f is presented to the network, and synapses are updated stochastically:

- for synapses such that $\xi_i^\mu = \xi_j^\mu = 1$:
  if $W_{ij}(\mu-1) = 0$, then $W_{ij}(\mu)$ is potentiated to 1 with probability q_+ ; and if $W_{ij}(\mu-1) = 1$ it stays at 1.
- for synapses such that $\xi_i^\mu \neq \xi_j^\mu$:
  if $W_{ij}(\mu-1) = 0$, then $W_{ij}(\mu)$ stays at 0 ; and if $W_{ij}(\mu-1) = 1$ it is depressed to 0 with probability q_-.
- for synapses such that $\xi_i^\mu = \xi_j^\mu = 0$, $W_{ij}(\mu) = W_{ij}(\mu-1)$.

The evolution of a synapse W_{ij} during learning can be described by the following Markov process :

$$\begin{bmatrix} \mathbb{P}(W_{ij}^{\mu+1}=0) \\ \mathbb{P}(W_{ij}^{\mu+1}=1) \end{bmatrix} = \begin{bmatrix} 1-a & b \\ a & 1-b \end{bmatrix} \times \begin{bmatrix} \mathbb{P}(W_{ij}^{\mu}=0) \\ \mathbb{P}(W_{ij}^{\mu}=1) \end{bmatrix} \tag{23}$$

where $a = f^2 q_+$ is the probability that a silent synapse is potentiated upon the presentation of pattern μ and $b = 2f(1-f)q_-$ is the probability that a potentiated synapse is depressed. After a sufficient number of patterns has been presented the distribution of synaptic weights in the network reaches a stationary state. We study the network in this stationary regime.

For the information capacity to be of order 1, the coding level has to scale as $\frac{\ln N}{N}$, as in the Willshaw model, and the effects of potentiation and depression have to be of the same order [17]. Thus we define the *depression-potentiation ratio* δ as,

$$\delta = \frac{2f(1-f)q_-}{f^2 q_+} \tag{24}$$

We can again use equation (9) and the saturated inequalities (19,20) to compute the maximal information capacity in the limit $N \to \infty$. This requires computing g and g_+, defined in the previous section, as a function of the different parameters characterizing the network. We track a pattern $\vec{\xi^{\mu_0}}$ that has been presented P time steps in the past. In the following we refer to P as the age of the pattern. In the sparse coding limit, g corresponds to the probability that a synapse is potentiated. It is determined by the depression-potentiation ratio δ,

$$g = \frac{1}{1+\delta} \tag{25}$$

and

$$g_+ = g + q_+(1-g)(1-a-b)^P \simeq g + q_+(1-g)\exp(-\frac{q_+\alpha}{g}) \text{ for } f \ll 1 \tag{26}$$

where $\alpha = Pf^2$. Our goal is to determine the age P of the oldest pattern that is still a fixed point of the network dynamics, with probability one. Note that in this network, contrary to the Willshaw model in which all patterns are equivalent, here younger patterns, of age $P' < P$, are more strongly imprinted in the synaptic matrix, $g_+(P') > g_+(P)$, and thus also stored with probability one.

Choosing an activation threshold and a coding level that saturate inequalities (19) and (20), information capacity can be expressed as :

$$\begin{aligned} i_{opt} &= \frac{g}{q_+} \ln\left[q_+ \frac{1-g}{g_+-g}\right]\left[g_+ \log_2 \frac{g_+}{g} + (1-g_+)\log_2 \frac{1-g_+}{1-g}\right] \\ &= \frac{\alpha}{1+\delta}\left[\left(1+\delta q_+ e^{-\alpha(1+\delta)q_+}\right)\log_2\left(1+\delta q_+ e^{-\alpha(1+\delta)q_+}\right) + \delta\left(1-q_+e^{-\alpha(1+\delta)q_+}\right)\log_2\left(1-q_+e^{-\alpha(1+\delta)q_+}\right)\right] \end{aligned} \tag{27}$$

The optimal information $i_{SP} = 0.083$ bits/synapse is reached for $q_+ = 1$, $\theta = 0.72$, $\beta = 2.44$, $\alpha = 0.14$, $\delta = 2.57$ which gives $g = 0.28$, $g_+ = 0.72$.

The dependence of i_{opt} on the different parameters is shown in figure 2. Panel *a* shows the dependence on g the fraction of activated synapses in the asymptotic learning regime. Panels b, c and d show the dependence on δ, β and q_+. Note from panel *c* that there is a broad

range of values of β that give information capacities similar to the optimal one. One can also observe that the optimal information capacity is about 9 times lower in the SP model than in the Willshaw model. This is the price one pays to have a network that is able to continuously learn new patterns. However, it should be noted that at maximal capacity, in the Willshaw model, every pattern has a vanishing basin of attraction while in the SP model, only the oldest stable patterns have vanishing basins of attraction. This feature is not captured by our measure of storage capacity.

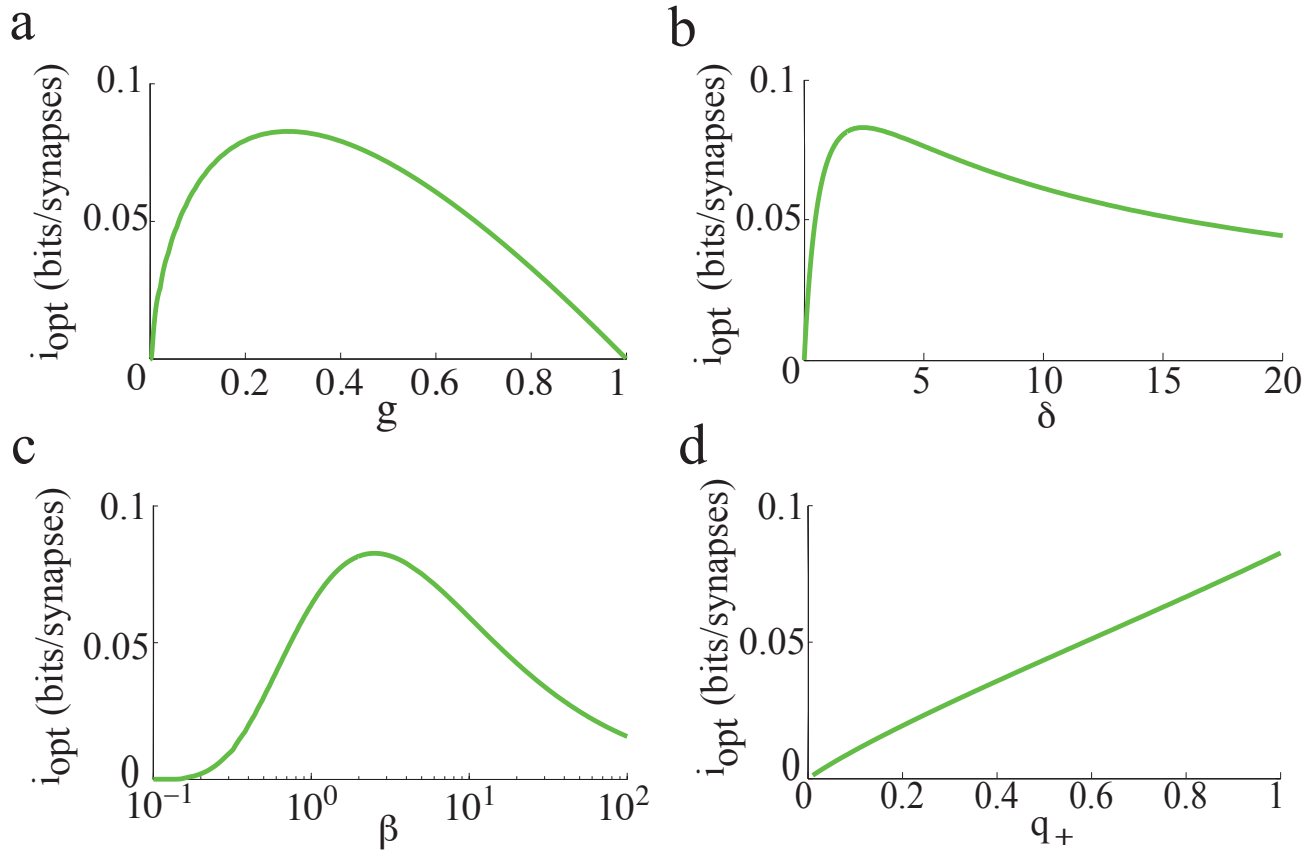

Figure 2: *Optimized information capacity for the SP model in the limit $N \to +\infty$. a. i_{opt} as a function of g, b. i_{opt} as a function of δ, the ratio between the number of depressing events and potentiating events at pattern presentation, c. i_{opt} as a function of $\beta = f\frac{N}{\ln N}$, d. i_{opt} as a function of the LTP transition probability q_+.*

4 Multiple presentations of patterns, slow learning regime

In the SP model, patterns are presented only once. Brunel et al [23] studied the same network of binary neurons with stochastic binary synapses but in a different learning context, where patterns are presented multiple times. More precisely, at each learning time step t, a noisy version $\vec{\xi}^{\mu(t),t}$ of one of the P prototypes $\vec{\xi}^{\mu}$ is presented to the network,

$$\begin{cases} \mathbb{P}(\xi_i^{\mu(t),t}=1)=1-(1-f)x \;\text{ and }\; \mathbb{P}(\xi_i^{\mu(t),t}=0)=(1-f)x \;\text{ for }\; \xi_i^{\mu(t)}=1 \\ \mathbb{P}(\xi_i^{\mu(t),t}=1)=fx \;\text{ and }\; \mathbb{P}(\xi_i^{\mu(t),t}=0)=1-fx \;\text{ for }\; \xi_i^{\mu(t)}=0 \end{cases} \tag{28}$$

