



Short-term memory by transient oscillatory dynamics in recurrent neural networks

Kohei Ichikawa  and Kunihiko Kaneko *

Graduate School of Arts and Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan



(Received 29 October 2020; accepted 30 July 2021; published 26 August 2021)

Despite the significance of short-term memory in cognitive function, the process of encoding and sustaining the input information in neural activity dynamics remains elusive. Herein, we unveiled the significance of transient neural dynamics to short-term memory. By training recurrent neural networks to short-term memory tasks and analyzing the dynamics, the characteristics of the short-term memory mechanism were obtained in which the input information was encoded in the amplitude of transient oscillations, rather than the stationary neural activities. This transient trajectory was attracted to a slow manifold, which permitted the discarding of irrelevant information. Additionally, we investigated the process by which the dynamics acquire robustness to noise. In this transient oscillation, the robustness to noise was obtained by a strong contraction of the neural states after perturbation onto the manifold. This mechanism works for several neural network models and tasks, which implies its relevance to neural information processing in general.

DOI: [10.1103/PhysRevResearch.3.033193](https://doi.org/10.1103/PhysRevResearch.3.033193)

I. INTRODUCTION

Short-term memory is essential for our cognitive activities [1]. Once an external signal is applied to an internal neuron, its information is encoded therein and saved for a certain time period. Some form of sustained neural activity is therefore necessary [2]. A conventional theory of working memory is based on the representation of each memory item as a different attractor of the considered system [3–5]. This theory is usually referred to as “memories as attractors”. However, this is not practical when a large number of objects has to be memorized, because it is difficult to prepare so many attractors. Additionally, if there is a need to memorize continuous information (e.g., length, frequency, amplitude of inputs), it becomes more difficult to form a memory using this attractor. Furthermore, extensive experimental reports suggest that while short-term memory is maintained, neural activities are not constant; they instead continue to change over time [6,7]. Indeed, the possibility of achieving short-term memory through sustained transient dynamics has recently been discussed [8–12]. Additionally, the general relevance of transient activities to neural information processing has been discussed [13].

It is unclear, however, what form of transient neural activity can afford short-term memory. In contrast to established studies on memories as attractors, the way external inputs are

encoded into transient dynamics is not well explored. Furthermore, in contrast to stationary neural activities at attractors, the way information is sustained in the transient process with time-varying neural activities remains elusive. Memory must be robust under noise to inputs or internal neural activities. From the view of memories as attractors, the stability of memory is supported as the state is attracted to the attractors (i.e., stationary states) after perturbation [14,15], whereas the robustness in the transient dynamics is not well explored. One must investigate how robustness to noise is achieved to understand the transient dynamics that support short-term memory.

To investigate the neural dynamics of short-term memory, we adopted a recurrent neural network (RNN) trained to solve a task that requires memorizing the input information for a given time span. We provided a task to compare two interspaced signals input with some time interval. The RNN is required to determine which of the two subsequent signals has the larger continuous characteristics (e.g., frequencies of the periodic signals) [16]. Here, to solve the comparison task, it is necessary for the neural dynamics to maintain the information of the first signal until the second signal input arrives. Thus, short-term memory is needed to solve this comparison task. In this RNN, the neural dynamics consist of activities entailing a large number of neurons connected via synapses whose weights are adjusted by the standard back-propagation method [17,18] to solve this task.

After the network is trained to solve the task successfully, we analyze the generated neural dynamics to uncover how the input information is encoded and memorized according to the dynamics of neural activities. We find that the neural activities exhibit a transient oscillation that endures for the time span between the first and second signals. Subsequently, we analyze how this oscillatory neural activity encodes the input information, provides short-term memory, and solves the task. We uncover that the information of the first signal (e.g.,

*Also at Research Center for Complex Systems Biology, Universal Biology Institute, University of Tokyo; kaneko@complex.c.u-tokyo.ac.jp

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

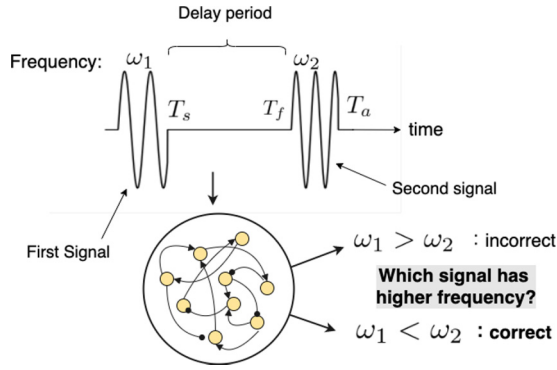


FIG. 1. Schematic diagram of frequency comparison task. The input consists of the first signal, a delay period during which there is no input signal, and the second signal. Both the first and second signals are represented by noisy sine waves. The task is to determine which of the frequencies of the first and second signals is higher. The frequencies of two signals satisfy $1 \leq \omega_1, \omega_2 \leq 5$. In addition, only in the training phase the difference of two signals satisfy $|\omega_1 - \omega_2| \geq 1$.

its frequency) is encoded and memorized by the amplitude of the transient oscillation. Then, we determine if and how this memory by the amplitude of transient oscillation is robust to noise using dynamical systems theory.

The remainder of this paper is organized as follows. In the next section, we introduce the task to compare the frequencies of two signals input with a given time interval as well as the RNN model used to solve it. After demonstrating that the trained RNN can solve the task, we analyze how the memory of the first input is maintained. We demonstrate that neural activities during the time span, while memory is maintained, exhibit a transient oscillation. Here, the input-signal information required to solve the task is encoded by the amplitude of this transient oscillation, whereas the irrelevant information in the input signal that is unrelated to solving the task is discarded. Then, the robustness of the memory to noise is analyzed. We also confirm that the mechanism for short-term memory presented in this study is valid for several different comparison tasks as well as in neural network models that take biological features into the account, such as excitatory-inhibitory balance or sparseness. Finally, the possible relevance of the presented mechanism for short-term memory to biological neural dynamics is discussed.

II. MODEL

A. Short-term memory task

As a task that requires short-term memory, we studied the frequency comparison task illustrated in Fig. 1, which is commonly adopted in the field of neuroscience [16]. In this task, the input signals consist of the first signal, a delay period, and the second signal. The objective of the task is to determine which frequency is higher: that of the first or the second signal.

Specifically, the first and second signals are chosen as noisy sine waves following $u_{1,2}(t) = \sin(\omega_{1,2}t + \phi) + \eta_{1,2}(t)$, where $\phi_{1,2}$ represents the phase of the signals, and $\eta_{1,2}(t)$ is a random Gaussian variable with average zero and standard deviation 0.05. During the delay period, there is no

TABLE I. Conditions of the short-term memory task.

