Repeated sequential learning increases memory capacity via effective decorrelation in a recurrent neural network

Tomoki Kurikawa*
Kansai Medical University, Shinmachi 2-5-1, Hirakata, Osaka, Japan

Omri Barak
Rappaport Faculty of Medicine and Network Biology Research Laboratories, Technion–Israeli Institute of Technology, Haifa, Israel

Kunihiko Kaneko
Graduate School of Arts and Sciences, University of Tokyo, Komaba 3-8-1, Meguro-ku, Tokyo, Japan

(Received 29 June 2019; accepted 12 May 2020; published 9 June 2020)

Memories in neural systems are shaped through the interplay of neural and learning dynamics under external inputs. This interplay can result in either overwriting or strengthening of memories as the system is repeatedly exposed to multiple input-output mappings, but it is unclear which effect dominates. By introducing a simple local learning rule to a neural network, we found that the memory capacity is drastically increased by sequentially repeating the learning steps of input-output mappings. We show that the resulting connectivity decorrelates the target patterns. This process is associated with the emergence of spontaneous activity that intermittently exhibits neural patterns corresponding to embedded memories. Stabilization of memories is achieved by a distinct bifurcation from the spontaneous activity under the application of each input.

DOI: 10.1103/PhysRevResearch.2.023307

I. INTRODUCTION

Time always moves forward. Accordingly, the brain learns to appropriately respond to various inputs by sequential exposure. In neural systems, synaptic connections are modified to shape neural dynamics such that the applied stimulus and desired response are adequately represented therein. After learning, the stimulus is represented according to the shaped neural dynamics [1–5]. How memories are successively embedded into neural dynamics through the interplay between the neural dynamics and learning process is a crucial question in neuroscience.

To understand the representation of memories in neural systems, associative memory models are often studied. In conventional models [6–8], multiple memories are embedded into corresponding attractors and are generated by a simple learning rule. In spite of their success, however, neural dynamics in these models are often decoupled from those of synapses—synapses are slowly modified according to the desired targets, and the faster neural dynamics of relaxation to memory attractors are studied independently [1,9] (but see also [10]).

In contrast, we previously proposed a novel associative memory model [11] in which input-output associations are learned on the background of chaotic spontaneous dynamics, while synaptic and neural dynamics coevolve. In that study, however, each pattern was only presented once during learning, and existing memories were gradually eroded as new patterns were learned.

In the present study, we first introduce a theoretical formulation for a sequential and repeated learning process. By studying this learning process, we investigated if all memories are able to be successfully stored by repeated learning. If so, we then address what kind of neural network emerges during this process and how memories are represented in neural dynamics upon input. We also study spontaneous dynamics without input, which were suggested to be involved in computations in neural systems [12–17].

II. RESULTS

A. Memory capacity

We consider a model that consists of N continuous rate-coding neurons to memorize M input-output (I/O) mappings (indexed by μ). The activity $\boldsymbol{x} = \{x_i\}$ $(i = 1, 2, \ldots, N)$ is set between -1 and 1 and evolves according to

$$\dot{x}_i = \tanh\left[\beta\left(\sum_{j\neq i}^{N} J_{ij}x_j + \gamma\eta_i^{\mu}\right)\right] - x_i, \qquad (1)$$

*kurikawt@hirakata.kmu.ac.jp

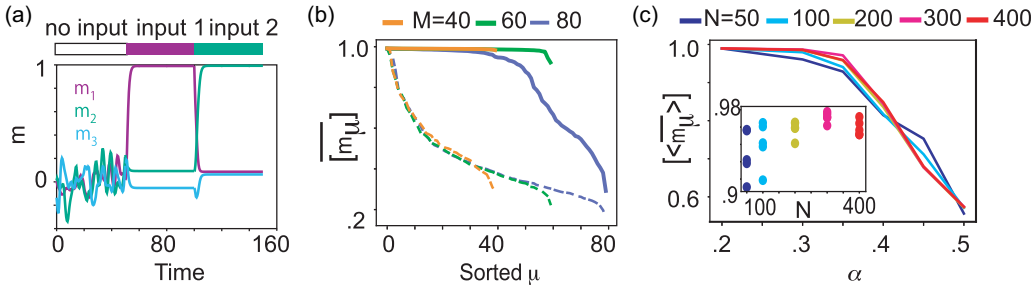

FIG. 1. Recall performance in trained networks. (a) Overlaps $m_\mu = \Sigma_i x_i \xi_i^\mu / N$ are shown for $\mu = 1, 2, 3$. Inputs 1 and then 2 are applied for $50 \leqq t < 100$ and $100 \leqq t < 150$, respectively. (b) Overlaps $[\overline{m_\mu}]$ averaged over time; networks and trials are plotted for three values of M, after learning every map once ($T = M$ steps, dashed) or many times ($T = 30M$, solid). Patterns are sorted according to decreasing overlaps (indicated by "sorted μ" in this and following figures). Network size is $N = 200$. (c) Scaling of memory capacity for $T = 30M$. Average overlap $[\langle \overline{m_\mu} \rangle]$ averaged over all patterns μ as a function of $\alpha = M/N$ for different network sizes. Inset: Overlap $\langle \overline{m_\mu} \rangle$ at $\alpha = 0.35$. Results for (b) and (c) are averages from five network realizations, and five trials in each realization.

where J_{ij} denotes a connection from the jth to the ith neuron, an N-dimensional vector $\boldsymbol{\eta}^\mu$ is an input pattern, and γ and β are input strength and activation function gain, respectively.

For each input $\boldsymbol{\eta}^\mu$, we set an N-dimensional vector $\boldsymbol{\xi}^\mu$ as a target. These input and target patterns are generated as random N-bit binary patterns, with probabilities $P(\xi_i = \pm 1) = P(\eta_i = \pm 1) = 1/2$. In the presence of each input $\boldsymbol{\eta}^\mu$, the corresponding target $\boldsymbol{\xi}^\mu$ is required to be recalled, i.e., an attractor matching $\boldsymbol{\xi}^\mu$ is generated. The learning process is required to modify the connectivity $\boldsymbol{J}$ such that the network recalls the targets.

Previously [11], we showed such a memory structure is formed through a simple learning rule: $\dot{J}_{ij} = (\epsilon/N)(\xi_i^\mu - x_i)x_j$. To enable repeated sequential learning, we added a decay term to the previous rule that maintains the norm of the connectivity constant despite the ongoing exposure to new stimuli:

$$\dot{J}_{ij} = (\epsilon/N)\left(\xi_i^\mu - x_i\right)(x_j - h_i J_{ij}), \tag{2}$$

where $h_i = \sum_{j \neq i} J_{ij} x_j$, and ϵ is the learning rate. Indeed, the norm changes according to $d(\sum_{j \neq i} J_{ij}^2)/dt \propto (1 - \sum_{j \neq i} J_{ij}^2)$. We thus initialize $\sum_{j \neq i} J_{ij}^2 = 1$ by choosing J_{ij} with the binary values $P[J_{ij} = \pm(N-1)^{-1/2}] = 1/2$, and the diagonal entries of $\boldsymbol{J}$ are kept at zero during the entire process. Learning stops automatically when the neural activity matches the target, because $\dot{J}_{ij} = 0$; otherwise, the learning process continues. Here we imposed M I/O maps successively in the following manner: An input $\boldsymbol{\eta}^\mu$ is applied until learning is completed (this is called a single learning step) and then the following input is applied to learn the following corresponding target. This is done for a total of $T(>M)$ steps, where the first M steps correspond to the ordered maps ($\mu = 1, 2, \ldots, M$), and the following $T - M$ steps are in random order.

Figure 1(a) shows the recall processes in response to two input patterns after learning. Without input, spontaneous activity has occasional similarities to the learned patterns. Once input $\boldsymbol{\eta}^1$ is applied, the neural dynamics are modified and rapidly converge to $\boldsymbol{\xi}^1$. Similarly, the later introduction of $\boldsymbol{\eta}^2$ leads to convergence to $\boldsymbol{\xi}^2$. In this and the following analysis we set $\gamma = 1.0$, $\beta = 4.0$, $N = 200$, and $\epsilon = 0.03$, unless otherwise stated.

We first analyzed how repeated learning enhances the memory capacity. For this purpose, we computed the temporal average of all overlaps $[\overline{m_\mu}] = [\Sigma_i \overline{x_i} \xi_i^\mu / N]$ in the presence of inputs $\boldsymbol{\eta}^\mu$ ($\mu = 1, \ldots, M$), where the symbols $\overline{\cdots}$ and $[\cdots]$ represent averages over time or over networks and trials, respectively. Figure 1(b) shows that after learning each map only once ($T = M$ learning steps, dashed lines), networks can recall only one or two targets perfectly and overlaps with other targets decrease rapidly, independent of M. After learning these targets more and more times ($T = 30M$, solid lines), however, recall performance increases and all of the targets are recalled perfectly for small M, with a decrease in performance as M increases. For the value of $N = 200$, we found the limit of memory between $M = 60$ and 80, namely, $\alpha = M/N = 0.3$–0.4.

