Extended Temporal Association Memory by Modulations of Inhibitory Circuits

Tatsuya Haga[1*] and Tomoki Fukai[1,2†]
[1]RIKEN Center for Brain Science, Wako, Saitama 351-0198, Japan
[2]Okinawa Institute of Science and Technology, Onna-son, Okinawa 904-0495, Japan

(Received 17 September 2018; revised manuscript received 24 May 2019; published 16 August 2019)

Hebbian learning of excitatory synapses plays a central role in storing activity patterns in associative memory models. Interstimulus Hebbian learning associates multiple items by converting temporal correlation to spatial correlation between attractors. Growing evidence suggests the importance of inhibitory plasticity in memory processing, but the consequence of such regulation in associative memory has not been understood. Noting that Hebbian learning of inhibitory synapses yields an anti-Hebbian effect, we show that the combination of Hebbian and anti-Hebbian learning can significantly increase the span of temporal association between correlated attractors as well as the sensitivity of these states to external input. Furthermore, these effects are regulated by changing the ratio of local and global recurrent inhibition after learning weights for excitation-inhibition balance. Our results suggest a nontrivial role of plasticity and modulation of inhibitory circuits in associative memory.

DOI: 10.1103/PhysRevLett.123.078101

Animals can recall memory from incomplete stimulus presentation; in other cases, the presentation of one item leads to the memory recall of a paired item. Such function is called associative memory. Hebb postulated that synchronous activation strengthens connections between neurons in the brain, and these strongly connected neuron ensembles (cell assemblies) are the basis of associative memory [1]. In the brain, this "Hebbian learning" of cell assemblies likely occurs through spike-timing-dependent plasticity (STDP) [2,3]. An attractor network model with Hebbian learning can recall activity patterns from incomplete external cues [4]. Today, Hebb's postulate is a widely accepted paradigm for memory processing in the brain.

A number of experiments suggest that association between items are represented by correlations between activity patterns in the brain [5–7]. One important finding was made in the investigation of prolonged activity patterns in the temporal cortex of monkeys performing a visual working memory task [8,9]. After uncorrelated visual stimuli were consecutively presented during training, those stimuli evoked mutually correlated activity patterns in the test phase although the presentation order was random. Griniasty *et al.* proposed a model that bridges Hebbian learning and this finding [10] by adding cross-stimulus terms to the local Hebbian connection matrix of the conventional associative memory model [4]. The extended model converts the sequence of uncorrelated stimulus patterns into correlations between attractors, which are significantly correlated up to a separation of five in the sequence. Notably, this span of temporal association is robust against variations in model parameters and is consistent with experimental observations [10–12].

While Hebbian learning is sufficient for supporting the correlated attractors, the role of inhibitory modulation for associative memory remains unclear. Actually, researchers are aware of the possible importance of inhibitory engrams in memory processing [7,13]. Experiments also have revealed that neuromodulators change the activity of cortical inhibitory neurons in various ways [14–20]. Here we show that the plasticity and modulation of inhibitory circuits induce previously unknown advantages in associative memory.

Let us assume a network of N neurons. Below, $S_i = 1, 0$ denotes activity of neuron i. Update of neural activity follows

$$S_i(t+\delta t) = \Theta\left(\sum_{j=1}^{N} J_{ij} S_j(t) + \theta_i\right), \tag{1}$$

where J_{ij} represents synaptic weights, θ_i is external inputs, and $\Theta(x)$ is Heaviside step function. We assume that neural activities are asynchronously (sequentially) updated, that is, only one neuron is chosen and updated at every time step. The network stores P random binary memory patterns $\xi_i^\mu (1 \le i \le N, 1 \le \mu \le P)$ that are biased as $E[\xi_i^\mu] = p (0 < p < 1)$ [21,22]. We define synaptic weights as

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} (c\hat{\xi}_i^\mu \hat{\xi}_j^\mu + \hat{\xi}_i^{\mu+1} \hat{\xi}_j^\mu + \hat{\xi}_i^\mu \hat{\xi}_j^{\mu+1}), \tag{2}$$

where $\hat{\xi}_i^\mu = \xi_i^\mu - p$ and $\xi_i^{P+1} = \xi_i^1$. The parameter c can be either positive or negative. When c is positive, this model is equivalent to that of Griniasty *et al.* [10]. On the other hand, negative c implies anti-Hebbian learning, which has not been extensively studied in associative memory.

