

Neuronal correlations shape the scaling behavior of memory capacity and nonlinear computational capability of reservoir recurrent neural networks

Shotaro Takasu¹* and Toshio Aoyagi¹

Graduate School of Informatics, *Kyoto University*, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan



(Received 29 April 2025; accepted 22 September 2025; published 21 October 2025)

Reservoir computing is a powerful framework for real-time information processing, characterized by its high computational ability and quick learning, with applications ranging from machine learning to biological systems. In this paper, we investigate how the computational ability of reservoir recurrent neural networks (RNNs) scales with an increasing number of readout neurons. First, we demonstrate that the memory capacity of a reservoir RNN scales sublinearly with the number of readout neurons. To elucidate this observation, we develop a theoretical framework for analytically deriving memory capacity that incorporates the effect of neuronal correlations, which have been ignored in prior theoretical work for analytical simplicity. Our theory successfully relates the sublinear scaling of memory capacity to the strength of neuronal correlations. Furthermore, we show this principle holds across diverse types of RNNs, even those beyond the direct applicability of our theory. Next, we numerically investigate the scaling behavior of nonlinear computational ability, which, alongside memory capacity, is crucial for overall computational performance. Our numerical simulations reveal that as memory capacity growth becomes sublinear, increasing the number of readout neurons successively enables nonlinear processing at progressively higher polynomial orders. Our theoretical framework suggests that neuronal correlations govern not only memory capacity but also the sequential growth of nonlinear computational capabilities. Our findings establish a foundation for designing scalable and cost-effective reservoir computing, providing insights into the interplay among neuronal correlations, linear memory, and nonlinear processing.

DOI: [10.1103/cwvm-s53p](https://doi.org/10.1103/cwvm-s53p)

I. INTRODUCTION

Over the past decade, deep learning has advanced significantly [1], with generative AI in particular beginning to transform many aspects of daily life [2]. Although deep neural networks possess the capacity to represent highly complex functions [3,4], they typically require substantial computational resources, large volumes of training data, and long training times [5]. These high computational demands pose a major obstacle to applying deep learning in real-time computing on edge-computing devices [6].

Reservoir computing (RC) is a machine learning framework for efficiently training large-scale recurrent neural networks (RNNs), referred to as “reservoirs” in the context of RC [7,8]. In contrast to conventional training methods such as backpropagation through time (BPTT), RC optimizes only readout weights and leaves the remaining weights fixed, which enables quick and low-cost learning. Despite its training simplicity, RC has high computational performance and a broad range of applications for processing time-series data [9]. RC is not restricted to RNNs, and a wide variety of dynamical systems can be utilized as reservoirs under appropriate conditions. In particular, RC using a real physical system such as

soft matter and optical systems, so called “physical reservoir computing,” has been intensively studied in recent years [10].

In a typical RC setting, the readout connections linking reservoir units to output units are sparse [10]. This means that the size of the reservoir (N) is substantially larger than the number of readout units (L) that connect to the output units. The performance of RC empirically improves as L increases because this increases the expressive power of the reservoir output [11]. In contrast, implementing a large number of readout connections can be resource intensive, especially in physical RC implementations. Consequently, understanding the relationship between L and the computational ability of a reservoir is essential for constructing cost-effective RC. Despite its practical importance, however, a guiding principle for determining an appropriate L has yet to be fully explored.

In this study, we aim to address this gap by elucidating how memory capacity (MC) [12] and nonlinear computational ability [13]—two key determinants of a reservoir’s computational power—scale with L , with particular focus on the role of neuronal correlations. We demonstrate, for the first time, that the memory capacity of an RNN increases monotonically as increasing L , with its growth rate progressively declining (Fig. 1). This *sublinear* scaling of memory capacity cannot be explained by the previous theoretical works assuming a scaling regime where L does not scale with N , i.e., $L \sim O(1)$ [14,15], because the memory capacity in this regime exhibits linear scaling, as will be shown later. To bridge the gap between our observations and existing theory, we develop a theory for analytically deriving memory capacity for $L \sim O(\sqrt{N})$. Using our theoretical framework, we demonstrate

*Contact author: takasu.shotaro.63x@st.kyoto-u.ac.jp

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

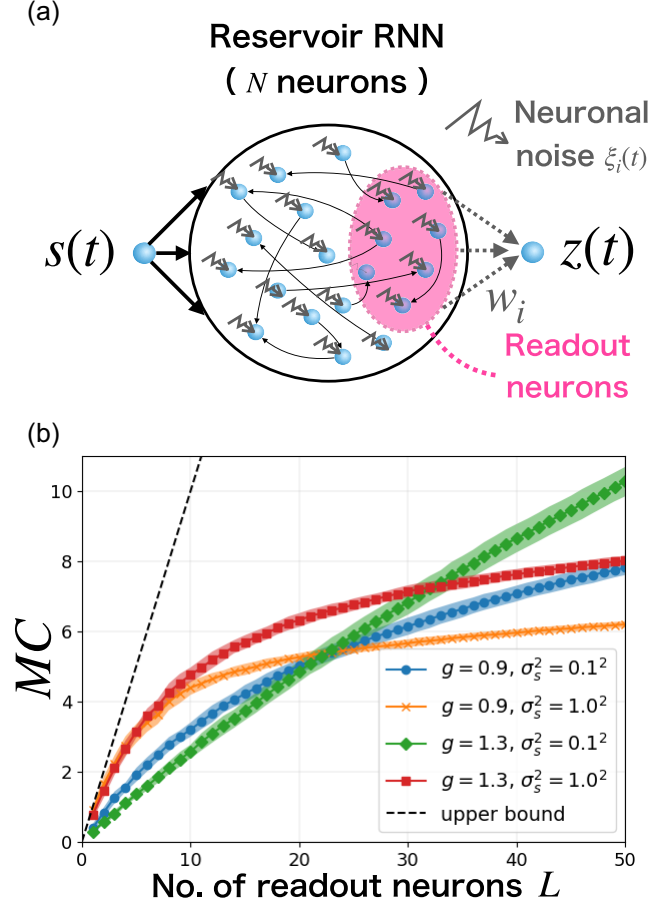


FIG. 1. (a) Overview of RC with a random RNN receiving input signals and inherent neuronal noise. The shaded region represents the readout neurons. (b) Numerical simulations for MC of the reservoir RNNs with network size $N = 10\,000$, noise intensity $\sigma_n^2 = 0.1^2$, and activation function $\phi(x) = \tanh(x)$. Shaded area represents mean $\pm$ std of direct numerical simulations for ten different network and input signal realizations. The dashed line indicates the theoretical upper bound of memory capacity, i.e., $MC = L$. Each MC curve shows a sublinear scaling with increasing L . In the simulations, the sum of M_d is calculated up to $d = 10^3$. Simulation time length is $T = 10^4$, and a washout period of $T_{\text{washout}} = 10^3$ steps is used to discard initial transients.

that the correlations between reservoir neurons, despite being very weak with a magnitude scaling as $O(1/\sqrt{N})$, play an important role in the declining growth rate of memory capacity. Furthermore, numerical simulations show that the ability to perform more complex nonlinear tasks emerges supralinearly and sequentially as L increases. These findings indicate that the number of readout connections should be tailored based on the specific memory and nonlinearity requirements of the task at hand.

II. MODEL SETUP

We investigate computational capacity of RC using a large random RNN as a reservoir [Fig. 1(a)], which is known as Echo State Network [7], one of the canonical models for RC [16,17]. The state of i th neuron at discrete time t is described

by the variable $x_i(t)$. The activation function ϕ is assumed to be an odd saturating sigmoid function satisfying $\phi'(0) = 1$, $\phi'(x) > 0$, $\phi''(x) \leq 0$ ($x \geq 0$), and $\phi(\pm\infty) = \pm 1$. The time evolution of a reservoir RNN is determined by the difference equation,

$$x_i(t) = \sum_{j=1}^N J_{ij} \phi(x_j(t-1)) + u_i s(t) + \xi_i(t), \quad (1)$$

where $N \gg 1$ indicates the total number of neurons. The recurrent weights, J_{ij} , and the input weights, u_i , are sampled independent and identically distributed (i.i.d) from Gaussian distributions, i.e., $J_{ij} \sim \mathcal{N}(0, g^2/N)$ and $u_i \sim \mathcal{N}(0, 1)$. The parameter g controls the strength of the recurrent weights. The input signal at time t , represented by $s(t)$, is Gaussian white noise with zero mean and variance σ_s^2 . To take into account inherent neuronal noise, each neuron is subjected to independent Gaussian white noise, $\xi_i(t)$, with zero mean and variance σ_n^2 . Notably, the input signal—shared across all neurons—induces correlations among reservoir neurons, whereas the independent neuronal noise acts to diminish them.

The macroscopic dynamical behavior of the large random RNN described by Eq. (1) has been extensively studied by means of dynamical mean-field theory (DMFT) [18]. In the absence of the inputs and noise, i.e., $\sigma_s^2 = \sigma_n^2 = 0$, the RNN exhibits phase transition from zero fixed point to chaotic dynamics at $g = 1$ in the large network size limit $N \rightarrow \infty$ [18,19]. In the presence of inputs ($\sigma_s^2 > 0$ or $\sigma_n^2 > 0$), chaotic variability is typically suppressed [14,15,19–22], while it has recently been shown that in RNNs incorporating gating mechanisms similar to those in gated recurrent units, external inputs can instead induce chaos [23].

The output of the reservoir RNN is defined as $z(t) = \sum_{i=1}^L w_i x_i(t)$, where L is the number of readout neurons. We assume that the readout neurons are sparse, i.e., $L \ll N$. According to the RC framework, the output weights, $\{w_i\}_{i=1}^L$, are optimized to minimize the time-averaged squared error between the output signals, $\{z(t)\}_{t=1}^T$, and the desired signals, $\{f(t)\}_{t=1}^T$. Since this is a least-squares problem, the solution for $\mathbf{w} \in \mathbb{R}^L$ is explicitly given by

$$\mathbf{w} = (X X^\top)^{-1} X \mathbf{f}, \quad (2)$$

where the matrix $X \in \mathbb{R}^{L \times T}$ contains the history of readout neuron activities throughout simulation time length T and the vector $\mathbf{f} \in \mathbb{R}^T$ denotes the target time series.

III. MEMORY CAPACITY

Memory capacity [12,13] is a commonly used benchmark for RC, quantifying how accurately a reservoir can reproduce its past Gaussian white noise inputs. Specifically, it is defined as $MC \equiv \sum_{d=0}^{\infty} M_d$, where

$$M_d \equiv 1 - \frac{\min_{\mathbf{w}} \langle (z(t) - s(t-d))^2 \rangle}{\langle s(t)^2 \rangle} \quad (3)$$

and the angular bracket denotes time averaging. Here, the optimized readout weights $\mathbf{w}$ are obtained by substituting $f(t) = s(t-d)$ into Eq. (2). A high value of M_d indicates that the reservoir can accurately output its input signal from d steps prior. It is shown that $0 \leq M_d \leq 1$ holds true [12,13].

Memory capacity is defined as the infinite sum of M_d , but it has been proven to be finite and bounded by the number of readout neurons, i.e., $0 \leq \text{MC} \leq L$ [12,13]. In particular, for RC with a linear activation function operating in noise-free conditions ($\sigma_n^2 = 0$), memory capacity is equivalent to L , independently of g and σ_s^2 [12].

Figure 1(b) illustrates the relationship between L and the memory capacity for the reservoir RNN described by Eq. (1), obtained through numerical simulations. As shown, memory capacity grows monotonically, but its growth rate gradually diminishes as L increases. Consequently, memory capacity can be characterized as a sublinear function of L .

IV. THEORETICAL ANALYSIS OF MEMORY CAPACITY

To elucidate the mechanism behind the decaying growth rate of memory capacity, we provide a theory for deriving an analytical solution of memory capacity. Our approach incorporates some concepts and methods from statistical physics. Hereafter, we make an additional assumption regarding the scale of the model parameters:

$$L = \alpha\sqrt{N}, \quad \sigma_s^2 = \frac{\tilde{\sigma}_s^2}{\sqrt{N}}, \quad \sigma_n^2 \sim O(1), \quad (4)$$

where both α and $\tilde{\sigma}_s^2$ are of $O(1)$. The parameter α represents the relative magnitude of L compared to $\sqrt{N}$. In contrast to previous studies [14,15,24], where $L \sim O(1)$, our model employs a significantly larger number of readout neurons, although the output connections remain sparse, as $L = \alpha\sqrt{N} \ll N$. It should be noted that while our theoretical analysis in this study is conducted under this assumption, our theoretical framework can be extended to a more general case where $L \sim O(N^c)$ and $\sigma_s^2 \sim O(1/N^c)$ for $0 \leq c \leq 1/2$ (see Appendix B 2).

Substituting the optimal solution for $\mathbf{w}$ from Eq. (2) into the definition of M_d [Eq. (3)] yields

$$M_d = \frac{\mathbf{a}_d^\top C^{-1} \mathbf{a}_d}{\langle s(t)^2 \rangle}, \quad (5)$$

where the elements of $\mathbf{a}_d \in \mathbb{R}^L$ and $C \in \mathbb{R}^{L \times L}$ are, respectively, $(\mathbf{a}_d)_i \equiv \langle s(t-d)x_i(t) \rangle$ and $C_{ij} \equiv \langle x_i(t)x_j(t) \rangle$ (i and j indicate the indices of the readout neurons) [13]. The inverse of the covariance matrix, C^{-1} , prevents us from making progress on analytical calculation. Previous theoretical studies circumvented this issue by assuming the off-diagonal entries of C are zero, that is, ignoring neuronal correlations [14,15,24]. However, as discussed later, this approximation results in linear scaling of memory capacity, $\text{MC} \propto L$, conflicting with our numerical simulations as shown in Fig. 1(b).

We notice that from the scale assumption [Eq. (4)], the diagonal entries of C are of $O(1)$, whereas the nondiagonal ones are of $O(1/\sqrt{N})$ (see Appendix A 2). Hence, the nondiagonal elements are much smaller than the diagonal ones, which motivates us to perform Neumann series expansion,

$$C^{-1} = (C_{\text{diag}} + C_{\text{nondiag}})^{-1} = \sum_{n=0}^{\infty} C_{\text{diag}}^{-1} (-C_{\text{nondiag}} C_{\text{diag}}^{-1})^n, \quad (6)$$

where $(C_{\text{diag}})_{ij} \equiv \delta_{ij} \langle x_i(t)^2 \rangle$ and $(C_{\text{nondiag}})_{ij} \equiv (1 - \delta_{ij}) \langle x_i(t)x_j(t) \rangle$. This approach allows us to circumvent the issue of the inverse matrix and enables an analytical calculation of M_d .

In the large network size limit, we can assume self-averaging for M_d , i.e., $\lim_{N \rightarrow \infty} M_d = \lim_{N \rightarrow \infty} [M_d]$, where the square bracket denotes the average over network realizations, known as *quenched average* [25]. Consequently, substituting Eq. (6) into Eq. (5) and taking quenched average of M_d , we analytically obtain memory capacity as

$$\begin{aligned} \lim_{N \rightarrow \infty} \text{MC}(L = \alpha\sqrt{N}) &= \lim_{N \rightarrow \infty} \sum_{d=0}^{\infty} [M_d(L = \alpha\sqrt{N})] \\ &= \sum_{d=0}^{\infty} \sum_{n=0}^{\infty} (-1)^n \left\{ \frac{\alpha \tilde{\sigma}_s^2}{[\langle x_i^2 \rangle]} \left(g \langle \phi'(x) \rangle_{x \sim \mathcal{N}(0, [\langle x_i^2 \rangle])} \right)^{2d} \right\}^{n+1}, \end{aligned} \quad (7)$$

where the value of $[\langle x_i^2 \rangle]$ can be obtained by solving a self-consistent equation (dynamical mean-field equation [18]),

$$[\langle x_i^2 \rangle] = \sigma_n^2 + g^2 \langle \phi(x)^2 \rangle_{x \sim \mathcal{N}(0, [\langle x_i^2 \rangle])}. \quad (8)$$

The memory capacity of a linear RNN with neuronal noise can be derived in a similar fashion, and this derivation also reveals a sublinear scaling (see Appendix B 4).

The derivation of these results requires calculating quenched averages of terms involving neuronal correlations, such as

$$[\langle s(t-d)x_i(t) \rangle \langle x_i(t)x_j(t) \rangle \langle s(t-d)x_j(t) \rangle] \quad (i \neq j). \quad (9)$$

However, these quantities cannot be evaluated using DMFT because it reduces the original N -body system to an effective single-body description, thereby neglecting interneuronal correlations. To overcome this limitation, we adopt the dynamical cavity method, a technique originally introduced by Clark *et al.* [26] from statistical physics to the analysis of neural networks. A detailed description of this theory is provided in Appendixes A 1 and A 2, followed by the full derivation of Eq. (7) in Appendix B 1. In the following, we use the shorthand notation $\langle f(x) \rangle_* \equiv \langle f(x) \rangle_{x \sim \mathcal{N}(0, [\langle x_i^2 \rangle])}$.

