Fundamental Limits to Expressive Capacity of Finitely Sampled Qubit-Based Systems

Fangjun Hu,[1,*] Gerasimos Angelatos,[1,*] Saeed A. Khan,[1] Marti Vives,[1,2] Esin Türeci,[3]
Leon Bello,[1] Graham E. Rowlands,[4] Guilhem J. Ribeill,[4] and Hakan E. Türeci[1]

[1]*Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544, USA*

[2]*Q-CTRL, Santa Monica, CA 90401, USA*

[3]*Department of Computer Science, Princeton University, Princeton, NJ 08544, USA*

[4]*Raytheon BBN, Cambridge, MA 02138, USA*

(Dated: January 3, 2023)

The expressive capacity for learning with quantum systems is fundamentally limited by the quantum sampling noise incurred during measurement. While studies suggest that noise limits the resolvable capacity of quantum systems, its precise impact on learning remains an open question. We develop a framework for quantifying the expressive capacity of qubit-based systems from finite numbers of projective measurements, and calculate a tight bound on the expressive capacity and the corresponding accuracy limit that we compare to experiments on superconducting quantum processors. We uncover the native function set a finitely-sampled quantum system can approximate, called eigentasks. We then demonstrate how low-noise eigentasks improve performance for tasks such as classification in a way that is robust to noise and overfitting. We also present experimental and numerical analyses suggesting that entanglement enhances learning capacity by reducing noise in eigentasks. Our results are broadly relevant to quantum machine learning and sensing applications.

I. INTRODUCTION

Learning with quantum systems is a promising application of near-term quantum processors, with several recent demonstrations in both quantum machine learning (QML) [1–5] and quantum sensing [6–8]. A broad class of such data-driven applications proceed by embedding data into the evolution of a quantum system, where the embedding, dynamics, and extracted outputs via measurement are all governed by a set of general parameters $\boldsymbol{\theta}$. Depending on the learning scheme, different components of this general framework may be trained for optimal performance of a given task. Irrespective of the scheme, however, the fundamental role of the quantum system is that of a high-dimensional feature generator. Given inputs $\boldsymbol{u}$, a set of frequencies for the occurrence of different measurement outcomes act as a feature vector to learn a function $f(\boldsymbol{u})$ that minimizes the chosen loss function (see Fig. 1). The relationship between the physical structure of the model and the function classes that can be expressed with high accuracy, referred to as *expressivity*, is a fundamental question of basic importance to the success of quantum models. Recent results have begun to shed light on this important question and provide guidance on the choice of parameterized quantum models [9–16]. Yet when it comes to experimental implementations, the presence of noise is found to substantially curtail theoretical expectations for performance [1–3].

Given an input $\boldsymbol{u}$ to a general dynamical system, we define its Expressive Capacity (EC) as a measure of the accuracy with which K linearly independent functions $\{f(\boldsymbol{u})\}$ of the input can be constructed from K readout features. This is a suitable generalization to noisy systems of the Information Processing Capacity introduced in Ref. [17]. A central challenge in determining the EC for *quantum* systems is the fundamentally stochastic nature of measurement outcomes. Even when technical noise due to system parameter fluctuations is minimized as in an error-corrected quantum computer, there is a fundamental level of noise, the quantum sampling noise (QSN), which cannot be eliminated in learning with quantum systems. QSN therefore sets a fundamental limit to the EC of any physical system. Although QSN is well-understood theoretically, a formulation of its impact on learning is a challenging task as it is strongly determined by the quantum state of the system relative to the measurement basis, and is highly correlated when entanglement is present. Consequently, the impact of QSN is often ignored [18–21] (with a few exceptions [14, 22]), even though it can place strong constraints on practical optimization [23] and performance [22]. In this article, we develop a mathematical framework to quantify the EC that exactly accounts for the structure of QSN, providing a tight bound for an L-qubit system under S measurements, and illustrate how a mathematical framework for its quantification can guide experimental design for QML applications.

Our work goes beyond simply defining the EC as a figure of merit, however. In particular, we offer a methodology to identify the native function set that is most accurately realizable by a given encoding under finite sampling. Equivalently, we show that this defines a construction of measured features that is optimally robust to noise in readout, thereby revealing how such a quantum system can be optimally employed for learning tasks. Finally, while the strength of the EC lies in its generality, we provide numerical examples that suggest that higher EC is typically indicative of improved performance on specific QML tasks. As such, the EC provides a metric whose optimization can be targeted for improved learning performance in a task-agnostic and parameter-independent manner.

This strategy for defining the noise-constrained EC natu-

* These two authors contributed equally

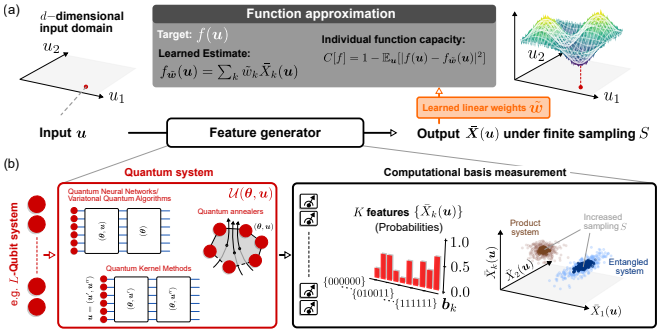

FIG. 1. (a) Representation of the learning framework considered in this work – inputs $\boldsymbol{u}$ are transformed to a set of outputs via a feature generator, here implemented using a finitely-sampled quantum system as shown in (b). Inputs are encoded in the state of a quantum system via a general quantum channel $\mathcal{U}$. Information is extracted from the quantum system via projective measurements in the computational basis. The geometric structure of the quantum sampling noise in the high-dimensional measured feature space can strongly depend on the encoding, and the degree of entanglement generated upon parametric evolution. The learning scheme discussed in the present work optimally leverages the geometric structure of correlated noise. This framework describes a wide range of practical quantum systems, from quantum circuits used in QML, to quantum annealers exhibiting continuous evolution, and beyond, all defined by general parameters $\boldsymbol{\theta}$. As shown in (a), learned estimates for desired functions are constructed via a trained linear estimator $\tilde{\boldsymbol{w}}$ applied to K measured observables $\bar{\boldsymbol{X}}$ of the quantum system, with a resolution limited by quantum sampling noise with finite shots S. Capacity then quantifies the error in the approximation of a target function via this scheme.

rally focuses on accessible noisy output features under a specified measurement scheme, as opposed to unmeasured degrees of freedom. This makes the EC an efficiently-computable quantity in practice, as we demonstrate using both numerical simulations and experiments on IBM Quantum's superconducting multi-qubit processors [24]. Our work also identifies enhancement in measurable quantum correlations as a general principle to increase the EC of quantum systems under finite sampling.

II. LEARNING WITH QUANTUM SYSTEMS

The most general approach to learning from data using a generic quantum system is depicted schematically in Fig. 1. A table with symbols and abbreviations used in the text can be found in Appendix A. For concreteness, we detail a specific realization for L-qubit systems that are measured projectively, which will be analyzed in the remainder of this work. Any learning scheme begins with embedding the data $\boldsymbol{u}$ through a quantum channel parameterized by $\boldsymbol{\theta}$ acting on a known initial state, $\hat{\rho}(\boldsymbol{u};\boldsymbol{\theta}) = \mathcal{U}(\boldsymbol{u};\boldsymbol{\theta})\hat{\rho}_0$. For an L-qubit quantum system, for example, we consider $\hat{\rho}_0 = |0\rangle\langle 0|^{\otimes L}$.

Any computation must be performed using outputs extracted from the quantum system via measurements in a specified basis parameterized by K operators $\{\hat{M}_k\}$, $k = 0, \cdots, K-1$. For a projectively measured L-qubit system, the measurement basis is defined by the $K = 2^L$ projectors $\hat{M}_k = |\boldsymbol{b}_k\rangle\langle\boldsymbol{b}_k|$ corresponding to measurement of bitstrings $\boldsymbol{b}_k$. A single measurement or "shot" yields a discrete outcome $\boldsymbol{b}^{(s)}(\boldsymbol{u})$ for each observable: if the outcome of shot s is state k, then $\boldsymbol{b}^{(s)}(\boldsymbol{u}) \leftarrow \boldsymbol{b}_k$. Measured features are then constructed by ensemble-averaging over S repeated shots: $\bar{X}_k(\boldsymbol{u}) = 1/S \sum_s \delta(\boldsymbol{b}^{(s)}(\boldsymbol{u}), \boldsymbol{b}_k)$. Hence $\bar{X}_k(\boldsymbol{u})$ in this case is the measured frequency of occurrence of the bitstring $\boldsymbol{b}_k$ in S repetitions of the experiment with the same input $\boldsymbol{u}$. These measured features are formally random variables that are unbiased estimators of the expected value of the corresponding observable as computed from $\hat{\rho}(\boldsymbol{u})$: explicitly,

$$\lim_{S\to\infty} \bar{X}_k(\boldsymbol{u}) \equiv x_k(\boldsymbol{u}) = \mathrm{Tr}\{\hat{M}_k \hat{\rho}(\boldsymbol{u};\boldsymbol{\theta})\}, \qquad (1)$$

so that x_k is the quantum mechanical probability of occurrence of the kth bitstring.

In QML theory, it is standard to consider the limit $S \to \infty$, and to thus use expected features $\{x_k(\boldsymbol{u})\}$ for learning. However, for any practical implementation, measured features $\{\bar{X}_k(\boldsymbol{u})\}$ must be constructed under finite S, in which case their fundamentally quantum-stochastic nature can no longer be ignored. This quantum sampling noise, like any other source of noise, can unsurprisingly limit the EC. Completely unlike classical noise sources however, the statistics of quantum sampling noise are strongly determined by the state of the quantum system itself. This leads to a rich noise structure that changes dramatically based on, for example, the entanglement of the generated quantum state, as depicted in Fig. 1. In this work, we exactly account for this structure of quantum sampling noise to quantify its fundamental impact on EC. By further leveraging the complexity and quantum state dependence of sampling noise, we provide a practical, experimentally applicable methodology that maximizes the capacity for learning functions using finitely-sampled quantum systems, and also avoids overfitting in ML tasks.

We begin by observing that $\bar{\boldsymbol{X}}$ are samples from a multinomial distribution with S trials and $K = 2^L$ categories, which can be decomposed into their expected value and an input-dependent sampling noise:

$$\bar{\boldsymbol{X}}(\boldsymbol{u}) = \boldsymbol{x}(\boldsymbol{u}) + \frac{1}{\sqrt{S}}\boldsymbol{\zeta}(\boldsymbol{u}), \qquad (2)$$

where $\boldsymbol{\zeta}(\boldsymbol{u})$ is a zero-mean random vector obeying multinomial statistics. As discussed in Appendix B and C, what makes quantum systems special is the fundamental relationship between the noise $\boldsymbol{\zeta}(\boldsymbol{u})$ and the 'signal' $\boldsymbol{x}(\boldsymbol{u})$. Precisely, the covariance $\boldsymbol{\Sigma}(\boldsymbol{u})$ of $\boldsymbol{\zeta}(\boldsymbol{u})$ depends on the generated quantum state: $\boldsymbol{\Sigma}_{kk'}(\boldsymbol{u}) = \mathrm{Tr}\{\hat{M}_k \hat{M}_{k'} \hat{\rho}(\boldsymbol{u})\} - \mathrm{Tr}\{\hat{M}_k \hat{\rho}(\boldsymbol{u})\}\mathrm{Tr}\{\hat{M}_{k'} \hat{\rho}(\boldsymbol{u})\}$. This *quantum covariance* of the measured observables therefore comprises non-linear functions of the signal $\boldsymbol{x}(\boldsymbol{u})$ itself; at a given S, we will show that this allows for more information to be extracted from systems with more quantum correlations between observables. Note that $\boldsymbol{\zeta}$ can be straightforwardly modified to include other

noise sources, such as gate or measurement errors (see Appendix B 2), with $1/\sqrt{S}$ then interpreted as a general noise strength. However our focus here remains on quantum sampling noise and its fundamental role in learning with quantum systems.

The use of such a quantum system for the learning of functions under finite sampling is then depicted schematically in Fig. 1. For a target function $f(\boldsymbol{u})$, an approximation $f_{\boldsymbol{W}}(\boldsymbol{u})$ is obtained via a linear (for reasons clarified shortly) estimator applied to readout features under finite S, $f_{\boldsymbol{W}}(\boldsymbol{u}) = \boldsymbol{W}\cdot\bar{\boldsymbol{X}}(\boldsymbol{u})$, where $\bar{\boldsymbol{X}} = (\bar{X}_0,\ldots,\bar{X}_{K-1})^T$. To quantify the fidelity of this approximation, we introduce the capacity [14, 17, 20] to construct the target function as the minimum achievable (normalized) mean squared error between the target and its estimate:

$$C[f] = 1 - \min_{\boldsymbol{W}\in\mathbb{R}^K} \frac{\mathbb{E}_{\boldsymbol{u}}[|f(\boldsymbol{u}) - f_{\boldsymbol{W}}(\boldsymbol{u})|^2]}{\mathbb{E}_{\boldsymbol{u}}[|f(\boldsymbol{u})|^2]}. \tag{3}$$

In the above, $\mathbb{E}_{\boldsymbol{u}}$ refers to the expected value with respect to an input distribution $p(\boldsymbol{u})$ over a compact input domain, which can be continuous or discrete: $\mathbb{E}_{\boldsymbol{u}}[f] \equiv \int \mathrm{d}\boldsymbol{u}\, p(\boldsymbol{u}) f(\boldsymbol{u}) \simeq \frac{1}{N}\sum_n f(\boldsymbol{u}^{(n)})$ for i.i.d. sampling obeying $\boldsymbol{u}^{(n)} \sim p(\boldsymbol{u})$ for all $n \in [N]$. Minimizing error in the approximation of $f(\boldsymbol{u})$ by $f_{\boldsymbol{W}}(\boldsymbol{u})$ over the input domain to determine capacity thus requires finding $\tilde{\boldsymbol{w}} = \mathrm{argmin}_{\boldsymbol{W}}\mathbb{E}_{\boldsymbol{u}}[|f - f_{\boldsymbol{W}}(\boldsymbol{u})|^2]$ (via a resource-efficient pseudoinverse). This capacity is constructed such that $0 \le C[f] \le 1$.

The choice of a linear estimator and a mean squared error loss function may appear restrictive at first glance, but the generality of our formalism averts such limitations. For example, the use of a linear estimator applied directly to readout features precludes classical nonlinear post-processing of measurements; however, this is simply to ensure the calculated capacity is a measure of the quantum system itself, and not of a classical nonlinear layer. Importantly, our formalism is general enough to incorporate such processing in a calculation of capacity, via a simple redefinition of readout features $\bar{\boldsymbol{X}}$. Hence the use of a linear estimator does not necessarily lose generality. Secondly, while higher-order loss functions may be used, the mean squared loss effectively describes the Taylor expansion of a wide range of loss functions (see Appendix C 5).

To extend the notion of capacity to a task-independent measure of the expressivity of a physical system, we can evaluate the function capacity over a complete orthonormal set of basis functions $\{f_\ell\}_{\ell\in\mathbb{N}}$, equipped with the inner product $\langle f_\ell, f_{\ell'}\rangle_p = \int_{-1}^{1} f_\ell(\boldsymbol{u}) f_{\ell'}(\boldsymbol{u}) p(\boldsymbol{u}) \mathrm{d}\boldsymbol{u} = \delta_{\ell\ell'}$. The *total Expressive Capacity* (EC) is then $C_T \equiv \sum_{\ell=0}^{\infty} C[f_\ell]$, which effectively quantifies how many linearly-independent functions can be expressed from a linear combination of $\{\bar{X}_k(\boldsymbol{u})\}$. Our main result, which is proven in Appendix C 4, is that the EC for an L-qubit system whose readout features are stochastic variables of the form of Eq. (2) is given by

$$C_T(\boldsymbol{\theta}) = \mathrm{Tr}\left(\left(\mathbf{G} + \frac{1}{S}\mathbf{V}\right)^{-1}\mathbf{G}\right) = \sum_{k=1}^{K} \frac{1}{1+\beta_k^2(\boldsymbol{\theta})/S}. \tag{4}$$

The first equality is written in terms of the expected feature Gram and covariance matrices $\mathbf{G} \equiv \mathbb{E}_{\boldsymbol{u}}[\boldsymbol{x}\boldsymbol{x}^T]$ and $\mathbf{V} \equiv \mathbb{E}_{\boldsymbol{u}}[\boldsymbol{\Sigma}]$ respectively; we later demonstrate that these expected quantities can be accurately estimated under finite S sampling. The second equality expresses the total capacity in a finite-dimensional linear space, in terms of the eigenvalues $\{\beta_k^2\}_{k\in[K]}$ satisfying the generalized eigenvalue problem $\mathbf{V}\boldsymbol{r}^{(k)} = \beta_k^2\mathbf{G}\boldsymbol{r}^{(k)}$. Inspecting this expression, we first note that it is independent of the particular set $\{f_\ell\}_{\ell\in\mathbb{N}}$ chosen, which would have required an evaluation over an infinite set of functions and its numerical evaluation therefore would be subject to inaccuracies due to truncation [17]. Instead, C_T can be interpreted as the sum of capacities to construct K individual functions living in an otherwise infinite-dimensional function space; the identity of these special functions is closely connected with the generalized eigenvectors $\{\boldsymbol{r}^{(k)}\}$, and will be clarified shortly. Secondly, in the absence of noise, $\lim_{S\to\infty} C_T = \mathrm{Rank}\{\mathbf{G}\} = K = 2^L$, provided no special symmetries exist (see Appendix C 6). Such theoretical exponential growth in expressive capacity with L is often cited as a motivator for ML on quantum systems [14, 20, 25]. From the perspective of infinite-shot capacity, this also implies that all L-qubit systems with K measured features are equivalent, regardless of encoding. Such universality has also been pointed out for classical dynamical systems subject to zero input and output noise [17].

However, our expression for C_T is also valid for any *noisy* physical system, corresponding to finite S. In particular, Eq. (4) shows that the EC of a qubit-based physical system satisfies $C_T \le K$ at finite S, and can be fully characterized in terms of the spectrum $\{\beta_k^2\}$, which is ultimately determined by parameters $\boldsymbol{\theta}$ governing the physical system and embedding via the Gram ($\mathbf{G}$) and covariance ($\mathbf{V}$) matrices. Related characterizations of noise-constrained capacity have been attempted for Gaussian quantum systems [22], but to our knowledge no precise formulation exists that also encompasses non-Gaussian systems such as qubit systems. Furthermore, from the perspective of capacity, what makes one embedding or physical system different from another is simply its ability to accurately express functions in the presence of noise. Our expression for C_T thus provides a general, comprehensive, and straightforward metric to assess and compare this capacity across physical systems and their associated embedding under finite S.

Furthermore, via the associated eigenvectors $\{\boldsymbol{r}^{(k)}\}$, our analysis uncovers a finite set of orthogonal functions native to a particular encoding that is maximally resolvable through S measurements. This set of K orthonormal functions, the *eigentasks* $y^{(k)}(\boldsymbol{u}) = \sum_j r_j^{(k)} x_j(\boldsymbol{u})$, can be estimated from measured readout features as described in Appendix D 1. The eigentasks characterize an ordered set of functions that can be constructed with mean squared error β_k^2/S, leading to a natural interpretation of β_k^2 as noise-to-signal (NSR) eigenvalues, determined by fundamental sampling noise. As we will show, this experimentally extractable information can be utilized for optimal learning (with minimal degrees of freedom) with a noisy quantum system.

III. EXPERIMENTAL RESULTS

To demonstrate the above results in practice, we now show how the spectrum $\{\beta_k^2\}$, the EC, and eigentasks can all be computed for real quantum devices in the presence of parameter fluctuations and device noise.

We emphasize at the outset that our approach for quantifying the EC of a quantum system is very general, and can be applied to a variety of quantum system models. For practical reasons, we perform experiments on IBM Quantum (IBMQ) processors, whose dynamics is described by a parameterized quantum circuit containing single and two-qubit gates. However, as an example of the general validity of our approach, in Appendix E we compute the EC for L-qubit quantum annealers via numerical simulations, governed by the markedly different model of continuous-time Hamiltonian dynamics.

On IBMQ devices, resource limitations restrict our computation of EC to 1D inputs u that are uniformly distributed, $p(u) = \mathrm{Unif}[-1, 1]$, see Fig. 2(a). We emphasize that this analysis can be straightforwardly extended to multi-dimensional and arbitrarily-distributed inputs given suitable hardware resources, without modifying the form of the Gram and covariance matrices.