We can biologically interpret this connectivity when we consider the weight [Eq. (2)] with $\tilde{\xi}_i^\mu = \xi_i^\mu - P^{-1}\sum_{\alpha=1}^P \xi_i^\alpha$ instead of $\hat{\xi}_i^\mu = \xi_i^\mu - p$. Here, p is substituted by $P^{-1}\sum_{\alpha=1}^P \xi_i^\alpha$ which converges to p in the limit of $P \to \infty$. Even when P is finite, the model behavior is similar to the original one (see Supplemental Material [23]). In the case of $\tilde{\xi}_i^\mu$, we can decompose the synaptic weight in our model into excitation and inhibition as

$$J_{ij} = J_{ij}^{\mathrm{E}} - J_{ij}^{\mathrm{I}}, \tag{3}$$

where

$$J_{ij}^{\mathrm{E}} = \frac{1}{N}\sum_{\mu=1}^{P}(2\xi_i^\mu\xi_j^\mu + \xi_i^{\mu+1}\xi_j^\mu + \xi_i^\mu\xi_j^{\mu+1}) \tag{4}$$

$$J_{ij}^{\mathrm{I}} = (2-c)\frac{1}{N}\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu + (2+c)\frac{1}{NP}\sum_{\mu=1}^{P}\xi_i^\mu\sum_{\mu=1}^{P}\xi_j^\mu. \tag{5}$$

We note that $J_{ij}^E \geq 0$ and $J_{ij}^{\mathrm{I}} \geq 0$ in the parameter region $-2 \leq c \leq 2$.

We briefly state the biological relevance of the connection matrices. First, the excitatory weights involve terms symmetric with respect to ξ^μ and $\xi^{\mu+1}$. In the cortex, consecutively presented stimuli can be associated with one another by certain mechanisms, as mentioned previously [9,24]. On the millisecond range timescale, these terms may emerge through a symmetric spike-timing-dependent plasticity with a broad time window. Such a STDP rule has been recently revealed in the hippocampal area CA3 [3]. Second, the inhibitory weights consist of two terms: the first term represents pattern-specific local inhibition and the second term is global inhibition proportional to the total activity of stored patterns. When c varies between -2 and 2, the balance of the two inhibition terms changes. Later, we will explore how this type of inhibitory modulation affects the associative memory performance of this model.

We analyze the attractors of this model by a similar procedure to the previous one [10]. For consistency with previous works, we use $\hat{\xi}_i^\mu = \xi_i^\mu - p$ in the analyses. We define a pattern overlap, which represents the degree of coincidence between the instantaneous activity and the μth memory pattern, as

$$m^\mu = \frac{1}{NB}\sum_{i=1}^{N}\hat{\xi}_i^\mu S_i, \tag{6}$$

where $B = p(1-p)$. Furthermore, we specifically consider the external inputs corresponding to a superposition of memory patterns ($\theta_i = \sum_\mu b^\mu \hat{\xi}_i^\mu$). Then, in the limit of $N \to \infty$, we can obtain the following mean-field time-evolution equations of overlaps from Eq. (1):

$$\dot{m}^\mu = -m^\mu + \frac{1}{B}\left\langle\!\!\left\langle \hat{\xi}^\mu \Theta\left(\sum_{\alpha=1}^{P}\hat{\xi}^\alpha[B(cm^\alpha + m^{\alpha+1} + m^{\alpha-1}) + b^\alpha]\right)\right\rangle\!\!\right\rangle, \tag{7}$$

where $\langle\!\langle\cdot\rangle\!\rangle$ denotes averaging over possible configurations of ξ^μ [25]. See the Supplemental Material for derivation [23].

We numerically calculate the overlaps at a fixed point ($\dot{m}^\mu = 0$). The initial condition is $m^\mu = 1$ for $\mu = \mu_{\mathrm{init}}$ and $m^\mu = 0$ otherwise. When the number of patterns is small, we can exactly evaluate the quantities over all possible combinations of $\{\xi^\mu\}$. However, when we increase the number of patterns, the number of possible configurations of ξ^μ (i.e., sublattices) rapidly diverges and becomes intractable. To overcome this difficulty, we perform the Monte Carlo approximation of the mean-field equation by sampling a finite but large enough number of $\{\xi^\mu\}$ (typically, 10^6 samples). Furthermore, $C(\nu)$, correlations between two attractors centered on the patterns μ_{init} and $\mu_{\mathrm{init}} + \nu$, were also calculated for vanishing external inputs ($b^\mu = 0$) (see Supplementary Material [23] for the procedure). We share PYTHON codes for these calculations in Ref. [26].