When the activation function is an error function, $\phi(x) = \int_0^x e^{-\frac{\pi}{4}t^2} dt$, we can analytically integrate both $\langle \phi'(x) \rangle_*$ and $\langle \phi(x)^2 \rangle_*$, yielding

$$\text{MC} = \sum_{d=0}^{\infty} \sum_{n=0}^{\infty} (-1)^n \left\{ \frac{\alpha \tilde{\sigma}_s^2}{[\langle x_i^2 \rangle]} \left(\frac{g^2}{1 + \frac{\pi}{2} [\langle x_i^2 \rangle]} \right)^d \right\}^{n+1}, \quad (10)$$

where $[\langle x_i^2 \rangle]$ is determined by solving

$$[\langle x_i^2 \rangle] = \sigma_n^2 + g^2 \left(-1 + \frac{4}{\pi} \arctan \sqrt{1 + \pi [\langle x_i^2 \rangle]} \right). \quad (11)$$

We confirm that the analytical values of MC are consistent with the numerical simulations, exhibiting a sublinear function of L , for L below a threshold, as illustrated in Fig. 2. When L exceeds the threshold, the analytical values rapidly deviate from the simulated ones and eventually diverge (dashed lines in Fig. 2). This divergence occurs because

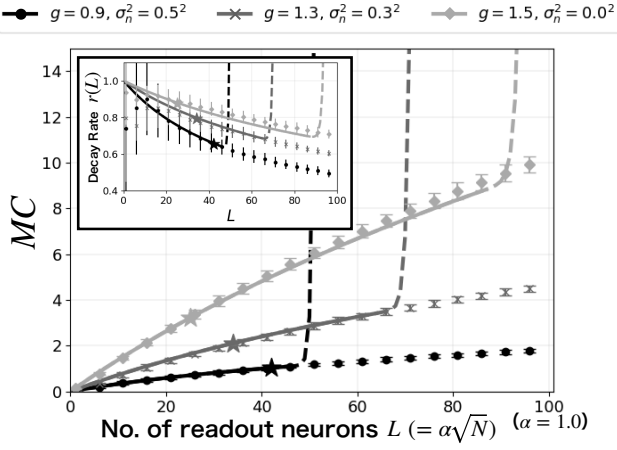


FIG. 2. Memory capacity as a function of L . An inset shows corresponding decay rate $r(L)$. Solid curves represent the analytical solutions given by Eqs. (10) and (11), while error bars indicate mean $\pm$ std for ten networks and inputs realizations obtained through numerical simulations. Star marks represent the upper bounds of the sufficient conditions for the convergence of the series expansion. Dashed lines indicate divergence of theoretical values. The theoretical values and simulation results show excellent agreement in the region where the analytical solution converges. The activation function is an error function, $\phi(x) = \int_0^x e^{-\frac{t^2}{2}} dt$. The input intensity is $\sigma_s^2 = 1.0^2/\sqrt{N}$ ($\bar{\sigma}_s^2 = 1.0^2$). The network size is $N = 10\,000$ and simulation time length is $T = 10^4$. The transient washout time is $T_{\text{washout}} = 10^3$.

the Neumann series expansion [Eq. (6)] fails to converge when the spectral norm $\|C_{\text{nondiag}}C_{\text{diag}}^{-1}\|$ exceeds one for large L . Evaluating this spectral norm to determine the convergence threshold for L is a challenging task. Instead, we note that the spectral norm is bounded by the Frobenius norm $\|\cdot\|_F$ and that $\|C_{\text{nondiag}}C_{\text{diag}}^{-1}\|_F$ can be obtained analytically. We therefore derive a sufficient condition for the series expansion to converge (see Appendix B 1). The corresponding upper bounds of L satisfying this condition are indicated by star marks in Fig. 2.

V. NEURONAL CORRELATIONS SHAPE THE SUBLINEAR SCALING OF MEMORY CAPACITY

Based on our analytical solution for memory capacity, we can now investigate the underlying reason for its sublinear scaling with L . To quantify the extent to which the growth rate of memory capacity decays, we define a *decay rate*,

$$r(L) \equiv \frac{\text{MC}(L)}{L \times \text{MC}(1)}. \quad (12)$$

By this definition, if $\text{MC}(L)$ scales linearly with L , then $r(L) = 1$ holds for all L . In contrast, $r(L)$ decays from unity when $\text{MC}(L)$ is a sublinear function of L . As depicted in the inset of Fig. 2, $r(L)$ consistently decays for various parameter combinations, confirming the sublinear scaling of memory capacity with respect to L .

From our analytical solution for memory capacity [Eq. (7)], we obtain the analytical form for the decay rate (see

Appendix B 3 for the derivation) as

$$\lim_{N \rightarrow \infty} r(L = \alpha\sqrt{N}) = 1 - \sum_{n=1}^{\infty} (-1)^{n-1} \left(\frac{\bar{\sigma}_s^2}{\langle x_i^2 \rangle} \alpha \right)^n \frac{1 - (g\langle \phi'(x) \rangle_*)^2}{1 - (g\langle \phi'(x) \rangle_*)^{2n+2}}. \quad (13)$$

Notably, the n th term corresponds exactly to the n th term in the Neumann series expansion, [Eq. (6)]. Therefore, under the assumption of vanishing neuronal correlations ($C_{\text{nondiag}} = 0$), the summation terms in Eq. (13) disappear, resulting in $r(L) = 1$, and thus, linear scaling. This implies that higher neuronal correlations contribute to the faster decay of $r(L)$ via the summation terms in Eq. (13). In fact, we confirm that $r(L)$ decays more rapidly for hyperparameters that lead to higher neuronal correlation, such as larger input signals ($\bar{\sigma}_s^2$), smaller recurrent weights (g), and lower noise intensity (σ_n^2), as shown in Fig. 3.

Importantly, while the slope of the memory capacity curve at the origin, which is proportional to $\text{MC}(1)$, reaches its maximum when the reservoir RNN operates near the edge of chaos [14,15], the extent of the decay of $r(L)$ varies monotonically with the hyperparameters g , σ_s^2 , and σ_n^2 , irrespective of whether the RNN is near the edge of chaos. This finding suggests a perspective: instead of proximity to the edge of chaos, neuronal correlation is a mechanism that governs the computational ability of reservoir computing.

We have thus far derived the memory capacity under the specific set of scaling assumptions [Eq. (4)]. However, practical applications often deviate from these conditions, raising the question of whether neuronal correlations remain a key determinant of sublinear memory capacity growth when these assumptions are relaxed. To explore this, Fig. 4(a) examines a broader range of model parameters, including larger input intensities, various noise levels, and different values of g . Here, we introduced a *half-life* of memory capacity growth, L_{half} , defined as the value of L at which $r(L)$ equals 0.5 [Fig. 4(a), inset]. A lower L_{half} indicates more rapid decay of $r(L)$. The overall strength of neuronal correlations is quantified by the root mean square (RMS) of pairwise Pearson correlation coefficients, $\sqrt{\langle \rho_{ij}^2 \rangle_{\text{pairwise}}}$, where

$$\rho_{ij} \equiv \frac{\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle}{\sqrt{(\langle x_i^2 \rangle - \langle x_i \rangle^2)(\langle x_j^2 \rangle - \langle x_j \rangle^2)}}.$$

Note that due to the symmetry of the weight distribution, the time-averaged value $\langle x_i \rangle$ almost vanishes for all neurons. We use the RMS value because the symmetry of the weight distribution creates a pairwise distribution of ρ_{ij} values that is symmetric around zero. Consequently, the simple average $\langle \rho_{ij} \rangle_{\text{pairwise}}$ is always zero and thus uninformative, whereas the RMS value properly quantifies the typical magnitude of the neuronal correlations.

As shown in Fig. 4(a), strong neuronal correlations are indeed closely associated with a more rapid decay of memory capacity growth, with L_{half} being inversely proportional to the level of neuronal correlations. This finding highlights the persistent influence of neuronal correlations even when our original scaling assumptions are relaxed.

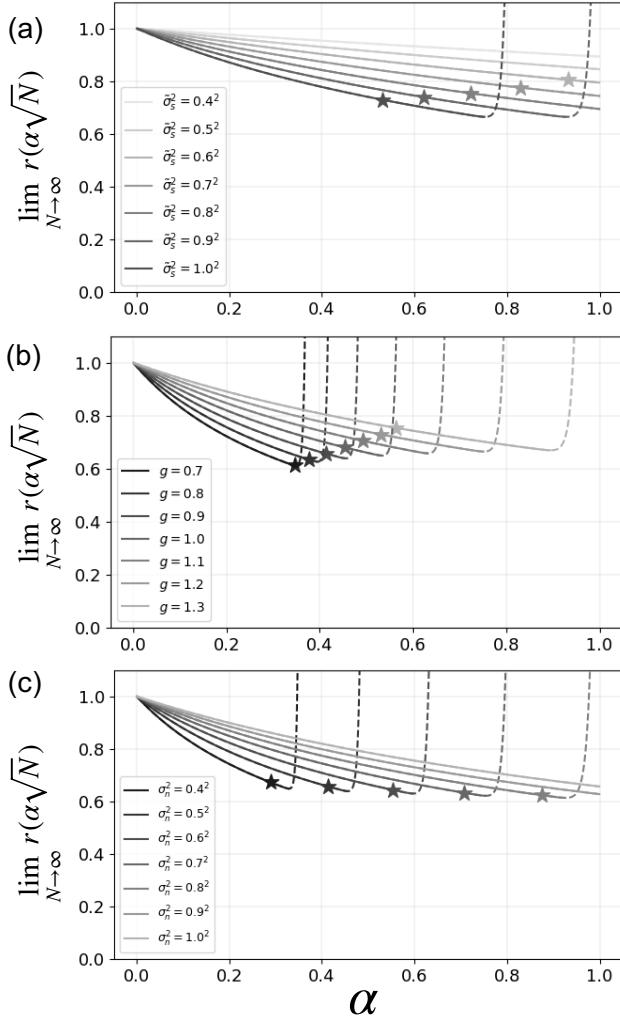


FIG. 3. Analytical solutions for the decay rate of memory capacity as a function of α . The activation function is an error function, $\phi(x) = \int_0^x e^{-\frac{\pi}{4}t^2} dt$. Each figure varies a single parameter: (a) input intensity $\tilde{\sigma}_s^2$, (b) recurrent weight scale g , and (c) noise intensity σ_n^2 , while the remaining parameters are fixed ($g = 1.2$, $\tilde{\sigma}_s^2 = 1.0^2$, $\sigma_n^2 = 0.5^2$). A gradient from darker to lighter gray lines indicates a decreasing level of neuronal correlations. The decay rate for systems with strong neuronal correlations decreases more rapidly.

VI. ROBUSTNESS OF OUR FINDINGS ACROSS OTHER TYPES OF RNNs

Thus far, we have examined the scaling behavior of reservoir RNNs with a sigmoid activation function and recurrent weights drawn i.i.d from a Gaussian distribution. We found that the level of neuronal correlations, which is controlled by hyperparameters such as g , σ_s^2 , and σ_n^2 in our model, determines the extent to which the growth rate of memory capacity decays. While our model is widely used in both theoretical [14,15,27] and practical [16] settings, a natural question arises: can these findings be generalized to other types of RNNs? Here, we hypothesize that our findings on the relationship between neuronal correlations and memory capacity hold more generally. To test this hypothesis, we numerically study three types of RNNs: those with a Rectified

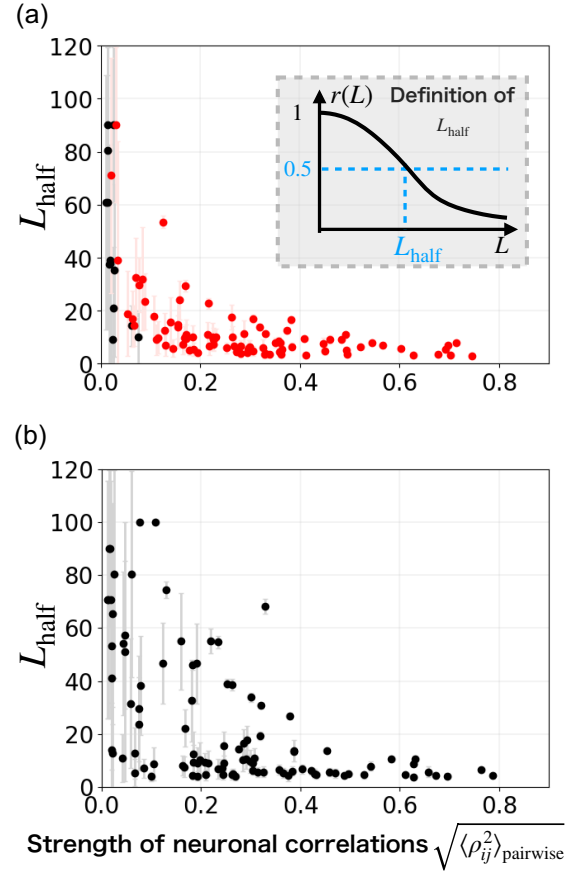


FIG. 4. Relationship between neuronal correlation and the half-life of memory capacity growth. The x axis represents the level of neuronal correlations quantified by the RMS of pairwise Pearson correlation coefficients. The y axis denotes L_{half} , the definition of which is graphically shown in the inset of panel (a). (a) Results for RNNs with $\phi = \tanh$. Each data point corresponds to a set of hyperparameters sampled from $g \sim U(0.2, 3)$, $\sigma_s \sim U(0.1, 3)$, and $\sigma_n \sim U(0, 3)$. Red points indicate parameter regimes beyond our theoretical framework, as determined by L_{half} surpassing the deviation threshold of the theoretical values of MC (Fig. 2). (b) Results for RNNs with $\phi(x) = \max(0, x)$ (ReLU). Each data point corresponds to a set of hyperparameters sampled from $g \sim U(0.2, \sqrt{2})$, $\sigma_s \sim U(0.1, 3)$, and $\sigma_n \sim U(0, 3)$. For both panels (a) and (b), a total of 100 samples are plotted. Each data point represents the mean value over ten network and input realizations. Error bars indicate the standard deviation for both L_{half} and $\sqrt{\langle \rho_{ij}^2 \rangle_{\text{pairwise}}}$, but most of the error bars for the latter are too small to be visible behind the data points. Overall, for both activation functions, L_{half} tends to be smaller for stronger neuronal correlations. All simulations use a network size of $N = 10\,000$ and a simulation time of $T = 10^4$. The transient washout time is $T_{\text{washout}} = 10^3$.

Linear Unit (ReLU) activation function, a heavy-tailed distribution of recurrent weights, and reciprocal motifs.

A. RNN with a ReLU activation function

We first study the neuronal correlations and memory capacity of an RNN described by Eq. (1) with $\phi(x) = \max(0, x)$ (ReLU). All parameter settings are the same as those described in Sec. II, with the exception that g must be lower

than $g_c = \sqrt{2}$ because the system diverges for $g > g_c$. To demonstrate this, we first derive the dynamical mean-field equation for this system in the limit of $N \rightarrow \infty$. The derivation is the same as in Appendix A 1 and yields the following equation:

$$[x_i(t+1)^2] = \frac{g^2}{2} [x_i(t)^2] + s(t+1)^2 + \sigma_n^2. \quad (14)$$

This self-consistent equation admits a stable solution of $[x_i(t)^2]$ only when $g < g_c = \sqrt{2}$. For $g > g_c$, the system diverges regardless of the values of $s(t)$ and σ_n .

Our theoretical framework for the analytical derivation of neuronal correlations and memory capacity cannot be applied to an RNN with the ReLU activation function. This limitation arises because our theory utilizes Price's theorem [Eq. (A27)], which cannot be applied to the ReLU function as it is non-smooth at the origin. Therefore, we numerically calculate neuronal correlations and memory capacity, and plot them to examine their relationship.

Following the same procedure as in the previous section, we numerically calculated the strength of neuronal correlations and L_{half} . The results, shown in Fig. 4(b), reveal a negative correlation between the level of neuronal correlations and L_{half} . Although this correlation is less pronounced than for the sigmoid network [Fig. 4(a)], the finding nonetheless supports our hypothesis that neuronal correlations govern the scaling behavior of memory capacity.

B. RNN with a heavy-tailed distribution of recurrent weights

We now turn to the case of an RNN where the distribution of recurrent weights is heavy tailed or scale-free. Recent experimental work has revealed that the distribution of synaptic strength in the brain follows a heavy-tailed distribution across a wide range of brain areas [28]. Several theoretical studies have also proposed potential roles for heavy-tailed synapse distributions in brain function [29,30].

The theoretical analysis of dynamics in large random RNNs, using methods like dynamical mean-field theory or the dynamical cavity method, typically relies on the central limit theorem. This allows the statistical behavior of neurons to be described as a Gaussian process (see Appendix A). However, this assumption breaks down for heavy-tailed weight distributions, where the central limit theorem is not guaranteed to hold. This leads to a fundamental difference in dynamical properties between conventional RNNs and those with heavy-tailed synapses. Although a theoretical framework to derive the dynamical mean-field equations for such RNNs has recently been proposed using the generalized central limit theorem [31], the analysis of neuronal correlations remains challenging. Consequently, we must rely on numerical simulations to investigate the neuronal correlations and memory capacity of these RNNs.

The model examined here is identical to that described in Sec. II, with the exception that the recurrent weights are i.i.d drawn from a Cauchy distribution [30]:

$$p(J_{ij}) = \frac{1}{\pi} \frac{\gamma/N}{(\gamma/N)^2 + J_{ij}^2}. \quad (15)$$

Here, the hyperparameter $\gamma > 0$ determines the strength of recurrent weights. In the limit of $N \rightarrow \infty$, the eigenvalue

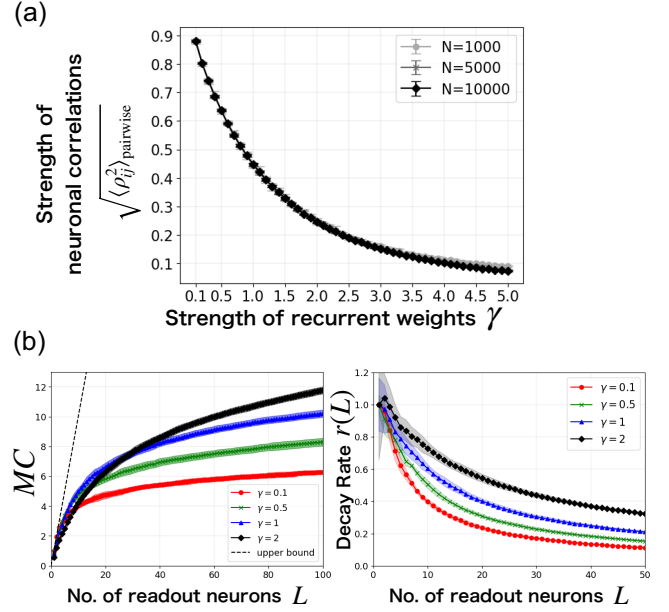


FIG. 5. Numerical simulations of neuronal correlations and memory capacity of RNNs with recurrent weights drawn from a Cauchy distribution. (a) The level of neuronal correlations, quantified by the root mean square of pairwise Pearson correlation coefficients, as a function of the scale parameter γ . Data are shown for network sizes $N = 1000, 5000$, and 10000 . (b) Left: MC as a function of the number of readout neurons, L . Right: Decay rate $r(L)$ as a function of L . Both plots show data for different values of γ with a fixed network size of $N = 10000$. Networks with a large γ exhibit weaker neuronal correlations and, consequently, a slower decay of $r(L)$. For both panels (a) and (b), the simulation time length is $T = 10^4$. The input and noise intensity are $\sigma_s^2 = 1.0$ and $\sigma_n^2 = 0.0$, respectively. The activation function is $\phi = \tanh$. Error bars and shaded areas represent mean $\pm$ std over ten network and input realizations.

spectrum of J is unbounded and its spectral radius is infinite for any $\gamma > 0$ [31], leading to chaotic dynamics in the RNN regardless of γ . On the other hand, when N is finite, whether the dynamics are chaotic or stable depends on the specific realization of J [32].

