Systems/Circuits

Long-Term Memory Stabilized by Noise-Induced Rehearsal

Yi Wei and Alexei A. Koulakov
Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724

Cortical networks can maintain memories for decades despite the short lifetime of synaptic strengths. Can a neural network store long-lasting memories in unstable synapses? Here, we study the effects of ongoing spike-timing-dependent plasticity (STDP) on the stability of memory patterns stored in synapses of an attractor neural network. We show that certain classes of STDP rules can stabilize all stored memory patterns despite a short lifetime of synapses. In our model, unstructured neural noise, after passing through the recurrent network connections, carries the imprint of all memory patterns in temporal correlations. STDP, combined with these correlations, leads to reinforcement of all stored patterns, even those that are never explicitly visited. Our findings may provide the functional reason for irregular spiking displayed by cortical neurons and justify models of system memory consolidation. Therefore, we propose that irregular neural activity is the feature that helps cortical networks maintain stable connections.

Introduction

Changing synaptic strengths is widely regarded as the mechanism by which long-term memory is encoded and stored in the brain (Martin et al., 2000). Long-term potentiation (LTP) and long-term depression (LTD) of synaptic conductances exhibit many features that make them candidates for the cellular mechanism of memory storage. By correlating presynaptic and postsynaptic activities, LTP/LTD can implement Hebbian plasticity that is at the basis of many learning and memory models (Abbott and Nelson, 2000). Because LTP is observed in many preparations, including freely behaving animals (Whitlock et al., 2006), and in many brain regions (Cooke and Bliss, 2006), it matches the description of the basic mechanism of learning and memory.

To fully satisfy the requirements for memory mechanism, the persistence of synaptic changes induced by LTP/LTD has to be reconciled with persistence of memory traces. Fundamentally, it is not clear whether these two timescales have to match. Although memories can be stored by the brain for dozens of years, the lifetime of LTP appears to be shorter. In hippocampal slice preparations, the persistence of synaptic changes is limited by several hours (Reymann et al., 1985), whereas in most cases *in vivo*, synaptic changes can last 4–5 weeks (Shors and Matzel, 1997; Abraham, 2003). In rare instances, synaptic changes can last for approximately a year; however, these examples require special conditions (Abraham et al., 2002). This occurs despite the observation that at least some components of consolidated long-term memory can be attributed to the hippocampal complex (Nadel and Moscovitch, 1997, 2001). In the cortex, LTP has not been demonstrated to last beyond a period of several weeks (Trepel and Racine, 1998; Ivanco and Racine, 2000). Although changes in structural connectivity can persist for more than a month (Grutzendler et al., 2002; Trachtenberg et al., 2002; Knott et al., 2006; Alvarez and Sabatini, 2007; Fu and Zuo, 2011), they may reflect ongoing changes in sensory inputs rather than carry memory traces. The same applies to other examples of cortical plasticity observed after sensory deprivation (Feldman, 2009). Whether synaptic strengths can persist throughout the lifetime is an open question. Cascade synaptic models (Fusi et al., 2005), for example, propose that individual synapses contain long-lasting internal states that are not directly related to synaptic strength. An alternative explanation is that robust long-term memories can somehow be maintained for decades without the requirement of stable synapses. This hypothesis is investigated here.

Here, we propose a mechanism for persistent memory storage that uses short-lived synapses. In our model, long-term memory can be stored in the network for a very long time despite a short time constant of LTP persistence. To be preserved, memory states do not have to be revisited. We analyze a simple mathematical model for attractor neural network, which includes several realistic elements such as stochastic neural noise, short synaptic lifetime, and ongoing synaptic plasticity described by spike-timing-dependent plasticity (STDP). The network activity resides near a set of states that represent activity relevant to its current environment. The average activity of neurons samples only these current memory states. However, we demonstrate that, because of the presence of noise, the correlations in neural activity carry imprints of all memory traces, including old ones. These correlations, under carefully chosen conditions, can allow the old traces to be rehearsed and maintained by the network even though they are not explicitly visited. We thus propose that old memory states can be reinforced by rehearsal, even though these memories are never visited or accessed. Because our rehearsal mechanism does not involve explicit reactivation of old memories, we call the proposed mechanism "implicit rehearsal." We show that for implicit rehearsal to be effective, STDP rules must satisfy certain strict conditions. This mechanism will work

Received Aug. 17, 2012; revised Sept. 4, 2014; accepted Sept. 28, 2014.

Author contributions: Y.W. and A.A.K. designed research; Y.W. and A.A.K. performed research; Y.W. and A.A.K. wrote the paper.

This work was supported by the Swartz Foundation and the National Institutes of Health (Grants NIH R01MH092928 and NIH R01DA036913).

The authors declare no competing financial interests.

Correspondence should be addressed to Alexei Koulakov, Cold Spring Harbor Laboratory, 1 Bungtown Rd., Cold Spring Harbor, NY 11724. E-mail: akula@cshl.edu.

DOI:10.1523/JNEUROSCI.3929-12.2014

with antisymmetric STDP that is often observed (Bi and Poo, 2001; Froemke and Dan, 2002), but does not work with the symmetric, non-negative form of LTP. We show therefore that neural noise combined with synaptic plasticity can lead to stability of old memory traces despite individual synapses being unstable. Our model has experimentally testable predictions.

Materials and Methods

Description of the model. Let N be the total number of neurons. We will assume that there are p patterns represented by N-dimensional vectors, the elements of which are ±1:

$$p_i^a = \pm 1,\ a = 1, \ldots, p,\ i = 1, \ldots, N. \tag{1}$$

Different patterns are chosen to be orthogonal to each other as follows:

$$\frac{1}{N}\,\mathbf{p}^{aT} \cdot \mathbf{p}^b = \delta^{ab}. \tag{2}$$

Equation 2 allows us to define p projection operators as follows:

$$P_{ij}^a = \frac{1}{N}\,p_i^a p_j^a, \ \text{ so that } \mathbf{P}^a \cdot \mathbf{P}^b = \delta^{ab}\,\mathbf{P}^a. \tag{3}$$

When projection operators number a are applied to a given activity vector, they result in an activity specific to the given pattern a.

Memories about the patterns are stored in a synaptic weight matrix using the conventional learning rule (Dayan and Abbott, 2001), as follows:

$$W_{ij}(t) = \sum_{a=1}^{p} c_a(t)\ P_{ij}^a, \tag{4}$$

where the set of coefficients, $c_a(t)$, represents the strengths of individual patterns.

In our model, the input current and firing rate of neurons i, $u_i(t)$, and $f_i(t)$, are related through the activation function F as follows:

$$f_i(t) = F(u_i). \tag{5}$$

Input currents are described by the following equation:

$$\tau\,\frac{d\,\mathbf{u}(t)}{dt} = -\,\mathbf{u}(t) + \mathbf{W}\,\mathbf{f}(t) + \boldsymbol{\xi}(t). \tag{6}$$

In Equation 6, τ is a constant that determines how rapidly the current varies and $\xi(t)$ is the Gaussian random white noise. We assume that $\xi(t)$ has the following properties:

$$\langle \xi_i(t) \rangle = 0,\ \langle \xi_i(t)\xi_j(t') \rangle = \xi^2 \delta_{ij}\delta(t - t') \tag{7}$$

where $\langle \ldots \rangle$ denotes the average over noise ensemble. Subsequently, we assume that the amplitude of noise, ξ, is very small and we use this fact along with the short timescale of ξ to treat noise as a perturbation.

Plasticity in the network is defined by spike-timing-dependent learning rules. For a pair of cells, the strength of synapses is updated with a rate of update that is dependent on presynaptic and postsynaptic activities as follows:

$$\tau_0\,\frac{dW_{ij}}{dt} = -\,W_{ij} + \gamma\int_{-\infty}^{t} dt_1 \langle f_i(t_1)\,K(t_1 - t)\,f_j(t)\rangle + \gamma\int_{-\infty}^{t} dt_2\,\langle f_i(t)\,K(t - t_2)\,f_j(t_2)\rangle. \tag{8}$$

Here, $f_i(t)$ is the firing rate of neuron number i at time t, and γ is the learning rate. Three terms in the r.h.s. of this equation describe the decay of synaptic strength with time and the modification due to presynaptic and postsynaptic firing, respectively. The relationship between learning in nonstationary rate-based model (Equation 8) and pairwise STDP in spiking models has been studies by Kempter et al. (1999). The STDP kernel, $K(\Delta t)$, in our model contains two components: short-range and long-range, $K(\Delta t) = K_s(\Delta t) + K_l(\Delta t)$. We assume that the short-range component varies within the timescale of several hundred milliseconds, which is defined as follows:

$$K_s(\Delta t) = \begin{cases} A_+\exp(\Delta t/\tau_+) & \Delta t < 0 \\ A_-\exp\,(-\Delta t/\tau_-) & \Delta t \geq 0. \end{cases} \tag{9}$$

The long-range STDP kernel, $K_l(\Delta t)$, is needed in our model to constrain the overall magnitude of firing rates and can originate from metabolic and other constraints. We assume that it varies very slowly on the timescales of the order of hours or more. The only constraint on $K_l(\Delta t)$ that is important in our model is that it makes the integral of the entire STDP kernel over time positive (i.e., $\int_{-\infty}^{\infty} K(t)dt > 0$) as discussed in more detail following Equation 14.

The STDP rule (Equation 8) can also be rewritten in an equivalent integral form as follows:

$$W_{ij}(t) = \frac{\gamma}{\tau_0}\int_{-\infty}^{t} dt_1 \int_{-\infty}^{t} dt_2\ e^{-\frac{t-\max(t_1-t_2)}{\tau_0}}\ \langle f_i(t_1)\,K(t_1 - t_2)\,f_j(t_2)\rangle. \tag{10}$$

This equation shows that τ_0 defines the forgetting time constant. The old memory is expected to decay after this time with the exception of memory that is rehearsed (i.e., relearned within the timescale τ_0). This rehearsal process is the topic of our present study.