Condition	Training phase	Test phase
Length of the input signal	$13 \leq T_s \leq 17$	$T_s = 15$
Length of the delay period	$25 \leq T_d \leq 35$	$T_d = 30$
Frequencies of signals	$1 \leq \omega_1, \omega_2 \leq 5$	$1 \leq \omega_1, \omega_2 \leq 5$
Difference of ω_1, ω_2	$ \omega_1 - \omega_2 \geq 1$	$ \omega_1 - \omega_2 \geq 0$

input signal. The duration of the first (second) signal, T_s (T_s') and the delay period, T_d , vary with each sample in the training phase. At T_s , the delay period starts, and at $T_f = T_s + T_d$, the delay period ends. At $T_a = T_s + T_d + T_s'$, the second signal completes, and the neural network is asked whether the first or second has the higher frequency. Specifically, T_s is homogeneously distributed as $T_s \in [13, 17]$, and T_d is homogeneously distributed as $T_d \in [25, 35]$ throughout the samples in the training phase (see Table I). During the test phase, to answer the task, T_s and T_d are fixed at 15 and 30, respectively. These conditions are fixed unless otherwise mentioned.

B. Recurrent neural network

In this study, a standard model [19] is adopted for neural activity dynamics as expressed in the equation

$$\dot{x}_i = -x_i + \sum_{j=1}^N J_{ij} \tanh(x_j) + W_i^{\text{in}} u, \quad (1)$$

where x_i represents the neural activity state of neuron i [or $(1 + \tanh(x_i))/2$ can be correlated with the firing rate].¹ Each neuron is recurrently connected to the others, where J_{ij} represents the strength of the connection from neuron j to i . Furthermore, u_i represents the input signal, which is defined by a cognitive task as described above and is projected onto the neurons by the input connection, W_i^{in} . The number of neurons, N , is set at 256. The initial state of neuron $x_i(t = 0; i = 1, \dots, N)$ is set to be a Gaussian random variable with average zero and standard deviation of 0.1. In this study, the Euler method is applied to simulate Eq. (1), and the discretized dynamics are subsequently calculated. A discretized Eq. (1) is adopted here as $x_i(t + 1) = (1 - \alpha)x_i(t) + \alpha\{\sum_{j=1}^N J_{ij} \tanh(x_j(t)) + W_i^{\text{in}} u(t)\}$. The time width, α , for discretization is set to 0.25; nonetheless, the choice of this specific value is not essential to the results.

The output of the RNN is determined by the weighted sum of the neural states, as stated in the equation,

$$z_i(t) = \sum_{j=1}^N W_{ij}^{\text{out}} x_j. \quad (2)$$

\mathbf{z} is chosen to be a two-component vector and $z_1 - z_2$ represents the LOGIT of the probability that the first signal has a

¹Because x_i takes both positive and negative values, it does not represent the neural activity *per se*. Instead, in the context of neuroscience, $[1 + \tanh(x_i)]/2$ is often regarded as the firing rate of the neuron.

higher frequency [20]. The calculation from \mathbf{z} to the probability is performed using the SOFTMAX function, $\text{Softmax}(z_i) = e^{z_i} / (e^{z_1} + e^{z_2})$.² Therefore, $\text{Softmax}(z_i)$ represents the probability that the RNN judge that the i th signal has the higher frequency; If $z_1 > z_2$, the RNN estimates the first signal to have a higher frequency, and vice versa.

The parameters of this RNN are composed of the weights of the vector, \mathbf{W}^{in} , and two matrices, \mathbf{J} and \mathbf{W}^{out} ; these weights are adjusted by training the RNN to solve the task introduced in Sec. II A. The specific learning procedure is described in the following Sec. II C. Although, in the reservoir computing scheme [21,22] training only changes the output matrix, \mathbf{W}^{out} , in this study, we adjust the weights of all three matrices.

C. Training

To train the RNN for the short-term memory task described above, we adopt a stochastic gradient descent scheme, which is commonly used in machine-learning communities [23]. In this scheme, the loss function is first defined to indicate how far the output of the current RNN is from the correct answer. This loss function is given as a function of the weights of the matrices, which provide the parameters of the RNN. The learning process is carried out by calculating the gradient and optimizing the parameters in the direction toward which the loss function becomes smaller. Although there are many optimization algorithms, the basic concept is given by the following equation:

$$W_{t+1} = W_t - \eta \nabla_w \mathcal{L}, \quad (3)$$

where W represents the parameters of the RNN (i.e., vector \mathbf{W}^{in} and two matrices \mathbf{J} and \mathbf{W}^{out}), and η represents the learning rate. $\nabla_w \mathcal{L}$ represents the gradient of the loss function and is calculated via back-propagation through time (BPTT) [18], in which the dynamics of the RNN are first unfolded in time; then, the derivatives of the loss function are calculated using the chain rule. In this study, we adopt the SOFTMAX cross-entropy loss function:

$$\mathcal{L}_{\text{CE}} = - \sum_{k=1}^2 \hat{z}_k \log z_k^{\text{softmax}}, \quad (4)$$

where $z_k^{\text{softmax}} \equiv e^{z_k} / (e^{z_1} + e^{z_2})$, and $\hat{\mathbf{z}}$ represents the target label of this task. If ω_1 is larger than ω_2 , $\hat{\mathbf{z}} = (1, 0)^T$; otherwise, $\hat{\mathbf{z}} = (0, 1)^T$. This target label is defined to satisfy the demand of this task. As explained, z_k^{softmax} gives the probability that the k th signal is recognized to have higher frequency by the

RNN. The SOFTMAX cross-entropy loss, therefore, represents the difference between the probability estimated by the RNN and the target probability. This loss function is widely used owing to its computational efficiency [24].

To make the training process stable, L2 norm regularization [20] is applied for the sum of the norm of the RNN weights. Finally, the loss function is defined as

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{L2}} \left(\sum_i W_{\text{in},i}^2 + \sum_{ij} J_{ij}^2 + \sum_{ij} W_{\text{out},ij}^2 \right), \quad (5)$$

where λ_{L2} is set to 0.0001. During the training phase, the loss function is summed up for 50 input samples; then, the weights of the matrices are adjusted by the gradient descent. This process is continued for 3000 iterations. Here we adopt ADAM [25] as the specific algorithm of optimization, which uses $\nabla_w \mathcal{L}$ calculated by BPTT, as is mentioned above. The learning rate, η , is set to 0.001. The training algorithm is implemented using PYTORCH [26], which is the framework for machine learning.