We are only now required to specify the model of the L-qubit system in Eq. (1), which has been left completely general thus far. The specific ansatz we consider is tailored to be natively implementable on IBMQ processors; more general ansatz can also be considered (see Appendix B). It consists of $\tau \in \mathbb{N}$ repetitions of the same input-dependent circuit block depicted in Fig. 2(a). The block itself is of the form $\mathcal{R}_x(\boldsymbol{\theta}^x/2)\mathcal{W}(J)\mathcal{R}_z(\boldsymbol{\theta}^z + \boldsymbol{\theta}^I u)\mathcal{R}_x(\boldsymbol{\theta}^x/2)$, where $\mathcal{R}_{x/z}$ are Pauli-rotations applied qubit-wise, e.g. $\mathcal{R}_z = \prod_l R_z(\theta_l^z + \theta_l^I u)$. The entangling gate acts between physically connected qubits in the device and can be written as $\mathcal{W}(J) = \prod_{\langle l,l' \rangle} \exp\{-i\frac{J}{2}\hat{\sigma}_l^z\hat{\sigma}_{l'}^z\}$.

Note that for this ansatz, the choice $J = 0 \pmod{\pi}$ yields either $\mathcal{W} = \hat{I}$ or $\hat{\sigma}^z \otimes \hat{\sigma}^z$, both of which ensure $\hat{\rho}(u)$ is a product state and measured features are simply products of uncorrelated individual qubit observables – equivalent to a noisy classical system. Starting from this *product system* (PS), tuning the coupling $J \neq 0 \pmod{\pi}$ provides a controllable parameter to realize an *entangled system* (ES). This control enables us to address a natural question regarding EC of quantum systems under finite S: what is the dependence of EC and realizable eigentasks on J, and hence on quantum correlations?

This calculation of EC requires extracting measured features from the quantum circuit under input u, one example of which is shown for the IBMQ *ibmq_perth* device in Fig. 2(a), for $S = 2^{14}$. For comparison, we also show ideal-device simulations (no device noise), where slight deviations are observed. The agreement with the experimental feature is improved when the effects of gate and readout errors, and qubit relaxation are included, hereafter referred to as "device noise" simulations, highlighting the non-negligible role of device errors.

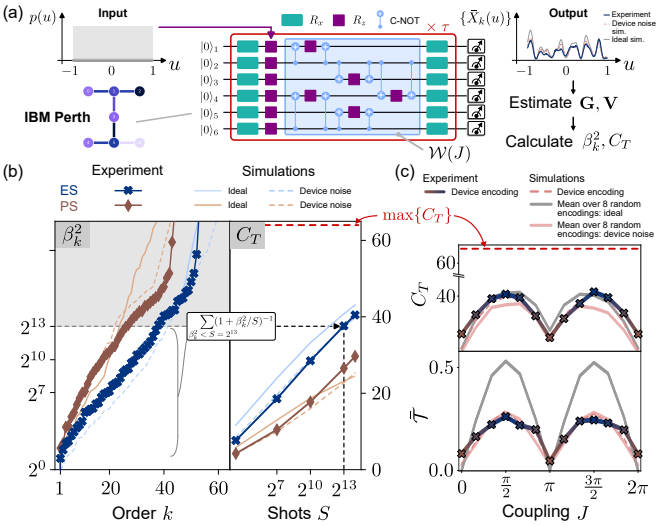

FIG. 2. (a) IBMQ Perth device and quantum circuit schematic for EC calculation, and classification task in Fig. 3. Here $\tau = 3$ layers, and random qubit rotation parameters are $\theta_l^{x/z} \sim \mathrm{Unif}[0, 2\pi]$ and $\theta_l^I \sim \mathrm{Unif}[0, 10\pi]$. On the right, the specific feature plotted is $\bar{X}_1(u) = P_{000001}(u)$ for $S = 2^{14}$ shots. (b) Left panel: Device NSR spectrum β_k^2 for ES, $J = \pi/2$ (blue crosses) and PS, $J = 0$ (brown diamonds). Ideal (solid) and device noise (dashed) simulations are also shown. Note the agreement between device and simulation, along with distortion from more direct exponential growth in β_k^2 with k in the ideal case, due to device errors. Right panel: C_T vs. S calculated from the left panel. At a given S, the C_T can be approximated by performing the indicated sum over all $\beta_k^2 < S$. (c) EC (top panel) and ETC (lower panel) under $S = 2^{14}$ from the IBM device, and device noise simulations (dashed peach). Average metrics over 8 random encodings for device noise (solid peach) and ideal (solid gray) simulations are also shown. The $S \to \infty$ EC of these encodings always attains the $\max\{C_T\} = 64$, indicated in dashed red.

The measured features under finite S are used to estimate the Gram and covariance matrices (see Appendix D), and to thus solve the eigenproblem for NSR eigenvalues $\{\beta_k^2\}$. Typical NSR spectra computed for two random encodings on the device are shown in Fig. 2(b), for $J = 0$ (PS) and $J = \pi/2$ (ES), together with spectra from device noise simulations, with which they agree well. We note that at lower k, the device NSR eigenvalues are larger than those from ideal simulations, due to device noise contributions. For larger k, device results deviate from the pure exponential increase (with order) seen in ideal simulations. The deviation is captured by device noise simulations and can therefore be attributed to device errors. The NSR spectra therefore can serve as effective diagnostic tools for quantum processors and encoding schemes. More examples will be provided later in the discussion.

The NSR spectra can be used to directly compute the EC of the corresponding quantum device for finite S, via Eq. (4). As a rule of thumb, at a given S only NSR eigenvalues $\beta_k^2 \lesssim S$ contribute substantially to the EC. An NSR spectrum with a flatter slope therefore has more NSR eigenvalues below S,

which gives rise to a higher capacity. Fig. 2(b) shows that the ES generally exhibits an NSR spectrum with a flatter slope than the PS, yielding a larger capacity for function approximation across all sampled S.

To more precisely quantify the role of entanglement and quantum correlations in EC, we introduce the *expected total correlation* (ETC) of the measured state over the input domain of u [26, 27],

$$\bar{\mathcal{T}} = \mathbb{E}_u \left[\sum_{l=1}^{L} \mathrm{S}(\hat{\rho}_l^M(u)) - \mathrm{S}(\hat{\rho}^M(u)) \right], \tag{5}$$

where $\hat{\rho}^M$ is the measured state: $\hat{\rho}^M(u) \equiv \sum_k \hat{\rho}_{kk}(u) |\boldsymbol{b}_k\rangle\langle\boldsymbol{b}_k|$ and S is the von Neumann entropy (see Appendix G). We now compute EC and ETC using $S = 2^{14}$ in Fig. 2(c) as a function of J, together with both ideal and device noise simulations of the same. We note that product states by definition have $\bar{\mathcal{T}} = 0$ [28]; this is seen in ideal simulations for $J = 0 \pmod{\pi}$. However, the actual device retains a small amount of correlation at this operating point, which is reproduced by device noise simulations. This can be attributed to gate or measurement errors as well as cross-talk, especially relevant for the transmon-based IBMQ platform with a parasitic always-on ZZ coupling.

With increasing J, $\bar{\mathcal{T}}$ increases and peaks around $J \sim \pi/2 \pmod{\pi}$; interestingly, C_T also peaks for the same coupling range. From the analogous plot of EC, we clearly see that at finite S, increased ETC appears directly correlated with higher EC. We have observed very similar behaviour using completely different models of quantum systems (see Appendix Fig. 5 [29, 30]). This indicates the utility of enhancing quantum correlations as a means of improving the general expressivity of quantum systems.

However, we see that at finite S, even with increased quantum correlations, the maximum EC is still substantially lower than the upper bound of $K = 64$. Note that this remains true even for ideal simulations, and over several random encodings, so the underperformance cannot be attributed to device noise or poor ansatz choice respectively. These results clearly indicate that the resulting sampling noise at finite S is the fundamental limitation for QML applications on this particular IBM device, rather than other types of noise sources and errors.

IV. A ROBUST APPROACH TO LEARNING

While we have demonstrated the EC as an efficiently-computable metric of general expressivity of a noisy quantum system, some important practical questions arise. First, does the general EC metric have implications for practical performance on *specific* QML tasks? Secondly, given the limiting – and unavoidable – nature of correlated sampling noise, does the EC provide any insights on optimal learning using a particular noisy quantum system and the associated embedding?

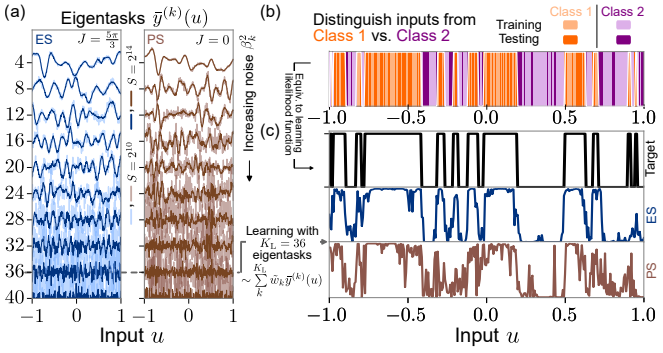

FIG. 3. (a) Device eigentasks for ES (left) and PS (right), constructed from noisy features at $S = 2^{10}$ and $S = 2^{14}$. (b) Classification demonstration on IBMQ Perth. Binary distributions to be classified over the input domain are shown. (c) The classification task can be cast as learning the likelihood function separating the two distributions; this target function is shown in the upper panel. Lower panels show the trained estimate of this target using outputs from the ES and PS respectively, using $K_\mathrm{L} = 36$ eigentasks with $S = 2^{14}$.

Our formulation addresses both these important questions naturally, as we now discuss. Beyond being a simple figure of merit, we show in the Appendix C that the EC is precisely the sum of capacities to approximate a particular set of orthogonal functions native to the given noisy quantum system: the eigentasks. Crucially, these eigentasks $\bar{y}^{(k)}(u) = \sum r_j^{(k)} \bar{X}_j(u)$ can be directly estimated from a noisy quantum system via the generalized eigenvectors $\{\boldsymbol{r}^{(k)}\}$, and are ordered by their associated NSR $\{\beta_k^2\}$. We show a selection of estimated eigentasks from IBMQ, for an ES ($J = 5\pi/3$) and PS ($J = 0$) in Fig. 3(a). For both systems, the increase in noise with eigentask order is apparent when comparing two sampling values, $S = 2^{10}$ and $S = 2^{14}$. Furthermore, for any order k, eigentasks for the PS are visibly noisier than the ES; this is consistent with NSR eigenvalues for PS being larger than those for ES, as seen in Fig. 2(b). This ability to more accurately resolve eigentasks provides a complementary perspective on the higher expressive capacity of ES in comparison to PS.

The resolvable eigentasks of a finitely-sampled quantum system are intimately related to its performance at specific QML applications. To demonstrate this result, we consider a concrete application: a binary classification task that is not linearly-separable. Samples $u^{(n)}$, $n \in [N]$, obeying the same distribution $p(u)$ for $u \in [-1, 1]$ as considered for the EC evaluation, are separated into two classes, as depicted in Fig. 3(b). A selection of $N_\mathrm{train} = 150$ total samples - with equal numbers from each class - are input to the IBMQ device, and readout features $\bar{\boldsymbol{X}}(u^{(n)})$ are extracted using $S = 2^{14}$ shots. A linear estimator applied to these features is then trained using logistic regression to learn the class label associated with each input. Finally, the trained IBMQ device is used to predict class labels of $N_\mathrm{test} = 150$ distinct input samples for testing.

This task can equivalently be cast as one of learning the likelihood function that discriminates the two input distribu-

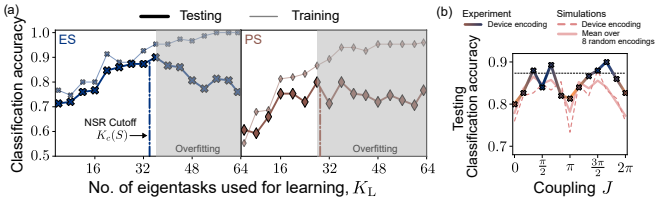

FIG. 4. (a) Training (light) and testing (dark) accuracy for an ES and PS in blue and brown respectively, as a function of number of eigentasks used in learning. The optimal test set performance is found near the NSR cutoff $K_c(S)$ (dash-dotted lines) informed by the quantum system's NSR spectra. In all figures, the IBMQ Perth device is sampled with $S = 2^{14}$, and the training and test sets consist of 150 random points. (b) Testing set classification accuracy as a function of J for our optimal learning method. The average of simulated encodings is shown in solid peach, and the horizontal line shows the best performance of a software neural network with $K_L = 36$ parameters for comparison.

tions, shown in Fig. 3(c), with minimum error. The set of up to K_L eigentasks $\bar{y}^{(k)}(u)$, where $K_L \leq K$, serves as the native basis of readout features used to approximate *any* target function using the quantum system. The noisier eigentasks of the PS therefore limit the accuracy with which it can be used to learn the target, in comparison to the ES. This is clear from the learned estimates shown in Fig. 3(c), using an equal number of $K_L = 36$ eigentasks to ensure a fair comparison. The higher approximation capacity translates to improved classification performance, as we show via the training and testing classification accuracy in Fig. 4(a) for both ES and PS. We plot both as a function of the number of eigentasks K_L used for learning, from which it is clear that the maximum testing accuracy using the ES exceeds that of the PS.

However, using eigentasks ordered by NSR reveals even more about learning using noisy quantum systems, and provides a path towards optimal learning. While intuition suggests that using more eigentasks can only be beneficial, weights learned when training with noisier eigentasks may not generalize well to unseen samples. For example, using all eigentasks ($K_L = K$) yields a test accuracy far lower than that found in training. The observed deviation is a distinct signature of overfitting: the optimized estimator learns noise in the training set, and thus loses generalizability in testing. Crucially, an optimal number of eigentasks clearly emerges, around $K_L \simeq K_c(S) = \max_k\{\beta_k^2 < S\}$, for which the NSR eigenvalue is closest to S. Eigentasks $k > K_c$ typically contribute more 'noise' to the function approximation task than 'signal'. Excluding these eigentasks therefore limits overfitting without adversely impacting performance.

Fig. 4(b) also shows the classification accuracy as J is varied, where we highlight the striking similarity with Fig. 2(c): encodings with larger quantum correlations and thus higher expressive capacity will perform generically better on learning tasks in the presence of noise, because they generate a larger set of eigentasks that can be resolved at a given sampling S. The NSR spectra and eigentasks therefore provide a natural truncation scheme to maximise testing accuracy, avoiding overfitting without any additional regularization (see also Appendix H and I).

V. DISCUSSION

We have developed a straightforward approach to quantify the expressive capacity of any qubit-based system in the presence of fundamental sampling noise. Our analysis is built upon an underlying framework that determines the native function set that can be most robustly realized by a finitely-sampled quantum system: its eigentasks. We use this framework to introduce a methodology for optimal learning using noisy quantum systems, which centers around identifying the minimal number of eigentasks required for a given learning task. The resulting learning methodology is resource-efficient and robust to overfitting. We demonstrate that eigentasks can be efficiently estimated from experiments on real devices using a limited number of training points and finite shots. We also demonstrate across two distinct qubit evolution ansätze that the presence of measured quantum correlations enhances expressive capacity. Our work has direct application to the design of circuits for learning with qubit-based systems. In particular, we propose the optimization of expressive capacity as a meaningful goal for the design of quantum circuits with finite measurement resources.

ACKNOWLEDGEMENT

This research was developed with funding from the DARPA contract HR00112190072, AFOSR award FA9550-20-1-0177, and AFOSR MURI award FA9550-22-1-0203. The views, opinions, and findings expressed are solely the authors and not the U.S. government.

[1] E. Grant, M. Benedetti, S. Cao, A. Hallam, J. Lockhart, V. Stojevic, A. G. Green, and S. Severini, Hierarchical quantum classifiers, npj Quantum Information **4**, 1 (2018).

[2] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, Supervised learning with quantum-enhanced feature spaces, Nature **567**, 209 (2019).

[3] F. Tacchino, P. Barkoutsos, C. Macchiavello, I. Tavernelli, D. Gerace, and D. Bajoni, Quantum implementation of an artificial feed-forward neural network, Quantum Science and Technology **5**, 044010 (2020).

[4] J. Chen, H. I. Nurdin, and N. Yamamoto, Temporal Information Processing on Noisy Quantum Computers, Physical Review Applied **14**, 024065 (2020).

[5] Y. Suzuki, Q. Gao, K. C. Pradel, K. Yasuoka, and N. Yamamoto, Natural quantum reservoir computing for temporal information processing, Scientific Reports **12**, 1353 (2022).

[6] J. J. Meyer, Fisher Information in Noisy Intermediate-Scale Quantum Applications, Quantum **5**, 539 (2021).

[7] Z. Ma, P. Gokhale, T.-X. Zheng, S. Zhou, X. Yu, L. Jiang, P. Maurer, and F. T. Chong, Adaptive Circuit Learning for Quantum Metrology, arXiv:2010.08702 [quant-ph] (2021).

[8] C. D. Marciniak, T. Feldker, I. Pogorelov, R. Kaubruegger, D. V. Vasilyev, R. van Bijnen, P. Schindler, P. Zoller, R. Blatt, and T. Monz, Optimal metrology with programmable quantum sensors, Nature **603**, 604–609 (2022).

[9] S. Sim, P. D. Johnson, and A. Aspuru-Guzik, Expressibility and Entangling Capability of Parameterized Quantum Circuits for Hybrid Quantum-Classical Algorithms, Advanced Quantum Technologies **2**, 1900070 (2019).

[10] R. LaRose and B. Coyle, Robust data encodings for quantum classifiers, Phys. Rev. A **102**, 032420 (2020).

[11] M. Schuld, R. Sweke, and J. J. Meyer, Effect of data encoding on the expressive power of variational quantum-machine-learning models, Physical Review A **103**, 032430 (2021).

[12] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner, The power of quantum neural networks, Nature Computational Science **1**, 403–409 (2021).

[13] Y. Wu, J. Yao, P. Zhang, and H. Zhai, Expressivity of quantum neural networks, Phys. Rev. Research **3**, L032049 (2021).

[14] L. G. Wright and P. L. McMahon, The Capacity of Quantum Neural Networks, arXiv:1908.01364 [quant-ph] (2019).

[15] Y. Du, M.-H. Hsieh, T. Liu, and D. Tao, Expressive power of parametrized quantum circuits, Physical Review Research **2**, 033125 (2020).

[16] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, PRX Quantum **3**, 010313 (2022).

[17] J. Dambre, D. Verstraeten, B. Schrauwen, and S. Massar, Information Processing Capacity of Dynamical Systems, Scientific Reports **2**, 514 (2012).

[18] K. Fujii and K. Nakajima, Harnessing Disordered-Ensemble Quantum Dynamics for Machine Learning, Physical Review Applied **8**, 024030 (2017).

[19] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, Parameterized quantum circuits as machine learning models, Quantum Science and Technology **4**, 043001 (2019).

[20] R. Martínez-Peña, J. Nokkala, G. L. Giorgi, R. Zambrini, and M. C. Soriano, Information processing capacity of spin-based quantum reservoir computing systems, Cognitive Computation (2020).

[21] M. Schuld and F. Petruccione, *Machine Learning with Quantum Computers*, Quantum Science and Technology (Springer International Publishing, 2021).

[22] J. García-Beni, G. L. Giorgi, M. C. Soriano, and R. Zambrini, Scalable photonic platform for real-time quantum reservoir computing, arXiv:2207.14031 [quant-ph] (2022).

[23] A. Arrasmith, M. Cerezo, P. Czarnik, L. Cincio, and P. J. Coles, Effect of barren plateaus on gradient-free optimization, Quantum **5**, 558 (2021).

[24] IBM Quantum, https://quantum-computing.ibm.com (2022).

[25] W. D. Kalfus, G. J. Ribeill, G. E. Rowlands, H. K. Krovi, T. A. Ohki, and L. C. G. Govia, Hilbert space as a computational resource in reservoir computing, Physical Review Research **4**, 033007 (2022).

[26] V. Vedral, The role of relative entropy in quantum information theory, Reviews of Modern Physics **74**, 197 (2002).

[27] K. Modi, T. Paterek, W. Son, V. Vedral, and M. Williamson, Unified view of quantum and classical correlations, Physical Review Letters **104**, 080501 (2010).

[28] M. A. Nielsen and I. Chuang, *Quantum computation and quantum information* (Cambridge University Press, 2010).

[29] V. Giovannetti, S. Lloyd, and L. Maccone, Quantum metrology, Physical Review Letters **96**, 010401 (2006).

