

Recurrent neural networks running on quantum spins: memory accuracy and capacity

Aki Kutvonen^{1,2,*}, Keisuke Fujii^{3,+}, and Takahiro Sagawa^{1,+}

¹Department of Applied Physics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

²COMP Center of Excellence, Department of Applied Physics, Aalto University School of Science, P.O. Box 11000, FI-00076 Aalto, Espoo, Finland

³Graduate School of Science, Kyoto University, Sakyo-ku, Kyoto, 606-8502, Japan

*aki.kutvonen@gmail.com

+these authors contributed equally to this work

ABSTRACT

Quantum computing and neural networks show great promise for the future of information processing. In this paper we study a quantum reservoir computer, a framework harnessing quantum dynamics and designed for fast and efficient solving of temporal machine learning tasks such as speech recognition, time series prediction and natural language processing. Specifically, we study memory capacity and accuracy of a quantum reservoir computer based on the fully connected transverse field Ising model by investigating different forms of inter-spin interactions and computing timescales. We show that variation in inter-spin interactions leads to a better memory capacity in general, by engineering the type of interactions the capacity can be greatly enhanced and there exists an optimal timescale at which the capacity is maximized. To connect computational capabilities to physical properties of the underlying system, we also study the out-of-time-ordered correlator and find that its faster decay implies a more accurate memory.

Introduction

Partly due to the increase in available data and computational speed, machine learning models based on neural and deep neural network architectures have recently been remarkably successful^{1–4}. Recurrent neural networks, which are designed for modeling problems with temporal aspects, have been successful in various end-to-end systems such as machine translation, speech interpretation/generation and time series prediction^{5–7}. Traditionally the neural networks have been trained using back propagation through time algorithm, in which the network is trained against loss function by gradient descent in parameter space⁸.

In this paper we consider an alternative way of utilizing recurrent neural networks, the framework of reservoir computing^{9–12}. Unlike in the common back propagation approach, in reservoir computing the network (reservoir) is fixed during the training and only the readout weights are optimized. The reservoir networks can be directly implemented to physical systems allowing extremely fast state-of-the-art information processing^{13–18}. Here we consider a quantum reservoir computer, where the network dynamics are governed by rich quantum mechanical dynamics. Specifically, we consider a model based on fully connected transverse Ising model, introduced in Ref.¹⁹. While quantum reservoir computing has demonstrated its computational capability in principle, it is still unknown how to enhance its computational power by engineering the underlying physical properties. In this work, we engineer quantum reservoirs which allow up to 50% memory performance boost compared to the previous setups. We note that the aim is not to build an universal quantum computer and run algorithms on it, rather we are interested in a solver for recurrent neural networks which takes the advantage of rich quantum mechanical dynamics and exponentially scaling Hilbert spaces.

One of key attributes of recurrent neural networks is the memory performance which determines the applicability of the network for various tasks²⁰. We measure the memory performance with the long short term memory test and show that the memory accuracy and length of a quantum reservoir computer can be optimized by choosing an optimal timescale for the process or altering the inter-spin interactions. In addition, to better understand the underlying dynamics and its relation to the memory performance, we investigate the speed of information encoding from the inputs to the internal state of the quantum reservoir computer by using the out-of-time-ordered correlator (OTOC)²¹.

Setup

We start by describing the standard protocol of recurrent neural networks and reservoir computing. Given the discrete splitting of time $\{t_i\}$, values of the input nodes $\bar{u}(t_i) = \{u_j(t_i)\}_{j=0}^{n_u}$, the network nodes $\bar{x}(t_i) = \{x_j(t_i)\}_{j=0}^{n_x}$ and the output nodes $\bar{y}(t) = \{y_j(t)\}_{j=0}^{n_y}$, the update equation for a simple recurrent neural network is given by

$$x_i(t_{i+1}) = f\left[\sum_j^{n_u} W_{x_i, u_j}^{\text{in}} u_j(t_{i+1}) + \sum_j^{n_x} W_{x_i, x_j}^{\text{int}} x_j(t_i)\right], \quad (1)$$

where f is some nonlinear differentiable function such as sigmoid function, W^{in} , W^{int} are the transition weights from the input nodes and hidden nodes, and n_u , n_x and n_y are the number of input, hidden and output nodes, respectively. A simple recurrent neural network is depicted in Fig. 1 (a). The network outputs are given by

$$y_i(t_{i+1}) = \sum_j^{n_{\text{out}}} W_{y_i, z_j}^{\text{out}} z_j(t_i), \quad (2)$$

where we assume that the readout signals are gathered from a subset $\{z_i\} \in \{x_i\}$ of size $n_{\text{out}} < n_x$ from hidden nodes. In general the output $y_i(t_{i+1})$ could have contribution from last inputs $\bar{u}(t_i)$ and outputs $\bar{y}(t_i)$ or the nodes itself could have more complex design²².