III. RESULTS

A. Short-term memory by transient oscillatory dynamics

After training, 50 pairs of signals having various ω_1 and ω_2 were input to the trained RNN. In Fig. 2(a), the fraction judged by the RNN to be ω_2 was higher than ω_1 plotted against $\omega_2 - \omega_1$. With the condition of $|\omega_2 - \omega_1| > 1$, the accuracy of the choice was greater than 95%. Hence we can see that the RNN correctly learned to solve the frequency comparison task. To more closely examine how the neural dynamics compared the frequencies, the neural states corresponding to various ω_1 and ω_2 at $t = T_a$ (i.e., at the end of the second signal) were plotted using the three principal components of $\{x_i\}$ [Fig. 2(b)] [27]. The state changed continuously in response to the difference in frequencies. Moreover, according to $\omega_1 > \omega_2$ or $\omega_1 \leq \omega_2$, the states could be separated by a plane. Hence, the frequency of the first signal was memorized over the delay period.

Here, the neural activities continued to change over time during the delay period, as illustrated in Supplemental Material, Fig. 1 [28]. It can now be confirmed that the first signal was memorized by transient dynamics rather than by the attractor. To further confirm this, the second signal was removed, and the long term behavior of neural dynamics after the delay period was observed. As shown in Fig. 3(a), when the delay period was prolonged, the neural states converged to a limit-cycle attractor independent of the signal frequency, ω_1 . After the first signal input, neural activities were attracted to a low-dimensional manifold at which the limit cycle was located.³ Then, the neural activities oscillated with slowly increasing amplitude towards the limit-cycle attractor. This increase, however, was so slow that the short-term memory was maintained for a sufficient amount of time. We analyze the duration of memory in detail below.

²The relationship between LOGIT $z_1 - z_2$ and the probability calculated by the SOFTMAX function is as follows. By definition of the SOFTMAX function, we have $p_1 = e^{z_1} / (e^{z_1} + e^{z_2}) = 1 / (1 + e^{-(z_1 - z_2)})$. Then, by the definition of LOGIT function, we get the LOGIT of p_1 as $\log[p_1 / (1 - p_1)] = \log(e^{z_1 - z_2}) = z_1 - z_2$. From the above explanation, the output of the SOFTMAX function can be treated as a probability because p_1 and p_2 satisfy $0 \leq p_1, p_2 \leq 1$, and $p_1 + p_2 = e^{z_1} / (e^{z_1} + e^{z_2}) + e^{z_2} / (e^{z_1} + e^{z_2}) = 1$. In this study, specifically, we interpret this probability as the probability that the first signal is recognized to have a higher frequency by the RNN.

³The fixed-point analysis [14] showed that there were a few slow points acting as pseudosaddles through which the neural activity passed after the first signal. See Supplemental Material, Fig. 3 [28], for the slow points plotted in the three-dimensional PC space.

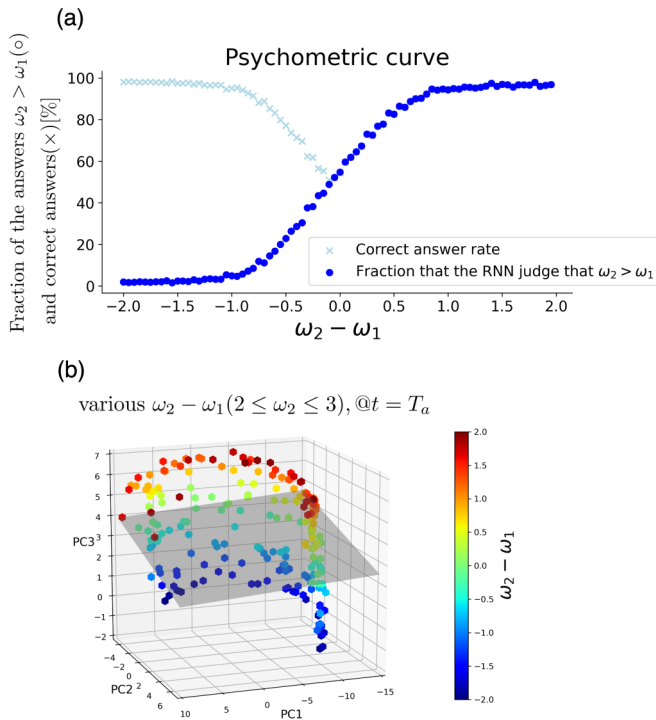


FIG. 2. (a) Fraction judged by the RNN judge to be $\omega_2 \geq \omega_1$ (\circ). The horizontal axis represents the difference between ω_1 and ω_2 . Hence, if the fraction is 100% (0%) for $\omega_2 - \omega_1 > 0$ (< 0), respectively, the correct answer rate (×) is 100%. As shown, if $|\omega_1 - \omega_2| \geq 1$, the correct answer rate is almost 100%, whereas it decreases as $|\omega_1 - \omega_2|$ decreases. (b) Neural states \mathbf{x} at the end of the second signal ($t = T_a$; timing to judge), plotted in the three-dimensional principal component (PC) space. The PC axis is computed from the data of the neural states at T_a for various frequency inputs. Colors represent $\omega_2 - \omega_1$. The plane in the center can separate these neural states into those satisfying $\omega_1 > \omega_2$.

Subsequently, we show how memory information (i.e., frequency of signals) was represented in the neural activity. In Fig. 3(a), we can see the trend in which the amplitude of the transient oscillation before the attraction to the limit cycle has monotonic dependence on ω_1 . At the end of the delay period ($t = T_f$), this trend is remarkable; there is strong negative correlation between the amplitude, $|\mathbf{x}(t = T_f)|$, and the first signal frequency, ω_1 [Fig. 3(d2)]. Hence, ω_1 is encoded by the amplitude of the transient oscillation [Fig. 3(c)]. Notably, this monotonic coding of the input frequency by the amplitude does not hold immediately after the input of the first signal. Indeed, at the beginning of the delay period ($t = T_s$), there was no such monotonic dependence on ω_1 [see Figs. 3(b), 3(d1)]. Therefore the neural dynamics during the delay period ($T_s \leq t \leq T_f$) shaped the manifold from Figs. 3(b) and 3(c).

This coding by amplitude is beneficial for discarding information irrelevant to the task. In the present task, the input signals included phases of the oscillation apart from the frequency. Hence the RNN should discriminate the frequency information from the phase information. Here, signals having different phases were mapped to different points on the same radius trajectory [Fig. 3(e)]. The neural dynamics during the delay period ($T_s \leq t \leq T_f$) dampened information on the signal other than its frequency, ω_1 .

