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

Shotaro Takasu* and Toshio Aoyagi
Graduate School of Informatics, Kyoto University,
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan

(Dated: March 20, 2025)

Reservoir computing is a machine learning framework characterized by its high computational ability and quick learning, making it well-suited for a wide range of applications for real-time computing. In this paper, we demonstrate that memory capacity of a reservoir recurrent neural network scales sublinearly with the number of readout neurons. To elucidate this phenomenon, we develop a theoretical framework for analytically deriving memory capacity, allowing us to attribute the decaying growth of memory capacity to neuronal correlations. In addition, numerical simulations reveal that the computational abilities required to solve increasingly complex nonlinear tasks emerge sequentially as the number of readout neurons increases. Furthermore, our theoretical framework suggests that the incremental increase of nonlinear computational capabilities is influenced by neuronal correlations in a manner similar to memory capacity. Our findings establish a foundation for designing scalable and cost-effective reservoir computing, providing novel insights into the interplay between neuronal correlations, linear memory, and nonlinear processing.

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 [1, 2]. In contrast to conventional training methods such as backpropagation through time, RC optimizes only read-out weights and leaves the remaining weights fixed, which enables quick and low-cost learning. In spite of its training simplicity, RC has high computational performance and a broad range of applications for processing time-series data [3]. RC is not restricted to RNNs, and a wide variety of dynamical systems can be utilized as reservoirs under appropriate conditions. Specifically, 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 [4].

In a typical RC setting, the readout connections linking reservoir units to output units are sparse [4]. 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. Empirically, the performance of RC improves as an increase in L. However, 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. In particular, performance in RC depends on both the reservoir's memory ability [5] and nonlinear processing capability [6], and the number of readout units must be carefully chosen to balance these two aspects according to the demands of a given task. Despite its practical importance, a systematic framework for determining an appropriate L remains largely unexplored. In this study, we aim to address this gap by elucidating how memory ability and nonlinear computational ability scale with L, with the particular focus on the role of neuronal correlations.

Here, we demonstrate for the first time that the the memory ability of an RNN increases monotonically with L, though its growth rate gradually declines (Fig.1). This *sublinear* scaling of memory ability cannot be explained by the previous theoretical works assuming a scaling regime where L is of order one relative to N, i.e. $L \sim O(1)$ because the memory ability in this regime is proportional to L [7, 8]. To bridge the gap between the observation and theory, we develop a novel theory for analytically deriving memory ability for $L \sim O(\sqrt{N})$. Using our theory, we demonstrate that the correlation between reservoir neurons plays an important role in the declining growth rate of memory ability, although the correlation diminishes as $O(1/\sqrt{N})$. Furthermore, numerical simulations show that the ability to perform more complex nonlinear tasks emerges sequentially as L increases. These findings indicate that the number of readout connections should be tailored to the specific memory and nonlinearity requirements of the task at hand.

We investigate computational capacity of RC using a large random RNN as a reservoir (Fig.1(a)), which is known as *Echo state network* (ESN) [1], one of the canonical model for RC [9, 10]. The state of ith neuron at discrete time t is described by the variable, $x_i(t)$. The activation function ϕ is assumed to be an odd saturated sigmoid function satisfying $\phi'(0) = 1$, $\phi'(x) > 0$, $\phi''(x) \leq 0$ $(x \geq 0)$ and $\phi(\pm\infty) = \pm 1$. The time evolution of a reservoir RNN is determined by the difference equation,

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

where $N \gg 1$ indicates the total number of neurons. The recurrent weights, J_{ij}, and the input weights, u_i,

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

are sampled 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.

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) [11]. 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 [11, 12]. In the presence of inputs ($\sigma_s^2 > 0$ or $\sigma_n^2 > 0$), chaotic variability is suppressed [7, 8, 12–15].

The output of the reservoir RNN is defined as $\hat{z}(t) = \sum_{i=1}^{L} w_i x_i(t)$, 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, $\{\hat{z}(t)\}_{t=1}^{T}$, and the desired signals, $\{z(t)\}_{t=1}^{T}$.

Memory capacity [5, 6] 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

$$M_d \equiv 1 - \frac{\min_{\boldsymbol{w}} \langle (\hat{z}(t) - s(t-d))^2 \rangle}{\langle s(t)^2 \rangle}, \tag{2}$$

where the angular bracket denotes time averaging. 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 [5, 6]. 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. $0 \leq MC \leq L$ [5, 6]. 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 [5].

Fig.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 exhibits monotonic growth, but its growth rate gradually diminishes as L increases. Consequently, memory capacity can be characterized as a sublinear function of L.

To elucidate the mechanism behind the decaying growth rate of memory capacity, we provide a novel 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:

$$L = \alpha\sqrt{N}, \quad \sigma_s^2 = \frac{\tilde{\sigma}_s^2}{\sqrt{N}}, \quad \sigma_n^2 \sim O(1), \tag{3}$$

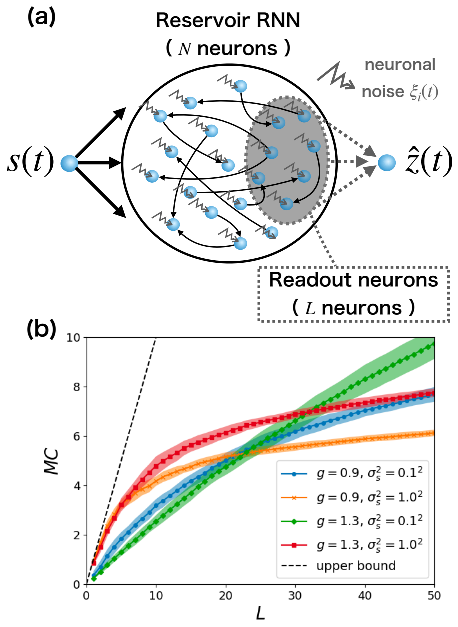

FIG. 1. (a) Overview of RC utilizing a random RNN receiving input signals and inherent neuronal noise. The shaded region represents the readout neurons. (b) Numerical simulations for memory capacity of the reservoir RNNs with network size $N = 1000$, noise intensity $\sigma_n^2 = 0.1^2$, and activation function $\phi(x) = \tanh(x)$. Shaded area represents mean±std of direct numerical simulations for 10 different network and input signal realizations. The sum of M_d is calculated up to $d = 1000$. The dashed line indicates the theoretical upper bound of memory capacity, $MC = L$. Simulation time length is $T = 10^4$.

where both α and $\tilde{\sigma}_s^2$ are $O(1)$. The parameter α represents the relative magnitude of L compared to $\sqrt{N}$. In contrast to previous models [7, 8, 16], where $L \sim O(1)$, our model employs a significantly larger number of readout neurons, although the output connections remain sparse, as $L = \alpha\sqrt{N} \ll N$

By simple calculation, M_d defined in Eq.(2) can be expressed as

$$M_d = \frac{\boldsymbol{a}_d^\top C^{-1} \boldsymbol{a}_d}{\langle s(t)^2 \rangle}, \tag{4}$$

where the elements of $\boldsymbol{a}_d \in \mathbb{R}^L$ and $C \in \mathbb{R}^{L \times L}$ are respectively $(\boldsymbol{a}_d)_i \equiv \langle s(t-d) x_i(t) \rangle$ and $C_{ij} \equiv \langle x_i(t) x_j(t) \rangle$ (i, j indicate the indices of the readout neurons) [6]. 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 [7, 8, 16]. However, this approximation results in linear scaling of memory capacity, $MC \propto L$, conflicting with our numerical simulations as shown in Fig.1(b).