Due to random noise, $\xi(t)$, the input currents fluctuate near constant values $\mathbf{u}(t) = \mathbf{u} + \delta\mathbf{u}(t)$. We assume that fluctuations are weak and can be treated as small perturbations (please see discussion after Equation 15 for the justification of this assumption). Therefore, by Equation 5, the firing rates fluctuate around stationary rates as follows:

$$\mathbf{f}(t) \approx \mathbf{f} + g\delta\mathbf{u}(t) \tag{11}$$

where $g = F'(u_i)$. Because neural noise is short range, its timescales are measured in milliseconds and we can decompose the dynamics of the system into two components: the fast-changing component, associated with noise, and the slowly varying component, determined by Hebbian learning. These two components are represented by two terms in Equation 11. The equation for fast-changing component is as follows:

$$\tau\,\frac{d\ \delta\mathbf{u}(t)}{dt} = -\delta\mathbf{u}(t) + g\mathbf{W}\ \delta\mathbf{u}\ (t) + \xi(t). \tag{12}$$

Our goal is to derive the contribution of the fast-changing component (i.e., noise) to the slowly varying component. This interplay could be interpreted as rehearsal. In the subsequent discussion, we treat noise as a small perturbation to the firing rates of the network; that is, we will assume that ξ^2 is small. The detailed conditions for the validity of this approximation can be found in the subsection below titled “Validity of approximations made in this study.”

For the given set of network weights, we assume that there are two types of attractors. One set of attractors is never visited by the network. We call these states implicit. The other set of attractors is explicitly visited by the system. Because our main goal is to consider the dynamics of implicit attractors (i.e., ones that are never visited), for simplicity, we assume that the explicit attractors are represented by only one attractor. Here, we discuss briefly the explicit attractor state and its stability with respect to learning.

Let us assume that the explicit attractor has an index $a = 1$. The stationary firing rates associated with this state are proportional to the pattern $\mathbf{p}^1$; that is, $\mathbf{f} = b\mathbf{p}^1$. Here, b is a constant determined by the function F (we assume that $F(x) = -F(-x)$; e.g., F is the sigmoid function). Assuming that the effects of noise are negligible, one can obtain the stationary value of the weight matrix that results from the explicit attractor. This contribution is present by virtue of the explicit attractor relearning, itself, through the STDP rules and could be viewed as resulting from explicit rehearsal (i.e., rehearsal of patterns that are currently in working memory). For this type of rehearsal, noise is not necessary. Equation 10 allows us to determine this component of the weight matrix as follows:

$$W_{ij} = c_1 P^1_{ij} + \delta W_{ij},$$

$$c_1 = \gamma b^2 N(A_+\tau_+ + A_-\tau_- + \Delta). \tag{13}$$

where $\delta W_{ij} = \sum_{a\neq 1} c_a P^a_{ij}$ is the component of the weight matrix as determined by other (implicit) patterns. From Equations 5 and 13, factor b is determined by the self-consistent equation as follows:

$$F\left[\gamma b^3(A_+\tau_+ + A_-\tau_- + \Delta)\right] = b. \tag{14}$$

Note that Equation 14 depends only on the total integral of the kernel. The temporal details of STDP do not affect the equation. Component $\delta W \to 0$ is in the network without noise according to Equation 10. Our goal here is to determine the behavior of δW (i.e., the contribution of implicit patterns that are never explicitly visited) in the network with noise.

The parameter $\Delta \equiv \int_{-\infty}^{\infty} K_l(t)dt$ makes the integral of STDP kernel positive so that $\int_{-\infty}^{\infty} K(t)dt = A_+\tau_+ + A_-\tau_- + \Delta > 0$. Because this is the only point at which the long-range STDP kernel $K_l(t)$ enters our model, we will not discuss this kernel further. From this point on, by STDP kernel we imply the short-range kernel, $K_s(t)$, that varies on the timescales of hundreds of milliseconds and is usually measured in LTP/LTD experiments (Abbott and Nelson, 2000).

According to Equation 10, without noise, the weight matrix contains only pattern P^1 with stationary strength c_1. All other coefficients $c_{a\neq 1}$ are zero. With Gaussian white noise included in Equation 6, other patterns (implicit) are represented in the activity of the network. Therefore, these patterns are present in the synaptic weight matrix $\delta W(t)$. Our goal is to find how coefficients $c_{a\neq 1}(t)$ evolve with time in this case.

To accomplish this goal, we consider noise a small perturbation and use the perturbation theory using the amplitude of noise, ξ^2, as a small parameter. Noise-induced firing rate fluctuation, $\delta\mathbf{u}(t)$, in the direction of pattern number, a, is $\delta\mathbf{u}^a(t) = \mathbf{P}^a \cdot \delta\mathbf{u}(t)$ and the corresponding component of random inputs is $\boldsymbol{\xi}^a(t) = \mathbf{P}^a \cdot \boldsymbol{\xi}(t)$. By applying projector $\mathbf{P}_a$, defined in Equation 3, to both sides of Equation 12, and because c_a changes much slower than $\delta\mathbf{u}(t)$, we find the following:

$$\delta\mathbf{u}^a(t) = \frac{1}{\tau}\int_{-\infty}^{t} dt' e^{-\frac{t-t'}{\tau}(1-gc_a)}\boldsymbol{\xi}^a(t'). \tag{15}$$

Next, we derive the condition for the validity of perturbation theory (i.e., the amplitude of fluctuations due to noise is smaller than the zero-th order solution obtained without noise). By choosing ξ to be small, we can make δu_i much smaller than u_i to allow us to treat noise as a perturbation in Equation 11. We give the detailed conditions for the validity of perturbation calculations in the subsection below titled "Validity of approximations made in this study" (Equation 30).

By Equations 3 and 7, we have the following:

$$\langle \boldsymbol{\xi}^a_i(t_1)\boldsymbol{\xi}^b_j(t_2)\rangle = \xi^2\, P^a_{ij}\delta^{ab}\delta(t_1 - t_2). \tag{16}$$

Using Equations 7 and 15, we obtain the average correlation function of fluctuations as follows:

$$\langle \delta u^a_i(t_1)\delta u^b_j(t_2)\rangle = \frac{\xi^2\delta^{ab}}{2\tau N}\frac{1}{1-gc_a}e^{-\frac{|t_1-t_2|}{\tau}(1-gc_a)}p^a_i p^a_j \tag{17}$$

Substituting Equation 11 and 13 into Equation 8 and using the correlation function in Equation 17, the fluctuating part of Equation 8 gives the equations that describe the dynamics of "unused" components of the weight matrix ($a = 2, \ldots, p$):

$$\tau_0\frac{dc_a}{dt} = -c_a + \frac{1}{1-gc_a}\left(\frac{A'_+}{\frac{1-gc_a}{\tau}+\frac{1}{\tau_+}} + \frac{A'_-}{\frac{1-gc_a}{\tau}+\frac{1}{\tau_-}}\right). \tag{18}$$

Coefficients $c_a(t)$ are defined in Equations 46 and 4. Here, we defined the parameters of the short range STDP kernel as follows:

$$A'_\pm = \frac{\gamma g^2\xi^2}{2\tau}A_\pm \tag{19}$$

Equation 18 describes how the strength of each unused pattern's representation in the network weights changes over time when neurons receive random noise. This equation is the main result of this study. The dependence of the right side of Equation 18 as a function of c_a is shown in Figures 2B and 4B. It is evident from Equation 18 that $c_a(t) = 1/g$ is a critical value at which the equation for $c_a(t)$ becomes singular.

The equation of evolution for the first (explicit) pattern is as follows:

$$\tau_0\frac{dc_1}{dt} = -c_1 + \frac{1}{1-gc_1}\left(\frac{A'_+}{\frac{1-gc_1}{\tau}+\frac{1}{\tau_+}} + \frac{A'_-}{\frac{1-gc_1}{\tau}+\frac{1}{\tau_-}}\right) + \gamma bh\,\frac{\xi^2}{2\tau}(A_+\tau_+ + A_-\tau_- + \Delta)\sum_{a=1}^{p}\frac{1}{1-gc_a}$$

where c_1 satisfies the self-consistency equation $F[bc_1] = b$ and $h = F''(u_i)$ is the second derivative of the activation function.

Firing rate model simulation. In the simulations, we construct a number of random patterns such that their elements are independent and take the value +1 or −1 with equal probability. These patterns are the memories that we store in the network. From these patterns, we choose an arbitrary one to be the explicit pattern that is stored in the network and constantly visited. Network parameters are chosen according to the bistability conditions given in the Results section. The explicit form of action function F is not important because all we need is its first-order derivative and value a specific point; for example, we can choose it to be a power function.

At each time step, we first generate random Gaussian white noise for each neuron. Then, using Equation 12, we calculate the changes to the fluctuating part of the input current ($\delta\mathbf{u}(t)$) due to noise. After updating $\mathbf{f}(t)$, we use Equation 12 to calculate the new firing rates $\mathbf{f}(t)$. Synaptic weights are updated according to Equation 8 but without averaging over noise.

Simulation consists of two phases. In the first phase, we prepare the network. We start with a synaptic weight matrix that contains only the explicit pattern with an arbitrary strength. Then, at each step, the weight matrix $\mathbf{W}(t)$ is updated according to Equation 8 and we stop when it stabilizes. The strength of the explicit pattern can also be read from the weight matrix. In the second phase, we test our results in the main text. First, introduce the implicit pattern(s) to the network. Then, let the network evolve according to Equation 8 and record the strength of patterns at each time step. We found that if the initial strength of the implicit pattern is set below a certain value (the transition point in Fig. 4), the implicit pattern decays with time. Conversely, if we set the initial strength of the implicit pattern above the transition point, then the implicit pattern is kept in the network for a long time. In this case, the strength of implicit pattern fluctuates around certain value (the second stable point in Fig. 4) above the transition point. In both cases, the strength of the explicit pattern never decays but fluctuates around its initial value, which is found when we prepare the network in the first phase. These observations are in good agreement with our model prediction (i.e., Equation 8 and Figure 4). Another test is whether when we turn off the noise, the implicit pattern always decays. This also agrees with our analysis in the main text.

In the simulations, we find that increasing τ_0 improves the performance of our model; that is, the implicit patterns are maintained for a longer time. This is in agreement with the analysis in the subsection below titled "Validity of approximations made in this study"; that is, in this limit, the mean field approximation (MFA) works better and therefore the strength of the implicit pattern is kept at the second stable point. However, to keep the running time of simulations feasible, we could not have τ_0 too large. In the simulations, we chose $\tau = 5$ ms, $\tau_+ = 50$ ms, $\tau_- = 100$ ms, and $\tau_0 = 2 \times 10^5$ ms. Other parameters are indicated in the captions of appropriate figures.