The neural state corresponding to ω_1 , encoded by the amplitude of transient oscillation, was used as the initial state for the response to the second signal. After the second signal was input, the neural states moved separately up and down along the PC3 axis according to the sizes of ω_2 and ω_1 . If $\omega_2 > \omega_1$, it moved in the positive direction along the PC3 axis, and vice versa. In Fig. 3(f), changes in the PC3 component of $\mathbf{x}(t)$ after the second signal with $\omega_2 = 3$ are plotted against the amplitude of $\mathbf{x}(t)$ oscillation at $t = T_f$. As shown, depending on the amplitude of the oscillation at $t = T_f$, PC3 moved upwards when $|\mathbf{x}(t = T_f)|$ was small (corresponding to $\omega_2 < \omega_1$), and it moved downwards when $|\mathbf{x}(t = T_f)|$ was large (corresponding to $\omega_2 > \omega_1$). This property separated the neural states in the direction of PC3, as shown in Fig. 2(b), and it enabled the RNN to correctly solve the frequency comparison task by using the short-term memory for ω_1 , as coded in the transient amplitude.

To verify the generality of short-term memory encoded by the amplitude of transient oscillation, we examined two other settings. First, we changed the length of the delay period as follows. During the training phase, T_d was homogeneously distributed as $T_d \in [75, 105]$, and during the test phase, T_d was fixed at 90. With this longer delay period setting, we confirmed that the present mechanism also worked (see Supplemental Material, Fig. 2 [28], for the neural activity of trained RNN with longer delay period). Second, we trained the RNN for two other comparison tasks, which requested we compare the velocity and noise variance. In the former task, the first and second signals were given by $u_{1,2}(t) = a_{1,2}t + b + \eta_{1,2}(t)$, and the RNN was required to determine which of the velocities, $a_{1,2}$, was larger. In the latter task, the first and second signals were given by $u_{1,2} = \sin(t + \phi) + \eta_{1,2}(t)$, where $\eta_{1,2}(t)$ was a random Gaussian variable having an average of zero and a standard deviation of $\sigma_{\text{ex}}^{1,2}$. The RNN was then required to determine which of the variances, $\sigma_{\text{ex}}^{1,2}$, was larger. For both the tasks, a monotonic dependence between the L2 norm of the neural states (the amplitude of transient oscillation) and the parameter to be compared was revealed at the end of the delay period ($t = T_f$) (see Supplemental Material, Figs. 4 and 5 [28], for the results of the different tasks). The information of the velocity, a_1 , (for the former) and the noise variance, σ_{ex}^1 , (for the latter) was memorized as the amplitude of transient oscillation.

B. Convergence to the limit cycle

As described previously, transient oscillatory dynamics eventually converged to the limit cycle. When this trajectory converged, the memory was forgotten, and the information of the input was lost. Hence, for this memory to work and to accomplish the task, the time to converge to the limit cycle must be sufficiently long. As a measure of convergence of $\mathbf{x}(t)$ to the limit cycle, we first introduced $\ell(t) = |\mathbf{x}(t) - \mathbf{x}(t + T_L)|$, where T_L is the period of the limit cycle. Subsequently, we estimated the convergence time to the limit cycle as the time, t , when $\ell(t)$ was sufficiently small (i.e., the time t at which $\ell(t)$ first satisfies $\ell(t) \leq 0.05$ for the first time). We computed the average convergence time over 100 trajectories, $\mathbf{x}(t)$, for different ω_1 values and noted that the convergence time was several times longer than T_f . We also found that

