

# Learning Nonlinear Input-Output Maps with Dissipative Quantum Systems

Jiayin Chen and Hendra I. Nurdin\*

*School of Electrical Engineering and Telecommunications,  
The University of New South Wales (UNSW), Sydney NSW 2052, Australia*

## Abstract

In this paper, we develop a theory of learning nonlinear input-output maps with fading memory by dissipative quantum systems, as a quantum counterpart of the theory of approximating such maps using classical dynamical systems. The theory identifies the properties required for a class of dissipative quantum systems to be *universal*, in that any input-output map with fading memory can be approximated arbitrarily closely by an element of this class. We then introduce an example class of dissipative quantum systems that is provably universal. Numerical experiments illustrate that with a small number of qubits, this class can achieve comparable performance to classical learning schemes with a large number of tunable parameters. Further numerical analysis suggests that the exponentially increasing Hilbert space presents a potential resource for dissipative quantum systems to surpass classical learning schemes for input-output maps.

---

\* h.nurdin@unsw.edu.au (corresponding author)

## I. INTRODUCTION

We are in the midst of the noisy intermediate-scale quantum (NISQ) technology era [39], marked by noisy quantum computers consisting of roughly tens to hundreds of qubits. Currently there is a substantial interest in early applications of these machines that can accelerate the development of practical quantum computers, akin to how the humble hearing aid stimulated the development of integrated circuit (IC) technology [30]. NISQ quantum computing machines will not be equipped with quantum error correction and are thus incapable of performing continuous quantum computation.

Several research directions are being explored for NISQ-class machines. One direction is to demonstrate so-called “quantum supremacy”, in which NISQ machines can perform computational tasks that are demonstrably out of the reach of the most powerful digital supercomputers. The computational tasks include sampling problems such as boson sampling [2, 27], instantaneous quantum polynomial (IQP) computation [11, 27], and sampling from random quantum circuits [8]. Recent works have also proposed quantum machine learning algorithms that offer provable speedups over their classical counterparts [7]. Another direction is the development of variational algorithms on hybrid classical-quantum machines to solve certain classes of optimization problems. Algorithms proposed include the quantum approximate optimization algorithm (QAOA) [15], the quantum variational eigensolver (QVE) [29, 38] and variations and generalizations thereof, e.g., [31, 47]. Experimental demonstration of QVE for calculating the ground-state energy of small molecules has been reported in [22], while the application of QAOA for unsupervised learning of a clustering problem can be found in [35].

An alternative paradigm to the quantum gate-based approaches above is to harness the computational capability of dissipative quantum systems. Dissipative quantum dynamics has been shown to be able to realize universal quantum computation [46] and has been applied in a time-delay fashion for supervised quantum machine learning without intermediate measurements [4]. Recently, quantum reservoir computers (QRCs) are introduced to harness the complex real-time quantum dissipative dynamics [17, 32]. This approach is essentially a quantum implementation of classical *reservoir computing* schemes, in which a dynamical system processes an input sequence and produces an output sequence that approximates a target sequence, see, e.g., [21, 26, 28]. The main philosophy in reservoir computing is that the dynamics in arbitrary naturally occurring or engineered dynamical systems could potentially be exploited for computational purposes. In particular, a dynamical system could be used for computation without precise tuning or optimization of its parameters. To possess temporal information, the systems are required to satisfy three properties [28]: the *convergence property* [36], the *fading memory property* [10] and form a family of systems with the *separation property*. The convergence property ensures that computations performed by a dynamical system are independent of its initial condition, and the fading memory property implies that outputs of a dynamical system stay close if the corresponding inputs are close in recent times. The separation property states that there should be a member in the family of systems with dynamics sufficiently rich to distinguish any two different input sequences. Classical reservoir computing has been realized as simple nonlinear photonic circuits with a delay line [5] and in neuromorphic computing based on nanoscale oscillators [43], and it has been demonstrated to achieve state-of-the-art performance on applications such as spoken digit recognition [43].

Nonlinear input-output (I/O) maps with fading memory can be approximated by a series

expansion such as the well-known Volterra series [10]. They can also be approximated by a family of classical nonlinear dynamical systems that have the three properties introduced in the previous paragraph. Such a family of dynamical systems is said to be *universal* (or possesses the universality property) for nonlinear I/O maps with fading memory. They include various classical reservoir computing schemes such as liquid state machines [28], echo-state networks (ESNs) [18], linear reservoirs with polynomial readouts (LRPO), and non-homogeneous state-affine systems (SAS) [19]. However, a theoretical framework for the learning of nonlinear fading memory I/O maps by quantum systems is so far lacking. Moreover, an extended investigation into the potential advantage quantum systems offer over classical reservoir computing schemes has not been conducted. The provision of such a learning theory, the demonstration of a class of quantum model that is provably universal, and a study of this model via numerical experiments are the main contributions of this paper.

The paper is organized as follows. In Sec. II, we formally define fading memory maps. In Sec. III, we formulate the theory of learning nonlinear fading memory maps with dissipative quantum systems. Sec. IV introduces a concrete universal class of dissipative quantum systems. Sec. V numerically demonstrates the emulation performance of the proposed universal class in the absence and presence of decoherence. The effect of different input encodings on the learning capability of this class is investigated. An in-depth comparison between this universal class and ESNs is also conducted. We conclude this section by discussing the potential of this universal class to surpass classical schemes when implemented on a NISQ machine. In Sec. VI, we discuss the feasibility of proof-of-principle experiments of the proposed scheme on existing NISQ machines. Detailed results and numerical settings are collected in and can be found in the Appendix.

## II. FADING MEMORY MAPS

Let  $\mathbb{Z}$  denote the set of all integers and  $\mathbb{Z}^- = \{\dots, -1, 0\}$ . Let  $u = \{\dots, u_{-1}, u_0, u_1, \dots\}$  be a real bounded input sequence with  $\sup_{k \in \mathbb{Z}} |u_k| < \infty$ . We say that a real output sequence  $y = \{\dots, y_{-1}, y_0, y_1, \dots\}$  is related to  $u$  by a time-invariant causal map  $M$  if  $y_k = M(u)_k = M(\tilde{u}_\ell)_k$  for any integer  $\ell$ , any  $k \leq \ell$ , and any sequence  $\tilde{u}_\ell$  such that  $\tilde{u}_\ell|_\ell = u|_\ell$ . Here,  $M(u)_k$  denotes the output sequence at time  $k$  given the input sequence  $u$ , and  $u|_k = \{\dots, u_{k-2}, u_{k-1}, u_k\}$  is the input sequence  $u$  truncated after time  $k$ .

For a fixed real positive constant  $L$  and a compact subset  $D \subseteq \mathbb{R}$ , we are interested in the set  $K_L(D)$  consisting of input sequences such that for all  $k \in \mathbb{Z}$ ,  $u_k \in D \cap [-L, L]$ . We say a time-invariant causal map  $M$  defined on  $K_L(D)$  has the fading memory property with respect to a decreasing sequence  $w = \{w_k\}_{k \geq 0}$ ,  $\lim_{k \rightarrow \infty} w_k = 0$  if, for any two input sequences  $u$  and  $v$ ,  $|M(u)_0 - M(v)_0| \rightarrow 0$  whenever  $\sup_{k \in \mathbb{Z}^-} |w_{-k}(u_k - v_k)| \rightarrow 0$ . In other words, if the elements of two sequences agree closely up to some recent past before  $k = 0$ , then their output sequences will also be close at  $k = 0$ .

## III. LEARNING NONLINEAR FADING MEMORY MAPS WITH DISSIPATIVE QUANTUM SYSTEMS

Since fading memory maps are time-invariant, any dynamical system that is used to approximate them must forget its initial condition. Classical dynamical systems with this

property are referred to as *convergent systems* in control theory [36], and the property is known as the *echo state property* in the context of ESNs [12, 21]. For dissipative quantum systems, this means that for the same input sequence, density operators asymptotically converge to the same sequence of density operators, independently of their initial values. We emphasize that the dissipative nature of the quantum system is *essential* for the learning task. Without it the system clearly cannot be convergent.

Consider a quantum system consisting of  $n$  qubits with a Hilbert space  $\mathbb{C}^{2^n}$  of dimension  $2^n$  undergoing the following discrete-time dissipative evolution:

$$\rho_k = T(u_k)\rho_{k-1}, \quad (1)$$

for  $k = 1, 2, \dots$ , with initial condition  $\rho(0) = \rho_0$ . Here,  $\rho_k = \rho(k\tau)$  is the system density operator at time  $t = k\tau$  and  $\tau$  is a (fixed) sampling time, and  $T(u_k)$  is a completely positive trace preserving (CPTP) map for each  $u_k$ . In this setting, the real input sequence  $\{u_1, u_2, \dots\}$  determines the system's evolution. The overall input-output map in the long time limit is in general non-linear. Let  $\|\cdot\|_p$  denote any Schatten  $p$ -norm for  $p \in [1, \infty)$  defined as  $\|A\|_p = \text{Tr}(\sqrt{A^*A})^{1/p}$ , where  $A$  is a complex matrix and  $*$  is the conjugate transpose operator. In Appendix [VIII A, Theorem 3], we show that if for all  $u_k \in D \cap [-L, L]$ , the CPTP map  $T(u_k)$  restricted on the hyperplane  $H_0(2^n)$  of  $2^n \times 2^n$  traceless Hermitian operators satisfies  $\|T(u_k)|_{H_0(2^n)}\|_{2 \rightarrow 2} := \sup_{A \in H_0(2^n), A \neq 0} \frac{\|T(u_k)A\|_2}{\|A\|_2} \leq 1 - \epsilon$  for some  $0 < \epsilon \leq 1$ , then under any input sequence  $u \in K_L(D)$ , it will forget its initial condition and is therefore convergent. This means that for any two initial density operators  $\rho_{j,0}$  ( $j = 1, 2$ ) and the corresponding density operators  $\rho_{j,k}$  at time  $t = k\tau$ , we will have that

$$\lim_{k \rightarrow \infty} \|\rho_{1,k} - \rho_{2,k}\|_2 = \lim_{k \rightarrow \infty} \left\| \left( \overleftarrow{\prod}_{j=1}^k T(u_j) \right) (\rho_{1,0} - \rho_{2,0}) \right\|_2 = 0,$$

where  $\overleftarrow{\prod}_{j=1}^k$  is a time-ordered composition of maps  $T(u_j)$  from right to left.

Let  $\mathcal{D}(\mathbb{C}^{2^n})$  denote the convex set of all density operators on  $\mathbb{C}^{2^n}$ . We introduce an output sequence  $\bar{y}$  in the form

$$\bar{y}_k = h(\rho_k), \quad (2)$$

where  $h : \mathcal{D}(\mathbb{C}^{2^n}) \rightarrow \mathbb{R}$  is a real functional of  $\rho_k$ . Eqs. (1) and (2) define a quantum dynamical system with input sequence  $u$  and output sequence  $\bar{y}$ . We now require the separation property. Consider a family  $\mathcal{F}$  of distinct quantum systems described by Eqs. (1) and (2), but possibly having differing number of qubits. Let  $u$  and  $u'$  be two input sequences in  $K_L(D)$  that are not identical,  $u_k \neq u'_k$  for at least one  $k$ , and let  $\bar{y}$  and  $\bar{y}'$  be the respective outputs of the quantum system for these inputs. We say that the family  $\mathcal{F}$  is *separating* if for any non-identical inputs  $u$  and  $u'$  in  $K_L(D)$ , there exists a member in this family with non-identical outputs  $\bar{y}$  and  $\bar{y}'$ . As stated in Appendix [VIII B, Theorem 9], any family of convergent dissipative quantum systems that implement fading memory maps with the separation property, and which forms an algebra of maps containing the constant maps, is universal and can approximate any I/O map with fading memory arbitrarily closely.

#### IV. A UNIVERSAL CLASS OF DISSIPATIVE QUANTUM SYSTEMS

We now specify a class of dissipative quantum systems that is provably universal in approximating fading memory maps defined on  $K_1([0, 1])$ . The class consists of systems

that are made up of  $N$  *non-interacting* subsystems initialized in a product state of the  $N$  subsystems, with subsystem  $K$  consisting of  $n_K + 1$  qubits,  $n_K$  “system” qubits and a single “ancilla” qubit. We label the qubits of subsystem  $K$  by an index  $i_j^K$  that runs from  $j = 0$  to  $j = n_K$ , with  $i_0^K$  labeling the ancilla qubit. The  $n_K + 1$  qubits interact via the Hamiltonian

$$H_K = \sum_{j_1=0}^{n_K} \sum_{j_2=j_1+1}^{n_K} J_K^{j_1, j_2} (X^{(i_{j_1}^K)} X^{(i_{j_2}^K)} + Y^{(i_{j_1}^K)} Y^{(i_{j_2}^K)}) + \sum_{j=0}^{n_K} \alpha Z^{(i_j^K)},$$

where  $J_K^{j_1, j_2}$  and  $\alpha$  are real-valued constants, while  $X^{(i_j^K)}$ ,  $Y^{(i_j^K)}$  and  $Z^{(i_j^K)}$  are Pauli  $X$ ,  $Y$  and  $Z$  operators of qubit  $i_j^K$ . The ancilla qubits for all subsystems are periodically reset at time  $t = k\tau$  and prepared in the input-dependent mixed state  $\rho_{i_0, k}^K = u_k|0\rangle\langle 0| + (1 - u_k)|1\rangle\langle 1|$  (with  $0 \leq u_k \leq 1$ ). The system qubits are initialized at time  $t = 0$  to some density operator. The density operator  $\rho_k^K$  of the  $K^{\text{th}}$  subsystem qubits evolves during time  $(k-1)\tau < t < k\tau$  according to  $\rho_k^K = T_K(u_k)\rho_{k-1}^K$ , where  $T_K(u_k)$  is the CPTP map defined by  $T_K(u_k)\rho_{k-1}^K = \text{Tr}_{i_0^K} \left( e^{-iH_K\tau} \rho_{k-1}^K \otimes \rho_{i_0, k}^K e^{iH_K\tau} \right)$  and  $\text{Tr}_{i_0^K}$  denotes the partial trace over the ancilla qubit of subsystem  $K$ . We now specify an output functional  $h$  associated with this system. We will use a single index to label the system qubits from the  $N$  subsystems, the ancilla qubits are not used in the output. Consider an individual system qubit with index  $j$ , with  $j$  running from 1 until  $n = \sum_{K=1}^N n_K$ . The output functional  $h$  is defined to be of the general form,

$$\bar{y}_k = h(\rho_k) = C + \sum_{d=1}^R \sum_{i_1=1}^n \sum_{i_2=i_1+1}^n \cdots \sum_{i_n=i_{n-1}+1}^n \sum_{r_{i_1}+\cdots+r_{i_n}=d} w_{i_1, \dots, i_n}^{r_{i_1}, \dots, r_{i_n}} \langle Z^{(i_1)} \rangle_k^{r_{i_1}} \cdots \langle Z^{(i_n)} \rangle_k^{r_{i_n}} \quad (3)$$

where  $C$  is a constant,  $R$  is an integer and  $\langle Z^{(i)} \rangle_k = \text{Tr}(\rho_k Z^{(i)})$  is the expectation of the operator  $Z^{(i)}$ . We note that the functional  $h$  (the right hand side of the above) is a multivariate polynomial in the variables  $\langle Z^{(i)} \rangle_k$  ( $i = 1, \dots, n$ ) and these expectation values depend on input sequence  $u = \{u_k\}$ . Thus computing  $\bar{y}_k$  only involves estimating the expectations  $\langle Z^{(i)} \rangle_k$  and the degree of the polynomial  $R$  can be chosen as desired. If  $R = 1$  then  $\bar{y}_k$  is a simple linear function of the expectations.

This family of dissipative quantum systems exhibits two important properties, see Appendix VIII C and VIII D for the proofs. Firstly, if for each subsystem  $K$  with  $n_K$  qubits and for all  $u_k \in [0, 1]$ ,  $\|T_K(u_k)\|_{H_0(2^{n_K})} \leq 1 - \epsilon_K$  for some  $0 < \epsilon_K \leq 1$ , then this family forms a polynomial algebra consisting of systems that implement fading memory maps. Secondly, a convergent single-qubit system with a linear output combination of expectation values (ie.  $n = 1$ ,  $N = 1$  and  $R = 1$ ), separates points of  $K_1([0, 1])$ . These two properties and an application of the Stone-Weierstrass Theorem [13, Theorem 7.3.1] guarantee the universality property.

The class specified above is a variant of the QRC model in [17] but is provably universal by the theory of the previous section. The differences are in the general form of the output and, in our model, the ancilla qubit is not used in computing the output. Also, we do not consider time-multiplexing. We remark that time-multiplexing can be in principle incorporated in the model using the same theory. However, this extension is more technical and will be pursued elsewhere.