Given the values of input nodes $\bar{u}(t_i)$ (input sequence), the training goal is to find a set of weights W , which minimize the error $\mathcal{E} = \sum_i \|\bar{y}^{\text{targ}}(t_i) - \bar{y}(t_i)\|$, where $\bar{y}^{\text{targ}}(t)$ is the target sequence and $\|\cdot\|$ denotes a differentiable distance function, such as the mean square distance or cross entropy between distributions. The standard method of back propagation for optimizing typically millions of parameters is to evaluate and numerically descend along the gradients $d\mathcal{E}/dW$ for all elements in W^{in} and W^{int} . Reservoir computing seeks to optimize the task with less computational power by training the output weights W^{out} alone. In the case of the mean square error the task reduces to the standard linear regression task, and the optimal configuration of weights, $W_{\text{opt}}^{\text{out}} = \arg \min_{W^{\text{out}}} \sum_i (\bar{y}^{\text{targ}}(t_i) - \bar{y}(t_i))^2$, is given by

$$W_{\text{opt}}^{\text{out}} = \mathcal{Z}^{-1} \mathcal{Y}, \quad (3)$$

where $\mathcal{Y}_{i,j} = y_j^{\text{targ}}(t_i)$ and \mathcal{Z}^{-1} denotes the Moore–Penrose pseudo inverse of matrix $\mathcal{Z}_{i,j} = z_j(t_i)$. Since in reservoir computing the inter network optimization is not necessary during the training, a physical system can drive the network dynamics directly.

The memory of the network is measured by the short term memory (STM) task defined as a task to remember the input d steps ago:

$$y_d^{\text{targ}}(t_i) = s_{i-d}. \quad (4)$$

The memory accuracy is defined as the correlation between the prediction and the target function:

$$\mathcal{R}(d) = \text{cov}^2(y, y_d^{\text{targ}}) / \sigma^2(y) \sigma^2(y_d^{\text{targ}}), \quad (5)$$

where cov denotes covariance and σ^2 variance. The memory capacity \mathcal{C} is defined as $\mathcal{C} = \sum_{d=1}^{d_c} \mathcal{R}(d)$, where d_c is a cut off distance, which we set to 100.

For efficient usage, reservoir computing requires complex dynamics, large network sizes and bounded dynamics. These requirements are naturally met by interacting quantum mechanical spin systems, whose state space scales exponentially with respect to the number of spins and the dynamics are governed by complex unitary dynamics. The notation here follows the standard quantum mechanics notation. The state of a spin is a two-dimensional complex vector spanned by the eigenstates of the Pauli Z operator, $\{|0\rangle, |1\rangle\}$. The total state space of N qubits is a tensor product space of the two-dimensional individual spin spaces. The total (pure) state of the system is represented by a 2^N -dimensional vector $|\Psi\rangle$. A statistical mixture of pure states can be described by $2^N \times 2^N$ -dimensional density matrix ρ . For a closed quantum system, the time evolution is governed by the time independent Schrödinger equation, which for the density matrix can be written as

$$\rho(t + \Delta t) = e^{-i\Delta t H} \rho(t) e^{i\Delta t H}, \quad (6)$$

where H is a $2^N \times 2^N$ -dimensional hermitian matrix, the Hamiltonian of the system, which defines the system dynamics. Here we consider an extensively studied model, the fully connected transverse field Ising model, whose Hamiltonian is given by