To evaluate this memory capacity more accurately, we calculated the averaged overlap $[\langle [m_\mu] \rangle_\mu]$ and plotted it for different N in Fig 1(c), where $\langle \cdots \rangle_\mu$ represents average over maps. After $T = M$ learning steps, the average overlap decreases rapidly, while, after $T = 30M$ learning steps, the overlap is maintained at around unity up to $\alpha = 0.35$ (see Fig. 6 in the Appendix for more details). Therefore, the capacity of the present model is estimated to be $\alpha_c = 0.35$ [18]. To explore dependence of the memory on T, we examined $[\langle m_\mu \rangle_\mu]$ for different T. We found that the memory capacity increases monotonically as T increases and saturates around $T = 20M$ (see Fig. 6). Thus, we studied the behavior for $T = M$ and $30M$ as typical samples in the earlier and later stages of learning and that for $\alpha = 0.3$ unless otherwise stated.

Enhancement in the memory capacity after iterative learning is not trivial, but depends on the learning rate ϵ, the activation function gain β, and input strength γ. As shown in Fig. 6, the memory capacity is decreased as ϵ increases and β decreases, as well as for much larger and much smaller γ. In particular, the memory capacity for $\epsilon \geqslant 1$ and $\beta \leqslant 1$ decreases to almost a single map, and cannot be increased by repeated learning. This result indicates that the nature of neural dynamics and a relation between the timescales of neural and learning dynamics, as well as input strength, are important to enhance memory capacity through repeated learning.

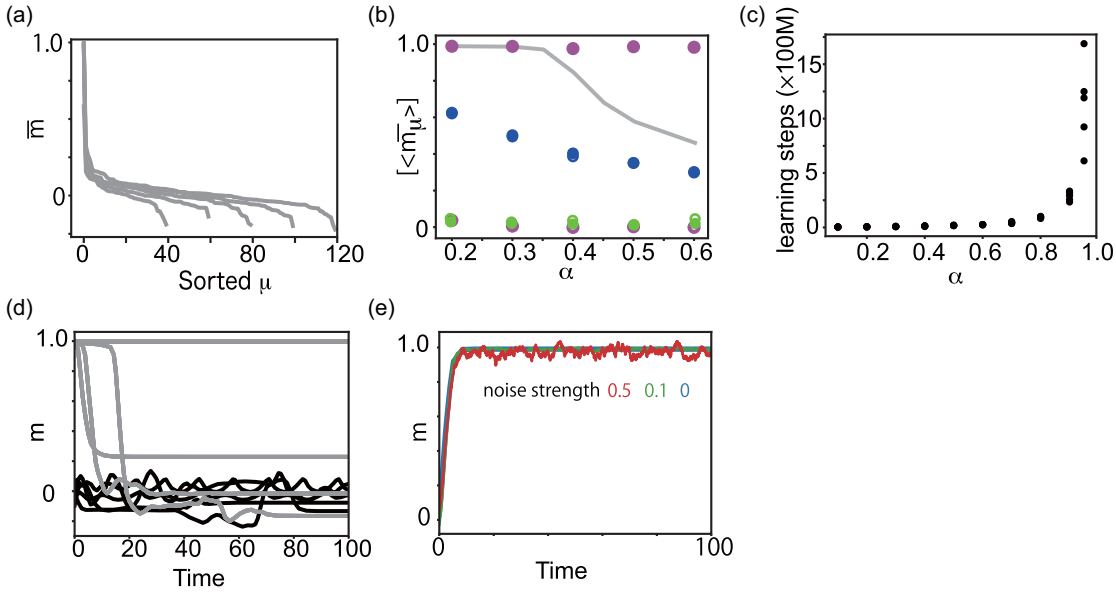

FIG. 2. Comparison of memory performance with other models. (a) Overlaps with targets are illustrated for $M = 40, 60, 80, 100, 120$ for the Perceptron model, same as that in Fig. 1(b). (b) Recall performance as a function of memory load for different models. Magenta, blue, and green circles indicate performance in the modified pseudoinverse model, Hopfield-type network, and Perceptron model, respectively. Filled and empty circles represent initial conditions near and far from the target, respectively. These two conditions are overlapping for Hopfield (blue) and Perceptron (green). Gray line is the performance in our model for reference. (c) The number of learning steps required for the modified pseudoinverse model to learn αN maps is plotted. (d) Stability of modified pseudoinverse attractors. Five trajectories were initialized either from the vicinity of the target (gray, $|\boldsymbol{x} - \boldsymbol{\xi}| < 0.001$) or from randomly chosen states (black). $\alpha = 0.3$. (e) Similar to (d), for our model. Distance from targets was 0, 0.1, or 0.5. $\alpha = 0.3$ and $N = 200$.

B. Comparison with other models

To put our learning rule's performance into perspective, we analyzed memory capacity in three different models: a Perceptron online learning rule, a pseudoinverse model without neural dynamics [19–21], and Hopfield-type connectivity [11,22], which is a function of the desired maps without online learning.

1. Perceptron model

As in our model, connectivity is modified online in parallel to the evolution of neural dynamics. Each learning step consists of presenting $\boldsymbol{\eta}^\mu$ and modifying connectivity until convergence according to

$$\dot{J}_{ij} = (\epsilon/N)\xi_i^\mu x_j. \tag{3}$$

After learning each map, $\boldsymbol{J}$ is normalized to $\Sigma_j J_{ij}^2 = 1$. Then, a new map is applied and a new learning step begins. Note that if the target pattern is stable in the presence of the corresponding input, neural dynamics rapidly converge to the target and the learning process is terminated, thus, modification of the connection is quite small.

We explored the performance of this model for various values of M after $10M$ learning steps. We found that networks could only recall the last presented map, forgetting all earlier ones [Fig. 2(a)]. This behavior was not improved for smaller ϵ. Thus, the performance of this model is much lower than ours [green circles in Fig. 2(b)].

We speculate that the postsynaptic factor in the Perceptron rule leads to larger changes to the weights, that destabilize previous memories.

2. Pseudoinverse model

In the pseudoinverse model, connectivity is modified according to

$$\Delta J_{ij} = (1/N)\left(\xi_i^\mu - u_i^\mu\right)\xi_j^\mu, \tag{4}$$

where the local field $u_i^\mu = \Sigma_{j\neq i} J_{ij}\xi_j^\mu + \eta_i^\mu$. Diagonal elements are kept at zero. We use the symbol ΔJ and not $\dot{J}$ to stress that connectivity is modified without neural dynamics. This is an adaptation of the original pseudoinverse model [19–21], to account for the heteroassociative nature of our task. During learning, neural dynamics do not run and $\boldsymbol{x}$ is quenched at $\boldsymbol{\xi}$. The learning process is maintained until the network can memorize all of the desired maps.

After learning with the modified rule, neural dynamics run according to Eq. (1). The original autoassociative rule was proved to achieve a capacity of $\alpha = 1.0$ [21]. We thus hypothesized that $\boldsymbol{\xi}^\mu$ is an attractor in the presence of $\boldsymbol{\eta}^\mu$ for all μ ($\alpha < 1.0$).

To verify this hypothesis, We tested the memory capacity of the modified pseudoinverse model. We plotted learning steps for a network required to learn all of αN maps for $N = 200$ in Fig. 2(c). As α increases, the number of required steps rapidly increases and diverges for $\alpha = 1.0$. Thus, the capacity of the modified pseudoinverse model is just below $\alpha = 1.0$.

We investigated the stability of target attractors after learning in the modified pseudoinverse model. Figure 2(d) shows two types of neural trajectories: one beginning from randomly chosen initial states (black) and the other beginning from the vicinity of the target (gray). We see that even neural states

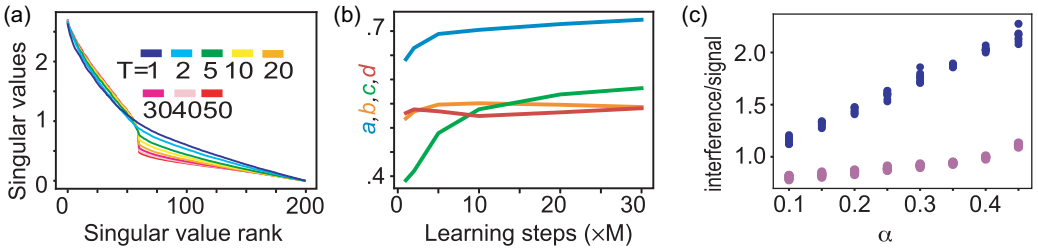

FIG. 3. (a) Evolution of singular values of the connectivity matrix through learning. Averages over five learned networks with $N = 200$ and $M = 60$. Note the discontinuity at the Mth singular value as learning progresses. (b) Evolution of connectivity components $\langle \Sigma_\mu (s^{\kappa,\mu})^2 \rangle_{1 \leqslant \kappa \leqslant M}$ for $s = a, b, c, d$ during the learning process with chance level $M/N = 0.3$. (c) Interference to signal ratio as a function of memory load α for early (blue, $T = M$) and late (magenta, $T = 30M$) learning stages.

beginning from the vicinity escape from the target, meaning that the basin of attraction of the target attractor is quite small. Thus, the learned attractors have a very small basin of attraction.

In contrast, the attractors corresponding to the targets formed through our learning process are very stable. Indeed, neural trajectories from randomly chosen initial states converge to the target even with noise as shown in Fig. 2(e).

To statistically confirm these observations, we analyzed and plotted recall performance in the modified pseudoinverse model with the random initial states condition (open circles in magenta) and the vicinity of the target condition (filled circles in magenta) in Fig. 2(b).