Here x is a noise level: if $x=0$, presented patterns are identical to the prototypes, while if $x=1$, the presented patterns are uncorrelated with the prototypes. As for the SP model this model achieves a finite non-zero information capacity i_{opt} in the large N limit if the depression-potentiation ratio δ is of order one, and if the coding level scales with network size as $f \propto \frac{\ln N}{N}$. If learning is slow, $q_+, q_- \ll 1$, and the number of presentations of patterns of each class become large the probabilities g and g_+ are [23]:

$$g=\sum_{\Pi=0}^{+\infty}\frac{(1-x)^2\Pi+\alpha x(2-x)}{(1-x)^2\Pi+\alpha(\delta+x(2-x))}\frac{\alpha^\Pi\exp(-\alpha)}{\Pi\,!} \tag{29}$$

and

$$g_+=\sum_{\Pi=0}^{+\infty}\frac{(1-x)^2(\Pi+1)+\alpha x(2-x)}{(1-x)^2(\Pi+1)+\alpha(\delta+x(2-x))}\frac{\alpha^\Pi\exp(-\alpha)}{\Pi\,!} \tag{30}$$

We inserted those expressions in Eqs. (19,20) to study the maximal information capacity of the network under this learning protocol. The optimal information $i_{MP}=0.69$ bits/synapse is reached at $x=0$ for $\theta\to 1\,,\ \beta\to 1.44\,,\ \delta\to 0\,,\ \alpha\to 0.69$ which gives $g\to\frac{1}{2}\,,\ g_+\to 1$. In this limit, the network becomes equivalent to the Willshaw model.

The maximal capacity is about 9 times larger than for a network that has to learn in one shot. On figure 3a we plot the optimal capacity as a function of g. The capacity of the slow learning network with multiple presentations is bounded by the capacity of the Willshaw model for all values of g, and it is reached when the depression-potentiation ratio $\delta\to 0$. For this value, no depression occurs during learning: the network loses palimpsest properties, i.e. the ability to erase older patterns to store new ones, and it is not able to learn if the presented patterns are noisy. The optimal capacity decreases with δ, for instance at $\delta=1$ (as many potentiation events as depression events at each pattern presentation), $i_{opt}=0.35$ bits/synapse. Figure 3c shows the dependence as a function of $\beta=f\frac{N}{\ln N}$. In figure 3d, we show the optimized capacity for different values of the noise x in the presented patterns. This quantifies the trade-off between the storage capacity and the generalization ability of the network [23].

5 Finite-size networks

The results we have presented so far are valid for infinite size networks. Finite-size effects can be computed for the three models we have discussed so far (see Methods B). The main result of this section is that the capacity of networks of realistic sizes is very far from the large N limit. We compute capacities for finite networks in the SP and MP settings, and we validate our finite size calculations by presenting the results of simulation of large networks of sizes $N=10,000$, $N=50,000$.

We summarize the finite size calculations for the SP model (a more general and detailed analysis is given in Methods B). In the finite network setting, conditional on the tested pat-

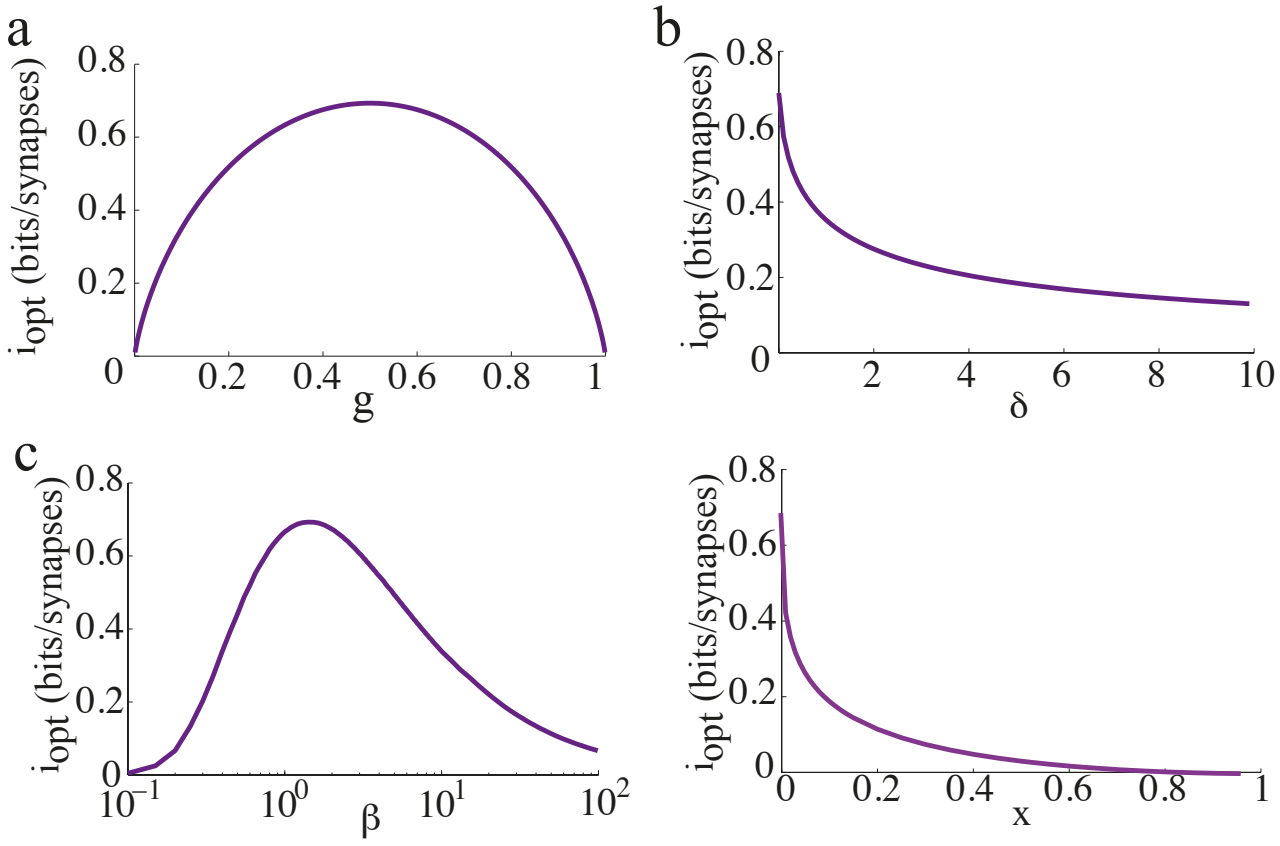

Figure 3: *Optimized information capacity for the MP model in the limit $N \to +\infty$. a. Optimal information capacity as a function of g, the average number of activated synapses after learning. Optimal capacity is reached in the limit $\delta \to 0$ and at $x = 0$ where the capacity is the same as for the Willshaw model. b. Dependence of information capacity on δ, the ratio between the number of depressing events and potentiating events at pattern presentation. c. Dependence on $\beta = f\frac{N}{\ln N}$. d. Dependence on the noise in the presented patterns, x. This illustrates the trade-off between the storage capacity and the generalization ability of the network.*

tern μ_0 having $M+1$ selective neurons, the probability of no error is $\mathbb{P}_{ne}$ is given by

$$\mathbb{P}_{ne} = \exp\left[-\exp\left(X_s\right) - \exp\left(X_n\right)\right]$$

with

$$\begin{aligned} X_s &= -\beta_M \Phi\left(g_+, \theta_M\right) \ln N + \frac{1}{2}\ln\ln N - \frac{1}{2}\ln\left[\frac{\left(1-\exp(\frac{\partial\Phi}{\partial\theta}(g_+,\theta_M))\right)^2 2\pi\theta_M(1-\theta_M)}{\beta_M}\right] + o(1) \\ X_n &= (-\beta_M\Phi(g,\theta_M)+1)\ln N - \frac{1}{2}\ln\ln N - \frac{1}{2}\ln\left[\left(1-\exp(-\frac{\partial\Phi}{\partial\theta}(g,\theta_M))\right)^2 2\pi\theta_M(1-\theta_M)\beta_M\right] + o(1) \end{aligned} \tag{31}$$

where $\beta_M = \frac{M}{\ln N}, \theta_M = \theta\frac{fN}{M}$ and Φ is given by Eq. (13). In the calculations for $N \to +\infty$ discussed in Sections 1-4 we kept only the dominant term in $\ln N$, which yields equations (19) and (20).

In the above equations, the first order corrections scale as $\frac{\ln\ln N}{\ln N}$, which has a dramatic effect on the storage capacity of finite networks. In figure 4a,b, we plot $\overline{\mathbb{P}}_{ne}$ (where the bar denotes an average over the distribution of M) as a function of the age of the pattern, and compare this with numerical simulations. It is plotted for $N = 10,000$ and $N = 50,000$ for learning and network parameters chosen to optimize the storage capacity of the infinite-size network (see Section 3). We show the result for two different approximations of the field distribution: a binomial distribution (magenta), as used in the previous calculations for infinite size networks ; and a gaussian (red) approximation (see Methods C for calculations) as used by previous authors [19, 20, 24]. For these parameters the binomial approximation gives an accurate estimation of $\overline{\mathbb{P}}_{ne}$, while the gaussian calculation overestimates it.