As demonstrated in Fig. 5(a), the level of neuronal correlations is controlled by the parameter γ , analogous to the role of the parameter g in our canonical Gaussian model. These neuronal correlations are considered to be of $O(1)$ with respect to the network size N , as indicated by the nearly constant values of $\sqrt{\langle \rho_{ij}^2 \rangle_{\text{pairwise}}}$ for increasing N .

Since we have observed that larger γ is associated with weaker neuronal correlations, our hypothesis predicts that the growth rate of memory capacity decays more slowly for larger γ . This prediction is supported by the results in Fig. 5(b), which shows that the decay rate $r(L)$ decreases more slowly as γ increases. Therefore, we conclude that our findings generalize to RNNs with a heavy-tailed distribution of recurrent weights.

C. RNN with reciprocal motifs

The models examined thus far assume that recurrent weights are mutually independent. However, this assumption

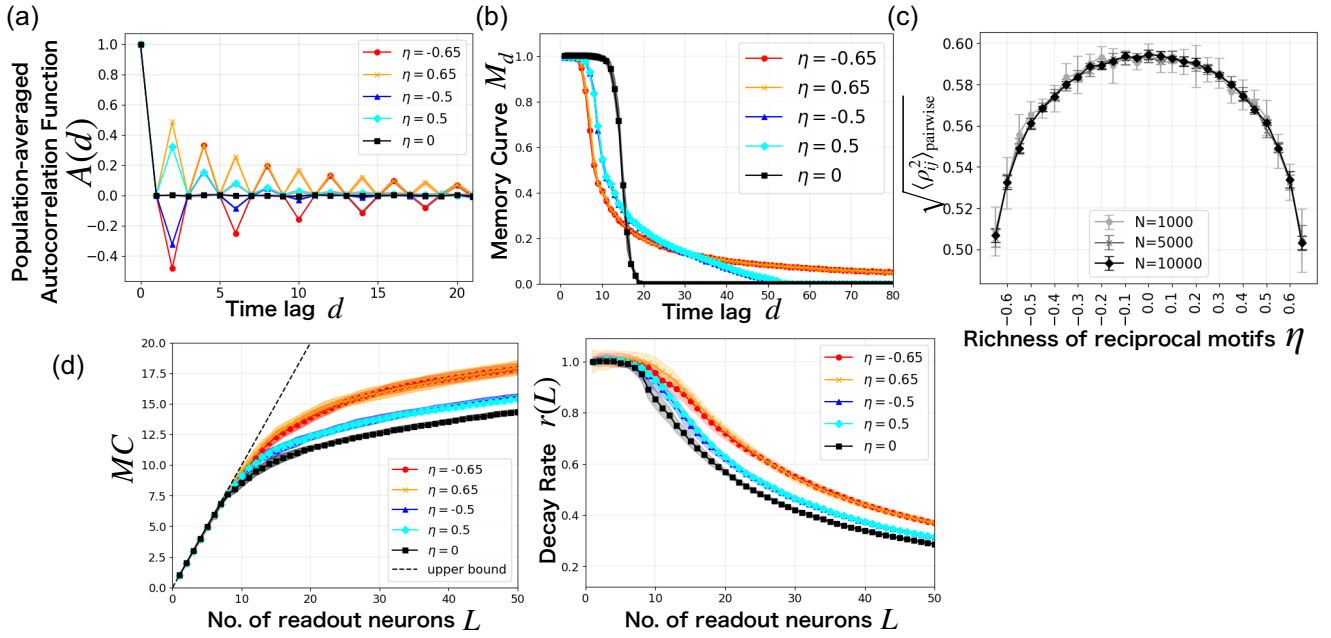


FIG. 6. Dynamical statistics and memory capacity of RNNs with reciprocal motifs. (a) Population-averaged autocorrelation function vs time lag d . A larger $|\eta|$ results in activity with a longer timescale. (b) Memory function M_d vs time lag d . A network with larger $|\eta|$ has better memory for long time lags d but worse memory for short ones. (c) RMS of pairwise Pearson correlation coefficients as a function of η . (d) Left: Memory capacity as a function of the number of readout neurons, L . Right: Decay rate as a function of L . Networks with larger $|\eta|$ exhibit weaker neuronal correlations and a correspondingly slower decay of $r(L)$. All numerical simulations are performed with input intensity $\sigma_s^2 = 0.1^2$, no neuronal noise $\sigma_n^2 = 0$, activation function $\phi = \tanh$, and simulation time length $T = 10^4$. The network size is $N = 10000$ for all panels except panel (c), which illustrates the results for $N = 1000, 5000, 10000$. Error bars and shaded areas represent mean $\pm$ std over ten network and input realizations.

is not always valid in real-world complex networks. In the brain, for example, certain connectivity motifs are observed more frequently than expected from random connections, which significantly impacts network dynamics [33]. Although neuronal correlations in such networks have yet to be studied, it is reasonable to assume that the prevalence of specific connectivity patterns also affects the strength of neuronal correlations. Therefore, we next investigate whether our findings on the relationship between neuronal correlations and memory capacity generalize to RNNs that incorporate such connectivity motifs.

The RNN model studied here is nearly identical to that described in Sec. II. The primary difference is that the recurrent weight matrix is now defined as gJ , where $g > 0$ is a gain parameter. The elements of the matrix J are drawn from a Gaussian distribution with zero mean, $1/N$ variance, and a specific correlation structure given by $[J_{ij}J_{ji}]_J = \eta/N$, following the approach of Marti *et al.* [34]. The parameter $\eta \in [-1, 1]$ controls the degree of (anti)symmetry of connections, i.e., the prevalence of reciprocal motifs. The model reduces to the canonical random RNN when $\eta = 0$, while $\eta = 1$ and $\eta = -1$ correspond to fully symmetric and fully antisymmetric connectivity matrices, respectively. It should be noted that we implement this RNN in a discrete-time model, unlike the original continuous-time formulation by Marti *et al.* [34], because memory capacity is not yet well defined for continuous-time systems.

In the absence of inputs and noise, the dynamics of the RNN can be analyzed through the eigenvalue spectrum of the

recurrent weight matrix gJ . According to the Elliptic Law from random matrix theory, the eigenvalues are uniformly distributed over an elliptical disk in the complex plane, with real and imaginary semiaxes of length, $g(1 + \eta)$ and $g(1 - \eta)$, respectively [35]. Therefore, the trivial fixed point at zero activity, $\mathbf{x} = \mathbf{0}$, is stable if and only if this entire eigenvalue spectrum lies within the unit circle. Otherwise, the system can potentially transition to other dynamical regimes, such as a nonzero fixed point, limit cycle, or chaos. While the phase diagram for the continuous-time version of this model was characterized by Marti *et al.* [34], the corresponding diagram for our discrete-time model is currently unknown. In fact, our own simulations suggest that its unstable regime contains a rich variety of dynamics, including complex oscillations and chaos (data not shown). Therefore, to ensure a clear interpretation of the results, we confine our numerical simulations to the parameter regime where the zero fixed point is stable. Specifically, we fix $g = 0.6$ and vary η in the range $-2/3 < \eta < 2/3$, which guarantees $g(1 \pm \eta) < 1$ is satisfied.

1. Autocorrelation function

First, we examine how the autocorrelation function of our model changes as a function of η . To do this, we numerically calculate the population-averaged autocorrelation function defined as $A(d) = \frac{1}{N} \sum_i \frac{\langle x_i(t)x_i(t-d) \rangle}{\langle x_i(t)^2 \rangle}$. As shown in Fig. 6(a), an RNN with a larger $|\eta|$ shows a stronger autocorrelation, $|A(d)|$. In contrast, the canonical random model ($\eta = 0$) shows a vanishing autocorrelation for all time lags

$d \geq 1$. This indicates that a higher density of reciprocal motifs promotes longer timescales in neuronal activity, a finding consistent with continuous-time models [34]. A key difference from the continuous-time model, however, is that $A(d)$ is nonzero only for even time lags and vanishes for odd d . Intuitively, this is because a perturbation to neuron i at time t can propagate back to itself at time $t + 2$ via a two-step path ($i \rightarrow j \rightarrow i$). The activity at time t is therefore correlated with the activity at $t + 2$ due to the underlying correlation between the synaptic weights J_{ij} and J_{ji} . It is possible to derive an analytical expression for $A(d)$ using the dynamical cavity method, but we leave this investigation for future work.

2. Memory function

An RNN with long-timescale activity is expected to be better at retaining inputs from the distant past. To test this prediction, we calculate M_d [defined in Eq. (3)] for all d , which is sometimes referred to as a *memory function*. Consistent with our prediction, a larger $|\eta|$ value, which yields a stronger autocorrelation, improves the recall of distant inputs, as shown in Fig. 6(b). However, these same RNNs perform worse at precisely recalling recent inputs, as indicated by their lower M_d values for short time lags. This happens because the long intrinsic neuronal timescales impose a slow timescale on the reservoir's output, preventing it from tracking rapidly changing Gaussian white noise inputs. We believe this trade-off between recent and distant memory is a general principle in reservoir computing. Indeed, a similar effect was reported in a previous study on a plastic spiking neural network reservoir [see Fig. 7(a) in Ref. [36]]. It is important to note that memory capacity, the primary focus of this study, is the sum of M_d over all d , and thus, a longer neuronal timescale does not necessarily guarantee a higher memory capacity.

3. Neuronal correlations

Before assessing memory capacity, we examine whether the parameter η affects the strength of neuronal correlations. Figure 6(c) demonstrates that as $|\eta|$ increases, the strength of neuronal correlations actually decreases. This result may seem counterintuitive. However, we confirmed that this observation is not a finite-size artifact, as the RMS correlation value remains constant for increasing network sizes N . Investigating whether this result is generalizable to other parameter sets and exploring its underlying mechanisms are promising directions for future work, potentially using the dynamical cavity method.

4. Scaling behavior of memory capacity

Given our observation that $|\eta|$ determines the level of neuronal correlations, our hypothesis suggests that it would also influence the decay rate of memory capacity. Our numerical simulations corroborate this prediction. As shown in Fig. 6(d), larger values of $|\eta|$, which produce weaker neuronal correlations, lead to a slower decrease in the decay rate $r(L)$. Notably, the value of η has no effect on the memory capacity of a single neuron, i.e., $MC(L = 1)$, while its impact becomes increasingly significant as the number of readout neurons grows. Therefore, the prevalence of reciprocal motifs affects the computational ability of the network as a whole, rather

than that of individual neurons, by modulating the structure of neuronal correlations.

VII. SCALING BEHAVIOR OF NONLINEAR COMPUTATIONAL CAPABILITY

While our previous analysis has focused on memory capacity, nonlinear computational ability is equally crucial for RC to address complex tasks. As established in previous works, there exists a trade-off between memory capacity and nonlinear computational ability for RC [13,27]. Thus, we anticipate that while the growth in memory capacity decelerates as L increases, the system's capacity for nonlinear computation shows a corresponding acceleration.

Information Processing Capacity (IPC) introduced by Dambre *et al.* [13] offers a task-independent metric for evaluating the performance of RC, enabling a comprehensive assessment of both linear and nonlinear computational capabilities. Since IPC theory is somewhat complicated, we provide detailed explanations in Appendix C 1. Put simply, the IPC for degree D , denoted by IPC_D , represents the reservoir's ability to approximate D th order polynomial functions of past white noise input signals. The IPC has two key properties. First, by definition, memory capacity is exactly equivalent to IPC_1 , and thus, IPC_D for $D \geq 2$ represents nonlinear computational ability. Second, it has been proven that the *total* IPC, defined as $IPC_{\text{total}} \equiv \sum_{D=1}^{\infty} IPC_D$, equals to the number of readout units, L , provided that the reservoir is a *fading memory* dynamical system [13]. A fading memory dynamical system is defined as a system whose state is uniquely determined solely by its input signals [13]. For example, a reservoir with chaotic dynamics or one subject to noise violates the fading memory property, because its state is determined not only by input signals but also by its initial state or by noise.

Figures 7(a)–7(c) illustrate how the values of IPC_D of the canonical random RNN, defined in Sec. II, change with increasing L across various hyperparameter combinations (g , σ_s^2 , and σ_n^2), as obtained through numerical simulations. Note that, due to the model's symmetry, IPC_D values for all even degrees D are identically zero. As shown, a decline in the growth of the IPC_1 (equivalent to memory capacity) is accompanied by an increase in the nonlinear computational ability (represented by IPC_D for $D \geq 3$), reflecting memory-nonlinearity trade-off.

Our finding is that, as L increases, the values of higher-degree IPC_D are activated *sequentially* from lower to higher orders, accompanied by their *supralinear* rises. This distinctive scaling behavior of nonlinear computational ability might be considered a trivial result in RNNs that satisfy the fading memory property, where the total capacity is constrained by $\sum_{D \geq 1} IPC_D = L$. In such cases, a sublinear growth in memory capacity ($= IPC_1$) necessarily implies that the remaining capacity grows supralinearly. However, the significance of our finding lies in its generality. This sequential, supralinear arising of nonlinear computational powers is observed not only in the RNNs that satisfy the fading memory property [Fig. 7(a)], but also in chaotic [Fig. 7(b)] and noisy [Fig. 7(c)] RNNs, where this property breaks down and thus the equality $MC + \sum_{D \geq 2} IPC_D = L$ does not hold. Furthermore, the phenomenon holds across various network architectures,