## V. NUMERICAL EXPERIMENTS

We demonstrate the emulation performance of the universal class introduced above in learning a number of benchmarking tasks. A random input sequence  $u^{(r)} = \{u_k^{(r)}\}_{k>0}$ , where each  $u_k^{(r)}$  is randomly uniformly chosen from  $[0, 0.2]$ , is applied to all computational tasks. We apply the multitasking method, in which we simulate the evolution of the quantum systems and record the expectations  $\langle Z^{(i)} \rangle_k$  for all timesteps  $k$  once, while the output weights  $C$  and  $w_{i_1, \dots, i_n}^{r_{i_1}, \dots, r_{i_n}}$  in Eq. (3) are optimized independently for each computational task.

The linear reservoirs with polynomial outputs (LRPO) implement a fading memory map, whose discrete-time dynamics is of the form [10, 19],

$$\begin{cases} x_k = Ax_{k-1} + cu_k \\ y_k = \hat{h}(x_k), \end{cases}$$

where we choose  $c \in \mathbb{R}^{1400}$  with elements randomly uniformly chosen from  $[0, 4]$  and  $\hat{h}$  to be a degree two multivariate polynomial, whose coefficients are randomly uniformly chosen from  $[-0.1, 0.1]$ . We choose  $A$  to be a diagonal block matrix  $A = \text{diag}(A_1, A_2, A_3)$ , where  $A_1, A_2$  and  $A_3$  are  $200 \times 200$ ,  $500 \times 500$  and  $700 \times 700$  real matrices, respectively. Elements of  $A_i$  ( $i = 1, 2, 3$ ) are randomly uniformly chosen from  $[0, 4]$ . To ensure the convergence and the fading memory property, the maximum singular value of each  $A_i$  is randomly uniformly set to be  $\sigma_{\max}(A_i) < 1$  [19]. In this setting, each linear reservoir defined by  $A_i$  evolves independently, while the output of the LRPO depends on all state elements  $x_k \in \mathbb{R}^{1400}$ .

It is interesting to investigate the performance of the universal class in learning tasks that do not strictly implement fading memory maps as defined here. We apply the universal class to approximate the outputs of a missile moving with a constant velocity in the horizontal plane [33] and the nonlinear autoregressive moving average (NARMA) models [6]. The nonlinear dynamics of the missile is given by

$$\begin{cases} \dot{x}_1 = x_2 - 0.1 \cos(x_1)(5x_1 - 4x_1^3 + x_1^5) - 0.5 \cos(x_1)\tilde{u} \\ \dot{x}_2 = -65x_1 + 50x_1^3 - 15x_1^5 - x_2 - 100\tilde{u} \end{cases}$$

where  $y = x_2$  is the output. We make a change of variable of the input  $\tilde{u} = 5u - 0.5$  so that the input range is the same as in [33]. The missile dynamics is simulated by the Runge-Kutta (4, 5) formula implemented by the ode45 function in MATLAB [14], with a sampling time of  $4 \times 10^{-4}$  seconds for a time span of 1 second, subject to the initial condition  $(x_1 \ x_2)^T = (0 \ 0)^T$ . We denote this task as Missile. The NARMA models are often used to benchmark algorithms for learning time-series. The outputs of each NARMA model depend on its time-lagged outputs and inputs, specified by a delay  $\tau_{\text{NARMA}}$ . We denote the corresponding task to be  $\text{NARMA}_{\tau_{\text{NARMA}}}$ .

We focus on members of the universal class with a single subsystem ( $N = 1$ ) and a small number of system qubits  $n = \{2, 3, 4, 5, 6\}$ , and denote this subset of the universal class as SA. We will drop the subsystem index  $K$  from now on. For all numerical experiments, the parameters of SA are chosen as follows. We introduce a scale  $S > 0$  such that the Hamiltonian parameters  $J^{j_1, j_2}/S$ ,  $\alpha/S = 0.5$  and  $\tau S = 1$  are dimensionless. As for the QRCs in [17], we randomly uniformly generate  $J^{j_1, j_2}/S$  from  $[-1, 1]$  and, to ensure convergence, select the resulting Hamiltonians for experiments if the associated CPTP map is convergent. We numerically test the convergence property by checking if 50 randomly generated initial

density operators converge to the same density operator in 500 timesteps under the input sequence  $u^{(r)}$ .

Each numerical experiment firstly washouts the effect of initial conditions of SA and all target maps with 500 timesteps. This is followed by a training stage of 1000 timesteps, where we optimize the output weights  $C$  and  $w_{i_1, \dots, i_n}^{r_{i_1}, \dots, r_{i_n}}$  of SA by standard least squares to minimize the error  $\sum_{k=501}^{1500} |y_k - \bar{y}_k|^2$  between the target output sequence  $y$ . In practical implementation, computation of the expectations  $\langle Z^{(j)} \rangle_k$  is offloaded to the quantum subsystem, and only a simple classical processing method is needed to optimize the output weights. For this reason, we associate the output weights  $C$  and  $w_{i_1, \dots, i_n}^{r_{i_1}, \dots, r_{i_n}}$  in Eq. (3) with *(classical) computational nodes*, with the number of such nodes being equal to the number of output weights. While the number of computational nodes for SA can be chosen arbitrarily by varying the degree  $R$  in the output, the state-space ‘size’ of the quantum system is  $2^n(2^n + 1) - 2^n = 4^n$ . This state-space size corresponds to the number of real variables needed to describe the evolution of elements of the system density operator. Note that since the density operator has unity trace, only up to at most  $4^n - 1$  of these nodes are linearly independent.

On the other hand, for ESNs [21], the number of computational nodes and the state-space size always differ by one (i.e. by the tunable constant output weight). For an ESN with  $m$  reservoir nodes ( $Em$ ), the number of computational nodes is  $m + 1$  and its state-space size is  $m$ . For the Volterra series [10] with kernel order  $o$  and memory  $p$  ( $Vo, p$ ), the number of computational nodes is  $(\frac{p^{o+1}-p}{p-1} + 1)$ . We select  $m$  and  $(o, p)$  such that the number of computational nodes is at most 801. This reduces the chance of overfitting for learning a sequence of length 1000 [25]. For detailed numerical settings for ESNs and the Volterra series, see Appendix VIIIE. We analyze the performance of all learning schemes during an evaluation phase consisting of 1000 timesteps, using the normalized mean-squared error  $\text{NMSE} := \sum_{k=1501}^{2500} |\bar{y}_k - y_k|^2 / \sum_{k=1501}^{2500} |y_k - \frac{1}{1000} \sum_{k=1501}^{2500} y_k|^2$ , where  $y$  is the target output and  $\bar{y}$  is the approximated output. For each task and each  $n$ , NMSEs of 100 convergent SA samples are averaged for analysis.

### A. Overview of SA learning performance

We present an overview of SA performance in learning the LRPO, Missile, NARMA15 and NARMA20 tasks. The degree of the multivariate polynomial output Eq. (3) is fixed to be  $R = 1$ , so that the number of computational is  $n + 1$  for each  $n$ . Fig. 1 shows the typical SA outputs for the LRPO, Missile, NARMA15 and NARMA20 tasks during the evaluation phase. It is observed that the SA outputs follow the LRPO outputs closely, while SA is able to approximate the Missile and NARMA tasks relatively closely. For all computational tasks, as the number of system qubits  $n$  increases, the SA outputs better approximate the target outputs. This is quantitatively demonstrated in Fig. 2, which plots the average SA NMSE as  $n$  increases.

From Fig. 2 we can see that the SA model with a small number of computational nodes performs comparably as ESNs and the Volterra series with a large number of computational nodes. For example, the average NMSE of 6-qubit SA with 7 computational nodes is comparable to the average NMSE of E100 with 101 computational nodes in the LRPO task. On average, 5-qubit SA with 6 computational nodes performs better than V2, 22 with 504 computational nodes in the Missile task. In the NARMA15 task, 4-qubit SA with 5 computational nodes outperforms V2, 4 with 21 computational nodes. In the NARMA20

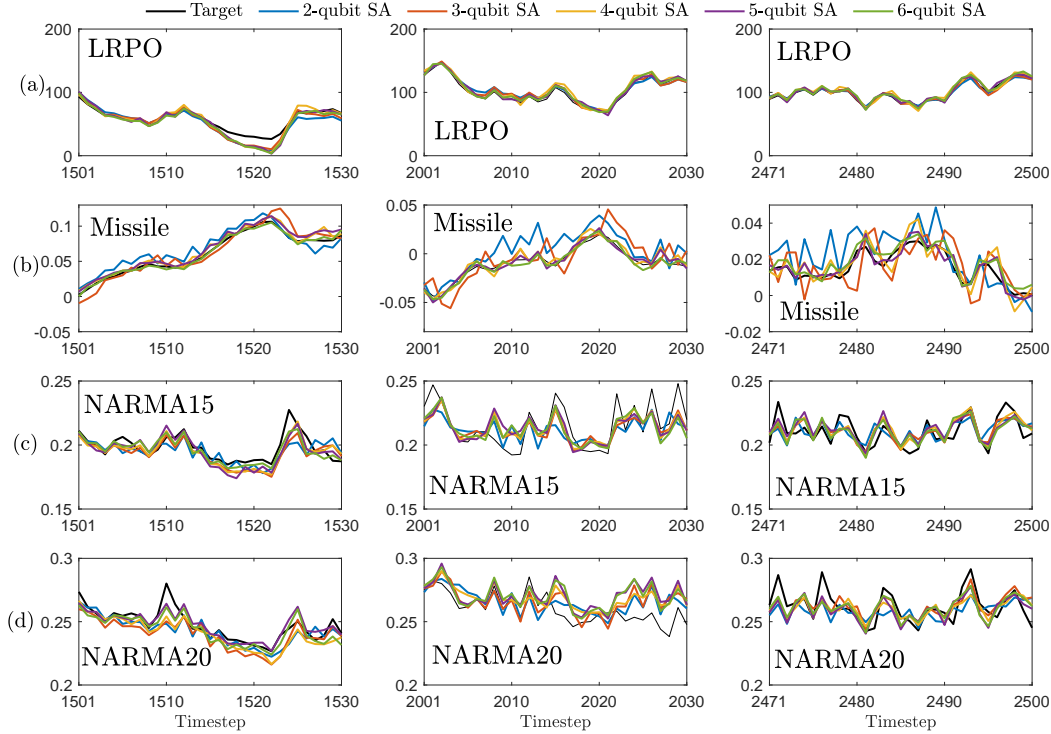


FIG. 1. Typical SA outputs during the evaluation phase, for the (a) LRPO, (b) Missile (c) NARMA15 and (d) NARMA20 tasks. The leftmost, middle and rightmost panels show the outputs for timesteps 1501-1530, 2001-2030 and 2471-2500, respectively

task, 5-qubit SA performs comparably as E600. Our results are similar to the performance of the QRCs with time multiplexing reported in [17], where the QRCs are demonstrated to perform comparably as ESNs with a larger number of trainable computational nodes. However, for the small number of qubits investigated, the rate of decrease in the average NMSE is approximately linear despite the dimension of the Hilbert space increases exponentially as  $n$  increases. For both the NARMA tasks, the average NMSEs for 2-qubit and 6-qubit SA are of the same order of magnitude. A larger number of additional system qubits are required to substantially reduce the SA task error.

## B. SA performance under decoherence

We further validate the feasibility of the SA model in the presence of the dephasing, decaying noise and the generalized amplitude damping (GAD) channel. We simulate the noise by applying the Trotter-Suzuki formula [42, 44], in which we divide the normalized time interval  $\tau S = 1$  into 50 small time intervals  $\delta_t = \tau S/50$ , and alternatively apply the unitary interaction and the Kraus operators  $\{M_l^{(j)}(\frac{\gamma}{S})\}_l$  of each noise type, each for a time duration of  $\delta_t$ . Each of the  $l$ -th Kraus operator is applied for all system and ancilla qubits  $j = 1, \dots, n+1$ , and  $\gamma/S$  denotes the noise strength. For all noise types, we apply the same noise strengths  $\gamma/S = \{10^{-4}, 10^{-3}, 10^{-2}\}$ , which are within the experimentally feasible range for systems like NMR ensembles [45] and some current superconducting NISQ machines [1]. Under the dephasing noise, the density operator  $\rho$  of the system and ancilla qubits evolves



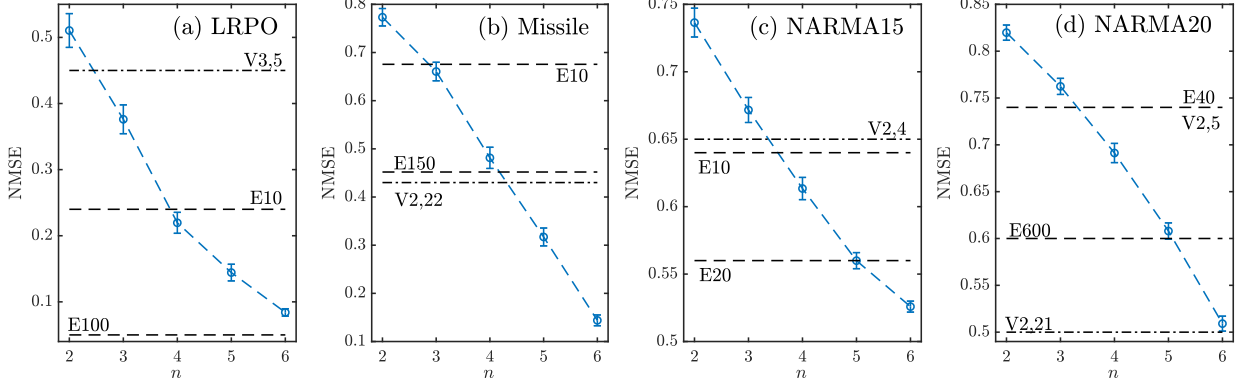


FIG. 2. Average SA NMSE for the (a) LRPO, (b) Missile, (c) NARMA15 and (d) NARMA20 tasks, the error bars represent the standard error. For comparison, horizontal dashed lines labeled with “Em” indicate the average performance of ESNs with  $m$  computational nodes, and horizontal dot-dashed lines labeled with “Vo,  $p$ ” indicates the performance of Volterra series with kernel order  $o$  and memory  $p$ . Overlapping dashed and dot-dashed lines are represented as dashed lines

according to  $\rho \rightarrow \frac{1+e^{-2\frac{\gamma}{S}\delta t}}{2}\rho + \frac{1-e^{-2\frac{\gamma}{S}\delta t}}{2}Z^{(j)}\rho Z^{(j)\dagger}$ , such that the diagonal elements in  $\rho$  remain invariant while the off-diagonal elements decay. The GAD channel gives rise to the evolution  $\rho \rightarrow \sum_{l=0}^3 M_l^{(j)}(\frac{\gamma}{S}, \lambda)\rho(M_l^{(j)}(\frac{\gamma}{S}, \lambda))^\dagger$ , where  $\dagger$  denotes the adjoint, and the Kraus operators  $M_l^{(j)}(\frac{\gamma}{S}, \lambda)$  ( $l = 0, 1, 2, 3$ ) depend on an additional finite temperature parameter  $\lambda \in [0, 1]$  [34]. When  $\lambda = 1$ , we recover the amplitude damping channel (the decaying noise), which takes a mixed state into the pure ground state  $|0\rangle\langle 0|$  in the long time limit. For  $\lambda \neq 1$ , we investigate the SA task performance under the GAD channel for  $\lambda = \{0.2, 0.4, 0.6, 0.8\}$ . The GAD channel affects both the diagonal and off-diagonal elements of the density operator.

Fig. 3 plots the average SA NMSE under the dephasing, decaying and GAD with  $\lambda = \{0.4, 0.6\}$  for all noise strengths. See Appendix VIII E 2 for the average NMSE under the GAD channel for all chosen temperature parameters. Fig. 3 indicates that for the same noise strength, different noise types affect the SA task performance in a similar manner. For noise strengths  $\gamma/S = \{10^{-4}, 10^{-3}\}$ , all noise types do not significantly degrade SA task performance for the computational tasks. However, the impact of the noise strength  $\gamma/S = 10^{-2}$  is more pronounced, particularly for a larger number of system qubits.

Changes in the SA task error under the effect of the decaying noise and the GAD channel are anticipated, since the expectations  $\langle Z^{(j)} \rangle_k$  in the output depend on the diagonal elements of the system density operator, which are affected by both of these noise types. However, the SA task performance is also affected by the dephasing noise, which does not change the diagonal elements. A possible explanation for this behavior is a loss of degrees of freedom, in the sense that off-diagonal elements of the density operator become smaller and the density operator looks more like a classical probability distribution. Alternatively, this could be viewed as the off-diagonal elements contributing less to the overall computation. To support this explanation, for the dephasing, decaying and the GAD with  $\lambda = \{0.4, 0.6\}$ , and for each  $n$ , we sum the complex modulus of off-diagonal elements in the system density operator for the 100  $n$ -qubit SA samples simulated above. The average of these 100 sums is plotted for the first 50 timesteps during the evaluation phase in Fig. 4. That is, Fig. 4 plots  $\frac{2}{n_s} \sum_{l=1}^{n_s} \sum_{r=1}^{2^n} \sum_{s=r+1}^{2^n} |\rho_{k,rs}^{(l)}|$ , where  $n_s = 100$  is the number of different random SA samples.