An unexpected finding is that negative values of c significantly expand the span of interstimulus association among correlated attractors. Figures 1(a) and 1(b) show solutions for $p = 0.5$ and $P = 21$ obtained without external inputs and the Monte Carlo approximation. When c is positive ($c = 1.5$), our model reproduces the result shown by Griniasty *et al.* [10] in which the neighboring attractors are significantly correlated up to the distance of five. In contrast, when c is negative ($c = -1.5$), the correlation distance easily extends beyond 10. To see how the correlation behaves at longer distances, we obtained solutions for $P = 71$ by using the Monte Carlo approximation [Figs. 1(c) and 1(d)]. The results show that the correlation between attractors extends up to the distance of 20, which is four times longer than that for $c = 1.5$.

We quantitatively study how the value of c changes attractors in our model by calculating approximate solutions in the range $-3 \leq c \leq 3$ for $p = 0.5$ and $P = 71$. We calculated two measures: the maximum overlap that indicates successful memory retrieval, and the span of correlation N_c defined as

$$N_c = \min\{\nu | C(\nu) < 10^{-2}\} - 1. \tag{8}$$

If only the nearest neighbor attractors have correlations greater than 10^{-2}, N_c is unity. The maximum overlap takes

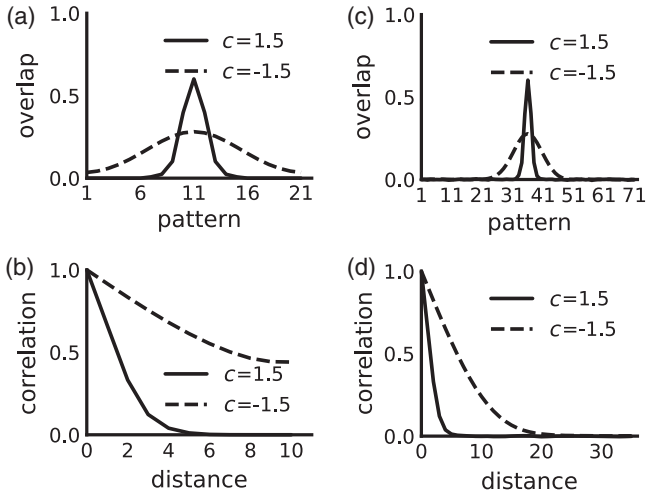

FIG. 1. Extended interstimulus association in anti-Hebbian learning. Overlaps between a reference attractor ($\mu_{\text{init}} = 11$) and memory patterns (a) and correlations between attractors (b) were calculated without the Monte Carlo approximation. Parameters are $P = 21$, $p = 0.5$. (c), (d) Overlaps and correlations were calculated with the Monte Carlo approximation for $\mu_{\text{init}} = 36$. Parameters are $P = 71$, $p = 0.5$.

non-zero values only for $c > -2$ [Fig. 2(a)]. The value of N_c is robustly around five for $0 < c < 2$ [Fig. 2(b)] and becomes 0 for $c > 2$ (that is, no correlated attractors exist in this range). Thus, for $c > 0$ the threshold value 10^{-2} reproduces the results obtained by Griniasty *et al.* ($N_c = 5$) [10]. By contrast, as c is decreased from 0 to -2, N_c gradually increases even beyond 20 [Fig. 2(b)]. We can observe a similar expansion of correlation for biased patterns ($p = 0.1$) [Figs. 2(c) and 2(d)]. In sum, the extended span of correlation is generally found in the range $-2 < c < 0$ regardless of the bias of memory patterns.

