# Precise Capacity Analysis in Binary Networks with Multiple Coding Level Inputs

**Yali Amit**
*amit@galton.uchicago.edu*
*Departments of Statistics and Computer Science, University of Chicago, Chicago, IL 60637, U.S.A.*

**Yibi Huang**
*yibih@uchicago.edu*
*Department of Statistics, University of Chicago, Chicago, IL 60637, U.S.A.*

**We compute retrieval probabilities as a function of pattern age for networks with binary neurons and synapses updated with the simple Hebbian learning model studied in Amit and Fusi (1994). The analysis depends on choosing a neural threshold that enables patterns to stabilize in the neural dynamics. In contrast to most earlier work, where selective neurons for each pattern are drawn independently with fixed probability *f*, here we analyze the situation where *f* is drawn from some distribution on a range of coding levels. In order to set a workable threshold in this setting, it is necessary to introduce a simple inhibition in the neural dynamics whose magnitude depends on the total activity of the network. Proper choice of the threshold depends on the value of the covariances between the synapses for which we provide an explicit formula. Retrieval probabilities depend on the distribution of the fields induced by a learned pattern. We show that the field induced by the first learned pattern evolves as a Markov chain during subsequent learning epochs, leading to a recursive formula for the distribution. Alternatively, the distribution can be computed using a normal approximation, which involves the value of the synaptic covariances. Capacity is computed as the sum of the retrival probabilities over all ages. We show through simulation that the chosen threshold enables retrieval with asynchronous dynamics even in the presence of significant noise in the initial state of the pattern. The computed probabilities with both methods are shown to be very close to probabilities estimated from simulation. The analysis is extended to randomly connected networks.**

## 1 Introduction

Amit and Fusi (1994) presented a simple model for Hebbian learning where both neurons and synapses are binary. When the presynaptic and postsynaptic neurons are on, there is a positive probability for the synapse to move from the depressed state to the potentiated state. If the presynaptic neuron is on but the postsynaptic is off, there is a positive probability for the synapse to move from the potentiated to the depressed state. This is the simplest possible dynamic synaptic modification rule that allows continuous learning where older patterns are gradually forgotten as new ones are learned. Patterns presented to the network are random and independent, and are produced as collections of independent binary variables—one for each neuron in the network. The capacity, the number of patterns the network can retain in memory, is analyzed in terms of the signal-to-noise ratio of the local fields produced by a pattern after a given number of learning steps. Variability in the fields is due to both the stochastic nature of the patterns and the stochastic nature of the synaptic updating rule. The capacity is shown to depend heavily on the coding level of the patterns, namely, the fraction of selective neurons or units equal to 1. The interesting conclusion is that if the coding level is on the order of $\log N/N$, where $N$ is the number of neurons, then asymptotically in $N$, the capacity of the network is on the order of $N^2/\log^2 N$. This analysis is extended in Romani, Amit, and Amit (2008), where a prescription for setting a threshold for the neuronal dynamics is provided (in terms of the parameters of the network) and capacity is subsequently defined as the number of patterns that can sustain themselves in the neuronal dynamics, after the initial stimulus is removed. The signal-to-noise estimates for capacity are then shown to be consistent with simulation. One basic assumption of the these two papers is that the expected coding level of the patterns is constant. This is also an assumption found in many other papers analyzing Hebbian learning such as (Tsodyks, 1990; Amit & Fusi, 1994; Brunel, Carusi, & Fusi, 1998; Amit & Mongillo, 2003; Del Giudice, Fusi, & Mattia, 2003; Senn & Fusi, 2005; Bernacchia & Amit, 2007; Fusi & Abbott, 2007; Ben Dayan Rubin & Fusi, 2007; Leibold & Kempter, 2008), as well as in other work analyzing network capacity independent of the learning mechanism (Willshaw, Buneman, & Longuet-Higgins, 1969; Gardner, 1986; Nadal & Toulouse, 1990). The stochastic nature of the inputs does allow for some variability in the coding level, and this does have a pronounced effect, as shown in Nadal (1991); however; the variability of the size of the selective set is small compared to the size itself.

It would appear that real inputs cannot be of such fixed coding levels, which motivates the first contribution of this letter, extending the analysis of the learning process to patterns deriving from a mixture distribution over coding levels. At each step, a pattern is chosen by first sampling a coding level from a distribution on coding levels and then sampling the inputs of the pattern based on the chosen coding level. Introducing a simple inhibition mechanism in the neural dynamics, and properly selecting a threshold, we show that the network can stably retrieve learned patterns from this richer population. We provide a recursive formula for the covariance of the synapses feeding into any given neuron. This quantity turns out to be most important in setting the threshold level to prevent the firing of too many nonselective neurons. An order of magnitude estimate of this covariance is provided in Amit and Fusi (1994) for low coding levels. Here, the precise value of the covariance as a function of pattern age is derived.

The analysis of capacity is based on a formula for the probability of a pattern to be retrieved in the dynamics as a function of the pattern's age. We show that the sum of synapses feeding into a neuron from a fixed set of other neurons evolves as a Markov chain, and we analyze its finite time and asymptotic properties. The probability of retrieval is obtained in terms of the distribution of these sums relative to the preset threshold that determines the dynamics. We also develop a simpler approximate formula for retrieval probability using normal approximations, which involves the synaptic covariance formulas. Ignoring these covariances can overestimate the retrieval probability.

Capacity can be redefined as the expected number of retrieved patterns over all ages, which is nothing but the sum of retrieval probabilities over all ages. The order of magnitude of the capacity can still be obtained using signal-to-noise considerations conditional on coding level, and using the predetermined threshold as a lower bound; however, precise probabilities would in principle enable comparison to psychometric experiments, which record retrieval as a function of pattern age. Finally the analysis is extended to noisy inputs in retrieval and to networks with random connectivity. We find that retrieval is very robust to rather high noise levels on the selective neurons (up to 50%). On the other hand, assuming synaptic connectivity is random, capacity drops significantly with the fraction of connected synapses.

The letter is organized as follows. In section 2 we define the basic network setup, the learning dynamics, and the network dynamics, including inhibition. In section 3 we introduce the new stimulus distribution, with multiple coding levels, analyze the corresponding Markov chain on synaptic weights, and provide closed-form formulas for means and variances of the fields induced by the first pattern. In section 3.3, we show how a threshold can be set to allow retrieval of patterns with different coding levels, as long as the proper level of inhibition is introduced. In section 4, we analyze the integer-valued Markov chain defined by the sum of all synapses feeding into a neuron from some fixed arbitrary set of other neurons. This yields a precise formula for retrieval probability as a function of age. We also show that this probability is well approximated through normal approximations to the sums. The analysis is extended to noisy inputs (in retrieval) and randomly connected networks. We demonstrate the accuracy of the predictions using simulations. We conclude in section 5 with a short discussion of the biological implication of the results.

## 2 Basic Network Setup

Consider a fully connected network of $N$ neurons with two states per synapse, $J_- = 0$ for depressed and $J_+ = 1$ for potentiated. The neurons are assumed to be binary as well—either activated or quiescent. Input stimuli are denoted $\xi = (\xi_1, \ldots, \xi_N)$, where $\xi_i = 0/1$. Those $\xi_i$ that are equal to one are called the *selective neurons* for the stimulus. We denote by $J_{ij}$ the state of the synapse for presynaptic neuron $j$ and postsynaptic neuron $i$. The synaptic process is determined by the stochastic sequence of patterns presented to the network.

**2.1 Hebbian Learning.** Upon the presentation of a stimulus, each synapse in the network updates its status by the following rules:

- If both presynaptic and postsynaptic neurons are activated and the synapse itself is in the depressed state, it will be potentiated with probability $q_+$.
- If the presynaptic neuron is on and the postsynaptic neuron is off and the synapse itself is potentiated, then it will be depressed with probability $q_-$.
- In all other cases, the synapse remains unchanged.

This is the simplest possible Hebbian-type learning rule that stays faithful to the assumption that synaptic updating is local; it depends only on the activity of the pre- and postsynaptic neurons, and that synaptic states are finite, discrete, and bounded.

**2.2 Dynamics and Retrieval.** In simulations we implement an asynchronous updating scheme. Given a fixed synaptic matrix $\{J_{ij}\}_{i,j=1}^{N}$ and a fixed threshold $\theta$, the neuronal configuration $\xi$ is repeatedly updated as follows:

1. Randomly choose a neuron $i$.
2. Compute the field induced by pattern $\xi$:
$$h_i(\xi) = \frac{1}{N} \sum_{j:j\neq i} J_{ij}\xi_j. \tag{2.1}$$
3. Update the status of neuron $i$ according to
$$\xi_i^{new} = \mathbf{1}[h_i \geq \theta], \quad \xi_j^{new} = \xi_j, \quad j \neq i. \tag{2.2}$$
4. Set $\xi = \xi^{new}$ and return to step 1.

If this dynamics stabilizes at a pattern similar to the initial pattern $\xi$, we say the network "retrieves" $\xi$ and can retain $\xi$ in working memory without the presence of external input.

Inhibition is introduced by adding a parameter $\eta$ for the strength of the inhibitory input into each neuron. We assume that inhibition is proportional to network activity. Let $\bar{\xi} = \sum_j \xi_j / N$. The neuronal dynamics is now given by

$$\xi_i^{new} = \mathbf{1}[h_i - \eta\bar{\xi} > \theta]. \tag{2.3}$$

The motivation for this dynamics is as follows. Assume $N_I$ inhibitory neurons all receive synaptic input of size 1 from all $N$ excitatory neurons in the network. Assume that the probability of firing for each inhibitory neuron is proportional to the input. Then if all inhibitory neurons feed into each excitatory neuron with strength $\eta$, the inhibitory input to each such neuron will be close to $\eta\bar{\xi}$.

## 3 Stimuli Distribution and Learning Dynamics

We train the network with a stream of temporally independent and homogeneous stimuli $\{\ldots, \xi^{(-1)}, \xi^{(0)}, \ldots, \xi^{(p)}, \ldots\}$, where $\xi^{(p)} = (\xi_1^{(p)}, \ldots, \xi_N^{(p)})$ is referred to as the $p$th stimulus or the $p$th pattern, and $\xi_i^{(p)} = 0$ or 1 is the indicator of whether neuron $i$ is activated by the $p$th stimulus, for all $p$ and $i = 1, \ldots, N$. To relate this to sensory inputs, we will often refer to the $\xi_i$ as features, and those that are activated as features present in the stimulus.