Validity of approximations made in this study. To analyze the behavior of our model, we used primarily the method based on analytical calculations. This method includes derivation of the results in the closed form that can be understood without the use of a computer. Therefore, our main result, Equation 18, describes the learning dynamics of the implicit component of memory and can be analyzed for various sets of parameters without computer simulations. The advantage of this method is that the dynamics of network weights can be understood without the limits on the network size and on the parameters used. To obtain these results, however, some approximations had to be made. Below, we derive the conditions under which our approximations can be considered valid and the effects of them are under control. Briefly, we assumed that the amplitude of fluctuations induced by noise is small, which allowed us to use an approximation called perturbation theory (Equation 11). The effects of noise on the network weights can still be large, however, because they are accumulated over time. Below, in this section of the Materials and Methods (Equation 30), we show that parameters of the model can be chosen so that both perturbation theory is valid and bistability of network weights exist, as described in Figure 4. We show, for example, that perturbation theory is valid for large firing rates, large neuronal gains, or large STDP time windows τ_+ and τ_-. The second approximation used by us is the MFA (Equation 8). MFA is often used in the network theory and has allowed to derive many important results, such as the memory capacity of the Hopfield model (Hertz et al., 1991). In the context of our model, MFA means that the instantaneous values of activity correlations entering STDP rules can be replaced by their average values. Below, we derive the conditions under which MFA is valid by analyzing the effects of relaxing this assumption. We show that MFA is accurate if the STDP time windows, τ_+ and τ_-, are substantially smaller than synaptic strength lifetime, τ_0 (Equation 31). Because the former set of timescales is approximately hundreds of milliseconds, whereas the latter is measured in weeks, this condition appears to be well valid in reality, thus motivating the use of MFA in our calculations. This comparison also discloses the challenges faced by realistic computer simulations in this setting. Because computer models have to integrate both millisecond neuronal timescales and long-term behaviors of the network lasting years, such simulations are challenging to even modern computers, especially because network size has to be kept large. Despite these challenges, we succeeded in reproducing computationally the predicted behavior of networks in keeping implicit memory states stable (Figs. 7, 8, 9). Cortical networks, however, can easily overcome these challenges due to their inherent parallelism and access the range of parameters only available in our analytical calculations.

Validity of perturbation theory approximation. In our study, we considered fluctuations induced by noise to be small (Equation 11). This means that noise was considered a small perturbation; that is, within a perturbation theory. In this section, we discuss the validity of this approximation. Although we used this approximation (perturbation theory) to solve equations of our model, our mechanism may take place even when the equations cannot be solved using this method.

More precisely, smallness of the amplitude of noise was needed when we used Taylor expansion around the value u_i in Equation 11. This equation does not include the second-order term $F''(u_i)\delta u_i^2/2$. This approximation is valid if:

$$|\delta u_i| \ll \left.\frac{F'(u)}{F''(u)}\right|_{u=F^{-1}(b)} \tag{20}$$

where b is given by Equation 14. Because $\delta u_i \sim \xi_i$, this condition imposes a constrain on the noise amplitude ξ. To see this, solving Equation 12, we get the following:

$$\delta\mathbf{u}(t) = \frac{1}{\tau}\int_{-\infty}^{t} dt' e^{-\frac{t-t'}{\tau}(1-g\mathbf{W})}\boldsymbol{\xi}(t'). \tag{21}$$

From this and Equation 8, we find the following:

$$\langle \delta u_i^2(t) \rangle = \frac{\xi^2}{\tau^2}\int_{-\infty}^{t} dt' \left[e^{-\frac{2(t-t')}{\tau}(1-g\mathbf{W})} \right]_{ii}. \tag{22}$$

As follows from this equation, this quantity averaged over neurons is as follows:

$$\overline{\langle \delta u^2(t) \rangle} = \frac{1}{N}\sum_i^N \langle \delta u_i^2(t) \rangle = \frac{1}{N}\frac{\xi^2}{\tau}\sum_i^N \frac{1}{2(1-gc_i)} \tag{23}$$

Here, c_i is the i-th eigenvalue of matrix W. Because there is only a small number of patterns with finite corresponding c_i, and most of the eigenvalues c_i are close to zero, we have the following:

$$\overline{\langle \delta u^2(t) \rangle} \approx \frac{\xi^2}{2\tau}. \tag{24}$$

Combining this with Equation 22, we find the condition for perturbation calculation to be valid is as follows:

$$\frac{\xi}{\sqrt{\tau}} \ll \left.\frac{F'(u)}{F''(u)}\right|_{u=F^{-1}(b)}. \tag{25}$$

The amplitude of noise is therefore limited by $\xi^2 \ll \tau(F'(u)/F''(u))^2_{u=F^{-1}(b)} \sim \tau u^2 \sim \tau f^2/g^2$. Here, u is a typical value of membrane voltage and f is the typical value of the firing rates. Therefore, the levels of noise have to be sufficiently low for the perturbation theory analysis to be valid.

For bistability, we need the value of noise to be larger than a certain threshold. The detailed conditions for this criterion are described in section titled "Conditions of bistability." Therefore, our analysis can be used when the level of noise is big enough for the bistability to exist and small enough for the Taylor expansion in Equation 11 to be valid. Can such a regime exist? Here, we will provide simple estimate for the existence of such a window of parameters. The perturbation theory is valid if noise is weak; that is:

$$\xi^2 \ll \tau f^2/g^2. \tag{26}$$

As follows from the discussion in this study, the bistability exists if, loosely speaking, the learning rate is sufficiently strong; that is:

$$\gamma g^3 \xi^2 \gg \frac{1}{A_\pm}\left(\frac{\tau}{\tau_\pm}\right)^2. \tag{27}$$

Both conditions can be satisfied, if the amplitude of noise ξ^2 lies within the range defined as follows:

$$\frac{1}{\gamma g^3 A_\pm}\left(\frac{\tau}{\tau_\pm}\right)^2 \ll \xi^2 \ll \frac{\tau f^2}{g^2} \tag{28}$$

This range exists if the boundaries for the rage differ in the correct direction; that is:

$$\frac{1}{\gamma g^3 A_\pm}\left(\frac{\tau}{\tau_\pm}\right)^2 \ll \frac{\tau f^2}{g^2}. \tag{29}$$

which implies the following:

$$\frac{\tau}{\tau_\pm^2} \ll \gamma A_\pm f^2 g \tag{30}$$

Therefore, if the learning rate $\gamma A_\pm$ is sufficiently big, both perturbation theory analysis (Taylor series expansion in Equation 11) is valid and bistability necessary for our mechanism exists. This occurred because the firing rate equations and, consequently, Taylor series expansion, do not depend on the learning rates. Therefore, learning rates can be used as an independent parameter to reach the conditions of bistability. In addition, the effects of noise can be big even though we assume weak noise in the perturbation theory. This is because our assumption of the weakness of noise only includes the validity of Taylor series expansion. Therefore, the overall impact of noise can be substantial despite its small amplitude.

Validity of the MFA. In this section, we show that the MFA calculations presented above in Materials and Methods are justified. In Equations 8 and 10, we assumed that the learning rates are determined by the averages of the firing rates over the ensemble of noise. In reality, these equations should be used without such averaging. To derive our results, we therefore used an approximation that could be called the MFA. At what condition can the instantaneous values of the pairwise products of the firing rates be replaced by their correlations? Below, we will show that this condition is determined by the timescale of synaptic modifications. In particular, it is determined by the time constant of synaptic decay τ_0. It is this timescale that determines the duration of time over which the firing rates are averaged in Equations 8 and 10). We will show that when the duration of STDP learning kernels $\tau_\pm$ (Equation 9) is much smaller than the forgetting timescale; that is: $\tau_\pm \ll \tau_0$, the MFA can accurately describe the behavior of the network. Because, in reality, the STDP learning kernel lasts ~100 ms whereas the forgetting timescale extends over several weeks, $\tau_0 \sim 10^9$ ms, the variance of the deviations from the mean field values are small, as follows:

$$\frac{\overline{(c - c_{MF})^2}}{c_{MF}^2} = \frac{\overline{\delta c^2}}{c_{MF}^2} \sim \frac{\tau_\pm}{\tau_0} \sim 10^{-7}. \tag{31}$$

This estimate argues that the MFA a valid method.

To derive the condition in Equation 31, we start from Equation 8. Without averaging of noise, Equation 8 has the following form:

$$\tau_0 \frac{dW_{ij}}{dt} = -W_{ij} + \gamma \int_{-\infty}^{t} dt_1 f_i(t_1) K(t_1 - t) f_j(t) + \gamma \int_{-\infty}^{t} dt_2 f_i(t) K(t - t_2) f_j(t_2). \tag{32}$$

Let $c_a(t)$ be the strength of implicit pattern a. In the main text, where we used MFA analysis, the equation for c_a is given in Equation 18. The quantity described by that equation will be called $c_{MF}(t)$. Here, we are interested in the difference between the mean field result and the result without averaging. To make notations simpler, we will omit the subscript a in the remaining part of this section and it is understood that our calculation is about a certain implicit patter a the strength of which is c.