3. *Hopfield-type model*

The Hopfield-type model [22] is a variant of the Hopfield network [7] in which the connectivity $J = \Sigma_\mu (\boldsymbol{\xi}^\mu - \boldsymbol{\eta}^\mu)(\boldsymbol{\xi}^\mu + \boldsymbol{\eta}^\mu)^t/N$ is given without a learning process. For this connectivity, we calculated recall performance with the neural dynamics in Eq. (1). Figure 2(b) (blue circles) shows that $[\langle \overline{m}_\mu \rangle_\mu]$ decreases rapidly as α increases. This is similar to the behavior of our model in the early phase ($T = M$).

To sum up, our learning rule was more efficient than the Perceptron, pseudoinverse, and Hopfield-type rules.

C. Decorrelation of inputs and targets through the learning process

What are the changes to the connectivity in our model, $\boldsymbol{J}$, that improve the model's performance with repeated learning? Motivated by the structure of connections in the Hopfield model [7] and our Hopfield-type models [11,22], we hypothesized that the connectivity matrix in our model consists mainly of linear combinations of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$. We thus followed the connectivity over the process of learning using a singular value decomposition: $\boldsymbol{J} = \boldsymbol{UPV}^t$, where $\boldsymbol{U}, \boldsymbol{V}$ are unitary matrices, and $\boldsymbol{P}$ is a diagonal matrix whose elements are the singular values. The singular values are plotted in the order of their magnitude in Fig. 3(a). They decrease continuously for earlier learning steps, while, after long learning, there appears a large discontinuity at M (60 maps in this example). This discontinuous drop at M at the late learning stage is also observed for different values of N. We conclude that M left and right singular vectors dominate the connectivity as learning progresses.

If our hypothesis is correct, these M vectors consist mainly of linear combinations of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ and the other $N - M$ left and right singular vectors are in the normal space to these combinations. To examine this point, we defined estimates for the contributions of $\boldsymbol{\xi}^{\boldsymbol{\mu}}$ and $\boldsymbol{\eta}^{\boldsymbol{\mu}}$ to $\boldsymbol{u}^\kappa$ ($\boldsymbol{v}^\kappa$) as $a^{\kappa,\mu}$ and $b^{\kappa,\mu}$ ($c^{\kappa,\mu}$ and $d^{\kappa,\mu}$), respectively [23]. These estimates are $a^{\kappa,\mu} = \Sigma_i u_i^\kappa \xi_i^\mu/N$, $b^{\kappa,\mu} = \Sigma_i u_i^\kappa \eta_i^\mu/N$, $c^{\kappa,\mu} = \Sigma_i v_i^\kappa \xi_i^\mu/N$, and $d^{\kappa,\mu} = \Sigma_i v_i^\kappa \eta_i^\mu/N$. Here, u_i^κ and v_i^κ are ith elements of κth left and right singular vectors, respectively.

Figure 3(b) shows $\langle \Sigma_\mu (a^{\kappa,\mu})^2 \rangle_{1 \leqslant \kappa \leqslant M}$, the average contribution of targets to all M dominant left singular vectors, as well as the corresponding quantities for b, c, and d. All of the values are much higher than chance level $M/N = 0.3$ meaning that the dominant M vectors mainly consist of targets and inputs. Furthermore, the contribution of targets (a, c) increases with learning.

We also found correlations between these contributions (Fig. 7 in Appendix), allowing us to estimate $b^{\kappa,\mu} \sim k^\mu a^{\kappa,\mu}$ and $d^{\kappa,\mu} \sim l^\mu c^{\kappa,\mu}$. Thus, the dominant M left and right singular vectors are decomposed as

$$\boldsymbol{u}^\kappa \sim \Sigma_\mu a^{\kappa,\mu}(\boldsymbol{\xi}^\mu + k^\mu \boldsymbol{\eta}^\mu), \quad \boldsymbol{v}^\kappa \sim \Sigma_\mu c^{\kappa,\mu}(\boldsymbol{\xi}^\mu + l^\mu \boldsymbol{\eta}^\mu), \quad (5)$$

Furthermore, we found that $a^{\kappa,\mu}$ is highly correlated with $c^{\kappa,\mu}$ across κ for a given μ, but not with $c^{\kappa,\nu}$ (for $\nu \neq \mu$, Fig. 7). Altogether, we can decompose $\boldsymbol{J}$ as

$$\boldsymbol{J} \sim \Sigma_{\mu\nu} S_{\mu\nu}(\boldsymbol{\xi}^\mu + k^\mu \boldsymbol{\eta}^\mu)(\boldsymbol{\xi}^\nu + l^\nu \boldsymbol{\eta}^\nu)^t, \quad (6)$$

where $S_{\mu\nu} = \Sigma_\kappa \rho^\kappa a^{\kappa,\mu} c^{\kappa,\nu}$ and ρ^κ is the κth singular value. To validate this approximation, we generated a new matrix consisting only of the M dominant singular vectors and evaluated its recall performance [Fig. 7(b)]. We found that this truncated matrix still allows good recall of a large number of patterns, while the overlaps are slightly reduced, probably due to the target and input components in the remaining singular vectors.

Equation (6) suggests that S controls which input is mapped to which target. We thus expected the off-diagonal terms to be small. Indeed, the lack of correlation between $a^{\kappa,\mu}$ and $c^{\kappa,\nu}$ shows that they are small. Additionally, these nondiagonal terms are reduced as learning progresses, as shown in Fig. 7(c).

It is important, however, that these nondiagonal terms are not zero. This is due to the correlations among the inputs and targets and between them. In the Hopfield-type model, $S_{\mu\nu} = \delta_{\mu\nu}$, and these correlations cause interference that limits capacity. Previously, this interference motivated the introduction

of an inverse correlation matrix into the connectivity, leading to increased capacity [19–21]. We thus wondered whether the remaining nondiagonal terms also confer some decorrelation in our case. To measure this, We define an interference term $\boldsymbol{O}^\lambda = \boldsymbol{J}\boldsymbol{\xi}^\lambda - S_{\lambda\lambda}(\boldsymbol{\xi}^\lambda + k^\lambda \boldsymbol{\eta}^\lambda)$. For a pure pseudoinverse matrix, the interference term is zero independent of α, while for the Hopfield-type case ($S_{\mu\nu} = \delta_{\mu\nu}$), this term is proportional to α. Figure 3(c) shows the interference-to-signal ratio, $1/\langle S_{\lambda\lambda}/|\boldsymbol{O}^\lambda|\rangle_\lambda$ for the connectivity shaped through learning in our model [24]. We found that the ratio increases with α in the earlier stage of learning (at $T = M$), but at the later stage of learning (at $T = 30M$), it stays below 1 up to $\alpha = 0.35$, which equals to the capacity, and then increases. These results (and further evidence in Fig. 7) show that indeed our learning rule reduces interference caused by correlation between targets and inputs.

This analysis raises a new question: What difference between our model and the modified pseudoinverse model causes the difference in the stability of memories? To answer this question, we analyzed the connectivity formed by the modified pseudoinverse model and calculated a, b, c, d in the same manner as done for our model. We found that the representation of inputs in the right singular vectors ($d^{\kappa,\mu}$) is around chance level [Fig. 6(f)], while in our model, it is much higher [Fig. 3(b)]. The higher $d^{\kappa,\mu}$, namely, the higher l^μ leads to $\boldsymbol{J}\boldsymbol{\eta}^\mu + \boldsymbol{\eta}^\mu \sim \boldsymbol{\xi}^\mu$ by omitting $O(N^{-1/2})$ according to Eq. (6), resulting in a flow in phase space from $\boldsymbol{x} = \boldsymbol{\eta}^\mu$ to $\boldsymbol{x} = \boldsymbol{\xi}^\mu$. Thus, the higher $d^{k\mu}$ likely contributes to the stability of the targets in our model compared to the pseudoinverse one.

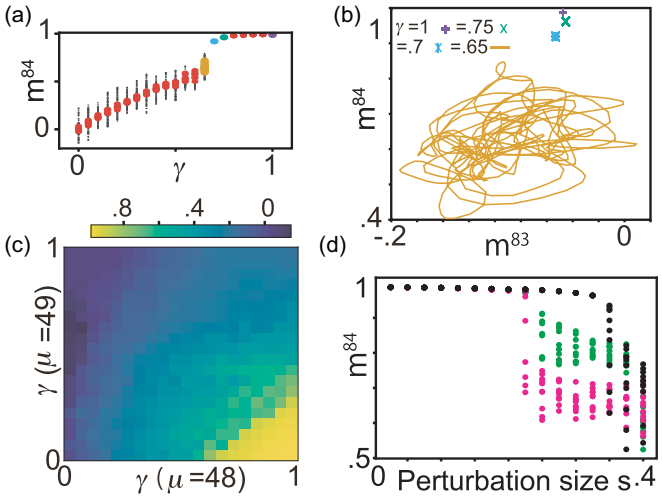

FIG. 4. (a) Bifurcation diagram showing the transition from a fixed point to chaotic oscillations as input strength decreases. Overlap with $\boldsymbol{\xi}^{\mathbf{84}}$ shown for $N = 300$, $T = 30M$, and $\alpha = 0.3$. Red symbols show the average over both time and ten trials, while black symbols show snapshots over time. Colored symbols refer to panel (b). (b) Change in qualitative behavior in phase space as input strength changes [different colors, indicated on panel (a)]. Neural activity is projected onto a two-dimensional plane defined by overlaps. (c) Bifurcation diagram for a mixture of two inputs (η^{48}, η^{49}). Color indicates overlap with $\overline{m_{48}}$. See also Fig. 8 in the Appendix. (d) Bifurcation diagram for three quenched perturbations of η^{84} (each in a different color). Overlap with the original pattern shown as a function of the scaled quenched perturbation pattern ζ.