We noticed that from the scale assumption Eq.(3), the diagonal entries of C are $O(1)$, whereas the non-diagonal ones are $O(1/\sqrt{N})$ (See Appendix.A). Hence, the non-diagonal elements are much smaller than the diagonal ones, which motivates us to perform Neumann series expansion,

$$C^{-1} = (D + \tilde{C})^{-1} = \sum_{n=0}^{\infty} D^{-1} \left(-\tilde{C} D^{-1} \right)^{n}, \qquad (5)$$

where $D_{ij} \equiv \delta_{ij} \langle x_i(t)^2 \rangle$ and $\tilde{C}_{ij} \equiv (1 - \delta_{ij}) \langle x_i(t) x_j(t) \rangle$. 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; $\lim_{N\to\infty} M_d = \lim_{N\to\infty} [M_d]$, where the square bracket denotes the average over network realizations, known as *quenched average* [17]. Consequently, substituting Eq.(5) into Eq.(4) and taking quenched average of M_d, we analytically obtain memory capacity:

$$\begin{aligned} \lim_{N\to\infty} MC(L = \alpha\sqrt{N}) &= \lim_{N\to\infty} \sum_{d=0}^{\infty} \left[M_d(L = \alpha\sqrt{N}) \right] \qquad (6) \\ &= \sum_{d=0}^{\infty} \sum_{n=0}^{\infty} (-1)^n \left\{ \frac{\alpha \tilde{\sigma}_s^2}{[\langle x_i^2 \rangle]} \left(g \langle \phi'(x) \rangle_{x\sim\mathcal{N}(0,[\langle x_i^2 \rangle])} \right)^{2d} \right\}^{n+1} \end{aligned}$$

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

$$[\langle x_i^2 \rangle] = \sigma_n^2 + g^2 \langle \phi(x)^2 \rangle_{x\sim\mathcal{N}(0,[\langle x_i^2 \rangle])}. \qquad (7)$$

In rigorously calculating quenched average, we employ the dynamical cavity method, first introduced by Clark et.al [18] from statistical physics to the analysis of neural networks. A detailed derivation is provided in Appendix.B 1. In the following, we use the shorthand notation, $\langle f(x) \rangle_* \equiv \langle f(x) \rangle_{x\sim\mathcal{N}(0,[\langle x_i^2 \rangle])}$. Specifically, 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 $\langle \phi(x)^2 \rangle_*$ (see Appendix.B 1).

We have confirmed that the analytical values of MC are consistent with the numerical simulations, exhibiting a sublinear function of L, for L below a threshold (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 is because Neumann series expansion, Eq.(5), does not converge as the spectrum norm $\|\tilde{C} D^{-1}\|$ exceeds one for large L. To evaluate the upper bound of L ensuring convergence of the series expansion, it is enough to evaluate $\|\tilde{C} D^{-1}\|$, but this is a challenging task. Instead, we derive the sufficient condition for the expansion series to converge (see Appendix.B 1), and the upper values of L for this sufficient condition are indicated by star marks in Fig.2.

Having obtained the analytical solution of memory capacity, we move on to elucidate the reason why memory capacity grows sublinearly with respect to L. To quantify the extent to which the growth rate of memory capacity decays, we define a *decay rate*,

$$r(L) \equiv \frac{MC(L)}{L \times MC(1)}. \qquad (8)$$

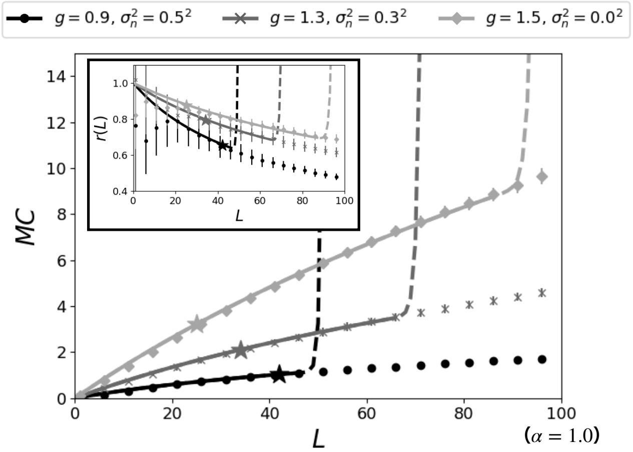

FIG. 2. Memory capacity as a function of L. An inset shows corresponding decay rate $r(L)$. Solid curves represent the analytical solutions, while error bars indicate mean±std for 10 networks and inputs realizations obtained through numerical simulations. Star marks represent the upper bounds of the sufficient conditions for the convergence of the expansion series. Dashed lines indicate explosion of theoretical values. The activation function is an error function, $\phi(x) = \int_0^x e^{-\frac{\pi}{4}t^2} dt$. The input intensity is $\sigma_s^2 = 1.0^2/\sqrt{N}$ ($\tilde{\sigma}_s^2 = 1.0^2$). The network size is $N = 10^4$, and simulation time length is $T = 10^4$. Note that for these parameter values, the input intensity is smaller than that in Fig.1 due to scaling assumptions. In numerical simulations, M_d values below a cutoff threshold ϵ determined by Eq.(C5) are set to zero to mitigate systematic positive errors arising from finite simulation time.

For $MC(L)$ scaling linearly with L, $r(L) = 1$ holds true for all L. In contrast, $r(L)$ decays from one when $MC(L)$ is a sublinear function of L. As depicted in Fig.2, $r(L)$ decays for various parameter combinations, confirming the sublinear scaling of memory capacity with respect to L.

Employing the solution of memory capacity (Eq.(6)), we obtain an analytical form of the decay rate (see Appendix.B 2 for derivation):

$$\begin{aligned} \lim_{N\to\infty} r(L = \alpha\sqrt{N}) = 1 - \sum_{n=1}^{\infty} (-1)^{n-1} \left(\frac{\tilde{\sigma}_s^2}{[\langle x_i^2 \rangle]} \alpha \right)^{n} \qquad (9) \\ \times \frac{1 - (g \langle \phi'(x) \rangle_*)^2}{1 - (g \langle \phi'(x) \rangle_*)^{2n+2}}. \end{aligned}$$

It is noteworthy that the n-th term corresponds exactly to the n-th term in the Neumann series expansion, Eq.(5). Therefore, under the assumption of vanishing neuronal correlations ($\tilde{C} = O$), the summation terms in Eq.(9) vanish, resulting in $r(L) = 1$. This implies that higher neuronal correlations contribute to the faster decay of $r(L)$ through the contributions of the summation terms in Eq.(9). In fact, we have confirmed that $r(L)$ rapidly decays for hyperparameters that lead to high neuronal correlation, such as large input signals ($\tilde{\sigma}_s^2$), small

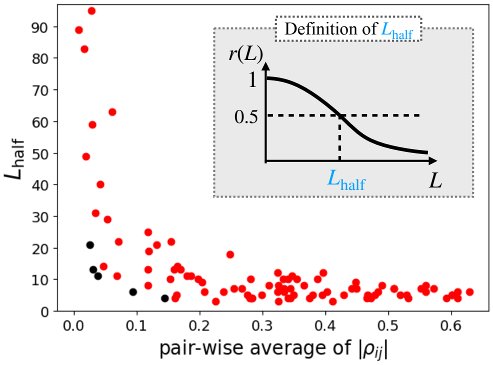

FIG. 3. Relation between the level of neuronal correlations and the half-life of memory capacity growth. Each data point corresponds to a set of hyperparameters sampled from $g \sim U(0.2, 2)$, $\sigma_s \sim U(0.1, 3)$ and $\sigma_n \sim U(0, 3)$ (100 samples are plotted). Red points indicate parameter regimes beyond our theoretical framework, as determined by L_{half} surpassing the deviation threshold of the theoretical values of MC (Fig.2). For each parameter set, the half-life of memory capacity growth, L_{half}, and pair-wise averaged absolute values of Pearson correlation coefficients of neuronal activity, ρ_{ij}, are calculated across 10 network and input realizations. The mean values obtained from these 10 realizations are displayed as a single plot in the figure. The activation function is a hyperbolic tangent function, $\phi(x) = \tanh(x)$. The network size is $N = 1000$, and the simulation time length is $T = 10^4$.

recurrent weights (g), and low noise intensity (σ_n^2), as depicted in Fig.6.

In the preceding sections, we derived the memory capacity under specific scaling assumptions (Eq.(3)). However, practical applications often deviate from these conditions, raising the question of whether neuronal correlations remain a key determinant of sublinear memory capacity growth in more general settings. To address this, Fig.3 explores a broader range of model parameters including larger input intensities and 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.3, inset). A lower L_{half} indicates more rapid decay of $r(L)$. Notably, as shown in Fig.3, strong neuronal correlations are closely associated with a faster decay of memory capacity growth, highlighting their persistent influence even when the original scaling assumptions are relaxed.

While our previous analysis has focused on memory capacity, non-linear 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 [6, 19]. Thus, we expect that nonlinear computational ability increases instead of a deceleration in the growth rate of memory capacity.

Information processing capacity (IPC) introduced by Dambre et al. [6] offers a task-independent metric for evaluating the performance of RC, enabling a comprehensive assessment of both linear and non-linear 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 Gaussian 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 non-linear computational ability. Second, it has been proven [6] that the *total IPC*, $IPC_{\text{total}} \equiv \sum_{D=1}^{\infty} IPC_D$, equals to the number of readout units, L, provided that the reservoir satisfies echo state property [1], which ensures that the reservoir's state is uniquely determined solely by input signals regardless of reservoir's initial state.