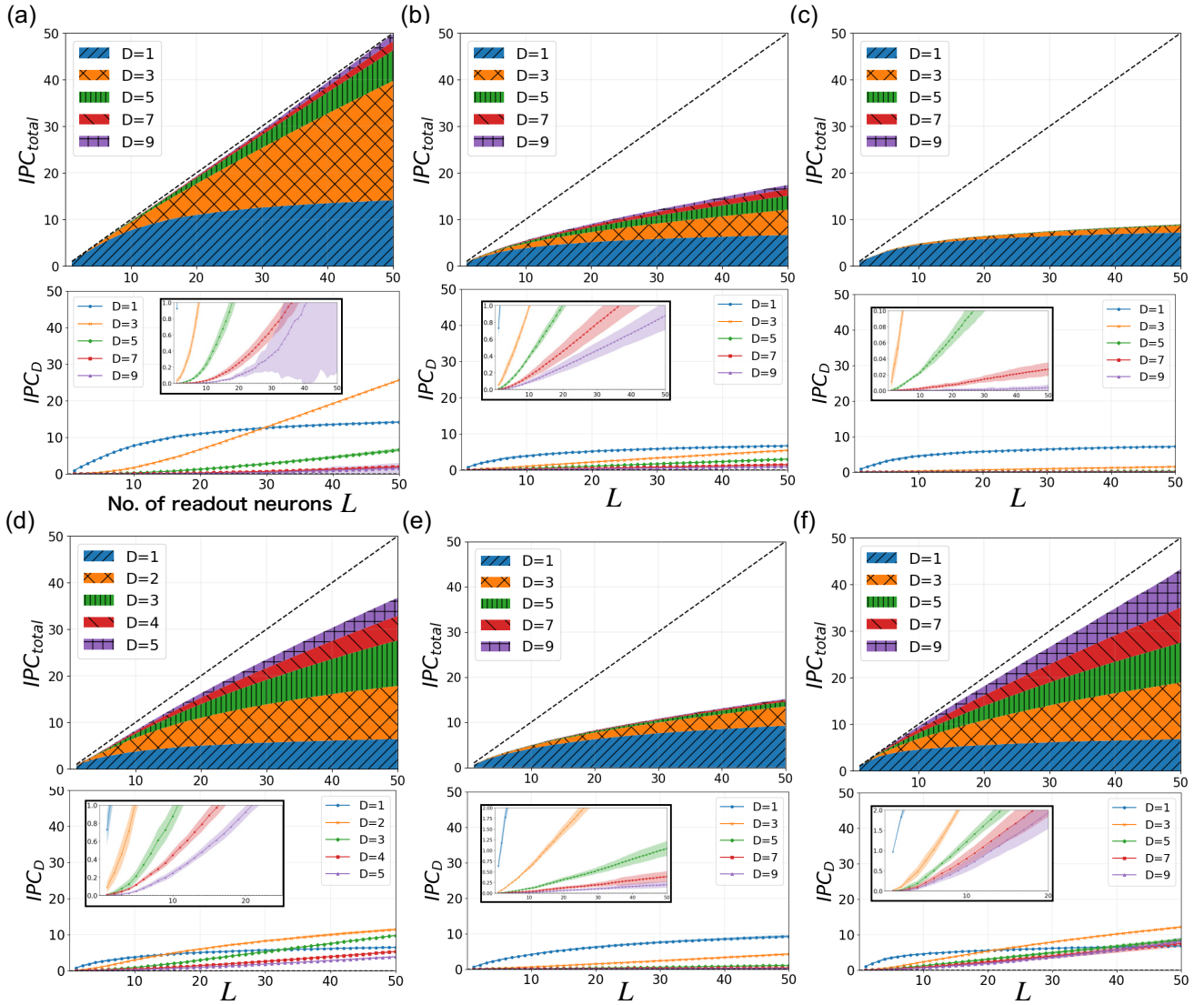


FIG. 7. Scaling behavior of IPC_D for various dynamical regimes and network architectures. In each main panel (a)–(f), the top plot shows a stacked representation of the IPC_D values for each degree D , while the bottom plot the individual IPC_D values as separate curves. The insets in the bottom plots provide an enlarged view of the smaller IPC values. Overall, the nonlinear capacity (IPC_D for $D \geq 2$) arises supralinearly, and the capacities are sequentially activated from lower to higher degrees D . These capacity values are computed for degrees up to $D = 9$, except in panel (d), where it is computed up to $D = 5$. (a) Canonical random RNN satisfying the fading memory property ($g = 0.9$, $\sigma_s^2 = 0.3^2$, $\sigma_n^2 = 0$). (b) Canonical random RNN in a chaotic regime ($g = 2.5$, $\sigma_s^2 = 2.0^2$, $\sigma_n^2 = 0$). (c) Canonical random RNN with neuronal noise ($g = 0.8$, $\sigma_s^2 = 0.3^2$, $\sigma_n^2 = 0.1^2$). The activation function for panels (a)–(c) is $\phi = \tanh$. (d) Canonical RNN with a ReLU activation function ($g = 1.0$, $\sigma_s^2 = 1.0$, $\sigma_n^2 = 0$). In contrast to the other cases, the IPC_D for even D does not vanish due to the asymmetry of the ReLU function. (e) Cauchy RNN with a tanh activation function ($\gamma = 2.0$, $\sigma_s^2 = 1.0$, $\sigma_n^2 = 0$). (f) RNN with reciprocal motifs and a tanh activation function ($g = 0.6$, $\eta = 0.5$, $\sigma_s^2 = 1.0$, $\sigma_n^2 = 0$). In panels (b), (c), and (e), the total measured capacity, $\sum_{D=1}^9 IPC_D$, falls significantly short of the theoretical limit L (dashed line). In contrast, the sum in panel (a) closely matches L . This discrepancy is attributed to a violation of the fading memory property. Specifically, the networks in panels (b) and (e) are chaotic, as confirmed by their positive maximum conditional Lyapunov exponents [21,37] ($\lambda_{\max} = 0.089 \pm 0.0041$ and $\lambda_{\max} = 0.23 \pm 0.011$, respectively); meanwhile, the network in panel (c) is subject to noise. Note that the networks in panels (d) and (f) satisfy the fading memory property, but the sums are slightly less than L due to the calculation being truncated at a finite degree ($D \leq 5$ and $D \leq 9$, respectively). Shaded areas represent mean $\pm$ std for ten different network and input signal realizations. All simulations are performed with a network size of $N = 1000$ and a simulation time of $T = 10^5$. The transient washout time is $T_{\text{washout}} = 10^3$.

including those with a ReLU activation function (Sec. VIA), heavy-tailed weight distributions (Sec. VIB), and reciprocal connectivity motifs (Sec. VIC), as shown in Figs. 7(d)–7(f).

As analytical derivation of higher-degree IPC_D is beyond our capability, the detailed mechanism underlying these observations remains unclear. However, it is plausible that the

neuronal correlation plays a pivotal role, analogous to its influence on memory capacity. This hypothesis is supported by speculation that each IPC_D would exhibit linear scaling with L if the effects of neuronal correlations were neglected, which contrasts with the observed supralinear and sequential emergence of IPC_D (see Appendix C2).

VIII. DISCUSSION

In the present study, we investigate how the computational capacity of large random RNN reservoirs scales with an increasing number of readout neurons, L . We develop a theoretical framework for deriving memory capacity that can incorporate the effect of neuronal correlations, which has been ignored for analytical simplicity. Our theory reveals that memory capacity scales *sublinearly* with L , and the extent of this sublinearity is significantly influenced by the strength of neuronal correlations, even though those correlations are quite weak. Although our theory is limited to cases where neuronal correlations scale as $O(1/\sqrt{N})$, a constraint imposed by specific assumptions about our model's hyperparameters [Eq. (4)], our numerical simulations confirm the robustness of this finding across a wider range of hyperparameters and in various network architectures. In addition, we numerically study the scaling behavior of nonlinear computational ability, showing that higher-order computational powers emerge *successively*, accompanied by their *supralinear* rises, enabling the reservoir to handle increasingly complex tasks.

The specific memory and nonlinearity requirements of a given task can be assessed using the framework proposed by Hülser *et al.* [38]. In this context, our results provide a practical principle for designing cost-effective RC systems: while increasing L raises computational and structural costs and yields diminishing returns for memory, it becomes essential when tasks demand higher-order nonlinear computations.

Beyond its practical usefulness, our work offers a lens for understanding RC performance. The computational ability of RC has been predominantly discussed from the perspective of proximity to the edge of chaos [14,15,39,40], typically for a fixed number of readout neurons [13,27]. To our knowledge, this is the first study to analyze the computational ability from the perspective of its scaling behavior with an increasing number of readout neurons and the corresponding role of neuronal correlations. Interestingly, a recent report has suggested that the edge of chaos is not always optimal [41], but the reason for this remains unclear. The perspective we propose may hold a crucial key to unraveling this puzzle.

It is also instructive to contrast our findings with other models of memory, such as Hopfield networks [42,43]. In the field of Hopfield networks, it is well established that making network activity sparse dramatically improves the capacity of associative memory [44,45]. Since sparse activity tends to weaken correlations between neurons [46], this implies that the strength of neuronal correlations is strongly associated with the memory ability of Hopfield networks. At first glance, our work might seem to merely replicate these prior findings. However, there is a fundamental difference between them. The mechanism by which Hopfield networks store information utilizes the fixed points of attractor dynamics, which can be described as static and long-term memory. RC, in contrast, takes the approach of temporarily retaining time-series information through reverberating neural activity within a reservoir network, which is a form of dynamic and short-term memory (or working memory). Given that their memory mechanisms are entirely different, there is no guarantee that a principle

governing static memory would apply equally to dynamic memory.

Our findings have implications that extend beyond the framework of RC. First, in neuroscience, the mammalian cortex in the awake state typically operates in an asynchronous regime characterized by near-zero neuronal correlations [47–49], while recent studies have shown the level of correlation is structured [50,51] and influenced by factors such as stimulus drive [52,53] and attention [54,55]. The functional impacts of the neuronal correlations have been extensively investigated through the lens of neural manifolds [56–58]. Within this framework, a neuronal network with a high level of correlated neuronal activity occupies a restricted region of N -dimensional phase space, the linear dimension of which is quantified by the effective dimension [59]. In the context of population coding, a well-established trade-off exists between representational richness and robustness: High-dimensional (i.e., low-correlation) activity enables complex and diverse information encoding, while low-dimensional (i.e., high-correlation) activity supports redundant encoding and noise robustness [57,58]. Existing theoretical frameworks, however, have primarily focused on the implications of neuronal correlations for static information processing, i.e., mapping static inputs to outputs. Our findings suggest a role for neuronal correlations in dynamic information processing, offering a hypothesis regarding their functional significance in the brain.

Second, in the field of deep learning theory, it has been established that neural networks trained via gradient-based methods, such as BPTT, can learn tasks both in the *lazy regime*, where their weights undergo only infinitesimal changes from their initial random values [60,61], and in the *rich regime*, where weights change substantially to extract task-relevant features [62–64]. Our insights may provide valuable understanding of the computational capacity of RNNs trained in the lazy regime, as their recurrent weights remain nearly identical to those of untrained, reservoir-style networks.

In conclusion, our study not only advances the understanding of RC but also sheds light on the broader implications of neuronal correlations and their effects on both biological and artificial systems. These insights offer directions for optimizing RNN architectures and exploring the fundamental nature of computation in complex networks.

ACKNOWLEDGMENTS

We thank Joni Dambre for providing the source code for computing information processing capacity and Tomoyuki Kubota for helpful discussions regarding its implementation. S.T. was supported by JSPS KAKENHI Grant No. JP22KJ1959. T.A. was supported by MEXT KAKENHI Grant No. 23H04467 and by JSPS KAKENHI Grants No. 24H00723 and No. 20K20520.

DATA AVAILABILITY

The data that support the findings of this article are openly available [65].

APPENDIX A: THEORETICAL ANALYSIS ON STATISTICS OF NEURONAL ACTIVITY

In this section, we derive statistics of neuronal activity based on dynamical cavity approach [26,66]. Several results and methodologies introduced here are frequently utilized in deriving memory capacity in Appendix B.

In deriving single-neuronal statistics such as the variance of $x_i(t)$ (see Appendix A 1) in the limit of $N \rightarrow \infty$, the same result can be obtained via dynamical mean-field theory [18,25]. However, since the dynamical mean-field theory transforms the N -body system into the effective single-body system, it is inherently unable to evaluate many-body neuronal statistics, such as neuronal correlations (see Appendix A 2) and memory capacity (see Appendix B 1). In contrast, the dynamical cavity method is, in principle, applicable to n -body neuronal statistics for any n . Furthermore, while the aforementioned limitation of dynamical mean-field theory has recently been addressed through subleading corrections to the saddle-point solution [67,68], the dynamical cavity method remains advantageous for its conceptual clarity.

In the following, we use the shorthand notation for brevity $x_{i,t} \equiv x_i(t)$, $s_t \equiv s(t)$, $\phi_{i,t} \equiv \phi(x_{i,t})$, and $\langle x_i x_j \rangle \equiv \langle x_i(t) x_j(t) \rangle$.

Before discussing the details, we first outline the necessity of employing the dynamical cavity method. Consider calculating the quenched variance of neuronal activity, $[x_{i,t}^2]$. For simplicity, we ignore input and noise signals. According to time evolution equation of neurons [Eq. (1)], $x_{i,t}$ is the sum of a large number ($N \gg 1$) of mutually independent quantities, $J_{ij}\phi(x_{j,t-1})$. This implies that $x_{i,t}$ follows a Gaussian distribution due to the central limit theorem, and thus, it is sufficient to determine the complete statistics of $x_{i,t}$ by calculating its mean and variance. Given the symmetry of the model, we have $[x_{i,t}^2] = 0$. On the other hand, we can intuitively calculate $[x_{i,t}^2]$ as

$$\begin{aligned} [x_{i,t}^2] &= \sum_{j,k}^N [J_{ij} J_{ik} \phi_{j,t-1} \phi_{k,t-1}] \\ &\stackrel{?}{=} \sum_{j,k}^N [J_{ij} J_{ik}] [\phi_{j,t-1} \phi_{k,t-1}] = g^2 [\phi_{j,t-1}^2] \\ &= g^2 \langle \phi(x)^2 \rangle_{x \sim \mathcal{N}(0, [x_{j,t-1}^2])}, \end{aligned} \quad (\text{A1})$$

yielding time series of $[x_{i,t}^2]$. Here, we assumed in the second equation that the recurrent weights, J_{ij} and J_{ik} , are independent of the neuronal activities, ϕ_j and ϕ_k . Fortunately, this assumption is validated in the limit of $N \rightarrow \infty$, as shown with the dynamical mean-field theory. However, this assumption is likely to break down when finite-size effects are taken into account. To circumvent this challenge, we need to employ the dynamical cavity method.

1. Single-neuronal statistics

We aim to derive the quenched average of the time variance of neuronal activity, $[x_i^2]$. Note that due to the self-averaging, it is equivalent to the population average of time variance to leading order, i.e., $[x_i^2] = \frac{1}{N} \sum_i^N \langle x_i^2 \rangle$. The symmetry of the model setting ensures that the quenched average of the time mean, $[x_i]$, vanishes.

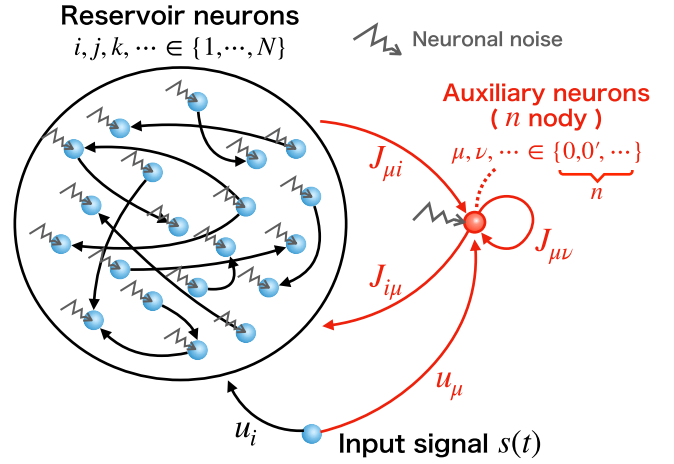


FIG. 8. Overview of dynamical cavity method. For simplicity, only one auxiliary neuron is illustrated, but in practice, n auxiliary neurons are introduced to calculate n -body statistics. Latin indices, $i, j, \dots$, are used to label reservoir neurons, while Greek indices, $\mu, \nu, \dots$, denote auxiliary neurons. The weights connecting auxiliary neurons to reservoir neurons, as well as the self-connections of auxiliary neurons, are sampled from the same distribution as the recurrent weights of the reservoir.

The overview of dynamical cavity method is described in Fig. 8. One auxiliary neuron, indexed by 0, is added to the original reservoir RNN whose neurons are indexed by $i = 1, \dots, N$. The original neuron is perturbed by the addition of the auxiliary neuron. By denoting the perturbation applied to the i th neuron at time t by $\delta_{i,t}$, we describe the dynamics of the auxiliary neuron as

$$\begin{aligned} x_{0,t+1} &= \sum_{i=1}^N J_{0i} \phi(x_{i,t} + \delta_{i,t}) + J_{00} \phi_{0,t} + u_0 s_{t+1} + \xi_{0,t+1} \\ &= \sum_{i=1}^N J_{0i} \phi_{i,t} + \sum_{i=1}^N J_{0i} \phi'_{i,t} \delta_{i,t} + J_{00} \phi_{0,t} \\ &\quad + u_0 s_{t+1} + \xi_{0,t+1}. \end{aligned} \quad (\text{A2})$$

Here, it should be noted that $x_{i,t}$ represents the original activity of a reservoir neuron before the introduction of the auxiliary neuron. Therefore, the value of J_{0i} is completely independent of reservoir activities, $x_{i,t}$ and $\phi_{i,t}$, in contrast to the intuitive argument in Eq. (A1).

We define $\chi_{ij,t-t-\tau} \equiv \delta x_{i,t} / \delta I_{j,t-\tau}$ as the linear susceptibility of the i th neuron at time t , induced by infinitesimal external inputs applied to the j th neuron at time $t - \tau$ ($\tau \geq 1$), $I_{j,t-\tau}$. Then, as the perturbation to i th neuron at time t , $\delta_{i,t}$, is the sum of all direct perturbations to $j \in \{1, \dots, N\}$ at time $t - \tau$ ($\tau \geq 1$) weighted by the corresponding susceptibility $\chi_{ij,t-t-\tau}$, we obtain

$$\delta_{i,t} = \sum_{\tau \geq 1} \sum_{j=1}^N \chi_{ij,t-t-\tau} J_{j0} \phi_{0,t-\tau-1}. \quad (\text{A3})$$

We substitute Eq. (A3) into Eq. (A2), yielding

$$\begin{aligned}
 x_{0,t+1} &= \underbrace{\sum_{i=1}^N J_{0i} \phi_{i,t}}_{\equiv \eta_t} + \sum_{\tau \geq 1} \underbrace{\sum_{i,j} \chi_{ij,t-\tau} J_{0i} J_{j0} \phi'_{i,t}}_{\equiv \kappa_\tau} \phi_{0,t-\tau-1} \\
 &\quad + J_{00} \phi_{0,t} + u_0 s_{t+1} + \xi_{0,t+1} \\
 &\equiv \eta_t + \sum_{\tau \geq 1} \kappa_\tau \phi_{0,t-\tau-1} + J_{00} \phi_{0,t} + u_0 s_{t+1} + \xi_{0,t+1}.
 \end{aligned} \tag{A4}$$

The physical interpretation of this equation is as follows. The first term, η_t , is the feedforward input from the reservoir. The second term, $\kappa_\tau \phi_{0,t-\tau-1}$, arises from the perturbation induced in the reservoir by the auxiliary neuron 0, which is read out by neuron 0. The third term originates from self-connection of the auxiliary neuron 0.