D. Representation of memories

Next, we analyzed how memories are represented after learning. Because the network operates with constant, and not transient, inputs, we explored how the phase space varies as the input is modified. We study both a transition from spontaneous activity ($\gamma = 0$) to one input, and from one input to another one.

Figure 4(a) shows a bifurcation diagram against γ for $T = 30M$. Neural activity for $\gamma = 0$, i.e., spontaneous neural activity, oscillates around the origin. As γ increases, it moves towards a target while maintaining the oscillation amplitude. At a certain strength, an attractor of the neural dynamics bifurcates from the oscillation to form a fixed point corresponding to the target. Neural dynamics projected onto a two-dimensional plane is plotted around the bifurcation point in Fig. 4(b). Neural activity with a large-amplitude oscillation reduces into a fixed point corresponding to the target between $\gamma = 0.65$ and 0.7. Beyond the bifurcation point, the fixed point stays around the target as γ is increased. Thus, neural activity corresponding to target recall is clearly distinguished from other activities through a bifurcation and is stable against a change in γ beyond the bifurcation point.

We next considered the transition between inputs, by using a mixture of two learned inputs. As an example, the phase diagram of $\overline{m}_{48}$ against the amplitude of two learned inputs ($\boldsymbol{\eta}^{48}$, $\boldsymbol{\eta}^{49}$) is shown in Fig. 4(c). There is a region around the pure input of $\boldsymbol{\eta}^{48}$ where the overlap with $\boldsymbol{\xi}^{48}$ remains high despite the change in input. Once this boundary is crossed, a bifurcation leads to oscillating dynamics (Fig. 8 in the Appendix; note the vertical spread in the middle region). As the input approaches $\boldsymbol{\eta}^{49}$, a symmetric image emerges with respect to $\boldsymbol{\xi}^{48}$ (Fig. 8). These results show that each target is represented as a globally attracting fixed point, separated by bifurcations from the other target attractors.

We also evaluated the sensitivity of the attractor state to modifications in the exact pattern of the inputs. To this end, we added a quenched perturbation ζ of size s to the patterns:

$$\boldsymbol{\eta}'^{\mu} = \boldsymbol{\eta}^\mu + s\boldsymbol{\zeta} \tag{7}$$

($\boldsymbol{\zeta} \in \mathcal{R}^N$, with independent and identically distributed elements from a uniform distribution over $[-1, 1]$). Figure 4(d) shows a bifurcation diagram for increasing s and three realizations of ζ. There is a range of s values for which the global attracting state persists and the network continues to generate the correct target. Beyond a bifurcation point, depending on ζ (roughly $s = 0.2$ for magenta and green, $s = 0.35$ for black), neural activity becomes oscillatory, and the target is no longer recalled. Thus, the target attractor is stable with respect to perturbed input patterns.

E. Spontaneous activity

To close the analysis of neural dynamics, we explored how spontaneous activity without input is related to recall performance through learning. In early learning stages, spontaneous activity shows chaotic dynamics that intermittently approach and depart from the targets [Fig. 9(a) in the Appendix]. Here, a few targets are approached much more

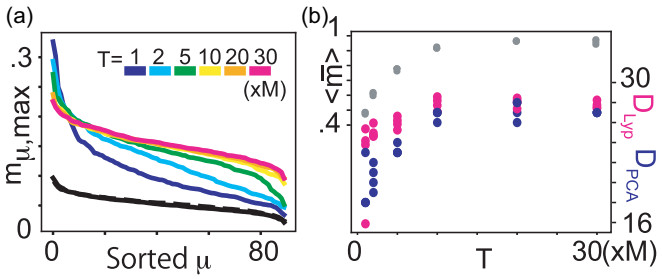

FIG. 5. Modification of spontaneous activity through learning for $M = 90$ and $N = 300$. (a) The maximal overlap of spontaneous activity with each target pattern during $t \in [0, 1000]$. Colors indicate different stages of learning. Black solid and dashed lines indicate overlaps with input and random patterns, respectively, for reference. (b) Recall performance and dimensionality of spontaneous activity as a function of learning steps T. Performance measured by average overlaps with targets in response to inputs (gray, left axis). Dimensionality measured by Lyapunov and PCA dimensions (magenta and blue, respectively, right axis). Lyapunov dimension (D_{Lyp}) is defined as $n + \Sigma_{i\leqslant n}\lambda_i/\lambda_{n+1}$; here, λ_i is the ith largest Lyapunov exponent and n is the largest index such that $\Sigma_{i\leqslant n}\lambda_i > 0$. PCA dimension ($D_{\mathrm{PCA}}$) is the number of PCs required to explain 0.8 of the variance. Four network realizations are shown for each measure.

than the rest [Fig. 5(a)], and these patterns (as well as their opposite patterns due to parity symmetry in our model) are also recalled better [Fig. 9(b)]. In later stages of learning, spontaneous activity approaches the targets in a more balanced manner [Fig. 5(a)], and performance is high for all targets [Fig. 9(b)]. We further analyzed neural dynamics by using principal components (PCs) analysis and measuring the Lyapunov dimension in Fig. 5(b). We found that the variation of the spontaneous activity is larger and more chaotic as learning progresses and recall performance is improved.

To confirm these relations between spontaneous activity and recall performance generally, we examined the spontaneous activity for different ϵ in Figs. 9(c)–9(g). For smaller ϵ, recall performance is higher and the spontaneous activity shows a high-dimensional activity which is close to all of the targets. As ϵ increases, recall performance decreases and the spontaneous activity turns to be low-dimensional, approaching only a few targets which are perfectly recalled. Finally, for a large value ($\epsilon = 5$), the input-less state converges to one of a few of fixed points which are target patterns and only these targets are successfully retrieved. These results support the relation between a rich spontaneous activity approaching the targets and heightened recall performance.

III. DISCUSSION

To sum, by studying neural networks that memorize I/O maps, we have shown how repeated learning stabilizes each memorized state and enhances memory capacity via the interplay between neural dynamics and learning. In usual sequential learning, e.g., gradient descent method [25,26] and palimpsest memory [27–29], connections are slowly shaped. The network's output moves in the direction of the desired target, but does not match it after a single step. In contrast, in the present study, connections are modified such that the network generates the correct target after each consolidation step. Thus, we can analyze how targets are embedded in neural dynamics and how the representation of these targets changes through learning. Interactions between neural dynamics and learning were investigated to reveal how neural representation is shaped in several studies [1,2,9,10,30–32]. These studies, however, did not focus on parametric effects of neural dynamics (e.g., the gain parameter, corresponding to β) and learning (e.g., the learning rate, corresponding to ϵ) on learning performance and representation of memories.

Spontaneous activity which intermittently reproduces stimulus-evoked patterns is commonly reported in visual [13,33] and auditory [34] cortices. Theoretical studies [9,35–38] demonstrated how the spontaneous activity is shaped through learning. Our study provides another simple learning rule to form such a spontaneous activity. A functional role for this phenomenon was suggested within a Bayesian inference framework [12–15]. In this view, spontaneous activity represents a prior distribution over stimuli, and neural dynamics utilize this prior to compute a posterior distribution from external stimuli. Our demonstration of a relation between the statistics of spontaneous activity and the functionality of recall performance is consistent with this approach.

More generally, properties of neural dynamics relevant for information processing were investigated [39–42], and the edge of chaos was suggested as an appropriate regime. Our model suggests that high-dimensional chaos with intermittent visits to learned patterns is suitable to produce appropriate targets in response to inputs. The role of such itinerant dynamics [43] has been discussed over decades [44–46], and the present study clearly demonstrates it.

Storing multiple patterns in the same circuit can cause interference, which can be alleviated by decorrelating the patterns via a pseudoinverse connectivity matrix [20]. Although calculating the inverse of a matrix requires global information, Diederich and Opper [21] showed that a local learning rule [Eq. (4)] can shape the inverse correlation matrix in the connectivity after repeated learning. In the Appendix, "Equivalence of our learning and pseudoinverse rules," we show that our rule approaches the Diederich rule in some limits, providing a partial explanation why our local, repeated learning shapes the connection matrix to decorrelate target patterns and enhances the memory capacity. Interestingly, our model provides more stable target attractors than the pseudoinverse model, while the capacity is decreased relative to that model. In our model, neural state evolves through phase space during learning and modifies the dynamics so that the current target is stable. Thus, learning a new memory has a larger effect on other areas of phase space, resulting in a smaller capacity compared to pseudoinverse learning. Note, however, that this evolution through phase space during learning enables global and robust attraction of the neural state to each target. Thus, there is a trade-off between capacity and stability.

In the present study, in contrast to the standard associative memory [7,8,27–29], each memory is recalled through an input-induced bifurcation from the spontaneous neural

activity. After the completion of repeated learning, gradually introducing the external input leads to a transition from a state of chaotic spontaneous activity to one with a globally attractive memory fixed point. This bifurcation has wide margins, implying that the recalled memory is stable under some perturbations to its input pattern. While several studies considered how spontaneous chaotic dynamics bifurcate to oscillatory dynamics as external input strengthens [47,48], we demonstrated generation of the stable memory through bifurcation from the spontaneous neural dynamics. The stability of memory against input strength was observed in auditory [49] and olfactory [50] cortices and in Hippocampus [51]. In these cortices, neural activity patterns are discretely switched between two memory states depending on the intensity of sensory inputs and/or ratio of mixture of two different inputs. Our model provides a simple learning rule to form such memory representations and gives a prediction in terms of spontaneous activity properties and memory performance.