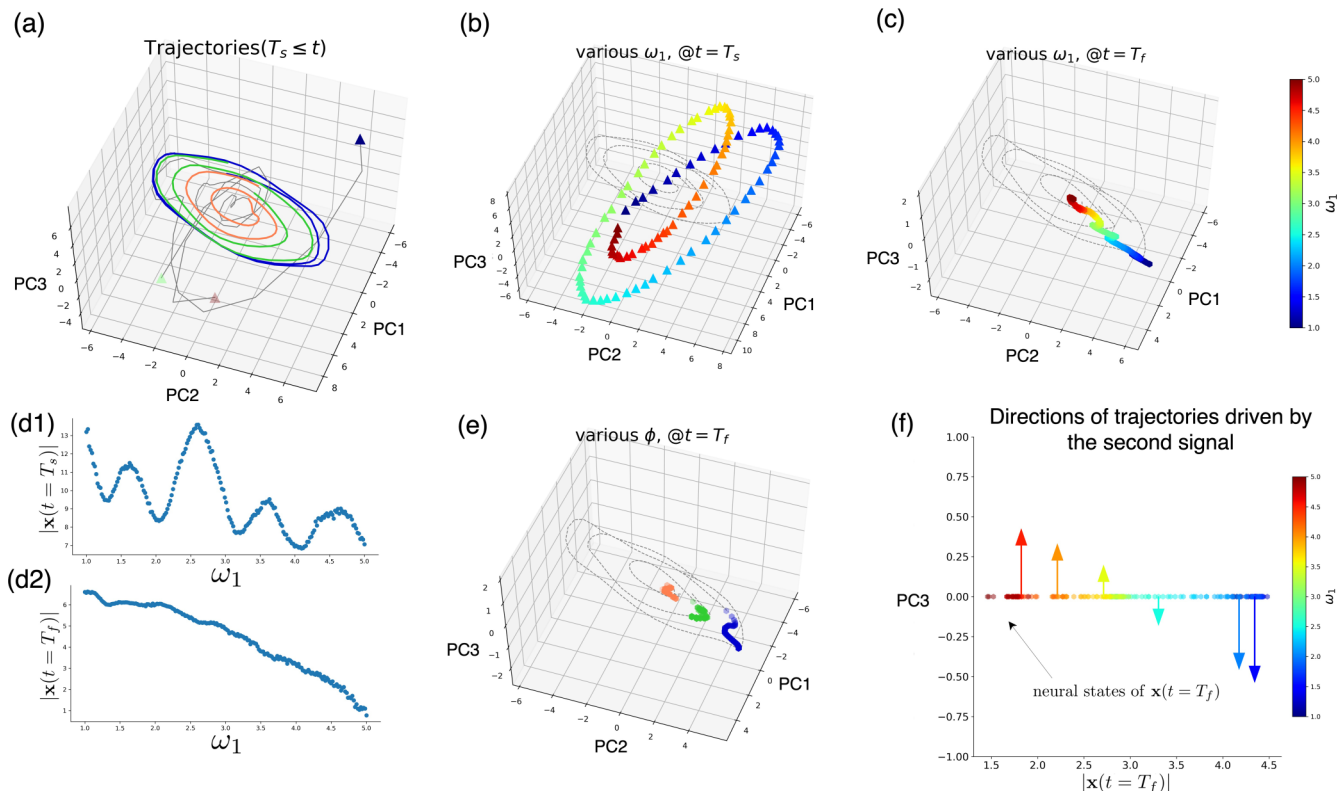


FIG. 3. (a) Trajectories of neural activities during a prolonged delay period ($T_s \leq t$, without second signal) were plotted in a three-dimensional principal component (PC) space. Trajectories from three different ω_1 , giving rise to different states at $t = T_s$, are given by triangle symbol. (b) Neural states \mathbf{x} in the PC space at the beginning of the delay period ($t = T_s$) for 50 different ω_1 's presented with different colors. (c) Neural states \mathbf{x} in the PC space at the end of the delay period ($t = T_f$) for ω_1 corresponding to (b). (d1) Scatter plot of the norm, $|\mathbf{x}(t = T_s)|$, against the frequency of the first signal, ω_1 . (d2) Scatter plot of the norm at the end of the delay period, $|\mathbf{x}(t = T_f)|$, against the frequency of the first signal, ω_1 . Monotonic dependence is discernible. (e) Neural states \mathbf{x} in the PC space at the end of the delay period ($t = T_f$) for $\omega_1 = 1.5, 3, 4.5$ with various $0 \leq \phi \leq \pi$. (f) Directions of neural trajectories driven by the second signal with $\omega_2 = 3$. The x axis shows the amplitude of the trajectory at $t = T_f$. Arrows show the direction and magnitude of the change in the PC3 component of $\mathbf{x}(T_f + 1) - \mathbf{x}(T_f)$ for the first input signal with $\omega_1 = 1.5, 2, 2.5, 3.5, 4, 4.5$.

during training, the convergence time increased alongside an increase in the accuracy of the task (see Supplemental Material, Fig. 6 [28], for the convergence time to the limit cycle). These results suggest that the short-term memory was maintained over a sufficiently long period, prolonged after the delay period, and the length of the convergence time was related to the performance of informational processing.

C. Robustness to noise

As described in Sec. I, the mechanism of the robustness to noise in the short-term memory by transient oscillation remains elusive. Hence, the robustness of the memory was examined by adding the noise term in Eq. (1) using the Langevin equation:

$$\dot{x}_i = -x_i + \sum_{j=1}^N J_{ij} \tanh(x_j) + W_i^{\text{in}} u + \xi_i, \quad (6)$$

where ξ_i is a random Gaussian variable with an average of zero and a standard deviation of σ_{neu} . As with the no-noise setting, we discretized the dynamics and obtained the following equation: $x_i(t + 1) = (1 - \alpha)x_i(t) + \alpha \{ \sum_{j=1}^N J_{ij} \tanh[x_j(t)] + W_i^{\text{in}} u(t) \} + \sqrt{\alpha} \xi_i(t)$. The RNN was

trained under a noise level, $\sigma_{\text{neu}}^{\text{train}}$, with regularization for the average squared norm of neural activity, $\frac{1}{T_a} \sum_t \sum_i x_i(t)^2$. Regularization was applied to the loss function in the form of

$$\mathcal{L}_{\text{noise}} = \mathcal{L} + \lambda_{\text{act}} \frac{1}{T_a} \sum_t \sum_i x_i(t)^2. \quad (7)$$

The reason that we introduced this regularization term is as follows. Notably, if the average squared norm, $\frac{1}{T_a} \sum_t \sum_i x_i(t)^2$, increases, robustness will increase because the derivative of $\tanh(x)$ will generally decrease. By using this regularization term to suppress the norm of neural activity, we can avoid such trivial robustness. We experimentally determined λ_{act} so that the norm of the internal dynamics of RNNs trained with noise ($\sigma_{\text{neu}}^{\text{train}} = 0.04$) would be comparable to those trained without noise ($\sigma_{\text{neu}}^{\text{train}} = 0$), and we adopted $\lambda_{\text{act}} = 30$. Subsequently, the trained network was tested for solving the task under the noise level, $\sigma_{\text{neu}}^{\text{test}}$.

As depicted in Fig. 4(a), the accuracy of the short-term memory task decreased with the applied noise, $\sigma_{\text{neu}}^{\text{test}}$, for the network trained without noise. In contrast, for the model trained with a sufficient noise level, $\sigma_{\text{neu}}^{\text{train}}$, the drop in the accuracy by the increase in noise level was suppressed, even

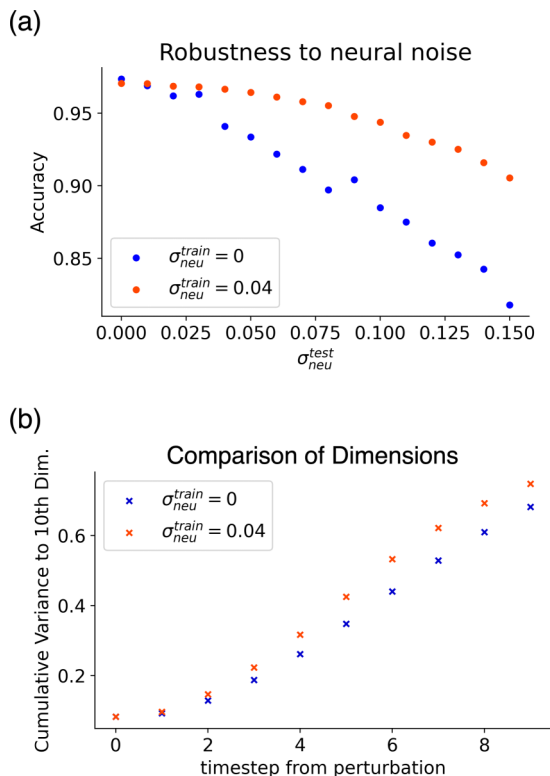


FIG. 4. (a) Accuracy of the frequency comparison task plotted against the noise level, σ_{neu}^{test} . Different colors show the RNN trained at different levels of noise ($\sigma_{neu}^{train} = 0, 0.04$). Bold lines show the mean score, and shaded areas show the standard deviations. (b) The cumulative contribution ratio of the perturbed trajectories up to the tenth principal component (i.e., the cumulative percentage of the eigenvalues corresponding to first-to-tenth eigenvectors $(\sum_{i=1}^{10} \lambda_i) / (\sum_{i=1}^{256} \lambda_i)$). These are plotted against the time following the application of perturbation for $\sigma_{neu}^{train} = 0$ (blue), and $\sigma_{neu}^{train} = 0.04$ (red). If it is larger, more points are restricted within the lower-dimensional manifold. Because the perturbation is completely random, the perturbation dimension is high immediately after it is applied. However, over time, it is attracted to a lower-dimensional manifold, especially for $\sigma_{neu}^{train} = 0.04$.

up to a noise level higher than σ_{neu}^{train} . The system gained a higher robustness to noise, which was beyond the level added during learning. Notably, the RNN trained with neural noise realized the same mechanism of the short-term memory with transient oscillation.