Fig.4 illustrates how the values of IPC_D for our RNN model (Eq.(1)) obtained numerically change as L increases across various hyperparameter combinations (g, σ_s, and σ_n). Note that, due to the symmetry of the model, IPC_D values for even-degree 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 non-linear computational ability represented by IPC_D for $D \geq 2$, reflecting memory-nonlinearity trade-off.

Our novel finding is the sequential emergence of higher-degree IPC with L, as well as the supralinear rise in nonlinear computational ability. This phenomenon is observed not only in an RNN satisfying the echo state property (Fig.4(a)), but also in a chaotic RNN (Fig.4(b)) and in an RNN subject to neuronal noise (Fig.4(c)), where the echo state property breaks down. Additionally, the same finding is observed for an RNN employing the ReLU activation function, although our theoretical framework for deriving memory capacity is not applicable to the ReLU function (see Appendix.C 2). As analytical derivation of higher-degree IPC 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 (see Appendix.C 3).

In the present study, we have investigated the relationship between the number of readout neurons and the computational capacity for large random RNN reservoirs. Through analytical and numerical approaches, we have demonstrated that the memory capacity grows sublinearly with L, and that nonlinear computational capabilities emerge incrementally with L, enabling the reservoir to handle increasingly complex tasks. These findings indicate that the number of readout connections should be tailored to the specific memory and nonlinearity requirements of the task at hand. The specific values of IPC required for processing a given task can be evaluated using framework proposed by [21]. By leveraging this frame-

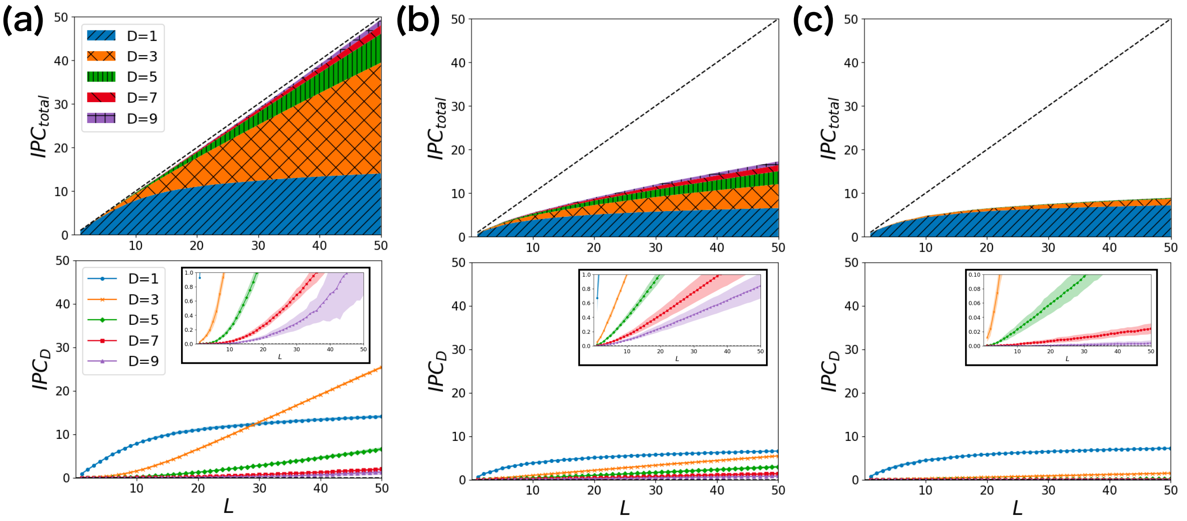

FIG. 4. IPC values for varying the number of readout neurons L of a reservoir RNN with activation function $\phi(x) = \tanh(x)$ and network size $N = 1000$. The top row panels show stacked representations of IPC values for each degree (IPC_D), while the bottom row panels provides a detailed breakdown of individual IPC_D values. An enlarged view of the smaller IPC values is shown in the insets in the bottom panels. The values of IPC for up to degree 9 are obtained. The model parameters (g, σ_s, σ_n) are $(0.9, 0.3, 0.0)$ for (a), $(2.5, 2.0, 0.0)$ for (b), and $(0.8, 0.3, 0.1)$ for (c), respectively. While $\sum_{D=1}^{9} IPC_D$ almost equals L for (a), it falls significantly short of L for (b) and (c) (a dashed line denotes $\sum_{D=1}^{\infty} IPC_D = L$). This is attributed to the violation of the echo state property. Specifically, the RNN for (b) is chaotic, as indicated by a positive maximum conditional Lyapunov exponent [14, 20] ($\lambda_{\max} = 0.087 \pm 0.0067$), while the RNN for (c) is subject to noise. Shaded area represents mean±std for 10 different network and input signal realizations. Simulation time length is $T = 10^5$.

work, it should be possible to systematically determine an optimal L that balances resource utilization and computational performance. In this way, our results provide practical guidelines for the design of RC systems.

We showed that the level of neuronal correlations influences both the growth rate of memory capacity and the incremental emergence of higher-order computational abilities in an RNN. In neuroscience, the dimensionality of neural representation is known to mediate the trade-off between separability and generality in population coding: high-dimensional neural representations, characterized by low neuronal correlations, facilitate the separation of more complex patterns, while low-dimensional representations, marked by higher neuronal correlations, enhance noise robustness [22–24]. Our study offers a novel perspective on the relationship between neuronal geometry and computation from the viewpoint of dynamical information processing.

We employ a large random RNN as a canonical model for RC, but the generalizability of our findings to other RNN architectures, such as spiking neural networks, gated neural networks [25], and broader classes of dynamical systems warrants further investigation. Nonetheless, our insights provide valuable guidance for the design of cost-effective RC and offer a novel perspective on neuronal correlations.

We thank Joni Dambre for providing the source code for computing information processing capacity. S.T. was supported by JSPS KAKENHI Grant No. JP22KJ1959. T.A. was supported by MEXT KAKENHI Grant Numbers 23H04467 and by JSPS KAKENHI Grant Numbers 24H00723, 20K20520.

[1] H. Jaeger, *The "echo state" approach to analysing and training recurrent neural networks*, Tech. Rep. GMD Report 148 (German National Research Center for Information Technology, 2001).

[2] W. Maass, T. Natschlager, and H. Markram, Real-time computing without stable states: A new frame- work for neural computation based on perturbations, Neural Comput. **14**, 2531 (2002).

[3] 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).

[4] K. Nakajima and I. Fischer, *Reservoir Computing: Theory, Physical Implementations, and Applications* (Springer, 2021).

[5] H. Jaeger, *Short term memory in echo state networks*, Tech. Rep. GMD Report 152 (German National Research Center for Information Technology, 2002).

[6] J. Dambre, D. Verstraeten, B. Schrauwen, and S. Massar, Information processing capacity of dynamical systems, Sci. Rep. **2** (2012).

[7] J. Schuecker, S. Goedeke, and M. Helias, Optimal se-

quence memory in driven random networks, Phys. Rev. X **8** (2018).

[8] T. Haruna and K. Nakajima, Optimal short-term memory before the edge of chaos in driven random recurrent networks, Phys. Rev. E **100** (2019).

[9] M. Lukoševičius, H. Jaeger, and B. Schrauwen, Reservoir computing trends, Kunstliche Intelligenz **26**, 365 (2012).

[10] D. Sussilo and L. F. Abbott, Generating coherent patterns of activity from chaotic neural networks, Neuron **63**, 544 (2009).

[11] H. Sompolinsky, A. Crisanti, and H. J. Sommers, Chaos in random neural networks, Phys. Rev. Lett. **61**, 259 (1988).

[12] L. Molgedey, J. Schuchhardt, and H. G. Schuster, Suppressing chaos in neural network by noise, Phys. Rev. Letters **69**, 3717 (1992).

[13] 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 Computational Biology **18**, 1 (2022).

[14] S. Takasu and T. Aoyagi, Suppression of chaos in a partially driven recurrent neural network, Phys. Rev. Res. **6**, 013172 (2024).

[15] M. Massar and M. Serge, Mean-field theory of echo state networks, Phys. Rev. E **87** (2013).

[16] T. Toyoizumi and L. F. Abbott, Beyond the edge of chaos: Amplification and temporal integration by recurrent networks in the chaotic regime, Physi. Revi. E **84** (2019).