Finally, we discuss the biological plausibility of our model. Our network receives two inputs: one from a lower cortical area (or sensory input) and the other from a higher cortical area (or top down signal). The trained network can predict the sensory input from a top down signal. Specifically, the desired neural activities ($\boldsymbol{\xi}$) could be interpreted as a sensory signal which is injected into each neuron in our network. The top down signal serves as the input to our network ($\boldsymbol{\eta}$). Thus, our network is trained to map between sensory and top down signals by using a Hebbian rule (correlation between ξ_j and x_i) and an anti-Hebbian rule (correlation between x_i and x_j). After training, the network infers the sensory signal when the trained top down signal is applied.

ACKNOWLEDGMENTS

We thank David Colliaux for fruitful discussion. This work was partly supported by JSPS KAKENHI (Grants No. 18K15343 and No. 20H00123) and Hitachi The University of Tokyo for funding. O.-B. is supported by the Israeli Science Foundation (Grant No. 346/16).

APPENDIX

1. Recall performance as N increases

To evaluate memory capacity adequately, we investigated recall performance $[\langle \overline{m_\mu} \rangle_\mu]$ as N increases in Fig. 1(c). We found that recall performance for $\alpha = 0.35$ increases to unity as N increases, but that for $\alpha = 0.4$ it does not. Here, we analyzed detailed behaviors for $\alpha = 0.3, 0.35$. In Fig. 6(a), we plotted $[\overline{m_\mu}]$ sorted by its value. For smaller N and $\alpha = 0.3$, some maps are not recalled perfectly, but, as N increases, almost all patterns are recalled perfectly [Fig. 6(a)]. For $\alpha = 0.35$, in contrast, although recall performance is saturated for $N = 300, 400$, a few of the maps are still not recalled [Fig. 6(b)]. These results indicate that our network can recall a little less than $0.35N$ maps.

2. Change in recall performance during learning process

We studied how memory performance is changed through learning. We plotted $[\overline{m^\mu}]$ against μ with the increase of T. It is sorted in the order of magnitude of the overlap in Fig. 6(c). For the early learning stage, only a few targets are stored, while, for the later stage, the number of targets perfectly recalled increases rapidly. Here, we measured $[\langle \overline{m_\mu} \rangle_\mu]$ as recall performance in Fig. 6(d). For $N = 200$ and $M = 60$, we plot the recall performance against learning step T. The capacity increases rapidly up to $T = 10M$ and almost saturates at $T = 20M$.

3. Dependence of recall performance on ϵ, β, and γ

For $\epsilon = 0.03$ and $\beta = 4$, the memory capacity is enhanced through repeated learning. We explored its dependence on different parameters, especially ϵ and β. ϵ is the timescale of the learning process relative to that of the neural dynamics. We plotted the capacity curve for various values of ϵ in Fig. 6(e). As ϵ increases, the number of patterns which are successfully recalled decreases and for $\epsilon > 1$, only one pattern is recalled. We also explored dependence of the recall performance on β. Generally, in randomly coupled neural network models, attractors change from fixed points to chaos with the increase in β. We plotted the recall performance for different β in Fig. 6(f). As β increases, the recall performance is increased. For $\beta < 1$, only one or two memories are recalled successfully.

We also examined the dependence of performance on the input strength γ. This value represents the balance between external and internal inputs. We trained networks for several γ and for $N = 100$, $\alpha = 0.3$ and then, we computed the averaged overlaps with all targets over time as shown in Fig. 6(g). We found that there is an optimal strength of the input around $0.2 \leqslant \gamma \leqslant 1.0$. $\gamma = 1.0$ is the value used for the main results in the paper. For too strong inputs, as well as too weak inputs, a network can memorize and recall only a few patterns.

4. Confining neural states around targets during learning

In the main text, we argued that one of the reasons our model has a decreased capacity relative to the pseudoinverse model is its increased stability. Specifically, our model converges to the target state from any random initial state. We checked whether capacity can be increased by relaxing this demand. To this end, we evaluated a modified learning model. In the original model, we set a random pattern as an initial state at the beginning of each learning step. Here, we set the desired target instead as an initial state. This is done throughout the learning process. By this modification, memory capacity is improved up to $\alpha = 0.45$ when initial states are confined in the vicinity of the targets [filled circles in Fig. 6(h)]. If we set randomly chosen patterns as initial states, memory capacity of this confined learning is worse than that in the original model [open circles in Fig. 6(h)]. Thus, there is a trade-off between stability and memory performance. Beyond the capacity, there appears no difference between memory performances in the original and modified learning rules. It is because the target is not stable even locally during the learning process and neural states escape from the target in the case of the target initial states. Thus, there is no difference between neural behaviors in both rules during

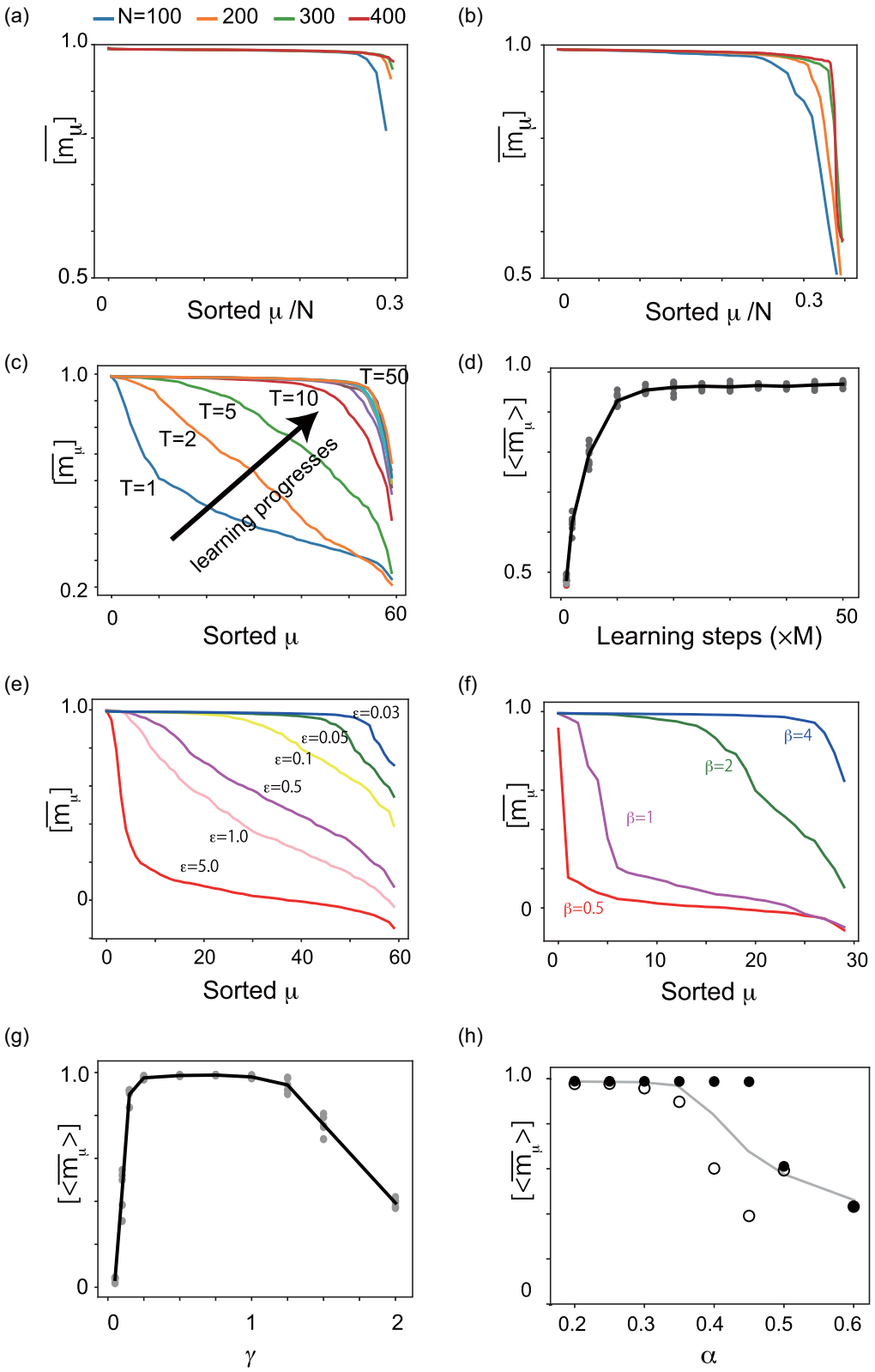

FIG. 6. (a),(b) $[\overline{m_\mu}]$ as a function of μ normalized by N is plotted for different N and for $\alpha = 0.3, 0.35$ in (a) and (b), respectively. (c) $[\overline{m_\mu}]$ as a function of μ is plotted for different T. $[\overline{m_\mu}]$ averaged over time, networks, and trials are plotted. (d) $[\langle\overline{m_\mu}\rangle_\mu]$ by averaging over μ as a function of learning steps T is plotted. (e) $[\overline{m_\mu}]$ is plotted for $T = 30M$ and for different ϵ. Neural dynamics for $N = 200$, $\alpha = 0.3$ in (c)–(e). (f) $[\overline{m_\mu}]$ is plotted for $T = 30M$, $\alpha = 0.3$ and for different β. (g) $[\langle\overline{m_\mu}\rangle_\mu]$ as a function of γ is plotted for $\alpha = 0.3$. (h) $[\langle\overline{m_\mu}\rangle_\mu]$ is plotted for the confining model for $\beta = 4$. Filled circles indicate capacity with beginning from the vicinity of the target. Open circles indicate capacity from randomly chosen patterns. Gray line indicates the capacity of the original learning model for reference. Neural dynamics for $N = 100$ in (f)–(h).