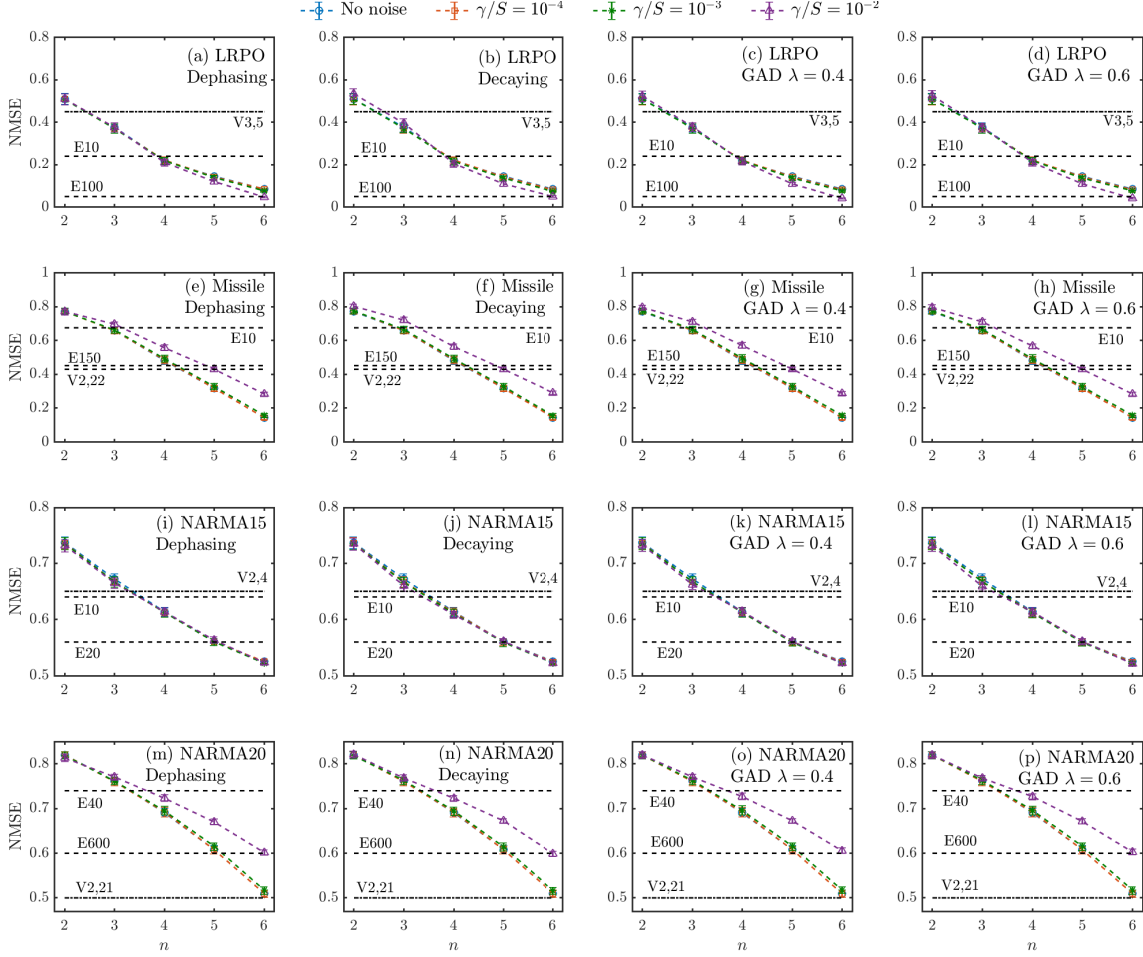


FIG. 3. Average SA NMSE for the LRPO, Missile, NARMA15 and NARMA20 tasks under decoherence. For comparison, the average SA NMSE without the effect of noise is also plotted. In all plots, the error bars represent the standard error

Here  $\rho_{k,rs}^{(l)}$  denotes the element of  $\rho_k^{(l)}$  in row  $r$  and column  $s$  (the superscript  $(l)$  indexing the SA sample).

Fig. 4 shows that as the noise strength increases, the average sum decreases, particularly with the noise strength  $\gamma/S = 10^{-2}$ . Similar trends are observed for the GAD channel for all the temperature parameters chosen, and the observed trend for the average sum persists as the timestep increases to 2500 (see Appendix VIII E 2). The results presented in Fig. 4 further indicate that though the output of SA depends solely on the diagonal elements of the density operator, nonzero off-diagonal elements are crucial for improving the SA emulation performance. This provides a plausible explanation for the improved performance achieved by increasing the number of qubits, thereby increasing Hilbert space size and the number of non-zero off-diagonal elements. Further investigation into this topic is presented in Sec. V D.

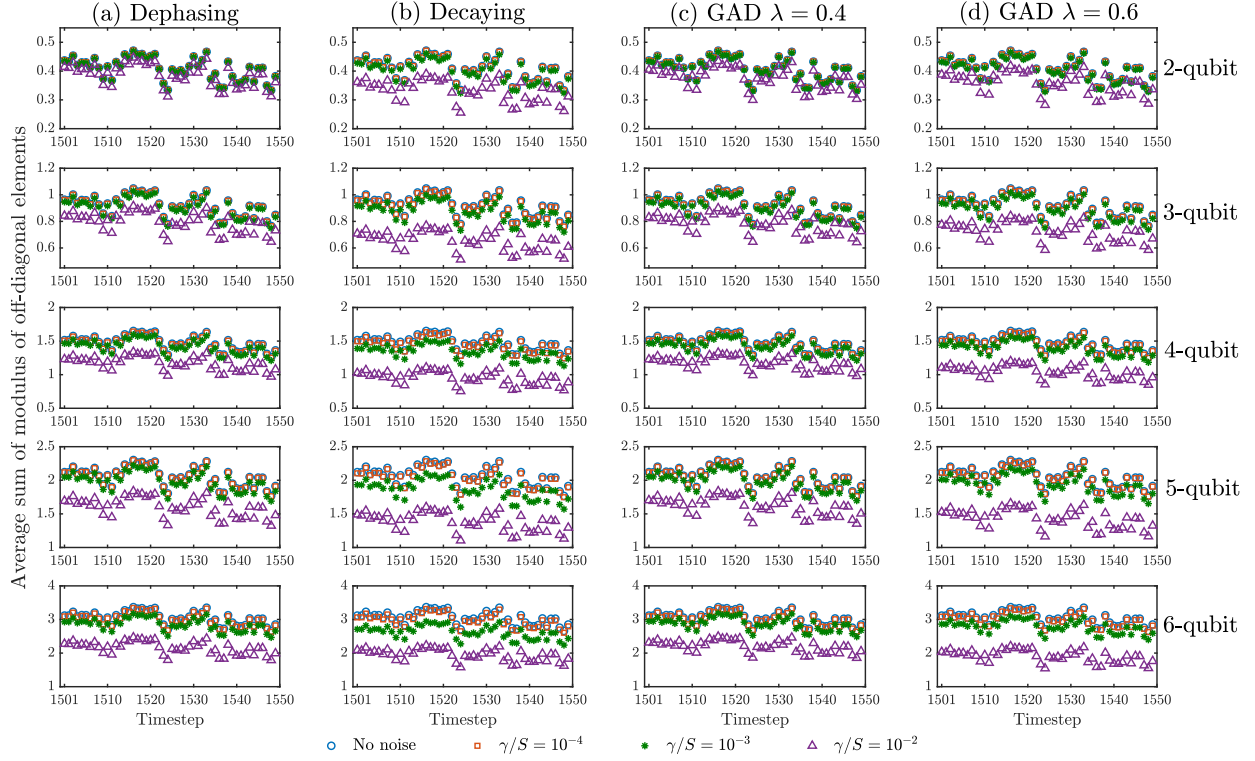


FIG. 4. Average sum of complex modulus of off-diagonal elements in the system density operator for timesteps 1501-1550, under the (a) dephasing noise, (b) decaying noise, (c) GAD with  $\lambda = 0.4$  and (d) GAD with  $\lambda = 0.6$ . Row  $n - 1$  in the figure corresponds to the average sum for  $n$ -qubit SA

### C. Effect of different input encodings

Our proposed universal class encodes the input  $u_k \in [0, 1]$  into the mixed state  $\rho_{i_0,k} = u_k|0\rangle\langle 0| + (1 - u_k)|1\rangle\langle 1|$ . Other input encoding possibilities include the pure state  $\rho_{i_0,k} = (\sqrt{u_k}|0\rangle + \sqrt{1 - u_k}|1\rangle)(\sqrt{u_k}\langle 0| + \sqrt{1 - u_k}\langle 1|)$  used in the QRC model [17], encoding the input into the phase  $\rho_{i_0,k} = \frac{1}{2}(|0\rangle + e^{-iu_k}|1\rangle)(\langle 0| + e^{iu_k}\langle 1|)$ , and encoding the input into non-orthogonal basis state  $\rho_{i_0,k} = u_k|0\rangle\langle 0| + \frac{1 - u_k}{2}(|0\rangle + |1\rangle)(\langle 0| + \langle 1|)$ . We denote these different input encodings as mixed, pure, phase and non-orthogonal. We emphasize that for the last three encodings the universality of the associated dissipative quantum system using these encodings has not been proven.

To investigate the impact of input encodings on the computational capability of quantum systems, the Hamiltonian parameters for all quantum systems simulated here are sampled from the same uniform distribution, and the resulting Hamiltonians are chosen if the associated CPTP map that implements the specified input-dependent density operator  $\rho_{i_0,k}$  is convergent. We again test the convergence property numerically by checking if 50 randomly generated initial density operators converge to the same density operator within 500 timesteps. The number of system qubits and the number of computational nodes for all input encodings are the same. For each input encoding, NMSEs of 100 convergent quantum systems are averaged for analysis. Fig. 5 shows that for all computational tasks, the mixed state encoding performs better than other encodings. However, the average NMSE

for different input encodings for all computational tasks are of the same order of magnitude. Moreover, as the number of system qubits increases, the errors of different input encodings decrease at roughly the same rate. This comparison indicates that the effect of different input encodings on the learning performance does not appear significant.

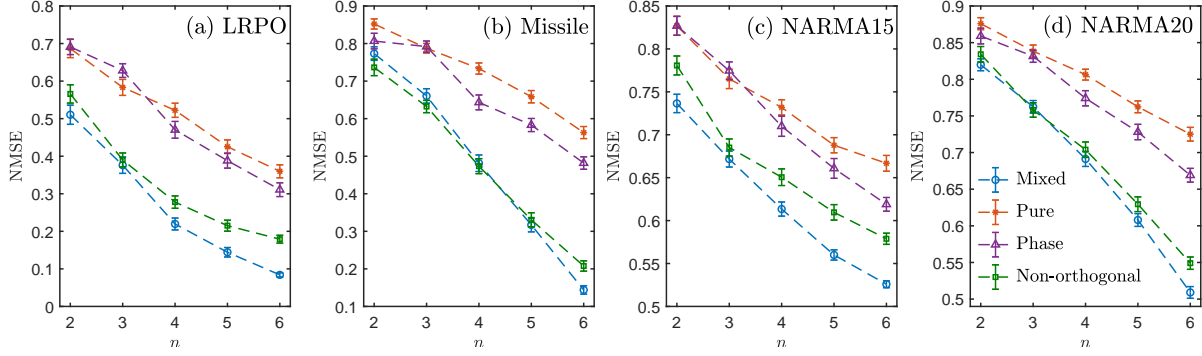


FIG. 5. Average NMSE for different input encodings, for approximating the (a) LRPO, (b) Missile (c) NARMA15 and (d) NARMA20 tasks. Error bars represent the standard error

#### D. Further comparison with ESNs

Our numerical results so far and the results shown in [17] both suggest that dissipative quantum systems with a small number of qubits achieve comparable performance to classical learning schemes with a large number of computational nodes. However, these comparisons may appear to be skewed favorably toward quantum dynamical systems, since it does not address their exponential state-space size. One can, for example, also increase the state-space size of ESNs and the number of computational nodes of SA, such that the state-space size and the number of computational nodes are similar for both models. Here we present a further comparison between the SA model and ESNs, and provide insights into the possible advantage the SA model might offer over its classical counterpart.

We focus on 4-qubit SA with a state-space size of 256. Setting  $R = 6$  in Eq. (3), the number of computational nodes for SA is 210. We compare this 4-qubit SA model's average task performance with the average E256 task performance in approximating the LRPO, Missile, NARMA15, NARMA20, NARMA30 and NARMA40 tasks. Here, the number of computational nodes for E256 is 257 and the average NMSE of 100 convergent E256s is reported. As shown in Table 1, for the Missile and all the NARMA tasks, the average NMSEs for both models are of the same order of magnitude, while E256 outperforms SA in the LRPO task. This comparison suggests that when the state-space size and the number of computational nodes for both models are similar, ESNs can outperform the SA model.

We further investigate under what conditions SA might offer a computational advantage. We observe that while the number of computational nodes is kept constant, increasing the state-space size of SA induces a considerable computational improvement. To demonstrate this, 4-, 5- and 6-qubit SA samples are simulated to perform all computational tasks mentioned above. For each  $n$ -qubit SA, we vary its output degree  $R$  such that its number of computational nodes ranges from 5 to 252. The chosen degrees for 4-qubit SA are  $R_4 = \{1, \dots, 6\}$ , for 5-qubit are  $R_5 = \{1, \dots, 5\}$ , and for 6-qubit SA are  $R_6 = \{1, \dots, 4\}$ .

TABLE 1. Average 4-qubit SA and E256 NMSE for the LRPO, Missile, NARMA15, NARMA20, NARMA30 and NARMA40 tasks. Results are rounded to two significant figures. The notation ( $\pm$  se) denotes the standard error

Task	SA NMSE ( $\pm$ se)	E256 NMSE ( $\pm$ se)
LRPO	$0.20 \pm 1.5 \times 10^{-2}$	$0.019 \pm 7.7 \times 10^{-4}$
Missile	$0.48 \pm 2.2 \times 10^{-2}$	$0.49 \pm 3.3 \times 10^{-3}$
NARMA15	$0.61 \pm 8.0 \times 10^{-3}$	$0.32 \pm 1.6 \times 10^{-4}$
NARMA20	$0.68 \pm 1.0 \times 10^{-2}$	$0.67 \pm 3.2 \times 10^{-4}$
NARMA30	$0.67 \pm 7.1 \times 10^{-3}$	$0.67 \pm 4.0 \times 10^{-4}$
NARMA40	$0.64 \pm 5.3 \times 10^{-3}$	$0.66 \pm 5.9 \times 10^{-4}$

For each  $n$ -qubit SA, the Hamiltonians are the same for all its chosen output degrees, and the task errors of 100 convergent SA samples are averaged for analysis.

For comparison, we simulate 100 convergent ESNs with reservoir size 256 to perform the same tasks. For  $n$ -qubit SA, let  $\mathcal{N}_n$  ( $n = 4, 5, 6$ ) denote the numbers of computational nodes corresponding to its output degrees  $R_n$ . The number of computational nodes  $\mathcal{C}$  for E256 is set to be elements in the set  $\mathcal{N}_4 \cup \mathcal{N}_5 \cup \mathcal{N}_6$ . We first optimize 257 output weights for E256 via standard least squares during the training phase. When  $\mathcal{C} < 257$  for E256, we select  $\mathcal{C} - 1$  computational nodes (excluding the tunable constant computational node) with the largest absolute values and their corresponding state elements. These  $\mathcal{C} - 1$  state elements are used to re-optimize  $\mathcal{C}$  computational nodes (including the tunable constant computational node) via standard least squares. During the evaluation phase, 256 state elements evolve; only  $\mathcal{C} - 1$  state elements and  $\mathcal{C}$  output weights are used to compute the E256 output.

Fig. 6 plots the 4-, 5-, and 6-qubit SA average NMSE as the number of computational nodes increases for all computational tasks. For comparison, the average E256 NMSE is also plotted. Two important observations are that increasing the number of computational nodes does not necessarily improve SA task performance, while increasing the state-space size induces a noticeable improvement. For example, for the NARMA20 task and 210 computational nodes, the average NMSE for 4-qubit SA is 0.68 while the average NMSE for 6-qubit SA is 0.48. When comparing to E256, we observe that for most tasks, despite 4-qubit SA might not perform better than E256, subsequent increases in the state-space size allow the SA model to outperform E256, without extensively increasing its number of computational nodes.