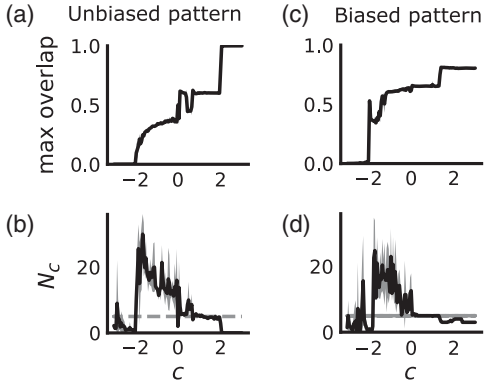

FIG. 2. Parameter dependence of the maximum overlaps (a) and N_c (b) for unbiased stimulus patterns ($P = 71$, $p = 0.5$). Black lines and gray areas show the mean and standard deviation over five different samples, respectively. The gray dashed line in (b) indicates $N_c = 5$ for comparison with the previous result [10]. (c), (d) Maximum overlaps and N_c for biased patterns ($P = 71$, $p = 0.1$).

To further explore the functional implications of this model, we investigated how the attractor states are modulated by external inputs at various values of c. We applied an external input with the amplitude $b^{\mu_{\text{input}}} = 0.1$ to the memory pattern μ_{input} ($\mu_{\text{input}} = \mu_{\text{init}} + 20$) in the attractor state centered at the pattern μ_{init} [see Fig. 1(a)]. Figure 3(a) shows the resultant attractor states. For $c = 1.5$, the input does not greatly affect the initial attractor state, creating a tiny additional peak representing the input pattern μ_{input}. In contrast, the attractor state completely shifts towards the input pattern for $c = -1.5$, indicating that negative c increases the sensitivity of the network to external input. We quantitatively clarified this effect by calculating the threshold for the shifts, namely, the minimum value of $b^{\mu_{\text{input}}}$ to shift the gravity center of the overlap distribution μ_{center} towards the input pattern

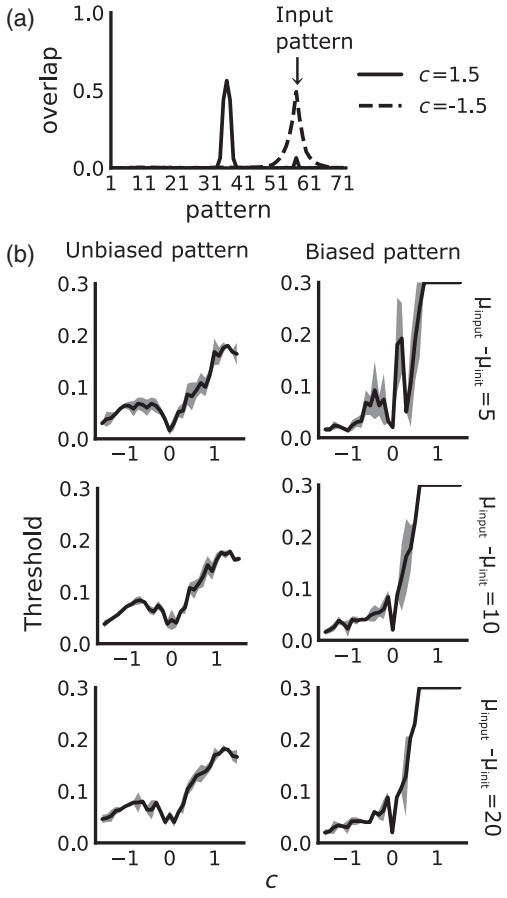

FIG. 3. The parameter dependence of sensitivity to external inputs. (a) Overlaps between a reference attractor ($\mu_{\text{init}} = 36$, $\mu_{\text{input}} = 56$) and memory patterns. (b) The relationship between the value of c and the threshold input strength for the attractor shift. The results are shown only for $b^{\mu_{\text{input}}} \leq 0.3$. When the attractor did not shift in $b^{\mu_{\text{input}}} \leq 0.3$, the threshold was plotted as $b^{\mu_{\text{input}}} = 0.3$. We used $p = 0.5$ for unbiased patterns, and $p = 0.1$ for biased patterns, and the number of patterns is $P = 71$. Black lines and gray areas show the mean and standard deviation, respectively, over five different samples.