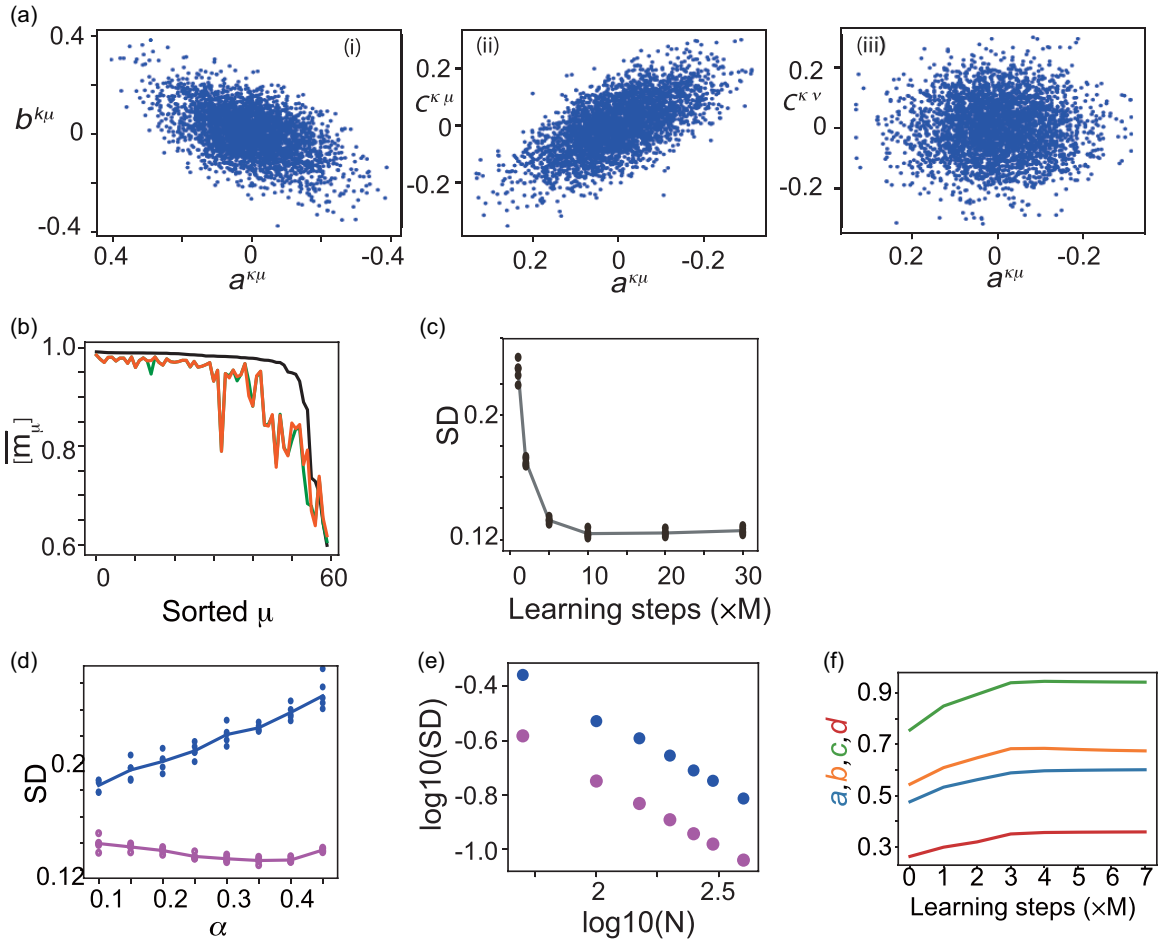

FIG. 7. (a) In (i), a scatter plot between $a_{\kappa,\mu}$ and $b_{\kappa,\mu}$ for $1 \leqslant \kappa, \mu \leqslant M$ and $M = 60$, $T = 30M$. In (ii), a scatter plot between $a_{\kappa,\mu}$ and $c_{\kappa,\mu}$ for the same parameters as (iii). In (iii), a scatter plot between $a_{\kappa,\mu}$ and $c_{\kappa,\nu}$ $(\mu \neq \nu)$ for the same parameters as left. We plot points for randomly selected 60 pairs of (μ, ν). (b) Recall performance for the matrix consisting of only M dominant singular vectors. Green line indicates the performance for this matrix, while the red one indicates the performance for the matrix which is rescaled so that $\sum_{j\neq i} J_{ij}^2 = 1$. Black line represents the performance for the original matrix for reference. (c) Standard deviation (SD) of nondiagonal elements of S during the learning process. Different points indicate different networks. (d),(e) SD of $\boldsymbol{\xi}^{\mu}\boldsymbol{J}\boldsymbol{\xi}^{\nu}/N$ is plotted for $T = M$ in blue and for $T = 30M$ in magenta. These values as a function of α are shown in (d), while those as a function of N are shown in (e). (f) $\langle \Sigma_{\mu}(s^{\kappa,\mu})^2 \rangle_{1\leqslant\kappa\leqslant M}$ for $s = a, b, c, d$ are plotted during the learning process in the modified pseudoinverse model corresponding to Fig. 3(b).

learning, resulting in the same memory performance beyond the capacity.

5. SVD analysis

We explored relationships between $a^{\kappa,\mu}$, $b^{\kappa,\mu}$, $c^{\kappa,\mu}$, and $d^{\kappa,\mu}$. A scatter plot of $(a^{\kappa,\mu}, b^{\kappa,\mu})$ is displayed in Fig. 7(a)(i). $a^{\kappa,\mu}$ is negatively correlated with $b^{\kappa,\mu}$. We also show a scatter plot of $(a^{\kappa,\mu}, c^{\kappa,\mu})$ in Fig. 7(a)(ii). $a^{\kappa,\mu}$ is positively correlated with $c^{\kappa,\mu}$. In Fig. 7(a)(iii), we plot $(a^{\kappa,\mu}, c^{\kappa,\nu})$ $(\mu \neq \nu)$. There is no correlation between them.

We showed that the learned connectivity $\boldsymbol{J}$ is approximately decomposed to the dominant M singular vectors as in Eq. (6). Here, we tried to confirm that a matrix consisting of only these dominant M vectors can recall memories, $\boldsymbol{J}' = \Sigma_{\mu\mu} S_{\mu\nu} (\boldsymbol{\xi}^{\mu} + k^{\mu}\boldsymbol{\eta}^{\mu})(\boldsymbol{\xi}^{\nu} + l^{\nu}\boldsymbol{\eta}^{\nu})^t$. We plotted recall performance of $\boldsymbol{J}'$ in Fig. 7(b). We found that almost all patterns are recalled by the reconstructed matrix $\boldsymbol{J}'$, although performance for some patterns is decreased. We also verified that this is not due to the norm of $\boldsymbol{J}'$.

To understand the improvement of capacity during learning from the viewpoint of the singular vectors, we calculated nondiagonal elements of S in Eq. (6). We plotted the standard deviation (SD) of these non-diagonal elements in Fig. 7(c). We found that the SD rapidly decreases during learning, but still is far away from zero, consistent with the decorrelating effect described in the main text.

6. Dependence of $\boldsymbol{\xi}\boldsymbol{J}\boldsymbol{\xi}^t$ on N and α

Here, we provided another support for the hypothesis that the connectivity obtained by our learning rule decorrelates target patterns by evaluating $\boldsymbol{\xi}^{\mu}\boldsymbol{J}\boldsymbol{\xi}^{\nu}/N$, in addition to the interference-to-signal ratio. We note that in the standard Hopfield network, corresponding to the case that S is a diagonal matrix, the standard deviation of $\boldsymbol{\xi}^{\mu}\boldsymbol{J}\boldsymbol{\xi}^{\nu}/N$ $(\mu \neq \nu)$ follows $O[(\alpha/N)^{1/2}]$, whereas it follows $O(N^{-1/2})$ for the pseudoinverse correlation matrix. If the shaped connectivity perfectly

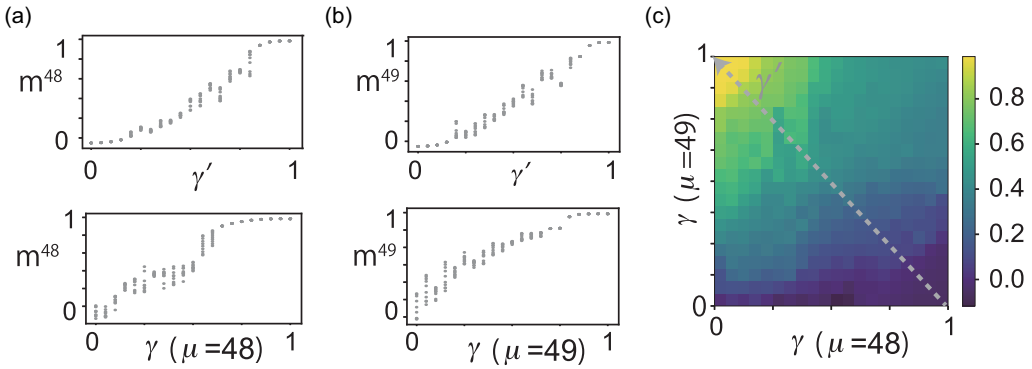

FIG. 8. (a) Bifurcation of the overlap with the target 48. In the upper panel, the bifurcation against $\gamma' \boldsymbol{\eta}^{48} + (1-\gamma')\boldsymbol{\eta}^{49}$ is plotted. In the lower panel, the bifurcation against $\gamma \boldsymbol{\eta}^{48}$ is plotted. (b) The same figures are shown as in panel (a) except the overlap with the target 49. (c) Phase diagram of the overlap with the target 49.