The order of the first term is evaluated as

$$\begin{aligned}
 [\eta_t^2] &= \sum_{i,j} [J_{0i} J_{0j} \phi_{i,t} \phi_{j,t}] \\
 &= \sum_{i,j} [J_{0i} J_{0j}] [\phi_{i,t} \phi_{j,t}] = g^2 [\phi_{i,t}^2] \sim O(1).
 \end{aligned} \tag{A5}$$

Here, we used in the second equation the fact that J_{0i} and J_{0j} are independent of $\phi_{i,t}$ and $\phi_{j,t}$ as mentioned above. The order of the third term is evaluated as

$$[J_{00} \phi_{0,t}]^2 \leq [J_{00}^2] \sim O(1/N). \tag{A6}$$

For evaluation of the second term in Eq. (A4), we assess the order of $\chi_{ij,t-\tau}$. We begin with $\tau = 1$. Since $\chi_{ij,t-1} = J_{ij} \phi'_{j,t-1}$, we find $[\chi_{ij,t-1}^2] \leq [J_{ij}^2] \sim O(1/N)$. Subsequently, for $\tau = 2$, $\chi_{ij,t-2} = \sum_k J_{ik} \phi'_{k,t-1} \chi_{kj,t-1}$, leading to $[\chi_{ij,t-2}^2] \leq \sum_k [J_{ik}^2] [(\chi_{kj,t-1})^2] \sim O(1/N)$, where we used the result for $\tau = 1$, $[(\chi_{kj,t-1})^2] \sim O(1/N)$. By inductively repeating this analysis, we conclude that $[\chi_{ij,t-\tau}^2] \sim O(1/N)$ for any $\tau \geq 1$. Therefore, the order of the second term is evaluated as

$$\begin{aligned}
 [\kappa_\tau^2] &= \sum_{i,j,k,l} [J_{0i} J_{j0} J_{0k} J_{l0}] [\phi'_{i,t} \phi'_{k,t} \chi_{ij,t-\tau} \chi_{kl,t-\tau}] \\
 &\leq \frac{g^4}{N^2} \sum_{i,j} [\chi_{ij,t-\tau}^2] \sim O(1/N).
 \end{aligned} \tag{A7}$$

Eventually, we obtain $[\kappa_\tau^2] \sim O(1/N)$ for $\tau \geq 1$.

Employing Cauchy-Schwarz inequality, it is shown that all cross terms between η_t , κ_τ , and $J_{00} \phi_{0,t}$ are $O(1/\sqrt{N})$.

Leveraging all these order estimates, from Eq. (A4), we obtain

$$\begin{aligned}
 [x_{0,t+1}^2] &= \left[\left(\eta_t + \sum_{\tau \geq 1} \kappa_\tau \phi_{0,t-\tau-1} + J_{00} \phi_{0,t} + u_0 s_{t+1} + \xi_{0,t+1} \right)^2 \right] \\
 &= [\eta_t^2] + [u_0^2] s_{t+1}^2 + [\xi_{0,t+1}^2] + O(1/\sqrt{N}) \\
 &= g^2 [\phi_{i,t}^2] + s_{t+1}^2 + \sigma_n^2 + O(1/\sqrt{N}).
 \end{aligned} \tag{A8}$$

Thus, to leading order, we can ignore the nontrivial terms arising from the effects of the perturbations induced by the auxiliary neuron. The statistical behavior of the auxiliary neuron is equivalent to that of the reservoir neurons, which enables the replacement of $[x_{0,t}^2]$ with $[\phi_{i,t}^2]$, resulting in

$$[x_{i,t+1}^2] = g^2 [\phi_{i,t}^2] + s_{t+1}^2 + \sigma_n^2 + O(1/\sqrt{N}). \tag{A9}$$

For the scaling assumption given by Eq. (4), we can assume the inputs to be subleading, obtaining

$$[x_{i,t+1}^2] = g^2 [\phi_{i,t}^2] + \sigma_n^2 + O(1/\sqrt{N}). \tag{A10}$$

Taking the time average and the network size limit ($N \rightarrow \infty$) yields a dynamical mean-field equation,

$$[\langle x_i^2 \rangle] = g^2 [\langle \phi_i^2 \rangle] + \sigma_n^2, \tag{A11}$$

which is identical to Eq. (8) in the main text.

For later use, we evaluate the order of the quenched variance of $\langle x_i^2 \rangle$, denoted by $\text{Var}[\langle x_i^2 \rangle]$, which is equivalent to the population variance because of the self-averaging property. Since the cavity method allows us to ignore the effects of perturbation induced by the auxiliary neuron, we easily calculate $[\langle x_0^2 \rangle^2]$ as

$$\begin{aligned}
 [\langle x_0^2 \rangle^2] &= \left[\left\langle \left(\sum_i J_{0i} \phi_{i,t-1} + u_0 s_t + \xi_{0,t} \right)^2 \right\rangle^2 \right] \\
 &= \frac{2g^4}{N^2} \sum_{i,j} [\langle \phi_i \phi_j \rangle^2] + \frac{g^4}{N^2} \sum_{i,j} [\langle \phi_i^2 \rangle \langle \phi_j^2 \rangle] \\
 &\quad + 2g^2 (\sigma_s^2 + \sigma_n^2) [\langle \phi_i^2 \rangle] + 3\sigma_s^4 + \sigma_n^4 \\
 &= 2g^4 [\langle \phi_i \phi_j \rangle^2]_{i \neq j} + \frac{3g^4}{N} [\langle \phi_i^2 \rangle^2] + g^4 [\langle \phi_i^2 \rangle \langle \phi_j^2 \rangle]_{i \neq j} \\
 &\quad + 2g^2 (\sigma_s^2 + \sigma_n^2) [\langle \phi_i^2 \rangle] + 3\sigma_s^4 + \sigma_n^4.
 \end{aligned} \tag{A12}$$

Employing Eqs. (A9) and (A12), we obtain

$$\begin{aligned}
 \text{Var}[\langle x_0^2 \rangle] &= [\langle x_0^2 \rangle^2] - [\langle x_0^2 \rangle]^2 \\
 &= g^4 ([\langle \phi_i^2 \rangle \langle \phi_j^2 \rangle]_{i \neq j} - [\langle \phi_i^2 \rangle]^2) + 2g^4 [\langle \phi_i \phi_j \rangle^2]_{i \neq j} \\
 &\quad + \frac{3g^4}{N} [\langle \phi_i^2 \rangle^2] + 2\sigma_s^4.
 \end{aligned} \tag{A13}$$

Specifically, when $\sigma_s^2 = \tilde{\sigma}_s^2/\sqrt{N}$, we have

$$\begin{aligned}
 \text{Var}[\langle x_0^2 \rangle] &= g^4 ([\langle \phi_i^2 \rangle \langle \phi_j^2 \rangle]_{i \neq j} - [\langle \phi_i^2 \rangle]^2) + 2g^4 [\langle \phi_i \phi_j \rangle^2]_{i \neq j} \\
 &\quad + \frac{3g^4}{N} [\langle \phi_i^2 \rangle^2] + \frac{2\tilde{\sigma}_s^4}{N}.
 \end{aligned} \tag{A14}$$

As shown in Appendix A2, we see $[\langle \phi_i \phi_j \rangle^2]_{i \neq j} \sim O(1/N)$. The self-averaging property allows us to evaluate the first term as

$$\begin{aligned}
 &[\langle \phi_i^2 \rangle \langle \phi_j^2 \rangle]_{i \neq j} - [\langle \phi_i^2 \rangle]^2 \\
 &\sim \frac{1}{N^2 - N} \sum_{i,j (i \neq j)} \langle \phi_i^2 \rangle \langle \phi_j^2 \rangle - \left(\frac{1}{N} \sum_i \langle \phi_i^2 \rangle \right)^2 \sim O\left(\frac{1}{N}\right).
 \end{aligned} \tag{A15}$$

Consequently, we obtain $\text{Var}[\langle x_i^2 \rangle] \sim O(1/N)$.

A direct consequence of the scaling $\text{Var}[\langle x_i^2 \rangle] \sim O(1/N)$ is the following equation to leading order:

$$\left[\frac{1}{\langle x_i^2 \rangle} \right] = \frac{1}{[\langle x_i^2 \rangle]}, \quad (\text{A16})$$

provided that $[\langle x_i^2 \rangle] \sim O(1)$.

2. Statistics for neuronal correlations

We aim to derive the quenched average of the squared neuronal correlation, $[\langle x_i x_j \rangle^2]$, which is equivalent to the population average to leading order, $[\langle x_i x_j \rangle^2] = \frac{1}{N^2 - N} \sum_{i,j(i \neq j)}^N \langle x_i x_j \rangle^2$. Similarly to the single-neuronal statistics, the model's symmetry ensures that $[\langle x_i x_j \rangle] = 0$.

As in Appendix A 1, we introduce two auxiliary neurons indexed by 0 and 0' to the original reservoir RNN. Following the notation in Clark's work [26], Latin indices are used to label reservoir neurons, while Greek indices denote auxiliary neurons. The dynamics of the auxiliary neurons $\mu \in \{0, 0'\}$ is described by

$$\begin{aligned} x_{\mu,t+1} &= \sum_i^N J_{\mu i} \phi(x_{i,t} + \delta_{i,t}) + \sum_v J_{\mu v} \phi_{v,t} \\ &\quad + u_{\mu} s_{t+1} + \xi_{\mu,t+1} \\ &= \sum_i^N J_{\mu i} \phi_{i,t} + \sum_i^N J_{\mu i} \phi'_{i,t} \delta_{i,t} + \sum_v J_{\mu v} \phi_{v,t} \\ &\quad + u_{\mu} s_{t+1} + \xi_{\mu,t+1}. \end{aligned} \quad (\text{A17})$$

The perturbation $\delta_{i,t}$ is given by

$$\delta_{i,t} = \sum_{\tau \geq 1} \sum_v \sum_j \chi_{ij,t-\tau} J_{jv} \phi_{v,t-\tau-1}. \quad (\text{A18})$$

Substituting Eq. (A18) into Eq. (A17), we obtain

$$\begin{aligned} x_{\mu,t+1} &= \underbrace{\sum_i^N J_{\mu i} \phi_{i,t}}_{\equiv \eta_{\mu,t}} + \sum_{\tau \geq 1} \sum_v \underbrace{\sum_{i,j}^N J_{\mu i} J_{jv} \phi'_{i,t} \chi_{ij,t-\tau} \phi_{v,t-\tau-1}}_{\equiv \kappa_{\mu v,t-\tau}} \\ &\quad + \sum_v J_{\mu v} \phi_{v,t} + u_{\mu} s_{t+1} + \xi_{\mu,t+1} \\ &\equiv \eta_{\mu,t} + \sum_v \sum_{\tau \geq 1} \kappa_{\mu v,t-\tau} \phi_{v,t-\tau-1} + \sum_v J_{\mu v} \phi_{v,t} \\ &\quad + u_{\mu} s_{t+1} + \xi_{\mu,t+1}. \end{aligned} \quad (\text{A19})$$

The physical interpretation of this equation is analogous to the case of single-neuronal statistics [Eq. (A4)]: $\eta_{\mu,t}$ denotes the direct feedforward input from the reservoir to the auxiliary neuron μ , and $\kappa_{\mu v,t-\tau} \phi_{v,t-\tau-1}$ arises from the perturbation induced in the reservoir by the auxiliary neuron v at time $t - \tau - 1$, which is read out by the auxiliary neuron μ at time t .

Expanding $[\langle x_0 x_{0'} \rangle^2]$ using Eq. (A19), we need to evaluate all cross terms. We can accomplish this task through direct calculation, analogous to the approach used in Appendix A 1. First,

$$\begin{aligned} [\langle \eta_{0,t} \eta_{0',t} \rangle^2] &= \sum_{i,j,k,l}^N [J_{0i} J_{0'j} J_{0k} J_{0'l} \langle \phi_i \phi_j \rangle \langle \phi_k \phi_l \rangle] \\ &= \sum_{i,j,k,l}^N [J_{0i} J_{0'j} J_{0k} J_{0'l}] [\langle \phi_i \phi_j \rangle \langle \phi_k \phi_l \rangle] \\ &= \frac{g^4}{N} [\langle \phi_i^2 \rangle^2] + g^4 [\langle \phi_i \phi_j \rangle^2]_{(i \neq j)}, \end{aligned} \quad (\text{A20})$$

where we used in the second equation the fact that the weights connecting to auxiliary neurons are independent from the reservoir neurons' activities. By making an ansatz, $[\langle \phi_i \phi_j \rangle^2]_{(i \neq j)} \sim O(1/N)$, we obtain $[\langle \eta_{0,t} \eta_{0',t} \rangle^2] \sim O(1/N)$.

Next, calculating the cross term between the second and third terms yields

$$\begin{aligned} &\left[\left\langle \sum_{\mu} \sum_{\tau \geq 1} \kappa_{0\mu,t-\tau} \phi_{\mu,t-\tau-1} \sum_v J_{0'v} \phi_{v,t} \right\rangle^2 \right] \\ &= \frac{g^2}{N} \sum_{\mu,v,\rho} \sum_{\tau,\tau'} [\langle \kappa_{0\mu,t-\tau} \phi_{\mu,t-\tau-1} \phi_{v,t} \rangle \langle \kappa_{0\rho,t-\tau'} \phi_{\rho,t-\tau'-1} \phi_{v,t} \rangle] \\ &= \frac{g^2}{N} \sum_{\mu,v,\rho} \sum_{\tau,\tau'} \sum_{i,j,k,l}^N [J_{0i} J_{j\mu} J_{0k} J_{l\nu}] [\langle \phi'_{i,t} \chi_{ij,t-\tau} \phi_{\mu,t-\tau-1} \phi_{v,t} \rangle \langle \phi'_{k,t} \chi_{kl,t-\tau'} \phi_{\rho,t-\tau'-1} \phi_{v,t} \rangle] \\ &= \frac{g^6}{N^3} \sum_{\mu} \sum_{\tau,\tau'} \sum_{i,j}^N [\langle \phi'_{i,t} \rangle^2 \langle \chi_{ij,t-\tau} \chi_{ij,t-\tau'} \rangle \langle \phi_{\mu,t-\tau-1} \phi_{\mu,t} \rangle \langle \phi_{\mu,t-\tau'-1} \phi_{\mu,t} \rangle] \\ &\sim O(1/N^2), \end{aligned} \quad (\text{A21})$$

where we used $\chi_{ij,t-\tau}^2 \sim O(1/\sqrt{N})$ for any $\tau \geq 1$, as shown before. Therefore, this cross term becomes subleading in $[\langle x_0 x_{0'} \rangle^2]$.

The cross term between the first and second terms vanishes as

$$\begin{aligned} & \left[\left\langle \eta_{0,t} \sum_{\mu} \sum_{\tau \geq 1} \kappa_{0'\mu,t} \phi_{\mu,t-\tau-1} \right\rangle^2 \right] \\ &= \left[\left(\sum_{i,j,k} \sum_{\mu} \sum_{\tau \geq 1} J_{0i} J_{0'j} J_{k\mu} \langle \phi_{i,t} \phi'_{j,t} \phi_{\mu,t-\tau-1} \rangle \right)^2 \right] \\ &= \left[\left(\sum_{i,j,k} \sum_{\mu} \sum_{\tau \geq 1} J_{0i} J_{0'j} J_{k\mu} \langle \phi_{i,t} \phi'_{j,t} \rangle \langle \phi_{\mu,t} \rangle \right)^2 \right] \\ &= 0, \end{aligned} \quad (\text{A23})$$

where the second equality holds because the activity of the reservoir neuron is independent of that of the auxiliary neurons.

Repeating these calculations, all cross terms are shown to scale as $O(1/N^2)$, except for the two $O(1/N)$ terms, $[\langle \eta_{0,t} \eta_{0',t} \rangle^2]$ and $[\langle (u_{0s_t})(u_{0's_t}) \rangle^2] = \tilde{\sigma}_s^4/N$. Consequently,

$$[\langle x_0 x_{0'} \rangle^2] = \frac{g^4}{N} [\langle \phi_i^2 \rangle^2] + g^4 [\langle \phi_i \phi_j \rangle^2]_{(i \neq j)} + \frac{\tilde{\sigma}_s^4}{N} + O\left(\frac{1}{N^2}\right). \quad (\text{A24})$$

Replacing $[\langle x_0 x_{0'} \rangle^2]$ with $[\langle x_i x_j \rangle^2]_{(i \neq j)}$, we obtain

$$[\langle x_i x_j \rangle^2] = \frac{g^4}{N} [\langle \phi_i^2 \rangle^2] + g^4 [\langle \phi_i \phi_j \rangle^2] + \frac{\tilde{\sigma}_s^4}{N} + O\left(\frac{1}{N^2}\right), \quad (\text{A25})$$

for $i \neq j$. From the ansatz, $[\langle \phi_i \phi_j \rangle^2]_{(i \neq j)} \sim O(1/N)$, we see that $[\langle x_i x_j \rangle^2]_{(i \neq j)} \sim O(1/N)$.

Consequently, similar to single-neuron statistics, the non-trivial perturbation induced by auxiliary neurons, i.e., the second and third terms in Eq. (A19), does not contribute to the leading order of neuronal correlations. In contrast, in the case of a continuous-time RNN, the perturbation induced by auxiliary neurons contributes to the leading order of neuronal correlations [26]. This discrepancy stems from the difference in the strength of the autocorrelation, $\langle x_i(t) x_i(t - \tau) \rangle_t$, which scales as $O(1)$ in the continuous-time model, but as $O(1/N^{\tau/2})$ in our discrete-time model. This property of negligible perturbation effects by auxiliary neurons in a discrete-time model greatly simplifies the analytical derivation of memory capacity (see Appendix B 1). Note that even in a discrete-time model, perturbation-induced effects contribute to the leading-order term of neuronal correlations if the RNN dynamics include a leak term, i.e., $x_i(t+1) = \gamma x_i(t) + \sum_j J_{ij} \phi_j(t)$ with $\gamma > 0$, which yields an $O(1)$ autocorrelation.