($|\mu_{\rm center}-\mu_{\rm init}|>|\mu_{\rm center}-\mu_{\rm input}|$). In both unbiased ($p=0.5$) and biased ($p=0.1$) patterns, this threshold is generally lower for $c<0$ than for $c>0$ for all choices of the initial distance to input ($\mu_{\rm input}-\mu_{\rm init}=5,\ 10,\ 20$) [Fig. 3(b)]. Thus, changing the balance of local and global inhibition (towards local inhibition) increases the sensitivity of correlated memory states to external events and hence lowers the threshold of the shift of the attractor.

We can qualitatively study these properties by means of the following energy function:

$$E=-\sum_{i,j}J_{ij}S_iS_j-\sum_i\theta_iS_i. \tag{9}$$

Sequential update of the neural activity by Eq. (1) monotonically decreases this energy [4]. Defining $\theta_i=\sum_\mu b^\mu\hat{\xi}_i^\mu$, we can rewrite the energy function in terms of pattern overlaps as

$$\begin{aligned}E&\propto -c\sum_{\mu=1}^{P}(m^\mu)^2-2\sum_{\mu=1}^{P}m^\mu m^{\mu+1}-N\sum_{\mu=1}^{P}b^\mu m^\mu\\&=-\sum_{\mu=1}^{P}(c'(m^\mu)^2+Nb^\mu m^\mu)+\sum_{\mu=1}^{P}(m^\mu-m^{\mu+1})^2,\end{aligned} \tag{10}$$

where $c'=c+2$. Note that $0<c'<4$ if $-2<c<2$. When $b^\mu=0$ and $c'<0$, this function is trivially minimized when all overlaps vanish. By contrast, if $c'>0$, energy minimization requires the maximization of $(m^\mu)^2$ under the penalty of $(m^\mu-m^{\mu+1})^2$. Without the penalty, the model is equivalent to the standard Hopfield model in which a single nonvanishing overlap minimizes the energy. However, the penalty term creates a broad distribution of nonvanishing overlaps for small values of c'. As c' increases, the relative contribution of the penalty term becomes smaller, narrowing the distribution. Decreases in c' also make the relative impact of b^μ greater, increasing input sensitivity.

Finally, we explored the role of inhibitory plasticity balancing inhibition with excitatory engrams. Excitatory synaptic weights [Eq. (4)] can arise from Hebbian learning, but whether experimentally suggested inhibitory learning rules [7,13] may create inhibitory weights [Eq. (5)] remains unclear. Here, we considered a network of $N=2000$ neurons storing $P=50$ sparse memory patterns ($p=0.05$) in excitatory synapses [Fig. 4(a)]. To clarify the distinct roles of inhibition, we divided the inhibitory network into local and global inhibition. The inhibition ratio a between the two inhibitory inputs was modulated in the range $0<a<1$, where $a=1$ and $a=0$ refer to fully local and fully global inhibition, respectively. This parameter plays a similar role to c in Eq. (5) and c' in Eq. (10). Inhibitory synaptic weights were initially zero, and the plasticity rule that imposes E-I balance on postsynaptic neurons was used as in [13]. See Supplemental Material SM [23] for the detail of the model.