[30] R. Martínez-Peña, G. L. Giorgi, J. Nokkala, M. C. Soriano, and R. Zambrini, Dynamical phase transitions in quantum reservoir computing, Phys. Rev. Lett. **127**, 100502 (2021).

Appendix A: Table of Symbols and Abbreviations

Abbreviations	
NISQ	Noisy Intermediate Scale Quantum
(Q)ML	(Quantum) Machine Learning
QSN	Quantum Sampling Noise
VQC	Variational Quantum Circuits
PS	Product System
ES	Entangled System
EC	Total Expressive Capacity, C_T
ETC	Expected Total Correlation, $\bar{\mathcal{T}}$
Symbols and notation	
S	Number of shots
N	Number of inputs
L	Number of qubits
K	$\equiv 2^L$, number of measured features
$\boldsymbol{u}$	Input
$\boldsymbol{\theta}$	Quantum system parameters
$\hat{\rho}$	Generated quantum state
$\hat{M}_k$	Measured observable
$\boldsymbol{W}$	Output weights (can be untrained)
$\tilde{\boldsymbol{w}}$	Optimal learned output weights on S-finite readout data
$\mathscr{L}$	Loss function
$\boldsymbol{b}_k$	Label for eigenstate of $\hat{M}_k$
$\boldsymbol{b}^{(s)}$	Measurement outcome for shot s
x_k	Expected features $\mathrm{Tr}\{\hat{M}_k\hat{\rho}\}$
$X_k^{(s)}$	Observed bit in shot s
$\bar{X}_k$	Empirical observed feature $1/S\sum_s \delta(\boldsymbol{b}^{(s)}, \boldsymbol{b}_k)$
ζ_k	Noise part in $\bar{X}_k$
$\mathbf{G}$	Gram matrix of expected features $\{x_k\}$
$\mathbf{V}$	Expected covariance matrix of random variable $X_k^{(s)}(\boldsymbol{u})$
$\mathbf{R}$	Noise-to-Signal matrix
β_k^2	Eigen-NSR
$y^{(k)}$	Principal feature
$\boldsymbol{r}^{(k)}$	Combination coefficients in $y^{(k)} = \sum_{k'} r_{k'}^{(k)} x_{k'}$
$\bar{y}^{(k)}$	$\equiv \sum_{k'} r_{k'}^{(k)} \bar{X}_{k'}$, noisy eigentask
$\xi^{(k)}$	$\equiv \sum_{k'} r_{k'}^{(k)} \zeta_{k'}$, noise part in $\bar{y}^{(k)}$
$\hat{O}_k$	$\equiv \sum_{k'} r_{k'}^{(k)} \lvert\boldsymbol{b}_{k'}\rangle\langle\boldsymbol{b}_{k'}\rvert$, optimal measurement basis
$\hat{\rho}^M$	$\equiv \sum_k \hat{\rho}_{kk}(u) \lvert\boldsymbol{b}_k\rangle\langle\boldsymbol{b}_k\rvert$, post-measurement state
$K_c(S)$	Cutoff index where β_k^2 reaches S
$\widetilde{(\cdot)}_N$	Quantity obtained from finite N sampling data
$\widetilde{(\cdot)}$	Large N limit, that is $\lim_{N\to\infty} \widetilde{(\cdot)}_N$

TABLE I. Table of notations.

Appendix B: Feature maps using quantum systems

1. Details of input encodings into quantum systems

In the main text, we introduce the idea of encoding inputs into the state of a quantum system via a parameterized quantum channel, reproduced below:

$$\hat{\rho}(\boldsymbol{u};\boldsymbol{\theta}) = \mathcal{U}(\boldsymbol{u};\boldsymbol{\theta})\hat{\rho}_0 \tag{B1}$$

Our analysis of EC presented in this work does not depend on the precise details of the quantum channel $\mathcal{U}$. For practical calculations, however, we have to consider concrete models, about which we provide more details in this section.

To describe these models, we begin by first limiting to 1-D inputs as analyzed in the main text; generalizations to multi-dimensional inputs $\boldsymbol{u}$ are straightforward. Then, we write Eq. (B1) in the form

$$\hat{\rho}(u;\boldsymbol{\theta}) = \mathcal{B}(u;\boldsymbol{\theta})\hat{\rho}_0\mathcal{B}^\dagger(u;\boldsymbol{\theta}) \tag{B2}$$

In the main text, we have considered a model for dynamics of an L-qubit quantum system that is natively implementable on modern quantum computing platforms: namely the ansatz of quantum circuits with single and two-qubit gates. In this case, which we refer to as the *circuit ansatz* (or *C-ansatz* for short), the operator $\mathcal{B}(u;\boldsymbol{\theta})$ takes the precise form

$$\mathcal{B}(u;\boldsymbol{\theta}) = \left[\mathcal{R}_x\left(\frac{\boldsymbol{\theta}^x}{2}\right)\mathcal{W}(J)\mathcal{R}_z\left(\boldsymbol{\theta}^z + \boldsymbol{\theta}^I u\right)\mathcal{R}_x\left(\frac{\boldsymbol{\theta}^x}{2}\right)\right]^\tau \qquad \textit{(C-ansatz)} \tag{B3}$$

For completeness, we recall that $\mathcal{R}_{x/z}$ are Pauli-rotations applied qubit-wise, e.g. $\mathcal{R}_z = \prod_l R_z(\theta_l^z + \theta_l^I u)$, while the entangling gate acts between physically connected qubits in the device and can be written as $\mathcal{W}(J) = \prod_{\langle l,l'\rangle}\exp\{-i\frac{J}{2}\hat{\sigma}_l^z\hat{\sigma}_{l'}^z\}$. We emphasize here again that $\tau \in \mathbb{N}^+$ is an integer, representing the number of repeated blocks in the C-ansatz encoding. We note that the actual operations implemented on IBMQ processors also include dynamics due to noise, gate, and measurement errors. As discussed in the main text, the EC of a quantum system can be computed in the presence of these more general dynamics, and is sensitive to the limitations introduced by them.

An alternative ansatz which we analyze in this SI, is where the operator $\mathcal{B}(u;\boldsymbol{\theta})$ describes continuous Hamiltonian dynamics. This ansatz is relevant to computation with general quantum devices, such as quantum annealers and more generally quantum simulators. In this case, which we refer to as the *Hamiltonian ansatz* (or *H-ansatz* for short),

$$\mathcal{B}(u;\boldsymbol{\theta}) = \exp\{-i\hat{H}(u)t\},\ \hat{H}(u) = \hat{H}_0 + u\cdot\hat{H}_1 \qquad \textit{(H-ansatz)} \tag{B4}$$

Here t is a continuous parameter defining the evolution time; and $\hat{H}_0 = \sum_{l,l'}^L J_{\langle l,l'\rangle}\hat{\sigma}_l^z\hat{\sigma}_{l'}^z + \sum_{l=1}^L h_l^x\hat{\sigma}_l^x + \sum_{l=1}^L h_l^z\hat{\sigma}_l^z$ and $\hat{H}_1 = \sum_{l=1}^L h_l^I\hat{\sigma}_l^z$. The transverse x-field strength $h_l^x = \bar{h}^x + \varepsilon_l^x$ and longitudinal z-drive strength $h_l^{z,I} = \bar{h}^{z,I} + \varepsilon_l^{z,I}$ are all randomly chosen and held fixed for a given realization of the quantum system,

$$\varepsilon_l^{x,z,I} \sim h_{\text{rms}}^{x,z,I}\,\mathcal{N}(0,1), \tag{B5}$$

where $\mathcal{N}(0,1)$ defines the standard normal distribution with zero mean and unit variance. We consider nearest-neighbor interactions $J_{l,l'}$, which can be constant $J_{l,l'} \equiv J$, or drawn from $J_{l,l'} \sim \text{Unif}[0, J_{\max}]$, where $\text{Unif}[a,b]$ is a uniform distribution with non-zero density within $[a,b]$.

As an aside, we note that the C-ansatz quantum channel described by Eq. (B3) can be considered a Trotterization-inspired implementation of the H-ansatz in Eq. (B4). In particular, if we set $\theta^{x/z/I} = h^{x/z/I}\Delta\cdot\tau$, where $t = \Delta\cdot\tau$, and consider the limit $\Delta \to 0$ while keeping t fixed, Eq. (B3) corresponds to a Trotterized implementation of Eq. (B4). This correspondence is chosen for practical reasons, but is not necessary in our analysis.

The parameterized quantum channel characterizes how information is injected into the quantum system and processed by it; however, to probe information from the quantum system, one must apply an appropriate and feasible quantum measurement. For extract information efficiently, we consider a wide family of observable $\hat{M}_k$: the only restriction of these observables is that they must be a product of local observables, $\hat{M}_k = \hat{o}_1 \otimes \cdots \otimes \hat{o}_L$, which *mutually commute* with each other (meaning they are are simultaneously measurable). We consider two general schemes. The first one is the probability representation $\hat{o}_l \in \{|0\rangle\langle 0|, |1\rangle\langle 1|\}$, while the second is the spin moments representation, $\hat{o}_l \in \{\hat{I}, \hat{\sigma}^z\}$; the former representation is used throughout the main text. We will show below that these two readout schemes are equivalent up to a unitary transformation.

2. Extracting output features under finite sampling: expressions for features and covariances

Following evolution of the quantum system under the input-dependent Hamiltonian given by Eq. (B4), we extract certain measurable observables that are used as outputs for any learning task. The form of observables is again chosen for compliance with measurement protocols native to near-term quantum computing implementations: we consider Pauli z basis measurements only (although this can be generalized easily). This means our algorithm has access only to diagonal terms in $\hat{\rho}(u)$. We abbreviate vectors $\vec{M}_k, \vec{\rho}(u) \in \mathbb{R}^K$ such that $(\vec{M}_{k'})_k = (\hat{M}_{k'})_{kk}$ and $(\vec{\rho}(u))_k = \hat{\rho}(u)_{kk}$. Then one can check for $\{+1,-1\}$ readout: $\vec{M}_k\cdot\vec{M}_{k'} = K\delta_{jj'}$, and the readout features can be expressed into dot product form $x_k(u) = \text{Tr}\left\{\hat{M}_k\hat{\rho}(u)\right\} = \vec{M}_k\cdot\vec{\rho}(u)$. In

QRC, we hope to make full use of all functions in family $\{(\vec{\rho}(u))_k\}_{k\in[K]}$ as readout features. The collection of all readout features

$$\boldsymbol{x}(u) = \begin{pmatrix} x_0(u) \\ x_1(u) \\ \vdots \\ x_{K-1}(u) \end{pmatrix} = \begin{pmatrix} \vec{M}_0^T \\ \vec{M}_1^T \\ \vdots \\ \vec{M}_{K-1}^T \end{pmatrix} \vec{\rho}(u) =: U\vec{\rho}(u), \tag{B6}$$

The orthonormality of $\{\vec{M}_k\}_{k\in[K]}$ implies that U is unitary up to an overall constant (in fact, $U = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^{\otimes L}$ is the *Hadamard matrix* [28]). This unitarity implies that the above transformation is information-preserving. In particularly, this guarantees the ability to reconstruct the *diagonal* QRC density matrix elements (via tomography), $\vec{\rho}(u) = U^{-1}\boldsymbol{x}(u)$, simply computing the required inverse via the numerically-robust relationship $U^{-1} = \frac{1}{K}U^T$.

If each qubit has a readout error ϵ, that is, it will flip $|0\rangle \leftrightarrow |1\rangle$. Then the transition probability of reading out $|\boldsymbol{b}_{k'}\rangle$ from $|\boldsymbol{b}_k\rangle$ will be $\epsilon^{d(\boldsymbol{b}_k,\boldsymbol{b}_{k'})}(1-\epsilon)^{L-d(\boldsymbol{b}_k,\boldsymbol{b}_{k'})}$ where $d(\boldsymbol{b}_k, \boldsymbol{b}_{k'})$ is the Hamming distance between $\boldsymbol{b}_k$ and $\boldsymbol{b}_{k'}$. Thus, readout errors can furthermore be mathematically modeled by one more transition matrix (more precisely, a *stochastic matrix*):

$$\boldsymbol{x}(u) = U \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}^{\otimes L} \vec{\rho}(u). \tag{B7}$$

The covariance of the $\boldsymbol{X}(u) \in \{+1, -1\}^L$ (the random features for individual shot $S = 1$) can also be expressed easily:

$$\mathbb{V}[\boldsymbol{X}(u)] = U\left(\mathrm{diag}(\vec{\rho}(u)) - \vec{\rho}(u)\cdot\vec{\rho}(u)^T\right)U^T \tag{B8}$$

where $\mathrm{diag}(\vec{v})$ is a diagonal matrix that has the elements of $\vec{v}$ as entries. To prove this expression, it suffices to verify that the second order moments are entries

$$\mathbb{V}[\boldsymbol{X}(u)]_{k_1k_2} \equiv \mathrm{Tr}\left\{\hat{M}_{k_1}\hat{M}_{k_2}\hat{\rho}(u)\right\} = \sum_{k=0}^{K-1}(\hat{M}_{k_1}\hat{M}_{k_2})_{kk}\,\hat{\rho}_{kk}(u) = \sum_{k=0}^{K-1}(U)_{k_1k}\,(U)_{k_2k}\,\hat{\rho}_{kk}(u) = \left(U\mathrm{diag}\left(\vec{\rho}(u)\right)U^T\right)_{k_1k_2}. \tag{B9}$$

Appendix C: Information capacity with quantum sampling noise

1. Definition of capacity for quantum systems with sampling noise

The function approximation universality (which will be formally stated in Appendix I), as a basic requirement of most neural network model can be made concrete by defining a metric to quantify how well a given quantum system (generalizable to *any* dynamical system) approximates general functions. Suppose an arbitrary probability distribution $p(u)$ for a random (scalar) variable u defined in $[-1, 1]$. This naturally defines a function space $L_p^2([-1, 1])$ containing all functions $f : [-1, 1] \to \mathbb{R}$ with $\int_{-1}^1 f^2(u)p(u)\mathrm{d}u < \infty$. The space is equipped with the inner product structure $\langle f_1, f_2\rangle_p = \int_{-1}^1 f_1(u)f_2(u)p(u)\mathrm{d}u$. A standard way to check the ability of fitting nonlinear functions by a physical system is the *information processing capacity* [17],

$$C[f_\ell] = 1 - \min_{\boldsymbol{W}_\ell\in\mathbb{R}^K} \frac{\int_{-1}^1 \left(\sum_{k=0}^{K-1} W_{\ell k}x_k(u) - f_\ell(u)\right)^2 p(u)\mathrm{d}u}{\int_{-1}^1 f_\ell(u)^2 p(u)\mathrm{d}u}, \tag{C1}$$

where functions $f_\ell(u)$ are orthogonal target functions $\langle f_\ell, f_{\ell'}\rangle_p = \int_{-1}^1 f_\ell(u)f_{\ell'}(u)p(u)\mathrm{d}u = 0$ for $\ell \neq \ell'$. The *total expressive capacity* is computing the limitation $C_T \equiv \sum_{\ell=0}^{\infty} C[f_\ell]$, capturing the ability of what type of function the linear combination of physical system readout features can produce. Dambre's argument claims that the total capacity must be upper bounded by the number of features $C_T \leq K$.

While Dambre's result is quite general [17], it neglects the limitations due to noise in readout features, a fact that is unavoidable when using quantum systems in the presence of finite computational and measurement resources. In this appendix section, we will focus on the impact of fundamental quantum readout noise on this upper bound under finite sampling S. Given u and S,

the quantum readout features $\bar{X}_k(u) = \frac{1}{S}\sum_{s=1}^{S} X_k^{(s)}(u)$ are stochastic variables (where $X_k^{(s)} \in \{-1, +1\}$ are binary random values). The expectation vector and covariance matrix of $\bar{\boldsymbol{X}}(u)$ can be expressed in terms of $\vec{\rho}(u)$, the diagonal entries of the density matrix (see Eq. (B8))

$$\mathbb{E}[\bar{\boldsymbol{X}}(u)] \equiv \boldsymbol{x}(u) = U\vec{\rho}(u), \tag{C2}$$

$$\mathbb{E}[\bar{\boldsymbol{X}}(u)\bar{\boldsymbol{X}}^T(u)] - \mathbb{E}[\bar{\boldsymbol{X}}(u)]\mathbb{E}[\bar{\boldsymbol{X}}(u)]^T \equiv \frac{1}{S}\boldsymbol{\Sigma}(u) = \frac{1}{S}U\left(\mathrm{diag}\left(\vec{\rho}(u)\right) - \vec{\rho}(u)\cdot\vec{\rho}(u)^T\right)U^T. \tag{C3}$$

The dependence of readout features $x_k(u)$ on the input u can always be written in the form of a Taylor expansion,

$$x_k(u) = \sum_{j=0}^{\infty}(\mathbf{T})_{kj}u^j \tag{C4}$$

where we define the *transfer matrix* $\mathbf{T}(\boldsymbol{\theta}) \equiv \mathbf{T} \in \mathbb{R}^{K\times\infty}$ that depends on the density matrix $\hat{\rho}(u)$, and in particular on parameters $\boldsymbol{\theta}$ characterizing the quantum system.

To determine the optimal capacity to compute an arbitrary normalized function $f(u) = \sum_{j=0}^{\infty}(\mathbf{Y})_j u^j$ using the noisy readout features $\bar{\boldsymbol{X}}(u)$ extracted from the quantum system, we need to find an optimal $\boldsymbol{W}$ such that

$$C[f] = 1 - \frac{\min_{\boldsymbol{W}}\int_{-1}^{1}\left(\sum_{k=0}^{K-1}W_k\bar{X}_k(u) - f(u)\right)^2 p(u)\mathrm{d}u}{\int_{-1}^{1}f(u)^2p(u)\mathrm{d}u} \tag{C5}$$

By expanding the numerator of the right-hand side for a given, finite number of shots S, we find

$$\begin{aligned}
&\int_{-1}^{1}f(u)^2p(u)\mathrm{d}u - \int_{-1}^{1}\left(\sum_{k=0}^{K-1}W_k\bar{X}_k(u) - f(u)\right)^2 p(u)\mathrm{d}u \\
&= -\sum_{k_1=0}^{K-1}\sum_{k_2=0}^{K-1}W_{k_1}W_{k_2}\int_{-1}^{1}\bar{X}_k(u)\bar{X}_{k_2}(u)p(u)\mathrm{d}u + 2\sum_{k=0}^{K-1}W_k\int_{-1}^{1}\bar{X}_k(u)f(u)p(u)\mathrm{d}u \\
&\approx -\frac{1}{N}\sum_{k_1=0}^{K-1}\sum_{k_2=0}^{K-1}W_{k_1}W_{k_2}\sum_{n=1}^{N}\bar{X}_{k_1}(u^{(n)})\bar{X}_{k_2}(u^{(n)}) + \frac{2}{N}\sum_{k=0}^{K-1}W_k\sum_{n=1}^{N}\bar{X}_k(u^{(n)})f(u^{(n)}).
\end{aligned} \tag{C6}$$

where we have approximated the integral over the input domain by a finite sum in the limit of a large number of inputs N. Next, note that if $n \neq n'$, then $X_{k_1}(u^{(n)})$ and $X_{k_2}(u^{(n')})$ are independent random variables (thought not necessarily identically distributed). The sums over N on the right hand side are therefore sums of bounded independent random variables. In the limit of large $N \gg 1$, the deviation between stochastic realizations of these sums and their expectation values is exponentially suppressed, as determined by the Hoeffding inequality. Then, with large probability, the sums over N may be replaced by their expectation values,

$$\begin{aligned}
&\int_{-1}^{1}f(u)^2p(u)\mathrm{d}u - \int_{-1}^{1}\left(\sum_{k=0}^{K-1}W_k\bar{X}_k(u) - f(u)\right)^2 p(u)\mathrm{d}u \\
&\approx -\frac{1}{N}\sum_{k_1=0}^{K-1}\sum_{k_2=0}^{K-1}W_{k_1}W_{k_2}\sum_{n=1}^{N}\mathbb{E}[\bar{X}_{k_1}(u^{(n)})\bar{X}_{k_2}(u^{(n)})] + \frac{2}{N}\sum_{k=0}^{K-1}W_k\sum_{n=1}^{N}\mathbb{E}[\bar{X}_k(u^{(n)})f(u^{(n)})] \\
&= -\frac{1}{N}\sum_{k_1=0}^{K-1}\sum_{k_2=0}^{K-1}W_{k_1}W_{k_2}\sum_{n=1}^{N}\left(x_{k_1}(u^{(n)})x_{k_2}(u^{(n)}) + \frac{1}{S}\boldsymbol{\Sigma}(u^{(n)})_{k_1k_2}\right) + \frac{2}{N}\sum_{k=0}^{K-1}W_k\sum_{n=1}^{N}x_k(u^{(n)})f(u^{(n)}) \\
&\approx -\sum_{k_1=0}^{K-1}\sum_{k_2=0}^{K-1}W_{k_1}W_{k_2}\int_{-1}^{1}\left(x_{k_1}(u)x_{k_2}(u) + \frac{1}{S}\boldsymbol{\Sigma}(u)_{k_1k_2}\right)p(u)\mathrm{d}u + 2\sum_{k=0}^{K-1}W_k\int_{-1}^{1}x_k(u)f(u)p(u)\mathrm{d}u.
\end{aligned} \tag{C7}$$

The first approximation above comes from the Hoeffding inequality, where terms that are dropped are proportional to $1/\sqrt{N}$. In going from the second to the third line, we have used Eq. (C3). The final expression is obtained by rewriting sums over u as integrals, with an error proportional to $1/\sqrt{N}$ once more. Thus we can say the original integral in Eq. (C5) is approximately equal to Eq. (C7) to $O(1/\sqrt{N})$.