To calculate $[\langle \phi_i \phi_j \rangle^2]$, we introduce Price's theorem [25]. For two Gaussian variables, $(z_1, z_2) \sim \mathcal{N}(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \tau \\ \tau & 1 \end{pmatrix})$, and any smooth activation function ϕ , we define $f_\phi(c_1, c_2, \tau) \equiv \langle \phi(z_1) \phi(z_2) \rangle$. Price's theorem then states

$$\partial_\tau f_\phi(c_1, c_2, \tau) = f_\phi'(c_1, c_2, \tau). \quad (\text{A26})$$

Iteratively applying this theorem yields $\partial_\tau^n f_\phi(c_1, c_2, \tau) = f_\phi^{(n)}(c_1, c_2, \tau)$, which allows us to expand $\langle \phi(z_1) \phi(z_2) \rangle$ around

$\tau = 0$, resulting in

$$\begin{aligned} \langle \phi(z_1) \phi(z_2) \rangle &= \sum_{n \geq 0} \frac{1}{n!} \tau^n f_\phi^{(n)}(c_1, c_2, 0) \\ &= \sum_{n \geq 0} \frac{1}{n!} \langle z_1 z_2 \rangle^n \langle \phi^{(n)}(z_1) \rangle \langle \phi^{(n)}(z_2) \rangle. \end{aligned} \quad (\text{A27})$$

Applying Eq. (A27) to $\langle \phi_i \phi_j \rangle$ yields

$$\begin{aligned} \langle \phi_i \phi_j \rangle &= \langle \phi'_{i,t} \rangle \langle \phi'_{j,t} \rangle \langle x_i x_j \rangle + \frac{1}{3!} \langle \phi_{i,t}^{(3)} \rangle \langle \phi_{j,t}^{(3)} \rangle \langle x_i x_j \rangle^3 \\ &\quad + \frac{1}{5!} \langle \phi_{i,t}^{(5)} \rangle \langle \phi_{j,t}^{(5)} \rangle \langle x_i x_j \rangle^5 + \dots, \end{aligned} \quad (\text{A28})$$

where we used $\langle \phi^{(2n)}(x) \rangle = 0$ since $\phi^{(2n)}$ is an odd function. Therefore, we obtain

$$[\langle \phi_i \phi_j \rangle^2] = [\langle \phi'_{i,t} \rangle^2] [\langle x_i x_j \rangle^2] + O\left(\frac{1}{N^2}\right). \quad (\text{A29})$$

Substituting Eq. (A29) into Eq. (A25) yields

$$\begin{aligned} [\langle x_i x_j \rangle^2] &= \frac{g^4}{N} \langle \phi(x)^2 \rangle_*^2 + g^4 \langle \phi'(x) \rangle_*^4 [\langle x_i x_j \rangle^2] \\ &\quad + \frac{\tilde{\sigma}_s^4}{N} + O\left(\frac{1}{N^2}\right), \end{aligned} \quad (\text{A30})$$

where $\langle f(x) \rangle_*$ denotes taking the average of $f(x)$ where $x \sim \mathcal{N}(0, [\langle x_i^2 \rangle])$. Therefore, we obtain

$$[\langle x_i x_j \rangle^2] = \frac{1}{N} \frac{g^4 \langle \phi(x)^2 \rangle_*^2 + \tilde{\sigma}_s^4}{1 - g^4 \langle \phi'(x) \rangle_*^4} + O\left(\frac{1}{N^2}\right), \quad (\text{A31})$$

which is self-consistent with the ansatz, $[\langle \phi_i \phi_j \rangle^2] \sim O(1/N)$.

3. Proof of $0 < g \langle \phi'(x) \rangle_* < 1$

We prove an inequality $0 < g \langle \phi'(x) \rangle_* < 1$ for any $g > 0$, $\sigma_s^2 \geq 0$, $\sigma_n^2 \geq 0$, and an odd saturated sigmoid function satisfying $\phi'(0) = 1$, $\phi'(x) > 0$, $\phi''(x) \leq 0$ ($x \geq 0$), and $\phi(\pm\infty) = \pm 1$. We use the abbreviation $K \equiv [\langle x_i^2 \rangle]$. It is evident that $g \langle \phi'(x) \rangle_* > 0$ because $g > 0$ and $\phi'(x) \geq 0$. We prove that $g \langle \phi'(x) \rangle_* < 1$ below.

$\langle \phi'(x) \rangle_*$ is a decreasing function of K , because

$$\begin{aligned} \frac{d}{dK} \langle \phi'(x) \rangle_* &= \frac{d}{dK} \int_{-\infty}^{\infty} \phi'(\sqrt{K}z) D\mathbf{z} \\ &= \frac{1}{\sqrt{K}} \int_0^{\infty} z \phi''(\sqrt{K}z) D\mathbf{z} \leq 0, \end{aligned} \quad (\text{A32})$$

where $D\mathbf{z}$ denotes a normal Gaussian measure. Since K increases with both σ_n^2 and σ_s^2 , it follows that $g \langle \phi'(x) \rangle_*$ decreases correspondingly. Therefore, it is enough to show that $g \langle \phi'(x) \rangle_* \leq 1$ only for the case of no input and noise, $\sigma_n^2 = \sigma_s^2 = 0$. In this case, from Eq. (8), K is determined by solving

$$K = g^2 \int \phi^2(\sqrt{K}z) D\mathbf{z}. \quad (\text{A33})$$

For $g \leq 1$, the only solution of this equation is $K = 0$, resulting in $g \langle \phi'(x) \rangle_* = g \phi'(0) \leq 1$. For $g > 1$, K is uniquely determined as a function of g , so that g can be assumed to be

a function of K ,

$$g^2 = \frac{K}{\int \phi^2(\sqrt{K}z)Dz}. \quad (\text{A34})$$

Therefore, it suffices to show the inequality

$$g^2 \left(\int \phi'(\sqrt{K}z)Dz \right)^2 = \frac{K \left(\int \phi'(\sqrt{K}z)Dz \right)^2}{\int \phi^2(\sqrt{K}z)Dz} \leq 1, \quad (\text{A35})$$

for any $K > 0$. Employing integration by parts and Cauchy-Schwarz inequality, we can prove this inequality as

$$K \left(\int \phi'(\sqrt{K}z)Dz \right)^2 = \left(\int z\phi(\sqrt{K}z)Dz \right)^2 \leq \int z^2 Dz \int \phi^2(\sqrt{K}z)Dz = \int \phi^2(\sqrt{K}z)Dz. \quad (\text{A36})$$

APPENDIX B: ANALYSIS ON MEMORY CAPACITY

1. Analytical derivation of memory capacity

In this subsection, we derive the analytical solution of memory capacity. By simple calculation, M_d defined in Eq. (3) can be expressed as

$$M_d = \frac{\mathbf{a}_d^\top C^{-1} \mathbf{a}_d}{\langle s(t)^2 \rangle}, \quad (\text{B1})$$

where the elements of $\mathbf{a} \in \mathbb{R}^L$ and $C \in \mathbb{R}^{L \times L}$ are, respectively, $(\mathbf{a}_d)_i \equiv \langle s(t-d)x_i(t) \rangle$ and $C_{ij} \equiv \langle x_i(t)x_j(t) \rangle$ (i, j indicate the indices of the readout neurons). Under the scaling assumption of model hyperparameters [Eq. (4)], the inverse of the covariance matrix C can be expanded using Neumann series expansion:

$$C^{-1} = \sum_{n=0}^{\infty} C_{\text{diag}}^{-1} (-C_{\text{nondiag}} C_{\text{diag}}^{-1})^n, \quad (\text{B2})$$

where $(C_{\text{diag}})_{ij} \equiv \delta_{ij} \langle x_i(t)^2 \rangle$ and $(C_{\text{nondiag}})_{ij} \equiv (1 - \delta_{ij}) \langle x_i(t)x_j(t) \rangle$.

This series expansion is ensured to converge when the spectral norm is less than 1, $\|C_{\text{nondiag}} C_{\text{diag}}^{-1}\| < 1$. However, calculating the exact value of the spectral norm is a challenging task. Employing the inequality $\|C_{\text{nondiag}} C_{\text{diag}}^{-1}\| \leq \|C_{\text{nondiag}} C_{\text{diag}}^{-1}\|_F$, where $\|\cdot\|_F$ denotes a Frobenius norm, we can obtain the sufficient condition for the convergence, $\|C_{\text{nondiag}} C_{\text{diag}}^{-1}\|_F < 1$. The value of $\|C_{\text{nondiag}} C_{\text{diag}}^{-1}\|_F$ is calculated as

$$\|C_{\text{nondiag}} C_{\text{diag}}^{-1}\|_F = \sum_{i,j(i \neq j)}^{\alpha\sqrt{N}} \frac{\langle x_i x_j \rangle^2}{\langle x_i^2 \rangle^2} = \alpha^2 N \left[\frac{\langle x_i x_j \rangle_{(i \neq j)}^2}{\langle x_i^2 \rangle^2} \right] = \alpha^2 N \frac{[\langle x_i x_j \rangle_{(i \neq j)}]^2}{[\langle x_i^2 \rangle]^2} = \frac{\alpha^2}{[\langle x_i^2 \rangle]^2} \frac{g^4 \langle \phi(x)^2 \rangle_*^2 + \tilde{\sigma}_s^4}{1 - g^4 \langle \phi'(x) \rangle_*^4}, \quad (\text{B3})$$

where the second equality employs the self-averaging property, and the final equality utilizes Eq. (A31). In addition, the third equality follows from Eq. (A16). Consequently, we obtain the sufficient condition

$$\alpha^2 < \frac{[\langle x_i^2 \rangle]^2}{\tilde{\sigma}_s^4 + ([\langle x_i^2 \rangle] - \sigma_n^2)^2} (1 - (g\langle \phi'(x) \rangle_*)^4), \quad (\text{B4})$$

where we used the equation, $g^2 \langle \phi(x)^2 \rangle_* = [\langle x_i^2 \rangle] - \sigma_n^2$, derived from the dynamical mean-field equation, Eq. (A11).

Substituting the series expansion expression for C^{-1} into Eq. (B1), we obtain

$$M_d = \frac{\sqrt{N}}{\tilde{\sigma}_s^2} \sum_i \frac{\langle s_{t-d} x_{i,t} \rangle^2}{\langle x_i^2 \rangle} - \frac{\sqrt{N}}{\tilde{\sigma}_s^2} \sum_{\substack{i,j \\ (i \neq j)}} \frac{\langle s_{t-d} x_{i,t} \rangle \langle x_i x_j \rangle \langle s_{t-d} x_{j,t} \rangle}{\langle x_i^2 \rangle \langle x_j^2 \rangle} + \frac{\sqrt{N}}{\tilde{\sigma}_s^2} \sum_{\substack{i,j,k \\ (i \neq j, j \neq k)}} \frac{\langle s_{t-d} x_{i,t} \rangle \langle x_i x_j \rangle \langle x_j x_k \rangle \langle s_{t-d} x_{k,t} \rangle}{\langle x_i^2 \rangle \langle x_j^2 \rangle \langle x_k^2 \rangle} - \dots, \quad (\text{B5})$$

where for simplicity we used the shorthand notation $x_{i,t} \equiv x_i(t)$, $s_t \equiv s(t)$, and $\langle x_i x_j \rangle \equiv \langle x_i(t)x_j(t) \rangle$.

In the large network size limit, we can assume self-averaging for M_d , that is, $\lim_{N \rightarrow \infty} M_d = [M_d]$, where the square bracket denotes the average over network realizations, known as “quenched average” [25]. In addition, as neurons exhibit statistically identical dynamics, the quenched averaging operation eliminates the dependence of each term in Eq. (3) on neuron indices $i, j, k, \dots$. Consequently, Eq. (B5) can be reduced to

$$[M_d] = \frac{\alpha N}{\tilde{\sigma}_s^2} \left[\frac{\langle s_{t-d} x_{i,t} \rangle^2}{\langle x_i^2 \rangle} \right] - \frac{\alpha^2 N^{\frac{3}{2}}}{\tilde{\sigma}_s^2} \left[\frac{\langle s_{t-d} x_{i,t} \rangle \langle x_i x_j \rangle \langle s_{t-d} x_{j,t} \rangle}{\langle x_i^2 \rangle \langle x_j^2 \rangle} \right]_{i \neq j} + \frac{\alpha^3 N^2}{\tilde{\sigma}_s^2} \left[\frac{\langle s_{t-d} x_{i,t} \rangle \langle x_i x_j \rangle \langle x_j x_k \rangle \langle s_{t-d} x_{k,t} \rangle}{\langle x_i^2 \rangle \langle x_j^2 \rangle \langle x_k^2 \rangle} \right]_{i \neq j, j \neq k} - \dots \quad (\text{B6})$$

For the solution of memory capacity, it is enough to calculate each quenched averaging term in Eq. (B6). We perform this calculation by the dynamical cavity method, similar to the approach described in Appendix A. As noted there, for a reservoir RNN with discrete time dynamics, perturbations induced by the auxiliary neurons do not contribute to the leading-order term, which simplifies the calculation.

To begin with, we calculate the first term. Employing Eq. (A16) and introducing an auxiliary neuron indexed by 0, it is enough to calculate $[\langle s_{t-d} x_{0,t} \rangle^2]$ as

$$\begin{aligned} [\langle s_{t-d} x_{0,t} \rangle^2] &= \left[\left\langle s_{t-d} \left(\sum_i^N J_{0i} \phi_{i,t-1} + u_0 s_t + \xi_{0,t} \right) \right\rangle^2 \right] = \sum_{i,j}^N [J_{0i} J_{0j} \langle s_{t-d} \phi_{i,t-1} \rangle^2] + [u_0^2] \langle s_{t-d} s_t \rangle^2 \\ &= \sum_{i,j}^N [J_{0i} J_{0j}] [\langle s_{t-d} \phi_{i,t-1} \rangle^2] + [u_0^2] \langle s_{t-d} s_t \rangle^2 = g^2 [\langle s_{t-(d-1)} \phi_{i,t} \rangle^2] + \delta_{d,0} \frac{\tilde{\sigma}_s^4}{N}, \end{aligned} \quad (\text{B7})$$

where we used the independence between $J_{0,i}$, $J_{0,j}$, and $\phi_{i,t}$ in the third equation. To proceed, we need to evaluate $[\langle s_{t-d} \phi(x_{i,t}) \rangle^2]$. Since $(s_{t-d}, x_{i,t})$ are correlated Gaussian variables, we can express them using auxiliary variables y and z :

$$\begin{aligned} s_{t-d} &= \sqrt{\sigma_s^2 - \frac{\langle s_{t-d} x_{0,t} \rangle^2}{\langle x_0^2 \rangle}} y + \frac{\langle s_{t-d} x_{0,t} \rangle}{\sqrt{\langle x_0^2 \rangle}} z, \\ x_{0,t} &= \sqrt{\langle x_0^2 \rangle} z. \end{aligned} \quad (\text{B8})$$

Using these expressions, we can calculate $\langle s_{t-d} \phi(x_{i,t}) \rangle$ as

$$\begin{aligned} \langle s_{t-d} \phi(x_{i,t}) \rangle &= \int Dy Dz \left(\sqrt{\sigma_s^2 - \frac{\langle s_{t-d} x_{0,t} \rangle^2}{\langle x_0^2 \rangle}} y + \frac{\langle s_{t-d} x_{0,t} \rangle}{\sqrt{\langle x_0^2 \rangle}} z \right) \phi \left(\sqrt{\langle x_0^2 \rangle} z \right) \\ &= \frac{\langle s_{t-d} x_{0,t} \rangle}{\sqrt{\langle x_0^2 \rangle}} \int Dz z \phi \left(\sqrt{\langle x_0^2 \rangle} z \right) \\ &= \langle s_{t-d} x_{0,t} \rangle \int Dz \phi' \left(\sqrt{\langle x_0^2 \rangle} z \right) \\ &= \langle s_{t-d} x_{0,t} \rangle \langle \phi'(x_0) \rangle, \end{aligned} \quad (\text{B9})$$

where Dy and Dz denote standard Gaussian measures and integration by parts is employed to derive the third line. Therefore, we obtain

$$[\langle s_{t-d} x_{0,t} \rangle^2] = g^2 [\langle \phi' \rangle^2] [\langle s_{t-(d-1)} x_{0,t} \rangle^2] + \delta_{d,0} \frac{\tilde{\sigma}_s^4}{N}, \quad (\text{B10})$$

where we restored $[\langle s_{t-(d-1)} x_{i,t} \rangle^2]$ to $[\langle s_{t-(d-1)} x_{0,t} \rangle^2]$. This equation is a recurrence formula for $[\langle s_{t-d} x_{0,t} \rangle^2]$, whose solution is

$$[\langle s_{t-d} x_{0,t} \rangle^2] = \frac{1}{N} \tilde{\sigma}_s^4 (g \langle \phi'(x) \rangle_*)^{2d}. \quad (\text{B11})$$

Consequently, we obtain

$$\text{first term of Eq. (B6)} = \frac{\alpha \tilde{\sigma}_s^2 (g \langle \phi'(x) \rangle_*)^{2d}}{[\langle x_i^2 \rangle]}. \quad (\text{B12})$$

Next, we move on to the calculation of the second term. We introduce the two auxiliary neurons indexed by 0 and 0', obtaining for $d \geq 1$,

$$\begin{aligned}
 [\langle s_{t-d}x_{0,t} \rangle \langle x_0x_{0'} \rangle \langle s_{t-d}x_{0',t} \rangle] &= \left[\sum_i^N J_{0i} \langle s_{t-d}\phi_{i,t-1} \rangle \sum_{j,k}^N J_{0j}J_{0'k} \langle \phi_j\phi_k \rangle \sum_l^N J_{0'l} \langle \phi_{l,t-1}s_{t-d} \rangle \right] \\
 &= \sum_{i,j,k,l}^N [J_{0i}J_{0j}J_{0'k}J_{0'l} \langle s_{t-d}\phi_{i,t-1} \rangle \langle \phi_j\phi_k \rangle \langle \phi_{l,t-1}s_{t-d} \rangle] \\
 &= \sum_{i,j,k,l}^N [J_{0i}J_{0j}J_{0'k}J_{0'l}][\langle s_{t-d}\phi_{i,t-1} \rangle \langle \phi_j\phi_k \rangle \langle \phi_{l,t-1}s_{t-d} \rangle] \\
 &= \frac{g^4}{N^2} \sum_{i,j}^N [\langle s_{t-d}\phi_{i,t-1} \rangle \langle \phi_i\phi_j \rangle \langle \phi_{j,t-1}s_{t-d} \rangle] \\
 &= \frac{g^4}{N} [\langle s_{t-(d-1)}\phi_{i,t} \rangle^2][\langle \phi_i^2 \rangle] + g^4 [\langle s_{t-(d-1)}\phi_{i,t} \rangle \langle \phi_i\phi_j \rangle \langle \phi_{j,t}s_{t-(d-1)} \rangle]_{(i \neq j)} \\
 &= \frac{g^4}{N} [\langle s_{t-(d-1)}\phi_{0,t} \rangle^2][\langle \phi_i^2 \rangle] + g^4 [\langle \phi'(x_i) \rangle^2]^2 [\langle s_{t-(d-1)}x_{0,t} \rangle \langle x_0x_{0'} \rangle \langle x_{0',t}s_{t-(d-1)} \rangle], \quad (\text{B13})
 \end{aligned}$$

where the last equality follows from Price's theorem. This equation is a recurrence formula for $[\langle s_{t-d}x_{0,t} \rangle \langle x_0x_{0'} \rangle \langle s_{t-d}x_{0',t} \rangle]$ with the initial condition ($d = 0$),

$$[\langle s_{t-d}x_{0,t} \rangle \langle x_0x_{0'} \rangle \langle s_{t-d}x_{0',t} \rangle] = \frac{\tilde{\sigma}_s^4}{N} [\langle x_0x_{0'} \rangle u_0 u_{0'}] = \frac{\tilde{\sigma}_s^4}{N} \left[u_0 u_{0'} \left(\sum_{i,j}^N J_{0i}J_{0'j} \langle \phi_i\phi_j \rangle + u_0 u_{0'} \langle s_t^2 \rangle \right) \right] = \tilde{\sigma}_s^6 N^{-3/2}. \quad (\text{B14})$$