The curves we get are far from the step functions predicted for $N \to +\infty$ by Eq. (45). To understand why, compare equations (15), and (31): finite size effects can be neglected when $|(-\beta\Phi(g_+,\theta))| \gg \frac{\ln\ln N}{\ln N}$ and $|(-\beta\Phi(g,\theta)+1)| \gg \frac{\ln\ln N}{\ln N}$. Because the finite size effects are of order $\frac{\ln\ln N}{\ln N}$, it is only for huge values of N that the asymptotic capacity can be recovered. For instance if we choose an activation threshold θ slightly above the optimal threshold given in Section 3 ($\theta = \theta_{opt} + 0.01 = 0.73$), then $-\beta\Phi(g,\theta)+1 = -0.06$, and for $N = 10^{100}$ we only have $|-\beta\Phi(g,\theta)+1| \simeq 3\frac{\ln\ln N}{\ln N}$. In figure 4c we plot $\mathbb{P}_{ne}$ as a function of $\frac{\alpha}{\alpha_{opt}}$ where $\alpha_{opt} = 0.14$ is the value of α that optimizes capacity in the large N limit, $\theta = 0.73$ and the other parameters are the one that optimizes capacity. We see that we are still far from the large N limit for $N = 10^{100}$. Networks of sizes $10^4 - 10^6$ have capacities which are only between 20% and 40% of the predicted capacity in the large N limit. Neglecting fluctuations in the number of selective neurons, we can derive an expression for the number of stored patterns P that includes the leading finite size correction for the SP model,

$$P(N) = c_1 \frac{N^2}{(\ln N)^2}\left[1 - c_2\sqrt{\frac{\ln\ln N}{\ln N}} + o\left(\sqrt{\frac{\ln\ln N}{\ln N}}\right)\right] \tag{32}$$

where c_1 and c_2 are two constants (see Methods B).

If we take fluctuations in the number of selective neurons into account, it introduces other finite-size effects as can be seen from equations (43) and (44) in the Methods section. These fluctuations can be discarded if $|(-\beta\Phi(g_+,\theta))| \gg \frac{\sqrt{\beta}}{\sqrt{\ln N}}\frac{1-\theta}{1-g_+}$ and $|(1-\beta\Phi(g,\theta))| \gg \frac{\sqrt{\beta}}{\sqrt{\ln N}}\frac{1-\theta}{1-g}$. In figure 4d we plot $\overline{\mathbb{P}}_{ne}$ for different values of N. We see that finite size effects are even stronger in this case.

To plot the curves of figure 4, we chose parameters to be those that optimize storage capacity for infinite network sizes. When N is finite, those parameters are no longer optimal. To optimize parameters at finite N, since the probability of error as a function of age is no longer a step function, it is not possible to find the last pattern stored with probability one. Instead we define the capacity P_c as the pattern age for which $\overline{\mathbb{P}}_{ne} = \frac{1}{2}$. Using equations (31) and performing an average over the distribution of M, we find parameters optimizing pattern capacity for fixed values of β. Results are shown on figure 5a,b for $N = 10,000$ and $N = 50,000$. We show the results for the different approximations used to model the neural fields: the blue line is the binomial approximation, the cyan line the gaussian approximation and the magenta one is a gaussian approximation with a covariance term that takes into account correlations between synapses (see Methods C and [19, 20]). For $f < \frac{1}{\sqrt{N}}$ the storage capacity of simulated networks (black crosses) is well predicted by the binomial approximation while the gaussian approximations over-estimates capacity. For $f > \frac{1}{\sqrt{N}}$, the

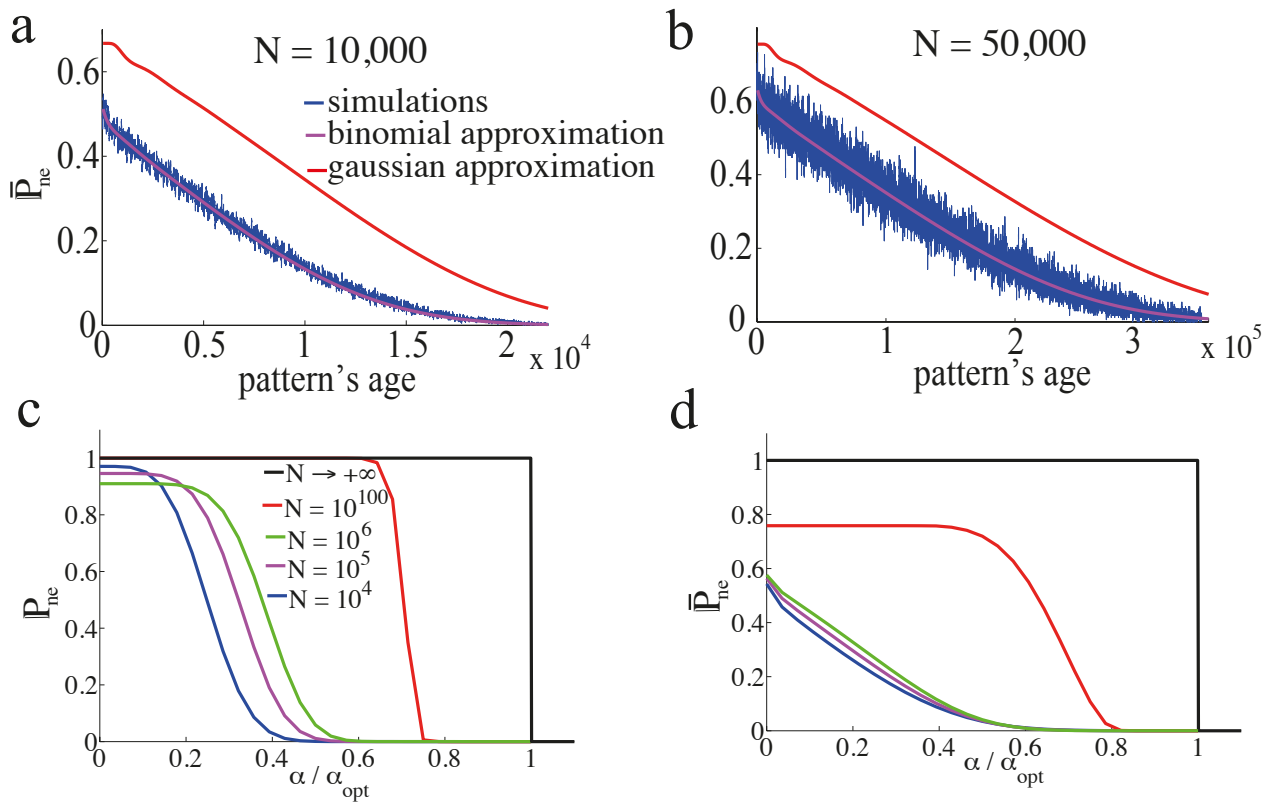

Figure 4: *Finite size effects. Shown is $\mathbb{P}_{ne}$, the probability that a tested pattern of a given age is stored without errors, for the SP model. a. $\mathbb{P}_{ne}$ as a function of the age of the tested pattern. Parameters are those optimizing capacity at $N \to +\infty$ (see Section 3), results are for simulations (blue line) and calculations with a binomial approximation of the fields distributions (magenta) and a gaussian approximation (red) ; $\mathbb{P}_{ne}$ is averaged over different value of M, the number of selective neurons in the tested pattern (magenta line). b Same for $N = 50,000$. c. $\mathbb{P}_{ne}$ as a function of a scaled version of pattern age (see text for details), fluctuations in M are discarded on this plot. d. Same as c with an average of $\mathbb{P}_{ne}$ over different M.*

correlations between synapses can no longer be neglected [17]. The gaussian approximation with covariance captures the drop in capacity at large f.

For $N = 10,000$, the SP model can store a maximum of $P_c = 7,800$ patterns at a coding level $f = 0.0015$ (see blue curve in figure 5c). As suggested in figures 4c,d, the capacity of finite networks is strongly reduced compare to the capacity predicted for infinite size networks. More precisely, if the network of size $N = 10,000$ had the same information capacity as the infinite size network (27), it would store up to $P = 70,000$ patterns at coding level $f = 0.0007$. Part of this decrease in capacity is avoided if we consider patterns that have a fixed number fN of selective neurons. This corresponds to the red curve in figure 4c. For fixed sizes the capacity is approximately twice as large. Note that finite-size effects tend to decrease as the coding level increases. In figure 5c, $f = 5.10^{-4}$, and the capacity is 3% of the value predicted by the large N limit calculation. The ratio of actual to asymptotic capacities increases to 10% at $f = 1.10^{-3}$ and 21% at $f = 1.10^{-2}$ In figure 5d, we do the same analysis

for the MP model with $N = 10,000$. Here we have also optimized all the parameters, except for the depression-potentiation ratio which is set to $\delta = 1$, ensuring that the network has the palimpsest property and the ability to deal with noisy patterns. For $N = 10,000$, the MP model with $\delta = 1$ can store up to $P_c = 70,000$ patterns, at $f = 0.001$ (versus $P_c = 7,800$ at $f = 0.0015$ for the SP model). One can also compute the optimized capacity for a given noise level. At $x = 0.1$, $P_c = 20,900$ for $f = 0.0012$ and $\delta = 4.3$ or at $x = 0.2$, $P_c = 8,900$ for $f = 0.0018$ and $\delta = 6.9$.

6 Storage capacity with errors

So far, we have defined the storage capacity as the number of patterns that can be perfectly retrieved. However, it is quite common for attractor neural networks to have stable fixed point attractors that are close to, but not exactly equal to, patterns that are stored in the connectivity matrix. It is difficult to estimate analytically the stability of patterns that are retrieved with errors as it requires analysis of the dynamics at multiple time steps. We therefore used numerical simulations to check whether a tested pattern is retrieved as a fixed point of the dynamics at a sufficiently low error level. To quantify the degree of error, we introduce the overlap $m\left(\vec{\sigma}^*, \vec{\xi}^{\mu_0}\right)$ between the network fixed point $\vec{\sigma}^*$ and the tested pattern $\vec{\xi}^{\mu_0}$, with M selective neurons

$$m\left(\vec{\sigma}^*, \vec{\xi}^{\mu_0}\right) = \frac{1}{M(1-f)} \sum_{i=1}^{N} \left(\xi_i^{\mu_0} - f\right) \sigma_i^* \tag{33}$$

In figure 6a we show $P_c(m)$, the number of fixed-point attractors that have an overlap larger than m with the corresponding stored pattern, for $m = 1$, $m = 0.99$ and $m = 0.7$. Note that only a negligible number of tested patterns lead to fixed points with m smaller than 0.7, for $N = 10,000$ neurons. Considering fixed points with errors leads to a substantial increase in capacity, e.g. for $f = 0.0018$ the capacity increases from $P_c(m = 1) = 7,800$ to $P_c(m = 0.7) = 10,400$. In figure 6b, we quantify the information capacity in bits stored per synapse, defined as in Eq. (6), $i = P_c\left(-f \log_2 f - (1-f)\log_2(1-f)\right)/N$. Note that in the situation when retrieval is not always perfect this expression is only an approximation of the true information content. The coding level that optimizes the information capacity in bits per synapse i is larger ($f_{opt} \simeq 0.003$) than the one that optimizes the number of stored patterns P_c ($f_{opt} \simeq 0.002$), since the information content of individual patterns decreases with f. Finally, note that the information capacity is close to its optimum in a broad range of coding levels, up to $f \sim 0.01$.