To examine the achievement of the stability of short-term memory, the following perturbation analysis was performed. We set $\sigma_{neu}^{test} = 0$, and subsequently, we introduced an instantaneous perturbation during the delay period [i.e., $T_s \leq t_{per} \leq T_f$, as $\mathbf{x}_{per}(t) = \mathbf{x}(t) + \delta \mathbf{x} \delta_{t,t_{per}}$], where $\delta \mathbf{x}$ is a random variable that follows a Gaussian distribution of mean zero and of variance σ_{per}^2 , and $\delta_{t,t_{per}}$ represents a Kronecker delta.

Intuitively, one might expect that the distance between the perturbed trajectory and the original trajectory would be decreased for the networks trained under noise. Here, we first computed the L2 norm distance between the perturbed and original trajectories: $L_{per} = |\mathbf{x}_{per}(t) - \mathbf{x}(t)|$. In contrast to the naive expectation, however, the distance does not depend substantially on the trained noise level, σ_{neu}^{train} . Hence, the distance

between the trajectories cannot explain the memory robustness for a system trained with noise σ_{neu}^{train} . To understand the robustness, the direction of the perturbation of the trajectories must be considered. Indeed, the shift of the trajectory along the original trajectory was not harmful to the memory discussed, because the amplitude of the transient oscillation was not affected by the shift. Hence, the dimension of the manifold spanned by the perturbed trajectories is more important to the robustness of memory than the distances between the perturbed trajectories and the original trajectories.

Accordingly, we estimated the dimension of the manifold spanned by a large number of perturbed trajectories. We adopted a principal component analysis (PCA) for an ensemble of these perturbed trajectories. If perturbation causes a shift mainly along the trajectory direction, perturbed trajectories should be restricted within a low-dimensional manifold along the original trajectory. Hence, the dimension of perturbed trajectories estimated by PCA will work as a measure of robustness in the directions of perturbed trajectories. To estimate the dimension, we computed the percentage of the variance corresponding the first-to-tenth eigenvectors for each time after the perturbation as $(\sum_{i=1}^{10} \lambda_i) / (\sum_{i=1}^{256} \lambda_i)$, where λ_i represents the eigenvalue of covariance matrix with an ordering to satisfy $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{256}$.

This percentage was larger when the perturbed trajectory was restricted to a lower-dimensional manifold. As shown in Fig. 4, for the RNN trained with noise σ_{neu} , the trajectory falls into a lower-dimensional manifold over time following the perturbation. The percentage for such robust RNN is larger than the RNN trained without neural noise. This result implies that the dynamics after the perturbation was more restricted to a lower-dimensional manifold for the model trained with noise. This low-dimensional compression is key to the robustness of the neural noise. Notably, another method of estimating the attractor dimension was proposed in Refs. [29,30], where the number of eigenvectors of the principal components required to achieve 90% of the total variance was calculated. We have confirmed that similar results have been obtained by applying this method (see Supplemental Material, Fig. 7 [28], for the estimated dimension of perturbed trajectories by the method in Refs. [29,30]).

To characterize the mechanism of this convergence of the transient trajectory to low-dimensional space, we focused on the Jacobi matrix, $\mathbf{G}(t)$, of this system. The Jacobi matrix is defined as the following equation:

$$G_{ij}(t) = J_{ij} \tanh'[x_j(t)]. \quad (8)$$

The perturbed dynamics convergence can be estimated by the eigenvalues of matrix $e^{\frac{1}{T} \int dt \mathbf{G}(t)}$ along the trajectory, as was adopted in the calculation of finite-time Lyapunov exponents [31]. Therefore, we investigated the difference of eigenvalues between the RNNs trained with $\sigma_{neu}^{train} = 0$ and those trained with $\sigma_{neu}^{train} = 0.04$. Because the value of the Jacobi matrix depends on the specific trajectory, $\mathbf{x}(t)$, we calculated the histograms of the eigenvalues for 50 samples of input signals (Fig. 5). We also adopted the discretization to calculate $e^{\frac{1}{T} \int dt \mathbf{G}(t)}$ as $\exp[\frac{1}{30} \sum_{t=15}^{45} \mathbf{G}(t)]$.

For the latter RNN, there were more eigenvalues with negative real parts. As shown in Fig. 5, the number of eigenvalues

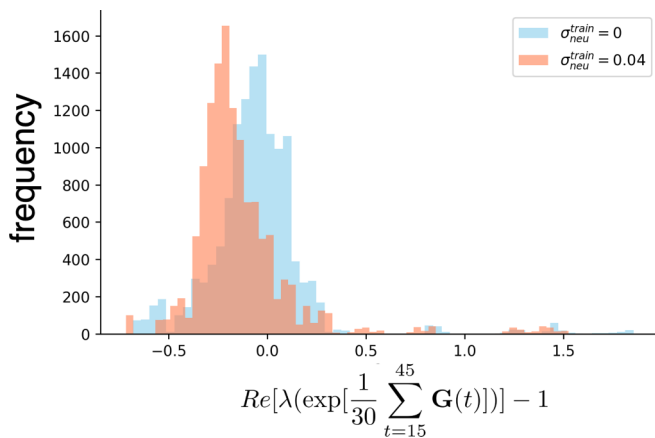


FIG. 5. Histogram of the frequency distribution of the real part of the eigenvalues of $\exp[\frac{1}{30} \sum_{t=15}^{45} \mathbf{G}(t)]$. We calculated $\mathbf{G}(t)$ for 50 samples of the input signals. Notably, there were 256 eigenvalues for one $\exp[\frac{1}{30} \sum_{t=15}^{45} \mathbf{G}(t)]$. Blue is the result for the RNN trained without noise, and red is the result for the RNN trained with noise ($\sigma_{\text{neu}}^{\text{train}} = 0.04$). In the latter, the eigenvalues were biased to the negative side overall.

with positive real parts was reduced. This result suggests that the restriction to a low-dimensional manifold in the robust RNN was caused by the compression of many modes of perturbed dynamics. This robustness caused by the contraction of transient trajectories was clearly distinguishable from that caused by the stability of fixed-point attractors.

D. Different types of neural networks

To investigate the range of validity of transient oscillatory short-term memory, we considered two different types of neural networks. First, we considered neural networks in which synapses obey Dale's law (i.e., the neural network consists of excitatory and inhibitory neurons, and synapses extended from a given neuron are either all excitatory, or inhibitory) [32,33]. Second, we considered neural networks in which synapses were sparse. We call these networks excitatory-inhibitory and sparse networks, respectively.