By projecting Equation 32 onto state a using operator $\mathbf{P}^a$, we obtain the following:

$$\tau_0 dc_a/dt + c_a = A(t) \tag{33}$$

where $A(t)$ is given by the following:

$$A(t) = \gamma g^2 \int_{-\infty}^{t} dt_1\, u(t_1) K(t_1 - t) u(t) + \gamma g^2 \int_{-\infty}^{t} dt_2 u(t) K(t - t_2) u(t_2). \tag{34}$$

where:

$$u(t) = N^{-1/2} \sum_i p_i^a \delta u_i(t) \tag{35}$$

is the projection of the membrane voltage onto state a. This quantity can be related to a Gaussian variable describing noise as follows:

$$u(t) = \frac{1}{\tau} \int_{-\infty}^{t} dt' e^{-\frac{t-t'}{\tau}(1-gc)} \xi(t') \tag{36}$$

where:

$$\xi(t) = N^{-1/2} \sum_i p_i^a \xi_i(t) \tag{37}$$

Here, it is easy to see that $\langle \xi(t) \rangle = 0$ and $\langle \xi(t)\xi(t') \rangle = \xi^2 \delta(t - t')$. It is also direct to show that $\langle u(t) \rangle = 0$ and:

$$\langle u(t)u(t') \rangle = \frac{\xi^2}{2\tau} \frac{1}{1 - gc} e^{-\frac{1-gc}{\tau}|t-t'|} \tag{38}$$

From Equation 18, we know that the following is true:

$$A_{MF} = \langle A(t) \rangle = \frac{1}{1 - gc} \left(\frac{A'_+}{\frac{1-gc}{\tau} + \frac{1}{\tau_+}} + \frac{A'_-}{\frac{1-gc}{\tau} + \frac{1}{\tau_-}} \right) \tag{39}$$

To determine how well we can approximate $A(t)$ by A, we need to calculate the variance of $A(t)$ as follows:

$$\langle (A(t) - A_{MF})(A(t') - A_{MF}) \rangle = \langle A(t)A(t') \rangle - A_{MF}^2 \tag{40}$$

In the calculation, we use the fact, which follows from the properties of Gaussian white noise, that:

$$\begin{aligned} &\langle u(t_1)u(t_2)u(t_3)u(t_4) \rangle \\ = \; &\langle u(t_1)u(t_2) \rangle \langle u(t_3)u(t_4) \rangle + \langle u(t_1)u(t_3) \rangle \langle u(t_2)u(t_4) \rangle \\ &+ \langle u(t_1)u(t_4) \rangle \langle u(t_2)u(t_3) \rangle. \end{aligned} \tag{41}$$

By straightforward calculations using Equations 34 and 41, we find that:

$$\frac{\langle (A(t) - A_{MF})(A(t') - A_{MF}) \rangle}{A_{MF}^2} = e^{-\frac{1-gc}{\tau}|t-t'|} \left(c_+ e^{-\frac{|t-t'|}{\tau_+}} + c_- e^{-\frac{|t-t'|}{\tau_-}} + c_0 e^{-\frac{1-gc}{\tau}|t-t'|} \right). \tag{42}$$

Here c_+, c_- and c_0 are all functions of $A_\pm$, $\tau_\pm$, τ, and g. To simplify the results, we define the following three variables as follows:

$$t_\pm = \frac{\tau_\pm}{\tau}(1 - gc) \text{ and } n = \left(\frac{A_+\tau_+}{1 + t_+} + \frac{A_-\tau_-}{1 + t_-} \right)^{-2}$$

With $t_\pm$ and n, different terms in Equation 42 can be written as follows:

$$c_+ = n \left\{ t_+ \left(1 + \frac{2}{1 + t_+} \right) \frac{A_+^2 \tau_+^2}{t_+^2 - 1} + \frac{A_+\tau_+}{t_+^2 - 1} \frac{A_-\tau_-}{1 + t_-} \times \left[2 \left(1 + \frac{\tau_+}{\tau_+ + \tau_-} \right) t_+ + \frac{2\tau_-}{\tau_+ + \tau_-} t_+^2 \right] \right\}$$

$$c_- = c_+(A_+ \leftrightarrow A_-,\ \tau_+ \leftrightarrow \tau_-,\ t_+ \leftrightarrow t_-)$$

$$c_0 = n \left\{ 2 \frac{A_+^2 \tau_+^2}{1 - t_+^2} - 2 \frac{A_+\tau_+}{t_+ - 1} \frac{A_-\tau_-}{1 + t_-} \right\} + (A_+ \leftrightarrow A_-,\ \tau_+ \leftrightarrow \tau_-,\ t_+ \leftrightarrow t_-)$$

From our analysis, we know that, near the second stable point, $t_\pm$ are both of order 1; that is, $t_\pm \sim O(1)$. Therefore, c_+, c_-, and c_0 are all of the order of 1.

From the previous discussion, we can write: $A(t) = A_{MF} + \delta A(t)$, such that $\langle \delta A(t) \rangle = 0$. To estimate $\delta A(t)$, notice the facts that τ is approximately a few milliseconds, $\tau_\pm$ are approximately a few hundred milliseconds, and τ_0 is approximately a few weeks; that is: $\tau \ll \tau_\pm \ll \tau_0$. By Equation 42, we have the following:

$$\langle \delta A(t)\delta A(t')\rangle \sim \tau_{\pm} A_{MF}^{2}\delta(t-t') \tag{43}$$

We can now write $c(t) = c_{MF}(t) + \delta c(t)$ such that $\tau_0 dc_{MF}/dt = -c_{MF} + A_{MF}$ and $\tau_0 d\delta c/dt = -\delta c(t) + \delta A(t)$. The first equation leads the mean-field solution that is presented in the main text. The second equation describes the fluctuations around the mean field results. Solving the second equation, we get the following:

$$\delta c(t) = \frac{1}{\tau_0}\int_{-\infty}^{t} dt' e^{-\frac{t-t'}{\tau_0}}\delta A(t'). \tag{44}$$

From Equation 43, we find the following:

$$\langle \delta c(t)\delta c(t')\rangle \sim \frac{\tau_{\pm}A_{MF}^{2}}{\tau_0} \sim \frac{\tau_{\pm}}{\tau_0}c_{MF}^{2} \tag{45}$$

from which Equation 31 follows directly. If we choose the synaptic decay time τ_0 to be 2 weeks, then $\tau_0 \sim 10^9$ms and the STDP window $\tau_{\pm}$ is a few hundred milliseconds; for example: $\tau_{\pm} \sim 100\ ms$ by Equation 45, we have $\delta c/c_{MF} \sim \sqrt{\tau_{\pm}/\tau_0} \sim 10^{-3}$; that is, the correction δc to the mean field solution A_{MF} is very small. This proves that we can approximate $c(t)$ with $c_{MF}(t)$ and the MFA calculations above (e.g., Equation 18), are indeed valid approximations.

Results

Patterns of neural activity stored in network weights correspond to network attractors

In this study, we analyze attractor neural networks with features similar to the continuous Hopfield model (Hopfield, 1984; Hertz et al., 1991). Such networks can exhibit two types of memory: long-term, contained in the recurrent network weights, and working memory, contained in the firing rates of neurons (Amit, 1989).

The network can store the long-term memory of a set of patterns in prespecified network weights (Hertz et al., 1991). If activity of a neuron number i that is associated with pattern number a is p_i^a, then, as within the conventional Hopfield model, the connection strength between two neurons i and j is given by the Hebbian-like learning rule:

$$W_{ij}(t) = \frac{1}{N}\sum_{a=1}^{p} c_a p_i^a p_j^a \tag{46}$$

This means that in the more patterns a given pair of neurons is coactive, the stronger the connection between these neurons. Here, N and p denote the total number of neurons and the number of stored patterns, respectively. We also introduced a set of coefficients, c_a, that describe how strongly a given pattern is included in the network connections. In the standard Hopfield model, these coefficients are initialized and remain equal to one. In this study, these coefficients are affected by the ongoing activity in the network. The goal of our study is to understand the long-term behavior of the strengths of the patterns $c_a(t)$ that result from ongoing learning.

In this network, patterns that are embedded in the recurrent weights, according to Equation 46, become network attractors (Hopfield, 1982, 1984). This means that if the activity of neurons matches one of the patterns at some moment in time, the pattern will be maintained by recurrent connections despite small perturbations and noise. Because several patterns are simultaneously embedded in the weights, the network may have several stable attractor states provided that the number of patterns, p, is not too large (Hertz et al., 1991). Because network firing rates can persist only near an attractor, in the absence of external inputs, the network must choose where to reside. This decision can be viewed as an implementation of short-term memory of the “which attractor I am near” kind. Therefore, Hopfield nets can support both short (working) and long-term memory (Bird and Burgess, 2008; Cowan, 2008) (Fig. 1).

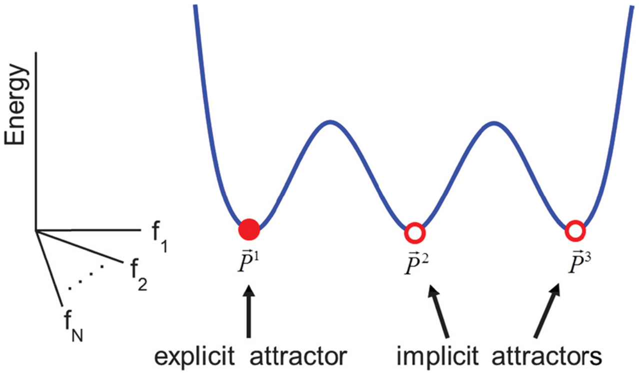

Figure 1. Attractor neural networks. Dynamics of neuronal firing rates are viewed as gradient descents with the landscape defined by an energy-like function (blue) (Hopfield, 1982, 1984). The landscape is defined in the space of activities of all neurons in the network ($f_1...f_N$). Position in the landscape is determined by the combined vector of neural activities. At the bottom of the landscape are the patterns that are stored in network weights according to Equation 46. These patterns represent network attractors that encode long-term memory stored in the network. In our study, attractors are divided into two classes: explicit and implicit. Explicit states (closed circles) are actively explored by the network and therefore can be reinforced by rehearsal. Implicit attractors (empty circles) have not been visited by the network within the time window of synaptic decay. Therefore, these patterns are expected to disappear because of the decay of synaptic strengths.

STDP rules applied to the attractor neural network lead to the deterioration of stored memories

What is the effect of synaptic plasticity on the attractors that are embedded into the network? From the point of view of plasticity, it is important to distinguish two types of attractors. First, there are attractors that represent memories that the network is constantly visiting. For example, because of external stimuli, the network can hop around states that are relevant to the particular task or environment. These attractor states represent recent memory. More precisely, recent states are defined as those that are visited within the time constant of synaptic decay. We call this type of states explicit attractors. The other type of state represents memories that were embedded into the network a long time ago and have not been accessed recently. These states will be called implicit (Fig. 1). Note that our terminology is somewhat different from the convention that uses the terms explicit/implicit to denote different classes of memory; that is, declarative versus procedural (Schacter, 1987).

