Quantum sequential scattering model for quantum state learning

Mingrui Jing [1,*] Geng Liu [2,3,*] Hongbin Ren [3] and Xin Wang [1,3,†]

[1]*Thrust of Artificial Intelligence, Information Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, China*

[2]*School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), Longgang, Shenzhen 518172, China*

[3]*Institute for Quantum Computing, Baidu Research, Beijing 100193, China*

(Received 4 December 2023; revised 14 March 2024; accepted 13 May 2024; published 18 June 2024)

Learning probability distribution is an essential framework in classical learning theory. As a counterpart, quantum state learning has spurred the exploration of quantum machine learning theory. However, as dimensionality increases, learning a high-dimensional unknown quantum state via conventional quantum neural network approaches remains challenging due to trainability issues. In this work we devise the quantum sequential scattering model, inspired by the classical diffusion model, to overcome this scalability issue. Training of our model could effectively circumvent the vanishing gradient problem for a large class of high-dimensional target states possessing polynomial-scaled Schmidt ranks. Theoretical analysis and numerical experiments provide evidence for our model's effectiveness in learning both physically and algorithmically meaningful quantum states and outperform the conventional approaches in training speed and learning accuracy. Our work indicates that an increasing entanglement (a property of quantum states) in the target states necessitates a larger-scaled model, which could reduce our model's learning performance and efficiency.

DOI: 10.1103/PhysRevA.109.062425

I. INTRODUCTION

Quantum computing, as a prospective advanced computational framework, is expected to provide significant inspiration in both theoretical and experimental aspects. Meanwhile, the remarkable achievements of influential classical machine learning models [1] have spurred the evolution of their quantum counterparts within the emerging field of quantum machine learning (QML) [2–5]. In particular, learning correlations between individual events and data is one of the crucial tasks in this area, where the correlations between random samples can be characterized by some probability distributions [6] of some stochastic process. Classical distribution learning aims to reconstruct a sample generator that could simulate such a correct stochastic process.

In quantum computing, the correlations between quantum data are encoded within the quantum states. Consequently, the task of learning an arbitrary quantum state bears a resemblance to classical distribution learning. Typical quantum state learning aims to efficiently reconstruct a complete representation of a given target state instead of numerically simulating it [7], which can be further used in quantum data encoding, data analysis, and quantum simulation [8–10].

A plethora of approaches and schemes have been designed to learn correlations for both scenarios. Classically, models such as continuous evolutionary algorithms [11,12] and supervised learning within the neural network framework, including the Boltzmann machine, graph neural network, and diffusion model [13–16], have been invented to determine distribution learning. Meanwhile, with the quick expansion of the field of noisy intermediate-scale quantum (NISQ) computing [17], the concept of quantum neural networks (QNNs), serving as quantum counterparts to classical neural networks, has emerged. The gradually reliable hybrid quantum-classical techniques make QML well suited for handling state learning tasks [18–21].

In this work we focus on the QNN-powered algorithms combining both classical and quantum computation. Utilizing parametrized quantum circuits, these algorithms explore the Hilbert space through classical optimization techniques involving gradient descent or gradient-free methods and then determine the optimal parameters [22,23]. Beyond our scope, schemes using shadow tomography [24,25] fulfill another category of state learning with the aim of characterizing the classical information of quantum states.

As the main solution to quantum state learning, the implementation of the QNN-based methods suffers challenges including scalability and training efficiency issue [26,27], dense local minima [28–30], and training phase transition [31,32]. Specifically, deep QNNs containing surplus circuit layers could ensure reachability of any unitary matrix in the Hilbert space. Such features will experience exponentially vanishing gradients during training, a phenomenon referred to as barren plateaus [26], leading to a flattening of the cost function landscape as the system scales up.

To overcome the above bottleneck, we drew inspiration from the classical diffusion model [33], which dilates information regarding Gaussian distribution using sequential diffusion layers. The main idea behind our work is to conduct state learning by progressively augmenting subsystems in a sequential manner, which is achieved by purifying subsystem states and layerwise training [34]. From an information theory perspective, the comprehensive learning procedure can

*These authors contributed equally to this work.