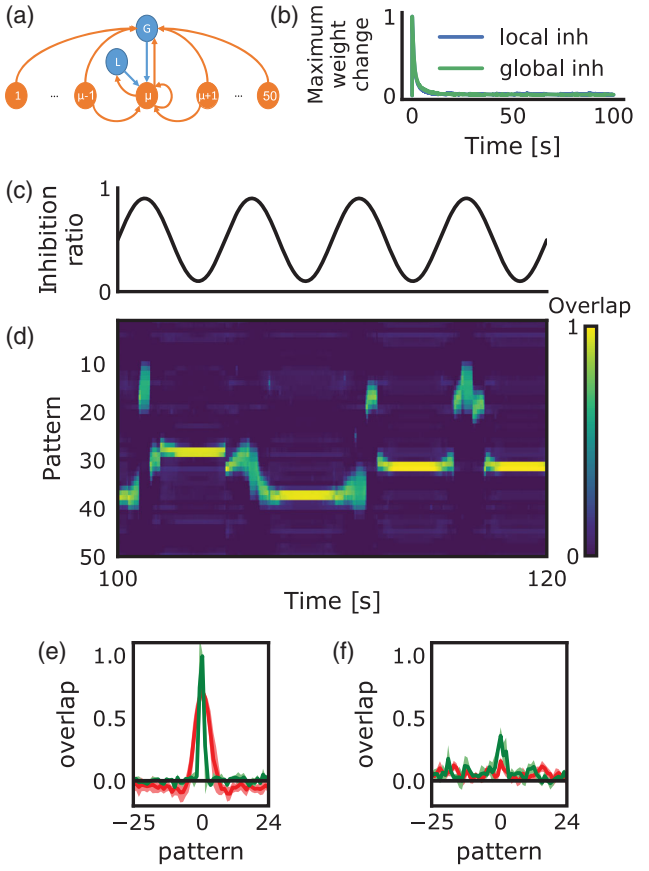

FIG. 4. Inhibitory regulations of correlated attractors. (a) A network model of excitatory neurons (orange) is regulated by local (L) and global (G) inhibitory neurons (blue). (b) The relaxation of synaptic weights of local and global inhibition during the initial adaptation period. We plotted the maximum weight changes every 5 ms $\{\max_{i,\mu}[w^{\rm L}_{i\mu}(t+5\ {\rm ms})-w^{\rm L}_{i\mu}(t)],\ \max_i[w^{\rm G}_i(t+5\ {\rm ms})-w^{\rm G}_i(t)]\}$, where the changes were normalized such that their peak is 1. The inhibition ratio was $a=0.9$. (c) The inhibition ratio was sinusoidally modulated during the test period (time >100 s). (d) Time evolution of overlaps between memory patterns and simulated neural activities is shown during the test period. Overlaps exceeding 1 or below 0 were truncated. We note that the overlap distributions are broadened and occasionally drifted at the peak times of the inhibition ratio (local-inhibition-dominant state). (e) Means (lines) and standard deviation (shaded areas) of overlap distributions were calculated for the epochs of low ($a<0.2$: green) and high ($a>0.8$: red) inhibition ratios. Memory patterns were renumbered such that memory pattern 0 takes the maximum value. (f) The same as (e), but after randomization of inhibitory synaptic weights.

The value of a was initially kept constant ($a=0.9$) to relax the weights of local and global inhibition onto the equilibrium values, which corresponds to the condition $c'\approx 0$ [Fig. 4(b)]. Then, we terminated the learning process and tested the responses of the network. In this test, we periodically modulated the value of a [Fig. 4(c)] and

intermittently stimulated neurons encoding a memory pattern, which was randomly selected for every stimulus. Driven by the external inputs, the center of correlated attractors drifted only at the peaks of a [Fig. 4(d)]. As shown in Fig. 4(e), the width of overlap distributions was also maximally broadened at the peaks. These changes of the width of overlaps and input sensitivity are consistent with the results of our mean-field analysis. Random changes in inhibitory synaptic weights within the $\pm 50\%$ of the learned values collapsed the attractor states [Fig. 4(f)], suggesting that the learned inhibitory weights are necessary for stabilizing the attractors. These results demonstrate the roles of inhibitory plasticity inducing E-I balance and the global-versus-local inhibition ratio for regulating attractor states.

We may speculate the biological mechanisms and functional implications of the regulations of local versus global inhibitory circuits. Parvalbumin- (PV+) and somatostatin-expressing (SOM+) interneurons are considered to regulate local and global inhibition, respectively [27,28]. Activity of SOM + interneurons are inhibited [14,15] and excitability of PV + interneurons is facilitated [16,17] by acetylcholine, respectively, implying that the cholinergic modulations of these neurons are candidate mechanisms to regulate the balance between the two inhibitory effects. There are, however, controversial experimental results [18–20], and further experimental clarification is necessary for the cholinergic mechanisms.

In our model, the extended span of simultaneously recalled memory patterns is accompanied by their increased sensitivity to external input, which can enhance the influence of context (sensory) information on the retrieval of extended (hence, uncertain) memory states. Related to this, acetylcholine level was proposed to reflect the uncertainty of top-down information in sensory processing [29]. Furthermore, the span of correlated attractors may determine the timescale of episodic events during encoding and retrieval. Actually, several cortical regions including the hippocampus are engaged in segregating and concatenating (i.e., chunking) episodic events on multiple timescales [30]. The roles of modifiable correlated attractors are open for future experimental and computational studies.