7 Increase in capacity with inhibition

As we have seen above, the fluctuations in the number of selective neurons in each pattern lead to a reduction in storage capacity in networks of finite size (e.g. figure 5c,d). The detrimental effects of these fluctuations can be mitigated by adding a uniform inhibition η to the network [19]. Using a simple instantaneous and linear inhibitory feed-back, the local

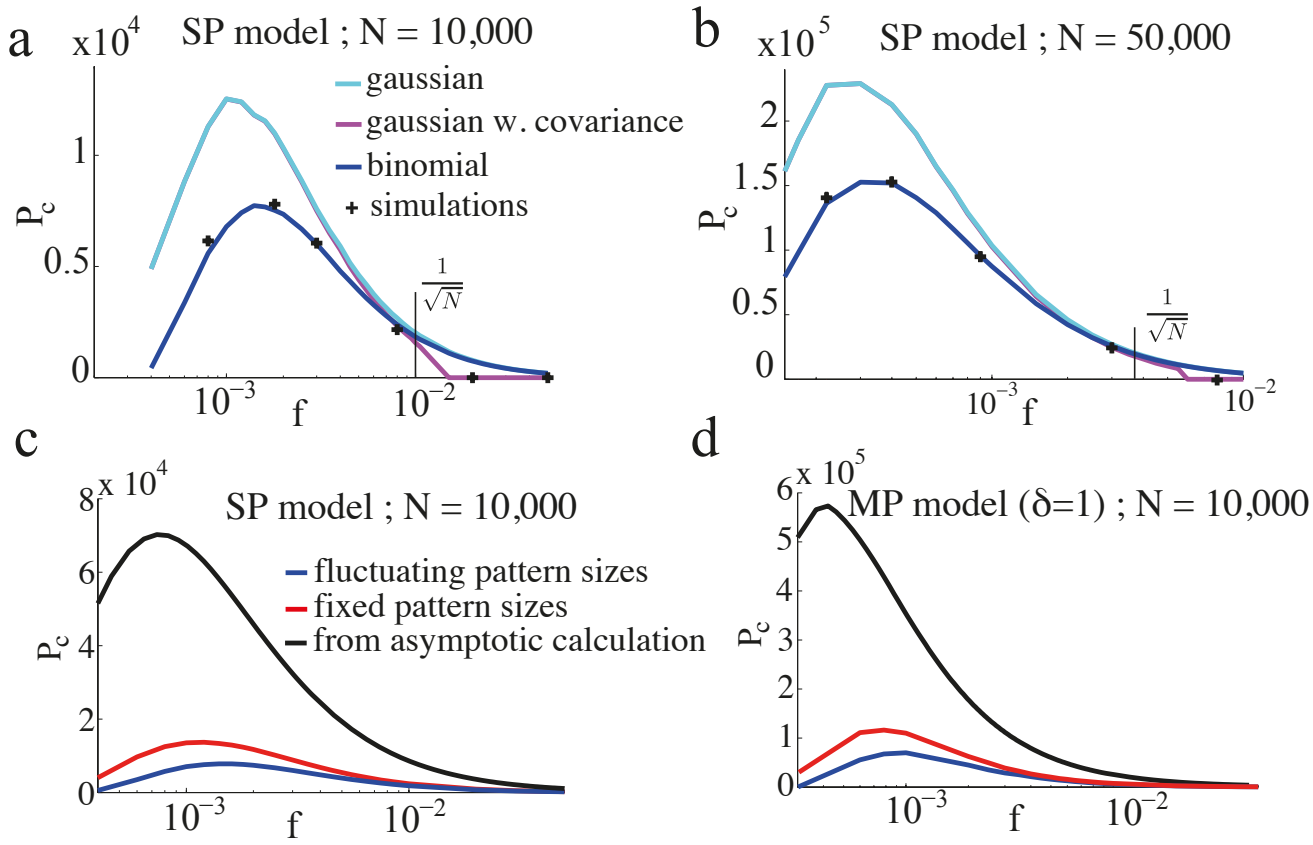

Figure 5: *Capacity at finite N. a,b.* P_c *as a function of* f *for the SP model and* $N = 10^4, 5.10^4$ *Parameters are chosen to optimize capacity under the binomial approximation. Shown are the result of the gaussian approximation without covariance (cyan) and with covariance (magenta) for these parameters. c. Optimized* P_c *as a function of* f *for the SP model at* $N = 10,000$*. The blue curve is for patterns with fluctuations in the number of selective neurons. The red curve is for the same number of selective neurons in all patterns. The black curve is the number of patterns that would be stored if the network were storing the same amount of information as in the case* $N \to +\infty$*. d. Same for the MP model, where parameters have been optimized, but the depression-potentiation ratio is fixed at* $\delta = 1$*.*

fields become

$$h_i = \sum_{k=1}^{N} W_{ik}\xi_k^{\mu_0} - \eta \sum_{k=1}^{N} \xi_k^{\mu_0} \tag{34}$$

For infinite size networks, adding inhibition does not improve storage capacity since fluctuations in the number of selective neurons vanish in the large N limit. However, for finite size networks, minimizing those fluctuations leads to substantial increase in storage capacity. When testing the stability of pattern $\vec{\xi}^1$, if the number of selective neurons is unknown, the variance of the field on non-selective neurons is $Nf(g-2\eta g+\eta^2)$, and $Nf(g_+-2\eta g_++\eta^2)$ for selective neurons (for small f). The variance for non-selective neurons is minimized if $\eta = g$, yielding the variance obtained with fixed sized patterns. The same holds for selective neurons at $\eta = g_+$. Choosing a value of η between g and g_+ brings the network capacity

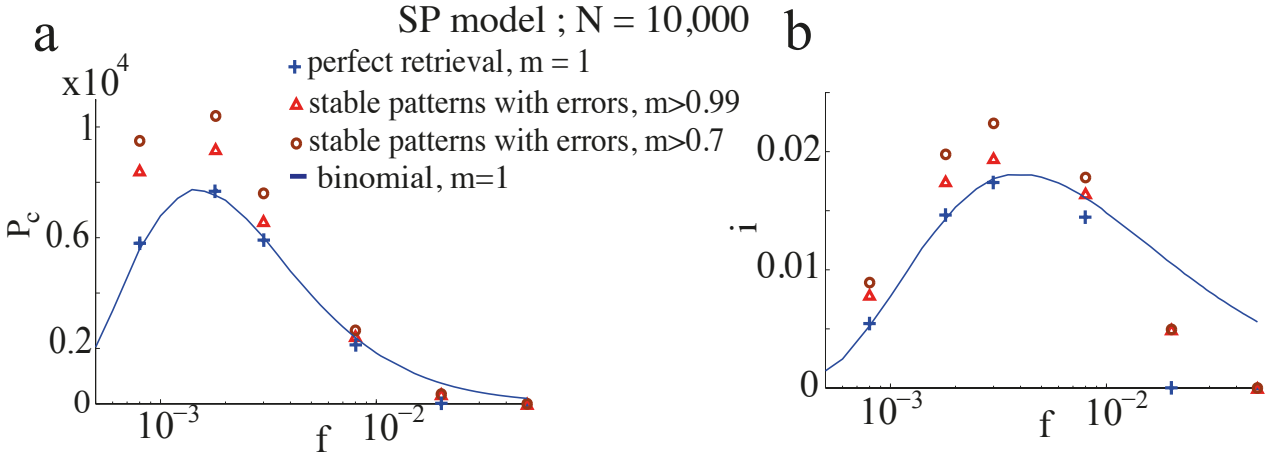

Figure 6: *Storage capacity with errors in the SP model. Instead of counting only patterns that are perfectly retrieved, patterns that lead to fixed points of the dynamics that overlap significantly with the tested memory pattern are also taken into account. Simulations are done with the same parameters as in figure 5a. a. Number of stored patterns P_c as a function of f. Blue crosses correspond to fixed points that are exactly the stored patterns. Red triangles correspond to fixed points that have an overlap larger than 0.99, and brown circles an overlap larger than 0.7. b. Information stored per synapse as a function of f.*

towards that of fixed size patterns. In figure 7a, we show the storage capacity as a function of f for these three scenarios. Optimizing the inhibition η increases the maximal capacity by 28% (green curve) compared to a network with no inhibition (blue curve). Red curve is the capacity without pattern size fluctuations. Inhibition increases the capacity from $P_c = 7,800$ at $f = 0.0018$ to $P_c = 12,000$. In figure 7b, information capacity measured in bits per synapse is shown as a function of f in the same three scenarios. Note again that for $f = \frac{1}{\sqrt{N}} = 0.01$, the capacity is quite close to the optimal capacity.