Contrary to the above observations for the SA model, increasing the reservoir size of ESNs while keeping the number of computational nodes fixed does not induce a significant computational improvement. To numerically demonstrate this, the reservoir size of ESNs is further increased to  $\{300, 400, 500\}$ . For each reservoir size, the number of computational nodes is set to be the same as that of E256. These computational nodes are chosen and optimized by the same method described above for E256. We average the task errors of 100 convergent ESNs for each reservoir size. As shown in Fig. 7, noticeable performance improvements for ESNs are only observed as the number of computational nodes increases, but not as the reservoir size varies. Another observation is that for the NARMA30 and NARMA40 tasks, the error increases as the number of computational nodes for ESNs increases. This could be due to overfitting, a condition occurs when too many adjustable parameters are trained on limited training data [16]. On the other hand, this observation is less significant for the SA

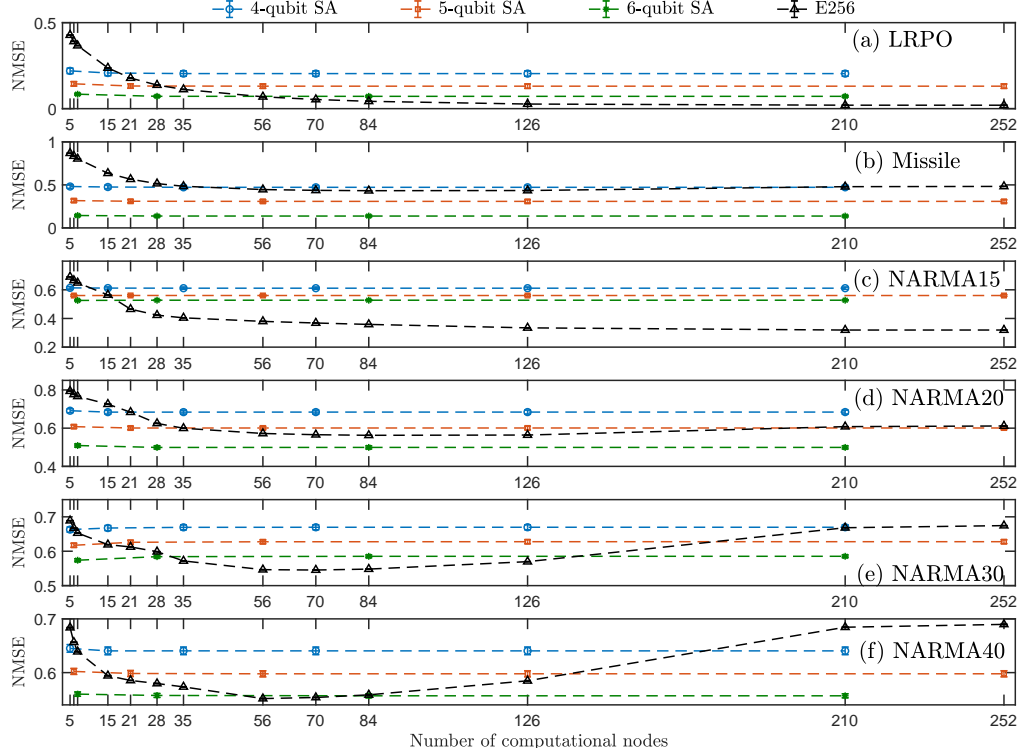


FIG. 6. Average SA NMSE as the state-space size and the number of computational nodes vary for all computational tasks. The average NMSE for E256 with the same number of computational nodes is plotted for comparison. The data symbols obscure the error bars, which represent the standard error

model. It would be interesting to conduct further investigation into this behavior in future work.

The above observations have several implications. To improve the computational capability of the SA model, one can take advantage of the exponentially increasing state-space size of the Hilbert space while only optimizing a polynomial number of computational nodes. On the contrary, to improve the computational capability of ESNs, one needs to increase the number of computational nodes, which is bounded by the reservoir size. Therefore, enhancing emulation performance of ESNs inevitably requires the state-space size to be increased. In the situation where the state-space size increases beyond what classical computers can simulate in a reasonable amount of time and with reasonable resources (such as memory), the computational capability of ESNs saturates, whereas the computational capability of the SA model could be further improved by increasing the number of qubits in a linear fashion. In this regime, the SA model could provide a potential computational advantage over its classical counterpart. To further verify the feasibility of this hypothesis, the learning capability of the SA model would need to be evaluated for a larger number of qubits on a physical quantum system. A possible implementation of this experiment is on NMR ensembles, as suggested in [17]. However, motivated by the availability of NISQ machines, a quantum circuit implementation of the SA model, using the schemes proposed in [9, 20], would be more attractive. This is another topic of further research continuing from this work.



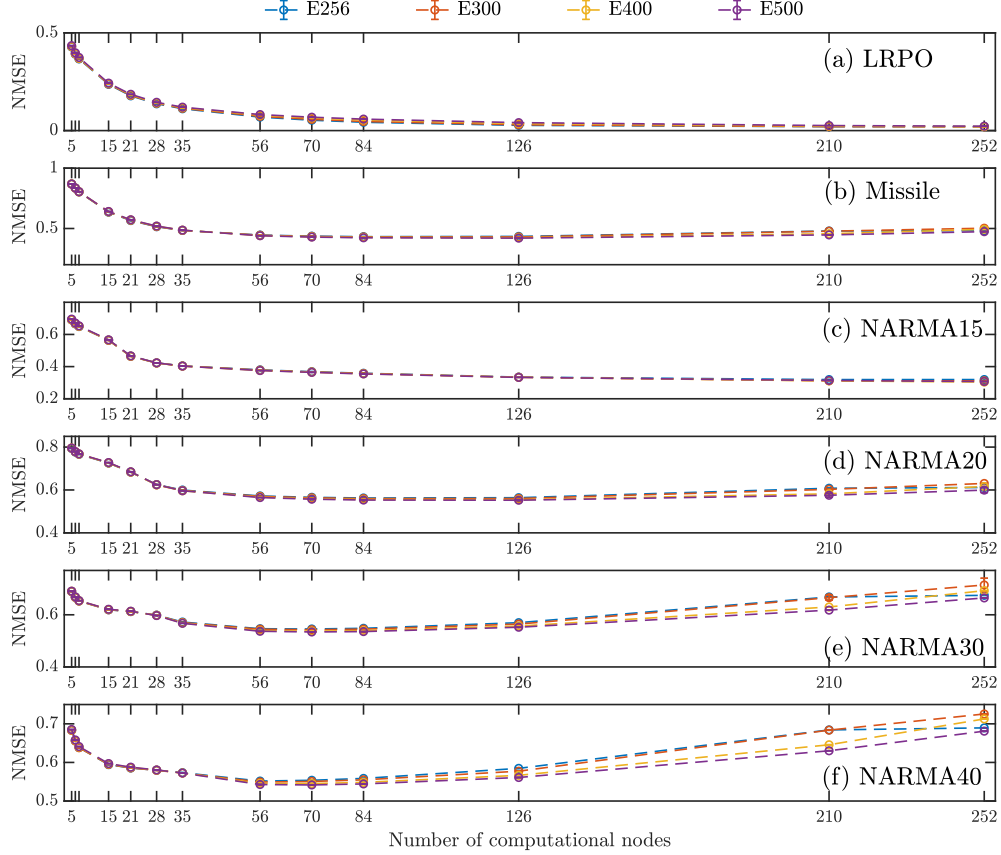


FIG. 7. Average ESNs NMSE as the state-space size and the number of computational nodes vary for all computational tasks. The data symbols obscure the error bars, which represent the standard error

## VI. DISCUSSION

We discuss the feasibility of realizing the proposed universal scheme in Section IV on the current most scalable NISQ quantum computers, such as quantum computers based on superconducting circuits or ion traps. We consider those that implement the quantum circuit model. Simulating the unitary interaction given by the Ising Hamiltonian  $H_K$  on a quantum circuit requires decomposition of the evolution using the Trotter-Suzuki product formula [42, 44]. Such a decomposition may require the sequential application of a large number of gates on NISQ machines, limiting the implementability of this family on current NISQ machines due to severe decoherence. However, it may be possible to engineer alternative families based on simpler unitary interactions between the subsystem and ancilla qubits (not of the Ising type), using only a short sequence of single-qubit and two-qubit gates, such that the associated CPTP maps possess the convergence and fading memory properties. A general framework for constructing such unitary interactions is the subject of on-going and future work continuing from this paper.

To realize the dissipative dynamics for a subsystem, we can construct a quantum circuit as follows. At each timestep  $k$ , the ancilla qubit is prepared as the input-dependent mixed state  $\rho_{i_0,k}^K$ . After the unitary interaction with the subsystem qubits, the partial trace over

the ancilla qubit can be performed by a projective measurement on the ancilla qubit and discarding the measurement outcome. At the next time step  $k + 1$ , the ancilla qubit is reset and prepared as  $\rho_{i_0, k+1}^K$ . To estimate the expectations  $\langle Z^{(j)} \rangle_k$ , we perform Monte Carlo estimation by running the circuit multiple times and measuring  $Z^{(j)}$  at time  $k$ , the average of measured results over these runs estimates  $\langle Z^{(j)} \rangle_k$ . If multiple NISQ machines can be run in parallel at the same time, the expectations  $\langle Z^{(j)} \rangle_k$  can be estimated in real time. In this setting, the number of qubits required to implement a dissipative subsystem for temporal learning is  $n_K + 1$ .

Some existing NISQ machines based on superconducting circuits are not capable of preparing mixed states or resetting qubits for reuse after measurement. To address the first limitation, we can approximate the ancilla input-dependent mixed state  $\rho_{i_0, k}^K$  by Monte Carlo sampling. That is, we construct  $M > 0$  quantum circuits as above, but for each circuit and at each timestep  $k$ , we prepare the ancilla qubit in  $|0\rangle$  with probability  $u_k$  or in  $|1\rangle$  with probability  $1 - u_k$ . We again remark that these  $M$  quantum circuits can be run in parallel, and therefore computations could be performed in real time if multiple circuits can be run at the same time. Not being able to reuse a qubit after a measurement means that each point in the input sequence must be encoded in a distinct qubit. This implies that the length of sequences that can be considered will be limited by the number of qubits available. Some of the qubits available will need to be assigned as the system qubits while all the other qubits as data carrying ancilla qubits. For instance, on a 20-qubit machine, one can use say 4 qubits as the system qubits and the remaining 16 qubits for carrying the input sequence. In this case the total input sequence length that can be processed for washout, learning and evaluation is only of length 16. Nevertheless, current high-performance quantum circuit simulators, such as the IBM Qiskit simulator (<https://qiskit.org/>) [3], are capable of simulating qubit reset and realistic hardware noise. We also anticipate that the qubit reset functionality on NISQ machines would be available in the near future, opening avenue for proof-of-principle experiments of the proposed scheme for input sequences of arbitrary length.

## VII. CONCLUSION AND OUTLOOK

We have developed a general theory for learning arbitrary I/O maps with fading memory using dissipative quantum systems. The attractiveness of the theory studied here is that it allows a dissipative quantum system (that meets certain requirements but is otherwise arbitrary) to be combined with a classical processor to learn I/O maps from sample I/O sequences. We apply the theory to demonstrate a universal class of dissipative quantum systems that can approximate arbitrary I/O maps with fading memory.

Numerical experiments indicate that even with only a small number of qubits and a simple linear output, this class can achieve comparable performance, in terms of the average normalized mean-squared error, to classical learning schemes such as ESNs and the Volterra series with a large number of tunable parameters. However, when the state-space sizes of the quantum subsystem and classical learning schemes are the same, and the number of computational nodes equals the number of nodes in the ESN plus one (for the constant term) and a similar number of the QRC, the quantum system does not demonstrate any computational advantage. Moreover, the numerical results for a small number of qubits indicate that increasing the dimension of the Hilbert space of the quantum system while fixing the number of computational nodes can still result in improved prediction performance on a number of benchmarking tasks, whereas increasing the state space of ESNs while fixing the



computational nodes does not lead to any noticeable improvement. This strongly suggests that the possibly very large Hilbert space of the quantum subsystem presents a potential resource that can be exploited in this approach. That is, for state-space dimensions beyond what can be simulated on a conventional digital computer. It remains to be investigated if this resource can indeed lead to a provable performance advantage over conventional classical learning approaches, and the circumstances where this will be the case.

## VIII. APPENDIX

### A. The convergence property

Recall from the main text that for a compact subset  $D \subseteq \mathbb{R}$  and  $L > 0$ ,  $K_L(D)$  denotes the set of all real sequences  $u = \{u_k\}_{k \in \mathbb{Z}}$  taking values in  $D \cap [-L, L]$ . Let  $K_L^-(D)$  and  $K_L^+(D)$  be subsets of input sequences in  $K_L(D)$  whose indices are restricted to  $\mathbb{Z}^- = \{\dots, -2, -1, 0\}$  and  $\mathbb{Z}^+ = \{1, 2, \dots\}$ , respectively. In the following, we write  $T$  for both input-independent and input-dependent CPTP maps. As in the main text, we write  $T(u_k)$  for a CPTP map that is determined by an input  $u_k$ , and  $\|\cdot\|_p$  for any Schatten  $p$ -norm for  $p \in [1, \infty)$ . All dissipative quantum systems considered here are finite-dimensional. We now state the definition of a convergent CPTP map with respect to  $K_L(D)$ .

**Definition 1** (Convergence). *An input-dependent CPTP map  $T$  is convergent with respect to  $K_L(D)$  if there exists a sequence  $\delta = \{\delta_k\}_{k>0}$  with  $\lim_{k \rightarrow \infty} \delta_k = 0$ , such that for all  $u = \{u_k\}_{k \in \mathbb{Z}^+} \in K_L^+(D)$  and any two density operators  $\rho_{j,k}$  ( $j = 1, 2$ ) satisfying  $\rho_{j,k} = T(u_k)\rho_{j,k-1}$ , it holds that  $\|\rho_{1,k} - \rho_{2,k}\|_2 \leq \delta_k$ . We call a dissipative quantum system whose dynamics is governed by a convergent CPTP map a convergent system.*

The convergence property can be viewed as an extension of the mixing property for a noisy quantum channel described by an input-independent CPTP map [40].

**Definition 2** (Mixing). *A  $n$ -qubit dissipative quantum system described by a CPTP map  $T$  is mixing if for all  $\rho_0 \in \mathcal{D}(\mathbb{C}^{2^n})$ , if there exists a unique density operator  $\rho_*$  such that,*

$$\lim_{k \rightarrow \infty} \left\| \left( \prod_{j=1}^k T(\rho_0) \right) - \rho_* \right\|_2 = 0.$$

We will see later that if an input-dependent CPTP map  $T(u_k)$  satisfies the sufficient condition in Theorem 3, then  $T(u_k)$  is mixing for each  $u_k \in D \cap [-L, L]$ .

**Theorem 3** (Convergence property). *A  $n$ -qubit dissipative quantum system governed by an input-dependent CPTP map  $T$  is convergent with respect to  $K_L(D)$  if, for all  $u_k \in D \cap [-L, L]$ ,  $T(u_k)$  on the hyperplane  $H_0(2^n)$  of  $2^n \times 2^n$  traceless Hermitian operators satisfies  $\|T(u_k)|_{H_0(2^n)}\|_{2 \rightarrow 2} := \sup_{A \in H_0(2^n), A \neq 0} \frac{\|T(u_k)A\|_2}{\|A\|_2} \leq 1 - \epsilon$  for some  $0 < \epsilon \leq 1$ . Moreover, any pair of initial density operators converge uniformly to one another under  $T$ .*

*Proof.* Let  $\rho_{1,0}$  and  $\rho_{2,0}$  be two arbitrary initial density operators,  $\rho_{1,0} - \rho_{2,0}$  is a traceless

Hermitian operator. We have,

$$\begin{aligned}
\|\rho_{1,k} - \rho_{2,k}\|_2 &= \left\| \left( \overleftarrow{\prod}_{j=1}^k T(u_j) \right) (\rho_{1,0} - \rho_{2,0}) \right\|_2 \\
&= \left\| \left( \overleftarrow{\prod}_{j=1}^k T(u_j)|_{H_0(2^n)} \right) (\rho_{1,0} - \rho_{2,0}) \right\|_2 \\
&\leq \overleftarrow{\prod}_{j=1}^k \|T(u_j)|_{H_0(2^n)}\|_{2 \rightarrow 2} \|\rho_{1,0} - \rho_{2,0}\|_2 \\
&\leq \overleftarrow{\prod}_{j=1}^k (1 - \epsilon) \|\rho_{1,0} - \rho_{2,0}\|_2 \\
&\leq \overleftarrow{\prod}_{j=1}^k (1 - \epsilon) (\|\rho_{1,0}\|_2 + \|\rho_{2,0}\|_2) \\
&\leq 2(1 - \epsilon)^k,
\end{aligned}$$

where the last inequality follows from the fact that for all  $\rho \in \mathcal{D}(\mathbb{C}^{2^n})$ ,  $\|\rho\|_2 \leq 1$ .  $\square$

We remark that for a  $n$ -qubit dissipative quantum system that satisfies the condition in Theorem 3, any initial density operator  $\rho_0$  reaches the state  $\lim_{k \rightarrow \infty} \left( \overleftarrow{\prod}_{j=1}^k T(u_j) \right) \left( \frac{I}{2^n} \right)$ . To see this, let

$$\rho_0 = \frac{I}{2^n} + \sum_{\substack{j_1, j_2, \dots, j_n = \{0, 1, 2, 3\} \\ j_1 j_2 \dots j_n \neq 0}} \alpha_{j_1 j_2 \dots j_n} \bigotimes_{i=1}^n \sigma_{j_i}^{(i)},$$

where  $\sigma_{j_i}^{(i)}$  denotes, for qubit  $i$ , the identity operator  $I$  if  $j_i = 0$ , the Pauli X operator if  $j_i = 1$ , the Pauli Y operator if  $j_i = 2$  and the Pauli Z operator if  $j_i = 3$ . Since  $\bigotimes_{i=1}^n \sigma_{j_i}^{(i)}$  for  $j_1 j_2 \dots j_n \neq 0$  are all traceless Hermitian operators, therefore as  $k \rightarrow \infty$ ,

$$\left\| \rho_k - \left( \overleftarrow{\prod}_{j=1}^k T(u_j) \right) \left( \frac{I}{2^n} \right) \right\|_2 \rightarrow 0.$$

## B. The universality property

We now show the universality property of convergent dissipative quantum systems. Let  $\mathbb{R}^{\mathbb{Z}}$  be the set of all real-valued infinite sequences. Consider a  $n$ -qubit convergent dissipative quantum system described by Eqs. (1) and (2) in the main text, whose dynamics and output are defined by a CPTP map  $T$  and a functional  $h : \mathcal{D}(\mathbb{C}^{2^n}) \rightarrow \mathbb{R}$ , respectively. We associate this quantum system with an induced filter  $M_h^T : K_L(D) \rightarrow \mathbb{R}^{\mathbb{Z}}$ , such that for any initial condition  $\rho_{-\infty} \in \mathcal{D}(\mathbb{C}^{2^n})$ , when evaluated at time  $t = k\tau$ ,

$$M_h^T(u)_k = h \left( \left( \overrightarrow{\prod}_{j=0}^{\infty} T(u_{k-j}) \right) \rho_{-\infty} \right),$$

where  $\overrightarrow{\prod}_{j=0}^{\infty} T(u_{k-j}) = \lim_{N \rightarrow \infty} \overleftarrow{\prod}_{j=0}^N T(u_{k+(j-N)}) = \lim_{N \rightarrow \infty} T(u_k) T(u_{k-1}) \dots T(u_{k-N})$ , and the limit is a pointwise limit. Lemma 4 states that this limit is well-defined.