We trained the RNN to solve the frequency comparison task while satisfying each condition. Under the excitatory-inhibitory networks condition, as in Ref. [34], hidden neurons were divided into excitatory and inhibitory neurons in a 4:1 ratio. \mathbf{J} was constrained so it can be decomposed as $\mathbf{J} = \mathbf{J}^+ \mathbf{D}$, where \mathbf{J}^+ is a matrix with all components greater than or equal to zero, and \mathbf{D} is a diagonal matrix that satisfies $D_{ii} = 1$ if $i \leq 180$ and $D_{ii} = -1$ if $i > 180$. With this constraint, it can be said that \mathbf{J} obeys Dale's law. During the training phase, \mathbf{D} was fixed, and \mathbf{W}^{in} , \mathbf{J}^+ , and \mathbf{W}^{out} were adjusted. In sparse networks, we set the percentage of synapses with nonzero weight to approximately 20%. We adopted the deep-rewiring technique [35] to train under constraints where only 20% of synapses have nonzero values, and all others are zero. As in Supplemental Material Figs. 8 and 9, after learning, the RNNs solved the frequency comparison task by maintaining short-term memory by transient oscillatory dynamics (see Supplemental Material, Figs. 8 and 9 [28], for the results

under the excitatory-inhibitory networks condition and the sparse networks condition).

IV. DISCUSSION

In this study, we uncovered a generic scheme for short-term memory sustained by transient oscillatory dynamics by training RNNs to achieve a task requiring short-term memory to compare two interspaced input signals [36,37]. We demonstrated that short-term memory was encoded in the amplitude of transient oscillations of neural activities. The neural state given by high-dimensional dynamical systems fell into a low-dimensional manifold [14] and exhibited transient oscillation, which slowly approached a limit-cycle attractor. With the passage of time, continuous information from the first signal input, such as the frequency and velocity of the input signal, was encoded in the amplitude of the transient oscillation and maintained as short-term memory. Other irrelevant information in the inputs (e.g., phase and noise) was discarded during the transient dynamics. Hence, short-term memory is encoded and maintained by transient dynamics and is robust to external noise.

The proposed mechanism, wherein the memory is encoded in transient oscillation, contrasts with the view of memory as fixed-point attractors (i.e., memories as attractors). When the RNN is trained to store discrete information (e.g., the possible input signal frequency candidates are limited to 1, 2, 3, 4, or 5 Hz), the memories are encoded by multiple fixed-point attractors (i.e., persistent activities) [38]. In the task adopted in this study, by contrast, storing continuous information was needed. Hence, encoding into the amplitude of transient oscillation occurred. Notably, as an alternative mechanism to encoding continuous information, a line attractor was proposed [39,40]. Here, attractors coexisted continuously on a line in the state space of neural activities, along which the fixed point was marginally stable [i.e., one of the eigenvalues of the Jacobi matrix (with the eigenvector along the line) should be zero]. However, in autonomous dynamical systems, such marginally stable attractors are not generic, and for their existence, special constraints are required. Indeed, in this study, such line attractors were not shaped by learning.⁴

As another mechanism of short-term memory, synaptic plasticity has been proposed [41–43]. In this case, the synaptic connections are modeled as time varying not only during training, but also during inferencing. The memory information is encoded in the time-varying strength of the synapses. In this study, because we focused on neural-based short-term memory, synaptic connections (i.e., the vector \mathbf{W}^{in} and two matrices \mathbf{J} and \mathbf{W}^{out}) change only during training and were fixed thereafter. In future research, we plan to investigate how short-term memory based on transient oscillation can be realized when considering synaptic plasticity after training.

It remains to be seen if the present scheme for short-term memory can be adopted biologically. The results suggest that

⁴For the possible applicability of chaotic attractors to short-term memory, see J. S. Nicolis and I. Tsuda, Chaotic dynamics of information processing: The magic number seven plus-minus two revisited, *Bull. Math. Bio.* **47**, 343 (1985).

the short-term memory encoded by the amplitude of transient oscillation works over a wide range of systems. This may be plausible, because the present mechanism is generally represented in terms of dynamical systems, and it is not difficult to generate oscillatory dynamics from an ensemble of neurons. Here, we offer two remarks. We expect that the present scheme is also valid therein, as it is robust to noise in neural dynamics. Next, oscillatory dynamics from certain modes of neural activities evoked by inputs were observed from neural data references [44,45]. Notably, such dynamics are not necessarily observed just by computing the average neural activity, which is often a rather stationary reference independent of inputs. Indeed, this is true in the proposed model. The prominent oscillation depending on the input, as shown in Fig. 2, is observed only by taking appropriate principal components. If we compute just the average activity over all neurons, it is almost stationary, and the oscillation is difficult to discern, which is consistent with experimental observations.

In this study, attraction to the low-dimensional manifold of perturbation allowed for the transient trajectory to be robust to noise. Recently, dynamic robustness has been regarded as an important property of information processing in the brain [46]. The present results provide general insight into the mechanism of this type of robustness. Furthermore, in cell and developmental systems, such dynamic robustness has also been discussed as homeorhesis [47–49], and the results may be applicable to such phenomena.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to Tomoki Kurikawa, Ichiro Tsuda, Hiromichi Suetani, and Tetsuhiro S. Hatakeyama for their comments and discussions. This research was partially supported by a Grant-in-Aid for Scientific Research (A) (20H00123) from the Japanese Society for the Promotion of Science (JSPS).