Discussion

We have presented an analytical method to compute the storage capacity of networks of binary neurons with binary synapses in the sparse coding limit. When applied to the classic Willshaw model, in the infinite limit, we find a maximal storage capacity of $\ln 2 = 0.69$ bits/synapse, the same than found in previous studies, although with a different definition adapted to recurrent networks, as discussed in the section 'Willshaw model'. We then used this method to study the storage capacity of a network with binary synapses and stochastic learning, in the single presentation (SP) scenario [17]. The main advantage of this model, compared to the Willshaw model, is its palimpsest property, that allows it to do on-line learning in an ever changing environment. Amit and Fusi showed that the optimal storage capacity was obtained in the sparse coding limit, $f \propto \frac{\ln N}{N}$ and with a balance between the effect of depression and potentiation. The storage capacity of this network has been further studied for finite size networks in [19, 20]. We have complemented this work by computing

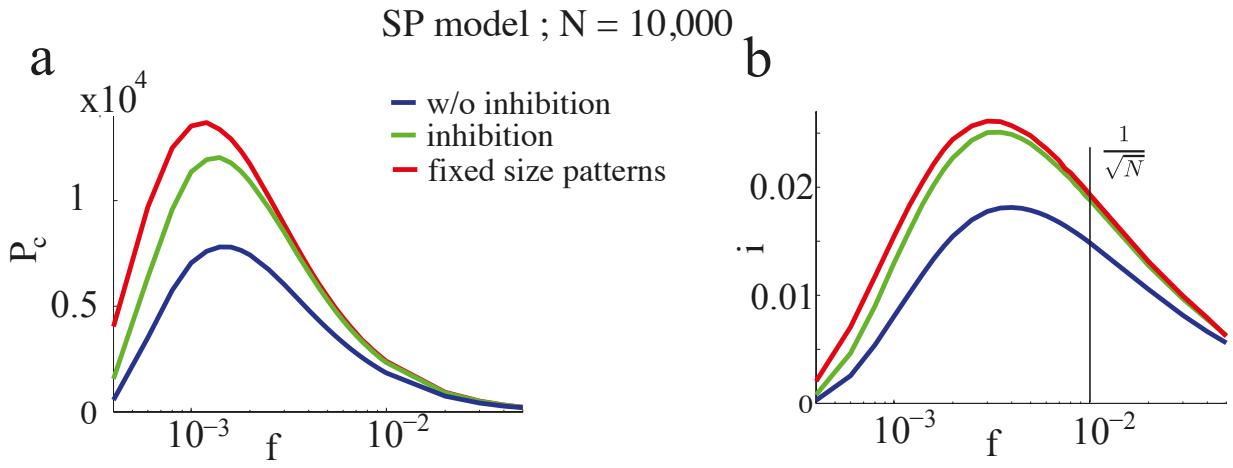

Figure 7: *Storage capacity optimized with inhibition in the SP model. Blue is for a fixed threshold and fluctuations in the number of selective neurons per pattern. Green, the fluctuations are minimized using inhibition. Red, without fluctuations in the number of selective neurons per pattern. a. Number of stored patterns as a function of the coding level f. b. Stored information in bits per synapse, as a function of f.*

analytically the storage capacity in the large N limit. The optimal capacity of the SP model is 0.083 bits/synapse, which is about 9 times lower than the one of the Willshaw model. This decrease in storage capacity is similar to the decrease seen in palimpsest networks with continuous synapses - for example, in the Hopfield model the capacity is about 0.14 bits/synapse, while in a palimpsest version the capacity drops to about 0.05 bits/synapse. The reason for this decrease is that the most recently seen patterns have large basins of attraction, while older patterns have smaller ones. In the Willshaw model, all patterns are equivalent, and therefore they all have vanishing basins of attraction at the maximal capacity.

We have also studied the network in a multiple presentation (MP) scenario, with in which patterns presented to the network are noisy versions of a fixed set of prototypes, in the slow learning limit in which transition probabilities go to zero [23]. In the extreme case in which presented patterns are the prototypes, all synaptic weights are initially at zero, and if the synapses do not experience depression, this model is equivalent to the Willshaw model with a storage capacity of 0.69 bits/synapse, which is about 9 times larger than the capacity of the SP model. A more interesting scenario is when depression is present. In this case then the network has generalization properties (it can learn prototypes from noisy versions of them), as well as palimpsest properties (if patterns drawn from a new set of prototypes are presented it will eventually replace a previous set with the new one). We have quantified the trade-off between generalization and storage capacity (see figure 3d). For instance, if the noisy patterns have 80% of their selective neurons in common with the prototypes to be learned, the storage capacity is decreased from 0.69 to 0.12bits/synapses.

A key step in estimating storage capacity is deriving an accurate approximation for the distribution of the inputs neurons receive. These inputs are the sum of a large number of binary variables, so the distribution is a binomial if one can neglect the correlations between these variables, induced by the learning process. Amit and Fusi [17] showed that

these correlations can be neglected when $f \ll 1/\sqrt{N}$. Thus, we expect the results with the binomial approximation to be exact in the large N limit. We have shown that a Gaussian approximation of the binomial distribution gives inaccurate results in the sparse coding limit, because the capacity depends on the tail of the distribution, which is not well described by a Gaussian. For larger coding levels ($f \sim 1/\sqrt{N}$), the binomial approximation breaks down because it does not take into account correlations between inputs. Following [19] and [20], we use a Gaussian approximation that includes the covariance of the inputs, and show that this approximation captures well the simulation results in this coding level range.

We computed storage capacities for two different learning scenarios. Both are unsupervised, involve a Hebbian-type plasticity rule, and allow for online learning (providing patterns are presented multiple times for the MP model). It is of interest to compare the performance of these two particular scenarios with known upper bounds on storage capacity. For networks of infinite size with binary synapses such a bound has been derived using the Gardner approach [25]. In the sparse coding limit, this bound is $\simeq 0.29$ bits/synapse with random patterns (in which fluctuations in the number of selective neurons per pattern fluctuates), and $\simeq 0.45$ bits/synapse if patterns have a fixed number of selective neurons [26]. We found a capacity of $i_{SP} = 0.083$bits/synapse for the SP model and $i_{MP} = 0.69$ bits/synapse for the MP model, obtained both for patterns with fixed and variable number of selective neurons. The result for the MP model seems to violate the Gardner bound. However, as noticed by Nadal [21], one should be cautious in comparing these results: in our calculations we have required that a given pattern is stored perfectly with probability one, while the Gardner calculation requires that **all** patterns are stored perfectly with probability one. As mentioned in the section 2, the capacity of the Willshaw and MP models drops to $i_{opt} = 0.23$bits/synapse in the case of fixed-size patterns, if one insists that **all** patterns should be stored perfectly, which is now consistent with the Gardner bound. This means that the MP model is able to reach a capacity which is roughly half the Gardner bound, a rather impressive feat given the simplicity of the rule. Note that supervised learning rules can get closer to these theoretical bounds [27].

We have also studied finite-size networks, in which we defined the capacity as the number of patterns for which the probability of exact retrieval is at least 50%. We found that networks of reasonable sizes have capacities that are far from the large N limit. For networks of sizes $10^4 - 10^6$ storage capacities are reduced by a factor 3 or more (see Fig. 4). These huge finite size effects can be understood by the fact that the leading order corrections in the large N limit are in $\frac{\ln(\ln N)}{\ln N}$ - and so can never be neglected unless N is an astronomical number (see Methods A). A large part of the decrease in capacity when considering finite-size networks is due to fluctuations in the number of selective neurons from pattern to pattern. In the last section, we have used inhibition to minimize the effect of these fluctuations. For instance, for a network of $N = 10,000$ neurons learning in one shot, inhibition allows to increase capacity from $P = 7,800$ to $P = 12,000$. For finite size networks, memory patterns that are not perfectly retrieved can still lead to fixed points where the activity is significantly correlated with the memory patterns. We have investigated with simulations how allowing errors in the retrieved patterns modifies storage capacity. For $N = 10,000$, the capacity increases from $P = 7,800$ to $P = 10,400$, i.e. by approximately 30%.

Our study focused on networks of binary neurons, connected through binary synapses, and storing very sparse patterns. These three assumptions allowed us to compute analytically the storage capacity of the network in two learning scenarios. An important ques-

tion is how far real cortical networks are from such idealized assumptions. First, the issue of whether real synapses are binary, discrete but with a larger number of states, or essentially continuous, is still unresolved, with evidence in favor of each of these scenarios [28, 29, 30, 31, 32]. We expect that having synapses with a finite number $K > 2$ of states will not modify strongly the picture outlined here [17, 33, 20]. Second, it remains to be investigated how these results will generalize to networks of more realistic neurons. In strongly connected networks of spiking neurons operating in the balanced mode [34, 35, 36, 37], the presence of ongoing activity presents strong constraints on the viability of sparsely coded selective attractor states. This is because 'non-selective' neurons are no longer silent, but are rather active at low background rates, and the noise due to this background activity can easily wipe out the selective signal [35, 38]. In fact, simple scaling arguments in balanced networks suggest the optimal coding level would become $f \sim 1/\sqrt{N}$ [3, 39]. The learning rules we have considered in this paper lead to a vanishing information stored per synapse with this scaling. Finding an unsupervised learning rule that achieves a finite information capacity in the large N limit in networks with discrete synapses for such coding levels remains an open question. However, the results presented here show that for networks of realistic sizes, the information capacity at such coding levels is in fact not very far from the optimal one that is reached at lower coding levels (see vertical lines in Fig. 5-7). Finally, the coding levels of cortical networks during delay period activity remain poorly characterized. Experiments in IT cortex [40, 41, 42] are consistent with coding levels of order 1%. Our results indicate that in networks of reasonable sizes, these coding levels are not far from the optimal values.