This work was partly supported by Grant-in-Aid for Specially Promoted Research (Grant No. 18H05213) and Grant-in-Aid for Scientific Research on Innovative Areas (Grant No. 19H04994) from Japan Society for the Promotion of Science (JSPS).

*tatsuya.haga@riken.jp
†tomoki.fukai@oist.jp

[1] D. O. Hebb, *The Organization of Behavior* (Wiley & Sons, New York, 1949).
[2] G. Q. Bi and M. M. Poo, J. Neurosci. **18**, 10464 (1998).
[3] R. K. Mishra, S. Kim, S. J. Guzman, and P. Jonas, Nat. Commun. **7**, 11552 (2016).
[4] J. J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).
[5] L. Deuker, J. L. Bellmund, T. Navarro Schröder, and C. F. Doeller, eLife **5**, 1 (2016).
[6] A. C. Schapiro, N. B. Turk-Browne, K. A. Norman, and M. M. Botvinick, Hippocampus **26**, 3 (2016).
[7] H. C. Barron, T. P. Vogels, T. E. Behrens, and M. Ramaswami, Proc. Natl. Acad. Sci. U.S.A. **114**, 201701812 (2017).
[8] Y. Miyashita, Nature (London) **335**, 817 (1988).
[9] V. Yakovlev, S. Fusi, E. Berman, and E. Zohary, Nat. Neurosci. **1**, 310 (1998).
[10] M. Griniasty, M. V. Tsodyks, and D. J. Amit, Neural Comput. **5**, 1 (1993).
[11] D. J. Amit, N. Brunel, and M. V. Tsodyks, J. Neurosci. **14**, 6435 (1994).
[12] L. F. Cugliandolo and M. V. Tsodyks, J. Phys. A **27**, 741 (1994).
[13] T. P. Vogels, H. Sprekeler, F. Zenke, C. Clopath, and W. Gerstner, Science **334**, 1569 (2011).
[14] Y. Fu, J. M. Tucciarone, J. S. Espinosa, N. Sheng, D. P. Darcy, R. A. Nicoll, Z. J. Huang, and M. P. Stryker, Cell **156**, 1139 (2014).
[15] H. J. Pi, B. Hangya, D. Kvitsiani, J. I. Sanders, Z. J. Huang, and A. Kepecs, Nature (London) **503**, 521 (2013).
[16] D. E. Pafundo, T. Miyamae, D. A. Lewis, and G. Gonzalez-Burgos, J. Physiol. **591**, 4725 (2013).
[17] F. Yi, J. Ball, K. E. Stoll, V. C. Satpute, S. M. Mitchell, J. L. Pauli, B. B. Holloway, A. D. Johnston, N. M. Nathanson, K. Deisseroth, D. J. Gerber, S. Tonegawa, and J. J. Lawrence, J. Physiol. **592**, 3463 (2014).
[18] N. Chen, H. Sugihara, and M. Sur, Nat. Neurosci. **18**, 892 (2015).
[19] H. J. Alitto and Y. Dan, Front. Syst. Neurosci. **6**, 1 (2013).
[20] J. Obermayer, T. S. Heistek, A. Kerkhofs, N. A. Goriounova, T. Kroon, J. C. Baayen, S. Idema, G. Testa-Silva, J. J. Couey, and H. D. Mansvelder, Nat. Commun. **9**, 4101 (2018).
[21] J. Buhmann, R. Divko, and K. Schulten, Phys. Rev. A **39**, 2689 (1989).
[22] M. Tsodyks and M. V. Feigel'man, Europhys. Lett. **6**, 101 (1988).
[23] See the Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.123.078101 for methodological details.
[24] N. Brunel, Neural Comput. **8**, 1677 (1996).
[25] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **32**, 1007 (1985).
[26] See https://github.com/TatsuyaHaga/antihebbhopfield.
[27] H. Adesnik, W. Bruns, H. Taniguchi, Z. J. Huang, and M. Scanziani, Nature **490**, 226 (2012).
[28] A. Litwin-Kumar, R. Rosenbaum, and B. Doiron, J. Neurophysiol. **115**, 1399 (2016).
[29] A. J. Yu and P. Dayan, Neural Netw. **15**, 719 (2002).
[30] C. Baldassano, J. Chen, A. Zadbood, J. W. Pillow, U. Hasson, and K. A. Norman, Neuron **95**, 709 (2017).