Synaptic learning leads to different outcomes for explicit and implicit memory states. Because explicit memories are replayed in network activities, they are constantly rehearsed and therefore their contribution within the weight matrix is stable. In the Materials and Methods section, we evaluate the component of the weight matrix that carries explicit states (Equation 13). Specifically, we show that this component does not decay with time and is reinforced by learning in the network.

The behavior of implicit memories is quite different. If the attractors that correspond to implicit patterns are not visited within the time window of the decay of synaptic strength, defined in our model by parameter τ_0, these memory states disappear from the network weight matrix (Fig. 2). This observation is not surprising because rehearsal that reinforces the explicit attractors

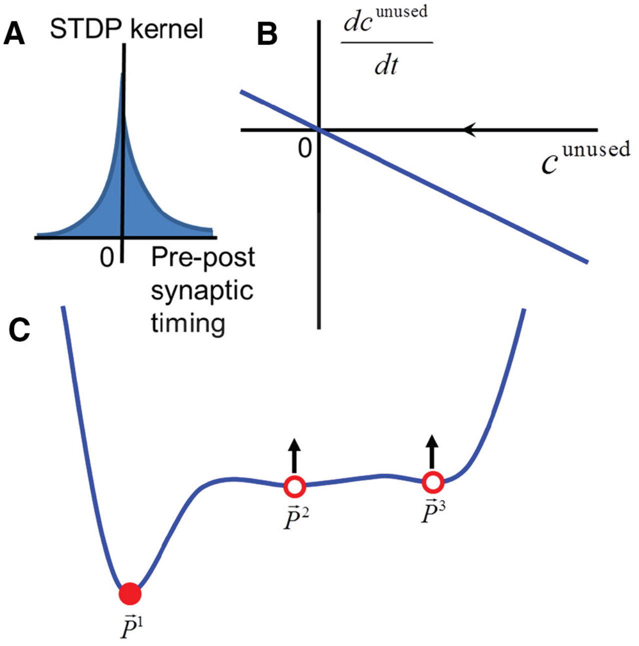

Figure 2. Synaptic plasticity implemented in the attractor neural network destroys implicit states. ***A***, STDP rules used. The rate of change of connection strength between two neurons (vertical axis) as a function of differences in spike time between presynaptic and postsynaptic cells. ***B***, Rate of change in the coefficient with which an implicit pattern enters the synaptic weight matrix, defined by Equation 46, as a function of the value of the coefficient. This coefficient describes the strength of the pattern in the weight matrix. For positive values of the coefficient, the rate of change is negative (left arrow), which implies decay. The decay is exponential $c^{\text{unused}} \sim \exp(-t/\tau_0)$, where τ_0 is the time constant of synaptic decay. ***C***, Illustration of the implications of decay of the coefficient for network attractors. Implicit attractors become less stable and disappear because they vanish from the weight matrix.

is not available for implicit attractor states. This is because the latter are not present in the network activity.

Noise added to the network can implement rehearsal of old (implicit) memory states

Next, we included noise in the inputs of neurons to determine whether noise can reinforce implicit memory states. We reasoned that, if white unstructured noise were added to the input of every neuron, the activity of the network would contain implicit memory states, which may potentially stabilize old memories through the process of rehearsal. We call the process of rehearsal that is based on random noise implicit rehearsal. This process is distinct from the rehearsal of explicit states that occurs due to the network actually visiting explicit attractors.

The dynamics of implicit rehearsal is as follows. Random unstructured noise is added to the inputs of every neuron in the network. The term unstructured implies that the amount of noise added to neurons does not contain the patterns being rehearsed. In this study, it is assumed to be the same for all neurons for simplicity. Because neurons are connected by recurrent weights that do contain implicit patterns, when noise passes through recurrent connections, it becomes structured. This means that neural activity acquires correlations that contain implicit patterns (Equation 17). This is because implicit states are amplified by positive feedback that is present within the recurrent weight matrix (Equation 46) and fluctuations along these directions are therefore amplified by recurrent connections. Therefore, despite the network staying near explicit attractors and never visiting the implicit states, the presence of implicit states in the weight matrix shapes network fluctuations along the directions that represent old memories. Implicit memory is contained in the correlations of network activity as opposed to the explicit memory that is contained in the mean firing rate.

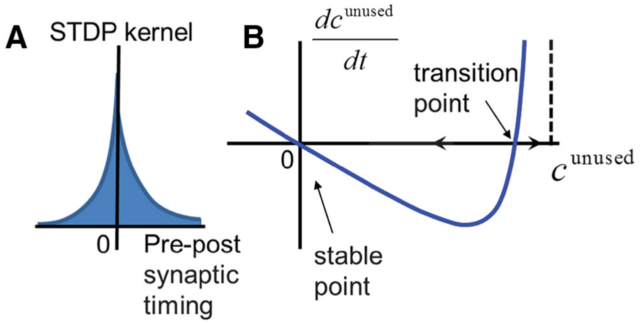

Figure 3. Unstructured neural noise does not stabilize implicit attractors in the case of the non-negative symmetric STDP rule. ***A***, Synaptic learning rule is the same in Figure 2*A*. ***B***, The rate of change in the contribution of the implicit attractor to the weight matrix contains only one stable point at $c^{\text{unused}} = 0$, which implies a lack of stability of unused memory. This dependence is represented by Equation 18.

Non-negative symmetric STDP rules applied to the network with white noise do not stabilize old (implicit) memory

What is the effect of implicit rehearsal in the case when STPD rules are ongoing in the network? For an STDP learning rule, a change in synaptic efficacy is dependent on the relative timing of presynaptic and postsynaptic action potentials for every synapse. In the simplest case, the synapse becomes stronger if both presynaptic spikes precede the action potential in the postsynaptic neuron and, in the opposite case, when postsynaptic spikes precede presynaptic spikes. We call this form of learning rule symmetric non-negative (Fig. 3). Our results show that this type of learning rule, applied in the presence of neural noise, does not make unused (implicit) memory more stable. The contribution of a memory state into the weight matrix is determined by coefficient, *c*, defined by Equation 46. The rate of change in this contribution defines the behavior of old memories with time. Figure 3*B* shows that, in the case of non-negative symmetric STDP, the only stable point for this coefficient is zero, which means that implicit memory is destined to disappear (for more detail, see Equation 18). Interestingly, if the strength of this coefficient is sufficiently large and if it passes the transition point in Figure 3*B*, the coefficient becomes unstable. This instability implies that the old pattern will emerge spontaneously in the network when the strength of the pattern is sufficiently large. In both regimes of small and unstable *c*, the network cannot maintain the old memory in a reliable manner.

Antisymmetric STDP rules combined with white noise can stabilize old memory states that are not revisited

We examined the stability of implicit states when STPD rules have a form that is often observed experimentally (i.e., antisymmetric) (Bi and Poo, 2001; Froemke and Dan, 2002; Sjöström et al., 2008). We assumed that if a presynaptic spike precedes the postsynaptic spike, the synapse is strengthened due to LTP. If the timing of the spikes is reversed, the synapse is weakened; that is, the contribution of such events to the synaptic strength are negative (Fig. 4*A*), which corresponds to LTD. We find that, in this case, an implicit (unused) memory state can have two stable points. The stable points are defined for the contribution of the

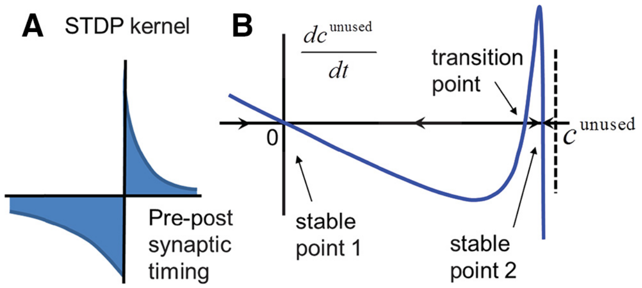

Figure 4. The stabilization of old memory states by the combination of unstructured noise and antisymmetric STDP learning rule. ***A***, Antisymmetric STDP learning rule. ***B***, Rate of change of the contribution of an unused state (dc^{unused}/dt) as a function of the contribution itself (c^{unused}). This dependence is represented by Equation 18. In this case, the contribution has two stable points near zero and at a finite value. The former/latter stable points correspond to the unused memory pattern being absent/present in the network connectivity, respectively. At a stable point, the rate of change of the pattern contribution is zero. In addition, small perturbations from the stable point will induce the rate of change that returns the system back. At the transition point, the rate of change is zero and unstable. Parameters used are: $A'_+ = 0.02$, $A'_- = -0.012$, $\tau_+ = 50$ ms, $\tau_- = 100$ ms, and $\tau = 5$ ms.

pattern to the weight matrix *c* (Equation 46). Stable points can be determined by examining the rate of change of this contribution (Fig. 4*B*) that is given by Equation 18). If, for a certain value of contribution, the rate of change is zero, this value is called the stationary point. If the contribution of a pattern is placed exactly into one of the stationary points, it will remain there because the rate of change of *c* is zero. Figure 4*B* shows three stationary points in this case. These three points differ in cases of small perturbations that deflect contribution, *c*, slightly from a stationary state. For two of the stationary states in Figure 4*B*, the resulting rate of change returns the contribution back to the state. This is illustrated by the arrows on the horizontal axis. Therefore, these two states are stable. The third stationary point is unstable and is called the transition point.

For two stable states, the contribution of the pattern is either low (stable point 1) or high (stable point 2). The former state corresponds to the weak representation of the pattern in the network that is indistinguishable from noise. The high contribution point (stable point 2) corresponds to the memory that is substantively present in the weight matrix. The system is capable of maintaining either high or low levels of a pattern in the weight matrix virtually indefinitely. This is despite the decay of synaptic strength that is ongoing in the system. Contribution is maintained because noise implements implicit rehearsal. Although the average values of firing rates are near the explicit attractor, the correlations in the firings rates between cells, induced by noise, carry information about other patterns that are not visited (Equation 17). Because learning rules are dependent on correlations, Hebbian learning is capable of maintaining implicit states in memory. This correlation-induced rehearsal results in the stability of patterns as a function of time. Although we presented the results for a single pattern, other implicit states are stabilized similarly due to their independence (Equation 2). Implicit rehearsal can stabilize several patterns simultaneously.