-
- [1] J. Jonides, R. L. Lewis *et al.*, The mind and brain of short-term memory, *Ann. Rev. Psychol.* **59**, 193 (2008).
- [2] C. L. Colby, J. R. Duhamel, and M. E. Goldberg, Visual, pre-saccadic, and cognitive activation of single neurons in monkey lateral intraparietal area, *J. Neurophysiol.* **76**, 2841 (1996).
- [3] S. Funahashi, C. Bruce, and P. Goldman-Rakic, Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex, *J. Neurophysiol.* **61**, 331 (1989).
- [4] J. Fuster and G. Alexander, Neuron activity related to short-term memory, *Science* **173**, 652 (1971).
- [5] X. J. Wang, Synaptic reverberation underlying mnemonic persistent activity, *Trends Neurosci.* **24**, 455 (2001).
- [6] B. Haider and D. A. McCormick, Rapid neocortical dynamics: Cellular and network mechanisms, *Neuron* **62**, 171 (2009).
- [7] E. H. Baeg, Y. B. Kim, K. Huh, I. Mook-Jung, H. T. Kim, and M. W. Jung, Dynamics of population code for working memory in the prefrontal cortex, *Neuron* **40**, 177 (2003).
- [8] G. Bondanelli and S. Ostojic, Coding with transient trajectories in recurrent neural networks, *PLoS Comput. Biol.* **16**, e1007655 (2020).
- [9] S. Druckmann and D. Chklovskii, Neuronal circuits underlying persistent representations despite time varying activity, *Current Biol.* **22**, 2095 (2012).
- [10] M. S. Goldman, Memory without feedback in a neural network, *Neuron* **61**, 621 (2009).
- [11] E. Orhan and X. Pitkow, Improved memory in recurrent neural networks with sequential non-normal dynamics, in *Proceedings of the International Conference on Learning Representations (ICLR, 2020)*.
- [12] E. Orhan and W. J. Ma, A diverse range of factors affect the nature of neural representations underlying short-term memory, *Nature Neurosci.* **22**, 275 (2019).
- [13] M. Rabinovich, R. Huerta, and G. Laurent, Transient dynamics for neural processing, *Science* **321**, 48 (2008).
- [14] D. Sussillo and O. Barak, Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural networks, *Neural Comput.* **25**, 626 (2013).
- [15] N. Maheswaranathan, A. Williams *et al.*, Universality and individuality in neural dynamics across large populations of recurrent networks, in *Advances in Neural Information Processing Systems 32* (Neural Information Processing Systems Foundation, San Diego, 2019), pp. 15629–15641.
- [16] R. Romo, C. Brody, A. Hernández, and L. L., Neuronal correlates of parametric working memory in the prefrontal cortex, *Nature (London)* **399**, 470 (1999).
- [17] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning representations by back-propagating errors, *Nature (London)* **323**, 533 (1986).
- [18] P. J. Werbos, Backpropagation through time: What it does and how to do it, *Proc. IEEE* **78**, 1550 (1990).
- [19] A. Rivkind and O. Barak, Local Dynamics in Trained Recurrent Neural Networks, *Phys. Rev. Lett.* **118**, 258101 (2017).
- [20] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, Berlin, 2006).
- [21] W. Maass, T. Natschläger, and H. Markram, Real-time computing without stable states: A new framework for neural computation based on perturbations, *Neural Comput.* **14**, 2531 (2002).
- [22] D. Sussillo and L. Abbott, Generating coherent patterns of activity from chaotic neural networks, *Neuron* **63**, 544 (2009).
- [23] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, *Nature (London)* **521**, 436 (2015).
- [24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, 2016).
- [25] D. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, in *International Conference on Learning Representations (ICLR, 2014)*.
- [26] A. Paszke, S. Gross *et al.*, PyTorch: An Imperative Style, High-Performance Deep Learning Library, in *Advances in Neural Information Processing Systems 32* (Neural Information Processing Systems Foundation, San Diego, 2019), pp. 8024–8035.
- [27] J. Lever, M. Krzywinski, and N. Altman, Principal component analysis, *Nature Meth.* **14**, 641 (2017).

- [28] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevResearch.3.033193> for additional figures.
- [29] L. Sirovich, Chaotic dynamics of coherent structures, *Physica D* **37**, 126 (1989).
- [30] S. Ciliberto and B. Nicolaenko, Estimating the number of degrees of freedom in spatially extended systems, *Europhys. Lett.* **14**, 303 (1991).
- [31] A. Crisanti, G. Paladin, and A. Vulpiani, Generalized Lyapunov exponents in high-dimensional chaotic dynamics and products of large random matrices, *J. Stat. Phys.* **53**, 583 (1988).
- [32] H. Dale, Pharmacology and nerve-endings (Walter Ernest Dixon Memorial Lecture): (Section of therapeutics and pharmacology), *Proc. R. Soc. Med.* **28**(3), 319 (1935).
- [33] N. Y. Masse, G. R. Yang, H. F. Song, X.-J. Wang, and D. J. Freedman, Circuit mechanisms for the maintenance and manipulation of information in working memory, *Nat. Neurosci.* **22**, 1159 (2019).
- [34] A. Orhan and W. Ma, Efficient probabilistic inference in generic neural networks trained with non-probabilistic feedback, *Nature Commun.* **8**, 138 (2017).
- [35] G. Bellec, D. Kappel, W. Maass, and R. A. Legenstein, Deep rewiring: Training very sparse deep networks, *International Conference on Learning Representations (ICLR)*, 2018).
- [36] O. Barak, Recurrent neural networks as versatile tools of neuroscience research, *Curr Opin Neurobiol.* **46**, 1 (2017).
- [37] B. Richards, T. Lillicrap, P. Beaudoin *et al.*, A deep learning framework for neuroscience, *Nature Neurosci.* **22**, 1761 (2019).
- [38] K. M. Christian, R. Ranulfo, and D. B. Carlos, Flexible control of mutual inhibition: A neural model of two-interval discrimination, *Science* **307**, 1121 (2005).
- [39] H. S. Seung, How the brain keeps the eyes still, *Proc. Natl. Acad. Sci.* **93**, 13339 (1996).
- [40] V. Mante *et al.*, Context-dependent computation by recurrent dynamics in prefrontal cortex, *Nature (London)* **503**, 78 (2013).
- [41] G. Mongillo, O. Barak, and M. Tsodyks, Synaptic theory of working memory, *Science* **319**, 1543 (2008).
- [42] Y. Mi, M. Katkov, and M. Tsodyks, Synaptic correlates of working memory capacity, *Neuron* **93**, 323 (2017).
- [43] H. Taher, A. Torcini, and S. Olmi, Exact neural mass model for synaptic-based working memory, *PLoS Comput. Biol.* **16**, e1008533 (2020).
- [44] L. Muller, F. Chavane, J. Reynolds *et al.*, Cortical travelling waves: mechanisms and computational principles, *Nature Rev. Neurosci.* **19**, 255 (2018).
- [45] H. Zhang, A. J. Watrous, A. Patel, and J. Jacobs, Theta and alpha oscillations are traveling waves in the human neocortex, *Neuron* **98**, 1269 (2018).
- [46] C. Warasinee, W. Xiao-Jing *et al.*, Computing by robust transience: How the fronto-parietal network performs sequential, category-based decisions, *Neuron* **93**, 1504 (2017).
- [47] C. H. Waddington, *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology* (Allen and Unwin, London, 1957).
- [48] Y. Matsushita and K. Kaneko, Homeorhesis in Waddington's landscape by epigenetic feedback regulation, *Phys. Rev. Research* **2**, 023083 (2020).
- [49] J. T. Young, T. S. Hatakeyama, and K. Kaneko, Dynamics robustness of cascading systems, *PLoS Comput. Biol.* **13**, e1005434 (2017).