In most previous theoretical work, the assumption is that that $\xi_i$ are independently set to 1 (selective) with some probability $f$, or that a fixed-size random subset of size $fN$ is sampled and assigned as the selective neurons of the pattern. In other words, the average coding level is assumed constant (see Nadal & Toulouse, 1990; Amit & Fusi, 1994; Brunel et al., 1998). 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, that is, neurons, that are activated. Since there is no specific preference to any of the inputs, we assume that the distribution of $\xi_j$, $j = 1, \ldots N$ is exchangeable in $j$. The joint distribution then depends on only the number of selective neurons but not on the specific set of active neurons. It can be summarized in terms of marginal probabilities,

$$p_{m,n} = \mathsf{P}(\text{the first } m \text{ neurons are 1, the next } n \text{ neurons are 0}), \tag{3.1}$$

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

$$p_{m,n} = \int_0^1 f^m (1 - f)^n \mu(df), \tag{3.2}$$

where $\mu$ is some distribution on the unit interval. Note that $p_{1,0}$ is the average coding level. In other words, at each step, draw a random variable

$F^{(p)}$ distributed according to $\mu$, and then draw $\xi_i^{(p)}$ independently $\mathsf{P}(\xi_i^{(p)} = 1) = F^{(p)}$. The case analyzed so far corresponds to $\mu = \delta_f$ for some value $f$.

Denote the synaptic efficacy after $p$-steps of learning as $J_{ij}^{(p)}$. If the network is presented with a sequence of independent samples from the above population of patterns, then $J_{ij}^{(p)}$ behaves as a Markov chain with the following transition matrix,

$$P = \begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix} = \begin{pmatrix} 1 - p_{2,0}q_+ & p_{2,0}q_+ \\ p_{1,1}q_- & 1 - p_{1,1}q_- \end{pmatrix} \tag{3.3}$$

where $p_{1,1}$ and $p_{2,0}$ are defined in equation 3.1.

Let $\alpha := p_{2,0}q_+$, $\beta := p_{1,1}q_-$. The subleading eigenvalue of the transition matrix is

$$\lambda := 1 - \alpha - \beta, \tag{3.4}$$

and the stationary distribution on the two states of the synapse is

$$(\pi_0, \pi_1) := \left(\frac{\beta}{\alpha + \beta}, \frac{\alpha}{\alpha + \beta}\right). \tag{3.5}$$

Let $\rho_{xy}^{(p)} := \mathsf{P}(J_{ij}^{(p)} = 1 \mid \xi_i^{(1)} = x, \xi_j^{(1)} = y)$ be the distribution of $J_{ij}^{(p)}$, conditional on the status of the pre- and postsynaptic neuron in the first pattern. Assuming the synapses are initialized at the stationary distribution ($\mathsf{P}(J_{ij}^{(0)} = 1) = \pi_1$), we have

$$\begin{aligned} \rho_{11}^{(p)} &= \pi_1 + \lambda^{p-1}\pi_0 q_+, \quad & \rho_{10}^{(p)} &= \pi_1, \\ \rho_{01}^{(p)} &= \pi_1 - \lambda^{p-1}\pi_1 q_-, \quad & \rho_{00}^{(p)} &= \pi_1. \end{aligned} \tag{3.6}$$

Note that asymptotically as $p \to \infty$, $\rho_{xy}^{(p)} \to \pi_1$.

We are interested in whether the network can retrieve the learned pattern $\xi^{(1)}$ after learning $\xi^{(2)}, \ldots, \xi^{(p)}$, that is, whether $\xi^{(1)}$ is stable with respect to the dynamics in a network with synaptic states given by $\{J_{ij}^{(p)}\}$. Denote by

$$h_i^{(p)} = \frac{1}{N} \sum_{j:j \neq i} (J_{ij}^{(p)} - \eta)\xi_j^{(1)}$$

the mean field induced by $\xi^{(1)}$ after $p$-steps of learning. The conditional mean $h_i^{(p)}$ of selective and nonselective neurons induced by pattern $\xi^{(1)}$, given the number of selective neurons in the first pattern (i.e., conditional on $\sum_j \xi_j^{(1)} = m$), is, respectively,

$$\mu_{m,1}^{(p)}(\eta) := \mathsf{E}\left[h_i^{(p)} \mid \xi_i^{(1)} = 1, \sum_j \xi_j^{(1)} = m\right] = \frac{m}{N}\mathsf{E}\big[(J_{ij}^{(p)} - \eta)\xi_j^{(1)} \mid \xi_i^{(1)} = 1\big]$$
$$= \frac{m}{N}\mathsf{E}\{\xi_j^{(1)} E\big[(J_{ij}^{(p)} - \eta) \mid \xi_i^{(1)} = 1, \xi_j^{(1)} = 1\big]\} = \frac{m}{N}(\rho_{11}^{(p)} - \eta) \tag{3.7}$$

$$\mu_{m,0}^{(p)}(\eta) := \mathsf{E}\left[h_i^{(p)} \mid \xi_i^{(1)} = 0, \sum_j \xi_j^{(1)} = m\right] = \frac{m}{N}\big(\rho_{01}^{(p)} - \eta\big). \tag{3.8}$$

### 3.1 Synaptic Covariances and Field Variance.

The signal-to-noise ratio analysis in Amit and Fusi (1994) and Romani et al. (2008) is performed using an approximation of the asymptotic variance of the fields, which ignores synaptic correlations. Although this is a reasonable approximation, we provide here a closed-form formula for these covariances, both for finite $p$ and as $p$ tends to infinity. This will prove important for the more refined analysis of retrieval probability developed in section 4.

Denote the covariance of two synapses $J_{ij}^{(p)}, J_{ik}^{(p)}$ conditional on $\xi^{(1)}$ as

$$\gamma_{xyz}^{(p)} := \mathsf{Cov}\big(J_{ij}^{(p)}, J_{ik}^{(p)} \mid \xi_i^{(1)} = x, \xi_j^{(1)} = y, \xi_k^{(1)} = z\big).$$

We have the following recursive formula for $\gamma_{xyz}^{(p)}$:

**Proposition 1.** *The conditional covariance of two synapses $J_{ij}^{(p)}, J_{ik}^{(p)}$ given $\xi_i^{(1)} = x, \xi_j^{(1)} = y, \xi_k^{(1)} = z$ satisfies the recursion*

$$\gamma_{xyz}^{(p)} = r\gamma_{xyz}^{(p-1)} + b_{xyz}^{(p)},$$

*where*

$$r = 1 - 2\alpha - 2\beta + p_{3,0}q_+^2 + p_{2,1}q_-^2$$
$$b_{xyz}^{(p)} = \big(1 - \rho_{xy}^{(p-1)}\big)\big(1 - \rho_{xz}^{(p-1)}\big)p_{3,0}q_+^2 + \rho_{xy}^{(p-1)}\rho_{xz}^{(p-1)}p_{2,1}q_-^2$$
$$-\big[\big(1 - \rho_{xy}^{(p-1)}\big)\alpha - \rho_{xy}^{(p-1)}\beta\big]\big[\big(1 - \rho_{xz}^{(p-1)}\big)\alpha - \rho_{xz}^{(p-1)}\beta\big]$$

*and*

$$\gamma := \lim_{p\to\infty} \gamma_{xyz}^{(p)} = \frac{\pi_0^2 q_+^2 p_{3,0} + \pi_1^2 q_-^2 p_{2,1}}{2\alpha + 2\beta - p_{3,0}q_+^2 - p_{2,1}q_-^2}$$

*Moreover, it can be shown that $\gamma = O(p_{1,0})$.*

**Proof.** See the appendix.

With proposition 1, one can calculate the conditional variance of the field at each step. It suffices to work out the case with no inhibition ($\eta = 0$) since the inhibition term $\eta(\sum_j \xi_j^{(1)}) = m\eta$ is simply a constant when $M^{(1)} \doteq \sum_j \xi_j^{(1)} = m$ is given. For selective neurons,

$$\begin{aligned}\left(R_{m,1}^{(p)}\right)^2 :=& \operatorname{Var}\left(h_i^{(p)} \mid \xi_i^{(1)} = 1, M^{(1)} = m\right)\\ =& \frac{m}{N^2}\operatorname{Var}\left(J_{ij}^{(p)} \mid \xi_i^{(1)} = \xi_j^{(1)} = 1\right)\\ &+ \frac{m^2 - m}{N^2}\operatorname{Cov}\left(J_{ij}^{(p)}, J_{ik}^{(p)} \mid \xi_i^{(1)} = \xi_j^{(1)} = \xi_k^{(1)} = 1\right)\\ =& \frac{m\rho_{11}^{(p)}\left(1 - \rho_{11}^{(p)}\right) + m(m-1)\gamma_{111}^{(p)}}{N^2},\end{aligned} \tag{3.9}$$

and similarly for nonselective neurons,

$$\begin{aligned}\left(R_{m,0}^{(p)}\right)^2 :=& \operatorname{Var}\left(h_i^{(p)} \mid \xi_i^{(1)} = 0, M^{(1)} = m\right)\\ =& \frac{m\rho_{01}^{(p)}\left(1 - \rho_{01}^{(p)}\right) + m(m-1)\gamma_{011}^{(p)}}{N^2}.\end{aligned} \tag{3.10}$$

If $\xi^{(1)}$ is known to be generated with coding level $f$, that is, $F^{(1)} = f$, but the pattern size $M^{(1)}$ is unknown, then $\mathsf{E}[M^{(1)}] = Nf$, and $\operatorname{Var}(M^{(1)}) = Nf(1-f)$. By equation 3.7, the mean field $h_i^{(p)}$ is

$$\begin{aligned}\mu_{f,x}^{(p)}(\eta) :=& \mathsf{E}\left[h_i^{(p)} \mid \xi_i^{(1)} = x, F^{(1)} = f\right]\\ =& \mathsf{E}\left[M^{(1)} \mid F^{(1)} = f\right]\left(\rho_{x1}^{(p)} - \eta\right)/N = f\left(\rho_{x1}^{(p)} - \eta\right),\end{aligned} \tag{3.11}$$

for $x = 0, 1$. By equations 3.9 and 3.10, the variance of a pattern of coding level $f$ is

$$\begin{aligned}\left(R_{f,x}^{(p)}\right)^2 :=& \operatorname{Var}\left(h_i^{(p)} \mid \xi_i^{(1)} = x, F^{(1)} = f\right)\\ =& \mathsf{E}\left[\operatorname{Var}\left(h_i^{(p)} \mid \xi_i^{(1)} = x, F^{(1)} = f, M^{(1)}\right)\right]\\ &+ \operatorname{Var}\left(\mathsf{E}\left[h_i^{(p)} \mid \xi_i^{(1)} = x, F^{(1)} = f, M^{(1)}\right]\right)\\ =& \frac{1}{N^2}\left\{\mathsf{E}\left[M^{(1)} \mid F^{(1)} = f\right]\rho_{x1}^{(p)}\left(1 - \rho_{x1}^{(p)}\right) + \mathsf{E}\left[M^{(1)}(M^{(1)} - 1) \mid\right.\right.\\ &\left.\left. F^{(1)} = f\right]\gamma_{x11}^{(p)}\right\} + \frac{1}{N^2}\operatorname{Var}\left(M^{(1)} \mid F^{(1)} = f\right)\left(\rho_{x1}^{(p)} - \eta\right)^2.\end{aligned}$$