Thus, the second term of Eq. (B13) for $d = 1$ is $O(N^{-3/2})$. On the other hand, from Eq. (B11), the first term of Eq. (B13) for $d = 1$ scales as $O(N^{-2})$. As a result, this term becomes subleading and negligible. Consequently, solving the recurrence formula yields

$$[\langle s_{t-d}x_{0,t} \rangle \langle x_0x_{0'} \rangle \langle s_{t-d}x_{0',t} \rangle] = \frac{\tilde{\sigma}_s^6 (g \langle \phi'(x) \rangle_*)^{4d}}{N^{3/2}} + \text{subleading terms} \quad (\text{B15})$$

and thus

$$\text{second term of Eq. (B6)} = -\frac{\alpha^2 \tilde{\sigma}_s^4 (g \langle \phi'(x) \rangle_*)^{4d}}{[\langle x_i^2 \rangle]^2} + \text{subleading terms}. \quad (\text{B16})$$

Subsequently, we proceed to calculate the third term in Eq. (B6). It is enough to consider the case for $i \neq k$ because the terms for $i = k$ have the same order as those for $i \neq k$, but the number of $i = k$ terms, $O(N^2)$, is much smaller than those of $i \neq k$ terms, $O(N^3)$, making their overall contribution subleading. We introduce three auxiliary neurons, labeled 0, 0', and 0'', obtaining for $d \geq 1$,

$$\begin{aligned}
 [\langle s_{t-d}x_{0,t} \rangle \langle x_0x_{0'} \rangle \langle x_{0'}x_{0''} \rangle \langle x_{0'',t}s_{t-d} \rangle] &= \left[\sum_i^N J_{0i} \langle s_{t-d}\phi_{i,t-1} \rangle \sum_{j,k}^N J_{0j}J_{0'k} \langle \phi_j\phi_k \rangle \sum_{l,m}^N J_{0'l}J_{0''m} \langle \phi_l\phi_m \rangle \sum_n^N J_{0''n} \langle \phi_{n,t-1}s_{t-d} \rangle \right] \\
 &= \frac{g^6}{N^3} \sum_{i,j,k}^N [\langle s_{t-d}\phi_{i,t-1} \rangle \langle \phi_i\phi_j \rangle \langle \phi_j\phi_k \rangle \langle \phi_{k,t-1}s_{t-d} \rangle] \\
 &= \frac{g^6}{N^2} [\langle \phi_i^2 \rangle]^2 [\langle s_{t-(d-1)}\phi_{i,t} \rangle^2] + \frac{2g^6}{N} [\langle \phi_i^2 \rangle][\langle s_{t-(d-1)}\phi_{0,t} \rangle \langle \phi_0\phi_{0'} \rangle \langle \phi_{0',t}s_{t-(d-1)} \rangle] \\
 &\quad + g^6 [\langle \phi_i^2 \rangle]^3 [\langle s_{t-(d-1)}x_{0,t} \rangle \langle \phi_0\phi_{0'} \rangle \langle x_{0'}x_{0''} \rangle \langle x_{0'',t}s_{t-(d-1)} \rangle], \quad (\text{B17})
 \end{aligned}$$

where we used Price's theorem in the last equality. This equation is a recurrence formula for $[\langle s_{t-d}x_{0,t} \rangle \langle x_0x_{0'} \rangle \langle x_{0'}x_{0''} \rangle \langle x_{0''},t s_{t-d} \rangle]$ with initial condition ($d = 0$),

$$\begin{aligned} [\langle s_t x_{0,t} \rangle \langle x_0 x_{0'} \rangle \langle x_{0'} x_{0''} \rangle \langle x_{0''},t s_t \rangle] &= \frac{\tilde{\sigma}^4}{N} [\langle x_0 x_{0'} \rangle \langle x_{0'} x_{0''} \rangle u_0 u_{0''}] \\ &= \frac{\tilde{\sigma}^4}{N} \left[u_0 u_{0''} \left(\sum_{i,j}^N J_{0i} J_{0'j} \langle \phi_i \phi_j \rangle + u_0 u_{0'} \langle (s_t)^2 \rangle \right) \left(\sum_{i,j}^N J_{0'i} J_{0''j} \langle \phi_i \phi_j \rangle + u_{0'} u_{0''} \langle (s_t)^2 \rangle \right) \right] \\ &= \tilde{\sigma}^8 N^{-2}. \end{aligned} \quad (\text{B18})$$

Therefore, the third term in Eq. (B17) for $d = 1$ is $O(1/N^2)$, while the first and second terms scale $O(1/N^3)$ and $O(1/N^{5/2})$, respectively. As a result, only third term contributes to the leading order. Solving the recurrence formula [Eq. (B17)] to leading order, we obtain

$$[\langle s_t x_{0,t} \rangle \langle x_0 x_{0'} \rangle \langle x_{0'} x_{0''} \rangle \langle x_{0''},t s_t \rangle] = \frac{\tilde{\sigma}_s^8 (g\langle \phi'(x) \rangle_*)^{6d}}{N^2} + \text{subleading terms} \quad (\text{B19})$$

and thus

$$\text{third term of Eq. (B6)} = \frac{\alpha^3 \tilde{\sigma}_s^6 (g\langle \phi'(x) \rangle_*)^{6d}}{[\langle x_i^2 \rangle]^3} + \text{subleading terms}. \quad (\text{B20})$$

The analogous calculation can be applied to the remaining terms, and we finally obtain the analytical solution for M_d and memory capacity:

$$M_d = \sum_{n=0}^{\infty} (-1)^n \left\{ \frac{\alpha \tilde{\sigma}_s^2}{[\langle x_i^2 \rangle]} (g\langle \phi'(x) \rangle_*)^{2d} \right\}^{n+1}, \quad \text{MC} = \sum_{d=0}^{\infty} \sum_{n=0}^{\infty} (-1)^n \left\{ \frac{\alpha \tilde{\sigma}_s^2}{[\langle x_i^2 \rangle]} (g\langle \phi'(x) \rangle_*)^{2d} \right\}^{n+1}. \quad (\text{B21})$$

2. Generalization of scaling assumptions

In this subsection, our aim is to identify broader scaling assumptions—beyond those specified in Eq. (4)—under which our theoretical framework for analytically deriving memory capacity remains applicable. To this end, we consider the following generalized scaling relations:

$$L = \alpha N^{c_1}, \quad \sigma_s^2 = \frac{\tilde{\sigma}_s^2}{N^{c_2}}, \quad \sigma_n^2 \sim O(1), \quad (\text{B22})$$

where $c_1, c_2 \geq 0$.

First, we require that the memory capacity remains nonvanishing in the limit $N \rightarrow \infty$. The first-order term in the Neumann series expansion of $[M_d]$ is given by

$$\frac{1}{\sigma_s^2} \sum_i^{\alpha N^{c_1}} \left[\frac{\langle s_{t-d} x_{i,t} \rangle^2}{\langle x_i^2 \rangle} \right] \sim O(N^{c_1 - c_2}). \quad (\text{B23})$$

To ensure that the memory capacity does not vanish as $N \rightarrow \infty$, it is necessary that $c_1 \geq c_2$.

Second, from the sufficient condition for the convergence of the Neumann series expansion, we obtain an additional constraint on c_1 and c_2 . Specifically, the Frobenius norm of the matrix $C_{\text{nondiag}} C_{\text{diag}}^{-1}$ must satisfy

$$\|C_{\text{nondiag}} C_{\text{diag}}^{-1}\|_F = \alpha^2 N^{2c_1} \frac{[\langle x_i x_j \rangle^2]_{(i \neq j)}}{[\langle x_i^2 \rangle]^2} \sim O(1). \quad (\text{B24})$$

The order of the neuronal correlations $[\langle x_i x_j \rangle^2]_{(i \neq j)}$ can be obtained using the same calculation as described in Appendix A 2, yielding $[\langle x_i x_j \rangle^2]_{(i \neq j)} \sim O(1/N^{\min(1, 2c_2)})$. Substituting this result into Eq. (B24), we obtain

$$\|C_{\text{nondiag}} C_{\text{diag}}^{-1}\|_F \sim O(N^{2c_1 - \min(1, 2c_2)}), \quad (\text{B25})$$

from which it follows that $2c_1 \leq \min(1, 2c_2)$ must be satisfied.

To summarize, the scaling exponents, c_1 and d_2 , must satisfy the following two conditions:

$$\begin{aligned} c_1 &\geq c_2, \\ 2c_1 &\leq \min(1, 2c_2), \end{aligned} \quad (\text{B26})$$

Combining these inequalities, we conclude that the allowed values are $0 \leq c_1 = c_2 \leq \frac{1}{2}$.

3. Analytical solution of the decay rate, $\lim_{N \rightarrow \infty} r(L)$

Noting that by applying Eq. (B21), the denominator in the definition of the decay rate [Eq. (12)] can be calculated as

$$\begin{aligned} \lim_{N \rightarrow \infty} L \times \text{MC}(1) &\stackrel{\text{Eq. (B21)}}{=} \lim_{N \rightarrow \infty} \alpha \sqrt{N} \times \sum_{d=0}^{\infty} \sum_{n=0}^{\infty} (-1)^n \left\{ \frac{1}{\sqrt{N}} \frac{\tilde{\sigma}_s^2}{[x_i^2]} (g\langle\phi'(x)\rangle_{x \sim \mathcal{N}(0, [x_i^2])})^{2d} \right\}^{n+1} \\ &= \frac{\alpha \tilde{\sigma}_s^2}{[x_i^2] (1 - g^2 \langle\phi'(x)\rangle_*^2)}, \end{aligned} \quad (\text{B27})$$

and the decay rate of memory capacity can be analytically derived as

$$\lim_{N \rightarrow \infty} r(L = \alpha \sqrt{N}) = \frac{[x_i^2] (1 - g^2 \langle\phi'(x)\rangle_*^2)}{\alpha \tilde{\sigma}_s^2} \sum_{d=0}^{\infty} \sum_{n=0}^{\infty} (-1)^n \left\{ \frac{\alpha \tilde{\sigma}_s^2}{[x_i^2]} (g\langle\phi'(x)\rangle_*)^{2d} \right\}^{n+1}. \quad (\text{B28})$$

The two infinite summations, $\sum_{d=0}^{\infty}$ and $\sum_{n=1}^{\infty}$, can be swapped under the condition [Eq. (B4)] because the infinite series is absolutely convergent:

$$\sum_{d=0}^{\infty} \sum_{n=0}^{\infty} \left\{ \frac{\alpha \tilde{\sigma}_s^2}{[x_i^2]} (g\langle\phi'(x)\rangle_*)^{2d} \right\}^{n+1} = \sum_{d=0}^{\infty} \frac{\frac{\alpha \tilde{\sigma}_s^2}{[x_i^2]} (g\langle\phi'(x)\rangle_*)^{2d}}{1 - \frac{\alpha \tilde{\sigma}_s^2}{[x_i^2]} (g\langle\phi'(x)\rangle_*)^{2d}} < \sum_{d=0}^{\infty} \frac{\alpha \tilde{\sigma}_s^2}{[x_i^2]} (g\langle\phi'(x)\rangle_*)^{2d} < \frac{\alpha \tilde{\sigma}_s^2}{[x_i^2]} \frac{1}{1 - (g\langle\phi'(x)\rangle_*)^2} < \infty. \quad (\text{B29})$$

Here, in the first inequality, we apply the result $0 < g\langle\phi'(x)\rangle_* < 1$ proven in Appendix A 3, and $\frac{\alpha \tilde{\sigma}_s^2}{[x_i^2]} < 1$ as shown below:

$$\frac{\alpha \tilde{\sigma}_s^2}{[x_i^2]} < \frac{\tilde{\sigma}_s^2}{[x_i^2]} \sqrt{\frac{[x_i^2]^2}{\tilde{\sigma}_s^4 + ([x_i^2] - \sigma_n^2)^2} (1 - (g\langle\phi'(x)\rangle_*)^4)} < \frac{\tilde{\sigma}_s^2}{[x_i^2]} \sqrt{\frac{[x_i^2]^2}{\tilde{\sigma}_s^4}} = 1, \quad (\text{B30})$$

where we used Eq. (B4) in the first inequality.

Consequently, by interchanging the two infinite summations, the decay rate can be represented as

$$\lim_{N \rightarrow \infty} r(L = \alpha \sqrt{N}) = 1 - \sum_{n=1}^{\infty} (-1)^{n-1} \left(\frac{\tilde{\sigma}_s^2}{[x_i^2]} \alpha \right)^n \frac{1 - (g\langle\phi'(x)\rangle_*)^2}{1 - (g\langle\phi'(x)\rangle_*)^{2n+2}}, \quad (\text{B31})$$

where the second term arises due to the neuronal correlations.

4. Decay rate for a linear RNN

We mention the decay rate of memory capacity for the reservoir RNN with a linear activation function, $\phi(x) = x$. In this case, the decay rate can be derived as

$$\begin{aligned} \lim_{N \rightarrow \infty} r(L = \alpha \sqrt{N}) \\ = 1 - \sum_{n=1}^{\infty} (-1)^{n-1} \left(\frac{\tilde{\sigma}_s^2}{\sigma_n^2} \alpha \right)^n \frac{(1 - g^2)^{n+1}}{1 - g^{2n+2}}, \end{aligned} \quad (\text{B32})$$

where the value of g must be less than 1 to ensure the stability of the reservoir RNN. Analogous to nonlinear RNNs, the memory capacity of the linear RNN with neuronal noise exhibits sublinear scaling with respect to L , as illustrated in Fig. 9. In contrast, for a linear RNN in the absence of noise, the memory capacity is known to be exactly L [12], resulting in $r(L) = 1$. It is intriguing that the presence or absence of noise greatly affects the scaling behavior of memory capacity with respect to L .

Here, we mention several points. First, the sublinear scaling of memory capacity for a linear RNN subject to neuronal noise is not accompanied by an increase in nonlinear computational capabilities, unlike in the case of a nonlinear RNN [Figs. 7(a) and 7(b)]. This is simply because a linear RNN lacks the ability to perform nonlinear computations. Consequently, in a linear RNN, the sublinear scaling of memory

capacity solely reflects a decline in overall computational performance caused by neuronal noise. Second, our theoretical analysis based on the Neumann series expansion is not applicable to the noise-free linear RNN. In this setting, the diagonal entries of covariance matrix C are comparable to its nondiagonal entries, both scaling as $O(1/\sqrt{N})$, which prevents us from employing the Neumann series expansion.

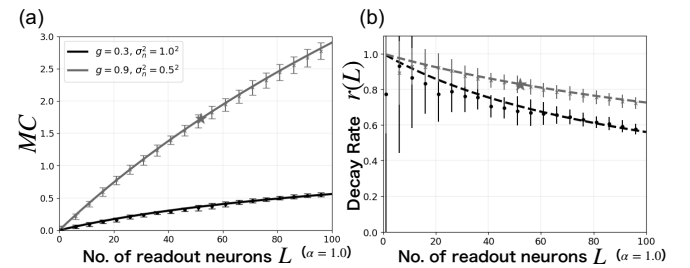


FIG. 9. (a) Memory capacity and (b) growth rate of memory capacity for linear RNNs. The figure's structure and content are the same as Fig. 2, with analytical and numerical results presented in the same format. The input intensity is $\sigma_s^2 = 1.0^2/\sqrt{N}$ ($\tilde{\sigma}_s^2 = 1.0^2$). In simulations, the network size is $N = 10\,000$ and simulation time length is $T = 10^4$.

APPENDIX C: INFORMATION PROCESSING CAPACITY THEORY

1. Overview of IPC theory

We provide a concise overview of the IPC, including its definition and key properties. We consider a reservoir receiving input signals $\{s(t)\}_t$. The general task is formulated as optimizing output weights such that the reservoir's output $z(t)$ approximates a given function of the inputs, $f[\dots, s(t-1), s(t)]$. For input signals drawn i.i.d from a standard Gaussian distribution, the orthonormal basis functions spanning the Hilbert space containing f are expressed as a formally infinite product of normalized Hermite polynomials:

$$y_d = \prod_{i \geq 0} \mathcal{H}_{d_i}(s(t-i)), \quad (\text{C1})$$

where $\mathcal{H}_{d_i}(\cdot)$ ($d_i \geq 0$) denotes the normalized Hermite polynomial of degree d_i and vector-form index $\mathbf{d} \equiv (d_i)_{i \geq 0}$ has only a finite number of nonzero elements. Since $\mathcal{H}_0 = 1$, the product in Eq. (C1) is effectively finite. The summation of d_i is equivalent to the degree of the polynomial y_d , defined as $\deg(y_d) \equiv \sum_{i \geq 0} d_i$.