**Lemma 4.** *The filter  $M_h^T : K_L(D) \rightarrow \mathbb{R}^{\mathbb{Z}}$  is well-defined. In particular, the limit  $\lim_{N \rightarrow \infty} T(u_k) T(u_{k-1}) \dots T(u_{k-N}) \rho_{-N}$  exists and is independent of  $\rho_{-N}$ .*

*Proof.* The set  $\mathcal{D}(\mathbb{C}^{2^n})$  equipped with the distance function induced by the norm  $\|\cdot\|_2$  is a complete metric space. Therefore, every Cauchy sequence converges to a point in  $\mathcal{D}(\mathbb{C}^{2^n})$  [41]. It remains to show that  $S_n = T(u_k)T(u_{k-1}) \cdots T(u_{k-n})\rho_{-n}$  is a Cauchy sequence. By hypothesis, for all  $u_k \in D \cap [-L, L]$ ,  $\|T(u_k)|_{H_0(2^n)}\|_2 \leq 1 - \epsilon$  for some  $0 < \epsilon \leq 1$ . For any  $\epsilon' > 0$ , let  $N > 0$  such that  $(1 - \epsilon)^N < \frac{\epsilon'}{2}$ . Then for all  $n, m > N$ , suppose that  $n \leq m$ ,

$$\begin{aligned} \|S_n - S_m\|_2 &= \|T(u_k)T(u_{k-1}) \cdots T(u_{k-n})(\rho_{-n} - T(u_{k-n-1}) \cdots T(u_{k-m})\rho_{-m})\|_2 \\ &\leq (1 - \epsilon)^{n+1} (\|\rho_{-n}\|_2 + \|(T(u_{k-n-1}) \cdots T(u_{k-m}))\rho_{-m}\|_2) \\ &\leq 2(1 - \epsilon)^N < \epsilon' \end{aligned}$$

□

This filter is causal since given  $u, v \in K_L(D)$  satisfying  $u_\tau = v_\tau$  for  $\tau \leq k$ ,  $M_h^T(u)_k = M_h^T(v)_k$ . For any  $\tau \in \mathbb{Z}$ , let  $M_\tau$  be the shift operator defined by  $M_\tau(u)_k = u_{k-\tau}$ . A filter is said to be time-invariant if it commutes with  $M_\tau$ . It is straightforward to show that  $M_h^T$  is time-invariant.

For a time-invariant and causal filter, there is a corresponding functional  $F_h^T : K_L^-(D) \rightarrow \mathbb{R}$  defined as  $F_h^T(u) = M_h^T(u)_0$  (see [10]). The corresponding filter can be recovered via  $M_h^T(u)_k = F_h^T(P \circ M_{-k}(u))$ , where  $P$  truncates  $u$  up to 0, that is  $P(u) = u|_0$ . We say a filter  $M_h^T$  has the fading memory property if and only if  $F_h^T$  is continuous with respect to a weighted norm defined as follows.

**Definition 5** (Weighted norm). *For a null sequence  $w = \{w_k\}_{k \geq 0}$ , that is  $w : \{0\} \cup \mathbb{Z}^+ \rightarrow (0, 1]$  is decreasing and  $\lim_{k \rightarrow \infty} w_k = 0$ , define a weighted norm  $\|\cdot\|_w$  on  $K_L^-(D)$  as  $\|u\|_w := \sup_{k \in \mathbb{Z}^-} |u_k| w_{-k}$ .*

**Definition 6** (Fading memory). *A time-invariant causal filter  $M : K_L(D) \rightarrow \mathbb{R}^{\mathbb{Z}}$  has the fading memory property with respect to a null sequence  $w$  if and only if its corresponding functional  $F : K_L^-(D) \rightarrow \mathbb{R}$  is continuous with respect to the weighted norm  $\|\cdot\|_w$ .*

To emphasize that the fading memory property is defined with respect to a null sequence  $w$ , we will say that  $M$  is a  $w$ -fading memory filter and the corresponding functional  $F$  is a  $w$ -fading memory functional. We state the following compactness result [19, Lemma 2] and the Stone-Weierstrass theorem [13, Theorem 7.3.1].

**Lemma 7** (Compactness). *For any null sequence  $w$ ,  $K_L^-(D)$  is compact with the weighted norm  $\|\cdot\|_w$ .*

We write  $(K_L^-(D), \|\cdot\|_w)$  to denote the space  $K_L^-(D)$  equipped with the weighted norm  $\|\cdot\|_w$ .

**Theorem 8** (Stone-Weierstrass). *Let  $E$  be a compact metric space and  $C(E)$  be the set of real-valued continuous functions defined on  $E$ . If a subalgebra  $A$  of  $C(E)$  contains the constant functions and separates points of  $E$ , then  $A$  is dense in  $C(E)$ .*

Let  $C(K_L^-(D), \|\cdot\|_w)$  be the set of continuous functionals  $F : (K_L^-(D), \|\cdot\|_w) \rightarrow \mathbb{R}$ . The following theorem is a result of the compactness of  $(K_L^-(D), \|\cdot\|_w)$  (Lemma 7) and the Stone-Weierstrass Theorem (Theorem 8).

**Theorem 9.** *Let  $w$  be a null sequence. For convergent CPTP maps  $T$ , let  $\mathcal{M}_w = \{M_h^T \mid h : \mathcal{D}(\mathbb{C}^{2^n}) \rightarrow \mathbb{R}\}$  be a set of  $w$ -fading memory filters. Let  $\mathcal{F}_w$  be the family of corresponding  $w$ -fading memory functionals defined on  $K_L^-(D)$ . If  $\mathcal{F}_w$  forms a polynomial algebra of  $C(K_L^-(D), \|\cdot\|_w)$ , contains the constant functionals and separates points of  $K_L^-(D)$ , then  $\mathcal{F}_w$  is dense in  $C(K_L^-(D), \|\cdot\|_w)$ . That is for any  $w$ -fading memory filter  $M_*$  and any  $\epsilon > 0$ , there exists  $M_h^T \in \mathcal{M}_w$  such that for all  $u \in K_L(D)$ ,  $\|M_*(u) - M_h^T(u)\|_\infty = \sup_{k \in \mathbb{Z}} |M_*(u)_k - M_h^T(u)_k| < \epsilon$ .*

*Proof.*  $\mathcal{F}_w$  is dense follows from Lemma 7 and Theorem 8. To prove the second part of the theorem, since  $\mathcal{F}_w$  is dense in  $C(K_L^-(D), \|\cdot\|_w)$ , for any  $w$ -fading memory functional  $F_*$  and any  $\epsilon > 0$ , there exists  $F_h^T \in \mathcal{F}_w$  such that for all  $u_- \in K_L^-(D)$ ,  $|F_*(u_-) - F_h^T(u_-)| < \epsilon$ . For  $u \in K_L(D)$ , notice that  $P \circ M_{-k}(u) \in K_L^-(D)$  for all  $k \in \mathbb{Z}$ , hence

$$|F_*(P \circ M_{-k}(u)) - F_h^T(P \circ M_{-k}(u))| = |M_*(u)_k - M_h^T(u)_k| < \epsilon.$$

Since this is true for all  $k \in \mathbb{Z}$ , therefore for all  $u \in K_L(D)$ ,  $\|M_*(u) - M_h^T(u)\|_\infty < \epsilon$ .  $\square$

### C. Fading memory property and polynomial algebra

Before we prove the universality of the family of dissipative quantum systems introduced in Sec. IV in the main text, we first show two important observations regarding to the multivariate polynomial output in Eq. (3).

We specify  $h$  to be the multivariate polynomial as in the right hand side of Eq. (3) in the main text. For ease of notation, we drop the subscript  $h$  in  $F_h^T$  and  $M_h^T$ . Let  $\mathcal{F} = \{F^T\}$  be the set of functionals induced from dissipative quantum systems given by Eqs. (1) and (3) in the main text. We will show in Lemma 10 that the convergence and continuity of  $T$  are sufficient to guarantee the fading memory property of  $F^T$ , and in Lemma 12 that  $\mathcal{F}$  forms a polynomial algebra, made of fading memory functionals. In the following, let  $\mathcal{L}(\mathbb{C}^{2^n})$  be the set of linear operators on  $\mathbb{C}^{2^n}$ , and for a CPTP map  $T$ , for all  $u_k \in D \cap [-L, L]$ , define  $\|T(u_k)\|_{2-2} := \sup_{A \in \mathcal{L}(\mathbb{C}^{2^n}), \|A\|_2=1} \|T(u_k)A\|_2$ .

**Lemma 10** (Fading memory property). *Consider a  $n$ -qubit dissipative quantum system with dynamics Eq. (1) and output Eq. (3). Suppose that for all  $u_k \in D \cap [-L, L]$ , the CPTP map  $T(u_k)$  satisfies the condition in Theorem 3, so that it is convergent. Moreover, for any  $\epsilon > 0$ , there exists  $\delta_T(\epsilon) > 0$  such that  $\|T(x) - T(y)\|_{2-2} < \epsilon$  whenever  $|x - y| < \delta_T(\epsilon)$  for  $x, y \in D \cap [-L, L]$ . Then for any null sequence  $w$ , the induced filter  $M^T$  and the corresponding functional  $F^T$  are  $w$ -fading memory.*

*Proof.* We first state the boundedness of CPTP maps [37, Theorem 2.1].

**Lemma 11.** *For a CPTP map  $T : \mathcal{L}(\mathbb{C}^{2^n}) \rightarrow \mathcal{L}(\mathbb{C}^{2^n})$ , we have  $\|T\|_{2-2} \leq \sqrt{2^n}$ .*

Moreover, recall that  $\text{Tr}(\cdot)$  is continuous, that is for any  $\epsilon > 0$ , there exists  $\delta_{\text{Tr}}(\epsilon) > 0$  such that  $|\text{Tr}(A - B)| < \epsilon$  whenever  $\|A - B\|_2 < \delta_{\text{Tr}}(\epsilon)$  for any complex matrices  $A, B$ . Note that here  $\|\cdot\|_2$  denotes the Schatten 2-norm or the Hilbert-Schmidt norm.

Let  $w$  be an arbitrary null sequence. We will show the linear terms  $L(u)$  in the functional  $F^T$  are continuous with respect to  $\|\cdot\|_w$ , and the continuity property of  $F^T$  follows from the fact that finite sums and products of continuous elements are also continuous.

For any  $u, v \in K_L^-(D)$ ,

$$|L(u) - L(v)| = \left| \text{Tr} \left( Z^{(i_1)} \left( \left( \vec{\prod}_{k=0}^{\infty} T(u_{-k}) \right) \rho_{-\infty} - \left( \vec{\prod}_{k=0}^{\infty} T(v_{-k}) \right) \rho_{-\infty} \right) \right) \right|.$$

Denote  $\rho_u = \left( \vec{\prod}_{k=N}^{\infty} T(u_{-k}) \right) \rho_{-\infty}$  and  $\rho_v = \left( \vec{\prod}_{k=N}^{\infty} T(v_{-k}) \right) \rho_{-\infty}$  for some  $0 < N < \infty$ ,

$$\begin{aligned} & \left\| Z^{(i_1)} \left( \left( \vec{\prod}_{k=0}^{\infty} T(u_{-k}) \right) \rho_{-\infty} - \left( \vec{\prod}_{k=0}^{\infty} T(v_{-k}) \right) \rho_{-\infty} \right) \right\|_2 \\ & \leq \|Z^{(i_1)}\|_2 \left( \left\| \vec{\prod}_{k=0}^{N-1} T(u_{-k}) - \vec{\prod}_{k=0}^{N-1} T(v_{-k}) \right\|_{2-2} \|\rho_u\|_2 + \left\| \left( \vec{\prod}_{k=0}^{N-1} T(v_{-k}) \right) (\rho_u - \rho_v) \right\|_2 \right). \end{aligned} \quad (4)$$

Since  $T$  satisfies conditions in Theorem 3, any two density operators converge uniformly to one another. Therefore, for any  $\epsilon > 0$ , there exists  $N(\epsilon) > 0$  such that for all  $N' > N(\epsilon)$ ,

$$\left\| \left( \vec{\prod}_{k=0}^{N'-1} T(v_{-k}) \right) (\rho_u - \rho_v) \right\|_2 < \frac{\delta_{\text{Tr}}(\epsilon)}{2 \|Z^{(i_1)}\|_2}. \quad (5)$$

Choose  $N' = N(\epsilon) + 1$  and bound the first term inside the bracket on the right hand side of Eq. (4) by rewriting it as a telescopic sum:

$$\begin{aligned} & \left\| \vec{\prod}_{k=0}^{N(\epsilon)} T(u_{-k}) - \vec{\prod}_{k=0}^{N(\epsilon)} T(v_{-k}) \right\|_{2-2} \\ & = \left\| \sum_{l=0}^{N(\epsilon)} (T(v_0) \cdots T(v_{-(l-1)}) T(u_{-l}) T(u_{-(l+1)}) \cdots T(u_{-N(\epsilon)}) \right. \\ & \quad \left. - T(v_0) \cdots T(v_{-(l-1)}) T(v_{-l}) T(u_{-(l+1)}) \cdots T(u_{-N(\epsilon)}) \right) \right\|_{2-2} \\ & \leq \sum_{l=0}^{N(\epsilon)} \|T(v_0) \cdots T(v_{-(l-1)})\|_{2-2} \|T(u_{-l}) - T(v_{-l})\|_{2-2} \|T(u_{-(l+1)}) \cdots T(u_{-N(\epsilon)})\|_{2-2} \\ & \leq 2^n \sum_{l=0}^{N(\epsilon)} \|T(u_{-l}) - T(v_{-l})\|_{2-2}, \end{aligned} \quad (6)$$

where the last inequality follows from Lemma 11. We claim that for any  $\epsilon > 0$ , if

$$\|u - v\|_w = \sup_{k \in \mathbb{Z}^-} |u_k - v_k| w_{-k} < \delta_T \left( \frac{\delta_{\text{Tr}}(\epsilon)}{2^{n+1} \|Z^{(i_1)}\|_2 (N(\epsilon) + 1)} \right) w_{N(\epsilon)}$$

then  $|L(u) - L(v)| < \epsilon$ . Indeed, since  $w$  is decreasing, the above condition implies that

$$\max_{0 \leq l \leq N(\epsilon)} |u_{-l} - v_{-l}| w_{N(\epsilon)} < \delta_T \left( \frac{\delta_{\text{Tr}}(\epsilon)}{2^{n+1} \|Z^{(i_1)}\|_2 (N(\epsilon) + 1)} \right) w_{N(\epsilon)}.$$

Since  $w_{N(\epsilon)} > 0$ , for all  $0 \leq l \leq N(\epsilon)$ ,

$$|u_{-l} - v_{-l}| < \delta_T \left( \frac{\delta_{\text{Tr}}(\epsilon)}{2^{n+1} \|Z^{(i_1)}\|_2 (N(\epsilon) + 1)} \right).$$