Dropping the terms of order $f^2/N$, we get

$$\left(R_{f,x}^{(p)}\right)^2 = \frac{f}{N}\left[\rho_{x1}^{(p)}\left(1-\rho_{x1}^{(p)}\right) + \left(\rho_{x1}^{(p)}-\eta\right)^2\right] + f^2\gamma_{x11}^{(p)}.$$

Taking the limit as $p \to \infty$, the dependence on $x$ disappears, and we have

$$R_f^2 \doteq \lim_{p\to\infty}\left(R_f^{(p)}\right)^2 = \frac{f}{N}[\pi_1\pi_0 + (\pi_1-\eta)^2] + f^2\gamma. \tag{3.12}$$

**3.2 Selection of the Threshold.** With low coding levels, the number of nonselective neurons is very large, and if a small although nonnegligible proportion of these neurons fires, the entire network will eventually be activated in the absence of inhibition, or in the presence of inhibition some arbitrary subset of the neurons will fire at each iteration. For this reason, $\theta$ should be set so as to control the probability of any nonselective neuron firing at a level $\delta$. Assuming the fields $h_i^{(p)}$ are independent of each other conditional on the pattern $\xi^{(1)}$, we write (omitting the conditioning):

$$\begin{aligned}&\mathsf{P}\left(h_i^{(p)} < \theta, \text{ for all nonselective } i\right)\\ &\qquad \approx \mathsf{P}\left(h_i^{(p)} < \theta\right)^{N(1-f)} \approx \mathsf{P}\left(h_i^{(p)} < \theta\right)^{N} = 1-\delta.\end{aligned} \tag{3.13}$$

Equations 3.11 and 3.12 with $x = 0$ provide the mean and variance of $h_i^{(p)}$ for nonselective neurons. Since we will be assuming $q_- = O(f) \ll 1$, these are very close to the asymptotic values $\mu_{f,0}^{\infty}(\eta) = f(\pi_1 - \eta)$ and $R_f^2$ even for $p = 1$.

Assume the distribution of $h_i^{(p)}$ for nonselective neurons is approximately normal, and define $C_{\delta,N}$ as the $(1-\delta)^{1/N}$-th quantile of the standard normal. Then, setting the threshold at

$$\theta_f = f(\pi_1 - \eta) + C_{\delta,N}R_f \tag{3.14}$$

will keep the probability that any nonselective neuron fires at $\delta$.

It is important to note that the threshold is determined by the properties of the nonselective neurons, in particular, the standard deviation $R_f$ of their fields. For fast learning—$q_+$ close to 1, the standard deviation $R_{f,1}^{(p)}$ of the selective neurons for moderately sized $p$ is much smaller than $R_f$.

If we assume that retrieval of a pattern of coding level $f$ requires the mean field of the selective neurons $\mu_{f,1}^{(p)}(\eta)$ to be at least $a\,R_f$ larger than $\theta_f$, we get

$$\frac{\mu_{f,1}^{(p)}(\eta) - \mu_{f,0}^{\infty}(\eta)}{R_f} = \frac{f\lambda^{(p-1)}\pi_0 q_+}{\sqrt{f[\pi_1\pi_0 + (\pi_1-\eta)^2]/N + f^2\gamma}} \sim C_{\delta,N} + a. \tag{3.15}$$

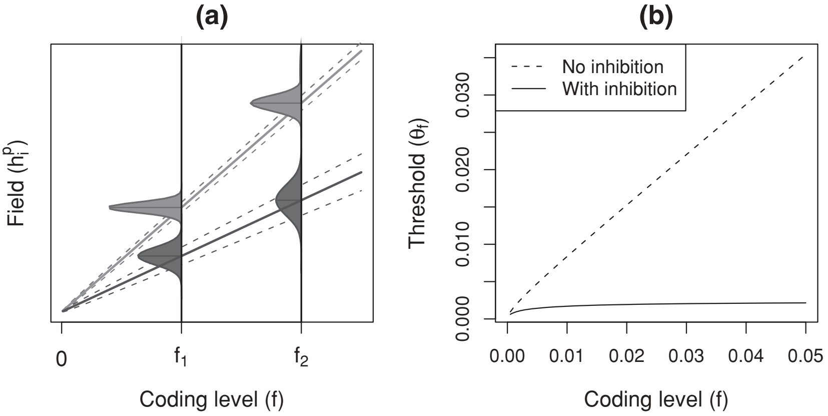


Figure 1: (a) The distribution of the fields of selective (gray) and nonselective (black) neurons of patterns at two different coding levels $f_1$ and $f_2$. The gray and black solid lines are, respectively, the mean field of selective and nonselective neurons as a function of coding levels, and the dashed lines are 1 SD away from the mean. There is no threshold to separate the selective and nonselective neurons of both coding levels. (b) Without inhibition ($\eta = 0$), the threshold $\theta_f$ is nearly linear in $f$; with inhibition ($\eta = \pi_1 + C_{\delta,N}\sqrt{\gamma}$), $\theta_f$ increases slightly with $f$ but is nearly constant.

This yields an estimate of the capacity

$$p \sim \frac{1}{-2\log\lambda}\log\left(\frac{Nf\pi_0^2 q_+^2}{(C_{\delta,N}+a)^2[\pi_1\pi_0 + (\pi_1-\eta)^2 + Nf\gamma]}\right),$$

which makes sense only if the argument of the logarithm is greater than 1. This requirement imposes a lower bound on the coding levels,

$$f > \frac{(C_{\delta,N}+a)^2(\pi_1\pi_0 + (\pi_1-\eta)^2)}{N\big[\pi_0^2 q_+^2 - (C_{\delta,N}+a)^2\gamma\big]},$$

and a constraint on $q_+$ as well: $\pi_0 q_+ > (C_{\delta,N}+a)\sqrt{\gamma}$. With only one coding level $f \sim \log N/N$ and $q_- = O(fq_+)$, then $-\log\lambda \sim f^2 = O(\log^2 N/N^2)$, and the memory capacity is on the order of $N^2/\log^2 N$.

**3.3 Multiple-Level Coding and Inhibition.** The challenge in the case of multiple-level coding is that both the mean and variance of the fields of selective and nonselective neurons grow linearly with $f$ (see Figure 1a). There is no threshold that can separate the fields of selective and nonselective neurons of the different coding levels simultaneously. If the threshold is set too high, no patterns from the lower coding level will be retrieved. If the threshold is set lower, some low-coding patterns might be successfully retrieved. However, for high-coding patterns, many nonselective neurons will be activated, leading to blow-up: full activation of all neurons in the network.

This problem can be solved by choosing an appropriate inhibition factor $\eta$. Setting

$$\eta = \pi_1 + C_{\delta,N}\sqrt{\gamma}, \tag{3.16}$$

then from equation 3.14

$$\begin{aligned}\theta_f &= f C_{\delta,N}\left[\sqrt{\frac{\pi_1\pi_0 + C_{\delta,N}^2\gamma}{Nf} + \gamma} - \sqrt{\gamma}\right]\\ &= \frac{C_{\delta,N}}{N}\frac{\pi_1\pi_0 + C_{\delta,N}^2\gamma}{\sqrt{\frac{\pi_1\pi_0+C_{\delta,N}^2\gamma}{Nf} + \gamma} + \sqrt{\gamma}}.\end{aligned}$$

As a function of $f$, the threshold $\theta_f$ grows very slowly and asymptotes at $C_{\delta,N}(\pi_1\pi_0 + C_{\delta,N}^2\gamma)/2N\sqrt{\gamma}$ (see Figure 1b). Taking $\theta = \theta_{f_{\max}}$, where $f_{\max}$ is the maximum coding level, is then sufficient to keep the nonselective neurons at all coding levels below threshold. This is our choice for all simulations we describe.

**3.4 Simulations.** To illustrate the distribution of the fields relative to the threshold, we ran a simulation with only two coding levels 0.02 and 0.04, $N = 5000$, $q_+ = 1$, $q_- = 0.04$, $\delta = 0.01$, which yields

$$\pi_1 = 0.463, \quad \lambda = 0.99784, \quad \gamma = 0.00250, \quad C_{\delta,N} = 4.61, \quad \theta = 0.0019.$$

Note that in each simulation, the synapses are initialized independently according to the corresponding stationary distribution, and then $P$ patterns are learned. After learning, we verify retrieval of the learned patterns using the asynchronous dynamics described in equation 2.3. Retrieval is achieved if more than $100(1 - e)\%$ of the selective remain active and no nonselective neurons are active. In Figure 2 we show the fields of selective (+) and nonselective (∘) for two very recent patterns of the two coding levels, as well as two older patterns. Thanks to the inhibition, the chosen threshold is able to separate selective from nonselective neurons for both coding levels.