The *capacity* for the reservoir to reconstruct the function of inputs, f , is defined as

$$C[f] \equiv 1 - \frac{\min_w \langle (z(t) - f(t))^2 \rangle_t}{\langle f(t)^2 \rangle_t}, \quad (\text{C2})$$

where $\langle \cdot \rangle_t$ denotes the time average. It is proven that $0 \leq C[f] \leq 1$ holds true for any function f . The *total* IPC is subsequently defined as the summation of capacities across all basis polynomials:

$$\text{IPC}_{\text{total}} \equiv \sum_{\mathbf{d}} C[y_{\mathbf{d}}]. \quad (\text{C3})$$

Notably, it has been established that $\text{IPC}_{\text{total}} = L$ holds, provided that the reservoir's activity is full rank and the reservoir satisfies the fading memory property [13], whereby the reservoir's state is uniquely determined solely by input signals, regardless of its initial state.

The total IPC can be decomposed based on the degree of the basis polynomials. We define the IPC for degree D as

$$\text{IPC}_D \equiv \sum_{\substack{\mathbf{d} \\ \text{s.t. } \deg(y_{\mathbf{d}}) = D}} C[y_{\mathbf{d}}]. \quad (\text{C4})$$

Crucially, IPC_1 is identical to memory capacity, as $C[y_{\mathbf{d}}]$ for $\deg(y_{\mathbf{d}}) = 1$ corresponds to Eq. (3). In contrast, IPC_D for $D \geq 2$ represents a nonlinear computational ability of the reservoir.

When calculating the capacities by numerical simulation, we must take care not to overestimate them, because for a finite simulation time T , they are subject to a systematic positive error. Following Dambre *et al.* [13], we determine the

threshold of the capacities, ϵ , as

$$\epsilon \equiv \frac{2\theta}{T}, \quad \theta \equiv \arg_{\theta} \{\mathbb{P}[\chi^2(L) \geq \theta] = p\}, \quad (\text{C5})$$

where $\chi^2(L)$ denotes a random variable that follows a chi-squared distribution with L degrees of freedom. The value of p represents the probability that a truly zero capacity is incorrectly assessed as nonzero. We consider a capacity to be zero if it falls below this threshold ϵ . Consistent with Dambre *et al.*, we set $p = 10^{-4}$ for all our numerical simulations of memory capacity ($= \text{IPC}_1$) and IPC, although the choice of p has a negligible impact on our results.

Note that calculating the IPC values of our reservoir RNN model requires a minor correction to the time evolution equation, Eq. (1), as

$$x_i(t) = \sum_{j=1}^N J_{ij} \phi(x_j(t-1)) + u_i \sigma_s s(t) + \xi_i(t),$$

$$s(t) \sim \mathcal{N}(0, 1), \quad (\text{C6})$$

where the standard deviation of the input signals, σ_s , is now incorporated as a scaling factor of the standard Gaussian noise inputs. This correction is necessary because IPC is defined for a reservoir receiving standard Gaussian noise inputs. Importantly, this modification does not alter the statistical behavior or computational ability of the original RNN model.

2. Linear scaling of IPC_D with L for vanishing neuronal correlations

In the main text, we claim that each IPC_D for $D \geq 1$ exhibits linear scaling with regard to L if we ignore neuronal correlations. The following provides a concise proof of this claim. First, similarly to Eq. (5), the capacity for a basis function $y_{\mathbf{d}}$ can be expressed as

$$C_T[y_{\mathbf{d}}] = \frac{\mathbf{a}^\top \mathbf{C}^{-1} \mathbf{a}}{\langle y_{\mathbf{d}}(t)^2 \rangle}, \quad (\text{C7})$$

where the elements of $\mathbf{a} \in \mathbb{R}^L$ are $a_i = \langle y_{\mathbf{d}}(t) x_i(t) \rangle$ and the matrix $\mathbf{C} \in \mathbb{R}^{L \times L}$ is the covariance matrix of the readout neurons. Assuming neuronal correlations vanish, the capacity for any $y_{\mathbf{d}}$ is proportional to L since

$$C_T[y_{\mathbf{d}}] \approx [C_T[y_{\mathbf{d}}]] \approx \left[\sum_i^L \frac{\langle y_{\mathbf{d}}(t) x_i(t) \rangle^2}{\langle x_i(t)^2 \rangle \langle y_{\mathbf{d}}(t)^2 \rangle} \right]$$

$$= L \left[\frac{\langle y_{\mathbf{d}}(t) x_i(t) \rangle^2}{\langle x_i(t)^2 \rangle \langle y_{\mathbf{d}}(t)^2 \rangle} \right] \propto L, \quad (\text{C8})$$

where we assume the self-averaging of the capacity in the first approximation equation, and in the second approximation equation, we ignore the correlation $\langle x_i x_j \rangle$.

- [1] I. D. Mienye and T. G. Swart, A comprehensive review of deep learning: Architectures, recent advances, and applications, *Information* **15**, 755 (2024).
- [2] S. S. Sengar, A. B. Hasan, S. Kumar, and F. Carroll, Generative artificial intelligence: A systematic review and applications, *Multimedia Tools Appl.* **84**, 23661 (2025).

- [3] G. Cybenko, Approximation by superpositions of a sigmoidal function, *Math. Control Signal Syst.* **2**, 303 (1989).
- [4] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, On the expressive power of deep neural networks, in *Proceedings of the 34th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research*, edited

- by D. Precup and Y. W. Teh (PMLR, Cambridge, MA, 2017), Vol. 70, pp. 2847–2854.
- [5] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, The computational limits of deep learning, [arXiv:2007.05558](#).
 - [6] A. Marchisio, M. A. Hanif, F. Khalid, G. Plastiras, C. Kyrkou, T. Theodorides, and M. Shafique, Deep learning for edge computing: Current trends, cross-layer optimizations, and open research challenges, in *2019 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)* (IEEE Computer Society, New York, 2019), pp. 553–559.
 - [7] H. Jaeger, The “echo state” approach to analysing and training recurrent neural networks, Technical Report, GMD Report 148, German National Research Center for Information Technology, 2001.
 - [8] W. Maass, T. Natschlager, and H. Markram, Real-time computing without stable states: A new framework for neural computation based on perturbations, *Neural Comput.* **14**, 2531 (2002).
 - [9] M. Yan, C. Huang, P. Bienstman, P. Tino, W. Lin, and J. Sun, Emerging opportunities and challenges for the future of reservoir computing, *Nat. Commun.* **15**, 2056 (2024).
 - [10] K. Nakajima and I. Fischer, *Reservoir Computing: Theory, Physical Implementations, and Applications* (Springer, Berlin, 2021).
 - [11] M. Cucchi, S. Abreu, G. Ciccone, D. Brunner, and H. Kleemann, Hands-on reservoir computing: A tutorial for practical implementation, *Neuromorph. Comput. Eng.* **2**, 032002 (2022).
 - [12] H. Jaeger, Short term memory in echo state networks, Technical Report, GMD Report 152, German National Research Center for Information Technology, 2002.
 - [13] J. Dambre, D. Verstraeten, B. Schrauwen, and S. Massar, Information processing capacity of dynamical systems, *Sci. Rep.* **2**, 514 (2012).
 - [14] J. Schuecker, S. Goedeke, and M. Helias, Optimal sequence memory in driven random networks, *Phys. Rev. X* **8**, 041029 (2018).
 - [15] T. Haruna and K. Nakajima, Optimal short-term memory before the edge of chaos in driven random recurrent networks, *Phys. Rev. E* **100**, 062312 (2019).
 - [16] M. Lukoševičius, H. Jaeger, and B. Schrauwen, Reservoir computing trends, *Künstliche Intell.* **26**, 365 (2012).
 - [17] D. Sussilo and L. F. Abbott, Generating coherent patterns of activity from chaotic neural networks, *Neuron* **63**, 544 (2009).
 - [18] H. Sompolinsky, A. Crisanti, and H. J. Sommers, Chaos in random neural networks, *Phys. Rev. Lett.* **61**, 259 (1988).
 - [19] L. Molgedey, J. Schuchhardt, and H. G. Schuster, Suppressing chaos in neural network by noise, *Phys. Rev. Lett.* **69**, 3717 (1992).
 - [20] R. Engelken, A. Ingrosso, R. Khajeh, S. Goedeke, and L. F. Abbott, Input correlations impede suppression of chaos and learning in balanced firing-rate networks, *PLoS Comput. Biol.* **18**, e1010590 (2022).
 - [21] S. Takasu and T. Aoyagi, Suppression of chaos in a partially driven recurrent neural network, *Phys. Rev. Res.* **6**, 013172 (2024).
 - [22] M. Massar and M. Serge, Mean-field theory of echo state networks, *Phys. Rev. E* **87**, 042809 (2013).
 - [23] K. Krishnamurthy, T. Can, and D. J. Schwab, Theory of gating in recurrent neural networks, *Phys. Rev. X* **12**, 011011 (2022).
 - [24] T. Toyozumi and L. F. Abbott, Beyond the edge of chaos: Amplification and temporal integration by recurrent networks in the chaotic regime, *Phys. Rev. E* **84**, 051908 (2019).
 - [25] M. Helias and D. Dahmen, *Statistical Field Theory for Neural Networks* (Springer, Berlin, 2020).
 - [26] D. G. Clark, L. F. Abbott, and A. Litwin-Kumar, Dimension of activity in random neural networks, *Phys. Rev. Lett.* **131**, 118401 (2023).
 - [27] M. Inubushi and K. Yoshimura, Reservoir computing beyond memory-nonlinearity trade-off, *Sci. Rep.* **7**, 10199 (2017).
 - [28] G. Buzsáki and K. Mizuseki, The log-dynamic brain: How skewed distributions affect network operations, *Nat. Rev. Neurosci.* **15**, 264 (2014).
 - [29] J.-n. Teramae, Y. Tsubo, and T. Fukai, Optimal spike-based communication in excitable networks with strong-sparse and weak-dense links, *Sci. Rep.* **2**, 485 (2012).
 - [30] Ł. Kuśmierz, S. Ogawa, and T. Toyozumi, Edge of chaos and avalanches in neural networks with heavy-tailed synaptic weight distribution, *Phys. Rev. Lett.* **125**, 028101 (2020).
 - [31] A. Wardak and P. Gong, Extended Anderson criticality in heavy-tailed neural networks, *Phys. Rev. Lett.* **129**, 048103 (2022).
 - [32] Y. Xie, S. Mihalas, and Łukasz Kuśmierz, Slow transition to low-dimensional chaos in heavy-tailed recurrent neural networks, [arXiv:2505.09816](#).
 - [33] Y. Shao, D. Dahmen, S. Recanatani, E. Shea-Brown, and S. Ostojic, Impact of local connectivity patterns on excitatory-inhibitory network dynamics, *PRX Life* **3**, 023008 (2025).
 - [34] D. Martí, N. Brunel, and S. Ostojic, Correlations between synapses in pairs of neurons slow down dynamics in randomly connected neural networks, *Phys. Rev. E* **97**, 062314 (2018).
 - [35] H. J. Sommers, A. Crisanti, H. Sompolinsky, and Y. Stein, Spectrum of large random asymmetric matrices, *Phys. Rev. Lett.* **60**, 1895 (1988).
 - [36] B. Cramer, D. Stöckel, M. Kreft, M. Wibral, J. Schemmel, K. Meier, and V. Priesemann, Control of criticality and computation in spiking neuromorphic networks with plasticity, *Nat. Commun.* **11**, 2853 (2020).
 - [37] A. Pikovsky and A. Politi, *Lyapunov Exponents: A Tool to Explore Complex Dynamics* (Cambridge University Press, Cambridge, UK, 2016).
 - [38] T. Hülser, F. Köster, K. Lüdge, and L. Jaurigue, Deriving task specific performance from the information processing capacity of a reservoir computer, *Nanophotonics* **12**, 937 (2023).
 - [39] N. Bertschinger and T. Natschlager, Real-time computation at the edge of chaos in recurrent neural networks, *Neural Comput.* **16**, 1413 (2004).
 - [40] R. Legenstein and W. Maass, Edge of chaos and prediction of computational performance for neural circuit models, *Neural Networks* **20**, 323 (2007).
 - [41] T. L. Carroll, Do reservoir computers work best at the edge of chaos?, *Chaos* **30**, 121109 (2020).
 - [42] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities., *Proc. Natl. Acad. Sci. USA* **79**, 2554 (1982).
 - [43] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Storing infinite numbers of patterns in a spin-glass model of neural networks, *Phys. Rev. Lett.* **55**, 1530 (1985).
 - [44] S. Amari, Characteristics of sparsely encoded associative memory, *Neural Networks* **2**, 451 (1989).

- [45] M. Okada, Notions of associative memory and sparse coding, *Neural Networks* **9**, 1429 (1996).
- [46] W. E. Vinje and J. L. Gallant, Sparse coding and decorrelation in primary visual cortex during natural vision, *Science* **287**, 1273 (2000).
- [47] C. van Vreeswijk and H. Sompolinsky, Chaos in neuronal networks with balanced excitatory and inhibitory activity, *Science* **274**, 1724 (1996).
- [48] A. Renart, J. de la Rocha, P. Bartho, L. Hollender, N. Parga, A. Reyes, and K. D. Harris, The asynchronous state in cortical circuits, *Science* **327**, 587 (2010).
- [49] D. Dahmen, S. Grün, M. Diesmann, and M. Helias, Second type of criticality in the brain uncovers rich multiple-neuron dynamics, *Proc. Natl. Acad. Sci. USA* **116**, 13051 (2019).
- [50] R. Rosenbaum, M. A. Smith, A. Kohn, J. E. Rubin, and B. Doiron, The spatial structure of correlated neuronal variability, *Nat. Neurosci.* **20**, 107 (2017).
- [51] S. Safavi, A. Dwarakanath, V. Kapoor, J. Werner, N. G. Hatsopoulos, N. K. Logothetis, and T. I. Panagiotaropoulos, Nonmonotonic spatial structure of interneuronal correlations in prefrontal microcircuits, *Proc. Natl. Acad. Sci. USA* **115**, E3539 (2018).
- [52] M. M. Churchland *et al.*, Stimulus onset quenches neural variability: A widespread cortical phenomenon, *Nat. Neurosci.* **13**, 369 (2010).
- [53] A. Y. Y. Tan, Y. Chen, B. Scholl, E. Seidemann, and N. J. Priebe, Sensory stimulation shifts visual cortex from synchronous to asynchronous states, *Nature (London)* **509**, 226 (2014).
- [54] M. R. Cohen and J. H. R. Maunsell, Attention improves performance primarily by reducing interneuronal correlations, *Nat. Neurosci.* **12**, 1594 (2009).
- [55] J. F. Mitchell, K. A. Sundberg, and J. H. Reynolds, Spatial attention decorrelates intrinsic activity fluctuations in macaque area v4, *Neuron* **63**, 879 (2009).
- [56] S. Fusi, E. K. Miller, and M. Rigotti, Why neurons mix: High dimensionality for higher cognition, *Curr. Opin. Neurobiol.* **37**, 66 (2016).
- [57] C. Stringer, M. Pachitariu, N. Steinmetz, M. Carandini, and K. D. Harris, High-dimensional geometry of population responses in visual cortex, *Nature (London)* **571**, 361 (2019).
- [58] M. Farrell, S. Recanatesi, T. Moore, G. Lajoie, and E. Shea-Brown, Gradient-based learning drives robust representations in recurrent neural networks by balancing compression and expansion, *Nat. Mach. Intell.* **4**, 564 (2022).
- [59] K. Rajan, L. Abbott, and H. Sompolinsky, Inferring stimulus selectivity from the spatial structure of neural network dynamics, in *Advances in Neural Information Processing Systems*, edited by J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (Curran Associates, Inc., Red Hook, NY, 2010), Vol. 23.
- [60] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2018), Vol. 31.
- [61] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, Wide neural networks of any depth evolve as linear models under gradient descent, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2019), Vol. 32.
- [62] G. Yang and E. J. Hu, Tensor programs IV: Feature learning in infinite-width neural networks, in Proceedings of the 38th International Conference on Machine Learning, in *Proceedings of Machine Learning Research*, edited by M. Meila and T. Zhang (PMLR, Cambridge, MA, 2021), Vol. 139, pp. 11727–11737.
- [63] B. Bodelon and C. Pehlevan, Self-consistent dynamical field theory of kernel evolution in wide neural networks, in *Advances in Neural Information Processing Systems*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc., Red Hook, NY, 2022), Vol. 35, pp. 32240–32256.
- [64] K. Fischer, J. Lindner, D. Dahmen, Z. Ringel, M. Krämer, and M. Helias, Critical feature learning in deep neural networks, in Proceedings of the 41st International Conference on Machine Learning, *Proceedings of Machine Learning Research*, edited by R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp (PMLR, Cambridge, MA, 2024), Vol. 235, pp. 13660–13690.
- [65] S. Takasu, Source code for figures, GitHub, 2025, <https://github.com/shtakasu/scalingofrc.git>.
- [66] W. Zou and H. Huang, Introduction to dynamical mean-field theory of randomly connected neural networks with bidirectionally correlated couplings, *SciPost Phys. Lect. Notes* **79** (2024).
- [67] D. G. Clark, O. Marschall, A. van Meegen, and A. Litwin-Kumar, Connectivity structure and dynamics of nonlinear recurrent neural networks, [arXiv:2409.01969](https://arxiv.org/abs/2409.01969).
- [68] M. Dick, A. van Meegen, and M. Helias, Linking network- and neuron-level correlations by renormalized field theory, *Phys. Rev. Res.* **6**, 033264 (2024).