[17] M. Helias and D. Dahmen, *Statistical Field Theory for Neural Networks* (Springer, 2020).

[18] D. G. Clark, L. F. Abbott, and A. Litwin-Kumar, Dimension of activity in random neural networks, Phys. Rev. Lett. **131**, 118401 (2023).

[19] M. Inubushi and K. Yoshimura, Reservoir computing beyond memory-nonlinearity trade-off, Sci. Rep. **7** (2017).

[20] A. Pikovsky and A. Politi, *Lyapunov Exponents: A Tool to Explore Complex Dynamics* (Cambridge University Press, 2016).

[21] 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).

[22] S. Fusi, E. K. Miller, and M. Rigotti, Why neurons mix: high dimensionality for higher cognition, Current Opinion in Neurobiology **37**, 66 (2016), neurobiology of cognitive behavior.

[23] C. Stringer, M. Pachitariu, N. Steinmetz, M. Carandini, and K. D. Harris, High-dimensional geometry of population responses in visual cortex, Nature **571**, 361 (2019).

[24] 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, Nature Machine Intelligence **4**, 564 (2022).

[25] K. Krishnamurthy, T. Can, and D. J. Schwab, Theory of gating in recurrent neural networks, Phys. Rev. X **12**, 011011 (2022).

[26] 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).

[27] D. G. Clark, O. Marschall, A. van Meegen, and A. Litwin-Kumar, Connectivity structure and dynamics of nonlinear recurrent neural networks (2024), arXiv:2409.01969 [q-bio.NC].

Supplementary Material for "Neuronal correlations shape the scaling behavior of memory capacity and nonlinear computational capability of recurrent neural networks"

Shotaro Takasu and Toshio Aoyagi
Graduate School of Informatics, Kyoto University,
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan

CONTENTS

Appendix A: THEORETICAL ANALYSIS ON STATISTICS OF NEURONAL ACTIVITY

In this section, we derive statistics of neuronal activity based on dynamical cavity approach [18, 26]. 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 $N \to \infty$, the same result can be obtained via dynamical mean-field theory [11, 17] . 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 2). In contrast, the dynamical cavity method is, in principle, applicable to n-body neuronal statistics for any n. Furthermore, although the aforementioned limitation of dynamical mean-field theory has been recently addressed through sub-leading corrections to saddle-point solution [27], the dynamical cavity method remains advantageous for its conceptual clarity and computational feasibility.

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 delving into 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 independent quantities, $J_{ij}\phi(x_{j,t-1})$, which implies that $x_{i,t}$ follows a Gaussian distribution due to the central limit theorem. We can intuitively calculate $[x_{i,t}^2]$ as:

$$[x_{i,t}^2] = \sum_{j,k}^{N} [J_{ij}J_{ik}\phi_{j,t-1}\phi_{k,t-1}] \overset{?}{=} \sum_{j,k}^{N} [J_{ij}J_{ik}]\,[\phi_{j,t-1}\phi_{k,t-1}] = g^2\,[\phi_{j,t-1}^2] = g^2\langle\phi(x)\rangle_{x\sim\mathcal{N}(0,[x_{i,t-1}^2])}, \tag{A1}$$

yielding time series of $[x_{i,t}^2]$. Here we assumed in the second equation that recurrent weights, J_{ij} and J_{ik}, are independent of neuronal activities, ϕ_j and ϕ_k. Fortunately, this assumption is validated in the limit of $N \to \infty$, 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.

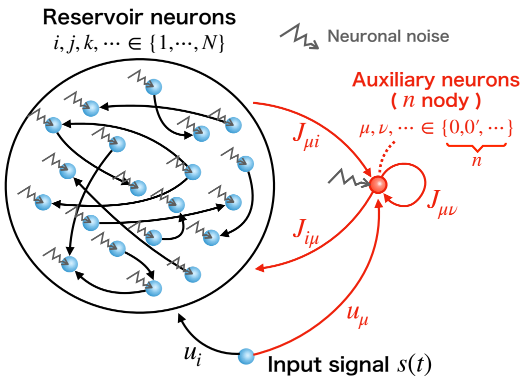

FIG. 5. Overview of dynamical cavity method. For simplicity, only one auxiliary neuron is illustrated, but in practice, n auxiliary neurons are introduced to calculate n-body statistics. Latin indices, $i, j, \cdots$, are used to label reservoir neurons, while Greek indices, $\mu, \nu, \cdots$, denote 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.

1. Single neuronal statistics

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

The overview of dynamical cavity method is described in Fig.5. One auxiliary neuron, indexed by 0, is added to the original reservoir RNN whose neurons are indexed by $i = 1, \cdots, N$. The original neuron is perturbated by the addition of the auxiliary neuron. By denoting the perturbation applied to the ith neuron at time t by $\delta_{i,t}$, we describe the dynamics of the auxiliary neuron as

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

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 $\chi_{ij,ts} \equiv \delta x_{i,t}/\delta I_{j,s}$ as the linear susceptibility of the ith neuron at time t, induced by an infinitesimal external inputs applied to the jth neuron at time $s \le t$, $I_{j,s}$. Then, as the perturbation to ith neuron at time t, $\delta_{i,t}$, is the sum of all direct perturbations to $j \in \{1, \cdots, N\}$ at time $s < t$ weighted by the corresponding susceptibility $\chi_{ij,ts}$, we obtain

$$\delta_{i,t} = \sum_{s<t}\sum_{j=1}^{N} \chi_{ij,ts} J_{j0}\phi_{0,s}. \tag{A3}$$

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

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

The physical interpretation of this equation is as follows. The first term, η_t, is the feedforward input from the reservoir. The second term, $\kappa_s\phi_{0,s}$, 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

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

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

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

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

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

Eventually, we obtain $[\kappa_s^2] \sim O(1/N)$ for $s < t$.

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

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

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

Thus, to leading order, we can ignore the non-trivial 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 $[x_{i,t}^2]$, 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 scailing assumption given by Eq.(3), 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 ($N \to \infty$) 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.(7) in the main text.

For later use, we evaluate the order of the quenched variance of $\langle x_i^2\rangle$, denoted by $\mathrm{Var}[\langle x_i^2\rangle]$, 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

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

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

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

Specifically, when $\sigma_s^2 = \tilde{\sigma}_s^2/\sqrt{N}$, $\mathrm{Var}[\langle x_0^2 \rangle] \sim O(1/N)$ is self-consistent as follows:

$$\mathrm{Var}[\langle x_0^2 \rangle]= g^4 \left([\langle \phi_i^2 \rangle\langle \phi_j^2 \rangle]_{i\neq j} - [\langle \phi_i^2 \rangle]^2\right) + 2g^4[\langle \phi_i\phi_j \rangle^2]_{i\neq j} + \frac{3g^4}{N}[\langle \phi_i^2 \rangle^2] + \frac{2\tilde{\sigma}_s^4}{N}. \tag{A14}$$

As shown in Appendix.A 2, 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

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

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

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

$$\left[\frac{1}{\langle x_i^2 \rangle}\right] = \frac{1}{[\langle x_i^2 \rangle]}, \tag{A16}$$

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, $[\langle x_i x_j \rangle^2] = 1/N^2 \sum_{i,j(i\neq j)} \langle x_i x_j \rangle^2$. 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 [18], 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

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

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

$$\delta_{i,t} = \sum_\nu J_{i\nu}\phi_{\nu,t} + \sum_{s<t}\sum_\nu\sum_j^N \chi_{ij,ts} J_{j\nu}\phi_{\nu,s}. \tag{A18}$$

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

$$\begin{aligned} x_{\mu,t+1} &= \underbrace{\sum_i^N J_{\mu i}\phi_{i,t}}_{\equiv \eta_{\mu,t}} + \sum_{s<t}\sum_\nu \underbrace{\sum_{i,j}^N J_{\mu i}J_{j\nu}\phi'_{i,t}\chi_{ij,ts}}_{\equiv \kappa_{\mu\nu,ts}}\,\phi_{\nu,s} + \sum_\nu J_{\mu\nu}\phi_{\nu,t} + u_\mu s_{t+1} + \xi_{\mu,t+1} \\ &\equiv \eta_{\mu,t} + \sum_\nu\sum_{s<t} \kappa_{\mu\nu,ts}\phi_{\nu,s} + \sum_\nu J_{\mu\nu}\phi_{\nu,t} + u_\mu s_{t+1} + \xi_{\mu,t+1}. \end{aligned} \tag{A19}$$