By continuity of  $T$ , we bound Eq. (6) by

$$2^n \sum_{l=0}^{N(\epsilon)} \|T(u_{-l}) - T(v_{-l})\|_{2-2} < 2^n \sum_{l=0}^{N(\epsilon)} \frac{\delta_{\text{Tr}}(\epsilon)}{2^{n+1} \|Z^{(i_1)}\|_2 (N(\epsilon) + 1)} = \frac{\delta_{\text{Tr}}(\epsilon)}{2 \|Z^{(i_1)}\|_2}. \quad (7)$$

Since  $\|\rho_u\|_2 \leq 1$ , Eqs. (4), (5) and (7) give

$$\begin{aligned} & \|Z^{(i_1)}\|_2 \left( \left\| \vec{\prod}_{k=0}^{N(\epsilon)} T(u_{-k}) - \vec{\prod}_{k=0}^{N(\epsilon)} T(v_{-k}) \right\|_{2-2} \|\rho_u\|_2 + \left\| \left( \vec{\prod}_{k=0}^{N(\epsilon)} T(v_{-k}) \right) (\rho_u - \rho_v) \right\|_2 \right) \\ & < \delta_{\text{Tr}}(\epsilon). \end{aligned}$$

The result now follows from the continuity of  $\text{Tr}(\cdot)$ .  $\square$

**Lemma 12** (Polynomial algebra). *Let  $\mathcal{F} = \{F^T\}$  be a family of functionals induced by dissipative quantum systems defined by Eqs. (1) and (3) in the main text. If for each member  $F^T \in \mathcal{F}$ ,  $T$  satisfies the conditions in Lemma 10, then for any null sequence  $w$ ,  $\mathcal{F}$  forms a polynomial algebra consisting of  $w$ -fading memory functionals.*

*Proof.* Consider two dissipative quantum systems described by Eqs. (1) and (3), with  $n_1$  and  $n_2$  system qubits respectively. Let  $\rho_k^{(m)} \in \mathcal{D}(\mathbb{C}^{2^{n_m}})$  be the state and  $T^{(m)}$  be the CPTP map of the  $m^{\text{th}}$  system. Let  $j_1 = 1, \dots, n_1$  and  $j_2 = 1, \dots, n_2$  be the respective qubit indices for the two systems. For the observable  $Z^{(j_m)}$  of qubit  $j_m$ , notice that

$$\begin{aligned} \text{Tr} \left( Z^{(j_1)} \rho_k^{(1)} \right) &= \text{Tr} \left( (Z^{(j_1)} \otimes I) (\rho_k^{(1)} \otimes \rho_k^{(2)}) \right), \\ \text{Tr} \left( Z^{(j_2)} \rho_k^{(2)} \right) &= \text{Tr} \left( (I \otimes Z^{(j_2)}) (\rho_k^{(1)} \otimes \rho_k^{(2)}) \right), \end{aligned}$$

where  $I$  is the identity operator. Therefore, we can relabel the qubit for the combined system described by the density operator  $\rho_k^{(1)} \otimes \rho_k^{(2)}$  by  $j$ , running from  $j = 1$  to  $j = n_1 + n_2$ . Using this notation, the above expectations can be re-expressed as

$$\begin{aligned} \text{Tr} \left( Z^{(j_1)} \rho_k^{(1)} \right) &= \text{Tr} \left( Z^{(j)} \rho_k^{(1)} \otimes \rho_k^{(2)} \right), \quad j = j_1 \\ \text{Tr} \left( Z^{(j_2)} \rho_k^{(2)} \right) &= \text{Tr} \left( Z^{(j)} \rho_k^{(1)} \otimes \rho_k^{(2)} \right), \quad j = n_1 + j_2. \end{aligned}$$

Following this idea, write out the outputs of two systems as follows,

$$\begin{aligned} \bar{y}_k^{(1)} &= C_1 + \sum_{d_1=1}^{R_1} \sum_{i_1=1}^{n_1} \cdots \sum_{i_{n_1}=i_{n_1-1}+1}^{n_1} \sum_{r_{i_1}+\cdots+r_{i_{n_1}}=d_1} w_{i_1, \dots, i_{n_1}}^{r_{i_1}, \dots, r_{i_{n_1}}} \langle Z^{(i_1)} \rangle_k^{r_{i_1}} \cdots \langle Z^{(i_{n_1})} \rangle_k^{r_{i_{n_1}}}, \\ \bar{y}_k^{(2)} &= C_2 + \sum_{d_2=1}^{R_2} \sum_{j_1=1}^{n_2} \cdots \sum_{j_{n_2}=j_{n_2-1}+1}^{n_2} \sum_{r_{j_1}+\cdots+r_{j_{n_2}}=d_2} w_{j_1, \dots, j_{n_2}}^{r_{j_1}, \dots, r_{j_{n_2}}} \langle Z^{(j_1)} \rangle_k^{r_{j_1}} \cdots \langle Z^{(j_{n_2})} \rangle_k^{r_{j_{n_2}}}. \end{aligned}$$

For any  $\lambda \in \mathbb{R}$ , let  $n = n_1 + n_2$  and  $k$  denote the qubit index of the combined system running from  $k = 1$  to  $k = n$ , and  $R = \max\{R_1, R_2\}$ , then

$$\bar{y}_k^{(1)} + \lambda \bar{y}_k^{(2)} = C_1 + \lambda C_2 + \sum_{d=1}^R \sum_{k_1=1}^n \cdots \sum_{k_n=k_{n-1}+1}^n \sum_{r_{k_1}+\cdots+r_{k_n}=d} \bar{w}_{k_1, \dots, k_n}^{r_{k_1}, \dots, r_{k_n}} \langle Z^{(k_1)} \rangle_k^{r_{k_1}} \cdots \langle Z^{(k_n)} \rangle_k^{r_{k_n}},$$

where the weights  $\bar{w}_{k_1, \dots, k_n}^{r_{k_1}, \dots, r_{k_n}}$  are changed accordingly. For instance, if all  $k_m \leq n_1$  for  $m = 1, 2, \dots, n$ , then  $\bar{w}_{k_1, \dots, k_n}^{r_{k_1}, \dots, r_{k_n}} = w_{i_1, \dots, i_{n_1}}^{r_{i_1}, \dots, r_{i_{n_1}}}$ , corresponding to the weights for the output  $\bar{y}_k^{(1)}$ . Similarly, let  $R = R_1 + R_2$ ,

$$\bar{y}_k^{(1)} \bar{y}_k^{(2)} = C_1 C_2 + \sum_{d=1}^R \sum_{k_1=1}^n \cdots \sum_{k_n=k_{n-1}+1}^n \sum_{r_{k_1}+\dots+r_{k_n}=d} \hat{w}_{k_1, \dots, k_n}^{r_{k_1}, \dots, r_{k_n}} \langle Z^{(k_1)} \rangle_k^{r_{k_1}} \cdots \langle Z^{(k_n)} \rangle_k^{r_{k_n}}.$$

Therefore,  $\bar{y}_k^{(1)} + \lambda \bar{y}_k^{(2)}$  and  $\bar{y}_k^{(1)} \bar{y}_k^{(2)}$  again have the same form as the right hand side of Eq. (3) in the main text. This implies that for any functionals  $F^{T^{(1)}}, F^{T^{(2)}} \in \mathcal{F}$ ,  $F^{T^{(1)}} + \lambda F^{T^{(2)}} \in \mathcal{F}$  and  $F^{T^{(1)}} F^{T^{(2)}} \in \mathcal{F}$ . Thus,  $\mathcal{F}$  forms a polynomial algebra.

It remains to show that for all  $u_k \in D \cap [-L, L]$ ,  $\|T(u_k)|_{H_0(2^n)}\|_{2-2} = \|(T^{(1)}(u_k) \otimes T^{(2)}(u_k))|_{H_0(2^n)}\|_{2-2} \leq 1 - \epsilon$  for some  $0 < \epsilon \leq 1$ . This will imply that  $F^{T^{(1)}} + \lambda F^{T^{(2)}}$  and  $F^{T^{(1)}} F^{T^{(2)}}$  are  $w$ -fading memory by Lemma 10, and that  $\mathcal{F}$  forms a polynomial algebra consisting of  $w$ -fading memory functionals. Suppose that for all  $u_k \in D \cap [-L, L]$ ,  $\|T(u_k)|_{H_0(2^{nm})}\|_{2-2} \leq 1 - \epsilon_m$  for  $m = 1, 2$ . Adopting the proof of [23, Proposition 3], let  $A = \sum_i A_i \otimes \tilde{A}_i$  be a traceless Hermitian operator. Without loss of generality, we assume that  $\{\tilde{A}_i\}$  is an orthonormal set with respect to the Hilbert-Schmidt inner product. Then  $\{A_i \otimes \tilde{A}_i\}$  and  $\{T^{(1)}(u_k)|_{H_0(2^{n_1})} A_i \otimes \tilde{A}_i\}$  are two orthogonal sets. By the Pythagoras theorem,  $T^{(1)}(u_k)|_{H_0(2^{n_1})} \otimes I$  on the hyperplane of traceless Hermitian operators satisfies

$$\begin{aligned} & \| (T^{(1)}(u_k)|_{H_0(2^{n_1})} \otimes I) \sum_i A_i \otimes \tilde{A}_i \|_2^2 = \sum_i \| T^{(1)}(u_k)|_{H_0(2^{n_1})} A_i \otimes \tilde{A}_i \|_2^2 \\ & = \sum_i \| T^{(1)}(u_k)|_{H_0(2^{n_1})} A_i \|_2^2 \| \tilde{A}_i \|_2^2 \leq \| T^{(1)}(u_k)|_{H_0(2^{n_1})} \|_{2-2}^2 \sum_i \| A_i \|_2^2 \| \tilde{A}_i \|_2^2 \\ & = \| T^{(1)}(u_k)|_{H_0(2^{n_1})} \|_{2-2}^2 \sum_i \| A_i \otimes \tilde{A}_i \|_2^2 = \| T^{(1)}(u_k)|_{H_0(2^{n_1})} \|_{2-2}^2 \| \sum_i A_i \otimes \tilde{A}_i \|_2^2. \end{aligned}$$

Therefore,  $\|T^{(1)}(u_k)|_{H_0(2^{n_1})} \otimes I\|_{2-2} \leq \|T^{(1)}(u_k)|_{H_0(2^{n_1})}\|_{2-2}$ . Similarly, a symmetric argument shows that  $\|I \otimes T^{(2)}(u_k)|_{H_0(2^{n_2})}\|_{2-2} \leq \|T^{(2)}(u_k)|_{H_0(2^{n_2})}\|_{2-2}$ . Therefore, when restricted to traceless Hermitian operators,

$$\begin{aligned} & \| (T^{(1)}(u_k) \otimes T^{(2)}(u_k))|_{H_0(2^n)} \|_{2-2} = \| (T^{(1)}(u_k)|_{H_0(2^{n_1})} \otimes I) (I \otimes T^{(2)}(u_k)|_{H_0(2^{n_2})}) \|_{2-2} \\ & \leq \| T^{(1)}(u_k)|_{H_0(2^{n_1})} \otimes I \|_{2-2} \| I \otimes T^{(2)}(u_k)|_{H_0(2^{n_2})} \|_{2-2} \\ & \leq \| T^{(1)}(u_k)|_{H_0(2^{n_1})} \|_{2-2} \| T^{(2)}(u_k)|_{H_0(2^{n_2})} \|_{2-2} \leq (1 - \epsilon_1)(1 - \epsilon_2). \end{aligned}$$

The convergence of  $T$  follows from Theorem 3.  $\square$

#### D. A universal class

We now prove the universality of the class of dissipative quantum systems introduced in the main text. Recall that this class consists of  $N$  non-interacting quantum subsystems initialized in a product state of the  $N$  subsystems, where the dynamics of subsystem  $K$  with  $n_K$  qubits is governed by the CPTP map:

$$T_K(u_k) \rho_{k-1}^K = \text{Tr}_{i_0^K} (e^{-iH_K \tau} \rho_{k-1}^K \otimes \rho_{i_0, k}^K e^{iH_K \tau}), \quad (8)$$

where

$$\rho_{i_0,k}^K = u_k|0\rangle\langle 0| + (1 - u_k)|1\rangle\langle 1|, \quad 0 \leq u_k \leq 1$$

$$H_K = \sum_{j_1=0}^{n_K} \sum_{j_2=j_1+1}^{n_K} J_K^{j_1,j_2} (X^{(i_{j_1}^K)} X^{(i_{j_2}^K)} + Y^{(i_{j_1}^K)} Y^{(i_{j_2}^K)}) + \sum_{j=0}^{n_K} \alpha Z^{(i_j^K)}, \quad (9)$$

with  $J_K^{j_1,j_2}$  and  $\alpha$  being real-valued constants and  $\text{Tr}_{i_0^K}$  denoting the partial trace over the ancilla qubit. Let  $\overline{H}_K = I \otimes \cdots \otimes H_K \otimes \cdots \otimes I$  with  $H_K$  in the  $K$ -th position, the total Hamiltonian of  $N$  subsystems is

$$H = \sum_{K=1}^N \overline{H}_K. \quad (10)$$

Writing  $\rho_k = \bigotimes_{K=1}^N \rho_k^K$ , the overall dynamics and the output are given by

$$\begin{cases} \rho_k = T(u_k) \rho_{k-1} = \bigotimes_{K=1}^N T_K(u_k) \rho_{k-1}^K \\ \bar{y}_k = h(\rho_k), \end{cases} \quad (11)$$

where  $h$  is the multivariate polynomial defined by the right hand side of Eq. (3) in the main text.

**Proposition 13.** *Let  $\mathcal{M}_S$  be the set of filters induced from dissipative quantum systems described by Eq. (11) such that each  $T_K$  ( $K = 1, \dots, N$ ) satisfies conditions in Theorem 3. Then for any null sequence  $w$ , the corresponding family of functionals  $\mathcal{F}_S$  is dense in  $C(K_1^-[0, 1]), \|\cdot\|_w$ .*

*Proof.* We first show  $T_K(x)$  satisfies the conditions in Lemma 10 for all  $x \in [0, 1]$ . Let  $x, y \in [0, 1]$  and  $Z$  be the Pauli Z operator. By definition,

$$\begin{aligned} \|T_K(x) - T_K(y)\|_{2-2} &= \sup_{\substack{A \in \mathcal{L}(\mathbb{C}^{2^n}) \\ \|A\|_2=1}} \|(T_K(x) - T_K(y))A\|_2 \\ &= \sup_{\substack{A \in \mathcal{L}(\mathbb{C}^{2^n}) \\ \|A\|_2=1}} \|\text{Tr}_{i_0}^K(e^{-iH_K\tau} A \otimes (x - y) Z e^{iH_K\tau})\|_2 \\ &= |x - y| \sup_{\substack{A \in \mathcal{L}(\mathbb{C}^{2^n}) \\ \|A\|_2=1}} \|\text{Tr}_{i_0}^K(e^{-iH_K\tau} A \otimes Z e^{iH_K\tau})\|_2 \\ &= |x - y| \|\tilde{T}\|_{2-2}, \end{aligned}$$

where  $\tilde{T}$  is an input-independent CPTP map.

Now, the same argument in the proof of Lemma 12 shows that  $T = T_1 \otimes \cdots \otimes T_N$  is convergent given the assumptions on each  $T_K$ . Furthermore, given two convergent systems whose dynamics are described by Eq. (11) with Hamiltonians  $H^{(1)}$  and  $H^{(2)}$ , the total Hamiltonian of the combined system is  $H = H^{(1)} \otimes I + I \otimes H^{(2)}$ , which again has the form Eq. (10). Therefore, by the above observation and Lemma 12,  $\mathcal{F}_S$  forms a polynomial algebra, consisting of  $w$ -fading memory functionals for any null sequence  $w$ .

It remains to show  $\mathcal{F}_S$  contains constants and separates points. Constants can be obtained by setting the weights  $w_{i_1, \dots, i_n}^{r_{i_1}, \dots, r_{i_n}}$  in the output to be zero. To show the family  $\mathcal{F}_S$  separates points, we state the following lemma for later use, whose proof can be found in [24, Theorem 3.2].