## 4 Refined Capacity Analysis

The signal-to-noise analysis provides verifiable order of magnitude predictions on capacity for one coding level and can be extended to multiple

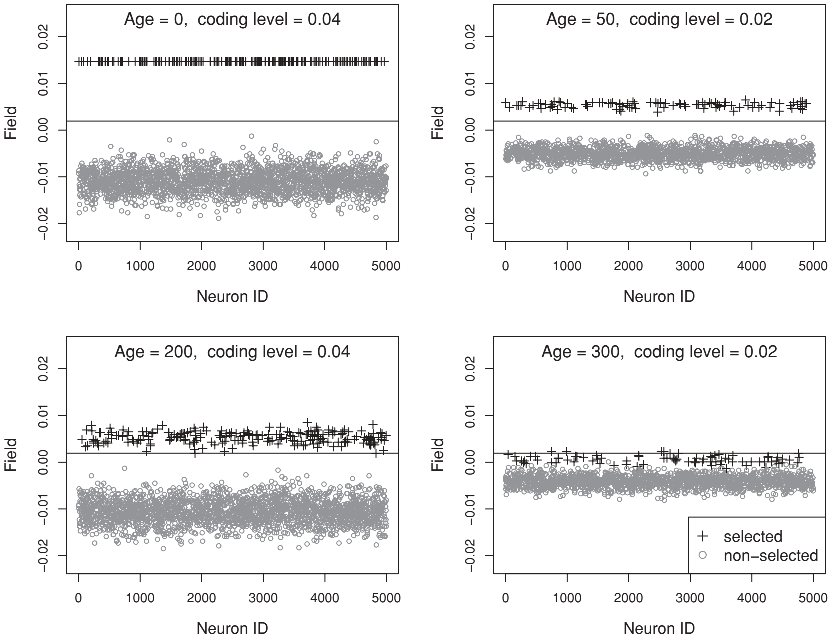


Figure 2: Scatter plots of fields for selective (+) and nonselective neurons (○) for patterns of two different coding levels at different ages. The horizontal line is the threshold $\theta = 0.0019$.

levels at least for small $f_{\max}$. Here we develop a far more detailed analysis, which provides a recursive formula for the retrieval probability at each age. We then show that this computation can be well approximated by a normal approximation, and both approximate the retrieval probability from simulations very closely. Recall that according to the learning dynamics described, the coding levels of incoming stimuli $F^{(p)}$, $p = 0, 1, 2, \ldots$, are independent and identically distributed (i.i.d.) with probability distribution $\mu(\cdot)$, and $F^{(p)}$ is independent of $J_{ij}^{(s)}$ for all $s < p$ and $1 \leq i, j \leq N$,

### 4.1 Markov Property of Synaptic Sums

**Proposition 2.** *Let $A$ be any set of indices of neurons not including $i$. Then $J_{iA}^{(p)} := \sum_{j \in A} J_{ij}^{(p)}$ is an irreducible Markov chain with state space $\{0, \ldots, |A|\}$, and*

$$J_{iA}^{(p)} = \begin{cases} Bin\big(J_{iA}^{(p-1)}, 1 - F^{(p)} q_-\big) & \text{if } \xi_i^{(p)} = 0 \\ J_{iA}^{(p-1)} + Bin\big(|A| - J_{iA}^{(p-1)}, F^{(p)} q_+\big) & \text{if } \xi_i^{(p)} = 1, \end{cases}$$

*where $F^{(p)}$ is the coding level of $\xi^{(p)}$.*

**Proof.** Let $m = J_{iA}^{(p-1)}$. According to the learning dynamics, if $\xi_i^{(p)} = 0$, only depression can occur on the $m$ incoming potentiated synapses, each independently with probability $F^{(p)}q_-$. If $\xi_i^{(p)} = 1$, potentiation can occur only on the $|A| - m$ incoming depressed synapses, each independently with probability $F^{(p)}q_+$. The chain is irreducible since all transitions have positive probability at any step.

As a special case, if $A$ is taken to be the set of selective neurons of $\xi^{(1)}$ ($A = \{j : \xi_j^{(1)} = 1, j \neq i\}$), then conditional on $\xi^{(1)}$, $\sum_j J_{ij}^{(p)}\xi_j^{(1)} = \sum_{j\in A} J_{ij}^{(p)}$ is a Markov chain.

Let $G(\cdot)$ be the cumulative distribution function of a probability distribution on the unit interval $[0, 1]$. ($G(u) = 0$ for $u \leq 0$ and $G(u) = 1$ for $u \geq 1$.) The corresponding binomial mixture on $N$ trials, denoted by $X \sim \text{BinMix}(N, G(\cdot))$, is defined as

$$P(X = m) = \binom{N}{m} \int_0^1 u^m (1-u)^{N-m} G(du).$$

We state the following result, which is proved in the appendix.

**Proposition 3.** *Using the assumptions and notation in proposition 2, if $J_{iA}^{(p-1)} \sim \nu_{p-1} = BinMix(|A|, G_{p-1}(\cdot))$, then $J_{iA}^{(p)} \sim \nu_p = BinMix(|A|, G_p(\cdot))$, where*

$$G_p(z) = \mathcal{R}(G_{p-1})(z) := \int f G_{p-1}\left(\frac{z - fq_+}{1 - fq_+}\right) + (1-f)G_{p-1}\left(\frac{z}{1 - fq_-}\right)\mu(df). \tag{4.1}$$

*Furthermore the iteration $\mathcal{R}$ is a contraction and has a unique fixed point $G_s$. Hence the stationary distribution $\nu$ of the Markov chain $J_{iA}^{(p)}$ is given by $BinMix(|A|, G_s)$.*

As an example, when there are $k$ levels $\{f_1, \ldots, f_k\}$ with weights $r_1, \ldots, r_k$, the evolution equation becomes

$$G_p(z) = \sum_{i=1}^{k} r_i \left[ f_i G_{p-1}\left(\frac{z - f_i q_+}{1 - f_i q_+}\right) + (1 - f_i) G_{p-1}\left(\frac{z}{1 - f_i q_-}\right)\right].$$

Using the same arguments from the previous proposition, we have the following:

**Proposition 4.** *Suppose at stage 0, the network is in the stationary state. Assume that $A$ is a subset of the selective neurons of the first learned pattern $\xi^{(1)}$. Then the distribution of $\tilde{h}_i^{(p)} = \sum_{j \in A \setminus i} J_{ij}^{(p)}$ is as follows:*

1. *For $p = 1$,*

$$\tilde{h}_i^1 \sim \begin{cases} BinMix(|A| - 1, G^1(\cdot)) & if\ \xi_i^{(1)} = 1 \\ BinMix(|A|,\ G^0(\cdot)) & if\ \xi_i^{(1)} = 0 \end{cases}$$

*where*

$$G^1(z) = \begin{cases} \mathbf{1}_{\{z \geq 1\}} & if\ q_+ = 1 \\ G_s(\frac{z - q_+}{1 - q_+}) & if\ q_+ < 1 \end{cases}$$

$$G^0(z) = G_s\left(\frac{z}{1 - q_-}\right) \text{ assuming } q_- < 1$$

*where $G_s(\cdot)$ is the limit point of $G_p$.*

2. *For $p = 2, 3, \ldots,$*

$$(\tilde{h}_i^{(p)} | \xi_i^{(1)} = x) \sim BinMix(|A| - x, G_p(\cdot)), x = 0, 1,$$

*where $G_p$ satisfies the recursion of equation 4.1 with $G_1 = G^1$ or $G^0$ defined above, depending on $\xi_i^{(1)} = 1$ or 0.*

**4.2 Computation of $G_p$.** The function $G_p(\cdot)$ is usually unavailable in closed form. Theoretically we can evaluate each $G_p$ exactly. However, after a few iterations, this becomes very complicated and hard to evaluate. Instead, we use the iterating equation 4.1 to evaluate $G_p(z)$ at $K + 1$ grid points,

$$z = \frac{i}{K}, \quad i = 0, 1, \ldots, K,$$

where $K$ is a large number. For nongrid points $z$, instead of tracing back to $G_{p-1}$ using equation 4.1, we linearly interpolate the two neighboring grid points of $z$ as an approximation. For example, when using equation 4.1 to evaluate $G_{p+1}(z)$ at grid $z$, $\frac{z - fq_+}{1 - fq_+}$ and $\frac{z}{1 - fq_-}$ are usually not grid points.

Once $G_p$ is approximated, the probability that $\mathsf{P}(\sum_{j=1}^{N} J_{ij}^{(p)} = m)$ is approximated by

$$\sum_{i=0}^{K} \binom{N}{m} \left(\frac{i}{K}\right)^m \left(1 - \frac{i}{K}\right)^{N-m} g_p\left(\frac{i}{K}\right),$$

where $g_p$ is the numerical differentiation of $G_p$,

$$g_p\left(\frac{i}{K}\right) := G_p\left(\frac{i}{K}\right) - G_p\left(\frac{i-1}{K}\right).$$

**4.3 Retrieval Probability.** A pattern $\xi$ is said to be retrieved with $e$-error if, after the dynamics runs for a period of time, the fraction of active selective neurons is at least $1-e$ and *no* nonselective neurons are active, at each step of the dynamics. Retrieval with 0 error is also called *perfect retrieval,* in which case the dynamics stabilizes at a pattern identical to $\xi$,

Retrieval is a function of the network dynamics. It is too difficult to fully analyze events involving the stable point of asynchronous dynamics. Instead, we analyze a simpler form of synchronous dynamics for which we can compute the retrieval probability. Empirically we observe that these computed probabilities are extremely close to the estimated probabilities obtained from simulating the asynchronous dynamics.

The modified form of synchronous dynamics proceeds as follows. Start at $\xi^{(1)}$, let $A^{(1)}$ be the set of selective neurons of stimulus 1, and write $m = |A^{(1)}|$, let $m_e = (1-e)m$. Given a state of the system $\xi^{old}$, let $H^{old}$ be the set of active neurons, and assume $|H^{old}| = m_e$. All neurons $i = 1, \ldots, N$ are updated based on the fields determined by $\xi^{old}$, yielding $\tilde{\xi}^{new}$. Now if the number of active neurons in $\tilde{\xi}^{new}$ is greater than $m_e$, leave only the first (based on the given ordering of the units) $m_e$ on and turn off the rest, yielding $\xi^{new}$. Let $H^{new}$ be the set of active neurons in $\xi^{new}$. Retrieval with $e$-error means that at each step, the set $H^{new}$ is of size $m_e$. This modification guarantees the same number of active neurons at each step and allows us to obtain a fixed inhibition level at each step.