The SP and MP models investigated in this paper can be thought of as minimal models for learning in hippocampus and neocortex. The SP model bears some resemblance to the function of hippocampus, which is supposed to keep a memory of recent episodes that are learned in one shot, thanks to highly plastic synapses. The MP model relates to the function of neocortex, where a longer-term memory can be stored, thanks to repeated presentations of a set of prototypes that occur repeatedly in the environment, and perhaps during sleep under the supervision of the hippocampus. The idea that hippocampal and cortical networks learn on different time scales has been exploited in several modeling studies [43, 44, 45], in which the memories are first stored in the hippocampus and then gradually transferred to cortical networks. It would be interesting to extend the type of analysis presented here to coupled hippocampo-cortical networks with varying degrees of plasticity.

Methods

A - Capacity calculation for infinite size networks

We are interested at retrieving pattern $\vec{\xi}^{\mu}$ that has been presented during the learning phase. We set the network in this state $\vec{\sigma} = \vec{\xi}^{\mu}$ and ask whether the network remains in this state while the dynamics (2) is running. At the first iteration, each neuron i is receiving a field

$$h_i = \sum_{j=1}^{N} W_{ij}\xi_j^{\mu} = \sum_{k=1}^{M} X_k^i \tag{35}$$

Where M+1 is the number of selective neurons in pattern $\vec{\xi}^\mu$, with $M = O(\ln N)$ [1] and $N \to +\infty$. We recall that $g_+ = \mathbb{P}(W_{ij} = 1 | \xi_i^\mu = \xi_j^\mu = 1)$ and $g = \mathbb{P}(W_{ij} = 1 | (\xi_i^\mu, \xi_j^\mu) \neq (1,1))$. Thus X_k^i is a binary random variable which is 1 with probability, either g_+ if i is a selective neuron (sites i such that $\xi_i^\mu = 1$), or g if i is a non-selective neuron (sites i such that $\xi_i^\mu = 0$). Neglecting correlations between W_{ij_1} and W_{ij_2} (it is legitimate in the sparse coding limit we are interested in, see [17]), the X_k^i's are independent and the distribution of the field on selective neurons can be written as

$$\begin{aligned} \mathbb{P}(h_i^s = S) &= \binom{M}{S} g_+^S (1-g_+)^{M-S} \\ &= \exp\left[-M\Phi\left(g_+, \frac{S}{M}\right) - \frac{1}{2}\ln\left(S\left(1-\frac{S}{M}\right)\right) - \frac{1}{2}\ln(2\pi) \right] \end{aligned} \tag{36}$$

where we used Stirling formula for $M, S \gg 1$, with Φ defined in (13). For non-selective neurons

$$\begin{aligned} \mathbb{P}\left(h_i^n = S\right) &= \binom{M}{S} g^S (1-g)^{M-S} \\ &= \exp\left[-M\Phi\left(g, \frac{S}{M}\right) - \frac{1}{2}\ln\left(S\left(1-\frac{S}{M}\right)\right) - \frac{1}{2}\ln(2\pi) \right] \end{aligned} \tag{37}$$

Now write

$$\begin{aligned} \mathbb{P}(h_i^s \leq \theta f N) &= \mathbb{P}(h_i^s = \theta f N) \sum_{S \leq \theta f N} \frac{\mathbb{P}(h_i^s = S)}{\mathbb{P}(h_i^s = \theta f N)} \\ \mathbb{P}(h_i^n \geq \theta f N) &= \mathbb{P}(h_i^n = \theta f N) \sum_{S \geq \theta f N} \frac{\mathbb{P}(h_i^n = S)}{\mathbb{P}(h_i^n = \theta f N)} \end{aligned} \tag{38}$$

In the limit $N \to +\infty$ we are considering in this section, and if $Mg < fN\theta < Mg_+$, the sums corresponding to the probabilities $\mathbb{P}(h_i^s \leq fN\theta), \mathbb{P}(h_i^n \geq fN\theta)$ are dominated by their first term (corrections are made explicit in the following section). Keeping only higher order terms in M in equations (36) and (37), we have:

$$\mathbb{P}(h_i^s \leq fN\theta) \simeq \exp(-M\Phi(g_+, \theta_M)) \tag{39}$$

and

$$\mathbb{P}(h_i^n \geq fN\theta) \simeq \exp(-M\Phi(g, \theta_M)), \tag{40}$$

yielding equation (15) with $\theta_M = \theta \frac{fN}{M} = O(1)$. Note that with the coding levels we are considering here ($f \propto \frac{\ln N}{N}$), M is of order $\ln N$. When the number of selective neurons per pattern is fixed at fN, we choose $M\theta$ for the activation threshold and these equations become:

$$\begin{aligned} X_s &= -\ln N \beta \Phi(g_+, \theta) + O(\ln\ln N) \\ X_n &= \ln N(-\beta\Phi(g, \theta) + 1) + O(\ln\ln N) \end{aligned} \tag{41}$$

[1] We use the standard 'Landau' notations, $a = O(F(N))$ means that $a/F(N)$ goes to a finite limit in the large N limit, while $a = o(F(N))$ means that $a/F(N)$ goes to zero in the large N limit.

where $\beta = f\frac{N}{\ln N}$

For random numbers of selective neurons we need to compute the average over M: $\overline{\mathbb{P}}_{ne}(N) = \sum_{M=0}^{N} \mathbb{P}(M)\mathbb{P}_{ne}(M,N)$. Since M is distributed according to a binomial of average Nf and variance $Nf(1-f) \simeq Nf$, for sufficiently large Nf, this can be approximated as $M = fN + z\sqrt{fN}$ where z is normally distributed:

$$\overline{\mathbb{P}}_{ne}(N) = \int_{-\infty}^{+\infty} dz \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} \exp(-\exp(X_s(z,N)) - \exp(X_n(z,N))) \tag{42}$$

with

$$\begin{aligned} X_s(z,N) &= -M\Phi\left(g_+, \frac{\theta}{1+\frac{z}{\sqrt{fN}}}\right) + O(\ln\ln N) \\ &\simeq -\beta \ln N \left[\Phi(g_+,\theta) + \frac{z}{\sqrt{fN}}\left(\Phi(g_+,\theta) - \theta\frac{\partial\Phi}{\partial\theta}(g_+,\theta)\right)\right] + O(\ln\ln N) \\ &\simeq -\beta \ln N \left[\Phi(g_+,\theta) + \frac{z}{\sqrt{fN}} \ln\frac{1-\theta}{1-g_+}\right] + O(\ln\ln N) \end{aligned} \tag{43}$$

and

$$\begin{aligned} X_n(z,N) &= -M\Phi\left(g, \frac{\theta}{1+\frac{z}{\sqrt{fN}}}\right) + \ln N + O(\ln\ln N) \\ &\simeq \ln N \left[1 - \beta\left(\Phi(g,\theta) + \frac{z}{\sqrt{fN}} \ln\frac{1-\theta}{1-g}\right)\right] + O(\ln\ln N) \end{aligned} \tag{44}$$

When N goes to infinity, we bring the limit into the integral in equation (42) and obtain

$$\begin{aligned} \lim_{N\to+\infty} \overline{\mathbb{P}}_{ne}(N) &= \int_{-\infty}^{+\infty} dz \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} \lim_{N\to+\infty} \exp\left[-\exp(X_s(z,N)) - \exp(X_n(z,N))\right] \\ &= \Theta(\Phi(g_+,\theta))\Theta(-\beta\Phi(g,\theta)+1) \end{aligned} \tag{45}$$

where Θ is the Heaviside function. Thus in the limit of infinite size networks, the probability of no error is a step function. The first Heaviside function implies that the only requirement to avoid errors on selective neurons is to have a scaled activation threshold θ below g_+. The second Heaviside function implies that, depending on β, θ has to be chosen far enough from g. The above equation allows to derive the inequalities (19) and (20).

B - Capacity calculation for finite-size networks.

We now turn to a derivation of finite-size corrections for the capacity. Here we show two different calculations. In the first calculation, we derive Eq. (32), taking into account the leading-order correction term in Eq. (43). This allows us to compute the leading-order correction to the number of patterns P that can be stored for a given set of parameters. However, it does not predict accurately the storage capacity of the large-size but finite networks that we simulated. In the second calculation presented, we focus on computing the probability of no error in a given pattern $\mathbb{P}_{ne}$, including a next-to-leading-order correction.

Equation (32) is derived for a fixed set of parameters, assuming that the set of active neurons have a fixed size, and that the activation threshold θ has been chosen large enough such that the probability to have non-selective neurons activated is small. From the Stirling expansion, adding the first finite-size correction term in Eq. (41), we get

$$X_s \simeq -\ln N \beta_M \Phi(g_+, \theta) + \frac{1}{2}\ln\ln N \tag{46}$$

with $\beta_M = M/\ln N$. For large N, the number of stored patterns P can be increased until $g_+(P) \gtrsim \theta$. Setting $g_+ = \theta + \epsilon$, an expansion of Φ in ϵ allows to write

$$X_s \simeq -\ln N \beta_M \frac{\epsilon^2}{2\theta(1-\theta)} + \frac{1}{2}\ln\ln N \tag{47}$$

The P patterns are correctly stored as long as $X_s \ll -1$. This condition is satisfied for $\epsilon < \sqrt{\frac{\theta(1-\theta)}{\beta_M}\frac{\ln\ln N}{\ln N}}$. For the SP model, we can deduce which value of P yields this value of ϵ (see Eq. (26)). This allows to derive Eq. (32),

$$P = \frac{g}{q_+\beta^2}\ln\left(\frac{q_+(1-g)}{\theta-g}\right)\frac{N^2}{(\ln N)^2}\left[1 - \frac{\sqrt{\theta(1-\theta)}}{\sqrt{\beta_M}(\theta-g)\ln\left(\frac{q_+(1-g)}{\theta-g}\right)}\sqrt{\frac{\ln\ln N}{\ln N}} + o\left(\frac{\ln\ln N}{\ln N}\right)\right] \tag{48}$$

We now turn to a calculation of the probability of no error on a given pattern $\mathbb{P}_{ne}$, taking into account the next-to-leading order correction of order one, in addition to the term of order $\ln\ln N$ in Eq. (41). This is necessary to predict accurately the capacity of realistic size networks (for instance for $N = 10,000$, $\ln\ln N \simeq 2 = O(1)$). $\mathbb{P}_{ne}(M)$ is computed for a memory pattern with M selective neurons. The estimation of $\overline{\mathbb{P}}_{ne}$ used in the figures is obtained by averaging over different values of M, with M drawn from a binomial distribution of mean fN.