The first term in Eq. (C7) does not depend explicitly on the function $f(u)$ being constructed, and introduces quantities that are determined entirely by the response of the quantum system of interest to inputs over the entire domain of u. In particular, we introduce the *Gram matrix* $\mathbf{G} \in \mathbb{R}^{K\times K}$ as

$$(\mathbf{G})_{k_1k_2} = \int_{-1}^{1} x_{k_1}(u)x_{k_2}(u)p(u)\mathrm{d}u = \sum_{j_1=0}^{\infty}\sum_{j_2=0}^{\infty}(\mathbf{T})_{k_1j_1}\left(\int_{-1}^{1}u^{j_1+j_2}p(u)\mathrm{d}u\right)(\mathbf{T})_{k_2j_2} \equiv (\mathbf{T}\mathbf{\Lambda}\mathbf{T}^T)_{k_1k_2} \tag{C8}$$

where in the second line we have also introduced the *generalized Hilbert matrix* $\mathbf{\Lambda} \in \mathbb{R}^{\infty\times\infty}$ as

$$(\mathbf{\Lambda})_{j_1j_2} = \int_{-1}^{1} u^{j_1+j_2}p(u)\mathrm{d}u. \tag{C9}$$

Secondly, we introduce the noise matrix $\mathbf{V} \in \mathbb{R}^{K\times K}$,

$$(\mathbf{V})_{k_1k_2} = \int_{-1}^{1}\mathbf{\Sigma}(u)_{k_1k_2}\,p(u)\mathrm{d}u = \int_{-1}^{1}(x_k(u) - x_{k_1}(u)x_{k_2}(u))p(u)\mathrm{d}u \equiv (\mathbf{D})_{k_1k_2} - (\mathbf{G})_{k_1k_2} \tag{C10}$$

for index k satisfying $\hat{M}_k = \hat{M}_{k_1}\hat{M}_{k_2}$. Here we have also introduced the *second-order-moment* matrix $\mathbf{D} \in \mathbb{R}^{K\times K}$ such that $(\mathbf{D})_{k_1k_2} = \int_{-1}^{1}x_k(u)p(u)\mathrm{d}u$. Then, the noise matrix simply defines the covariance of readout features, and is therefore given by $\mathbf{V} = \mathbf{D} - \mathbf{G}$.

The second term in Eq. (C7) depends on $f(u)$ and can be simplified using the $\mathbf{\Lambda}$ matrix as well,

$$\int_{-1}^{1}x_k(u)f(u)p(u)\mathrm{d}u = \sum_{j_1=0}^{\infty}\sum_{j_2=0}^{\infty}(\mathbf{T})_{kj_1}\left(\int_{-1}^{1}u^{j_1+j_2}p(u)\mathrm{d}u\right)(\mathbf{Y})_{j_2} = (\mathbf{T}\mathbf{\Lambda}\mathbf{Y})_k. \tag{C11}$$

With these definitions, Eq. (C5) can be compactly written in matrix form as a Tikhonov regularization problem:

$$C[f] = \max_{\boldsymbol{W}}\left(\frac{-\boldsymbol{W}^T\left(\mathbf{T}\mathbf{\Lambda}\mathbf{T}^T + \frac{1}{S}\mathbf{V}\right)\boldsymbol{W} + 2\boldsymbol{W}^T\mathbf{T}\mathbf{\Lambda}\mathbf{Y}}{\mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y}}\right) = 1 - \min_{\boldsymbol{W}}\left(\frac{\left\|\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{T}^T\boldsymbol{W} - \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Y}\right\|^2 + \frac{1}{S}\boldsymbol{W}^T\mathbf{V}\boldsymbol{W}}{\mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y}}\right). \tag{C12}$$

The least-squares form ensures that the optimal value (argmin) $\widetilde{\boldsymbol{w}}$ of $\boldsymbol{W}$ has closed form

$$\widetilde{\boldsymbol{w}} = \left(\mathbf{T}\mathbf{\Lambda}\mathbf{T}^T + \frac{1}{S}\mathbf{V}\right)^{-1}\mathbf{T}\mathbf{\Lambda}\mathbf{Y}. \tag{C13}$$

Substituting $\boldsymbol{w}$ into the expression for C, we obtain the optimal capacity with which a function f can be constructed, which takes the form of a *generalized Rayleigh quotient*

$$C[f] = \frac{\mathbf{Y}^T\mathbf{\Lambda}\mathbf{T}^T\left(\mathbf{G} + \frac{1}{S}\mathbf{V}\right)^{-1}\mathbf{T}\mathbf{\Lambda}\mathbf{Y}}{\mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y}}. \tag{C14}$$

2. Eigentasks

Eq. (C14) defines the optimal capacity of approximating an arbitrary function $f(u) = \sum_{j=0}^{\infty}(\mathbf{Y})_j u^j$. We can therefore naturally ask which functions f maximise this optimal capacity. To this end, we first note that the denominator of Eq. (C14) is simply a normalization factor that can be absorbed into the definition of the function $f(u)$ being approximated, without loss of generality. More precisely, we consider:

$$\langle f, f\rangle_p = 1 = \left(\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Y}\right)^T\left(\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Y}\right) = \mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y}. \tag{C15}$$

Then, we can rewrite the optimal capacity from Eq. (C17) as

$$C[f] = \mathbf{Y}^T\mathbf{\Lambda}^{\frac{1}{2}}\left(\mathbf{Q}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Y}\right) \tag{C16}$$

Here we have defined the matrix $\mathbf{Q} \in \mathbb{R}^{\infty\times\infty}$ as

$$\mathbf{Q} = \mathbf{B}\left(\mathbf{I} + \frac{1}{S}\mathbf{R}\right)^{-1}\mathbf{B}^T, \tag{C17}$$

by introducing the matrix square root of $\mathbf{G} = \mathbf{G}^{\frac{1}{2}}\mathbf{G}^{\frac{1}{2}}$, where $\mathbf{G}^{\frac{1}{2}} \in \mathbb{R}^{K\times K}$. Then, $\mathbf{R} = \mathbf{G}^{-\frac{1}{2}}\mathbf{V}\mathbf{G}^{-\frac{1}{2}}$ becomes the *noise-to-signal* matrix, while the matrix $\mathbf{B}$ is given by

$$\mathbf{B} = \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{T}^T\mathbf{G}^{-\frac{1}{2}}, \tag{C18}$$

The decomposition in Eq. (C17) may be verified by direct substitution into Eq. (C16). The ability to calculate matrix powers and in particular the inverse of $\mathbf{G}$ requires constraints on its rank, which we show are satisfied in Appendix C 6.

We now consider the measure-independent part of the eigenvectors of $\mathbf{Q}$, indexed $\mathbf{Y}^{(k)}$, satisfying the standard eigenvalue problem:

$$\mathbf{Q}\left(\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Y}^{(k)}\right) = C_k\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Y}^{(k)}. \tag{C19}$$

where $k = 0, \cdots, K-1$. From Eq. (C16), it is clear that these eigenvectors have a particular meaning. Consider the function $y^{(k)}(u)$ defined by the eigenvector $\mathbf{Y}^{(k)}$, namely

$$y^{(k)}(u) = \sum_{j=0}^{\infty}\mathbf{Y}_j^{(k)}u^j, \tag{C20}$$

which we will refer to from now on as *eigentasks*. Suppose we wish to construct the function $y^{(k)}(u)$ using outputs obtained from the physical system defined by $\mathbf{Q}$ in the $S \to \infty$ limit (namely, with *deterministic* outputs). At a first glance, before we dive into solving the eigenproblem Eq.(C19), we do not know any relationship between $y^{(k)}$ and $\boldsymbol{x}(u)$.The rest part of this subsection is aiming to prove that $y^{(k)}$ must be a specific linear combination of features $\boldsymbol{x}(u)$. Then, the physical system's capacity for this construction is simply given by the corresponding eigenvalue C_k, as may be seen by substituting Eq. (C19) into Eq. (C16). Formally, the $y^{(k)}(u)$ serves as the *critical point* (or *stationary point*) of the generalized Rayleigh quotient in Eq. (C14). Consequently, the function that is constructed with largest capacity then corresponds to the nontrivial eigenvector with largest eigenvalue.

To obtain these eigentasks, we must solve the eigenproblem defined by Eq. (C19). Here, the representation of $\mathbf{Q}$ in Eq. (C17) becomes useful, as we will see that the eigensystem of $\mathbf{Q}$ is related closely to that of the noise-to-signal matrix $\mathbf{R}$. In particular, we first define the eigenproblem of $\mathbf{R}$,

$$\mathbf{R}\left(\mathbf{G}^{\frac{1}{2}}\boldsymbol{r}^{(k)}\right) = \beta_k^2\mathbf{G}^{\frac{1}{2}}\boldsymbol{r}^{(k)} \tag{C21}$$

with NSR eigenvalues β_k^2 and corresponding eigenvectors $\boldsymbol{r}^{(k)}$, which satisfy the orthogonality relation $\boldsymbol{r}^{(k')T}\mathbf{G}\boldsymbol{r}^{(k)} = \delta_{k,k'}$. Here the $\boldsymbol{r}^{(k)}$ is equivalent to be defined as the solution to generalized eigen-problem:

$$\mathbf{V}\boldsymbol{r}^{(k)} = \beta_k^2\mathbf{G}\boldsymbol{r}^{(k)}. \tag{C22}$$

This is because $\mathbf{V}\boldsymbol{r}^{(k)} = \mathbf{G}^{\frac{1}{2}}\mathbf{R}\mathbf{G}^{\frac{1}{2}}\boldsymbol{r}^{(k)} = \beta_k^2\mathbf{G}^{\frac{1}{2}}\mathbf{G}^{\frac{1}{2}}\boldsymbol{r}^{(k)} = \beta_k^2\mathbf{G}\boldsymbol{r}^{(k)}$. The prefactor $\mathbf{G}^{\frac{1}{2}}$ is introduced for later convenience. Eq. (C21) then allows us to define the related eigenproblem

$$\left(\mathbf{I} + \frac{1}{S}\mathbf{R}\right)^{-1}\mathbf{G}^{\frac{1}{2}}\boldsymbol{r}^{(k)} = \left(1 + \frac{\beta_k^2}{S}\right)^{-1}\mathbf{G}^{\frac{1}{2}}\boldsymbol{r}^{(k)} \tag{C23}$$

Next, we note that $\mathbf{Q}$ is related to the matrix in brackets above via a *generalized* similarity transformation defined by $\mathbf{B}$, Eq. (C17). In particular, $\mathbf{B}^T\mathbf{B} = \mathbf{G}^{-\frac{1}{2}}\mathbf{G}\mathbf{G}^{-\frac{1}{2}} = \mathbf{I} \in \mathbb{R}^{K\times K}$, while we remark that $\mathbf{B}\mathbf{B}^T \neq \mathbf{I}$ since it is in $\mathbb{R}^{\infty\times\infty}$. This connection allow us to show that

$$\mathbf{Q}\left(\mathbf{B}\mathbf{G}^{\frac{1}{2}}\boldsymbol{r}^{(k)}\right) = \mathbf{B}\left(\mathbf{I} + \frac{1}{S}\mathbf{R}\right)^{-1}\mathbf{B}^T\mathbf{B}\mathbf{G}^{\frac{1}{2}}\boldsymbol{r}^{(k)} = \frac{1}{1+\beta_k^2/S}\mathbf{B}\mathbf{G}^{\frac{1}{2}}\boldsymbol{r}^{(k)}. \tag{C24}$$

Comparing with Eq. (C19), we can now simply read off both the eigenvalues and eigenvectors of $\mathbf{Q}$,

$$\left.\begin{aligned} C_k &= \tfrac{1}{1+\beta_k^2/S} \\ \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Y}^{(k)} &= \mathbf{B}\mathbf{G}^{\frac{1}{2}}\boldsymbol{r}^{(k)} \end{aligned}\right\} \implies \mathbf{Y}^{(k)} = \mathbf{T}^T\boldsymbol{r}^{(k)} \tag{C25}$$

where we have used the definition of $\mathbf{B}$ from Eq. (C18). The functions defined by the eigenvectors $\mathbf{Y}^{(k)}$ are automatically orthonormalized:

$$\left\langle y^{(k_1)}, y^{(k_2)} \right\rangle_p = \left(\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Y}^{(k_1)}\right)^T\left(\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Y}^{(k_2)}\right) = \boldsymbol{r}^{(k_1)T}\mathbf{G}^{\frac{1}{2}}\mathbf{B}^T\mathbf{B}\mathbf{G}^{\frac{1}{2}}\boldsymbol{r}^{(k_2)} = \boldsymbol{r}^{(k_1)T}\mathbf{G}\boldsymbol{r}^{(k_2)} = \delta_{k_1k_2}. \tag{C26}$$

3. Noisy eigentasks from readout features

We can now also discuss the interpretation of $\{\beta_k^2\}$ for a physical system - in this case a quantum circuit - for which $\{\boldsymbol{r}^{(k)}\}$ are known. Consider a single run of the quantum system under finite shots S, which yields a single instance of the readout features $\bar{\boldsymbol{X}}(u)$. We can simply read off that an *noisy* version of the kth eigentask, $\bar{y}^{(k)}(u)$ can be constructed as

$$\bar{y}^{(k)}(u) = \sum_{k'=0}^{K-1} r_{k'}^{(k)}\bar{X}_{k'}(u) \tag{C27}$$

which is equivalent to requiring the output weights $\boldsymbol{W} = \boldsymbol{r}^{(k)}$.The corresponding set of noisy function is also orthogonal, this is because $\mathbf{V}\boldsymbol{r}^{(k)} = \beta_k^2\mathbf{G}\boldsymbol{r}^{(k)}$ implies $\boldsymbol{r}^{(k)T}\mathbf{V}\boldsymbol{r}^{(k')} = \beta_k^2\delta_{k,k'}$ and hence

$$\left\langle \bar{y}^{(k_1)}, \bar{y}^{(k_2)} \right\rangle_p = \boldsymbol{r}^{(k_1)T}\left(\mathbf{G} + \frac{1}{S}\mathbf{V}\right)\boldsymbol{r}^{(k_2)} = \left(1 + \frac{\beta_k^2}{S}\right)\delta_{k_1k_2} \tag{C28}$$

This equation can be further decomposed into two parts. Let the linear transformation of noise $\boldsymbol{\xi}(u)$ by defining $\xi^{(k)}(u) = \sum_{k=0}^{K-1} r_{k'}^{(k)}\zeta_{k'}(u)$

$$\mathbb{E}_u[y^{(k_1)}y^{(k_2)}] = \left\langle y^{(k_1)}, y^{(k_2)} \right\rangle_p = \boldsymbol{r}^{(k_1)T}\mathbf{G}\boldsymbol{r}^{(k_2)} = \delta_{k_1k_2}, \tag{C29}$$

$$\mathbb{E}_u[\xi^{(k_1)}\xi^{(k_2)}] = \left\langle \xi^{(k_1)}, \xi^{(k_2)} \right\rangle_p = \frac{1}{S}\boldsymbol{r}^{(k_1)T}\mathbf{V}\boldsymbol{r}^{(k_2)} = \frac{\beta_{k_1}^2}{S}\delta_{k_1k_2}. \tag{C30}$$

It means that the combination $\{\boldsymbol{r}^{(k)} \in \mathbb{R}^K\}_{k\in[K]}$ not only produces orthogonal eigentasks $\{y^{(k)}(u)\}$ for signal, but also induces a set of orthogonal noise functions $\{\xi^{(k)}(u)\}$.

If the quantum circuit can be run multiple times for a given S, multiple instances of $\bar{\boldsymbol{X}}(u)$ can be obtained, from each of which an estimate of the kth eigentask $\bar{y}^{(k)}(u)$ can be constructed. The expectation value of these estimates then simply yields

$$\mathbb{E}[\bar{y}^{(k)}(u)] = \sum_{k'=0}^{K-1} r_{k'}^{(k)}\mathbb{E}[\bar{X}_{k'}(u)] = \sum_{k'=0}^{K-1} r_{k'}^{(k)}x_{k'}(u) = y^{(k)}(u) \tag{C31}$$

If we have access to only a single instance of $\bar{\boldsymbol{X}}(u)$, however, and thus only one estimate $\bar{y}^{(k)}(u)$ (as $y^{(k)}(u)$ and $\bar{y}^{(k)}(u)$ depicted in Fig. 7), it is useful to know the expected error in this estimate. This error can be extracted from Eq. (C12). In particular, requiring $\mathbf{Y}^{(k)} = \mathbf{T}^T\boldsymbol{r}^{(k)}$, we have

$$\frac{\left\|\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{T}^T\boldsymbol{r}^{(k)} - \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Y}^{(k)}\right\|^2 + \frac{1}{S}\boldsymbol{r}^{(k)T}\mathbf{V}\boldsymbol{r}^{(k)}}{\mathbf{Y}^{(k)T}\mathbf{\Lambda}\mathbf{Y}^{(k)}} = \frac{1}{S}\boldsymbol{r}^{(k)T}\mathbf{V}\boldsymbol{r}^{(k)} = \frac{\beta_k^2}{S}. \tag{C32}$$

This mean squared error in using $\bar{y}^{(k)}(u)$ to estimate $y^{(k)}(u)$ over the domain of u decreases to zero for $S \to \infty$ as expected, since the noise in $\bar{\boldsymbol{X}}$ decreases with S. However, β_k^2 defines the S-independent contribution to the error. In particular, this indicates that at a given S, certain functions with lowers NSR eigenvalues β_k^2 may be better approximated using this physical system than others. We present in Fig. 7 the measured features $\bar{\boldsymbol{X}}$, the eigentasks $\boldsymbol{y}$ and their S-finite version $\bar{\boldsymbol{y}}$ in a 6-qubit Hamiltonian based system. The associated eigen-NSR spectrum, expressive capacity, and total correlations are also depicted for both ES $J \neq 0$ and PS $J = 0$.