$$H = \sum_{i,j} J_{i,j} X_i X_j + h_i Z_i, \quad (7)$$

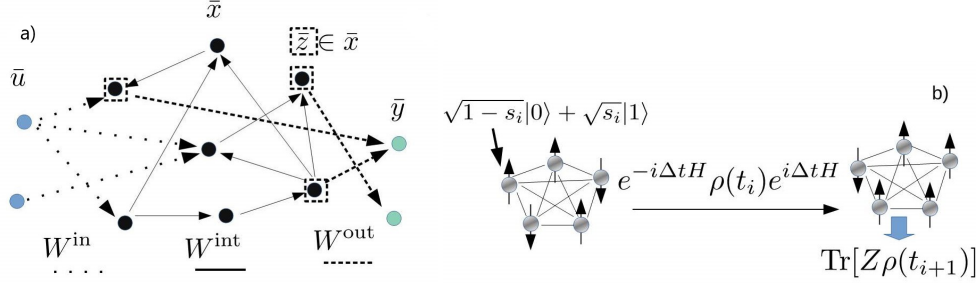


Figure 1. a) In reservoir computing the input data is fed from the input nodes \bar{u} with weights W^{in} to the network hidden nodes \bar{x} . The internal immutable dynamics are governed by W^{int} . Outputs are obtained by weighting the readout nodes \bar{z} , with the readout weights W^{out} . b) In quantum reservoir computing an input s_i is fed to a spin by setting its state to $|\Psi_{s_i}\rangle = \sqrt{1-s_i}|0\rangle + \sqrt{s_i}|1\rangle$. After free time evolution the readout node values are obtained as the ensemble averages of the spins $\langle Z \rangle$.

, where X_i and Z_i are the Pauli X and Z operators at site i . In this model all the spins interact with each other in x -direction and are coupled to an external magnetic field in z -direction.

We consider one-dimensional inputs $\bar{s}(t_i) = s_i$ and outputs $\bar{y}(t_i) = y_i$. At time t_i the input s_i is fed to the system by setting the state of the first spin to $|\Psi_{s_i}\rangle = \sqrt{1-s_i}|0\rangle + \sqrt{s_i}|1\rangle$. The density matrix of the system then becomes $\rho \mapsto |\Psi_{s_i}\rangle \langle \Psi_{s_i}| \otimes \text{Tr}_1[\rho]$, where Tr_1 denotes the trace over the first spin degree of freedom.

After the input is set, the system continues evolving itself for time Δt . During this time the dynamics are governed by Eq. (6) and the information encoded in the first spin will spread through the system. During $[t_i, t_{i+1}]$ we measure the average spin values in the z -direction, $\langle Z_i \rangle = \text{Tr}[Z_i \rho]$, which corresponds to the readout nodes in our setup. After gathering a sufficient amount of input-output signal pairs, the readout weights W^{out} are trained according to Eq. (3) and the corresponding outputs and the future predicted outputs are obtained from Eq. (2). The input-readout loop is illustrated in Fig. 1 (b).

Since the number of spins in our setup, N_s , is fairly small, we also gather the readout signals from intermediate times between $[t, t + \Delta t]$. The input timescale Δt is divided into N_v time steps and thus the intermediate times are given by $t_i^k = t_i + k\Delta t/N_v$, where $k = 1, \dots, N_v$. The readout signals are gathered as $\text{Tr}[Z_j \rho(t_i^k)]$ for all the spins $j = 1, \dots, N_s$ and the intermediate times t_i^k . Thus the total number of the output signals is $N_{\text{spins}} N_v$.

We note that an experimental realization of the setup requires multiple copies of the same system in order to make the effect of measurement back-action irrelevant. This is feasible by using nuclear magnetic resonance systems, where a huge ensemble of identical molecules can be utilized. Further details on the input-readout loop with practical implementation considerations are available in Ref. ¹⁹.

Results

We set the number of spins and intermediate times to $N_s = 6$ and $N_v = 10$, respectively, which gives a relatively good performance at reasonable computational cost¹⁹. The input sequences were random binary inputs, $s_i = \{0, 1\}$, sampled for each simulation separately. Initial state was set to the maximally entangled state, where all the possible states $|\Psi\rangle$ appear with the same probability $1/2^{N_s}$. All of the simulations consisted of an initial equilibration phase of 2000 time steps during which the inputs were injected but no training was done. During following 3000 time steps the input and readout signals were gathered and the output weights were trained according to Eq. (3). In the testing phase of 1000 steps the output was generated according to Eq. (2) and compared to the target output of Eq. (4).