The physical interpretation of this equation is analogous to the case of single neuronal statistics (Eq.A4): $\eta_{\mu,t}$ denotes direct feedforward input from the reservoir to the auxiliary neuron μ, and $\kappa_{\mu\nu,ts}\phi_{\nu,s}$ arises from the perturbation induced in the reservoir by the auxiliary neuron ν at time s, which is readout 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 such as $[\langle \eta_{0,t}\eta_{0',t} \rangle^2]$ and $[\langle \kappa_{0\mu,ts}\kappa_{0'\nu,ts} \rangle^2]$. We can accomplish this task through direct calculation, analogous to the approach used in Appendix.A 1. First,

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

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, in the same manner, we obtain

$$[\langle\kappa_{0\mu,ts}\kappa_{0'\nu,ts}\rangle^2] = \frac{g^8}{N^4}\sum_{i,j,k,l}^{N}[\langle\phi'_{i,t}\phi'_{k,t}\chi_{ij,ts}\chi_{kl,ts}\rangle^2] \leq g^8[\langle(\phi'_{i,t})^2\rangle]^2[\langle\chi^2_{ij,ts}\rangle]^2 \sim O(1/N^2). \tag{A21}$$

Repeating the same calculations, all cross terms except for the two $O(1/N)$ terms, $[\langle\eta_{0,t}\eta_{0',t}\rangle^2]$ and $[\langle(u_0 s_t)(u_{0'}s_t)\rangle^2] = \tilde{\sigma}_s^4/N$, are shown to scale as $O(1/N^2)$. Consequently,

$$[\langle x_0 x_{0'}\rangle^2] = \frac{g^4}{N}[\langle\phi_i^2\rangle^2] + g^4[\langle\phi_i\phi_j\rangle^2]_{(i\neq j)} + \frac{\tilde{\sigma}_s^4}{N} + O\left(\frac{1}{N^2}\right). \tag{A22}$$

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

$$[\langle x_i x_j\rangle^2] = \frac{g^4}{N}[\langle\phi_i^2\rangle^2] + g^4[\langle\phi_i\phi_j\rangle^2] + \frac{\tilde{\sigma}_s^4}{N} + O\left(\frac{1}{N^2}\right), \tag{A23}$$

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 [18]. This discrepancy stems from the difference in the strength of the autocorrelation, $\langle x_i(t)x_i(t-\tau)\rangle_t$, which is $O(1)$ in the continuous-time model but only $O(1/\sqrt{N})$ 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. $x_i(t+1) = \gamma x_i(t) + \sum_j^N J_{ij}\phi_j(t)$ with $\gamma > 0$, which yields an $O(1)$ autocorrelation.

To calculate $[\langle\phi_i\phi_j\rangle^2]$, we introduce Price's theorem [17]. For two Gaussian variables, $\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} c_1 & \tau \\ \tau & c_2 \end{pmatrix}\right)$, 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:

$$\partial_\tau f_\phi(c_1, c_2, \tau) = f_{\phi'}(c_1, c_2, \tau) \tag{A24}$$

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

$$\langle\phi(z_1)\phi(z_2)\rangle = \sum_{n\geq 0}\frac{1}{n!}\tau^n f_{\phi^{(n)}}(c_1, c_2, 0) = \sum_{n\geq 0}\frac{1}{n!}\langle z_1 z_2\rangle^n\langle\phi^{(n)}(z_1)\rangle\langle\phi^{(n)}(z_2)\rangle. \tag{A25}$$

Applying Eq.(A25) to $\langle\phi_i\phi_j\rangle$ yields

$$\langle\phi_i\phi_j\rangle = \langle\phi'_{i,t}\rangle\langle\phi'_{j,t}\rangle\langle x_i x_j\rangle + \frac{1}{3!}\langle\phi^{(3)}_{i,t}\rangle\langle\phi^{(3)}_{j,t}\rangle\langle x_i x_j\rangle^3 + \frac{1}{5!}\langle\phi^{(5)}_{i,t}\rangle\langle\phi^{(5)}_{j,t}\rangle\langle x_i x_j\rangle^5 + \cdots, \tag{A26}$$

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

$$[\langle\phi_i\phi_j\rangle^2] = [\langle\phi'_{i,t}\rangle^2]^2[\langle x_i x_j\rangle^2] + O\left(\frac{1}{N^2}\right). \tag{A27}$$

Substituting Eq.(A27) into Eq.(A23) yields

$$[\langle x_i x_j\rangle^2] = \frac{g^4}{N}\langle\phi(x)^2\rangle_*^2 + g^4\langle\phi'(x)\rangle_*^4[\langle x_i x_j\rangle^2] + \frac{\tilde{\sigma}_s^4}{N} + O\left(\frac{1}{N^2}\right), \tag{A28}$$

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

$$[\langle x_i x_j\rangle^2] = \frac{1}{N}\frac{g^4\langle\phi(x)^2\rangle_*^2 + \tilde{\sigma}_s^4}{1 - g^4\langle\phi'(x)\rangle_*^4} + O\left(\frac{1}{N^2}\right), \tag{A29}$$

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 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

$$\frac{d}{dK}\langle\phi'(x)\rangle_* = \frac{d}{dK}\int_{-\infty}^{\infty}\phi'(\sqrt{K}z)Dz = \frac{1}{\sqrt{K}}\int_0^{\infty} z\phi''(\sqrt{K}z)Dz \leq 0, \tag{A30}$$

where Dz 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.(7), K is determined by solving

$$K = g^2\int\phi^2(\sqrt{K}z)Dz. \tag{A31}$$

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,

$$g^2 = \frac{K}{\int\phi^2(\sqrt{K}z)Dz}. \tag{A32}$$

Therefore, it suffices to show the inequality,

$$g^2\left(\int\phi'(\sqrt{K}z)Dz\right)^2 = \frac{K\left(\int\phi'(\sqrt{K}z)Dz\right)^2}{\int\phi^2(\sqrt{K}z)Dz} \leq 1, \tag{A33}$$

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

$$K\left(\int\phi'(\sqrt{K}z)Dz\right)^2 = \left(\int z\phi(\sqrt{K}z)Dz\right)^2 \leq \int z^2Dz\int\phi^2(\sqrt{K}z)Dz = \int\phi^2(\sqrt{K}z)Dz. \tag{A34}$$

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.(2) can be expressed as

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

where the elements of $\boldsymbol{a} \in \mathbb{R}^L$ and $C \in \mathbb{R}^{L\times L}$ are respectively $(\boldsymbol{a}_d)_i \equiv \langle s(t-d)x_i(t)\rangle$ 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.(3), the inverse of the covariance matrix C can be expanded using Neumann series expansion:

$$C^{-1} = \sum_{n=0}^{\infty} D^{-1}\left(-\tilde{C}D^{-1}\right)^n, \tag{B2}$$

where $D_{ij} \equiv \delta_{ij}\langle x_i(t)^2\rangle$ and $\tilde{C}_{ij} \equiv (1-\delta_{ij})\langle x_i(t)x_j(t)\rangle$.

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

$$\|\tilde{C}D^{-1}\|_F = \sum_{i,j(i\neq j)}^{\alpha\sqrt{N}}\frac{\langle x_ix_j\rangle^2}{\langle x_i^2\rangle^2} = \alpha^2N\left[\frac{\langle x_ix_j\rangle^2_{(i\neq j)}}{\langle x_i^2\rangle^2}\right] = \alpha^2N\frac{[\langle x_ix_j\rangle^2]_{(i\neq j)}}{[\langle x_i^2\rangle]^2} = \frac{\alpha^2}{[\langle x_i^2\rangle]^2}\frac{g^4\langle\phi(x)^2\rangle_*^2+\tilde{\sigma}_s^4}{1-g^4\langle\phi'(x)\rangle_*^4}, \tag{B3}$$

where the second equality employs the self-averaging property, and the final equality utilizes Eq.(A29). In addition, the third equality follows from Eq.(A16). Consequently, we obtain the sufficient condition,

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

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

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

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, $\lim_{N\to\infty} M_d = [M_d]$, where the square bracket denotes the average over network realizations, known as "quenched average" [17]. In addition, as neurons exhibit statistically identical dynamics, the quenched averaging operation eliminates the dependence of each term in Eq.(2) on neuron indices $i, j, k, \cdots$. Consequently, Eq.(B5) can be reduced to

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

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:

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

where we used the independence between $J_{0,i}$, $J_{0,j}$ and $\phi_{i,t}$ in the third equation. Since $(s_t, x_{i,t})$ are Gaussian variables, we apply Price's theorem, resulting in

$$[\langle s_{t-d}x_{0,t}\rangle^2] = g^2[\langle\phi_i'\rangle^2][\langle s_{t-(d-1)}x_{0,t}\rangle^2] + \delta_{d,0}\frac{\tilde{\sigma}_s^4}{N}, \tag{B8}$$

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

$$[\langle s_{t-d}x_{0,t}\rangle^2] = \frac{1}{N}\tilde{\sigma}_s^4\left(g\langle\phi'(x)\rangle_*\right)^{2d}. \tag{B9}$$

Consequently, we obtain

$$\text{1st term of Eq.(B6)} = \frac{\alpha\tilde{\sigma}_s^2\left(g\langle\phi'(x)\rangle_*\right)^{2d}}{[\langle x_i^2\rangle]}. \tag{B10}$$

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

$$
\begin{aligned}
[\langle s_{t-d}x_{0,t}\rangle\langle x_0x_{0'}\rangle\langle s_{t-d}x_{0',t}\rangle] &= \left[\sum_{i}^{N} J_{0i}\langle s_{t-d}\phi_{i,t-1}\rangle \sum_{j,k}^{N} J_{0j}J_{0'k}\langle\phi_j\phi_k\rangle \sum_{l}^{N} J_{0'l}\langle\phi_{l,t-1}s_{t-d}\rangle\right] \\
&= \sum_{i,j,k,l}^{N} [J_{0i}J_{0j}J_{0'k}J_{0'l}\langle s_{t-d}\phi_{i,t-1}\rangle\langle\phi_j\phi_k\rangle\langle\phi_{l,t-1}s_{t-d}\rangle] \\
&= \sum_{i,j,k,l}^{N} [J_{0i}J_{0j}J_{0'k}J_{0'l}]\,[\langle s_{t-d}\phi_{i,t-1}\rangle\langle\phi_j\phi_k\rangle\langle\phi_{l,t-1}s_{t-d}\rangle] \\
&= \frac{g^4}{N^2}\sum_{i,j}^{N} [\langle s_{t-d}\phi_{i,t-1}\rangle\langle\phi_i\phi_j\rangle\langle\phi_{j,t-1}s_{t-d}\rangle] \\
&= \frac{g^4}{N}[\langle s_{t-(d-1)}\phi_{i,t}\rangle^2][\langle\phi_i^2\rangle] + g^4\left[\langle s_{t-(d-1)}\phi_{i,t}\rangle\langle\phi_i\phi_j\rangle\langle\phi_{j,t}s_{t-(d-1)}\rangle\right]_{(i\neq j)} \\
&= \frac{g^4}{N}[\langle s_{t-(d-1)}\phi_{0,t}\rangle^2][\langle\phi_i^2\rangle] + g^4[\langle\phi'(x_i)\rangle^2]^2\left[\langle s_{t-(d-1)}x_{0,t}\rangle\langle x_0x_{0'}\rangle\langle x_{0',t}s_{t-(d-1)}\rangle\right],
\end{aligned}
\tag{B11}
$$

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 ($d=0$),

$$
[\langle s_t x_{0,t}\rangle\langle x_0x_{0'}\rangle\langle s_t x_{0',t}\rangle] = \frac{\tilde{\sigma}_s^4}{N}[\langle x_0x_{0'}\rangle u_0u_{0'}] = \frac{\tilde{\sigma}_s^4}{N}\left[u_0u_{0'}\left(\sum_{i,j}^{N} J_{0i}J_{0'j}\langle\phi_i\phi_j\rangle + u_0u_{0'}\langle s_t^2\rangle\right)\right] = \tilde{\sigma}_s^6 N^{-3/2}. \tag{B12}
$$

Thus, the second term of Eq.(B11) for $d=1$ is $O(N^{-3/2})$. On the other hand, from Eq.(B9), the first term of Eq.(B11) for $d=1$ scales as $O(N^{-2})$. As a result, this term becomes subleading and negligible. Consequently, solving the recurrence formula yields

$$
[\langle s_{t-d}x_{0,t}\rangle\langle x_0x_{0'}\rangle\langle s_{t-d}x_{0',t}\rangle] = \frac{\tilde{\sigma}_s^6\left(g\langle\phi'(x)\rangle_*\right)^{4d}}{N^{3/2}} + \text{subleading terms} \tag{B13}
$$

and thus

$$
\text{2nd term of Eq.(B6)} = -\frac{\alpha^2\tilde{\sigma}_s^4\left(g\langle\phi'(x)\rangle_*\right)^{4d}}{[\langle x_i^2\rangle]^2} + \text{subleading terms} \tag{B14}
$$

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 $d\geq 1$

$$
\begin{aligned}
[\langle s_{t-d}x_{0,t}\rangle\langle x_0x_{0'}\rangle\langle x_{0'}x_{0''}\rangle\langle x_{0'',t}s_{t-d}\rangle] &= \left[\sum_{i}^{N} J_{0i}\langle s_{t-d}\phi_{i,t-1}\rangle\sum_{j,k}^{N} J_{0j}J_{0'k}\langle\phi_j\phi_k\rangle\sum_{l,m}^{N} J_{0'l}J_{0''m}\langle\phi_l\phi_m\rangle\sum_{n}^{N} J_{0''n}\langle\phi_{n,t-1}s_{t-d}\rangle\right] \\
&= \frac{g^6}{N^3}\sum_{i,j,k}^{N}[\langle s_{t-d}\phi_{i,t-1}\rangle\langle\phi_i\phi_j\rangle\langle\phi_j\phi_k\rangle\langle\phi_{k,t-1}s_{t-d}\rangle] \\
&= \frac{g^6}{N^2}[\langle\phi_i^2\rangle]^2[\langle s_{t-(d-1)}\phi_{i,t}\rangle^2] + \frac{2g^6}{N}[\langle\phi_i^2\rangle]\left[\langle s_{t-(d-1)}\phi_{0,t}\rangle\langle\phi_0\phi_{0'}\rangle\langle\phi_{0',t}s_{t-(d-1)}\rangle\right] \\
&\qquad + g^6[\langle\phi_i'\rangle^2]^3[\langle s_{t-(d-1)}x_{0,t}\rangle\langle\phi_0\phi_{0'}\rangle\langle x_{0'}x_{0''}\rangle\langle x_{0'',t}s_{t-(d-1)}\rangle],
\end{aligned}
\tag{B15}
$$

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 ($d=0$),

$$
\begin{aligned}
[\langle s_t x_{0,t}\rangle\langle x_0x_{0'}\rangle\langle x_{0'}x_{0''}\rangle\langle x_{0'',t}s_t\rangle] &= \frac{\tilde{\sigma}^4}{N}[\langle x_0x_{0'}\rangle\langle x_{0'}x_{0''}\rangle u_0u_{0''}] \\
&= \frac{\tilde{\sigma}^4}{N}\left[u_0u_{0''}\left(\sum_{i,j}^{N} J_{0i}J_{0'j}\langle\phi_i\phi_j\rangle + u_0u_{0'}\langle(s_t)^2\rangle\right)\left(\sum_{i,j}^{N} J_{0'i}J_{0''j}\langle\phi_i\phi_j\rangle + u_{0'}u_{0''}\langle(s_t)^2\rangle\right)\right] \\
&= \tilde{\sigma}^8 N^{-2}.
\end{aligned}
\tag{B16}
$$

Therefore, the third term in Eq.(B15) for $d = 1$ is $O(1/N^2)$, while the first and second term 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.(B15), to leading order, we obtain

$$[\langle s_t x_{0,t}\rangle\langle x_0 x_{0'}\rangle\langle x_{0'} x_{0''}\rangle\langle x_{0'',t} s_t\rangle] = \frac{\tilde{\sigma}_s^8 \left(g\langle\phi'(x)\rangle_*\right)^{6d}}{N^2} + \text{subleading terms} \tag{B17}$$

and thus

$$\text{3rd term of Eq.(B6)} = \frac{\alpha^3\tilde{\sigma}_s^6 \left(g\langle\phi'(x)\rangle_*\right)^{6d}}{[\langle x_i^2\rangle]^3} + \text{subleading terms} \tag{B18}$$