4. Expressive capacity

Given an arbitrary set of complete orthonormal basis functions $f_\ell(u) = \sum_{j=0}^{\infty} (\mathbf{Y}_\ell)_j u^j$,

$$\langle f_\ell, f_{\ell'} \rangle_p = \left(\mathbf{\Lambda}^{\frac{1}{2}} \mathbf{Y}_\ell\right)^T \left(\mathbf{\Lambda}^{\frac{1}{2}} \mathbf{Y}_{\ell'}\right) = \delta_{\ell\ell'}. \tag{C33}$$

The total capacity is independent of the basis choice

$$\begin{aligned} C_T(S) &= \sum_{\ell=0}^{\infty} C[f_\ell] = \sum_{\ell=0}^{\infty} \mathbf{Y}_\ell^T \mathbf{\Lambda}^{\frac{1}{2}} \left(\mathbf{\Lambda}^{\frac{1}{2}} \mathbf{T}^T \left(\mathbf{T}\mathbf{\Lambda}\mathbf{T}^T + \frac{1}{S}\mathbf{V}\right)^{-1} \mathbf{T}\mathbf{\Lambda}^{\frac{1}{2}}\right) \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{Y}_\ell \\ &= \mathrm{Tr}\left(\mathbf{\Lambda}^{\frac{1}{2}} \mathbf{T}^T \left(\mathbf{T}\mathbf{\Lambda}\mathbf{T}^T + \frac{1}{S}\mathbf{V}\right)^{-1} \mathbf{T}\mathbf{\Lambda}^{\frac{1}{2}}\right) = \mathrm{Tr}\left(\left(\mathbf{G} + \frac{1}{S}\mathbf{V}\right)^{-1} \mathbf{G}\right) = \sum_{k=0}^{K-1} \frac{1}{1 + \frac{\beta_k^2}{S}}. \end{aligned} \tag{C34}$$

5. Estimation in case of nonlinear functions after linear output layer

Usually, instead of taking the linear transformation $\boldsymbol{W} \cdot \bar{\boldsymbol{X}}$, the training process can involve some complicated nonlinear activation functions or classical kernel, which may also be fed into a non-quadratic nonlinear loss function afterwards. These two processes can be unified to be $\sigma_{\mathrm{NL}}(\bar{\boldsymbol{X}}(u))$ with any smooth function σ_{NL}. In this subsection, we show how to translate our result obtaining from quadratic nonlinear function Eq. (C5) into a more general loss function with form of

$$\mathscr{L} = \mathbb{E}_u[\sigma_{\mathrm{NL}}(\bar{\boldsymbol{X}})] \tag{C35}$$

Now let us first transform all noisy measured features $\{\bar{X}_k\}$ into the naturally orthogonal basis of signal $\{y^{(k)}\}$ and noise $\{\xi^{(k)}\}$.

$$\bar{X}_{k'}(u) \equiv \sum_{k=0}^{K-1} \Gamma_{k'k}(y^{(k)}(u) + \xi^{(k)}(u)), \tag{C36}$$

such transformation of $\mathbf{\Gamma} \in \mathbb{R}^{K\times K}$ must uniquely exist, this is because all K of $\{\boldsymbol{r}^{(k)}\}$ are linearly independent. Recall Eq. (C30) claims that $\mathbb{E}_u[\xi^{(k)}] = 0$ and $\mathbb{E}_u[\xi^{(k)}\xi^{(k')}] = \beta_k^2 \delta_{kk'}/S$, we can deal with the nonlinearity by taking the quadratic expansion, where , we get

$$\begin{aligned} \mathscr{L} &= \mathbb{E}_u[\sigma_{\mathrm{NL}}(\bar{\boldsymbol{X}})] = \mathbb{E}_u[\sigma_{\mathrm{NL}}(\mathbf{\Gamma}\bar{\boldsymbol{y}})] = \mathbb{E}_u\left[\sigma_{\mathrm{NL}}\left(\sum_k \Gamma_{0,k}(y^{(k)} + \xi^{(k)}), \cdots, \sum_k \Gamma_{K-1,k}(y^{(k)} + \xi^{(k)})\right)\right] \\ &\approx \mathbb{E}_u[\sigma_{\mathrm{NL}}(\mathbf{\Gamma}\boldsymbol{y})] + \sum_{k=0}^{K-1} \mathbb{E}_u\left[\frac{\partial \sigma_{\mathrm{NL}}}{\partial y^{(k)}} \xi^{(k)}\right] + \frac{1}{2} \sum_{k_1=0}^{K-1} \sum_{k_2=0}^{K-1} \mathbb{E}_u\left[\frac{\partial^2 \sigma_{\mathrm{NL}}}{\partial y^{(k_1)} \partial y^{(k_2)}} \xi^{(k_1)} \xi^{(k_2)}\right] \\ &= \mathbb{E}_u[\sigma_{\mathrm{NL}}(\mathbf{\Gamma}\boldsymbol{y})] + \frac{1}{2} \sum_{k_1=0}^{K-1} \sum_{k_2=0}^{K-1} \mathbb{E}_u\left[\frac{\partial^2 \sigma_{\mathrm{NL}}}{\partial y^{(k_1)} \partial y^{(k_2)}} \xi^{(k_1)} \xi^{(k_2)}\right], \end{aligned} \tag{C37}$$

where the first order terms vanish due to Hoeffding inequality again. We then make a further approximation of Eq. (C37) by replacing the $\xi^{(k_1)}\xi^{(k_2)}$ with its u-average $\mathbb{E}_u[\xi^{(k_1)}\xi^{(k_2)}] = \delta_{k_1 k_2}\beta_{k_1}^2/S$:

$$\mathscr{L} \approx \mathbb{E}_u[\sigma_{\mathrm{NL}}(\mathbf{\Gamma}\boldsymbol{y})] + \sum_{k=0}^{K-1} \frac{\beta_k^2}{S} \cdot \mathbb{E}_u\left[\frac{\partial^2 \sigma_{\mathrm{NL}}}{(\partial y^{(k)})^2}\right]. \tag{C38}$$

In fact, any of the second terms can be further simplified by chain rule: $\mathscr{L} \approx \mathbb{E}_u[\sigma_{\mathrm{NL}}(\mathbf{\Gamma}\boldsymbol{y})] + \sum_k \frac{\beta_k^2}{S} \cdot \mathbb{E}_u[(\mathbf{\Gamma}^T \nabla_{\boldsymbol{x}}^2 \sigma_{\mathrm{NL}} \mathbf{\Gamma})_{kk}]$. The approximation in Eq. (C38) is rough, but it still gives us a sufficient reason to do the following manipulation: for optimized $\mathscr{L}$, the dependence on $y^{(k)}$ with $\beta_k^2/S > 1$ will be strongly suppressed in large-N limit, hence we can pre-exclude the eigentasks whose $\beta_k^2/S > 1$.

Let us use one typical example, the widely used logistic regression in classification, to illustrate our argument here. As what we will introduce in Appendix I, the target function is the conditional probability distribution $f(u) := \Pr[u \in C_1 | u]$ in such

classification model (see Eq. (I4)), and then there is one more layer of softmax and cross-entropy function acting on linear map $\mathscr{L} = \mathbb{E}_u[\mathrm{H}(f(u), \sigma(\boldsymbol{W}\cdot\bar{\boldsymbol{X}}(u)))]$ where σ is sigmoid function (*e.g.* softmax function $\sigma(z) = 1/(1+\exp(-z))$), and $\mathrm{H}(p,q) = -p\ln q - (1-p)\ln(1-q)$ is the cross-entropy. Especially, any linear combination of $\{\bar{X}_k\}$ can be translated into linear combination

$$\boldsymbol{W}\cdot\bar{\boldsymbol{X}}(u) \equiv \sum_{k=0}^{K-1} \Omega_k \cdot (y^{(k)}(u) + \xi^{(k)}(u)), \tag{C39}$$

Again, such vector $\boldsymbol{\Omega} = \boldsymbol{\Gamma}^T\boldsymbol{W}$ must also uniquely exist. For any $\sigma_{\mathrm{NL}} = g(\boldsymbol{W}\cdot\boldsymbol{x})$, one always have $\boldsymbol{\Gamma}^T\nabla^2_{\boldsymbol{x}}\sigma_{\mathrm{NL}}\boldsymbol{\Gamma} = g''(\boldsymbol{\Omega}\cdot\boldsymbol{y})\boldsymbol{\Omega}^T\boldsymbol{\Omega}$:

$$\mathscr{L} \approx \mathbb{E}_u[\mathrm{H}(f, \sigma(\boldsymbol{\Omega}\cdot\boldsymbol{y}))] + \left(\sum_{k=0}^{K-1}\frac{\beta_k^2}{S}\Omega_k^2\right)\cdot\mathbb{E}_u[\sigma(\boldsymbol{\Omega}\cdot\boldsymbol{y})(1-\sigma(\boldsymbol{\Omega}\cdot\boldsymbol{y}))]\,. \tag{C40}$$

It helps us read from the prefactor β_k^2/S induces a natural regularization on Ω_k in loss function, in addition to the S-infinity term $\lim_{S\to\infty}\mathscr{L} = \mathbb{E}_u[\mathrm{H}(f,\sigma(\boldsymbol{\Omega}\cdot\boldsymbol{y}))]$. We will leave the detailed discussion of this important application in Appendix H and Appendix I.

6. Proof that the Gram matrix G is full rank

Recall that before we analytically find the eigenvectors of $\mathbf{Q}$, we first show that the matrix $\mathbf{G}$ is invertible. It comes from that all K readout features $\{x_k(u)\}_{k\in[K]}$ being linear independent is entirely equivalent to the full-rankness of the corresponding Gram matrix $\mathrm{Rank}(\mathbf{G}) = K$. Thanks to the linearity of readout, we can show such linear independence by contradiction. Suppose on the contrary there exists coefficients $\{c_k\}_{k\in[K]}$ such that

$$\sum_{k=0}^{K-1} c_k x_k(u) = \mathrm{Tr}\left\{\left(\sum_{k=0}^{K-1} c_k\hat{M}_k\right)\mathcal{U}(u)\hat{\rho}_0\right\} = 0. \tag{C41}$$

However, this means that the quantum observable $\sum_{k=0}^{K-1} c_k\hat{M}_k$ is a zero-expectation readout-qubit quantity for any state $\mathcal{U}(u)\hat{\rho}_0$ under arbitrary input u, which is impossible. This shows the linear independence. Furthermore, we then argue that it ensures $\mathbf{G}$ has no non-trivial null space. This is because that any $\{c_k\}_{k\in[K]}$ will satisfy

$$\sum_{k_1,k_2=1}^{K} c_{k_1}c_{k_2}(\mathbf{G})_{k_1,k_2} = \int_{-1}^{1}\left(\sum_{k_1=1}^{K} c_{k_1}x_{k_1}(u)\right)\left(\sum_{k_2=1}^{K} c_{k_2}x_{k_2}(u)\right)p(u)\mathrm{d}u = \left\langle\sum_{k=0}^{K-1} c_k x_k, \sum_{k=0}^{K-1} c_k x_k\right\rangle_p. \tag{C42}$$

where the RHS is the norm of function $\sum_{k=0}^{K-1} c_k x_k(u)$. The summation $\sum_{k_1,k_2=1}^{K} c_{k_1}c_{k_2}(\mathbf{G})_{k_1,k_2} = 0$ vanishes if and only if function $\sum_{k=0}^{K-1} c_k x_k(u)$ is a zero function. That is why the linear independence of features $\{c_k\}_{k\in[K]}$ is equivalent to that symmetric matrix $\mathbf{G}$ has no zero eigenvalues, namely $\mathrm{Rank}(\mathbf{G}) = K$. Numerically speaking, this relation always holds in general as long as assuming this is for the case where $N \gg K$.

7. Simplifying the noise-to-signal matrix and its eigenproblem

We have shown that the problem of obtaining the eigentasks for a generic quantum system, and deducing its expressive capacity under finite measurement resources, can be reduced simply to solving the eigenproblem of its noise-to-signal matrix $\mathbf{R}$, Eq. (C21). Note that constructing $\mathbf{R} = \mathbf{G}^{-\frac{1}{2}}\mathbf{V}\mathbf{G}^{-\frac{1}{2}}$ requires computing the inverse of $\mathbf{G}$. However, $\mathbf{G}$ can have small (although always nonzero) eigenvalues, especially for larger systems, rendering it ill-conditioned and making the computation of $\mathbf{R}$ numerically unstable. Fortunately, certain simplifications can be made to derive an equivalent eigenproblem that is much easier to solve.

To begin, we first note that so far, we have placed no requirements on the specific form of measurement operators $\{\hat{M}_k\}$, and thus the readout features $x_k(u) = \mathrm{Tr}\{\hat{M}_k\hat{\rho}(u)\}$ are also unspecified. Our analysis thus far holds for any set of measurement operators that describe a complete set of commuting observables. However, specific choices of measurement operators can

simplify the form of the matrices $\mathbf{G}$ and $\mathbf{V}$ involved. In particular, if one chooses $\hat{M}_k$ to be the projections onto the computational basis, $\hat{M}_k = |\boldsymbol{b}_k\rangle\langle\boldsymbol{b}_k|$, then according to Eq. (B8), by setting $U = I$ we have $\boldsymbol{x}(u) \equiv \vec{\rho}(u)$, which we refer to as the *probability representation* of readout features. Practically, the probability representation is native to measurement schemes in contemporary quantum processors, and therefore minimizes the required post-processing of readout features obtained from a real device. More importantly, although it is related to any other readout feature representation via a unitary transformation, the strength of the probability representation lies in the fact that it renders the second-order moment matrix $\mathbf{D}$ diagonal. In particular,

$$(\mathbf{D})_{k_1k_2} = \begin{cases} \sum_{k=0}^{K-1}(\mathbf{G})_{kk_1}, & \text{if } k_1 = k_2 \\ 0, & \text{if } k_1 \neq k_2 \end{cases} \quad \text{(in probability representation of readout features)} \tag{C43}$$

Using $\mathbf{V} = \mathbf{D} - \mathbf{G}$, we can rewrite the eigenproblem for $\mathbf{R}$,

$$\begin{aligned} \mathbf{R}\left(\mathbf{G}^{\frac{1}{2}}\boldsymbol{r}^{(k)}\right) &= \beta_k^2\mathbf{G}^{\frac{1}{2}}\boldsymbol{r}^{(k)} \\ \implies \mathbf{G}^{-\frac{1}{2}}(\mathbf{D}-\mathbf{G})\mathbf{G}^{-\frac{1}{2}}\left(\mathbf{G}^{\frac{1}{2}}\boldsymbol{r}^{(k)}\right) &= \beta_k^2\mathbf{G}^{\frac{1}{2}}\boldsymbol{r}^{(k)} \\ \implies \mathbf{G}^{-1}\mathbf{D}\boldsymbol{r}^{(k)} &= (1+\beta_k^2)\boldsymbol{r}^{(k)} \end{aligned} \tag{C44}$$

Finally, considering the inverse of the matrix on the left hand side, we obtain the simplified eigenproblem for the matrix $\mathbf{D}^{-1}\mathbf{G}$,

$$\mathbf{D}^{-1}\mathbf{G}\boldsymbol{r}^{(k)} = (1+\beta_k^2)^{-1}\boldsymbol{r}^{(k)} \equiv \alpha_k\boldsymbol{r}^{(k)}, \tag{C45}$$

which shares eigenvectors with $\mathbf{R}$, and whose eigenvalues are a simple transformation of the NSR eigenvalues β_k^2. Importantly, constructing $\mathbf{D}^{-1}\mathbf{G}$ no longer requires calculating any powers of $\mathbf{G}$, and when further choosing readout features in the probability representation, it relies only on the inversion of a simple diagonal matrix $\mathbf{D}$.

The matrix $\mathbf{D}^{-1}\mathbf{G}$ has significance in spectral graph theory, when interpreting the Gram matrix $\mathbf{G}$ as the adjacency matrix of a weighted graph. This connection is elaborated upon in Appendix C 8.

8. Connections to spectral graph theory

Let us have a small digression to the graphic theoretic meaning of $\mathbf{G}$ and $\mathbf{D}^{-1}\mathbf{G}$. Now we consider a weighted graph with adjacency matrix $\mathbf{G}$. In spectral graph theory, the matrix $\mathbf{D}^{-1}\mathbf{G}$ is exactly the random walk matrix associated with graph $\mathbf{G}$, and then the second order matrix $\mathbf{D}$ happens to be the *degree matrix* of this graph since $(\mathbf{D})_{kk} = \sum_{k'=0}^{K-1}(\mathbf{G})_{kk'}$. Then the eigentask combination coefficient $\boldsymbol{r}^{(k)}$ is precisely the right eigenvector of random walk matrix. Another concept associated with a graph is $\mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{G}\mathbf{D}^{-\frac{1}{2}}$, the *normalized Laplacian matrix* of $\mathbf{G}$, while the matrix $\mathbf{D}^{-\frac{1}{2}}\mathbf{G}\mathbf{D}^{-\frac{1}{2}}$ is always referred to be *normalized adjacency matrix* in many literatures. The eigenproblem of normalized adjacency matrix can also be solved easily, because

$$\mathbf{D}^{-\frac{1}{2}}\mathbf{G}\mathbf{D}^{-\frac{1}{2}}\left(\mathbf{D}^{\frac{1}{2}}\boldsymbol{r}^{(k)}\right) = \mathbf{D}^{\frac{1}{2}}\mathbf{D}^{-1}\mathbf{G}\boldsymbol{r}^{(k)} = \alpha_k\left(\mathbf{D}^{\frac{1}{2}}\boldsymbol{r}^{(k)}\right). \tag{C46}$$

From perspective of spectral graph theory, roughly speaking, a reservoir with stronger ability to resist noise are those who has more "bottlenecks" in graph $\mathbf{G}$'s connectivity. The extreme case is supposing that $\alpha_k = 1$ (or $1 - \alpha_k = 0$) for all k. According the basic conclusion in spectral graph theory, the normalized Laplacian matrix has K zero eigenvalues iff the graph $\mathbf{G}$ is fully disconnected. This gives us the condition when noisy information capacity obtain its upper bound K: there exists a partition $\{\mathrm{Dom}_k\}_{k\in[K]}$ of domain $\mathrm{Dom} = [-1, 1]$ such that $\hat{\rho}_{kk}(u) = 1$ iff $u \in \mathrm{Dom}_k$.

Appendix D: Spectral analysis based on finite statistics

While Eq. (C45) is a numerically simpler eigenproblem to solve than Eq. (C21), it still requires the approximation of $\mathbf{G}$ (recall that $\mathbf{D}$ can be obtained from $\mathbf{G}$) from readout features $\bar{\boldsymbol{X}}(u)$ under finite sampling, due to the finiteness of shots S, the number of input points N, and also the number of realizations of readout features for a given S. In what follows, we show how an approximation $\widetilde{\mathbf{G}}_N$ of $\mathbf{G}$ can be constructed from finitely-sampled readout features, as relevant for practical quantum devices. Secondly, we also describe an approach to obtain the eigentasks $y^{(k)}(u)$ and corresponding NSR eigenvalues β_k^2 that avoids explicit construction of the Gram matrix, and is thus even more numerically robust.

1. Approximating eigentasks and NSR eigenvalues under finite S and N

For practical computations, readout features $\bar{\boldsymbol{X}}(u)$ from the quantum system for finite S can be computed for a discrete set of $u^{(n)} \in [-1, 1]$ for $n = 1, \ldots, N$. Labelling the corresponding readout features $\bar{\boldsymbol{X}}(u^{(n)})$, we can define the *regression matrix* constructed from these readout features,

$$\widetilde{\mathbf{F}}_N \equiv (\bar{\boldsymbol{X}}(u^{(1)}), \bar{\boldsymbol{X}}(u^{(2)}), \cdots, \bar{\boldsymbol{X}}(u^{(N)}))^T = \begin{pmatrix} \bar{X}_0(u^{(1)}) & \cdots & \bar{X}_{K-1}(u^{(1)}) \\ \vdots & & \vdots \\ \bar{X}_0(u^{(N)}) & \cdots & \bar{X}_{K-1}(u^{(N)}) \end{pmatrix}. \tag{D1}$$

Here, $\widetilde{\mathbf{F}}_N \in \mathbb{R}^{N\times K}$, with subscript N indicating its construction from a finite set of N inputs, is a random matrix due to the stochasticity of readout features; in particular it can be written as:

$$\widetilde{\mathbf{F}}_N = \mathbf{F}_N + \frac{1}{\sqrt{S}}\mathbf{Z}(\mathbf{F}_N) \tag{D2}$$

where $(\mathbf{F}_N)_{nk} = \mathbb{E}[\bar{X}_k(u^{(n)})] = x_k(u^{(n)})$, and $\mathbf{Z}$ is the centered multinomial stochastic process, so that $\mathbb{E}[\widetilde{\mathbf{F}}_N] = \mathbf{F}_N$.