We considered various values of magnetic coupling h , inter-spin interactions J (Eq. (7)) and input intervals Δt . Values for both h and J were sampled from 0.5 centered uniform distribution $\mathcal{U}[0.5 - h_s/J_s, 0.5 + h_s/J_s]$. The scale parameters h_s and J_s simulated were $h_s, J_s \in \{0.01, 0.1, 0.25, 0.50, 2, 3\}$ and the number of simulations for each setup was 25. We also sampled J from Beta-distribution with values $(\alpha, \beta) = \{(0.1, 0.1), (0.7, 0.7), (0.9, 0.9), (1.1), (2, 2), (9, 9), (100, 100)\}$ while keeping the value of $h = 0.5$ fixed (see Supplementary Fig. S1 online). Figure 2 shows the memory capacity as a function of the input timescale Δt for selected scale parameters h_s and J_s . For all the parameters tested, larger deviation in J , corresponding to larger scale values of J_s and smaller values of (α, β) , resulted in larger maximum capacity. However, the given boost in performance saturated around $J_s = 0.5$ and $(\alpha, \beta) = (1, 1)$, corresponding to uniform $[0, 1]$ sampling. Within the parameters sampled, the values of h did not have considerable effect on the memory capacity.

In all setups there existed an optimal input timescale Δt_{opt} at which the memory capacity is maximized as shown in Fig. 2. The STM task measures how well the past input signals can be decoded from readout signals $\{\langle Z_i \rangle\}_{i=1}^6$ during time window

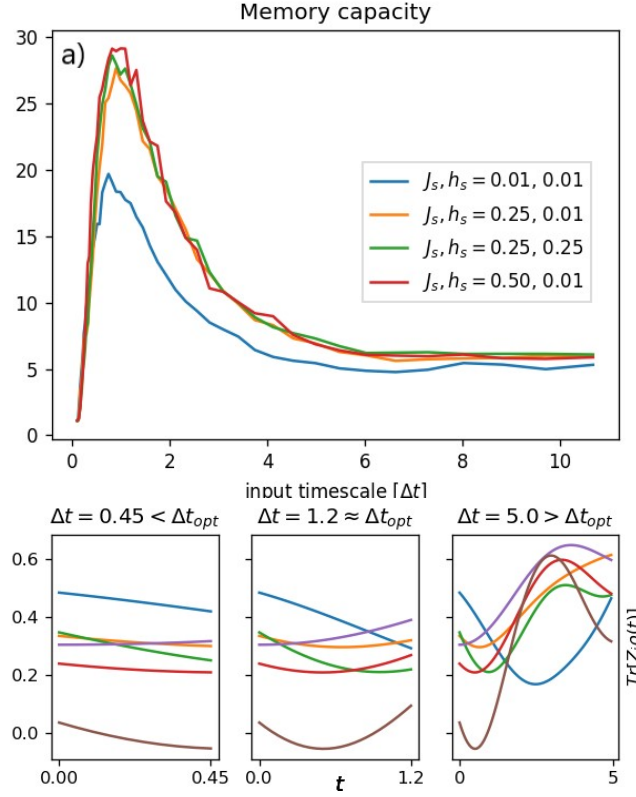


Figure 2. a) Memory capacity shown as a function of the input timescale Δt . The values for h and J are sampled from $\mathcal{U}[0.5 - h_s, 0.5 + h_s]$ and $\mathcal{U}[0.5 - J_s, 0.5 + J_s]$, respectively. Deviation in J leads to better performance but the deviation in h is less important. Furthermore, for all parameters simulated, there exists an optimal timescale for Δt , which maximizes the capacity. b) Readout signals $\text{Tr}[Z_i \rho(t)]$ for three different time windows $[t_i, t_i + \Delta t]$ for parameters $h_s = 0.01, J_s = 0.5$. The leftmost, center and the rightmost panels corresponds to timescales shorter, equal and longer than the optimal timescale Δt_{opt} , respectively.

$[t_i, t_{i+1}] = [t_i, t_i + \Delta t]$, using the simple linear map of Eq. (2). On too short timescales the dynamics are almost linear at $t \in [t_i, t_{i+1}]$ and thus there is no prediction power outside the linear signal regime. On too long timescales the signals are too chaotic during $t \in [t_i, t_{i+1}]$ giving rise to sub-optimal performance. Thus at optimal timescales the dynamics are not completely linear but not yet too chaotic either. Signals at different timescales are illustrated in the lower panel of Fig. 2.