Conditions of bistability

For the pattern contribution to have two stable points, several conditions have to be met. First, the integral of STDP kernel (Fig. 4*A*) must be negative. This implies that the LTD part of the STDP curve is stronger than the LTP part. Second, we need the following equation:

$$\gamma g^2 \xi^2 (A_+ + A_-) > \frac{2\tau^2}{\tau_+ \tau_-},$$

which can be satisfied if timescales of STDP, $\tau_\pm$, are larger than that of firing rate, τ. This condition guaranties that $\frac{dc_a}{dt}$, given by Equation 18, has one local minimum and one local maximum on the region $0 \le c_a(t) < 1/g$. To have the second stable point at finite $c_a(t)$, we also need the local maximum to be positive. In the case that $\gamma g^2 \xi^2 (A_+ + A_-)$ is much larger than $\frac{2\tau^2}{\tau_+ \tau_-}$, this condition can be met when:

$$\frac{\gamma g^2 \xi^2}{2}\left(\frac{A_+ + A_-}{3}\right)^3 > \left(\frac{\tau}{2\tau_+ \tau_-}(A_+ \tau_+ + A_- \tau_-)\right)^2$$

Figure 4*B* shows the typical behavior of $\frac{dc_a}{dt}$ and a set of values for parameters.

Our model predicts correlations between network weights and neural noise

What are the implications of our findings for an individual synapse? To maintain a set of memories in the network, the synapses have to preserve their strength. Because, in our model, synaptic strength decays with time, it has to be maintained at a constant level by the correlations in the presynaptic and postsynaptic activities. This means that stronger synapses have higher correlations in presynaptic and postsynaptic activities. Therefore, the form of rehearsal proposed here can be detected by measuring the correlations in activity for individual synapses and observing their relationships with synaptic strength. A more precise definition of this relationship is given by Equation 10.

Computer simulations

We have presented the results of our analyses that can be described by equations in the closed form, such as Equation 8. These results have the advantage because we can immediately see what combination of parameters can implement the proposed mechanism, as shown in Figure 4. We also can analyze the behavior of very large networks with an unlimited number of neurons. The mathematical methods used in the previous sections, including the separation of short timescales (STDP ~0.1 s) and long timescales (synaptic decay time ~month), are often used in the studies of dynamical stabilization of mechanical, atomic, and plasma confinement (Landau and Lifshits, 1976; Paul, 1990; Hoang et al., 2013). However, to derive these results, some assumptions had to be made (see Materials and Methods). One of the assumptions is that learning in the network is determined by average noise correlations, rather than instantaneous values of noise (Equation 8). This assumption is generally called the mean-field approximation (Hertz et al., 1991). The second assumption is that the patterns stored in the network are strictly orthogonal. Finally, to derive our results, we assumed that the timescale of noise is much faster that the rate of learning, which allowed us to consider noise a perturbation. To validate these assumptions, in this section, we present the results of computer modeling of the implicit attractors in the network of neurons described by the firing rate model.

First, we start with the simplest case where the network has only one explicit pattern (an active memory) and one implicit pattern (memory that is never reactivated). Our simulations confirm that without noise the implicit pattern will decay with time and only the explicit pattern survives (Fig. 5).

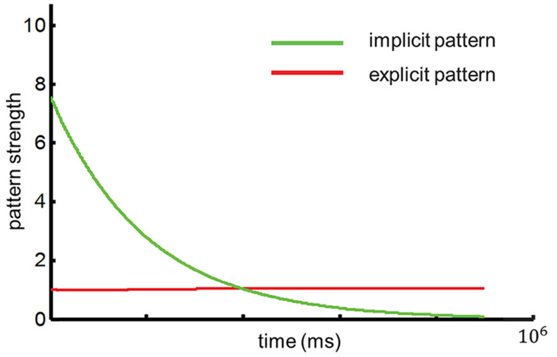

Figure 5. Network with one explicit pattern and one implicit pattern in the case of no noise. The strength of the explicit pattern (red line) maintains a constant value. The implicit pattern (green line) decays to zero. The network parameters used in this simulation are as follows: $A_+ = 2, A_- = -1.2, \tau_+ = 50$ ms, $\tau_- = 100$ ms, $\tau = 5$ ms, $\gamma = 9000$ ms^{-1}, $g = 0.1$, $\tau_0 = 2 \times 10^5$ ms and the total number of neurons is 1024.

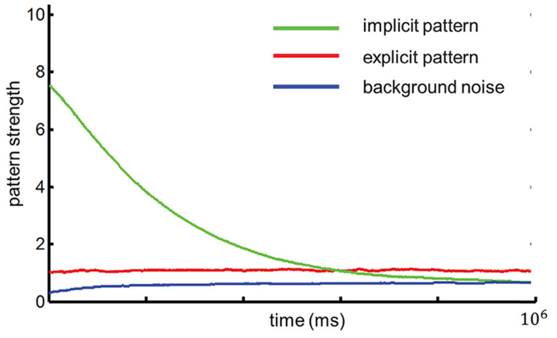

Figure 6. Network with one explicit pattern and one implicit pattern. The strength of the explicit pattern (red line) fluctuates around a constant value. The initial strength of the implicit pattern is set below the transition point in Figure 4*B* and is therefore in the first stable region. The strength of the implicit pattern (green line) decays to the first stable point, which is the noise level (blue line). The network parameters used in this simulation are as follows: $A_+ = 2$, $A\_ = -1.2$, $\tau_+ = 50$ ms, $\tau_- = 100$ ms, $\tau = 5$ ms, $\gamma = 9000$ ms^{-1}, $g = 0.1$, $\tau_0 = 2 \times 10^5$ ms, $\xi = 0.1118$ and the total number of neurons is 1024.

With random noise and the antisymmetric STDP rules (Fig. 6), we observe that the network behaves as suggested by our previous analysis: the strength of the explicit pattern fluctuates around a constant value and never decays. The implicit pattern may decay to the noise level, however, defined as the amplitude of an arbitrary pattern contained in the random noise, if the strength of pattern is set at a value lower than the transition point in Figure 4*B*. Conversely, if the strength of the implicit pattern is initially set to a value higher than the transition point, it will fluctuate around the second stable point of Figure 4*B* for a longer time (Fig. 7). In this case, the implicit memory is rehearsed by random noise and is kept in the network for a time much longer than the synaptic decay time τ_0.

Next, we simulate the dynamics of a network with one explicit pattern and multiple implicit patterns. Similar behaviors are observed: the strength of the explicit pattern fluctuates around a constant value; implicit patterns are maintained by random noise. When the initial strength of implicit patterns are set above the transition point, the implicit memories can be maintained for a period of time considerably longer than the typical decay-time τ_0. When the number of implicit patterns increases, some patterns may start to decay earlier than others (Fig. 8).

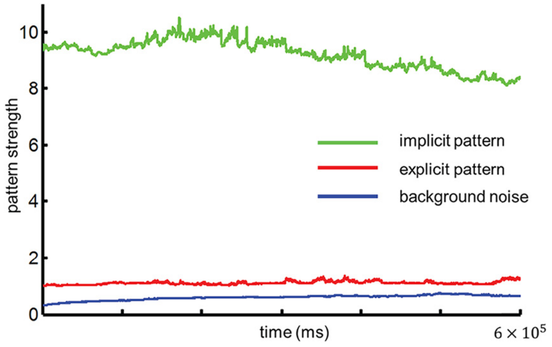

Figure 7. Network with one explicit pattern and one implicit pattern. The strength of the explicit pattern (red line) fluctuates around a constant value. When the initial strength is set above the transition point, and thus in the second stable region, the implicit pattern (green line) fluctuates around the second stable point. The blue line shows the noise level. The network parameters are the same with those in Figure 6.

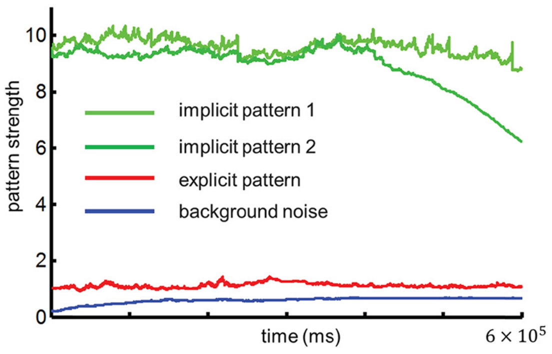

Figure 8. Network with one explicit pattern (red) and two implicit patterns (bright and dark green). The blue line shows the noise level. One of the implicit pattern starts to decay at $\sim T = 4 \times 10^5$ ms. The network parameters are the same as in Figure 6. Decay of the implicit pattern starts when its strength falls below the transition point. Since the strength of implicit pattern evolves as a random walk with drift (the drift can be either positive or negative, depending on the pattern strength; Fig. 4*B*), it can drop below the transition point. From Figure 6 and Figure 8, we estimate that the transition point for the set of network parameters used in these simulations is ~9. The upper bound of pattern strength for the memory to be stable is $g^{-1} = 10$. Therefore, the implicit memory can be maintained in the network if its initial strength is set between 9 and 10.

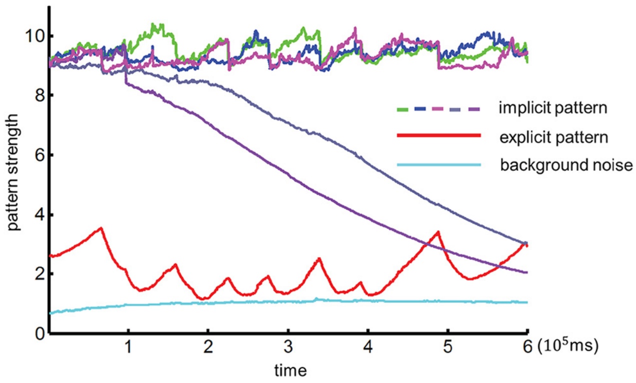

Figure 9. Network with one explicit pattern (red) and five implicit patterns (bright green, blue, and pink). The bright blue line shows the noise level. The network parameters are similar to those is Figure 6, except with larger noise amplitude $\xi = 0.125$.