Using this regression matrix $\widetilde{\mathbf{F}}_N$, we can obtain an estimation of the Gram matrix and second order moment matrix, which we denote $\widetilde{\mathbf{G}}_N$ and $\widetilde{\mathbf{D}}_N$, and whose matrix elements are defined via

$$(\widetilde{\mathbf{G}}_N)_{k_1k_2} \equiv \frac{1}{N}\sum_{n=1}^{N} \bar{X}_{k_1}(u^{(n)})\bar{X}_{k_2}(u^{(n)}) = \frac{1}{N}(\widetilde{\mathbf{F}}_N^T\widetilde{\mathbf{F}}_N)_{k_1k_2} \approx \int_{-1}^{1} \bar{X}_{k_1}(u)\bar{X}_{k_2}(u)p(u)\mathrm{d}u, \tag{D3}$$

$$(\widetilde{\mathbf{D}}_N)_{k_1k_2} \equiv \delta_{k_1,k_2}\frac{1}{N}\sum_{n=1}^{N} \bar{X}_{k_1}(u^{(n)}) \approx \delta_{k_1,k_2}\int_{-1}^{1} \bar{X}_{k_1}(u)p(u)\mathrm{d}u. \tag{D4}$$

While the quantities $\widetilde{\mathbf{G}}_N$ and $\widetilde{\mathbf{D}}_N$ are computed from stochastic readout features, their stochastic contributions are suppressed in the large N limit by the Hoeffding inequality for sums of bounded stochastic variables. In particular, we can define their deterministic limit for $N \to \infty$, according to Eq. (C7), as

$$\widetilde{\mathbf{G}} \equiv \lim_{N\to\infty}\frac{1}{N}(\widetilde{\mathbf{F}}_N^T\widetilde{\mathbf{F}}_N)_{k_1k_2} = \mathbf{G} + \frac{1}{S}\mathbf{V} = \mathbf{G} + \frac{1}{S}(\mathbf{D} - \mathbf{G}), \tag{D5}$$

$$\widetilde{\mathbf{D}} \equiv \lim_{N\to\infty}\widetilde{\mathbf{D}}_N = \mathbf{D}. \tag{D6}$$

Inverting the above expressions allow us to express the Gram matrix $\mathbf{G}$ and second-order moment matrix $\mathbf{D}$ in terms of the estimates $\widetilde{\mathbf{G}}$ and $\widetilde{\mathbf{D}}$ computed using a finite number of shots S,

$$\mathbf{G} = \frac{S}{S-1}\widetilde{\mathbf{G}} - \frac{1}{S-1}\widetilde{\mathbf{D}}, \tag{D7}$$

$$\mathbf{D} = \widetilde{\mathbf{D}}. \tag{D8}$$

We see that to lowest order in $\frac{1}{S}$, $\mathbf{G} \approx \widetilde{\mathbf{G}}$ and $\mathbf{D} \approx \widetilde{\mathbf{D}}$, which is what one might expect naively. However, we clearly see that the estimation of $\mathbf{G}$ can be improved by including a higher-order correction in $\frac{1}{S}$. This contribution arises due to the highly-correlated nature of noise and signal for quantum systems: we are able to estimate the noise matrix $\widetilde{\mathbf{G}}$ and $\widetilde{\mathbf{D}}$ using knowledge of the readout features, and correct for the contribution to $\widetilde{\mathbf{G}}$ and $\widetilde{\mathbf{D}}$ that arises from this noise matrix. We will see that this contribution will be important in more accurately approximating quantities of interest derived from $\mathbf{G}$, $\mathbf{D}$.

To this end, we recall that our ultimate aim is not just to estimate $\mathbf{G}$ and $\mathbf{D}$, but to solve the eigenproblem of Eq. (C45). Using the above relation, we can then establish $\widetilde{\mathbf{D}}^{-1}\widetilde{\mathbf{G}} = \frac{S-1}{S}\mathbf{D}^{-1}\mathbf{G} + \frac{1}{S}\mathbf{I}$, and write Eq. (C45) in a form entirely in terms of $\widetilde{\mathbf{G}}$ and $\widetilde{\mathbf{D}}$,

$$\begin{aligned} &\mathbf{D}^{-1}\mathbf{G}\boldsymbol{r}^{(k)} = (1+\beta_k^2)^{-1}\boldsymbol{r}^{(k)}, \\ \implies &\widetilde{\mathbf{D}}^{-1}\widetilde{\mathbf{G}}\boldsymbol{r}^{(k)} = \left[\frac{S-1}{S}(1+\beta_k^2)^{-1} + \frac{1}{S}\right]\boldsymbol{r}^{(k)}. \end{aligned} \tag{D9}$$

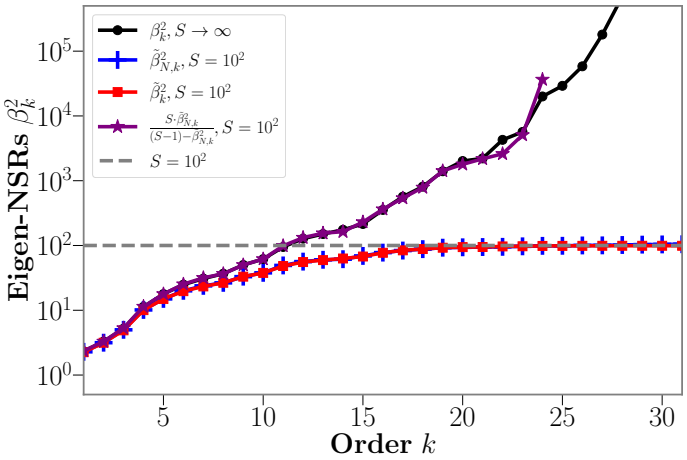

FIG. 5. Eigen-analysis in $L=5$ H-ansatz system by taking $S=10^2$ shots on each of $N=10^4$ samples, with true eigen-NSRs β_k^2 (black), S-finite sampled $\tilde\beta_{N,k}^2$ (blue) and corrected $(S\cdot\tilde\beta_{N,k}^2)/((S-1)-\tilde\beta_{N,k}^2)$ (purple). $\tilde\beta_k^2$, the large N limit of $\tilde\beta_{N,k}^2$ is also plotted in red for comparison. The data correction is necessary since all $\tilde\beta_{N,k}^2$ are below the $S=10^2$, and the corrected data show much better performance even if $\beta_k^2\gg S$. The estimated line (in purple) are cutoff at $k=25$ since all sampled $\tilde\beta_{N,k}^2$ after that are larger the $S-1$ so that they are not correctable.

Note that the final form is conveniently another eigenproblem, now for the finite-S matrix $\widetilde{\mathbf{D}}^{-1}\widetilde{\mathbf{G}}$:

$$\widetilde{\mathbf{D}}^{-1}\widetilde{\mathbf{G}}\tilde{\boldsymbol{r}}^{(k)}=(1+\tilde\beta_k^2)^{-1}\tilde{\boldsymbol{r}}^{(k)}\equiv\tilde\alpha_k\tilde{\boldsymbol{r}}^{(k)},\tag{D10}$$

whose eigenvalues and eigenvectors can be easily related to the desired eigenvalues β_k^2 and eigenvectors $\boldsymbol{r}^{(k)}$ of Eq. (C45). Following some algebra, we find:

$$\beta_k^2=\frac{S}{(S-1)-\tilde\beta_k^2}\cdot\tilde\beta_k^2=\tilde\beta_k^2+\sum_{j=1}^{\infty}\tilde\beta_k^2\left(1+\tilde\beta_k^2\right)^j\left(\frac{1}{S}\right)^j,\tag{D11}$$

$$\boldsymbol{r}^{(k)}=\tilde{\boldsymbol{r}}^{(k)}.\tag{D12}$$

From Eq. (D11), we see that to lowest order in $\frac{1}{S}$, $\beta_k^2\approx\tilde\beta_k^2$. However, this expression also supplies corrections to higher orders in $\frac{1}{S}$, which are non-negligible even for $\beta_k^2<S$, as we see in example of Fig. 5. In contrast, the estimated eigenvectors $\tilde{\boldsymbol{r}}^{(k)}$ to *any* order in $\frac{1}{S}$ equal the desired eigenvectors $\boldsymbol{r}^{(k)}$ without any corrections.

Of course, in practice we do not have access to the matrices $\widetilde{\mathbf{G}}$ and $\widetilde{\mathbf{D}}$, as these are only defined precisely in the limit where $N\to\infty$. However, for large enough N, we can approximate these matrices to lowest order by their finite N values, $\widetilde{\mathbf{G}}=\widetilde{\mathbf{G}}_N+\mathcal{O}\left(\frac{1}{N}\right)$ and $\widetilde{\mathbf{D}}=\widetilde{\mathbf{D}}_N+\mathcal{O}\left(\frac{1}{N}\right)$. Then, the eigenproblem in Eq. (D10) can be expressed in the final form,

$$\widetilde{\mathbf{D}}_N^{-1}\widetilde{\mathbf{G}}_N\tilde{\boldsymbol{r}}_N^{(k)}=(1+\tilde\beta_{N,k}^2)^{-1}\tilde{\boldsymbol{r}}_N^{(k)}\equiv\tilde\alpha_{N,k}\tilde{\boldsymbol{r}}_N^{(k)},\tag{D13}$$

where the eigenvalues $\tilde\beta_{N,k}^2,\tilde\alpha_{N,k}$ and eigenvectors $\tilde{\boldsymbol{r}}_N^{(k)}$ in the large N limit must satisfy

$$\lim_{N\to\infty}\tilde\beta_{N,k}^2=\tilde\beta_k^2,\quad\lim_{N\to\infty}\tilde\alpha_{N,k}=\tilde\alpha_k,\quad\lim_{N\to\infty}\tilde{\boldsymbol{r}}_N^{(k)}=\tilde{\boldsymbol{r}}^{(k)}\equiv\boldsymbol{r}^{(k)}.\tag{D14}$$

Here the invertibility of the empirically-computed matrix $\widetilde{\mathbf{D}}_N$ required for Eq. (D13) is numerically checked, based on which we can establish a better numerical method in Appendix D 2.

Eq. (D13) represents the eigenproblem whose eigenvalues $\tilde\beta_{N,k}^2$ and eigenvectors $\tilde{\boldsymbol{r}}_N^{(k)}$ we actually calculate. For large enough N and under finite S, we can use these as valid approximations to the eigenvalues and eigenvectors of Eq. (D10). This finally enables us to directly estimate the $N,S\to\infty$ quantities β_k^2 and $\boldsymbol{r}^{(k)}$ using Eqs. (D11), (D12):

$$\beta_k^2\approx\frac{S\cdot\tilde\beta_{N,k}^2}{(S-1)-\tilde\beta_{N,k}^2}=\frac{1-\tilde\alpha_{N,k}}{\tilde\alpha_{N,k}-\frac{1}{S}},\tag{D15}$$

$$\boldsymbol{r}^{(k)}\approx\tilde{\boldsymbol{r}}_N^{(k)}.\tag{D16}$$

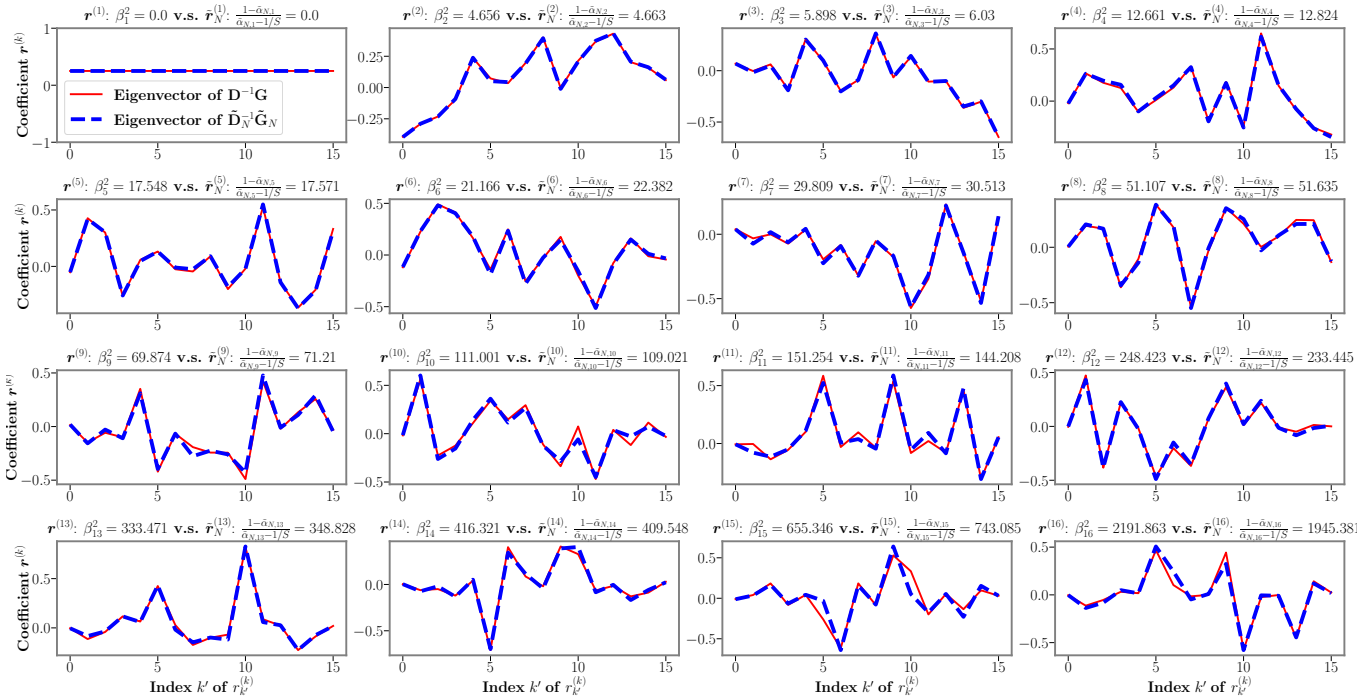

FIG. 6. Estimating NSR eigenvalues and corresponding eigentask coefficients under finite statistics ($N = 300, S = 1000$) in a 4-qubit H-encoding system, and comparison with theoretical value for $N \to \infty, S \to \infty$.

It is clear that the approximation of β_k^2 to lowest order will be an underestimate, as the contribution of order $\frac{1}{S}$ is positive. In Fig. 6, we plot the estimated eigenvectors $\tilde{\boldsymbol{r}}_N^{(k)}$ computed under finite statistics ($N = 300, S = 1000$, where these two numbers are relevant for IBM quantum processors) in H-encoding, together with the $N, S \to \infty$ eigenvectors $\boldsymbol{r}^{(k)}$, and the estimated eigenvalues.

2. Gram matrix-free construction to approximate eigentasks and NSR eigenvalues

If we consider Eq. (D13) and multiply through by $\mathbf{D}_N^{-\frac{1}{2}}$, the resulting equation can be written as an equivalent eigenproblem,

$$\frac{1}{N}\widetilde{\mathbf{D}}_N^{-\frac{1}{2}}\widetilde{\mathbf{F}}_N^T\widetilde{\mathbf{F}}_N\widetilde{\mathbf{D}}_N^{-\frac{1}{2}}\left(\widetilde{\mathbf{D}}_N^{\frac{1}{2}}\tilde{\boldsymbol{r}}_N^{(k)}\right) = \tilde{\alpha}_{N,k}\left(\widetilde{\mathbf{D}}_N^{-\frac{1}{2}}\tilde{\boldsymbol{r}}_N^{(k)}\right) \tag{D17}$$

where we have also written $\widetilde{\mathbf{G}}_N = \frac{1}{N}\widetilde{\mathbf{F}}_N^T\widetilde{\mathbf{F}}_N$ as in the previous section. Note that as written above, the eigenproblem is entirely equivalent to obtaining the singular value decomposition of the matrix $\frac{1}{\sqrt{N}}\widetilde{\mathbf{D}}_N^{-\frac{1}{2}}\widetilde{\mathbf{F}}_N^T$. This particular normalization factor $\frac{1}{\sqrt{N}}\widetilde{\mathbf{D}}_N^{-\frac{1}{2}}$ is different from the standard z-score of principal components analysis. To obtain the combination coefficients $\boldsymbol{r}^{(k)}$, let $\boldsymbol{t}^{(k)} \in \mathbb{R}^K$ be the left singular vector of $\frac{1}{\sqrt{N}}\widetilde{\mathbf{D}}_N^{-\frac{1}{2}}\widetilde{\mathbf{F}}_N^T$ (which is also the eigenvector of $\frac{1}{N}\widetilde{\mathbf{D}}_N^{-\frac{1}{2}}\widetilde{\mathbf{F}}_N^T\widetilde{\mathbf{F}}_N\widetilde{\mathbf{D}}_N^{-\frac{1}{2}} \approx \mathbf{D}^{-\frac{1}{2}}\widetilde{\mathbf{G}}\mathbf{D}^{-\frac{1}{2}}$ in the large N limit). Then $\boldsymbol{r}^{(k)} = \widetilde{\mathbf{D}}_N^{-\frac{1}{2}}\boldsymbol{t}^{(k)} \in \mathbb{R}^K$ can be treated as the combination prefactor of $\hat{M}_k$, to obtain the observables which correspond to the eigentasks. The merit of SVD analysis of $\frac{1}{\sqrt{N}}\widetilde{\mathbf{D}}_N^{-\frac{1}{2}}\widetilde{\mathbf{F}}_N^T$ is that we only need to work with a K-by-N matrix of features $\widetilde{\mathbf{F}}_N$, which is numerically cheaper than further constructing a Gram matrix $\frac{1}{N}\widetilde{\mathbf{F}}_N^T\widetilde{\mathbf{F}}_N$. We will explore more about the usage of our technique in sense of PCA in Appendix H.

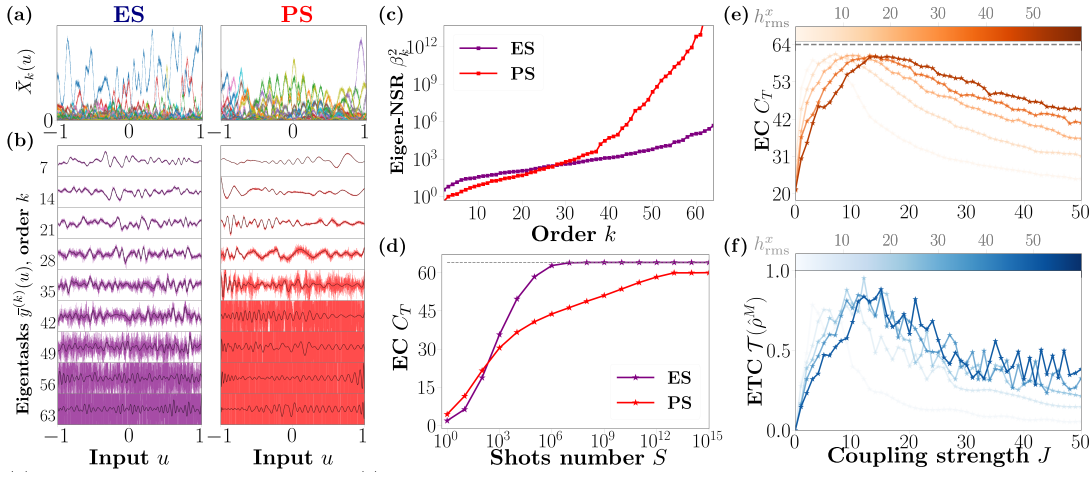

FIG. 7. Eigen analysis in a 6-qubit H-ansatz system (with $N = 5000$ and $S = 1000$) forming a 1D ring. The Hamiltonian parameters are chosen randomly with zero-mean and variance $(h^x_{\rm rms}, h^z_{\rm rms}, h^I_{\rm rms}) = (20, 5, 5)$, and $t = 5$ (See Appendix B 1 for details). Coupling strength is uniformly $J \neq 0$ (ES) or $J = 0$ (PS). (a) All $2^L = 64$ noisy features $\bar{X}_k(u)$ and (b) noisy eigentasks $\bar{y}^{(k)}(u) = \boldsymbol{r}^{(k)} \cdot \bar{\boldsymbol{X}}(u)$ for selected k from the features in (a), as well as their expected values $y^{(k)}(u) = \lim_{S\to\infty} \bar{y}^{(k)}(u) = \boldsymbol{r}^{(k)} \cdot \boldsymbol{x}(u)$ (black). (c) NSR spectrum β^2_k and (d) C_T vs shots S for both ES and PS encodings. (e) C_T at $S = 10^5$ and (f) ETC, $\bar{\mathcal{T}}(\hat{\rho}^M)$ in representative random 6-qubit H-ansatz, as a function of coupling strength J. The peaks of capacity and correlation coincide, around $J \sim h^x_{\rm rms}$.

Appendix E: H-ansatz quantum systems: NSR spectra, expressive capacity, and eigentasks

In this section, we evaluate the EC for quantum systems described by the H-ansatz introduced in Appendix B 1, as an example of how EC can be efficiently computed for a variety of general quantum systems, and is not just restricted to parameterized quantum circuits. The results of the analysis are compiled in Fig. 7, and discussed below.

Fig. 7(a) presents the set of features $\{\bar{X}_k(u)\}$ for typical $L = 6$ qubit ES and PS at $S = 1000$ with randomly chosen parameters (referred to as encodings, see caption). The resultant noisy eigentasks $\{\bar{y}^{(k)}(u)\}$ and NSR spectra $\{\beta^2_k\}$ extracted via the eigenvalue analysis are shown in Figs. 7(b) and 7(c) respectively. In the side-by-side comparison in Fig. 7(b), we clearly see the $J = 0$ ansatz transitioning to a regime with more noise at much lower k than the $J \neq 0$ ansatz. This is reflected in Fig. 7(c), the β^2_k spectrum, having a much flatter slope for larger k (note the plot is semilog). Finally, Fig. 7(d) shows the EC of both systems as a function of S. EC rapidly rises for small S for both systems, but the rise of the $J = 0$ system is steeper. After a certain threshold in S, however, the ES grows more rapidly, approaching the upper bound $2^6 = 64$ with $S \sim 10^8$; in contrast, the PS has a significantly lower C_T.