†Corresponding author: felixxinwang@hkust-gz.edu.cn

be likened to a quantum information diffusion mechanism across a cascade of system dilations, analogous to the classical diffusion model.

In this work we devise a quantum sequential scattering model (QSSM) integrating Uhlmann's purification theorem [35] into the training of QNNs. Our main contributions involve (i) conceptually combining quantum information diffusion and adaptive quantum state learning, (ii) technically devising a quantum neural network model, namely, the QSSM, and the state learning algorithm via a sequentially subsystem-learning strategy, (iii) theoretically proving the effectiveness of the state learning algorithm and a polynomial-scaled gradient variance of the QSSM which indicates an avoidance of barren plateaus for rank-restricted state learning, and (iv) numerically demonstrating our results on learning different quantum states involving the noise effects. We compare the QSSM directly to the conventional QNN model for handling state learning tasks and showcase its enhancement in both training efficiency and learning accuracy.

II. MAIN RESULT

A. Quantum sequential scattering model

We present a sketch of the design for the quantum sequential scattering model, which could efficiently accomplish the state learning tasks. Given an n-qubit quantum state ρ represented by certainly ordered qubits denoted by $q_1, q_2, \ldots, q_n$, a kth partition of ρ separates the state into bipartite subsystems $\mathcal{A}_k$ and $\bar{\mathcal{A}}_k$ covering the first k qubits and the remaining qubits, respectively, where $1 \leqslant k \leqslant n$. For $k = n$, $\bar{\mathcal{A}}_k$ becomes trivial and $\mathcal{A}_k = \mathbb{C}^{2^n}$. We denote by ρ_k the partial state on system $\mathcal{A}_k$, i.e., $\rho_k = \mathrm{Tr}_{\bar{\mathcal{A}}_k}(\rho)$, where $\mathrm{Tr}_B(\cdot)$ denotes the partial trace operation on subsystem B. By fixing the number of qubits in the system, the dimension of $\mathcal{A}_k$ increases as k grows and the dimension of $\bar{\mathcal{A}}_k$ decreases.

The fundamental idea of the QSSM is to construct ρ_k for each k. In contrast, traditional QNN learning handles the entire system at a time. Suppose we have access to the copies of a pure target state $\rho = |\phi\rangle\langle\phi|$ from some quantum instances and we denote by k the kth learning step. The model aims to construct a purification $|\psi_k\rangle\langle\psi_k|$ for the kth partition target state ρ_k of the first k qubits via training the corresponding kth scattering layer, denoted by $U_k(\boldsymbol{\theta}_k)$. The scattering layer indicates the information flow driven by $U_k(\boldsymbol{\theta}_k)$ during training, which is a parametrized circuit acting partially on the entire system. The kth scattering layer is applied on w_k qubits indexed from q_k to q_{k+w_k-1}, which preserves previous learning results from ρ_k. As a model iterative variable, w_k controls the dimensionality of the purification at the kth learning step generated from the model. For sufficiently large w_k the reachability of the purification $|\psi_k\rangle\langle\psi_k|$ is ensured so that the state $|\psi_k\rangle$ is represented in $k + w_k - 1$ qubits. The optimization on $U_k(\boldsymbol{\theta}_k)$ is accomplished by minimizing an adaptive kth step cost function as

$$C_k(\boldsymbol{\theta}_k) \coloneqq \mathcal{D}(\sigma_k(\boldsymbol{\theta}_k), \rho_k), \tag{1}$$

where σ_k represents the kth partition of $|\psi_k\rangle\langle\psi_k|$, i.e., $\sigma_k = \mathrm{Tr}_{\bar{\mathcal{A}}_k}(|\psi_k\rangle\langle\psi_k|)$, and $\mathcal{D}(\cdot)$ denotes a distance measured between two density operators. By hierarchically conducting the scattering layers, we could then construct the entire target through our QSSM. We summarize our quantum state learning algorithm via the QSSM in Algorithm 1. A flowchart of QSSM state learning is illustrated in Fig. 1.

ALGORITHM 1. QSSM for (pure) state learning.