Instead of sampling J and h from a distribution, we considered engineered forms of J while keeping the value of h fixed to 0.5. The coupling between the spins was set to

$$J_{i,j}^k = (i+j)^k / c_k, \quad (8)$$

where i and j are the spin indexes, k is a scaling parameter and $c_k = \sum_{i,j} 2N_s^2 (i+j)^k$ is a k -dependent constant which ensures comparable energy scales between setups with the mean interaction $\mathbb{E}[J^k] = 0.5$. In addition to the fixed mean, the form of J^k above has the benefit of inducing deviation in J with simple parametrization k without additional noise induced by random sampling. Furthermore, all the spins have different average couplings to the other spins, $\sum_j J_{i,j}^k < \sum_j J_{i+1,j}^k, i \in 1, 2, \dots, 5$. We expect that this in general results into more rich and divergent signals increasing the prediction power of Eq. (2). In fact, compared to the random sampled J , the memory capacity can be increased up to 50% by using J^k , as shown in Fig. 3. Maximum memory capacity has a non-monotonic dependence on k and the best performance is obtained with $k = 4$. Furthermore, the larger the k , the larger the optimal input interval Δt_{opt} . The number of simulations was 10 for each of the value $k \in \{1, 2, 4, 6, 8\}$.

For practical purposes, the memory accuracy $\mathcal{R}(d)$ (Eq. (5)) may be more important than the overall capacity. Figure 3 shows the memory accuracy as a function of the delay d for selected values of k and input intervals Δt . On short Δt the memory extends up to delays 100 but is inaccurate throughout the delay range. On the contrary at long intervals Δt the memory is shorter but more accurate. As the capacity $\mathcal{C} = \sum_d \mathcal{R}(d)$ measures the sum of accuracies over memory length, the largest \mathcal{C} is obtained somewhere in the intermediate timescales where the correlation is relatively good over all the delays up to the cutoff delay.

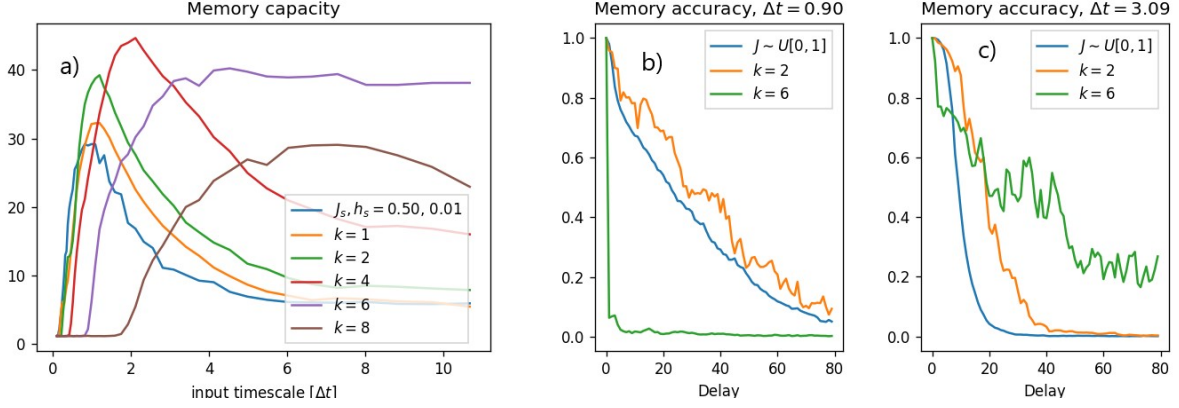


Figure 3. a) Memory capacity for different inter-spin couplings $J_{i,j}^k = (i+j)^k/c_k$. The maximum capacity has a non-monotonic behaviour in k , while the location of the maximum shifts towards larger Δt in increasing k . b) Memory accuracy at the short time interval. The interval Δ is too short for $k = 6$ while the other setups exhibit the long but inaccurate memory. c) At long time intervals the setups exhibit the short but accurate memory. Higher k values result into a longer but inaccurate of memory.

It is also interesting to compare the memory type to the signals given in the lower panel of Fig 2. On longer intervals Δt the time window seems to give a good view of the last couple of time windows but the prediction power cannot extend far due to the chaotic nature of the signals, corresponding to the short but accurate memory in the STM test. Vice versa, at shorter intervals Δt the linear prediction extends longer in the past but it is never quite accurate.