For $J \to \infty$ we also have $\bar{\mathcal{T}} = 0$ because $\hat{\rho}_0 = |0\rangle\langle 0|^{\otimes L}$ is an eigenstate of the encoding ($\hat{\rho}(u) = \hat{\rho}_0$). This implies there must be a peak at some intermediate J, which for both EC and ETC occurs when the coupling is proportional to the transverse field $J \sim h^x$.

Our results elucidate the same kind of improvement, as can be observed when we consider how the EC C changes with J, and compare it to the total correlation ETC $\bar{\mathcal{T}}$, as shown in Fig. 7(f). For $J \to 0$ we have a PS with $\bar{\mathcal{T}} = 0$, whereas in the $J \to \infty$ we also have $\bar{\mathcal{T}} = 0$ because $\hat{\rho}_0 = |0\rangle\langle 0|^{\otimes L}$ is an eigenstate of the encoding ($\hat{\rho}(u) = \hat{\rho}_0$). This implies there must be a peak at some intermediate J, which for both EC and ETC occurs when the coupling is proportional to the transverse field $J \sim h^x$. At finite S, increased ETC is directly related to a higher EC.

Another interesting aspect is the clear trend seen in the maximization of EC around $J \sim h^x_{\rm rms}$ for various $h^x_{\rm rms}$, possibly hinting at the role of increased entanglement around the MBL phase transition in random spin systems [30]. This trend is consistent with results in quantum metrology – in general, the SNR obtained from averaging L uncorrelated probes scales as $1/\sqrt{L}$. This scaling can become favorable in the presence of entanglement and other non-classical correlations, in which case the scaling of the SNR can show up to a quadratic improvement $1/L$ [29]. For even larger J, we find that $\hat{\rho}(u) \to \hat{\rho}_0 = |0\rangle\langle 0|^{\otimes L}$, which clearly reduces $\bar{\mathcal{T}}$, but also C_T as the quantum system state becomes u-independent.

Appendix F: Scaling with quantum system size

An important question in quantum machine learning applications is the possible advantage of using larger quantum systems for information processing. In this section, we present preliminary results of scaling with quantum system size. The left panel of Fig. 8 shows EC vs L at select S values for H-ansatz, while the right panel shows two encodings in the C-ansatz device, as well as their noisy simulations. In both plots, the dashed line indicates the $S \to \infty$ result $C_T = 2^L$. We see that the EC increases when adding more qubits into the Ising chain for the H-ansatz, or when increasing the number of circuit qubits L for the C-ansatz. Note, however, that at any finite S the noise-constrained EC falls off the exponential bound for $S \to \infty$. The dropoff is particularly severe for the IBMQ device, where we are limited to just $S \sim 10^4$, which significantly suppresses the EC even for $L = 7$ qubits. Note, however, that even if one is well below $C_T = 2^L$ due to this finite sampling constraint, increasing the dimension of the quantum system is always an effective way to increase the EC, particularly when compared to the logarithmic growth with S of Fig. 2 of Main Text.

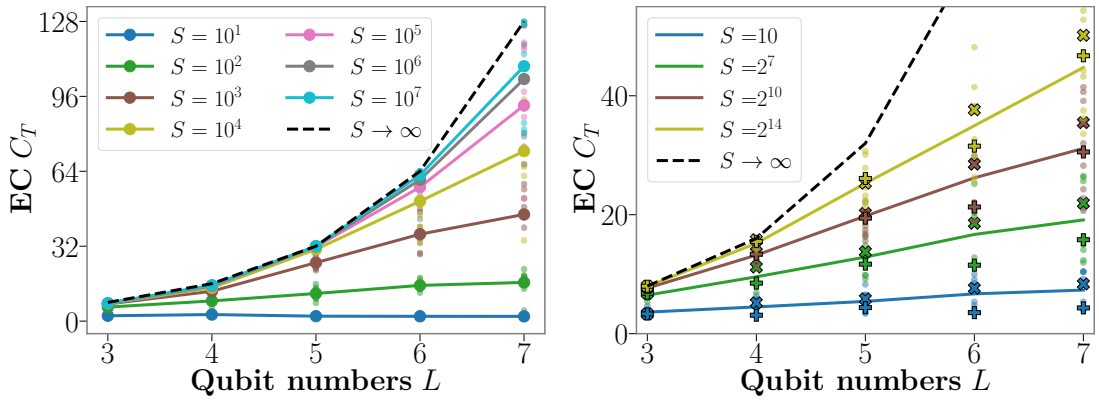

FIG. 8. (a) H-ansatz and (b) C-ansatz at finite S as a function of qubit number L. Various colours indicate different S values, with the $S \to \infty$ bound in dashed black. Individual noisy simulations are indicated in small and transparent dots, with their average as a thick line, and the EC of the C-ansatz device for encoding 1 and 2 are indicated with '$\times$' and '$+$' respectively.

Appendix G: Quantum correlation metrics

There is no one standard metric to quantify entanglement or correlation in a many-body state. The metric we introduce here, the *quantum total correlation*, is a quantity inspired by the classical total correlation of L random variables $(b_1, \cdots, b_L)$, that is $\sum_{l=1}^{L} \mathrm{H}(b_l) - \mathrm{H}(b_1, \cdots, b_L)$. Using chain rule of Shannon entropy $\mathrm{H}(b_1, b_2, \cdots, b_L) = \mathrm{H}(b_1) + \mathrm{H}(b_2|b_1) + \cdots + \mathrm{H}(b_L|b_1, b_2, \cdots, b_{L-1})$

$$\sum_{l=2}^{L} \mathrm{H}(b_l) - \mathrm{H}(b_1, b_2, \cdots, b_L) = \sum_{l=1}^{L} \mathrm{H}(b_l) - \sum_{l=1}^{L} \mathrm{H}(b_l|b_1, b_2, \cdots, b_{l-1}) = \sum_{l=2}^{L} \mathrm{I}(b_1, \cdots, b_{l-1}; b_l) \in [0, L-1], \tag{G1}$$

we can see that the classical total correlation tells us how a set of random variables reveals information of each other. Similarly, quantum total correlation can be defined as [26, 27]

$$\mathcal{T}(\hat{\rho}) = \sum_{l=1}^{L} \mathrm{S}(\hat{\rho}_l) - \mathrm{S}(\hat{\rho}) \tag{G2}$$

where S is von Neumann entropy and $\hat{\rho}_l := \mathrm{Tr}_{[L]\setminus\{l\}}\{\hat{\rho}\}$ is the subsystem state at qubit l. Due to the subadditivity of von-Neumann entropy $\sum_{l=1}^{L} \mathrm{S}(\hat{\rho}_l) \geq \mathrm{S}(\hat{\rho})$, we conclude that the quantum total correlation is non-negative, and is zero iff the state $\hat{\rho} = \bigotimes_{l=1}^{L} \hat{\rho}_l$ is a product state.

In this paper's measurement scheme, the specific readout POVMs are the projectors onto the computational states $\{|\boldsymbol{b}_k\rangle\langle\boldsymbol{b}_k|\}_{k\in[K]}$. Thus, we are in particular interested in analyzing the post-measurement state $\hat{\rho}^M(u) = \sum_k \rho_{kk}(u) |\boldsymbol{b}_k\rangle\langle\boldsymbol{b}_k|$

whose subsystems are correspondingly in states $\hat{\rho}_l^M(u) = \mathrm{Tr}_{[L]\setminus\{l\}}\left\{\hat{\rho}^M(u)\right\}$. We compute the *average* quantum total correlation over the input domain u with respect to the input probability distribution $p(u)$:

$$\bar{\mathcal{T}}\left(\hat{\rho}^M\right) = \mathbb{E}_u\left[\sum_{l=1}^{L}\mathrm{S}(\hat{\rho}_l^M(u)) - \mathrm{S}(\hat{\rho}^M(u))\right] = \mathbb{E}_u\left[\sum_{l=1}^{L}\mathrm{H}(b_l(u)) - \mathrm{H}(b_1(u),\cdots,b_L(u))\right] \tag{G3}$$

where the second equality comes from the diagonal nature of post-measurement state which reduces the quantum total correlation to a normal classical total correlation.

The post-measurement quantum total correlation always reaches its maximum $L-1$ when the diagonal terms of the state is a GHZ-type state. Also as a comparison, for a W-state $|W\rangle = \frac{1}{\sqrt{L}}\left(|10\cdots0\rangle + |01\cdots0\rangle + \cdots + |00\cdots1\rangle\right)$, then post-measurement quantum total correlation $\mathrm{T}(|W\rangle)$ is

$$L\left(-\left(\frac{1}{L}\right)\log_2\left(\frac{1}{L}\right) - \left(\frac{L-1}{L}\right)\log_2\left(\frac{L-1}{L}\right)\right) - L\left(-\left(\frac{1}{L}\right)\log_2\left(\frac{1}{L}\right)\right) = (L-1)\log_2\left(\frac{L}{L-1}\right). \tag{G4}$$

which is upper bounded by $\lim_{L\to\infty}\mathcal{T}(|W\rangle) = \frac{1}{\ln(2)} \approx 1.443$.

Appendix H: Guidance from EC theory: principal component analysis with respect to quantum noise

Another fundamental use of the capacity spectrum analysis we propose is giving a natural truncation of eigentask. In machine learning theory, the technique of projection of a high-dimensional data to a far lower subspace is called *principal component analysis*. Within the computing architecture we are discussing, we are trying to use some K'-dimensional data where $K' \ll K$ to approximate the original data as much as possible. More specifically, consider a given function $f(u)$, we hope to find K' functions $\{G^{(k)}(u)\}_{k\in[K']}$ where $G^{(k)}(u) = \sum_{k'=0}^{K-1} g_{k'}^{(k)} x_{k'}(u)$ lies in the space spanned by measured features $G^{(k)}(u) \in \mathrm{Span}\{\boldsymbol{x}\}$, such that the relative mean square error

$$\min_{\boldsymbol{W}} \frac{\mathbb{E}_u\left[\left|f - \sum_{k=1}^{K'} W_k\left(\sum_{k'=0}^{K-1} g_{k'}^{(k)}\bar{X}_{k'}\right)\right|^2\right]}{\mathbb{E}_u[|f|^2]} \tag{H1}$$

is much smaller as possible. According to Appendix C, the solution to $\{\boldsymbol{g}^{(k)}\}_{k\in[K']}$ is exactly $\boldsymbol{g}^{(k)} = \boldsymbol{r}^{(k)}$. Fig. 9 supplies a concrete example of fitting linear function $f(u) = u$, by setting $K' = 40$ in a 6-qubit system (and thus $K = 64$).

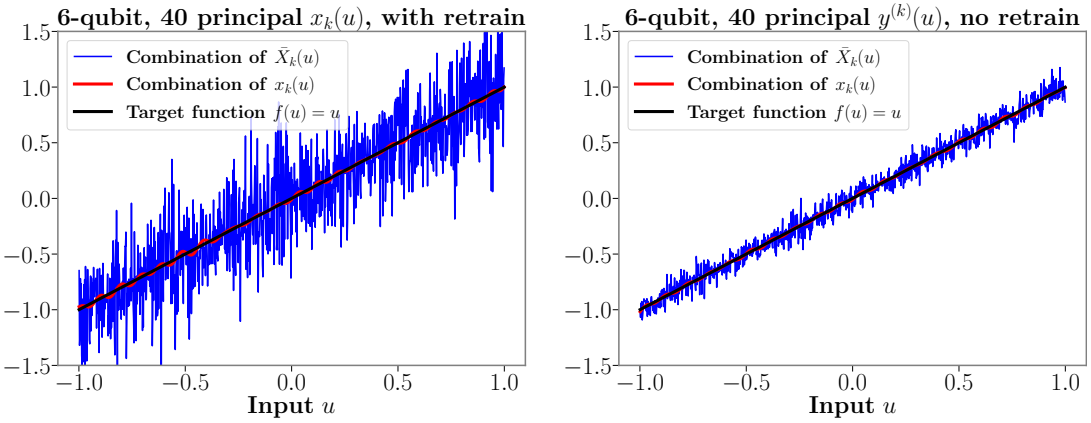

FIG. 9. Projection onto 40-dimensional space spanned by 40 principal $x_k(u)$ vs. spanned by 40 principal $y^{(k)}$, in a 6-qubit H-encoding system. The number of shots is fixed as $S = 5000$.

Fig. 9(a) shows the projection onto the space spanned by the dominant 40 readout features. Here, by "dominant" we mean one can first train by least square regression to get an output weight $\boldsymbol{w} \in \mathbb{R}^K$, and then select corresponding w_k with the leading K' largest $w_k^2 \cdot \mathbb{E}_u[|x_k|^2]$. Then we need to use these K' features to retrain and obtain a new output weight $\boldsymbol{w}' \in \mathbb{R}^{K'}$. In such particular example, $\boldsymbol{g}^{(k)}$ are some one-hot vectors where the index of 1 are chosen by the sorting K' largest $w_k^2 \cdot \mathbb{E}_u[|x_k|^2]$ as we

described before. We can compare the the relative mean square error with the case of $\boldsymbol{g}^{(k)} = \boldsymbol{r}^{(k)}$, the eigentasks. The latter one shows an approximation function with conspicuously much smaller relative mean square error.

One fundamental question is: what will be an appropriate selection of K' in practice. In Appendix D we claim that those β_k^2 has stronger noise than signal itself, which should be excluded when taking the linear combination of measured features (or equivalently taking the linear combination of eigentasks). Namely we should defined the cut-off $K_c(S)$ such that

$$K_c(S) = \max_{\beta_k^2 < S} k. \tag{H2}$$

Based on this observation, we can further explore the trend of $K_c(S)$ when qubit number L is scaled. As we showed in main text. The eigen-NSR spectra growth much slower when L increases. Then the quantum system is able to provide much more eigentasks with more signal than noise. Fig. 10(a) shows spectrum in H-encoding quantum system with size $L = 3 \sim 8$ with fixed hyperparameters. Notice that shot number $S = 5000$ here is not a larger number, which means that we cannot sample enough shots so that features converges to its mean in $2^8 = 256$ dimensional Hilbert space. But applying eigentasks analysis in this example still shows a fast decay of relative error $\min_{\boldsymbol{W}} \mathbb{E}_u[|f - \sum_{k=0}^{K_c(S)} W_k \bar{y}^{(k)}|^2]/\mathbb{E}_u[|f|^2]$ until the fitting accuracy saturates at $L = 8$.

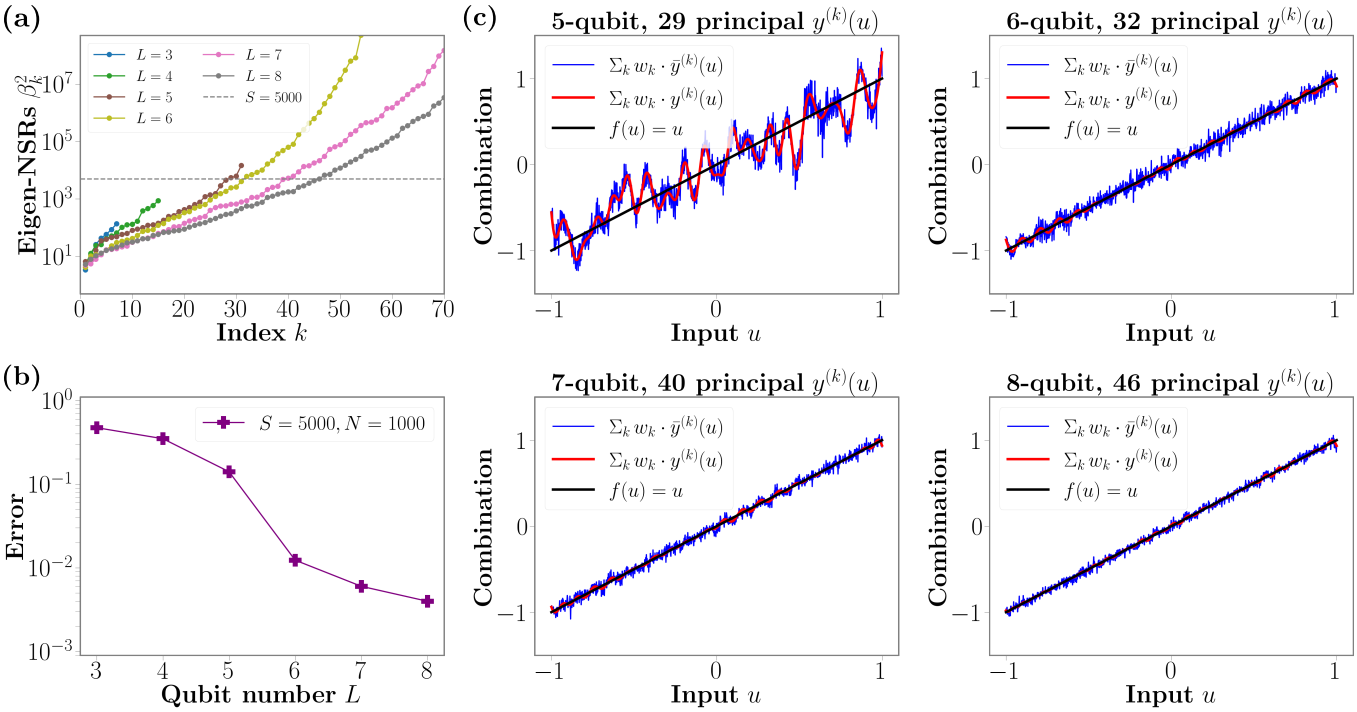

FIG. 10. PCA for different ES H-encoding system size $L = 3, 4, 5, 6, 7, 8$ with fixed hyperparameters and $S = 5000$. (a) Eigen-NSRs spectrum of different sized system. (b) Relative error $\min_{\boldsymbol{W}} \mathbb{E}_u[|f - \sum_{k=1}^{K_c} W_k \bar{y}^{(k)}|^2]/\mathbb{E}_u[|f|^2]$ for fitting $f(u) = u$, where K_c can be read out from (a). (c) Combination of K_c eigentasks $\sum_{k=0}^{K_c(S)} w_k y^{(k)}(u)$ and noisy eigentasks $\sum_{k=0}^{K_c(S)} w_k \bar{y}^{(k)}(u)$ in $L = 5, 6, 7, 8$ qubits system.

Appendix I: Quantum-noise-PCA in classification problem

The highly nonlinear readout feature $x_k(u)$ should have Taylor expansion $x_k(u) = \sum_j^{\infty} (\mathbf{T})_{kj} u^j$. Such complicated functions will span a certain functional space. One fundamental question is what the limit of approximation ability based on the architecture we proposed. Hereby we first show that this architecture *under infinite sampling* is capable of approximating *any* continuous function on the domain $[-1, 1]$ to arbitrary precision. Furthermore, the linearity of quantum moment readout and complexity of quantum evolution will help us to understand why such a quantum system has capability to approximate a highly nonlinear function, under finite and bounded computational resources. Exploring the capacity for function approximation under finite measurement resources, as is done in the main text and Appendix C, highlights the fundamental limitations places by quantum noise on computation using the QRC.

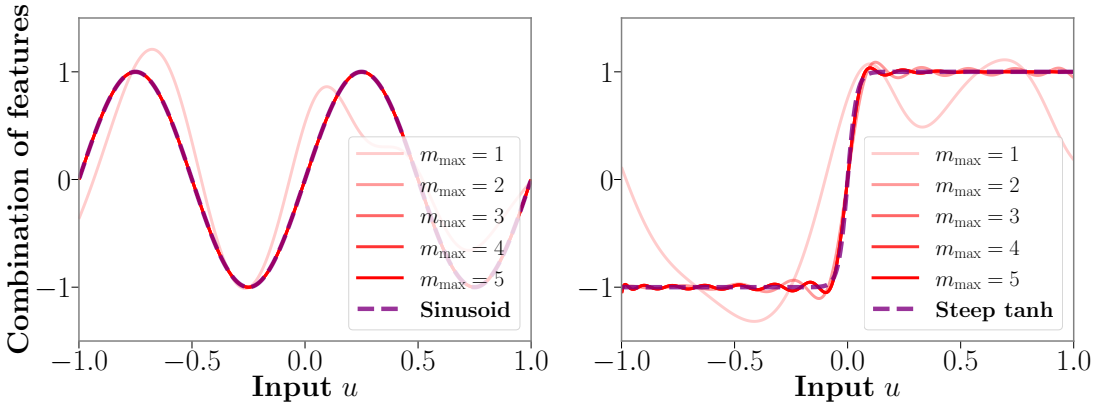

FIG. 11. Function approximation by using $y = \sum_{k=0}^{K-1} w_k x_k(u)$ (solid red lines) to approximate sine function and steep tanh function (dashed purple lines) in a 5-qubit quantum annealing system, where $K_{\text{eff}} = \sum_{m=0}^{m_{\max}} \binom{L}{m}$ depends on different quantum moment thresholds $m_{\max} = 1, 2, 3, 4, 5$. The hyperparameters are $(J_{\max}; \bar{h}^x, h^x_{\text{rms}}; \bar{h}^I, h^I_{\text{rms}}) = (1; 3, 1; 5, 2)$ in unit $1/t$ and no h^z field. This simulation shows that for some simple functions, it is sufficient to merely use lower order moments, *e.g.*, $m_{\max} = 2$ in sine function and $m_{\max} = 3$ in steep tanh function.