Figure 9 demonstrates behavior of the network that contains one explicit and five implicit attractors. During the simulation shown, two of the implicit attractor states decayed to the baseline, indicating that these implicit states have been forgotten. Forgetting was initiated when their strength fell below the transition point briefly due to a fluctuation. This behavior is expected in our simulations because we used the timescale of synaptic decay equal to 200 s. This choice was forced by the limits on the amount of time needed to run these simulations that also included the millisecond timescales. We anticipate that if the synaptic lifetime is close to several weeks, as in biological networks, the implicit patterns are more stable (for more detail, see Materials and Methods, the section "Validity of the MFA," and Equation 31). Therefore, although the computer simulations attempted here can validate the behaviors described by the analytic calculations presented in the previous sections, the computer model is constrained to overestimate the effects of global fluctuations leading to transitions in some implicit states.

Discussion

In this study, we examined the behavior of memory states stored in the weights of an attractor neural network. Our network included some realistic features, such as STDP learning rules, neural noise, and limited LTP/LTD lifetime. We assumed that learning occurs in the network on a continuous basis; that is, the weights are continuously updated to reflect ongoing activity. In these conditions, the network weights should reflect the ongoing activity that we described by the term explicit attractors. The other set of states, which we called implicit attractors, represent memories that were stored at some point in the past and that have not been revisited recently or within the lifetime of synaptic strengths. Such states are expected to disappear from the network weight matrix because of the decay of synaptic strengths. How can the network maintain implicit memory states despite synaptic decay?

We show that unstructured noise can substantially alter the dynamics of forgetting. Although input noise is unstructured in our model; that is, it does not contain stored patterns—when noise passes through recurrent synaptic weight matrix, it becomes colored. This means that the correlations in neural activity that are induced by noise reflect all memory patterns stored in the network, both explicit and implicit. This is because the memory states represent the directions in neural activities that are amplified by the positive feedback present in the recurrent network. Therefore, although the average neural activity represents recent states only, the correlations reflect the entirety of memories, including the old ones (i.e., implicit). Because synaptic learning is dependent on correlations, in principle, it can reinforce implicit memory states, when certain conditions are met. We show that the antisymmetric STDP rule, which contains both positive and negative components (i.e., both LTP and LTD; Bi and Poo, 2001; Froemke and Dan, 2002), can reinforce old memory traces without explicitly visiting them (Fig. 4). In contrast, non-negative symmetric STDP cannot stabilize old memories (Fig. 3).

In our model, the old memory traces (implicit) are never visited or accessed by the network. The network always resides near the set of newer states that are relevant to current behavior and, therefore, are called explicit. However, we propose that implicit states can be rehearsed (Fig. 10). The rehearsal occurs not because the average activity, but rather fluctuations reflect the old states. We call this form of rehearsal, in which the old memory is never directly accessed, implicit rehearsal.

A candidate mechanism making memory more robust involves rehearsals whereby old memories are constantly revisited and relearned via an ongoing process. This mechanism has been proposed to resolve both the problem of unstable synapses (Wittenberg et al., 2002) and the catastrophic interference problem (McClelland et al., 1995; Robins, 1996; Robins and McCallum, 1998). Because all old memory states must be explicitly visited within the time window of LTP decay, presumably in the sleep, it is unclear whether such a mechanism is realistic, especially if the number of patterns is large. Within our classification, the class of models proposed by Wittenberg et al. (2002) could be called explicit rehearsal networks.

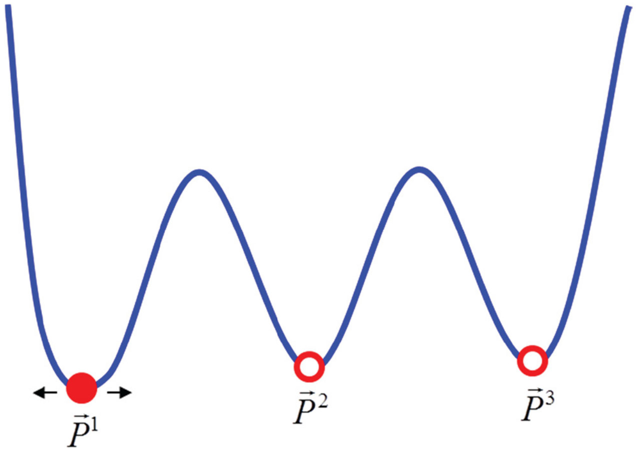

Figure 10. Illustration of implicit rehearsal. Implicit (unused and never visited) attractors 2 and 3 do not vanish with time, but rather remain stable due to the fluctuations (arrows) around explicit attractor 1. Although the average activity remains near the explicit attractor, the fluctuations are biased toward the direction of implicit states, which leads to their rehearsal.

Our model suggests the functional reason for the high degree of irregularity observed in firing of cortical neurons (Softky and Koch, 1993). In our model, irregular neuronal firing implements rehearsal of old memory patterns. When neural noise is passed through the network weight matrix, it captures the information about patterns stored in these connections. The ongoing synaptic plasticity can subsequently reinforce the stored patterns. Neural noise plays an essential role in generating correlations in neural activity. Similar roles can be played by the unreliable nature of synaptic vesicle release (Sudhof, 2004). Therefore, we propose that unreliable neural activity is the feature that helps cortical networks maintain stable connections.

In our model, both network weights and firing rates exhibit attractor behaviors. Firing rates can have several discrete states that are robust with respect to small perturbations and noise and are called network attractors. The identity of these states depends on the strengths of recurrent connections between neurons (Amit, 1989; Amit et al., 1994). In addition, these states represent long-term memories that are stored in the network weight matrix. In our model, network weights also exhibit attractor behaviors. We show that, because of ongoing synaptic plasticity, a weight matrix can have self-maintaining stable states that could also be called attractors. In Figure 4*B*, we show that synaptic weight matrix can have two stable states that correspond to a given memory pattern being present or absent from the network connectivity. Once the weight matrix is placed near the state that includes a given memory pattern, it will stay there for a long time, which ensures the stability of the memory of the pattern. The attractors of the weight matrix, in our model, are maintained by ongoing neural activity generated by network noise. As such, neural activity helps synaptic weights form stable states (i.e., attractors). In our model, firing rates and synaptic connections form two dual systems of attractors: synaptic weights help firing rates to exhibit discrete stable states and firing rates stabilize discrete self-maintaining states within network connections.

Our model can provide a rationale to the standard model of systems memory consolidation (Dudai, 2004). We proposed here that certain memory traces can be maintained in stable states over long periods by implicit rehearsal. The problem of placing the network into these states is not addressed here. However, we notice that the regions of stability surrounding stable memory states are narrow (stable point 2 in Fig. 4*B*). The parameters of network weights that describe the contribution of a given memory pattern must be tuned to a relatively precise value for the pattern to be stable. We argue that the function of placing the network in a narrow parameter range, which is necessary for long-term storage, is performed by memory consolidation. Once consolidated, that is, placed near stable point 2 (Fig. 4*B*), a memory pattern can be maintained by implicit rehearsal. Therefore, we argue that the functional role of system memory consolidation is to place the network weights within the narrow range of parameters where the memory trace can persist for a long time.

Our study provides experimentally testable predictions. Within the implicit rehearsal mechanism, synaptic strength is larger for synapses with stronger correlations between presynaptic and postsynaptic activities (Equation 10). The remainder of the network produces these correlations, which then reinforce the synaptic strength. This prediction can be tested if synaptic strength is measured simultaneously with ongoing neural activity for individual synapses. In doing so, one should isolate correlations of activity induced by measured synapse and the remainder of the network. This could be done pharmacologically or by including correlations over certain timescales, such as one temporal semi-axis for unidirectional synapses. Specific STDP kernel could also be surmised based on studies of synaptic plasticity or could be derived from the best match between synaptic strength and activity correlations. Overall, we propose that synaptic strengths are maintained by ongoing irregular spiking, which can be tested experimentally.

In our model, individual synapses are unstable, which is described by a “forgetting” term in Equation 9. This feature limits the strength of synapses for a given value of activity correlations, thus introducing a soft bound on synaptic strength. Soft bound implies that synaptic strengths are constrained but the synapses are not bound by a specific value. The behavior of individual synapses in our model is not expected to be different from the models that include hard limits on synaptic strengths (Amit and Fusi, 1994; Fusi, 2002; Fusi et al., 2005; Fusi and Abbott, 2007). In contrast to the models with a hard limit on synaptic strengths, in our model, strong synapses are possible, but their existence is less likely. In adopting this assumption, we were motivated by the observation of log-normal distribution of synaptic strength (Song et al., 2005; Koulakov et al., 2009; Mizuseki and Buzsáki, 2013), which implies that strong synapses are quite possible.

A related question has been addressed in the attempt to build molecular models of LTP (Crick, 1984). Although LTP lifetime, in most cases, is measured in weeks (Abraham, 2003), it is believed that molecules in synapses undergo turnover every several days (Lisman and Hell, 2008). Therefore, the persistence of LTP has to be reconciled with the dynamics of molecules that have relatively short lifetimes. Several studies have proposed how short-lived molecules can build a lasting synapse, including bistability (Lisman and Zhabotinsky, 2001; Miller et al., 2005) and self-sustaining molecular clusters (Shouval, 2005). This problem has many parallels with the question studied here because, in these models, relatively stable synapses result from activities of unstable molecules.

A related question, known in computational literature as the plasticity-stability dilemma, poses that a memory system must evolve to be able to both store new memories promptly and retain old information (Grossberg, 1987; Abraham and Robins, 2005). As a consequence of these contradicting requirements, the neural networks have to overcome what is known as a catastrophic forgetting or catastrophic interference phenomenon whereby old memories are continually overwritten by new ones (Mézard et al., 1986; Nadal et al., 1986; McCloskey and Cohen, 1989). Some solutions to catastrophic interference have been proposed (Carpenter and Grossberg, 1987). Here, we argue that, even without the challenge from novel memories, the known lifetime of LTP is not in agreement with the long-lasting nature of long-term memory.