Denote by $h_i^{(e,p)} = \frac{1}{N}\sum_{j\in H^{(old)}} J_{ij}^{(p)} - \eta m_e$. The threshold determined in equation 3.14 depended on the probability $\delta$ of any nonselective neuron firing. Given this threshold, we are interested in the conditional probability given $A^{(1)}$ that the fields generated by $H^{old}$ are above $\theta$ for at least $m_e$ of the neurons in $A^{(1)}$ and below $\theta$ for all nonselective neurons. Assuming all fields are independent conditional on the set $A^{(1)}$ and using $\theta$ defined in equation 3.14, we have

$$\begin{aligned}
\mathsf{P}_e^{(p)}(A^{(1)}) &= (1-\delta)\mathsf{P}\left(\left|\{i \in A^{(1)} : h_i^{(e,p)} \geq \theta\}\right| \geq m_e \middle| A^{(1)}\right)\\
&= (1-\delta) \sum_{S\subset A^{(1)}:|S|>m_e}\\
&\quad \times \left[\prod_{i\in S} P\left(h_i^{(e,p)} > \theta \mid A^{(1)}\right) \prod_{i\notin S} P\left(h_i^{(e,p)} < \theta \mid A^{(1)}\right)\right]\\
&= (1-\delta)\sum_{k=m_e}^{m} \binom{m}{k} \Psi_e^{(p)}(\theta)^k \left(1-\Psi_e^{(p)}(\theta)\right)^{m-k}, \qquad (4.2)
\end{aligned}$$

where

$$\Psi_e^{(p)}(\theta) = \mathsf{P}(h_i^{(e,p)} > \theta).$$

This probability can be computed using the formulas of proposition 4 or using the following approximation of the sums in terms of the normal distribution. Using the normal approximation to $h_i^{(e,p)}$, we write, conditional on $A^{(1)}$,

$$Nh_i^{(e,p)} \sim \mathcal{N}\big(\mu_{e,p}, \sigma_{e,p}^2\big),$$

where

$$\mu_{e,p} = m_e \rho_{11}^{(p)} - \eta m_e \quad \text{and} \quad \sigma_{e,p}^2 = m_e \rho_{11}^{(p)}\big(1 - \rho_{11}^{(p)}\big) + \big(m_e^2 - m_e\big)\gamma_{111}^{(p)}. \tag{4.3}$$

(see equation 3.9). Consequently,

$$\Psi_e^{(p)}(\theta) \approx \tilde{\Phi}\left(\frac{N\theta - \mu_{e,p}}{\sigma_{e,p}}\right), \tag{4.4}$$

where $\tilde{\Phi} = 1 - \Phi$ is the tail of the normal distribution. Since $\mathsf{P}_e^{(p)}(A^{(1)})$ depends on $A^{(1)}$ only through its size, we write $\mathsf{P}_e^{(p)}(m) = \mathsf{P}_e^{(p)}(A^{(1)})$. The distribution of $|A^{(1)}|$ is

$$\mathsf{P}\big(|A^{(1)}| = m\big) = \int \binom{N}{m} u^m (1-u)^{N-m} \mu(du),$$

and we finally approximate the retrieval probability as

$$\mathsf{P}_e^{(p)} \approx \int \sum_{m=0}^{N} \binom{N}{m} \mathsf{P}_e^{(p)}(m) u^m (1-u)^{N-m} \mu(du).$$

With retrieval probabilities at hand, an alternative way to obtain capacity is to compute the expected number of retrievable patterns, which is nothing but the sum of the retrieval probabilities over all ages:

$$\mathcal{C} = \sum_{p=1}^{\infty} P_e^{(p)}. \tag{4.5}$$

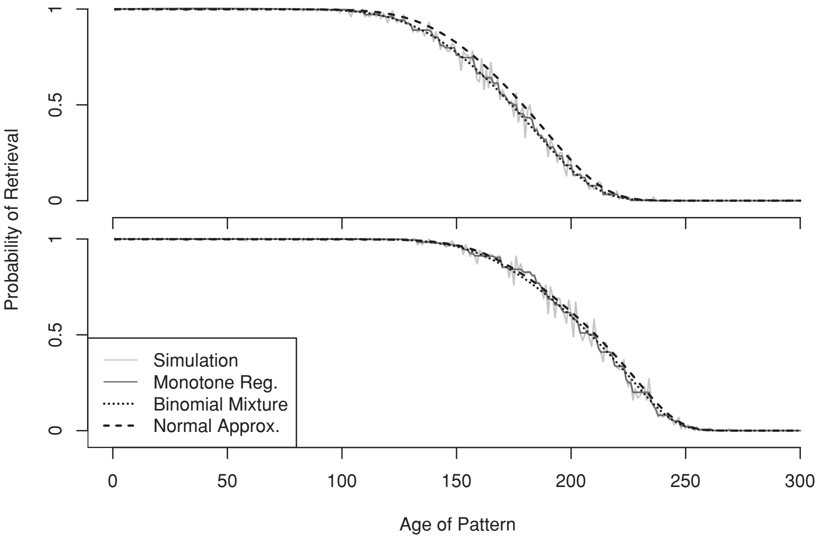


Figure 3: Predicted and simulated retrieval probabilities. The upper panel is for perfect retrieval ($e = 0$), and the lower panel allows for 5% error ($e = .05$). The jagged gray line is an average over 100 runs of a simulation for each age. The smooth black line is a monotonic regression (the nonincreasing line that best fits the observations) to the gray line. Dotted line: Binomial mixture prediction. Dashed line: Normal approximation prediction.

**4.4 Simulations.** The retrieval probabilities are shown in a number of scenarios. First, for coding levels uniformly distributed between 0.02 and 0.04, with $q_- = 0.04$, $q_+ = 1$, $N = 5000$, $\delta = 0.01$,

$$\pi_1 = 0.0445,\ \lambda = 0.9979,\ \gamma = 0.0023,\ C_{\delta,N} = 4.61,\ \theta = \theta_{0.04} = 0.0021.$$

The upper panel in Figure 3 shows predicted and actual retrieval rates with zero error ($e = 0$), whereas the lower panel shows the same for $e = .05$. The dotted curve is based on the recursive computation of the binomial mixture. The dashed curve is based on the normal approximation. The jagged gray lines are averages of 100 runs of the simulation, and the black is the nondecreasing line that best fits the gray. With the tools developed in section 4, we can predict the retrieval rates for much larger networks for which simulation is not practical. When we raise the network size $N$ to 5000, 50,000, and 150,000, and decrease the coding levels accordingly at rate $\log N/N$, we see a perfectly quadratic increase in capacity (see Figure 4).

Finally, we note that ignoring the covariance leads to overestimating the retrieval probabilities. This is not a major problem for $q_+ = 1$. But taking

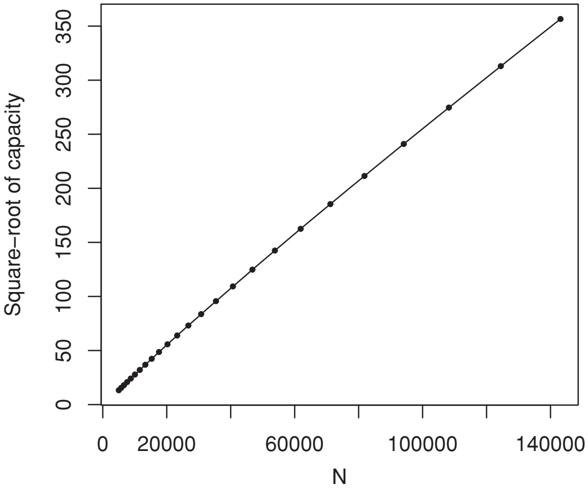
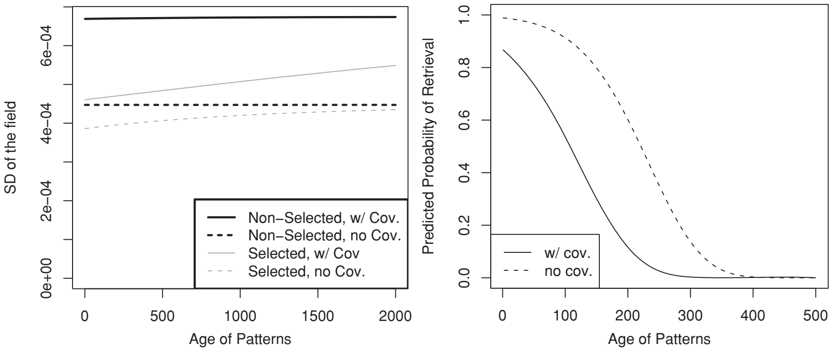


Figure 4: Square root capacity computed from equation 4.5 as a function of $N$, with $f \propto \log N/N$. The straight line verifies the nearly quadratic increase $N^2/\log^2 N$ of capacity in $N$.



(a) (b)

Figure 5: (a) Standard deviations for selective and nonselective neurons computed with and without the covariance term for different ages. (b) Retrieval probability predictions with and without covariance term. Solid line with covariance term, dashed line without. One coding level $f = .02, q_+ = .5, N = 25{,}000$.

$q_+ = .5$ in a network of 25,000 neurons, Figure 5 shows the difference in the variances on the right panel and the difference in the predicted probabilities on the left.

**4.5 Retrieval with Input Noise.** The stimuli initiating the network dynamics could be noisy versions of the learned patterns. Of interest is how the retrieval probability depends on the level of noise. We denote by $\tilde{\xi}^{(1)}$ the noisy version of the first pattern, with the following noise model:

$$\mathsf{P}\big(\tilde{\xi}_i^{(1)} = 1 \mid \xi_i^{(1)} = 1\big) = 1 - \epsilon_+ \quad \mathsf{P}\big(\tilde{\xi}_i^{(1)} = 1 \mid \xi_i^{(1)} = 0\big) = \epsilon_-.$$

Let $A_+ = \{i : \xi_i^{(1)} = 1, \tilde{\xi}_i^{(1)} = 1\}$ and $A_- = \{i : \xi_i^{(1)} = 0, \tilde{\xi}_i^{(1)} = 1\}$. Then the field induced by $\tilde{\xi}^{(1)}$ is $Nh_i = J_{iA_+}^{(p)} + J_{iA_-}^{(p)}$. The distributions of $J_{iA_+}^{(p)}, J_{iA_-}^{(p)}$ are again given by proposition 4. For the mean and variance conditional, we have

$$\mathsf{E}\big(J_{iA_+}^{(p)} \mid \xi_i^{(1)} = x, A_+\big) = |A_+|\rho_{x1}^{(p)}, \text{ and } \mathsf{E}\big(J_{iA_-}^{(p)} \mid \xi_i^{(1)} = x, A_-\big) = |A_-|\pi_1,$$

As for the variances,

$$\mathsf{Var}\big(J_{iA_+}^{(p)} \mid \xi_i^{(1)} = x, A_+, A_-\big) = |A_+|\rho_{x1}^{(p)}\big(1 - \rho_{0x1}^{(p)}\big) + |A_+|(|A_+| - 1)\gamma_{x11}^{(p)}$$

and

$$\mathsf{Var}\big(J_{iA_-}^{(p)} \mid \xi_i^{(1)} = x, A_+, A_-\big) = |A_-|\pi_1\pi_0 + |A_-|(|A_-| - 1)\gamma.$$

A similar computation for the covariance between the two terms yields

$$\mathsf{Cov}\big(J_{iA_+}^{(p)}, J_{iA_-}^{(p)} \mid \xi_i^{(1)} = x, A_+, A_-\big) = |A_+||A_-|\gamma_{x10}^{(p)}.$$

Thus, setting $m = |A^{(1)}|$, the mean and variance from equation 4.3 become

$$\mu_p = m(1 - \epsilon_+)\rho_{11}^{(p)} + (N - m)\epsilon_-\pi_1. \tag{4.6}$$

And using the variance decomposition,

$$\begin{aligned}\sigma_p^2 = {} & m\epsilon_+\rho_{11}^{(p)}\big(1 - \rho_{11}^{(p)}\big) + m(m-1)\epsilon_+^2\gamma_{111}^{(p)} + m\epsilon_+(1 - \epsilon_+)\big(\rho_{11}^{(p)}\big)^2 \\ & + (N - m)\epsilon_-\pi_1\pi_0 + (N - m)(N - m - 1)\epsilon_-^2\gamma \\ & + (N - m)\epsilon_-(1 - \epsilon_-)\gamma + m(N - m)\epsilon_+\epsilon_-\gamma_{110}^{(p)}.\end{aligned}$$

To obtain quantities of reasonable order of magnitude, we must set $\epsilon_- = \epsilon_+ p_{1,0}/(1 - p_{1,0})$.