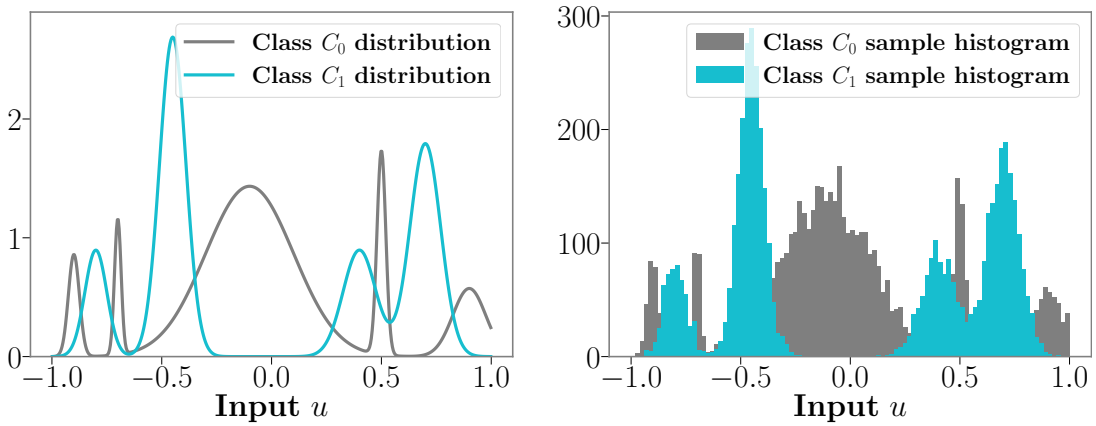

FIG. 12. (Left) Distribution $p_0(u)$ and $p_1(u)$ for classes C_0 and C_1, respectively. (Right) The histogram of C_0 and C_1. Each class contains 5000 samples.

1. Function approximation universality

A very general question is that what type of functions can such a single-step quantum approximate. One conclusion which can be drawn is the *function approximation universality*. That is, give any continuous function from space of continuous functions on domain $[-1, 1]$, *i.e.* $\phi \in \mathscr{C}([-1, 1], \mathbb{R})$, for any given error $\varepsilon > 0$, there always exists a function $\varphi(u) = \boldsymbol{w} \cdot \boldsymbol{x}(u)$ such that

$$|\varphi(u) - \phi(u)| \leq \varepsilon \tag{I1}$$

for any input $u \in [-1, 1]$. The proof is also employing the well-known Stone-Weierstrass theorem. For our particular architecture, $D = [-1, 1]$ is obviously a compact space, while point-separation can also be trivially fulfilled by a single qubit system ($L = 1$). The subalgebra structure of the function family generated by quantum systems is automatically satisfied in representation of moment in family of all product systems.

2. 1D classification as function approximation for noiseless measured features

In this section, we will show how the function approximation universality of architecture described in Appendix I 1 enables it to perform – among others – paradigmatic machine learning tasks such as classification.

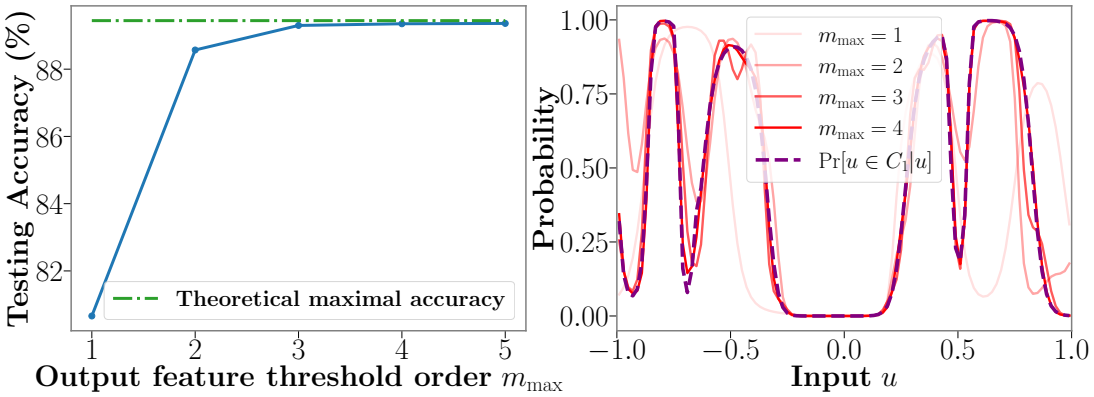

FIG. 13. 1D classification as function approximation in a 5-qubit quantum system with full connectivity. The hyperparameters are $(J_{\max}; \bar{h}^x, h^x_{\rm rms}; \bar{h}^I, h^I_{\rm rms}) = (1; 3, 1; 8, 5)$ in unit $1/t$ and no h^z field. (Left) Testing accuracy as a function highest order $m_{\max}$ of moment feature. (Right) Conditional distribution $\Pr[u \in C_1|u]$ (purple dashed line) vs. readout features $\sigma(\boldsymbol{w}\cdot\boldsymbol{x}(u))$ with $m_{\max} = 1, 2, 3, 4$ (red solid line). $m_{\max} = 4$ saturates the approximation accuracy.

Suppose two classes C_0 and C_1 of samples, each of which is generated from distributions $p_0(u)$ and $p_1(u)$ respectively. The probability of occurrence of C_0 and C_1 are both 50%, and we simply let each class equally contain 5000 samples and thus $N = 10000$ samples in total. Both distribution are artificially defined by summing several Gaussian distributions with different amplitudes and widths together. Domain of both distributions are restricted in $[-1, 1]$ and both distributions are also normalized. Due to the overlap of two distributions, there is some theoretical maximal classical accuracy to distribution whether a given u belongs to either C_0 or C_1.

During the training, we feed each sample $u^{(n)}$ (belonging to class $C_{c^{(n)}}$) into a 5-qubit quantum system. The quantum system will be read out with $K_{\rm eff} = \sum_{m=0}^{m_{\max}} \binom{L}{m}$ different features $\{x_k(u^{(n)})\}_{k\in[K_{\rm eff}]}$. Then features of N sample forms the regressor matrix. According to the standard supervised learning procedure, we simply train based on $(\boldsymbol{x}(u^{(n)}), c^{(n)})$ by logistics regression where one should minimize the cross-entropy loss

$$\mathscr{L}(\boldsymbol{W}) = \frac{1}{N}\sum_{n=1}^{N}\Big[-c^{(n)}\log\Big(\sigma(\boldsymbol{W}\cdot\boldsymbol{x}(u^{(n)}))\Big) - \Big(1-c^{(n)}\Big)\log\Big(1-\sigma(\boldsymbol{W}\cdot\boldsymbol{x}(u^{(n)}))\Big)\Big] \tag{I2}$$

where σ is the sigmoid function $\sigma(y) = \frac{1}{1+e^{-y}}$. A small L_2 penalty $\lambda\|\boldsymbol{W}\|^2$ (where $\lambda = 10^{-6}$) is added to Eq. (I2) for preventing overfitting. The optimal $\boldsymbol{W}$ is then simply the set of weights that minimizes this cost function,

$$\boldsymbol{w} = \mathrm{argmin}_{\boldsymbol{W}}\,\{\mathscr{L}(\boldsymbol{W})\} \tag{I3}$$

We test the fidelity of learning the classification task by determining the accuracy of classification on a testing set formed by drawing $N = 10000$ new samples (independent of the training set) as a function of the order of output moments extracted, $m_{\max} = 1, 2, 3, 4, 5$, corresponding to reading out $K_{\rm eff} = 6, 16, 26, 31, 32$ features respectively. The resulting testing accuracy is plotted in the left panel of Fig. 13). We see that the testing accuracy converges to the theoretical maximal accuracy (dashed green) with increase in readout features.

Importantly, one can show that this improvement in learning performance coincides with training of optimal weights $\boldsymbol{w}$ such that the QRC is able to approximate the conditional distribution $\Pr[u \in C_1|u]$ of the two classes with increasing accuracy (lower error). To verify this, we first numerically compute all $K = 32$ readout feature functions $\boldsymbol{x}(u)$ of the system, by sweeping 500 equidistant values of $u \in [-1, 1]$. Effectively learning the conditional distribution means that $\sigma(\boldsymbol{w}\cdot\boldsymbol{x}(u)) \approx \Pr[u \in C_1|u]$. It is equivalent to use $\boldsymbol{w}\cdot\boldsymbol{x}(u)$ to approximate the following function:

$$\boldsymbol{w}\cdot\boldsymbol{x}(u) \approx \sigma^{-1}(\Pr[u \in C_1|u]). \tag{I4}$$

We therefore see that the function approximation universality property of the architecture discussed in Appendix I 1 enables its use as a generic classifier.

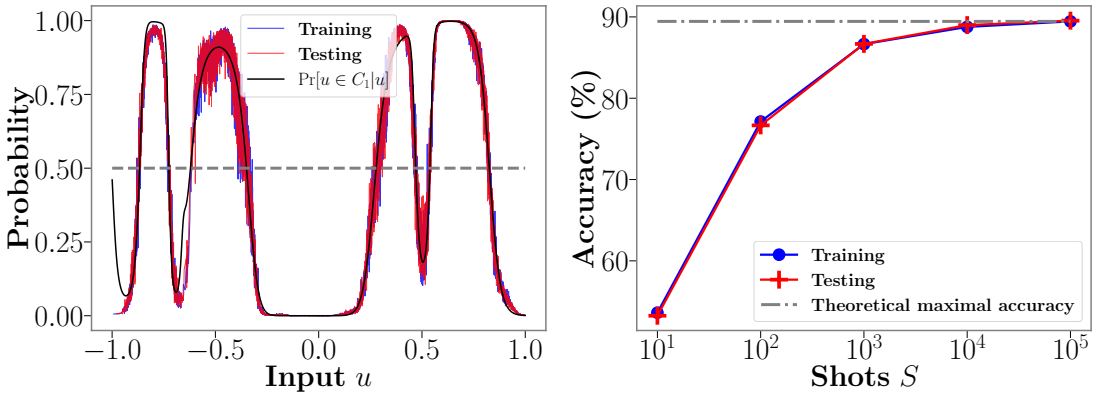

FIG. 14. (Left) The linear combination with sigmoid activation, that is the stochastic function $\sigma\left(\sum_{k'=1}^{K_c(S)} w_{k',\mathrm{Train}}(\tilde{\boldsymbol{r}}_N^{(k)}\cdot\bar{\boldsymbol{X}}_{\mathrm{Train}})_{k'}\right)$ (blue line) and $\sigma\left(\sum_{k'=1}^{K_c(S)} w_{k',\mathrm{Train}}(\tilde{\boldsymbol{r}}_N^{(k)}\cdot\bar{\boldsymbol{X}}_{\mathrm{Test}})_{k'}\right)$ (red line), compared with the true conditional probability $\Pr[u\in C_1|u]$ (black line). (Right) Training accuracy and testing accuracy. They saturate the theoretical maximal accuracy as S reaches $10^4\sim10^5$. Their agreement shows the quantum measurement noise serves well as a regularizer.

3. Solving classification problem by quantum-noise-PCA

Now we can solve the classification task above by using the quantum-noise princilpal component analysis we learn from capacity analysis. Suppose a physical system with $L=5$ qubits and ring connectivity, we choose the hyperparameter to be $J=2$, $h^x_{\mathrm{rms}}=h^z_{\mathrm{rms}}=h^I_{\mathrm{rms}}=5$ and $t=3$. In this H-encoding scheme, we can obtain $K=32$ measured features on each of $N=10^5$ samples $\{u^{(n)}\}$ (5000 in class C_0 and 5000 in class C_1). We emphasize here that the underlying marginal distribution $p(u)$ is no longer uniform here, and it will make both $\{\beta_k^2\}$ and $\{\boldsymbol{r}^{(k)}\}$ very different.

Given the number of shots $S\in[10^1,10^5]$, we can still compute the empirical $\tilde{\boldsymbol{r}}_N^{(k)}$ and estimating β_k^2 by using the correction techniques we used in Appendix D. By comparing the estimated $(1-\tilde{\alpha}_{N,k})/(\tilde{\alpha}_{N,k}-\frac{1}{S})$ and S, we can figure out the cutoff order $K_c(S)$ and combination coefficients $\tilde{\boldsymbol{r}}_N^{(k)}$, based on which we can define a set of observables

$$\hat{O}_k=\sum_{k'=0}^{K-1}\tilde{\boldsymbol{r}}^{(k)}_{N,k'}\hat{M}_{k'}\quad k=0,1,\cdots,K_c(S). \tag{I5}$$

It is equivalent to say, by measuring $\hat{O}_k$, we can effectively obtain eigentasks $\tilde{\boldsymbol{r}}_N^{(k)}\cdot\bar{\boldsymbol{X}}_{\mathrm{Train}}$. Then we can apply standard logistics regression on those eigentasks as we did in Eq. I2. The only difference is we no longer need any regularization term as penalty like $\lambda\|\boldsymbol{W}\|^2$. The training procedure eventual yield $\boldsymbol{w}_{\mathrm{Train}}\in\mathbb{R}^{K_c(S)}$, together with $\tilde{\boldsymbol{r}}_N^{(k)}$ and $K_c(S)$.

Now we generate a totally new and independent set of u's for testing purpose. By measuring $\hat{O}_k$, one get eigentasks $\tilde{\boldsymbol{r}}_N^{(k)}\cdot\bar{\boldsymbol{X}}_{\mathrm{Test}}$. By plugging $\boldsymbol{w}_{\mathrm{Train}}\in\mathbb{R}^{K_c(S)+1}$, together with $\tilde{\boldsymbol{r}}_N^{(k)}$ and $K_c(S)$ in training, we can achieve the testing accuracy. The agreement between training and testing accuracy show that the quantum measurement noise effectively works as a regularizer, and do a pretty good job (see Fig. 14).

Appendix J: Finite sampling bound and uncertainty propagation

We conclude that the principle advantage brought about by entanglement in this sections. There we observe that for certain inputs u (that depend on the input encoding) the measurement of an ES when mapped into the moment space can generate distributions that can be highly anisotropic at finite S. While for PS these distributions are generally isotropic unless they are close to the boundaries of the output domain (when the encoding produces outputs that are eigenstates of the measurement basis). We observe that this trend is also present in the experimental system despite non-idealities. The origin of higher expressive capacity at large S provided by ESs can be traced back to this basic feature. To be more specific, let $\hat{M}_k=\hat{\sigma}^z_{l_1}\hat{\sigma}^z_{l_2}\cdots\hat{\sigma}^z_{l_m}$, and $\bar{X}_k(u)$ be empirical mean based on S sampling. Notice that the variance of $\bar{X}_k$ is

$$\mathrm{Var}[\bar{X}_k]=\frac{1}{S}\left(\langle(\hat{\sigma}^z_{l_1}\hat{\sigma}^z_{l_2}\cdots\hat{\sigma}^z_{l_m})^2\rangle-\langle\hat{\sigma}^z_{l_1}\hat{\sigma}^z_{l_2}\cdots\hat{\sigma}^z_{l_m}\rangle^2\right)=\frac{1}{S}(1-x_k^2(u)). \tag{J1}$$

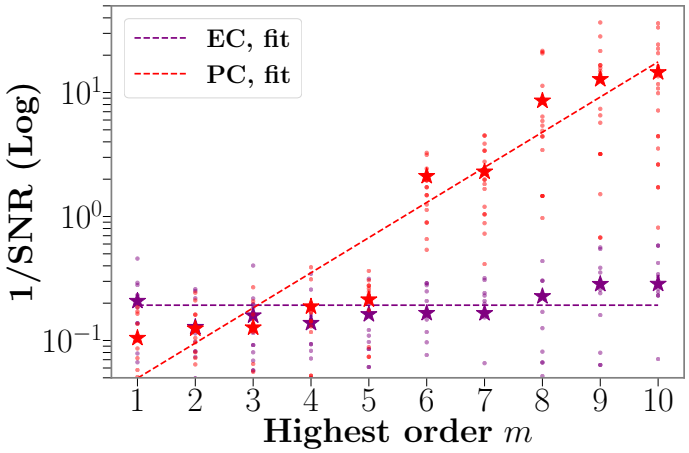

FIG. 15. NSR of ES vs PS in a 10-qubit quantum annealing system with shot number $S = 1000$ by feeding $u = 1/2$. The hyperparameters are chosen to be $(\bar{h}^x, h^x_{\rm rms}; \bar{h}^z, h^z_{1,\rm rms}) = (8, 2; 3, 2)$ in unit $1/t$. The purple and red colors correspond to coupling being switched on and off, respectively; and the coupling hyperparameter in ES is $J_{\max} = 2/t$. For each m, the $N = 30$ dots are relative error $x^{(r)}_k(u)/x_k(u) - 1$ of 30 repetitions $r = 1, 2, \cdots, 30$. The standard deviation of those relative errors (namely NSR) are also plotted. The ES NSR (purple stars) is well fitted by $O(1/\sqrt{S})$ (purple dashed line) while the PS NSR (red stars) scales exponentially as $O(2^m/\sqrt{S})$ (purple dashed line). We take y-axis being log-scale, and one may find in these regime ES 1/SNR grows exponentially faster than PS NSR (red stars) and hence PS readout scheme will be less powerful in sense of quantum sampling noise resistant.

By central limit theorem,

$$\bar{X}_k(u) = x_k(u) + \delta_k(u) = x_k(u) + \frac{1}{\sqrt{S}}\zeta_k(u), \tag{J2}$$

where random sampling noise $\zeta_k(u) \approx \sqrt{1 - x_k^2(u)}\epsilon$ and $\epsilon \sim \mathcal{N}(0, 1)$ is standard Gaussian. For quantum moment readout, the amplitude of relative error is

$$\left|\frac{\delta_k(u)}{x_k(u)}\right| \approx \sqrt{\frac{1 - x_k^2(u)}{x_k^2(u)}}\frac{1}{\sqrt{S}} \propto \frac{1}{\sqrt{S}}. \tag{J3}$$

For classical polynomial readout the amplitude of relative error is obtained by rule of uncertainty propagation

$$\begin{aligned}
&\left|\frac{(x_{l_1}(u) + \delta_{l_1}) \cdots (x_{l_m}(u) + \delta_{l_m}) - x_{l_1}(u) \cdots x_{l_m}(u)}{x_{l_1}(u) \cdots x_{l_m}(u)}\right| \approx \left|\frac{\delta_{l_1}}{x_{l_1}(u)} + \cdots + \frac{\delta_{l_m}}{x_{l_m}(u)}\right| \\
&\approx \left(\sqrt{\frac{1 - x_{l_1}^2(u)}{x_{l_1}^2(u)}} + \cdots + \sqrt{\frac{1 - x_{l_m}^2(u)}{x_{l_m}^2(u)}}\right) \times \frac{1}{\sqrt{S}} \propto m \times \frac{1}{\sqrt{S}}.
\end{aligned} \tag{J4}$$

If there is no entanglement in quantum system, then the readout features for both quantum moment readout and classical polynomial readout are the same $\langle \hat{\sigma}^z_{l_1} \hat{\sigma}^z_{l_2} \cdots \hat{\sigma}^z_{l_m} \rangle = \langle \hat{\sigma}^z_{l_1} \rangle \langle \hat{\sigma}^z_{l_2} \rangle \cdots \langle \hat{\sigma}^z_{l_m} \rangle$. However, even if the expectations under infinite sampling limit $S \to \infty$ are the same, the measurement noise under finite sampling are still different. For classical polynomial readout, the scaling of still follows the simple additivity relation of uncertainty propagation in Eq. (**??**). But now the noise of $x_{l_1}(u) \cdots x_{l_m}(u)$ in quantum moment readout will be very strong, this is because $x_{l_1}(u) \cdots x_{l_m}(u)$ is now close to zero, thus

$$\left|\frac{\delta_k}{x_k(u)}\right| \approx \frac{1}{x_k(u)}\frac{1}{\sqrt{S}} = \frac{1}{x_{l_1}(u) \cdots x_{l_m}(u)}\frac{1}{\sqrt{S}} \propto 2^m \times \frac{1}{\sqrt{S}}. \tag{J5}$$