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

$$M_d = \sum_{n=0}^{\infty}(-1)^n \left\{\frac{\alpha\tilde{\sigma}_s^2}{[\langle x_i^2\rangle]}\left(g\langle\phi'(x)\rangle_*\right)^{2d}\right\}^{n+1}, \qquad MC = \sum_{d=0}^{\infty}\sum_{n=0}^{\infty}(-1)^n \left\{\frac{\alpha\tilde{\sigma}_s^2}{[\langle x_i^2\rangle]}\left(g\langle\phi'(x)\rangle_*\right)^{2d}\right\}^{n+1}. \tag{B19}$$

Specifically, 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 $\phi'(x)$ and $\phi(x)^2$, obtaining

$$MC = \sum_{d=0}^{\infty}\sum_{n=0}^{\infty}(-1)^n \left\{\frac{\alpha\tilde{\sigma}_s^2}{[\langle x_i^2\rangle]}\left(\frac{g^2}{1+\frac{\pi}{2}[\langle x_i^2\rangle]}\right)^d\right\}^{n+1}, \tag{B20}$$

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

$$[\langle x_i^2\rangle] = \sigma_n^2 + g^2\left(-1 + \frac{4}{\pi}\arctan\sqrt{1+\pi[\langle x_i^2\rangle]}\right). \tag{B21}$$

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

Noting that by applying Eq.(B19), the denominator in the definition of the decay rate, Eq.(8), can be calculated as

$$\begin{aligned}\lim_{N\to\infty} L \times MC(1) &\overset{\text{Eq.}(B19)}{=} \lim_{N\to\infty}\alpha\sqrt{N}\times\sum_{d=0}^{\infty}\sum_{n=0}^{\infty}(-1)^n\left\{\frac{1}{\sqrt{N}}\frac{\tilde{\sigma}_s^2}{[\langle x_i^2\rangle]}\left(g\langle\phi'(x)\rangle_{x\sim\mathcal{N}(0,[\langle x_i^2\rangle])}\right)^{2d}\right\}^{n+1} \\ &= \frac{\alpha\tilde{\sigma}_s^2}{[\langle x_i^2\rangle]\left(1-g^2\langle\phi'(x)\rangle_*^2\right)},\end{aligned} \tag{B22}$$

the decay rate of memory capacity can be analytically derived as

$$\lim_{N\to\infty} r(L=\alpha\sqrt{N}) = \frac{[\langle x_i^2\rangle]\left(1-g^2\langle\phi'(x)\rangle_*^2\right)}{\alpha\tilde{\sigma}_s^2}\sum_{d=0}^{\infty}\sum_{n=0}^{\infty}(-1)^n\left\{\frac{\alpha\tilde{\sigma}_s^2}{[\langle x_i^2\rangle]}\left(g\langle\phi'(x)\rangle_*\right)^{2d}\right\}^{n+1}. \tag{B23}$$

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:

$$\sum_{d=0}^{\infty}\sum_{n=0}^{\infty}\left\{\frac{\alpha\tilde{\sigma}_s^2}{[\langle x_i^2\rangle]}\left(g\langle\phi'(x)\rangle_*\right)^{2d}\right\}^{n+1} = \sum_{d=0}^{\infty}\frac{\frac{\alpha\tilde{\sigma}_s^2}{[\langle x_i^2\rangle]}\left(g\langle\phi'(x)\rangle_*\right)^{2d}}{1-\frac{\alpha\tilde{\sigma}_s^2}{[\langle x_i^2\rangle]}\left(g\langle\phi'(x)\rangle_*\right)^{2d}} < \sum_{d=0}^{\infty}\frac{\alpha\tilde{\sigma}_s^2}{[\langle x_i^2\rangle]}\left(g\langle\phi'(x)\rangle_*\right)^{2d} < \frac{\alpha\tilde{\sigma}_s^2}{[\langle x_i^2\rangle]}\frac{1}{1-\left(g\langle\phi'(x)\rangle_*\right)^2} < \infty. \tag{B24}$$

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

$$\frac{\alpha\tilde{\sigma}_s^2}{[\langle x_i^2\rangle]} < \frac{\tilde{\sigma}_s^2}{[\langle x_i^2\rangle]}\sqrt{\frac{[\langle x_i^2\rangle]^2}{\tilde{\sigma}_s^4+\left([\langle x_i^2\rangle]-\sigma_n^2\right)^2}\left(1-\left(g\langle\phi'(x)\rangle_*\right)^4\right)} < \frac{\tilde{\sigma}_s^2}{[\langle x_i^2\rangle]}\sqrt{\frac{[\langle x_i^2\rangle]^2}{\tilde{\sigma}_s^4}} = 1, \tag{B25}$$

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

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

$$\lim_{N\to\infty} r(L=\alpha\sqrt{N}) = 1 - \sum_{n=1}^{\infty}(-1)^{n-1}\left(\frac{\tilde{\sigma}_s^2}{[\langle x_i^2\rangle]}\alpha\right)^n \frac{1-(g\langle\phi'(x)\rangle_*)^2}{1-(g\langle\phi'(x)\rangle_*)^{2n+2}}, \tag{B26}$$

where the second term arises due to the neuronal correlations.

Fig.6 illustrates the dependence of the theoretically derived decay rate on the model parameters, $\tilde{\sigma}_s$, g, and σ_n. As can be seen, the decay rate decreases rapidly for large inputs ($\tilde{\sigma}_s$), small recurrent weights (g), and small neuronal noise (σ_n). Notably, these changes in hyperparameters result in increased neuronal correlations, which can be confirmed by Eq.(A29). Consequently, we can expect the second term in Eq.(B26), which originates from neuronal correlations, to grow as the neuronal correlations increase, leading to a faster decline in the decay rate.

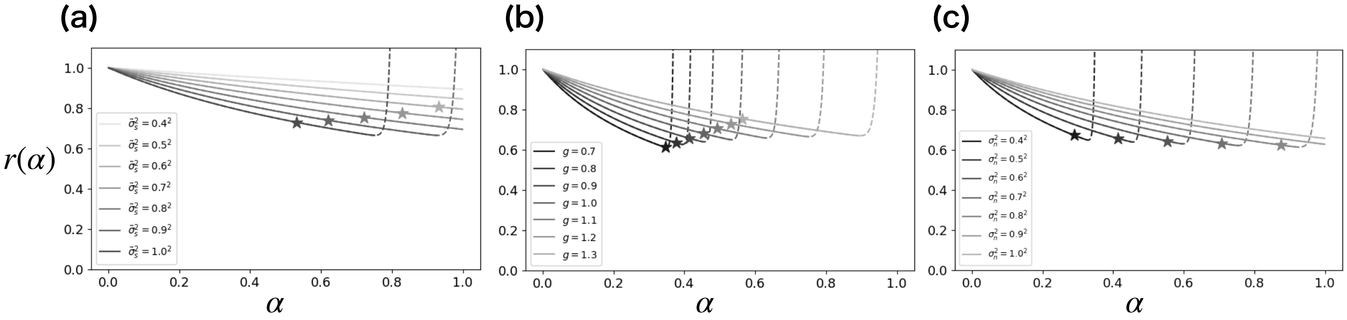

FIG. 6. Analytical solutions for the decay rate of memory capacity as a function of α. The activation function is an error function, $\phi(x) = \int_0^x e^{-\frac{\pi}{4}t^2}dt$. Each figure varies a single parameter: (a) input intensity $\tilde{\sigma}_s^2$, (b) recurrent weight scale g, (c) noise intensity 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.

3. 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

$$\lim_{N\to\infty} r(L=\alpha\sqrt{N}) = 1 + \sum_{n=1}^{\infty}\left(\frac{\tilde{\sigma}_s^2}{\sigma_n^2}\alpha\right)^n \frac{(1-g^2)^n}{1-g^{2n+2}}, \tag{B27}$$

where the value of g must be less than one to ensure the stability of the reservoir RNN. Analogous to non-linear RNNs, the memory capacity of the linear RNN with neuronal noise exhibits sublinear scaling with respect to L, as illustrated in Fig.7. In contrast, for a linear RNN in the absence of noise, the memory capacity is known to be exactly L [5], 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 (Fig.4(a)(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 cannot be applied to the noise-free linear RNN. In this setting, the diagonal entries of covariance matrix C are comparable to its non-diagonal entries, both scaling as $O(1/\sqrt{N})$, which prevents us from employing the Neumann series expansion.