An interesting observation is that stability of long-term memory in our model seems to imply stability of individual synapses, which is in contrast to the fleeting nature of synaptic strengths discussed in the introduction. Although our model does stabilize memory states, it also allows individual synapses to be unstable. Because, in our study, memory is delocalized and each memory trace is represented by nearly all synapses in the network, variability in individual synapses does not imply memory decay. Our model offers two distinct scenarios for how memory states are corrupted. First, they can completely disappear through a discontinuous jump between two stable points, as in Figure 4*B*. Second, they can slowly change by changing each individual synapse at a time, whereby a memory state morphs into something else. Because the number of synapses involved in each trace is large, this process may progress for a long time without substantial change in the representation of memory. Overall, our model predicts that synaptic lifetime observed in reduced preparations, such as in slices, should be shorter than *in vivo*, because the latter lifetime is improved by the ongoing activity. Rare instances when synapses show stability (Abraham et al., 2002) could be attributed to the mechanism of stabilization proposed here. *In vivo*, synaptic persistence can be shorter than memory retention time because memory is a collective property of a large ensemble of synapses.

Conclusions

Here, we studied the stability of long-term memory patterns stored in a recurrent neural network. Our model includes ongoing synaptic plasticity regulated by STDP rules. We show that old memory traces can be stabilized by fluctuations of neural activity when STDP rules satisfy certain constraints. Old memory patterns become stable self-maintaining and persistent states of the network weight matrix. Our model provides a mechanism for the extension of memory lifetime via the combination of ongoing synaptic plasticity and neural noise.

References

Abbott LF, Nelson SB (2000) Synaptic plasticity: taming the beast. Nat Neurosci 3:1178–1183. CrossRef Medline

Abraham WC (2003) How long will long-term potentiation last? Philos Trans R Soc Lond B Biol Sci 358:735–744. CrossRef Medline

Abraham WC, Robins A (2005) Memory retention–the synaptic stability versus plasticity dilemma. Trends Neurosci 28:73–78. CrossRef Medline

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

Alvarez VA, Sabatini BL (2007) Anatomical and physiological plasticity of dendritic spines. Annu Rev Neurosci 30:79–97. CrossRef Medline

Amit DJ (1989) Modelling brain function: the world of attractor neural networks. New York: Cambridge University.

Amit DJ, Fusi S (1994) Learning in neural networks with material synapses. Neural Computing 6:957–982. CrossRef

Amit DJ, Brunel N, Tsodyks MV (1994) Correlations of cortical Hebbian reverberations: theory versus experiment. J Neurosci 14:6435–6445. Medline

Bi G, Poo M (2001) Synaptic modification by correlated activity: Hebb's postulate revisited. Annu Rev Neurosci 24:139–166. CrossRef Medline

Bird CM, Burgess N (2008) The hippocampus and memory: insights from spatial processing. Nat Rev Neurosci 9:182–194. CrossRef Medline

Carpenter GA, Grossberg S (1987) Art-2: self-organization of stable category recognition codes for analog input patterns. Applied Optics 26: 4919–4930. CrossRef Medline

Cooke SF, Bliss TV (2006) Plasticity in the human central nervous system. Brain 129:1659–1673. CrossRef Medline

Cowan N (2008) What are the differences between long-term, short-term, and working memory? Essence of Memory 169:323–338. CrossRef

Crick F (1984) Memory and molecular turnover. Nature 312:101. CrossRef Medline

Dayan P, Abbott LF (2001) Theoretical neuroscience: computational and mathematical modeling of neural systems. Cambridge, MA: MIT.

Dudai Y (2004) The neurobiology of consolidations, or, how stable is the engram? Annu Rev Psychol 55:51–86. CrossRef Medline

Feldman DE (2009) Synaptic mechanisms for plasticity in neocortex. Annu Rev Neurosci 32:33–55. CrossRef Medline

Froemke RC, Dan Y (2002) Spike-timing-dependent synaptic modification induced by natural spike trains. Nature 416:433–438. CrossRef Medline

Fu M, Zuo Y (2011) Experience-dependent structural plasticity in the cortex. Trends Neurosci 34:177–187. CrossRef Medline

Fusi S (2002) Hebbian spike-driven synaptic plasticity for learning patterns of mean firing rates. Biol Cybern 87:459–470. CrossRef Medline

Fusi S, Abbott LF (2007) Limits on the memory storage capacity of bounded synapses. Nat Neurosci 10:485–493. CrossRef Medline

Fusi S, Drew PJ, Abbott LF (2005) Cascade models of synaptically stored memories. Neuron 45:599–611. CrossRef Medline

Grossberg S (1987) Competitive learning: from interactive activation to adaptive resonance. Cognitive Science 11:23–63. CrossRef

Grutzendler J, Kasthuri N, Gan WB (2002) Long-term dendritic spine stability in the adult cortex. Nature 420:812–816. CrossRef Medline

Hertz J, Krogh A, Palmer RG (1991) Introduction to the theory of neural computing. Cambridge, MA: Westview.

Hoang TM, Gerving CS, Land BJ, Anquez M, Hamley CD, Chapman MS (2013) Dynamic stabilization of a quantum many-body spin system. Phys Rev Lett 111:090403. CrossRef Medline

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

Hopfield JJ (1984) Neurons with graded response have collective computational properties like those of two-state neurons. Proc Natl Acad Sci U S A 81:3088–3092. CrossRef Medline

Ivanco TL, Racine RJ (2000) Long-term potentiation in the reciprocal corticohippocampal and corticocortical pathways in the chronically implanted, freely moving rat. Hippocampus 10:143–152. CrossRef Medline

Kempter R, Gerstner W, von Hemmen JL (1999) Hebbian learning and spiking neurons. Physical Review E 59:4498–4514. CrossRef

Knott GW, Holtmaat A, Wilbrecht L, Welker E, Svoboda K (2006) Spine growth precedes synapse formation in the adult neocortex in vivo. Nat Neurosci 9:1117–1124. CrossRef Medline

Koulakov AA, Hromádka T, Zador AM (2009) Correlated connectivity and the distribution of firing rates in the neocortex. J Neurosci 29:3685–3694. CrossRef Medline

Landau LD, Lifshits EM (1976) Mechanics, Ed 3. New York: Pergamon.

Lisman J, Hell JW (2008) Long-term potentiation. In: Structural and functional organization of the synapse (Hell JW, Ehlers MD, eds.), pp 501–534. New York: Springer.

Lisman JE, Zhabotinsky AM (2001) A model of synaptic memory: a CaMKII/PP1 switch that potentiates transmission by organizing an AMPA receptor anchoring assembly. Neuron 31:191–201. CrossRef Medline

Martin SJ, Grimwood PD, Morris RG (2000) Synaptic plasticity and memory: an evaluation of the hypothesis. Annu Rev Neurosci 23:649–711. CrossRef Medline

McClelland JL, McNaughton BL, O'Reilly RC (1995) Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychol Rev 102:419–457. CrossRef Medline

McCloskey M, Cohen N (1989) Catastrophic interference in connectionist networks: the sequential learning problem. In: the psychology of learning and motivation (Bower GH, ed), pp 109–164. San Diego: Academic.

Mézard M, Nadal JP, Toulouse G (1986) Solvable models of working memories. Journal de Physique 47:1457–1462.

Miller P, Zhabotinsky AM, Lisman JE, Wang XJ (2005) The stability of a stochastic CaMKII switch: dependence on the number of enzyme molecules and protein turnover. PLoS Biol 3:e107. CrossRef Medline

Mizuseki K, Buzsáki G (2013) Preconfigured, skewed distribution of firing rates in the hippocampus and entorhinal cortex. Cell Rep 4:1010–1021. CrossRef Medline

Nadal JP, Toulouse G, Changeux JP, Dehaene S (1986) Networks of formal neurons and memory palimpsests. Europhysics Letters 1:535–542. CrossRef

Nadel L, Moscovitch M (1997) Memory consolidation, retrograde amnesia and the hippocampal complex. Curr Opin Neurobiol 7:217–227. CrossRef Medline

Nadel L, Moscovitch M (2001) The hippocampal complex and long-term memory revisited. Trends Cogn Sci 5:228–230. CrossRef Medline

Paul W (1990) Electromagnetic traps for charged and neutral particles. Reviews of Modern Physics 62:531–540. CrossRef

Reymann KG, Malisch R, Schulzeck K, Brödemann R, Ott T, Matthies H (1985) The duration of long-term potentiation in the CA1 region of the hippocampal slice preparation. Brain Res Bull 15:249–255. CrossRef Medline

Robins A (1996) Consolidation in neural networks and in the sleeping brain. Connection Science 8:259–276. CrossRef

Robins A, McCallum S (1998) Catastrophic forgetting and the pseudorehearsal solution in Hopfield-type networks. Connection Science 10: 121–135. CrossRef

Schacter DL (1987) Implicit memory: history and current status. Journal of Experimental Psychology: Learning, Memory, and Cognition 13:501–518. CrossRef

Shors TJ, Matzel LD. Long-term potentiation: what's learning got to do with it? Behav Brain Sci 20:597–614, 1997; discussion 614–555.

Shouval HZ (2005) Clusters of interacting receptors can stabilize synaptic efficacies. Proc Natl Acad Sci U S A 102:14440–14445. CrossRef Medline

Sjöström PJ, Rancz EA, Roth A, Häusser M (2008) Dendritic excitability and synaptic plasticity. Physiol Rev 88:769–840. CrossRef Medline

Softky WR, Koch C (1993) The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. J Neurosci 13: 334–350. Medline

Song S, Sjöström PJ, Reigl M, Nelson S, Chklovskii DB (2005) Highly nonrandom features of synaptic connectivity in local cortical circuits. PLoS Biol 3:e68. CrossRef Medline

Sudhof TC (2004) The synaptic vesicle cycle. Annu Rev Neurosci 27:509–547. CrossRef Medline

Trachtenberg JT, Chen BE, Knott GW, Feng G, Sanes JR, Welker E, Svoboda K (2002) Long-term in vivo imaging of experience-dependent synaptic plasticity in adult cortex. Nature 420:788–794. CrossRef Medline

Trepel C, Racine RJ (1998) Long-term potentiation in the neocortex of the adult, freely moving rat. Cereb Cortex 8:719–729. CrossRef Medline

Whitlock JR, Heynen AJ, Shuler MG, Bear MF (2006) Learning induces long-term potentiation in the hippocampus. Science 313:1093–1097. CrossRef Medline

Wittenberg GM, Sullivan MR, Tsien JZ (2002) Synaptic reentry reinforcement based network model for long-term memory consolidation. Hippocampus 12:637–647. CrossRef Medline