decorrelates targets as well as the pseudoinverse matrix, the standard deviation should follow $O(N^{-1/2})$. We thus measured $\boldsymbol{\xi}^{\mu} \boldsymbol{J} \boldsymbol{\xi}^{\nu}/N$ $(\mu \neq \nu)$ and estimated its dependence on α and N in Figs. 7(d) and 7(e), respectively, for the connection matrix shaped by learning. We found that the standard deviation at the earlier stage of learning ($T = M$) follows $O[(\alpha/N)^{1/2}]$, but at the later stage of learning it turns to follow $O(N^{-1/2})$ (for $T = 30M$). This result supports our hypothesis that the connectivity shaped through our learning rule effectively decorrelates targets to optimally reduce interference.

7. Calculation of contributions of the inputs and targets in singular vectors of the pseudoinverse model

We calculated components $a^{\kappa,\mu}$, $b^{\kappa,\mu}$, $c^{\kappa,\mu}$, and $d^{\kappa,\mu}$, in singular vectors of the pseudoinverse model in the same manner as that in the paper. In Fig. 7(f), component d (coefficient of input in right singular vectors) is around chance level ($0.3 = M/N$) and component c (coefficient of targets in right singular vectors) is beyond 0.9. Values of these components are quite different from those in our model [Fig. 3(b)].

8. Response to input mixtures

Figure 8 shows more details of the response to input mixtures. Panel (c) repeats the same analysis of the main text, but showing the overlap with pattern 49. Panels (a) and (b) show the behavior for horizontal [(a), bottom], vertical [(b), bottom], and diagonal [(a),(b), top] lines in this space. All the results support that recall of the target pattern is represented as a distinctive phase of the corresponding fixed-point attractor and as separated from oscillating neural activity.

9. Spontaneous activity

We analyzed how the nature of spontaneous activity is changed through learning. Spontaneous activity shows chaotic behavior intermittently approaching some targets in Fig. 9(a). For earlier learning, we found a clear correlation between recall performance $\langle \overline{m} \rangle$ and maximum overlap $m_{\mu,\max} = \langle \max_{0<t<1000} m_{\mu}(t) \rangle_{\mu}$ as shown in Fig. 9(b). A few targets which show nearly perfect recall performance are closely approached by the spontaneous activity. For later learning, in contrast, there appears no clear correlation. Almost all targets show perfect recall performance and their maximum overlap is more evenly distributed.

Next, we explored spontaneous activity for different ϵ. As ϵ decreases and recall performance increases [Figs. 6(e) and 9(e)], the spontaneous activity is distributed broader [Fig. 9(d)] and more chaotic [Fig. 9(f)]. In Fig. 9(c) [similar to Fig. 9(b)], a few targets which show nearly perfect recall performance are closely approached for $\epsilon = 1.0$, while there appears no clear correlation for $\epsilon = 0.03$. This relation between the spontaneous activity and recall performance is consistent with that for different learning steps. For quite larger ϵ, some fixed points, instead of chaotic dynamics, are shaped, one of which corresponds to the latest trained network in Fig. 9(g).

10. Equivalence of our learning and pseudoinverse rules

We show that the our learning rule is reduced to the pseudoinverse rule [21] in the presence of input $\boldsymbol{\eta}$ [same as in Eq. (4)]:

$$\Delta J_{ij} = (1/N)\left(\xi_i^{\mu} - u_i^{\mu}\right)\xi_j^{\mu} \tag{A1}$$

under limited conditions. Here, $\boldsymbol{u}^{\mu} = \boldsymbol{J}\boldsymbol{\xi}^{\mu} + \boldsymbol{\eta}^{\mu}$. According to [21], we consider map dynamics of binary neurons,

$$s_i(t+1) = \Theta\left[\Sigma_j J_{ij} s_j(t) + \eta_i^{\mu}\right], \tag{A2}$$

where $\Theta(x) = 1$ for $0 \leqslant x$, -1 for otherwise, and discretized temporarily our learning rule without decay term as

$$\Delta J_{ij} = (1/N)(\xi_i^{\mu} - s_i)s_j. \tag{A3}$$

Now, we consider the following situations: First, $\boldsymbol{\xi}^{\mu}$ is almost a fixed point after some learning, i.e., for a small residual $\boldsymbol{q}$,

$$\Theta(\boldsymbol{u}^{\mu}) = \boldsymbol{\xi}^{\mu} + \boldsymbol{q} \tag{A4}$$

and we define $\boldsymbol{\xi}'\boldsymbol{\mu} \equiv \boldsymbol{\xi}^{\mu} + \boldsymbol{q}$. Next, $\boldsymbol{s}$ approaches $\boldsymbol{\xi}'\boldsymbol{\mu}$:

$$\boldsymbol{s} = \boldsymbol{\xi}'\boldsymbol{\mu} + \boldsymbol{q}' = \boldsymbol{\xi}^{\mu} + \boldsymbol{q} + \boldsymbol{q}' \tag{A5}$$

for small residuals $\boldsymbol{q}'$. Here, $q_i, q_i' \in \{-2, 0, 2\}$. We denote $\boldsymbol{u}^{\mu} - \boldsymbol{\xi}'$ by $\boldsymbol{r}$ ($|r_i| < 1$). The numbers of nonzero elements of $\boldsymbol{q}$ and $\boldsymbol{q}'$ are denoted as n and m, respectively. We assumed the residuals $\boldsymbol{q}$, $\boldsymbol{q}'$, and $\boldsymbol{r}$ are sufficiently small compared to N with keeping that $\boldsymbol{q}$ is larger than $\boldsymbol{q}'$, i.e., $1 \gg n/N \gg m/N$ and $1 \gg \Sigma_i |r_i|/N$.