In Figure 6, we show that retrieval is robust to a significant amount of noise $\epsilon_+ = .1$ in the input. Assuming the learned pattern is the correct

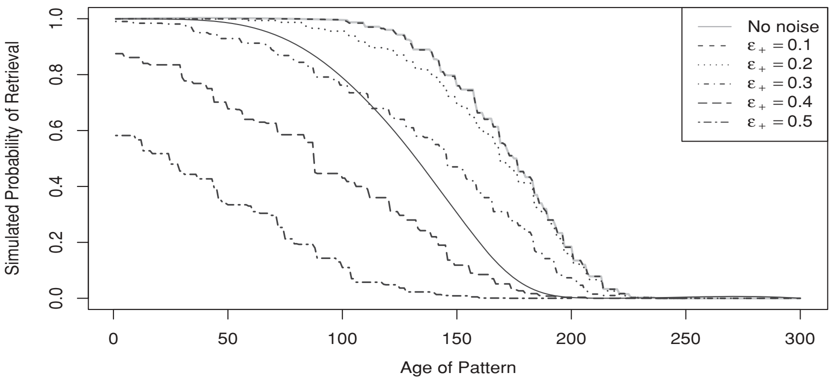


Figure 6: The simulated retrieval rates (averaging over 100 runs) with $\epsilon_+ = 0 - .5$, $\epsilon_- = \epsilon_+ p_{1,0}/(1 - p_{1,0})$ noise in the input. The parameters are as in section 4.4. Solid black: the predicted retrieval rates for $\epsilon_+ = 0.1$ based on the normal approximation and the mean and variances computed in section 4.5. The dashed line ($\epsilon_+ = 0.1$) is nearly indistinguishable from the solid gray line (no noise) and is higher than the predicted retrieval rates.

prototype, when a noisy version of this prototype is presented to the network, because of the asynchronous updating, the noise will gradually be removed. Neurons that are updated later will see a cleaner input. For this reason, the predicted retrieval probability with noise, which cannot take into account the asynchronous updating, is actually higher than the probability observed in simulation. With inhibition and using asynchronous updating, the network can recover the true pattern without error for a significant fraction of patterns, even when the noise level for selective neurons is $\epsilon_+ = .5$, $\epsilon_- = .01$.

**4.6 Randomly Connected Network.** If the network is randomly connected, the field becomes

$$h_i^{(p)} = \frac{1}{N}\left[\sum_{j \neq i} J_{ij}^{(p)} C_{ij} \xi_j^{(1)} - \eta_c \sum_j \xi_j^{(1)}\right].$$

Here $C_{ij}$ are independent Bernoulli($c$), which determine the existence of the synapse from neuron $j$ to neuron $i$. Assume the $C_{ij}$'s remain fixed throughout learning and pattern retrieval. Given $F^{(1)} = f$, $\xi_i^{(1)} = x$, using the same arguments as in section 3,

$$\begin{aligned}\mu_{f,x}^{(p)}(\eta_c) &:= \mathsf{E}\big[h_i^{(p)} \mid \xi_i^{(1)} = x, F^{(1)} = f\big] \\ &= f\big(c\rho_{x1}^{(p)} - \eta_c\big) \to f\big(c\pi_1 - \eta_c\big)\end{aligned}$$

$$\big(R_{f,x}^{(p)}\big)^2 := \frac{f}{N}\Big[c\rho_{x1}^{(p)}\big(1-c\rho_{x1}^{(p)}\big) + \big(c\rho_{x1}^{(p)}-\eta_c\big)^2\Big] + f^2c^2\gamma_{x11}^{(p)}.$$
$$\xrightarrow[p\to\infty]{} \frac{f}{N}[c\pi_1(1-c\pi_1) + (c\pi_1-\eta_c)^2] + f^2c^2\gamma$$

The inhibition factor $\eta$ from section 3.3 is scaled to give $\eta_c = c\eta$, and the new threshold is

$$\theta_{f,c} = fcC_{\delta,N}\left[\sqrt{\frac{\pi_1(1/c-\pi_1)+C_{\delta,N}^2\gamma}{Nf}+\gamma} - \sqrt{\gamma}\right].$$

The new SNR for a single coding level is

$$\frac{\mu_{f,1}^{(p)}(\eta_c) - \mu_{f,0}^{(p)}(\eta_c)}{R_f} = \frac{\lambda^{(p-1)}\pi_0 q_+}{\sqrt{\frac{\pi_1(1/c-\pi_1)+C_{\delta,N}^2\gamma}{Nf}+\gamma}}.$$

Compared with the fully connected SNR, equation 3.15, the network size $N$ has to increase by a factor of $(1/c-\pi_1)/(1-\pi_1) > 1/c$ to maintain the memory capacity at the same level as the fully connected network. For example, if $\pi_1 = 1/2$, $c = 0.1$, $N$ needs to be 19 times larger.

The plot in Figure 7 shows predicted and estimated retrieval probabilities for $c = 0.6$ using the methods of section 4.3. Note that the terms in equation 4.3 become

$$\mu_{e,p} = cm_e\big(\rho_{11}^{(p)} - \eta\big) \text{ and } \sigma_{e,p}^2 = m_e c\rho_{11}^{(p)}\big(1-c\rho_{11}^{(p)}\big) + (m_e^2 - m_e)c^2\gamma_{111}^{(p)}.$$

All parameters are the same as in section 4.4. The gray lines are for perfect retrieval, and the black lines allow 5% error. Again we see that the predicted probabilities are very close to those estimated from simulation.

Finally, in Figure 8, we show plots of capacity as a function of coding level for a number of values of the connectivity $c$ assuming $cN = 10{,}000, 20{,}000$. Note that as $c$ decreases, even if $Nc$ is held constant while $N$ increases, the capacity decreases. This is a result of the additional variability in the field induced by the random connectivity. Assuming for simplicity that $\gamma = 0$, the normalized ratio from equation 4.4 becomes

$$t_p(c) = \frac{C_{\delta,N}\sqrt{fN\pi_1(1/c-\pi_1)} - m_e\big(\rho_{11}^{(p)}-\eta\big)}{\sqrt{m_e\rho_{11}^{(p)}(1/c-\rho_{11}^{(p)})}}.$$

Retrieval is possible when $t_p(c)$ is negative of large modulus. Clearly when $c$ grows, the right-hand term in the numerator grows, and this is lost, irrespective of the magnitude of $N$. This problem could be alleviated in

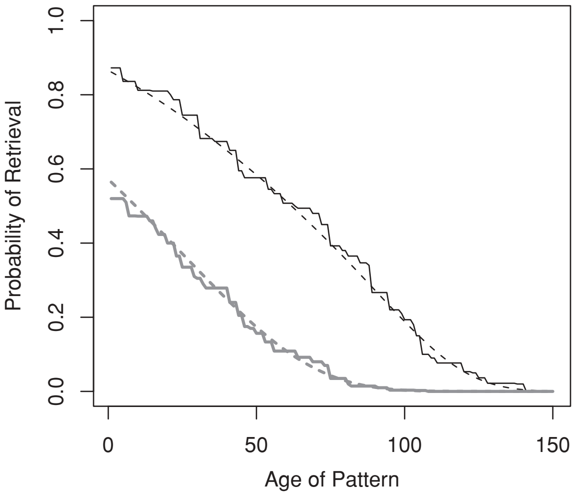


Figure 7: Simulated (solid lines) and predicted (dashed lines) retrieval probabilities for a randomly connected network with $N = 5000$ neurons and $c = .6$. The gray lines are for perfect retrieval, and the black lines allow 5% error ($e = .05$).

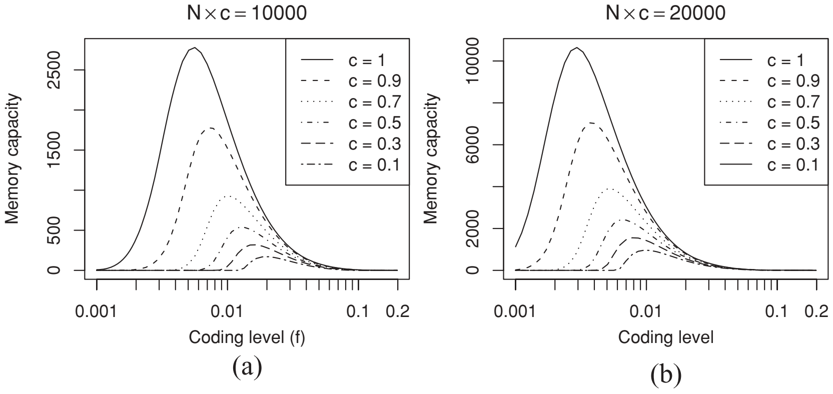


Figure 8: Capacities as a function of coding level for fixed $cN$. (a) $cN = 10{,}000$. (b) $cN = 20{,}000$. Same parameter values as for Figure 7.

an artificial way if the fraction of existing synapses among the potentiated synapses feeding into a neuron was always exactly $c$. In this case, retrieval would depend on only $cN$ and $f$.