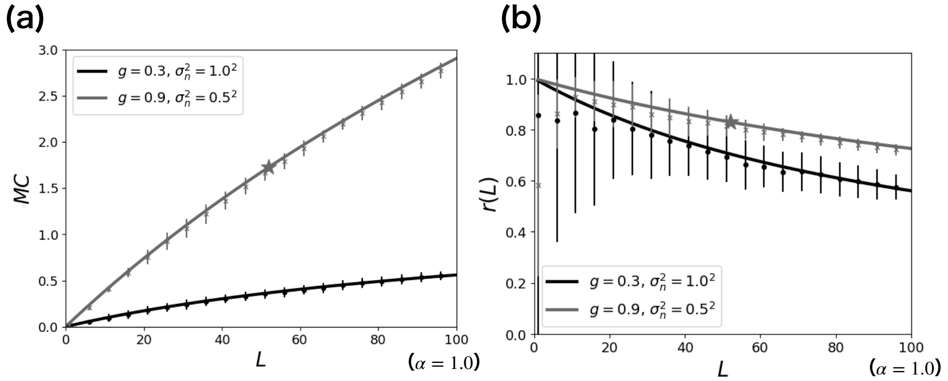

FIG. 7. (a) Memory capacity and (b) growth rate of memory capacity for linear RNNs. The figure's structure and content are the same as Fig.2, with analytical and numerical results presented in the same format. The input intensity is $\sigma_s^2 = 1.0^2/\sqrt{N}$ ($\tilde{\sigma}_s^2 = 1.0^2$). The activation function is an error function, $\phi(x) = \int_0^x e^{-\frac{\pi}{4}t^2}dt$. In simulations, the network size is $N = 10^4$, 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 $\hat{z}(t)$ approximates a given function of the inputs, $f[\cdots, s(t-1), s(t)]$. 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:

$$y_{\boldsymbol{d}} = \prod_{i\geq 0} \mathcal{H}_{d_i}\left(s(t-i)\right), \tag{C1}$$

where $\mathcal{H}_{d_i}(\cdot)$ ($d_i \geq 0$) denotes the normalized Hermite polynomial of degree d_i, and vector-form index $\boldsymbol{d} \equiv (d_i)_{i\geq 0}$ has only a finite number of non-zero 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 $y_{\boldsymbol{d}}$, defined as $deg(y_{\boldsymbol{d}}) \equiv \sum_{i\geq 0} d_i$.

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

$$C_T[f] \equiv 1 - \frac{\min_{\boldsymbol{w}} \langle(\hat{z}(t) - f(t))^2\rangle_T}{\langle f(t)^2\rangle_T}, \tag{C2}$$

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

$$IPC_{\text{total}} \equiv \sum_{\boldsymbol{d}} C_T[y_{\boldsymbol{d}}]. \tag{C3}$$

It has been established that $\lim_{T\to\infty} IPC_{\text{total}}$ is equivalent to the number of readout units, L, provided that the reservoir satisfies echo state property [1], whereby the reservoir's state is uniquely determined solely by input signals.

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

$$IPC_D \equiv \sum_{\substack{\boldsymbol{d} \\ \text{s.t. } deg(y_{\boldsymbol{d}})=D}} C_T[y_{\boldsymbol{d}}]. \tag{C4}$$

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

When calculating the capacities by numerical simulation, we must take care not to overestimate them, because for finite simulation time T, they are subject to a systematic positive error. Following the approach of Dambre et al. [6], we determine the threshold of the capacities, ϵ, as

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

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 non-zero. We assume $C_T = 0$ if C_T is lower than a threshold value ϵ. Following Dambre et al., we set $p = 10^{-4}$ for all our numerical simulations of memory capacity ($= IPC_1$) 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

$$x_i(t) = \sum_{j=1}^{N} J_{ij}\phi(x_i(t-1)) + u_i\sigma_s s(t) + \xi_i(t), \quad s(t) \sim \mathcal{N}(0,1), \tag{C6}$$

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. Simulation results for IPC for an RNN with the ReLU activation function

Our theoretical framework for the analytical derivation of memory capacity cannot be applied to the RNNs with the ReLU activation function. This limitation arises because Price's theorem (Eq.(A25)) cannot be used for the ReLU function, as it is non-smooth at the origin. Therefore, we numerically obtain memory capacity and IPC of a ReLU RNN, as shown in Fig.8. Similarly to the RNNs employing the sigmoid function, the memory capacity (IPC_1) increases sublinearly with L, while non-linear computational capacities (IPC_D for $D \geq 2$) emerge supralinearly and sequentially.

Note that unlike the sigmoid RNNs, the ReLU RNNs do not exhibit chaotic behavior. Instead, they diverge when the parameter g exceeds a critical value, g_c. To demonstrate this, substituting $\phi(x) = \max(0, x)$ into the dynamical mean-field equation (Appendix.A 1), we obtain

$$[x_i(t+1)^2] = \frac{g^2}{2}[x_i(t)^2] + s(t+1)^2 + \sigma_n^2 + O(1/\sqrt{N}). \tag{C7}$$

This self-consistent equation admits a stable solution only when $g < g_c = \sqrt{2}$. For $g > g_c$, the system diverges regardless of the parameters σ_s and σ_n. Consequently, the numerical simulations presented in Fig.8 are conducted with g values strictly less than $\sqrt{2}$.

3. 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 L if we ignore neuronal correlations. The following provides a concise proof of this claim. First, similarly to Eq.(4), the capacity for a basis function $y_{\boldsymbol{d}}$ can be expressed as

$$C_T[y_{\boldsymbol{d}}] = \frac{\boldsymbol{a}^\top C^{-1}\boldsymbol{a}}{\langle y_{\boldsymbol{d}}(t)^2\rangle}, \tag{C8}$$

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

$$C_T[y_{\boldsymbol{d}}] \approx [C_T[y_{\boldsymbol{d}}]] \approx \left[\sum_{i}^{L} \frac{\langle y_{\boldsymbol{d}}(t)x_i(t)^2\rangle}{\langle x_i(t)^2\rangle\langle y_{\boldsymbol{d}}(t)^2\rangle}\right] = L\left[\frac{\langle y_{\boldsymbol{d}}(t)x_i(t)\rangle^2}{\langle x_i(t)^2\rangle\langle y_{\boldsymbol{d}}(t)^2\rangle}\right] \propto L, \tag{C9}$$

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

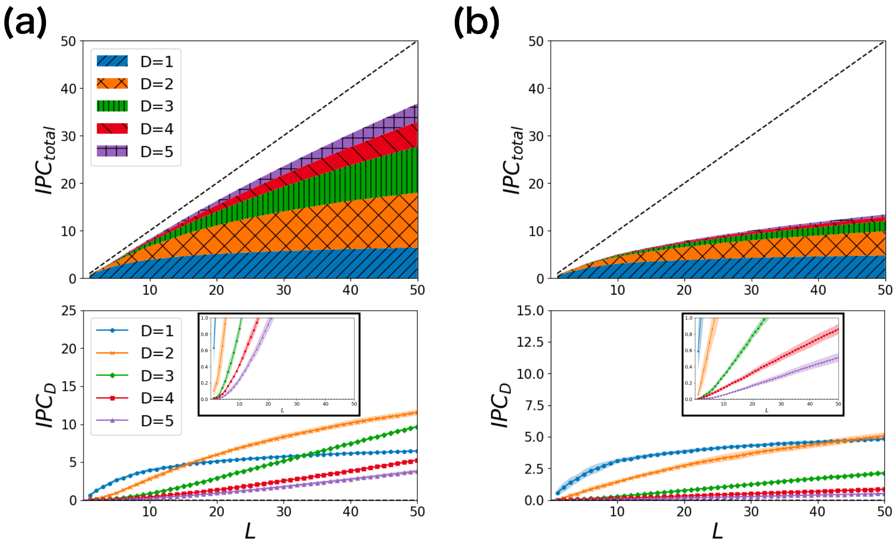

FIG. 8. IPC values for varying the number of readout neurons L of a reservoir RNN with the ReLU activation function $\phi(x) = \max(0, x)$ and network size $N = 1000$. The figure's structure and content mirror Fig.4. The values of IPC for up to degree 5 are calculated. Simulation time length is $T = 10^4$. The model parameters (g, σ_s, σ_n) are $(1.0, 1.0, 0.0)$ for (a) and $(1.0, 1.0, 0.3)$ for (b). The RNN for (a) satisfies the echo state property since its positive maximum conditional Lyapunov exponent is negtive ($\lambda_{\max} = -0.35 \pm 0.007$), while the property breaks down for (b) as the RNN receives neuronal noise.