When comparing the different J^k at the same input intervals, the larger k exhibits the longer but less accurate memory and smaller k values are associated with the shorter but more accurate memory. Thus, like in the case of the input interval Δt , also the intermediate k value results in the highest memory capacity (Fig. 3). We note that the short but accurate type of memory associated with timescales $\Delta t > \Delta t_{\text{opt}}$ and lower values of k might be preferred for practical applications where a high enough accuracy is needed.

In order to investigate the information spreading from the input spin to the other spins, we studied the out-of-time-ordered correlator (OTOC) defined as:

$$O(t) = \frac{1}{N_s} \sum_{j=2}^{N_s} \text{Tr}[Z_1(t_0)Z_j(t_0+t)Z_1(t_0)Z_j(t_0+t)]. \quad (9)$$

The OTOC measures delocalization of information and is previously used in the context of fast scrambling of black hole and thermalization of closed quantum systems^{21,23}. We ran 60 simulations, each consisting of an initial washout phase of 20000 time steps with time step $\Delta\tau = 0.05$. Figure 4 shows the OTOC for different J^k couplings and the case that J is sampled from $\mathcal{U}[0, 1]$. The larger the value of k is, the slower the OTOC decays, signaling slower decay of information from the first spin to the other spins. This is in agreement with the notion that the average interaction of spin 1 to the others, $\sum_j J_{1,j}/N_s = \sum (1+j)^k/(c_k N_s)$, is a decreasing function in k . Larger values of k means that the first spin is more weakly coupled to the other spins resulting into a slower decay of information.

Comparison of the memory capacity and the OTOC, Figs. 3 and 4, shows that the slower decay of OTOC implies a larger optimal timescale for the STM task. Thus the system needs more time for effective distribution of information (encoding) to the other spins. These systems exhibit the long but inaccurate type of memory. The encoding takes time and the encoded information can be decoded back after a long time, but the decoding is never particularly efficient. In contrary, the random sampled and low k setups are able to encode the information to the other spins fast and the decoding is efficient, but the memory does not extend as far. We can further synthesize desired memory properties by spatially multiplexing distinct quantum reservoirs²⁴.

Summary

We studied the memory capacity and accuracy of a quantum reservoir computer based on the fully connected transverse field Ising model. We observed that the input data interval Δt has a system dependent optimal value Δt_{opt} which leads to the maximum capacity. Furthermore, our results suggest that the deviation in inter-spin interaction strengths leads to improved

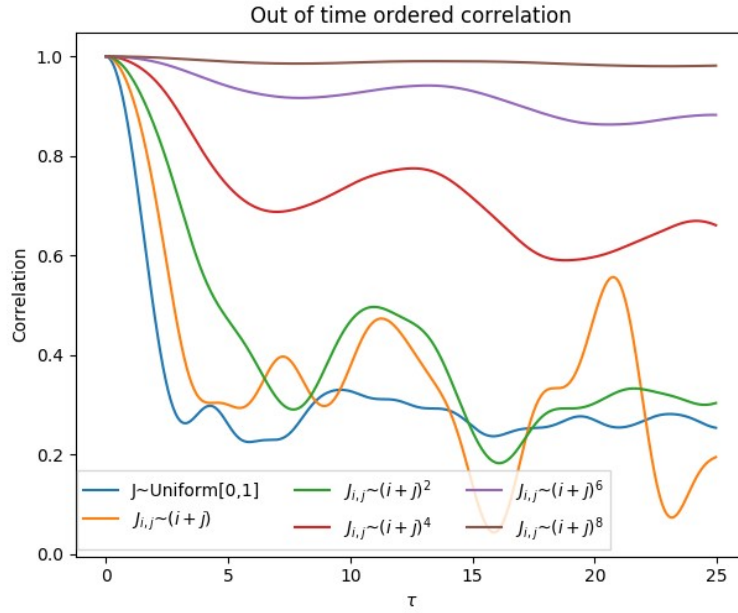


Figure 4. The OTOC between the input spin (site $i = 1$) and the other spins for different couplings $J_{i,j}^k = (i + j)^k / c_k$. The larger the k , the slower the information decays from the input spin. This is in agreement with the notion that the input spin is more weakly coupled to the other spins with higher values of k .

memory capacity up to some limit. While the maximum capacity is a sign of performance, in practical applications we may be interested in the memory accuracy, which is in general improved by using a slightly longer input time interval than Δt_{opt} . In addition, our results suggest that the fast decay of the OTOC is a sign of fast encoding of information from the input spins to the others. These setups exhibit the accurate memory at the cost of memory length. Possible future research directions include more advanced data input-output strategies, which could implement e.g., multi-dimensional inputs and outputs, mixture of multiple readout timescales and efficient strategies for determining optimal timescales for given tasks.