## 5 Discussion

We have shown that with simple linear inhibition and setting the threshold appropriately, a network can retrieve noisy versions of learned patterns with a wide range of low coding levels. Representing the field produced by a learned pattern as a Markov chain that evolves through subsequent learning epochs yields an accurate formula to predict the probability of retrieval as a function of pattern age. These predictions are nonasymptotic and valid for a wide range of parameter values.

Retrieval probabilities and capacity in a recurrent network depend heavily on the threshold, which in turn must be set to prevent even moderate numbers of nonselective neurons from firing, because such errors propagate in the neural dynamics and all trace of the initial pattern will disappear. The variance of the fields of nonselective neurons, which is very close to the field variance in the stationary state, is essential in determining the threshold. This requires computing the covariances between the synapses.

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. Thus, the most meaningful capacity estimates are those shown in the last section. We see that the loss of full connectedness leads to a significant loss in capacity. Indeed according to Figure 8, for a network with $10^5$ neurons and $10^4$ synaptic connections per neuron ($c = .1$) the capacity is 0 for $f = .01$ and about 250 for $f = .02$, whereas for a fully connected network of $10^4$ neurons, it is approximately 1500. For smaller values of $c$, retrieval effectively disappears. We also note that for higher coding levels around $f = .1$, retrieval disappears as well. And yet in areas such as prefrontal cortex, people have reported such high coding levels during working memory.

Ben Dayan Rubin and Fusi (2007) and Fusi and Abbott (2007) attempt to remedy this problem by introducing metastates in each synapse, which are used to slow both learning and forgetting. The authors are also motivated by the fact that some regions, where working memory is observed, have higher coding rates (on the order of .1), which further decreases the capacity for simple binary synapses. The authors report that with metaplasticity, for a significant period of time, the rate of decay of the signal-to-noise ratio is polynomial as opposed to exponential. Moreover realizing the importance of synaptic covariances in Ben Dayan Rubin and Fusi (2007), a learning rule is proposed that guarantees synaptic independence. (We note that with multiple coding levels, this method to generate independence does not work, nor do the symmetry assumptions for potentiation seem to be realistic.) The use of metastates is further analyzed in Leibold and Kempter (2008).

The measure of capacity is somewhat different in these two papers. In Ben Dayan Rubin and Fusi (2007), the question is how much information is retained in the actual synapses, which leads to very high capacity (order $10^7$ with $f \sim .1$ and around 20 metastates in a network with $10^5$ neurons and $10^4$ synapses per neuron), although the question of how this information can be expressed in neural activity is not addressed. In Leibold and Kempter (2008), the error in neural output is evaluated in a feedforward setting. They observe that higher capacities are achieved with sparser networks and binary states. However, both papers conclude that for high coding levels, the networks with more metastates achieve higher capacity. In contrast, we have studied retrieval in neural output in a recurrent setting. Capacity results are lower since the signal-to-noise ratio constraints are more severe; even moderate errors in nonselective neurons can destroy retrieval. Analyzing networks with metastates in the recurrent setting, with actual network dynamics, is beyond the scope of this letter, but we intend to try to extend the results here to this more general setting.

However, we raise the possibility that working memory actually employs much more extensive areas of the brain than localized cortical columns and that the recurrent network pools activity of multiple local networks that are interconnected. For example, in recent years, a growing number of experiments show fMRI traces of working memory in retinotopic layers of visual cortex including primary visual cortex (Cox & Savoy, 2003; Serences, Ester, Vogel, & Awh , 2008; Harrison & Tong, 2009). The signals in each voxel are very weak and insignificant; however, using classification algorithms on the ensemble of voxels covering a visual area, one can get a strong distinction between different states of the brain. It is also well known that there are massive feedback connections between IT and lower-level retinotopic visual areas, which play much more than a passive feedforward role in processing information (Bullier, 2001; Kveraga, Ghuman, & Bar, 2007). It may be that because the fraction of neurons in V1 involved in a particular memory is so small, it cannot be picked up by the fMRI signal in individual voxels and is not easy to detect with single or even multiple electrode recordings.

If memory is indeed retained in a network consisting of multiple areas, including the retinotopic areas of the visual cortex, then the number of neurons available increases dramatically, and this may offer a way to remedy the disappointing capacity results for recurrent retrieval. Even if these speculative comments turn out to be correct, the global network is not a homogeneous one as analyzed in this letter or others. Rather, it consists of interacting relatively homogeneous modules with different coding levels and different internal and external conductivities. Such recurrent networks pose an interesting and important challenge from a modeling point of view.

## Appendix: Cumulative Distribution Function

**Proof of Proposition 1.** The learning mechanism can be summarized as

$$J_{ij}^{(p)} - J_{ij}^{(p-1)} = \begin{cases} B_{ij+}^{(p)} \quad \text{if } \xi_i^p = 1, \quad \xi_j^p = 1 \quad \text{and} \quad J_{ij}^{(p-1)} = 0 \\ -B_{ij-}^{(p)} \quad \text{if } \xi_i^p = 0, \quad \xi_j^p = 1 \quad \text{and} \quad J_{ij}^{(p-1)} = 1 \\ 0 \quad \text{otherwise} \end{cases}$$

$$= \xi_i^{(p)} \xi_j^{(p)} \left(1 - J_{ij}^{(p-1)}\right) B_{ij+}^{(p)} - (1 - \xi_i^{(p)}) \xi_j^{(p)} J_{ij}^{(p-1)} B_{ij-}^{(p)}, \quad \text{(A.1)}$$

where $B_{ij+}^{(p)}$ are i.i.d Bernoulli($q_+$) and $B_{ij-}^{(p)}$ are i.i.d Bernoulli($q_-$). All $B_{ij+}^{(p)}$, $B_{ij-}^{(p)}$, and $\xi$'s are mutually independent. One can further rewrite equation A.1 as

$$J_{ij}^{(p)} = D_{ij+}^{(p)} + \big(1 - D_{ij+}^{(p)} - D_{ij-}^{(p)}\big)J_{ij}^{(p-1)}, \tag{A.2}$$

where

$$D_{ij+}^{(p)} = \xi_i^{(p)}\xi_j^{(p)}B_{ij+}^{(p)},\ \ D_{ij-}^{(p)} = \big(1 - \xi_i^{(p)}\big)\xi_j^{(p)}B_{ij-}^{(p)}.$$

Note that $\{D_{ij+}^{(p)}, D_{ij-}^{(p)}\}$ are independent of $J_{ij}^{(p-1)}$. By equation A.2,

$$\gamma_{xyz}^{(p)} = A_1 + A_2 + A_3 + A_4,$$

where

$$\begin{aligned}
A_1 &= \mathrm{Cov}\big(D_{ij+}^{(p)}, D_{ik+}^{(p)}\big),\\
A_2 &= \mathrm{Cov}\big(D_{ij+}^{(p)}, (1 - D_{ik+}^{(p)} - D_{ik-}^{(p)})J_{ik}^{(p-1)}\big),\\
A_3 &= \mathrm{Cov}\big((1 - D_{ij+}^{(p)} - D_{ij-}^{(p)})J_{ij}^{(p-1)}, D_{ik+}^{(p)}\big),\\
A_4 &= \mathrm{Cov}\big((1 - D_{ij+}^{(p)} - D_{ij-}^{(p)})J_{ij}^{(p-1)}, (1 - D_{ik+}^{(p)} - D_{ik-}^{(p)})J_{ik}^{(p-1)}\big).
\end{aligned}$$

Here, all the covariances are conditioned on $\xi_i^{(1)} = x,\ \xi_j^{(1)} = y,\ \xi_k^{(1)} = z$, since $\{D_{ij+}^{(p)}, D_{ij-}^{(p)}\}$ and $J_{ij}^{(p-1)}$ depends only on $\{\xi^{(p)}, B_{ij+}^{(p)}, B_{ij-}^{(p)}\}$ and $\{\xi^{(k)}, B_{ij+}^{(k)}, B_{ij-}^{(k)} :\ k \le p-1\}$, respectively, and hence are independent. Using the fact that

$$\begin{aligned}
\mathrm{Cov}(AX, B) &= \mathrm{E}(X)\mathrm{Cov}(A, B)\\
\mathrm{Cov}(AX, BY) &= \mathrm{E}(AB)\mathrm{Cov}(X, Y) + \mathrm{E}(X)\mathrm{E}(Y)\mathrm{Cov}(A, B),
\end{aligned}$$

when $(A, B)$ and $(X, Y)$ are independent random variables, we have

$$\begin{aligned}
A_2 &= -\rho_{xz}^{(p)}\mathrm{Cov}\big(D_{ij+}^{(p)}, D_{ik+}^{(p)} + D_{ik-}^{(p)}\big),\\
A_3 &= -\rho_{xy}^{(p)}\mathrm{Cov}\big(D_{ij+}^{(p)} + D_{ij-}^{(p)}, D_{ik+}^{(p)}\big),\\
A_4 &= \mathrm{E}\big[\big(1 - D_{ij+}^{(p)} - D_{ij-}^{(p)}\big)\big(1 - D_{ik+}^{(p)} - D_{ik-}^{(p)}\big)\big]\gamma_{xyz}^{(p-1)}\\
&\quad + \rho_{xy}^{(p)}\rho_{xz}^{(p)}\mathrm{Cov}\big(D_{ij+}^{(p)} + D_{ij-}^{(p)}, D_{ik+}^{(p)} + D_{ik-}^{(p)}\big).
\end{aligned}$$

To sum up,

$$\gamma_{xyz}^{(p)} = A_1 + A_2 + A_3 + A_4 = r\,\gamma_{xyz}^{(p-1)} + b_{xyz}^{(p)} \tag{A.3}$$

where

$$\begin{aligned} r &= \mathsf{E}\big[\big(1 - D_{ij+}^{(p)} - D_{ij-}^{(p)}\big)\big(1 - D_{ik+}^{(p)} - D_{ik-}^{(p)}\big)\big] \\ b_{xyz}^{(p)} &= \mathsf{Cov}\big(\big(1 - \rho_{xy}^{(p)}\big)D_{ij+}^{(p)} - \rho_{xy}^{(p-1)}D_{ij-}^{(p)},\ \big(1 - \rho_{xz}^{(p)}\big)D_{ik+}^{(p)} - \rho_{xz}^{(p-1)}D_{ik-}^{(p)}\big). \end{aligned}$$