We first provide a more detailed expansion of the sums in equation (38). Setting $S = fN\theta + k$, with the Taylor expansions:

$$M\Phi\left(g, \theta_M + \frac{k}{M}\right) = M\Phi(g,\theta_M) + k\frac{\partial\Phi}{\partial\theta}(g,\theta_M) + \frac{k^2}{2M}\frac{\partial^2\Phi}{\partial\theta^2}(g,\theta_M) + O\left(\frac{1}{M^2}\right) \tag{49}$$

$$\ln\left(S\left(1-\frac{S}{M}\right)\right) = \ln(M\theta_M(1-\theta_M)) + \frac{k}{M}\Delta\theta_M^{-1} + O\left(\frac{1}{M^2}\right) \tag{50}$$

where $\theta_M = \theta\frac{fN}{M}$ and $\Delta\theta_M^{-1} = \frac{1}{\theta_M} - \frac{1}{1-\theta_M}$. Using (37) we can rewrite:

$$\sum_{S\geq fN\theta}\frac{\mathbb{P}(h_i^n = S)}{\mathbb{P}(h_i^n = fN\theta)} = \sum_{k=0}^{M-fN\theta}\exp\left[-k\frac{\partial\Phi}{\partial\theta}(g,\theta_M) - \frac{1}{M}\left(\frac{k^2}{2}\frac{\partial^2\Phi}{\partial\theta^2}(g,\theta_M) - k\Delta\theta_M^{-1}\right) + O\left(\frac{1}{M^2}\right)\right] \tag{51}$$

In the cases we consider, we will always have $\frac{\partial\Phi}{\partial\theta}(g,\theta_M) \neq 0$ so that we can consider only the term of order 1 in M. The sum is now geometric, and we obtain

$$\sum_{S\geq fN\theta}\frac{\mathbb{P}(h_i^n = S)}{\mathbb{P}(h_i^n = fN\theta)} = \frac{1}{1-\exp\left(-\frac{\partial\Phi}{\partial\theta}(g,\theta_M)\right)} + o(1) \tag{52}$$

The same kind of expansion can be applied for the selective neurons. Again if we are in a situation where $\frac{\partial \Phi}{\partial \theta}(g_+, \theta_M) \neq 0$,

$$\sum_{S \leq fN\theta} \frac{\mathbb{P}(h_i^s = S)}{\mathbb{P}(h_i^s = fN\theta)} = \frac{1}{1 - \exp\left(\frac{\partial \Phi}{\partial \theta}(g_+, \theta_M)\right)} + o(1) \tag{53}$$

When g_+ close to θ and thus $\frac{\partial \Phi}{\partial \theta}(g_+, \theta_M) \simeq 0$, we are then left with:

$$\begin{aligned}
& \sum_{k=0}^{\theta M} \exp\left[-\frac{1}{M}\left(\frac{k^2}{2}\frac{\partial^2 \Phi}{\partial \theta^2}(g_+, \theta_M) - k\Delta\theta_M^{-1}\right)\right] & (54)\\
= \; & \exp\left[\frac{1}{8M}\frac{\partial^2 \Phi}{\partial \theta^2}(g_+, \theta_M)(\Delta\theta_M^{-1})^2\right] \sum_{k=0}^{+\infty} \exp\left[-\frac{(k - \Delta\theta_M^{-1})^2}{2M}\frac{\partial^2 \Phi}{\partial \theta^2}(g_+, \theta_M)\right] + o(1) \\
= \; & \int_0^{+\infty} dt e^{-\frac{(t-(\Delta\theta_M^{-1}))^2}{2M}\frac{\partial^2 \Phi}{\partial \theta^2}(g_+, \theta_M)} + o(1) \\
= \; & \sqrt{\frac{\pi}{2}\frac{M}{\frac{\partial^2 \Phi}{\partial \theta^2}(g_+, \theta_M)}} + o(1) & (55)
\end{aligned}$$

When g_+ is too close to θ, which is the case for the optimal parameters in the large N limit, we need to use (55). It only contributes a term of order $\ln \ln N$ in X_s and does not modify our results. In the figures of Sections 6 and 7, we use (53), which gives from (38) and (36), (37) and (53),(52):

$$\mathbb{P}(h_i^s \leq fN\theta) = \exp\left[\ln N(-\beta_M \Phi(g_+, \theta_M)) - \frac{1}{2}\ln \ln N - \frac{1}{2}\ln\left(2\pi\theta_M(1-\theta_M)[1 - \exp(\frac{\partial \Phi}{\partial \theta}(g_+, \theta_M))]^2\right)\right] \tag{56}$$

$$\mathbb{P}(h_i^n \geq fN\theta) = \exp\left[\ln N(-\beta_M \Phi(g, \theta_M)) - \frac{1}{2}\ln \ln N - \frac{1}{2}\ln\left(2\pi\theta_M(1-\theta_M)[1 - \exp(-\frac{\partial \Phi}{\partial \theta}(g, \theta_M))]^2\right)\right] \tag{57}$$

The probability of no error is

$$\begin{aligned}
\mathbb{P}_{ne} &= (1 - \mathbb{P}(h_i^s \leq fN\theta))^M (1 - \mathbb{P}(h_i^n \geq fN\theta))^{N-M} \\
&= \exp(-\exp X_s - \exp X_n) & (58)
\end{aligned}$$

which leads to equations (31)

$$\begin{aligned}
X_s &= -\beta_M \Phi\left(g_+, \theta_M\right) \ln N + \frac{1}{2}\ln \ln N - \frac{1}{2}\ln\left[\frac{\left(1 - \exp(\frac{\partial \Phi}{\partial \theta}(g_+, \theta_M))\right)^2 2\pi\theta_M(1-\theta_M)}{\beta_M}\right] + o(1) \\
X_n &= (-\beta_M \Phi(g, \theta_M) + 1) \ln N - \frac{1}{2}\ln \ln N - \frac{1}{2}\ln\left[\left(1 - \exp(-\frac{\partial \Phi}{\partial \theta}(g, \theta_M))\right)^2 2\pi\theta_M(1-\theta_M)\beta_M\right] + o(1)
\end{aligned}$$

C - Gaussian approximation of the fields distribution.

For a fixed number $M+1$ of selective neurons in pattern $\vec{\xi}^1$, approximating the distribution of the fields on background neurons h_i^n and selective neurons h_i^s with a gaussian distribution gives:

$$\mathbb{P}^G(h_i^n = S) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(S-\mu_b)^2}{2\sigma_n^2}\right) \tag{59}$$

where

$$\mu_b = Mg\,,\ \sigma_n^2 = Mg(1-g) \tag{60}$$

and

$$\mathbb{P}^G(h_i^s = S) = \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left(-\frac{(S-\mu_f)^2}{2\sigma_s^2}\right) \tag{61}$$

where

$$\mu_f = Mg_+\,,\ \sigma_s^2 = Mg_+(1-g_+) \tag{62}$$

The probability that those fields are on the wrong side of the threshold are:

$$\mathbb{P}^G(h_i^n \geq fN\theta) = \int_{fN\theta}^{+\infty} P^G(h_i^n = z)dz \tag{63}$$

and

$$\mathbb{P}^G(h_i^s \leq fN\theta) = \int_{-\infty}^{fN\theta} P^G(h_i^s = z)dz \tag{64}$$

Following the same line of calculation than in Methods A, and keeping only terms that are relevant in the limit $N \to +\infty$, the probability that there is no error is given by:

$$\Theta(\Phi^G(g_+,\theta))\Theta(-\beta\Phi^G(g,\theta)+1) \tag{65}$$

where the rate function Φ^G is

$$\Phi^G(x,\theta) = \frac{(\theta-x)^2}{2x(1-x)} \tag{66}$$

Calculations with the binomial versus the gaussian approximation differ only in the form of Φ. Finite size terms can be taken into account in the same way it is done in Methods B for the binomial approximation.

In all above calculations we assumed that fields are sums of independent random variables (35). For small f correlations are negligible [17, 19]. It is possible to compute the covariances between the terms of the sum (see Eq. (3.9) in [19]), and take them into account in the gaussian approximation. This can be done using

$$\sigma_n^2 = Mg(1-g) + M(M-1)\gamma \tag{67}$$

$$\sigma_s^2 = Mg_+(1-g_+) + M(M-1)\gamma \tag{68}$$

in Eqs. (59),(61), where

$$\gamma = f\frac{\delta^2}{2(1+\delta)^3} \tag{69}$$

Acknowledgements

We would like to thank Stefano Fusi for his comments on a first version of the manuscript.

References

[1] Hopfield JJ (1982) Neural Networks and Physical Systems with Emergent Collective Computational Abilities. PNAS, 79, 2554-2558.

[2] Amit DJ (1989) Modeling Brain Function: The World of Attractor Neural Networks. Cambridge University Press.

[3] Brunel N (2004) Network Models of Memory, in Methods and Models in Neurophysics , C. Chow, B. Gutkin, D. Hansel, C. Meunier and J. Dalibard Eds., Elsevier.

[4] Fuster JM, Alexander G (1971) Neuron activity related to short-term memory. Science 173, 652-654.

[5] Miyashita Y (1993) Inferior Temporal Cortex: where visual perception meets memory. Ann. Rev. Neurosci. 16, 245-263.