Require: Copies of the n-qubit target state $\rho = |\phi\rangle\langle\phi|$, initialize the model to $|0\rangle^{\otimes n}$ with qubit labels $q_1, q_2, \ldots, q_n$.
Ensure: All layer parameters are randomly initialized regarding uniform distribution of $[0, 2\pi)$.
Let $k = 1$.
The step scattered layer width $w_k = k + 1$.
while $k \leqslant n$ **do**
  if $k \leqslant \lfloor n/2 \rfloor$ **then**
    $w_k = k + 1$.
  else if $k > \lfloor n/2 \rfloor$ **then**
    $w_k = n - k + 1$.
  end if
  Random initialize $U_k(\boldsymbol{\theta}_k)$ acting on qubits $q_k \sim q_{k+w_k-1}$.
  Minimize $C_k(\boldsymbol{\theta}_k)$ via classical optimization algorithm.
  $k = k + 1$.
end while
Store all optimized $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ in classical memory.
return reconstructed representation $|\psi\rangle = U_n \cdots U_1 |0\rangle^{\otimes n} \approx |\phi\rangle$.

Remark 1. QSSM state learning applies to a given mixed target state ρ acting on A by learning the purification $|AR\rangle$ of ρ, where R is the ancillary system.

B. Cost function evaluation

As a hybrid quantum-classical model, declaring the practical realization of the model is necessary. In our context, we use the square of the Schatten 2-norm to define the distance metric, or the kth step cost, as shown in

$$C_k(\boldsymbol{\theta}_k) = \mathrm{Tr}\{[\sigma_k(\boldsymbol{\theta}_k) - \rho_k][\sigma_k(\boldsymbol{\theta}_k) - \rho_k]^\dagger\}. \tag{2}$$

We have chosen the cost function of the form (2) due to its convenience and efficiency of evaluations, which could possibly be performed on near-term quantum hardware. We rearrange Eq. (2) as

$$C_k(\boldsymbol{\theta}_k) = \mathrm{Tr}\left[\sigma_k^2(\boldsymbol{\theta}_k)\right] + \mathrm{Tr}\left(\rho_k^2\right) - 2\,\mathrm{Tr}[\sigma_k(\boldsymbol{\theta}_k)\rho_k]. \tag{3}$$

The terms involving state overlaps can be evaluated via a SWAP test [36], which has been experimentally demonstrated on real quantum devices [37,38]. In addition, in the case of exponentially vanishing overlap between states which fails the SWAP test, the estimation of the overlap can rely on the strategy with collective measurement [39]. The training of the kth layer can be described as finding the kth step optimal parameters $\boldsymbol{\theta}_k^{\mathrm{opt}}$ so that $C_k(\boldsymbol{\theta}_k^{\mathrm{opt}})$ is minimized to approximately zero.

C. Analytic gradient

We now show that the analytical gradients of the cost function in Eq. (2) can also be computed efficiently, making the gradient-based scheme a prospective candidate for the training processes. The analytic gradient of C_k can be evaluated according to [21,40–42]. Suppose the kth layer U_k consists of the gates satisfying the parameter-shift rule [40,41] and