**Lemma 14.** *Let  $f(\theta) = \sum_{n=0}^{\infty} x_n \theta^n$  be a non-constant real power series, having a non-zero radius of convergence. If  $f(0) = 0$ , then there exists  $\beta > 0$  such that  $f(\theta) \neq 0$  for all  $\theta$  with  $|\theta| \leq \beta$  and  $\theta \neq 0$ .*

Consider a single-qubit system interacting with a single ancilla qubit whose dynamics is governed by Eq. (11). Order an orthogonal basis of  $\mathcal{L}(\mathbb{C}^2)$  as  $\mathcal{B} = \{I, Z, X, Y\}$ . Recall that the normal representations of a CPTP map  $T$  and a density operator  $\rho$  are given by

$$\overline{T}_{i,j} = \frac{\text{Tr}(B_i T(B_j))}{2} \quad \text{and} \quad \overline{\rho}_i = \frac{\text{Tr}(\rho B_i)}{2},$$

where  $B_i \in \mathcal{B}$ . Without loss of generality, let  $\tau = 1$  and set  $J_1^{j_1, j_2} = J \in \mathbb{R}$  for all  $j_1, j_2$  in the Hamiltonian given by Eq. (9). We obtain the normal representation of the CPTP map defined in Eq. (8) as

$$\overline{T}(u_k) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \sin^2(2J)(2u_k - 1) & \cos^2(2J) & 0 & 0 \\ 0 & 0 & \cos(2J) \cos(2\alpha) & -\cos(2J) \sin(2\alpha) \\ 0 & 0 & \cos(2J) \sin(2\alpha) & \cos(2J) \cos(2\alpha) \end{pmatrix}.$$

When restricted to the hyperplane of traceless Hermitian operators,

$$\overline{T}|_{H_0(2)} = \begin{pmatrix} \cos^2(2J) & 0 & 0 \\ 0 & \cos(2J) \cos(2\alpha) & -\cos(2J) \sin(2\alpha) \\ 0 & \cos(2J) \sin(2\alpha) & \cos(2J) \cos(2\alpha) \end{pmatrix}$$

with  $\|\overline{T}|_{H_0(2)}\|_{2-2} = \sigma_{\max}(\overline{T}|_{H_0(2)}) = |\cos(2J)|$ . Here,  $\|\cdot\|_{2-2}$  is the matrix 2-norm and  $\sigma_{\max}(\cdot)$  is the maximum singular value. Choose  $J \neq \frac{z\pi}{2}$  for  $z \in \mathbb{Z}$ , then  $|\cos(2J)| \leq 1 - \epsilon$  for some  $0 < \epsilon \leq 1$ . By Theorem 3,  $T$  is convergent and we choose an arbitrary initial density operator  $\overline{\rho}_{-\infty} = (1/2 \ 1/2 \ 0 \ 0)^T$ , corresponding to  $\rho_{-\infty} = |0\rangle\langle 0|$ . If we only take the expectation  $\langle Z \rangle$  in the output Eq. (3) by setting the degree  $R = 1$ , then this single-qubit dissipative quantum system induces a functional

$$F^T(u) = w \left[ \left( \prod_{j=0}^{\infty} \overline{T}(u_{-j}) \right) \overline{\rho}_{-\infty} \right]_2 + C,$$

for all  $u \in K_1^-([0, 1])$ . Here,  $[\cdot]_2$  refers to the second element of the vector, corresponding to  $\langle Z \rangle$  given the order of the orthogonal basis elements in  $\mathcal{B}$ . Given two input sequences  $u \neq v$  in  $K_1^-([0, 1])$ , consider two cases:

(i) If  $u_0 \neq v_0$ , choose  $J = \frac{\pi}{4}$  such that  $\cos^2(2J) = 0$  and  $\sin^2(2J) = 1$ . Then

$$F^T(u) - F^T(v) = w(u_0 - v_0) \neq 0.$$

(ii) If  $u_0 = v_0$ ,

$$F^T(u) - F^T(v) = w \sin^2(2J) \sum_{j=0}^{\infty} (\cos^2(2J))^j (u_{-j} - v_{-j}).$$

Let  $\theta = \cos^2(2J)$ , then given our choice of  $J$ ,  $0 \leq \theta \leq 1 - \epsilon$  and  $\sin^2(2J) \geq \epsilon$  for some  $0 < \epsilon \leq 1$ . Consider the power series

$$f(\theta) = \sum_{j=0}^{\infty} \theta^j (u_{-j} - v_{-j}),$$

since  $|u_{-j} - v_{-j}| \leq 1$ ,  $f(\theta)$  has a non-zero radius of convergence  $R$  such that  $(-1, 1) \subseteq R$ . Moreover,  $f(\theta)$  is non-constant and  $f(0) = 0$ . The separation of points follows from invoking Lemma 14.

Finally, the universality property of  $\mathcal{F}_S$  follows from Theorem 9.  $\square$

## E. Detailed numerical experiment settings

In this section, we describe detailed formulas for the NARMA tasks, simulation of decoherence and experimental conditions for ESNs and the Volterra series.

### 1. The NARMA task

The general  $m$ th-order NARMA I/O map is described as [6]:

$$y_k = 0.3y_{k-1} + 0.05y_{k-1} \left( \sum_{j=0}^{\tau_{\text{NARMA}}-1} y_{k-j-1} \right) + 1.5u_{k-\tau_{\text{NARMA}}}u_k + \gamma.$$

where  $\gamma \in \mathbb{R}$ . In the main text, we consider  $\tau_{\text{NARMA}} = \{15, 20, 30, 40\}$ . For  $\tau_{\text{NARMA}} = \{15, 20\}$ , we set  $\gamma = 0.1$ . For  $\tau_{\text{NARMA}} = \{30, 40\}$ ,  $\gamma$  is set to be 0.05 and 0.04 respectively. A random input sequence  $u^{(r)}$ , where each  $u_k^{(r)}$  is randomly uniformly chosen from  $[0, 0.2]$ , is deployed for all the computational tasks. This range is chosen to ensure stability of the NARMA tasks.

### 2. Decoherence

We consider the dephasing, decaying and generalized amplitude damping (GAD) noise, which are of experimental importance. The dephasing noise has the Kraus operators [34]:

$$M_0 = \sqrt{\frac{1 + \sqrt{1-p}}{2}} I, M_1 = \sqrt{\frac{1 - \sqrt{1-p}}{2}} Z,$$

where  $\sqrt{1-p} = e^{-2\frac{\gamma}{S}\delta t}$ . Therefore, we implement single-qubit phase-flip for all  $n$  system and ancilla qubits. That is for  $j = 1, \dots, n+1$  the density operator  $\rho$  for the system and ancilla qubits undergoes the evolution:

$$\rho \rightarrow \frac{1 + e^{-2\frac{\gamma}{S}\delta t}}{2} \rho + \frac{1 - e^{-2\frac{\gamma}{S}\delta t}}{2} Z^{(j)} \rho Z^{(j)},$$

where  $Z^{(j)}$  denotes the Pauli  $Z$  operator for qubit  $j$ .

The generalized amplitude damping (GAD) channel captures the effect of dissipation to an environment at a finite temperature  $\lambda \in [0, 1]$ . Its Kraus operators are defined by

$$M_0 = \sqrt{\lambda} \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-p} \end{pmatrix}, M_2 = \sqrt{\lambda} \begin{pmatrix} 0 & \sqrt{p} \\ 0 & 0 \end{pmatrix},$$

$$M_3 = \sqrt{1-\lambda} \begin{pmatrix} \sqrt{1-p} & 0 \\ 0 & 1 \end{pmatrix}, M_4 = \sqrt{1-\lambda} \begin{pmatrix} 0 & 0 \\ \sqrt{p} & 0 \end{pmatrix}.$$

When  $\lambda = 1$ , the GAD channel corresponds to the amplitude damping channel (decaying noise). We simulate the generalized amplitude damping channel for  $\lambda = \{0.2, 0.4, 0.6, 0.8\}$ . To implement the GAD channel with the same noise strengths as the dephasing channel, we set  $\sqrt{1-p} = e^{-2\frac{\gamma}{S}\delta t}$ ,  $\sqrt{p} = \sqrt{1 - e^{-4\frac{\gamma}{S}\delta t}}$  to be the same as the dephasing noise.

Following the discussion in Sec. VB, Fig. 8 plots the average SA NMSE for the LRPO, Missile, NARMA15 and NARMA20 tasks under the GAD channel for all the chosen temperature parameters. Fig. 9 and Fig. 10 plot the average sum of modulus of off-diagonal elements in the system density operator, for the last 50 timesteps of the SA samples, under all noise types discussed above.

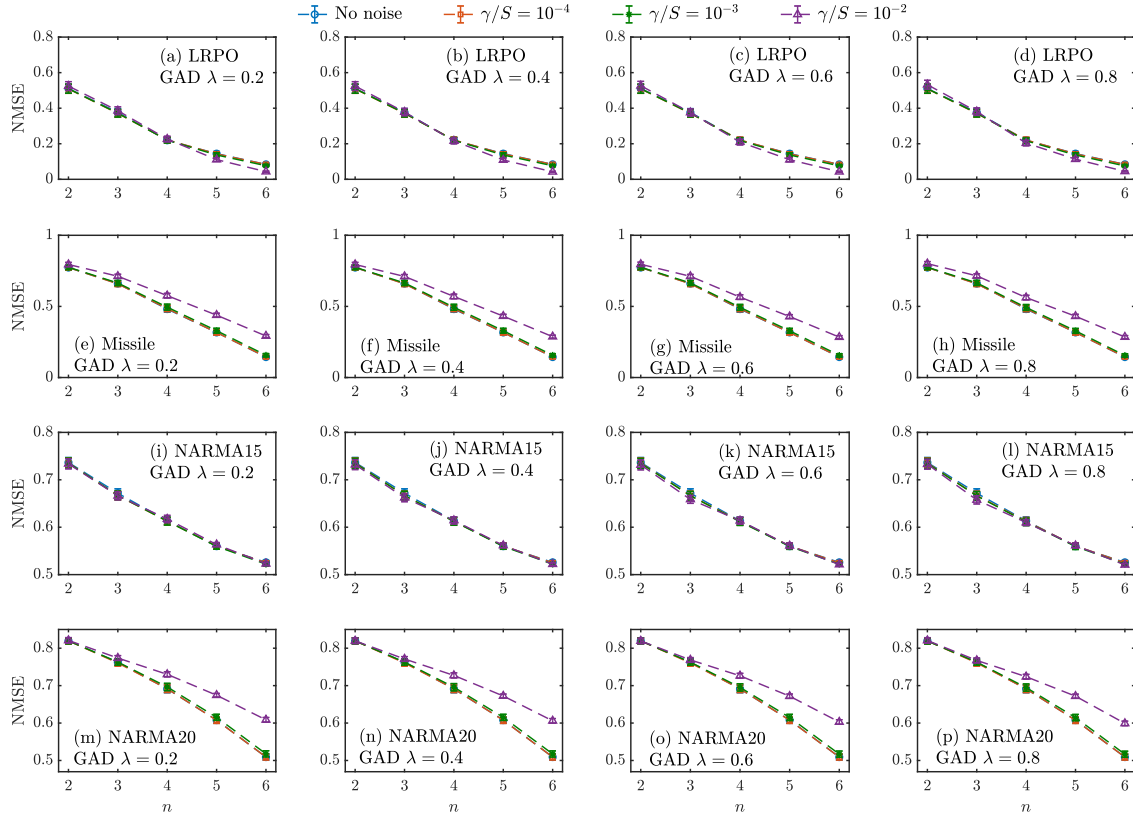


FIG. 8. Average SA NMSE for the LRPO, Missile, NARMA15 and NARMA20 tasks under GAD for  $\lambda = \{0.2, 0.4, 0.6, 0.8\}$

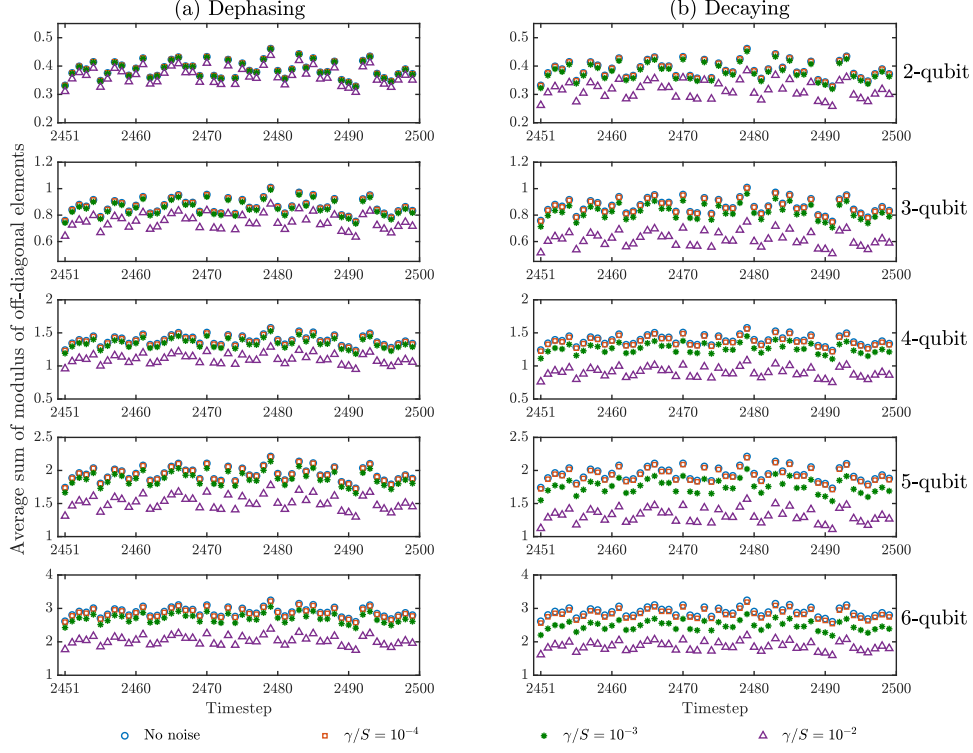


FIG. 9. Average sum of modulus of off-diagonal elements in the density operator, for the last 50 timesteps of the SA samples, under the (a) dephasing noise and (b) decaying noise

### 3. The echo state networks

An ESN with  $m$  reservoir nodes is a type of recurrent neural network with a  $m \times 1$  input matrix  $W_i$ , a  $m \times m$  reservoir matrix  $W_r$  and an  $1 \times m$  output matrix  $W_o$ . The state evolution and output are given by [21]

$$\begin{cases} x_k = \tanh(W_r x_{k-1} + W_i u_k) \\ \hat{y}_k = W_o x_k + w_c, \end{cases}$$

where  $w_c$  is a tunable constant and  $\tanh(\cdot)$  is an element-wise operation.

In the numerical examples, lengths of washout, learning and evaluation phases for ESNs and SA are the same. Given an output sequence  $y$  to be learned, the output weights  $w_c$  and  $W_o$  are optimized via standard least squares to minimize  $\sum_k |y_k - \hat{y}_k|^2$ , for timesteps  $k$  during the training phase. We now detail the experimental conditions for ESNs in various subsections of the numerical experiments (Sec. V).

For the comparison given in Subsection V A, we set the reservoir size to be  $m \in \mathcal{M} = \{10, 20, 30, 40, 50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 800\}$ . Here, the number of computational nodes is  $m + 1$  for each  $m$ . For each computational task and each  $m$ , the average NMSE of 100 ESNs is reported. The average NMSE for ESNs is obtained as follows. For each reservoir size  $m$ , we prepare 100 ESNs with elements of  $W_r$  randomly uniformly chosen  $[-2, 2]$ . Let  $\mathcal{S}$  denote the set of 10 points evenly spaced between  $[0.01, 0.99]$ . For each of the 100 ESNs, we scale the maximum singular value of  $W_r$  to  $\sigma_{\max}(W_r) = s$  for all  $s \in \mathcal{S}$ . This ensures the convergence and fading memory property of ESNs [18]. For each of the chosen  $s$ ,

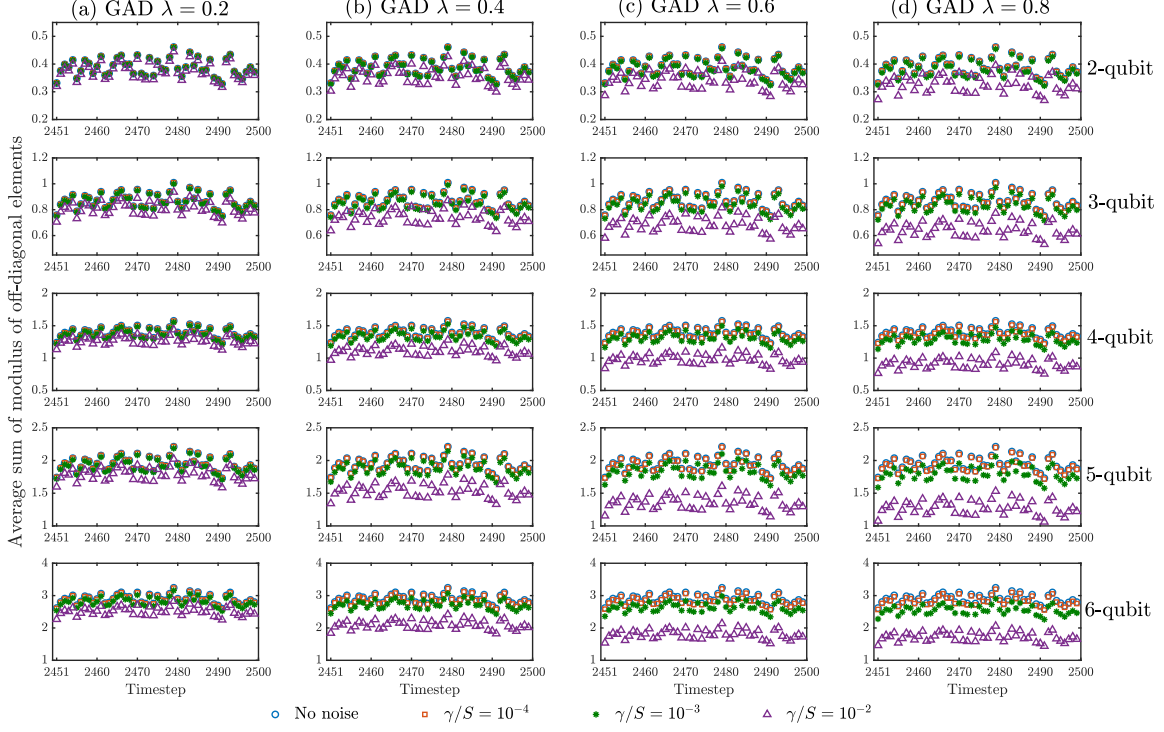


FIG. 10. Average sum of modulus of off-diagonal elements in the density operator, for the last 50 timesteps of the SA samples, under GAD for (a)  $\lambda = 0.2$ , (b)  $\lambda = 0.4$ , (c)  $\lambda = 0.6$  and (d)  $\lambda = 0.8$

the elements of  $W_i$  are randomly uniformly chosen within  $[-\delta, \delta]$ , where  $\delta$  is chosen from the set  $\mathcal{I}$  of 10 points evenly spaced between  $[0.01, 1]$ . Now, for the  $i$ -th ( $i = 1, \dots, 100$ ) ESN with parameter  $(m, s, \delta)$ , we denote its associated NMSE to be  $\text{NMSE}_{(m,s,\delta,i)}$ . For each reservoir size  $m$ , the average NMSE is computed as  $\frac{1}{|\mathcal{S}|} \frac{1}{|\mathcal{I}|} \frac{1}{100} \sum_{s \in \mathcal{S}} \sum_{\delta \in \mathcal{I}} \sum_{i=1}^{100} \text{NMSE}_{(m,s,\delta,i)}$ . Fig. 11 summarizes the average ESNs NMSE for the LRPO, Missile, NARMA15 and NARMA20 tasks.