[6] Fuster JM (1995) Memory in the cerebral cortex. MIT press.

[7] Goldman-Rakic PS (1995) Cellular basis of working memory. Neuron 14, 477-485.

[8] Amit DJ, Gutfreund H, Sompolinsky H (1987) Statistical mechanics of neural networks near saturation. Annals of Physics 173, 30-67.

[9] Sompolinsky H (1986) Neural networks with nonlinear synapses and a static noise. Physical Review A 34, 2571-2574.

[10] Gardner E (1988) The space of interactions in neural network models. Journal of Physics A: Mathematical and General 21, 257.

[11] Tsodyks M, FeigelMan M (1988) The enhanced storage capacity in neural networks with low activity level. EPL (Europhysics Letters) 6, 101.

[12] Sejnowski TJ (1977) Storing covariance with nonlinearly interacting neurons. Journal of Mathematical Biology 4, 303-321.

[13] Amit DJ, Tsodyks MV (1991) Quantitative study of attractor neural networks retrieving at low spike rates: II. Low-rate retrieval in symmetric networks. Network: Computation in Neural Systems 2, 275-294.

[14] Nadal JP, Toulouse G, Changeux JP, Dehaene S (1986) Networks of formal neurons and memory palimpsests. Europhys. Lett. 1, 535-542

[15] Parisi G (1986). A memory which forgets. Journal of Physics A: Mathematical and General 19, L617.

[16] Tsodyks M (1990) Associative Memory in Neural Networks with Binary Synapses. Modern Physics Letters B, 4, 713.

[17] Amit DJ, Fusi S (1994) Learning in neural networks with material synapses. Neural Computation 6, 957-982.

[18] Willshaw DJ, Buneman OP, Longuet-Higgins HC (1969) Non-Holographic Associative Memory. Nature 222, 960-962.

[19] Amit Y, Huang Y (2010) Precise capacity analysis in binary networks with multiple coding level inputs. Neural Computation 22, 660-688.

[20] Huang Y, Amit Y (2011) Capacity analysis in multi-state synaptic models: a retrieval probability perspective. Journal of Computational Neuroscience 30, 699-720.

[21] Nadal JP (1991) Associative memory: on the (puzzling) sparse coding limit. Journal of Physics A: Mathematical and General 24, 1093.

[22] Knoblauch A, Palm G, Sommer FT (2010) Memory capacities for synaptic and structural plasticity. Neural Computation 22, 289-341.

[23] Brunel N, Carusi F, Fusi S (1998) Slow stochastic Hebbian learning of classes of stimuli in a recurrent neural network. Network: Computation in Neural Systems 9, 123-152.

[24] Leibold C, Kempter R (2008) Sparseness constrains the prolongation of memory lifetime via synaptic metaplasticity. Cerebral Cortex 18, 67-77.

[25] Gutfreund H, Stein Y (1990) Capacity of neural networks with discrete synaptic couplings. Journal of Physics A: Mathematical and General 23, 2613.

[26] Brunel N (1994) Storage capacity of neural networks: effect of the fluctuations of the number of active neurons per memory. Journal of Physics A: Mathematical and General 27, 4783.

[27] Baldassi C, Braunstein A, Brunel N, Zecchina R (2007) Efficient supervised learning in networks with binary synapses. PNAS, 104, 11079-11084.

[28] Petersen CC, Malenka RC, Nicoll RA, Hopfield JJ (1998) All-or-none potentiation at CA3-CA1 synapses. PNAS, 95, 4732-4737.

[29] Montgomery JM, Madison DV (2004) Discrete synaptic states define a major mechanism of synapse plasticity. Trends in Neurosciences 27(12), 744-750.

[30] OConnor DH, Wittenberg GM, Wang SSH (2005) Graded bidirectional synaptic plasticity is composed of switch-like unitary events. PNAS, 102, 9679-9684.

[31] Enoki R, Hu YL, Hamilton D, Fine A (2009) Expression of long-term plasticity at individual synapses in hippocampus is graded, bidirectional, and mainly presynaptic: optical quantal analysis. Neuron 62(2), 242-253.

[32] Loewenstein Y, Kuras A, Rumpel S (2011) Multiplicative dynamics underlie the emergence of the log-normal distribution of spine sizes in the neocortex in vivo. Journal of Neuroscience 31(26), 9481-9488.

[33] Barrett AB, van Rossum MC (2008) Optimal learning rules for discrete synapses. PLoS Computational Biology 4(11), e10000230.

[34] Van Vreeswijk C, Sompolinsky H (1996) Chaos in neuronal networks with balanced excitatory and inhibitory activity. Science 274:1724-1726.

[35] Amit DJ, Brunel N (1997) Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. Cerebral Cortex 7, 237-252.

[36] van Vreeswijk CA, Sompolinsky H (1998) Chaotic Balanced State in a Model of Cortical Circuits. Neural Comp. 10:1321-1372.

[37] Brunel N (2000) Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. Journal of Computational Neuroscience 8, 183208.

[38] Roudi Y, Latham PE (2007) A Balanced Memory Network. PLoS Computational Biology 3, e141.

[39] van Vreeswijk CA, Sompolinsky H (2004) Irregular activity in large networks of neurons, in Methods and Models in Neurophysics , C. Chow, B. Gutkin, D. Hansel, C. Meunier and J. Dalibard Eds., Elsevier.

[40] Miyashita Y (1988) Neuronal correlate of visual associative long-term memory in the primate temporal cortex. Nature, 335, 817-820.

[41] Miyashita Y, Chang HS (1988) Neuronal correlate of pictorial short-term memory in the primate temporal cortex. Nature, 331, 68-70.

[42] Nakamura K, Kubota K (1995) Mnemonic firing of neurons in the monkey temporal pole during a visual recognition memory task. Journal of Neurophysiology, 74, 162-178.

[43] Alvarez P, Squire LR (1994) Memory consolidation and the medial temporal lobe: a simple network model. PNAS, 91, 70417045.

[44] Kàli S, Dayan P (2004) Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. Nature Neuroscience 7, 3, 286-294.

[45] Roxin A, Fusi S (2013) Efficient Partitioning of Memory Systems and Its Importance for Memory Consolidation. PLoS Computational Biology 9, e1003146.

Figure Legends

Figure 1: *Optimized information capacity of the Willshaw model in the limit $N \to +\infty$. Information is optimized by saturating (19) ($\theta = 1$) and (20): a. i_{opt} as a function of g, b. i_{opt} as a function of $\beta = fN/\ln N$.*

Figure 2: *Optimized information capacity for the SP model in the limit $N \to +\infty$. a. i_{opt} as a function of g, b. i_{opt} as a function of δ, the ratio between the number of depressing events and potentiating events at pattern presentation, c. i_{opt} as a function of $\beta = f\frac{N}{\ln N}$, d. i_{opt} as a function of the LTP transition probability q_+.*

Figure 3: *Optimized information capacity for the MP model in the limit $N \to +\infty$. a. Optimal information capacity as a function of g, the average number of activated synapses after learning. Optimal capacity is reached in the limit $\delta \to 0$ and at $x = 0$ where the capacity is the same as for the Willshaw model. b. Dependence of information capacity on δ, the ratio between the number of depressing events and potentiating events at pattern presentation. c. Dependence on $\beta = f\frac{N}{\ln N}$. d. Dependence on the noise in the presented patterns, x. This illustrates the trade-off between the storage capacity and the generalization ability of the network.*

Figure 4: *Finite size effects. Shown is $\mathbb{P}_{ne}$, the probability that a tested pattern of a given age is stored without errors, for the SP model. a. $\mathbb{P}_{ne}$ as a function of the age of the tested pattern. Parameters are those optimizing capacity at $N \to +\infty$ (see Section 3), results are for simulations (blue line) and calculations with a binomial approximation of the fields distributions (magenta) and a gaussian approximation (red) ; $\mathbb{P}_{ne}$ is averaged over different value of M, the number of selective neurons in the tested pattern (magenta line). b Same for $N = 50,000$. c. $\mathbb{P}_{ne}$ as a function of a scaled version of pattern age (see text for details), fluctuations in M are discarded on this plot. d. Same as c with an average of $\mathbb{P}_{ne}$ over different M.)*

Figure 5: *Capacity at finite N. a,b. P_c as a function of f for the SP model and $N = 10^4, 5.10^4$ Parameters are chosen to optimize capacity under the binomial approximation. Shown are the result of the gaussian approximation without covariance (cyan) and with covariance (magenta) for these parameters. c. Optimized P_c as a function of f for the SP model at $N = 10,000$. The blue curve is for patterns with fluctuations in the number of selective neurons. The red curve is for the same number of selective neurons in all patterns. The black curve is the number of patterns that would be stored if the network were storing the same amount of information as in the case $N \to +\infty$. d. Same for the MP model, where parameters have been optimized, but the depression-potentiation ratio is fixed at $\delta = 1$.*

Figure 6: *Storage capacity with errors in the SP model. Instead of counting only patterns that are perfectly retrieved, patterns that lead to fixed points of the dynamic overlaping significantly (see text for the definition of the overlap) with the tested memory pattern are also counted. Simulations are done with the same parameters as in figure 5a. a. P_c as a function of f. Blue crosses correspond to fixed points that are exactly the stored patterns. Red triangles correspond to fixed points that have an overlap larger than 0.99, and brown circles an overlap larger than 0.7. b. Same as a. but instead of quantifying storage capacity with P_c, it is done with $i = \frac{P_c(-f\log_2 f-(1-f)\log_2(1-f))}{N}$.*

Figure 7: *Storage capacity optimized with inhibition in the SP model. Blue is for a fixed threshold and fluctuations in the number of selective neurons per pattern. Green, the fluctuations are minimized using inhibition. Red, without fluctuations in the number of selective neurons per pattern. a. Number of stored patterns as a function of the coding level f. b. Stored information in bits per synapse, as a function of f.*