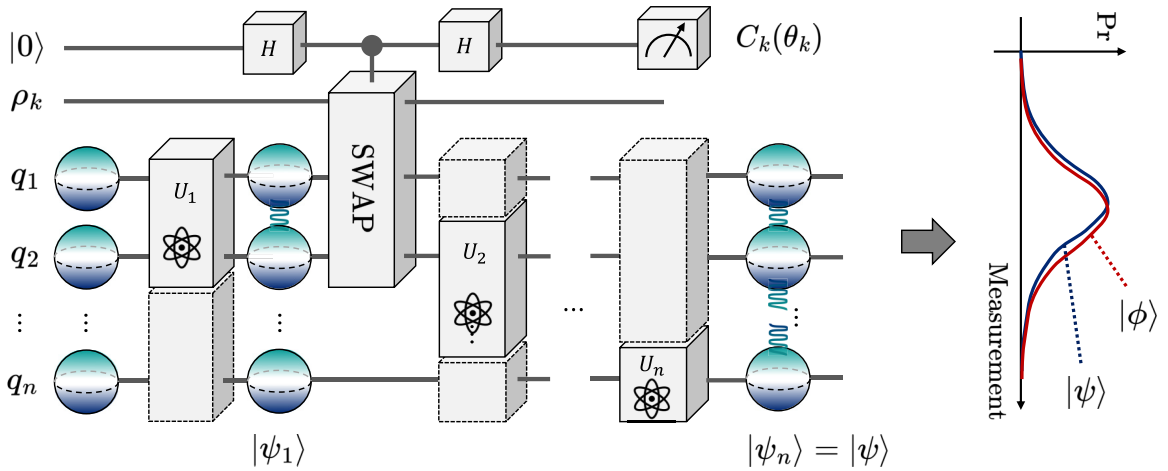


FIG. 1. Conceptual diagram of QSSM state learning. Starting with a full tensor product state (e.g., $|0\rangle^{\otimes n}$) initially, each QSSM layer U_k produces a purification $|\psi_k\rangle$ of the reduced density ρ_k of $|\phi\rangle$. At each step, the cost $C_k(\boldsymbol{\theta}_k)$ can be estimated via the SWAP test [36] shown in the diagram. After all n training steps, the entire trained model produces a complete circuit representation $|\psi\rangle$ approaching the target $|\phi\rangle$. The state $|\psi\rangle$ therefore carries almost the same stochastic behavior as $|\phi\rangle$ and can be regenerated conveniently for further computational assignments.

contains m trainable parameters. Each optimization iteration is driven by the estimations of the cost gradient given by

$$\nabla_{\boldsymbol{\theta}_k} C_k(\boldsymbol{\theta}_k) = (\partial_1 C_k(\boldsymbol{\theta}_k), \ldots, \partial_m C_k(\boldsymbol{\theta}_k)), \tag{4}$$

where $\partial_\mu := \frac{\partial}{\partial \theta_k^\mu}$, indicating the partial derivative with respect to a fixed θ_k^μ in the kth layer. In particular, we derive the analytic gradient of C_k as follows:

$$\partial_\mu C_k^* = \langle G_k^* \rangle_{\left(\theta_k^\mu\right)^* + \pi/2} - \langle G_k^* \rangle_{\left(\theta_k^\mu\right)^* - \pi/2}. \tag{5}$$

The asterisk indicates the corresponding quantity evaluated at $\boldsymbol{\theta}_k = \boldsymbol{\theta}_k^*$. Here G_k is a Hermitian operator involving both σ_k and ρ_k and is expressed as

$$G_k(\boldsymbol{\theta}_k) := \Delta_k(\boldsymbol{\theta}_k) \otimes \Gamma_k, \tag{6}$$

where $\Delta_k(\boldsymbol{\theta}_k) = \sigma_k(\boldsymbol{\theta}_k) - \rho_k$ represents the kth step state difference and Γ_k is the maximally mixed state I/d, with I the identity operator of dimension $d = 2^{w_k - 1}$; $\Gamma_k = 1$ when $w_k = 1$. The bra-ket operation is in the analytic form $\langle A \rangle_\alpha = \langle \psi_{k-1} | U_k^\dagger(\boldsymbol{\theta}_k) A U_k(\boldsymbol{\theta}_k) | \psi_{k-1} \rangle$ for some Hermitian operator A evaluated at $\theta_k^\mu = \alpha$. This quantity of G_k in Eq. (5) indicates the expectation value of G_k regarding the kth step variational ansatz $|\psi_k\rangle$ evaluated at $(\theta_k^\mu)^* \pm \pi/2$, where all other scattering layers remain unchanged. The detailed derivation of these definitions and forms can be found in Appendix C.

To summarize, each partial derivative of C_k at $\boldsymbol{\theta}_k^*$ can be explicitly determined by Eq. (5), which can be efficiently computed on NISQ devices via shifting the corresponding parameter. The gradient-based optimization could be applied to the cost by specifically updating the parameters $\boldsymbol{\theta}_k$ in the kth layer as

$$\boldsymbol{\theta}_k \leftarrow \boldsymbol{\theta}_k^* - \eta \nabla_{\boldsymbol{\theta}_k} C_k(\boldsymbol{\theta}_k^*), \tag{7}$$

where η is the learning rate determined for the classical optimizers defining the iteration step size. Apart from the plain gradient decent, classical gradient-based and gradient-free methods, such as Adam and COBYLA [43,44], can be used during optimizations. By repeating the step (7), the cost function will possibly converge to the optimal minimum. We then iterate the above procedures for each kth layer to complete the model training with a final output circuit representation $\boldsymbol{U}(\boldsymbol{\theta}^{\text{opt}}) = U_n(\boldsymbol{\theta}_n^{\text{opt}}) \cdots U_1(\boldsymbol{\theta}_1^{\text{opt}})$ for generating the target state.