Note that $\mathsf{E}[D_{ij+}^{(p)}] = p_{2,0}q_+ = \alpha$, $\mathsf{E}[D_{ij-}^{(p)}] = p_{1,1}q_- = \beta$, $D_{ij+}^{(p)}D_{ik-}^{(p)} = D_{ij-}^{(p)}D_{ik+}^{(p)} = 0$, and

$$\begin{aligned} \mathsf{E}\big[D_{ij+}^{(p)}D_{ik+}^{(p)}\big] &= \mathsf{E}\big[\xi_i^{(p)}\xi_j^{(p)}\xi_i^{(p)}\xi_k^{(p)}B_{ij+}^{(p)}B_{ik+}^{(p)}\big] = p_{3,0}q_+^2, \\ \mathsf{E}\big[D_{ij-}^{(p)}D_{ik-}^{(p)}\big] &= \mathsf{E}\big[\big(1 - \xi_i^{(p)}\big)\xi_j^{(p)}\big(1 - \xi_i^{(p)}\big)\xi_k^{(p)}B_{ij-}^{(p)}B_{ik-}^{(p)}\big] = p_{2,1}q_-^2. \end{aligned}$$

Thus,

$$\begin{aligned} r &= 1 - 2\alpha - 2\beta + p_{3,0}q_+^2 + p_{2,1}q_-^2 \\ b_{xyz}^{(p)} &= \big(1 - \rho_{xy}^{(p-1)}\big)\big(1 - \rho_{xz}^{(p-1)}\big)p_{3,0}q_+^2 + \rho_{xy}^{(p-1)}\rho_{xz}^{(p-1)}p_{2,1}q_-^2 \\ &\quad - \big[\big(1 - \rho_{xy}^{(p-1)}\big)\alpha - \rho_{xy}^{(p-1)}\beta\big]\big[\big(1 - \rho_{xz}^{(p-1)}\big)\alpha - \rho_{xz}^{(p-1)}\beta\big]. \end{aligned}$$

Since $\rho_{xy}^{(p)} \to \pi_1$ as $p \to \infty$ and $\pi_0\alpha = \pi_1\beta$, we have

$$\lim_{p\to\infty} b_{xyz}^{(p)} = \pi_0^2 p_{3,0}q_+^2 + \pi_1^2 p_{2,1}q_-^2 := b.$$

Furthermore from equation 3.2, it follows that $p_{3,0} < p_{2,0}$ and $p_{2,1} < p_{1,1}$ so that $|r| < 1$. Also since the coding levels are small (the support of $\mu(df)$ is bounded away from 1), clearly $0 < r < 1$. Consequently by the recursion A.3, it follows that

$$\gamma := \lim_{p\to\infty} \gamma_{xyz}^{(p)} = \frac{b}{1-r} = \frac{\pi_0^2 q_+^2 p_{3,0} + \pi_1^2 q_-^2 p_{2,1}}{2\alpha + 2\beta - p_{3,0}q_+^2 - p_{2,1}q_-^2}.$$

If at stage 0, the network is stationary, then the six different values of $\gamma_{xyz}^{(1)}$ are

| $\gamma_{111}^{(1)}$ | $\gamma_{110}^{(1)}$ | $\gamma_{100}^{(1)}$ | $\gamma_{011}^{(1)}$ | $\gamma_{010}^{(1)}$ | $\gamma_{000}^{(1)}$ |
|---|---|---|---|---|---|
| $(1-q_+)^2\gamma$ | $(1-q_+)\gamma$ | $\gamma$ | $(1-q_-)^2\gamma$ | $(1-q_-)\gamma$ | $\gamma.$ |

**Proof of Proposition 3.** Note that given $0 < s < 1$, if $G(u)$ is a CDF of a distribution on the unit interval, so are

$$\hat{G}^{(a,b)}(u) = G\left(\frac{u-b}{a-b}\right). \tag{A.4}$$

We first prove one identity. Let $0 \le a,\ b \le 1$. If $X \mid Y \sim \mathrm{Bin}(Y, a) + \mathrm{Bin}(N - Y, b)$ and $Y \sim \mathrm{BinMix}(N, G(\cdot))$, then $X \sim \mathrm{BinMix}(N, \hat{G}^{(a,b)})$.

The conditional characteristic function of $X \mid Y$ is $\mathsf{E}[e^{itX} \mid Y] = (e^{it}a + 1 - a)^{Y}(e^{it}b + 1 - b)^{N-Y}$. Thus, the unconditional characteristic function of $X$ is

$$\begin{aligned}
\mathsf{E}[e^{itX}] &= (e^{it}b + 1 - b)^N \mathsf{E}\left(\frac{e^{it}a + 1 - a}{e^{it}b + 1 - b}\right)^Y \\
&= (e^{it}b + 1 - b)^N \sum_{k=0}^{N} \int \binom{N}{k} \left(\frac{e^{it}a + 1 - a}{e^{it}b + 1 - b}\right)^k u^k (1-u)^{N-k} G(du) \\
&= (e^{it}b + 1 - b)^N \int \left[\left(\frac{e^{it}a + 1 - a}{e^{it}b + 1 - b}\right) u + 1 - u\right]^N G(du) \\
&= \int [e^{it}(b + (a-b)u) + (1 - b + (a-b)u]^N G(du) \\
&\quad (\text{let } z = b + (a-b)u) \\
&= \int [e^{it}z + 1 - z]^N \hat{G}^{(a,b)}(dz).
\end{aligned}$$

Assuming $J_{iA}^{(p-1)} \sim \mathrm{BinMix}(|A|, G_{p-1}(\cdot))$, then by proposition 2 and the identity above, conditional on $F^{(p)} = f$, we have

$$\begin{aligned}
J_{iA}^{p}|\xi_i^{(p)} &= 0 \sim \mathrm{BinMix}(|A|, \hat{G}_{p-1}^{(1-fq_-,0)}) \\
J_{iA}^{p}|\xi_i^{(p)} &= 1 \sim \mathrm{BinMix}(|A|, \hat{G}_{p-1}^{(1,fq_+)}).
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathsf{P}(J_{iA}^{p} = k) &= \int \mathsf{P}(J_{iA}^{p} = k|F^{(p)} = f)\mu(df) \\
&= \int (1-f)\mathsf{P}(J_{iA}^{p} = k|F^{(p)} = f, \xi_i^{(p)} = 0) \\
&\quad + f\mathsf{P}(J_{iA}^{p} = k|F^{(p)} = f, \xi_i^{(p)} = 1)\mu(df)
\end{aligned}$$

$$
\begin{aligned}
&= \int\!\!\int \binom{|A|}{k} z^k (1-z)^{|A|-k} \\
&\quad \times \left[ (1-f)\hat{\mathrm{G}}_{p-1}^{(1-fq_-,0)}(dz) + f\tilde{\hat{\mathrm{G}}}_{p-1}^{(1,fq_+)}(dz) \right] \mu(df) \\
&= \int \binom{|A|}{k} z^k (1-z)^{|A|-k} G_p(z)\, dz.
\end{aligned}
$$

The evolution equation, 4.1, has a unique fixed point given that the initial $G_1(\cdot)$ is the CDF of a probability distribution on the unit interval [0, 1]. That is, $G_1(u) = 0$ for $u \leq 0$ and $G_1(u) = 1$ for $u \geq 1$.

Let $H(z)$ and $G(z)$ be any CDFs on [0, 1] and $\mathcal{R}$(G) the functional defined in equation 4.1,

$$
\begin{aligned}
&\int_0^1 |\mathcal{R}(G)(z) - \mathcal{R}(H)(z)| dz \\
&\leq \int \left[ f \int_0^1 \left| G\left(\frac{z - fq_+}{1 - fq_+}\right) - H\left(\frac{z - fq_+}{1 - fq_+}\right) \right| dz + \right. \\
&\quad \left. (1-f) \int_0^1 \left| G\left(\frac{z}{1 - fq_-}\right) - H\left(\frac{z}{1 - fq_-}\right) \right| dz \right] \mu(df) \\
&= \lambda \int_0^1 |G(u) - H(u)| du,
\end{aligned}
$$

where the equality comes from change of variables $u = \frac{z}{1-fq_-}$ and $u = \frac{z-fq_+}{1-fq_+}$ and

$$
\lambda = \int [f(1 - fq_+) + (1-f)(1 - fq_+)] \mu(df) = 1 - p_{2,0}q_+ - p_{1,1}q_- < 1
$$

is as defined in equation 3.4. This proves the map $\mathcal{R}$ is contracting, and the claim follows from the Banach fixed point theorem.

## Acknowledgments

Y. A. is indebted to Sandro Romani for discussions on the role of inhibition. Y. A. is supported in part by NSF ITR DMS-0706816.

## References

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

Amit, D. J., & Mongillo, G. (2003). Spike-driven synaptic dynamics generating working memory states. *Neural Comp.*, *15*, 565–596.

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

Bernacchia, A., & Amit, D. J. (2007). Impact of spatiotemporally correlated images on the structure of memory. *Proc. Natl. Acad. Sci. U.S.A.*, *104*(9), 3544–3549.

Brunel, N., Carusi, F., & Fusi, S. (1998). Slow stochastic Hebbian learning of classes of stimuli in a recurrent neural network. *Network*, *9*, 123–152.

Bullier, J. (2001). Feedback connections and conscious vision. *Trends in Cognitive Sciences*, *5*, 369–370.

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (FMRI) "brain reading": Detecting and classifying distributed patterns of FMRI activity in human visual cortex. *NeuroImage*, *19*, 261–270.

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. *J. Physiol. Paris*, *97*(4–6), 659–681.

Fusi, S., & Abbott, L. F. (2007). Limits on the memory storage capacity of bounded synapses. *Nat. Neurosci.*, *10*, 485–93.

Gardner, E. (1986). Structure of metastable states in the Hopfield model. *Phys. A: Math. Gen.*, *19*, L1047–1052.

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*, 632–635.

Kveraga, K., Ghuman, A. S., & Bar, M. (2007). Top-down predictions in the cognitive brain. *Brain and Cognition*, *65*(2), 145–168.

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

Nadal, J. (1991). Associative memory: On the (puzzling) sparse coding limit. *J. Phys. A Math. Gen.*, *24*, 1093–1101.

Nadal, J.-P., & Toulouse, G. (1990). Information storage in sparsely coded memory nets. *Network*, *1*, 61–74.

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

Senn, W., & Fusi, S. (2005). Convergence of stochastic learning in perceptrons with binary synapses. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, *71*(6 Pt. 1), 061907.

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2008). Stimulus-specific delay activity in human primary visual cortex. *Psychological Science*, *20*, 207–214.

Tsodyks, M. (1990). Associative memory in neural networks with binary synapses. *Modern Physics Letters B*, *4*, 713–716.

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

Received February 20, 2009; accepted June 1, 2009.