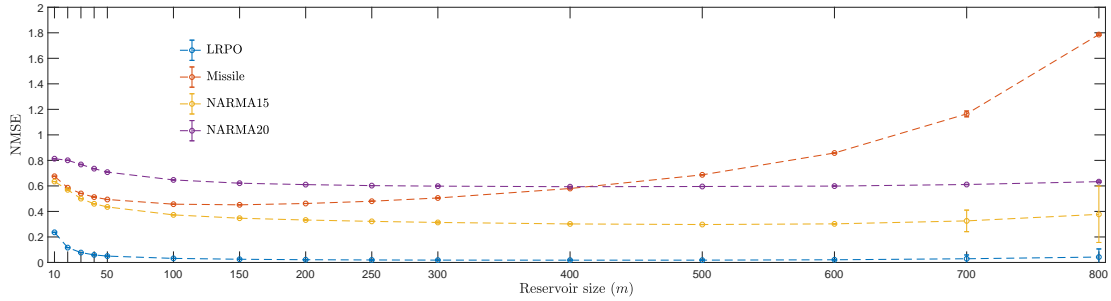


FIG. 11. Average NMSE of ESNs for the LRPO, Missile, NARMA15 and NARMA20 tasks. The data symbols obscure the error bars, which represent the standard error

For the further comparison in Subsection VD, ESNs are simulated to approximate the LRPO, Missile, NARMA15, NARMA20, NARMA30 and NARMA40 tasks. The reservoir size of ESNs for each task is set to be  $m \in \mathcal{M} = \{256, 300, 400, 500\}$ . For each  $m$ , the

number of computational nodes  $\mathcal{C}$  for ESNs is

$$\mathcal{C} \in \mathcal{N}_4 \cup \mathcal{N}_5 \cup \mathcal{N}_6 = \{5, 6, 7, 15, 21, 28, 35, 56, 70, 84, 126, 210, 252\},$$

where  $\mathcal{N}_n$  denotes the chosen numbers of computational nodes for  $n$ -qubit SA defined as follows. Recall that in this experiment, 4-, 5- and 6-qubit SA with varying degrees  $R$  in the output are chosen. For 4-qubit SA,  $R_4 = \{1, \dots, 6\}$  correspond to the number of computational nodes  $\mathcal{N}_4 = \{5, 15, 35, 70, 126, 210\}$ . For 5-qubit SA,  $R_5 = \{1, \dots, 5\}$ , such that  $\mathcal{N}_5 = \{6, 21, 56, 126, 252\}$ . For 6-qubit SA,  $R_6 = \{1, \dots, 4\}$ , such that  $\mathcal{N}_6 = \{7, 28, 84, 210\}$ . To compute the output weights  $W_o$  and  $w_c$  when  $\mathcal{C} < m + 1$ , we first optimize  $W_o$  and  $w_c$  by standard least squares. Then choose  $\mathcal{C} - 1$  elements of  $W_o$  with the largest absolute values and their corresponding elements  $x'_k$  from the state  $x_k$ . These  $\mathcal{C} - 1$  state elements  $x'_k$  are used to re-optimize  $\mathcal{C} - 1$  elements  $W'_o$  of  $W_o$  and  $w'_c$  via standard least squares. At each timestep  $k$ , the full state  $x_k$  evolves, while the output is computed as  $\hat{y}' = W'_o x'_k + w'_c$ . For this numerical experiment, the chosen parameters  $\mathcal{S}$  and  $\mathcal{I}$  of ESNs are the same as above. For the  $i$ -th ESN with parameter  $(m, s, \delta)$ , the number of computational nodes  $\mathcal{C}$  varies. Let  $\text{NMSE}_{(m, \mathcal{C}, s, \delta, i)}$  denotes the corresponding NMSE. For each  $m$  and each  $\mathcal{C}$ , we report the average NMSE computed as  $\frac{1}{|\mathcal{S}|} \frac{1}{|\mathcal{I}|} \frac{1}{100} \sum_{s \in \mathcal{S}} \sum_{\delta \in \mathcal{I}} \sum_{i=1}^{100} \text{NMSE}_{(m, \mathcal{C}, s, \delta, i)}$ .

#### 4. The Volterra series

The discrete-time finite Volterra series with kernel order  $o$  and memory  $p$  is given by [10]

$$\hat{y}_k = h_0 + \sum_{i=1}^o \sum_{j_1, \dots, j_i=0}^{p-1} h_i^{j_1, \dots, j_i} \prod_{l=1}^i u_{k-j_l},$$

where  $u_{k-j}$  is the delayed input,  $h_0$  and  $h_i^{j_1, \dots, j_i}$  are real-valued kernel coefficients (or output weights in our context). Notice that when memory  $p = 1$ , the Volterra series is a map from the current input  $u_k$  to the output  $\hat{y}_k$ . The kernel coefficients are optimized via linear least squares to minimize  $\sum_k |y_k - \hat{y}_k|^2$  during the training phase, where  $y$  is the target output sequence to be learned.

The number of computational nodes, that is the number of kernel coefficients  $h_0$  and  $h_i^{j_1, \dots, j_i}$ , is given by  $(p^{o+1} - p)/(p - 1) + 1$ . We vary the parameters of the Volterra series as follows: for each  $o = \{2, \dots, 8\}$ , choose  $p$  from  $\{2, \dots, 27\}$  such that the maximum number of computational nodes does not exceed 801. Note that for  $o = 1$ , the output of the Volterra series is a linear function of delayed inputs. Since we are interested in nonlinear I/O maps, we choose  $o \geq 2$ . Table 2 summarizes the number of computational nodes as  $o$  and  $p$  vary. Fig. 12 shows the Volterra series NMSE according to the kernel order and memory.

It is observed in Fig. 12 that as the kernel order increases, the Volterra series task performance does not improve. On the other hand, as the memory increases for kernel order 2, the Volterra series task performance improves. The improvement is particularly significant as the memory  $p$  coincides with the delay for NARMA tasks, that is when  $p = \tau_{\text{NARMA}} + 1$ .

---

[1] “IBM Q 20 Tokyo,” <https://www.research.ibm.com/ibm-q/technology/devices/>, Accessed: 2019-04-10.

TABLE 2. Values of  $o$  and  $p$  for the Volterra series and the corresponding number of computational nodes. The empty entries indicate that for the chosen  $o$  and  $p$ , the number of computational nodes exceeds 801

$o \backslash p$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
2	7	13	21	31	43	57	73	91	111	133	157	183	211	241	273	307	343	381	421	463	507	553	601	651	703	757
3	15	40	85	156	259	400	585																			
4	31	121	341	781																						
5	63	364																								
6	127																									
7	255																									
8	511																									

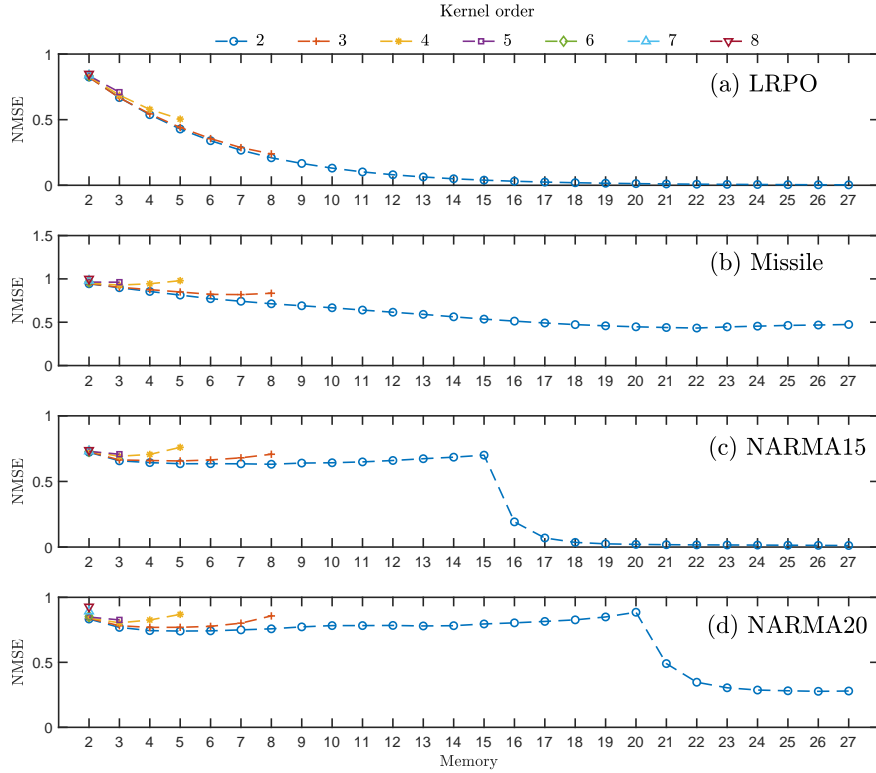


FIG. 12. NMSE of the Volterra series according to kernel order and memory, for the (a) LRPO, (b) Missile, (c) NARMA15 and (d) NARMA20 tasks

- [2] Aaronson, S. and Arkhipov, A., in *Proceedings of the 43rd ACM Symposium on Theory of Computing (STOC)* (2011) pp. 333–342.
- [3] Aleksandrowicz, G. *et al.*, “Qiskit: An open-source framework for quantum computing,” (2019).
- [4] Alvarez-Rodriguez, U., Lamata, L., Escandell-Montero, P., Martín-Guerrero, J. D., and Solano, E., *Scientific reports* **7**, 13645 (2017).
- [5] Appellant, L. *et al.*, *Nat. Commun.* **2**, 468 (2011).
- [6] Atiya, A. F. and Parlos, A. G., *IEEE Trans. Neural Netw.* **11**, 697 (2000).
- [7] Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., and Lloyd, S., *Nature* **549**, 195 (2017).

- [8] Boixo, S., Isakov, S. V., Smelyanskiy, V. N., Babbush, R., Ding, N., Jiang, Z., Bremner, M. J., Martinis, J. M., and Neven, H., *Nature Physics* **14**, 595 (2018).
- [9] Bouten, L., van Handel, R., and James, M. R., *SIAM Rev.* **51**, 239 (2009).
- [10] Boyd, S. and Chua, L., *IEEE Trans. Circuits Syst.* **32**, 1150 (1985).
- [11] Bremner, M. J., Jozsa, R., and Shepherd, D. J., *Proc. Royal Soc. A* **467**, 459 (2010).
- [12] Buehner, M. and Young, P., *IEEE Trans. Neural Netw.* **17**, 820 (2006).
- [13] Dieudonné, J., *Foundations of Modern Analysis* (Read Books Ltd, 2013).
- [14] Dormand, J. R. and Prince, P. J., *Journal of computational and applied mathematics* **6**, 19 (1980).
- [15] Farhi, E., Goldstone, J., and Gutmann, S., “A quantum approximate optimization algorithm,” (2014), arXiv preprint. [Online] Available: <https://arxiv.org/abs/1411.4028>.
- [16] Friedman, J., Hastie, T., and Tibshirani, R., *The elements of statistical learning*, Vol. 1 (Springer series in statistics New York, 2001).
- [17] Fujii, K. and Nakajima, K., *Phys. Rev. Appl.* **8**, 024030 (2017).
- [18] Grigoryeva, L. and Ortega, J.-P., *Neural Networks* **108**, 495 (2018).
- [19] Grigoryeva, L. and Ortega, J.-P., *The Journal of Machine Learning Research* **19**, 892 (2018).
- [20] Gross, J. A., Caves, C. M., Milburn, G. J., and Combes, J., *Quantum Science and Technology* **3**, 024005 (2018).
- [21] Jaeger, H. and Haas, H., *Science* **304**, 5667 (2004).
- [22] Kandala, A., Mezzacapo, A., Temme, K., Takita, M., Brink, M., Chow, J. M., and Gambetta, J. M., *Nature* **549**, 242 (2017).
- [23] Kubrusly, C. S., *Far East Journal of Mathematical Sciences* **22**, 137 (2006).
- [24] Lang, S., *Complex Analysis*, Graduate Texts in Mathematics (Springer-Verlag, 1985).
- [25] Lukoševičius, M., in *Neural networks: Tricks of the trade* (Springer, 2012) pp. 659–686.
- [26] Lukoševičius, M. and Jaeger, H., *Computer Science Review* **3**, 127 (2009).
- [27] Lund, A. P., Bremner, M. J., and Ralph, T. C., *npj Quantum Information* **3**, 15 (2017).
- [28] Maass, W., Natschläger, T., and Markram, H., *Neural Computation* **14**, 2531 (2002).
- [29] McClean, J. R., Romero, J., Babbush, R., and Aspuru-Guzik, A., *New J. Phys.* **18** (2016).
- [30] Mills, M., *IEEE Annals of the History of Computing* **22**, 24 (2011).
- [31] Mitarai, K., Negoro, M., Kitagawa, M., and Fujii, K., *Physical Review A* **98**, 032309 (2018).
- [32] Nakajima, K., Fujii, K., Negoro, M., Mitarai, K., and Kitagawa, M., *Physical Review Applied* **11**, 034021 (2019).
- [33] Ni, X., Verhaegen, M., Krijgsman, A. J., and Verbruggen, H. B., *Engineering Applications of Artificial Intelligence* **9**, 231 (1996).
- [34] Nielsen, M. A. and Chuang, I., “Quantum computation and quantum information,” (2002).
- [35] Otterbach, J. S. *et al.*, “Unsupervised machine learning on a hybrid quantum computer,” (2017), arXiv preprint. [Online] Available: <https://arxiv.org/abs/1712.05771>.
- [36] Pavlov, A., van de Wouw, N., and Nijmeijer, H., in *Control and Observer Design for Nonlinear Finite and Infinite Dimensional Systems*, Lecture Notes in Control and Information Science, Vol. 322, edited by T. Meurer, K. Graichen, and E. D. Gilles (Springer, 2005) pp. 131–146.
- [37] Perez-Garcia, D., Wolf, M. M., Petz, D., and Ruskai, M. B., *Journal of Mathematical Physics* **47**, 083506 (2006).
- [38] Peruzzo, A., McLean, J., Shadbolt, P., Yung, M., Zhou, X., Love, P. J., Aspuru-Guzik, A., and O’Brien, J. L., *Nature Comms* **5** (2013).
- [39] Preskill, J., “Quantum computing in the NISQ era and beyond,” (2018), arxiv preprint, [Online] Available: <https://arxiv.org/abs/1801.00862>.



- [40] Richter, S. and Werner, R. F., Journal of Statistical Physics **82**, 963 (1996).
- [41] Rudin, W. *et al.*, *Principles of mathematical analysis*, Vol. 3 (McGraw-hill New York, 1964).
- [42] Suzuki, M., Progress of Theoretical Physics **46**, 1337 (1971).
- [43] Torrejon, J. *et al.*, Nature **547**, 428 (2017).
- [44] Trotter, H. F., Proceedings of the American Mathematical Society **10**, 545 (1959).
- [45] Vandersypen, L. M., Steffen, M., Breyta, G., Yannoni, C. S., Sherwood, M. H., and Chuang, I. L., Nature **414**, 883 (2001).
- [46] Verstraete, F., Wolf, M. M., and Cirac, J. I., Nature physics **5**, 633 (2009).
- [47] Wang, D., Higgott, O., and Brierley, S., “A generalised variational quantum eigensolver,” (2018), arXiv preprint. [Online] Available: <https://arxiv.org/abs/1802.00171>.