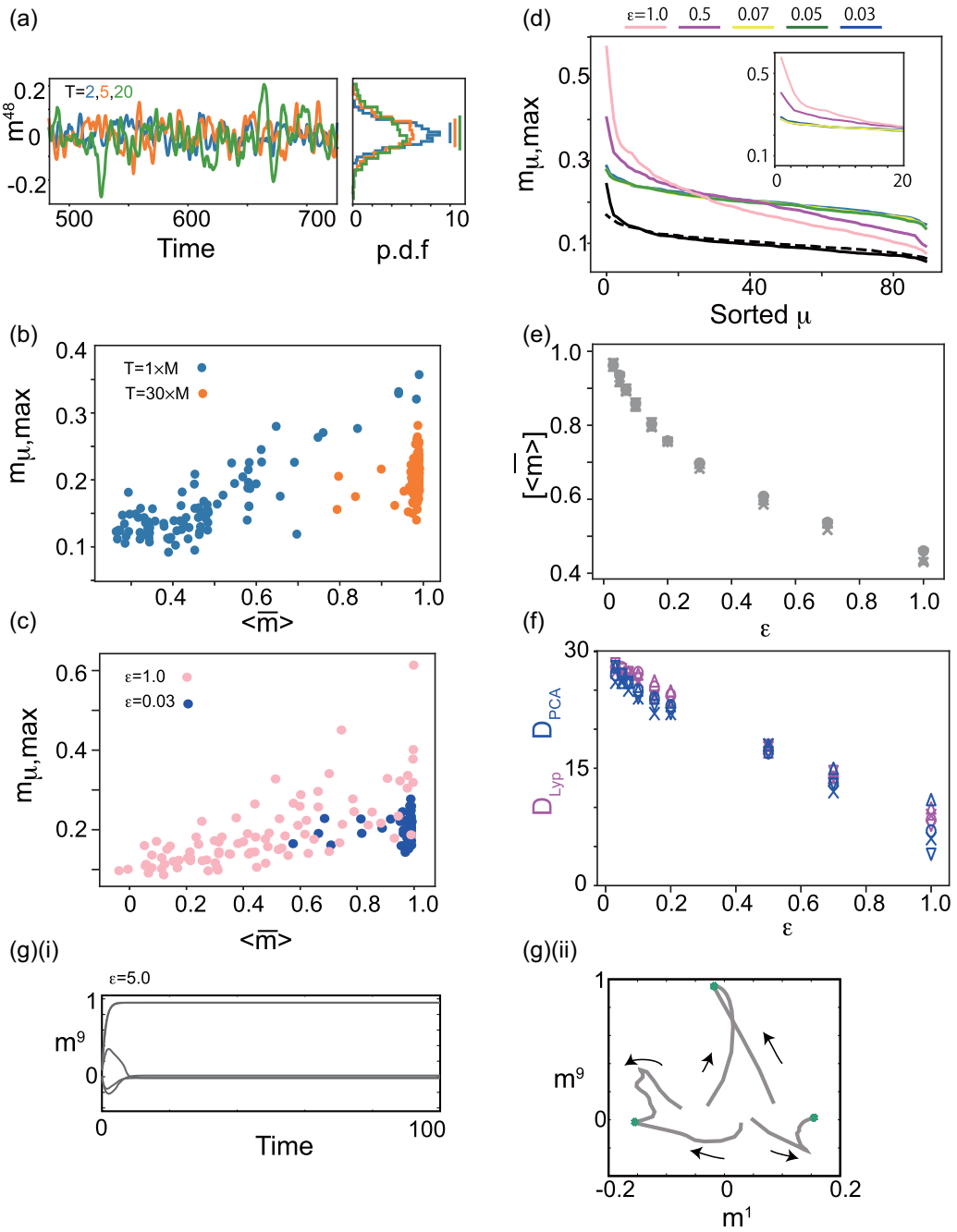

FIG. 9. (a) Overlaps of spontaneous activities without input m_{45} are plotted from $t = 500$ to 700 for $T = 2M$, $5M$ and $T = 20M$ in blue, orange, and green, respectively, in the left panel. Here, $N = 300$ and $M = 90$ as well as in the following panels. In the right panel, probability density functions (p.d.f.) of these overlaps in longer intervals ($0 < t < 1000$) and their standard deviations are also plotted as bars. (b) Scatter plot of maximum overlap of the spontaneous activity with a target against its recall performance. Dots in blue and orange are for $T = M$ and $T = 30M$, respectively. (c) Scatter plot same as in (b). Dots in blue and pink are for $\epsilon = 0.03$ and $\epsilon = 1.0$, respectively. (d) Maximum overlaps of the spontaneous activity with targets are plotted for different ϵ. Black solid and dashed lines indicate overlap with input and random patterns, respectively, for reference. (e),(f) Recall performance in (e) and Lyapunov dimension and PC_{80} in (f) are plotted. (g) Spontaneous activity for $\epsilon = 5.0$. (i) Five time series of the overlap with target 9 from five random initial points, which is the latest trained pattern. (ii) The same neural dynamics with upper panels projected into two-dimensional space.

Based on this setting, we substitute Eqs. (A4) and (A5) into Eq. (A3) and get

$$(\xi_i^\mu - s_i)s_j/N = \left(\xi_i^\mu - u_i^\mu\right)\xi_j^\mu/N + D_{ij}/N, \qquad \text{(A6)}$$

where $D_{ij} = r_i\xi_j - q_i'\xi_j + (q_i + q_i')(q_j + q_j')$ is the difference between our learning and pseudoinverse rules. The average amplitude of D_{ij} is evaluated as $\Sigma_{ij}|D_{ij}|/N^2 < \Sigma_i|r_i|/N + 2m/N + (2m + 2n)^2/N^2$, while those of $(\xi_i^\mu - u_i^\mu)\xi_j^\mu$ are evaluated as n/N. Thus, the averaged difference goes to zero much faster than the average of $(\xi_i^\mu - u_i^\mu)\xi_j^\mu$ as N goes to infinity.

[1] B. Blumenfeld, S. Preminger, D. Sagi, and M. Tsodyks, Neuron **52**, 383 (2006).

[2] A. Bernacchia and D. J. Amit, Proc. Natl. Acad. Sci. USA **104**, 3544 (2007).

[3] S. McKenzie, N. T. M. Robinson, L. Herrera, J. C. Churchill, and H. Eichenbaum, J. Neurosci. **33**, 10243 (2013).

[4] J. E. Dunsmoor, V. P. Murty, L. Davachi, and E. A. Phelps, Nature (London) **520**, 345 (2015).

[5] L. N. Driscoll, N. L. Pettit, M. Minderer, S. N. Chettih, and C. D. Harvey, Cell **170**, 986 (2017).

[6] S.-I. Amari, Biol. Cybernetics **26**, 175 (1977).

[7] J. J. Hopfield, Proc. Natl. Acad. Sci. USA **81**, 3088 (1984).

[8] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Ann. Phys. **173**, 30 (1987).

[9] A. Bernacchia, Front. Synaptic Neurosci. **6**, 1 (2014).

[10] L. Saglietti, F. Gerace, A. Ingrosso, C. Baldassi, and R. Zecchina, Int. Focus **8**, 20180033 (2018).

[11] T. Kurikawa and K. Kaneko, PLoS Comput. Biol. **9**, e1002943 (2013).

[12] G. Orbán, P. Berkes, J. Fiser, and M. Lengyel, Neuron **92**, 530 (2016).

[13] P. Berkes, G. Orbán, M. Lengyel, and J. Fiser, Science **331**, 83 (2011).

[14] L. Buesing, J. Bill, B. Nessler, and W. Maass, PLoS Comput. Biol. **7**, e1002211 (2011).

[15] W. J. Ma, J. M. Beck, P. E. Latham, and A. Pouget, Nat. Neurosci. **9**, 1432 (2006).

[16] A. Litwin-Kumar and B. Doiron, Nat. Neurosci. **15**, 1498 (2012).

[17] G. Hennequin, Y. Ahmadian, D. B. Rubin, M. Lengyel, and K. D. Miller, Neuron **98**, 846 (2018).

[18] In our model, we set a random pattern as an initial state at the beginning of each learning step. If we set the desired target instead, memory capacity improves up to $\alpha = 0.45$ as shown in Fig. 6.

[19] L. Personnaz, I. Guyon, and G. Dreyfus, Phys. Rev. A **34**, 4217 (1986).

[20] I. Kanter and H. Sompolinsky, Phys. Rev. A **35**, 380 (1987).

[21] S. Diederich and M. Opper, Phys. Rev. Lett. **58**, 949 (1987).

[22] T. Kurikawa and K. Kaneko, Europhys. Lett. **98**, 48002 (2012).

[23] Our estimate is based on $O(1)$ and we discarded terms of $O(N^{-1/2})$, since targets and inputs are not exact normal orthogonal basis $[\Sigma_i \xi_i^\mu \eta_i^\nu / N = O(N^{-1/2})]$.

[24] For large α and $T = M$, fluctuations in S sometimes leads to negative values, while $\boldsymbol{O}$ is relatively confined around 1.0. To avoid huge fluctuations in the interference-to-signal ratio, we first evaluated $\langle S_{\lambda\lambda}/|\boldsymbol{O}^\lambda|\rangle_\lambda$ and then used its inverse.

[25] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning internal representations by error propagation, Technical Report No. ICS-8506. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[26] R. J. Williams and D. Zipser, Neural Comput. **1**, 270 (1989).

[27] D. J. Amit and S. Fusi, Neural Comput. **6**, 957 (1994).

[28] N. Brunel, F. Carusi, and S. Fusi, Network (Bristol, England) **9**, 123 (1998).

[29] S. Fusi and L. F. Abbott, Nat. Neurosci. **10**, 485 (2007).

[30] B. Siri, H. Berry, B. Cessac, B. Delord, and M. Quoy, Neural Comput. **20**, 2937 (2008).

[31] M. N. Galtier, O. D. Faugeras, and P. C. Bressloff, Neural Comput. **24**, 2346 (2011).

[32] Y. Kim, B. B. Vladimirskiy, and W. Senn, Frontiers Comput. Neurosci. **2**, 1 (2008).

[33] T. Kenet, D. Bibitchkov, M. Tsodyks, A. Grinvald, and A. Arieli, Nature (London) **425**, 954 (2003).

[34] A. Luczak, P. Bartho, and K. D. Harris, Neuron **62**, 413 (2009).

[35] F. Zenke, E. J. Agnes, and W. Gerstner, Nat. Commun. **6**, 6922 (2015).

[36] C. Hartmann, A. Lazar, B. Nessler, and J. Triesch, PLoS Comput. Biol. **11**, e1004640 (2015).

[37] T. Miconi, J. L. McKinstry, and G. M. Edelman, Nat. Commun. **7**, 13208 (2016).

[38] A. Litwin-Kumar and B. Doiron, Nat. Commun. **5**, 5319 (2014).

[39] T. Toyoizumi and L. F. Abbott, Phys. Rev. E **84**, 051908 (2011).

[40] N. Bertschinger and T. Natschläger, Neural Comput. **16**, 1413 (2004).

[41] R. Legenstein and W. Maass, Neural Networks **20**, 323 (2007).

[42] D. Sussillo and L. F. Abbott, Neuron **63**, 544 (2009).

[43] K. Kaneko and I. Tsuda, Chaos **13**, 926 (2003).

[44] I. Tsuda, Neural Networks **5**, 313 (1992).

[45] C. A. Skarda and W. J. Freeman, Behav. Brain Sci. **10**, 161 (1987).

[46] M. I. Rabinovich, R. Huerta, P. Varona, and V. S. Afraimovich, PLoS Comput. Biol. **4**, e1000072 (2008).

[47] A. Minai and T. Anand, Biol. Cybern. **79**, 87 (1998).

[48] K. Rajan, L. F. Abbott, and H. Sompolinsky, Phys. Rev. E **82**, 011903 (2010).

[49] B. Bathellier, L. Ushakova, and S. Rumpel, Neuron **76**, 435 (2012).

[50] J. Niessing and R. W. Friedrich, Nature (London) **465**, 47 (2010).

[51] T. J. Wills, C. Lever, F. Cacucci, N. Burgess, and J. O'Keefe, Science (NY) **308**, 873 (2005).