References

1. William, C., Navdeep, J., Quoc, V. & Oriol, V. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *ICASSP* (2016).
2. Hirschberg, J. & Manning, C. Advances in natural language processing. *Sci.* **349**, 261–266 (2015).
3. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nat.* **521**, 436 (2015).
4. Mnih, V. e. a. Human-level control through deep reinforcement learning. *Nat.* **518**, 529 (2015).
5. Chung, J. e. a. A recurrent latent variable model for sequential data. In *NIPS*, vol. 28, 2980–2988 (2015).
6. Mikolov, T. e. a. Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, vol. 2, 1045–1048 (2010).
7. Sutskever, I., Vinyals, O. & V.L., Q. Sequence to sequence learning with neural networks. In *NIPS*, vol. 27, 3104–3112 (2014).
8. Rumelhart, D., Hinton, G. & R.J., W. Learning representations by back-propagating errors. *Nat.* **323**, 533 (1986).
9. Jaeger, H. & Haas, H. Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. *Sci.* **304**, 78 (2004).
10. Maass, W., Natschlanger, T. & Markram, H. Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput.* **14**, 2531 (2002).
11. Verstraeten, D., Schrauwen, B. & D’Haene, M. . S. D. An experimental unification of reservoir computing methods. *Neural Netw.* **20**, 391 (2007).

12. Lukoševičius, M. & Jaeger, H. Reservoir computing approaches to recurrent neural network training. *Comput. Sci. Rev.* **3**, 127 – 149 (2009).
13. Laurent, L. e. a. High-speed photonic reservoir computing using a time-delay-based architecture: Million words per second classification. *Phys. Rev. X* **7**, 011015 (2017).
14. Dambre, J. e. a. Information processing capacity of dynamical systems. *Sci. Rep.* **2**, 514 (2012).
15. Woods, D. & Naughton, T. Photonic neural networks. *Nat. Phys.* **8**, 257 (2012).
16. Brunner, D. e. a. Parallel photonic information processing at gigabyte per second data rates using transient states. *Nat. Commun.* **4**, 1364 (2013).
17. Vandoorne, K. e. a. Experimental demonstration of reservoir computing on a silicon photonics chip. *Nat. Commun.* **5**, 3541 (2014).
18. Chao, D. e. a. Reservoir computing using dynamic memristors for temporal information processing. *Nat. Commun.* **8**, 2204 (2017).
19. Fujii, K. & Nakajima, K. Harnessing disordered-ensemble quantum dynamics for machine learning. *Phys. Rev. Appl.* **8**, 024030 (2017).
20. Nakajima, K. e. a. Exploiting short-term memory in soft body dynamics as a computational resource. *J. R. Soc. Interface* **11**, 20140437 (2014).
21. Ruihua, F., Pengfei, Z., Huitao, S. & Hui, Z. Out-of-time-order correlation for many-body localization. *Sci. Bull.* **62**, 707 – 711 (2017).
22. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735–80 (1997).
23. Maldacena, J., Shenker, S. H. & Stanford, D. A bound on chaos. *J. High Energy Phys.* **2016**, 106 (2016).
24. Nakajima, K., Fujii, K., Negoro, M., Mitarai, K. & Kitagawa, M. Boosting computational power through spatial multiplexing in quantum reservoir computing. *ArXiv e-prints* (2018). [1803.04574](https://arxiv.org/abs/1803.04574).

Acknowledgements

A.K. is supported by Jenny and Antti Wihuri Foundation through Council of Finnish Foundations' Post Doc Pool, T.S. by JSPS KAKENHI Grant Number JP16H02211 and K.F. by JST PRESTO Grant Number JPMJPR1668, JST ERATO Grant Number JPMJER1601, and JST CREST Grant Number JPMJCR1673. We thank Eiki Iyoda for useful discussions.

Author contributions statement

A.K. conducted the simulations and wrote the first draft of the manuscript. All authors analyzed the results and reviewed the manuscript.

Additional information

The authors declare no competing interests.