III. THEORETICAL PERFORMANCE ANALYSIS

In this section we elaborate on the theoretical effectiveness of QSSM state learning for arbitrary quantum states. Then we explore the model's trainability, which strongly corresponds to the scattering layers' maximum width.

A. Effectiveness of the QSSM

One of the implications of Uhlmann's theorem is that it ensures the degrees of freedom for quantum state purification [35]. Given a mixed state ρ with a purification $|AR\rangle$, one can always find a local unitary U_R acting on the ancillary system R such that $|AR'\rangle = (I_A \otimes U_R)|AR\rangle$ forms another purification of ρ. Based on that, the theoretical guarantee of QSSM state learning is stated in the following lemma.

Lemma 1. Given a target state ρ acting on the system of A and B, suppose its purification $|\psi\rangle$ is on system ABE, where E is an ancillary system, such that

$$\text{Tr}_{BE}(|\psi\rangle\langle\psi|) = \text{Tr}_B(\rho). \tag{8}$$

There always exists a local unitary U_{BE} acting on the composite system BE such that

$$\text{Tr}_E[(I_A \otimes U_{BE})|\psi\rangle\langle\psi|(I_A \otimes U_{BE}^\dagger)] = \rho. \tag{9}$$

The proofs and details of Lemma 1 can be found in Appendix D. We now apply Lemma 1 to our QSSM in order to demonstrate the effectiveness of perfectly learning a target pure state. Consider an n-qubit pure target $\rho = |\phi\rangle\langle\phi|$ where we wish to reconstruct it using an n-qubit circuit representation. Apart from these n qubits and the ancillary systems for cost and gradient evaluations, there is no need for additional ancillary qubits for the entire learning process. At the kth step, we suppose that the purification $|\psi_k\rangle$ has been perfectly learned and represented on the first $k + w_k - 1$ qubits so that

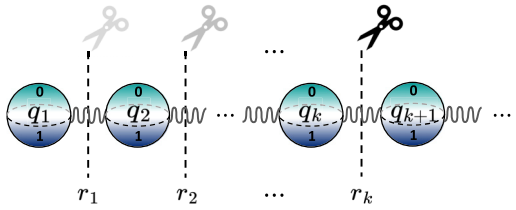


FIG. 2. The kth partition of an ordered qubit system. The corresponding element r_k is the rank of the reduced state $\rho_k = \mathrm{Tr}_{\bar{\mathcal{A}}_k}(\rho)$.

the following equation holds:

$$\sigma_k = \mathrm{Tr}_{\bar{\mathcal{A}}_k}(|\psi_k\rangle\langle\psi_k|) = \mathrm{Tr}_{\bar{\mathcal{A}}_k}(\rho) = \rho_k. \tag{10}$$

We call this the kth perfect learning condition of QSSM state learning. Then, by Lemma 1 there exists a local unitary such that

$$\mathrm{Tr}_{\bar{\mathcal{A}}_{k+1}}[(I_k \otimes U_{k+1})|\psi_k\rangle\langle\psi_k|(I_k \otimes U_{k+1}^\dagger)] = \mathrm{Tr}_{\bar{\mathcal{A}}_{k+1}}(\rho), \tag{11}$$

where the existence of U_{k+1} ensures the effectiveness of the QSSM. We call it a perfect learning assumption of QSSM state learning if all the kth perfect learning can be achieved.

One important point to note here is that when the rank of ρ_k for $1 \leqslant k < n$ in the target is bounded from above, an upper bound for each layer width w_k can be determined while maintaining perfect learning. We introduce the definition of rank sequence of any quantum state characterizing the Schmidt rank distributions within the state.

Definition 1. Given an n-qubit pure quantum state ρ represented by certainly ordered qubits, the (Schmidt) rank sequence is an ordered list $\mathcal{R}_\rho$,

$$\mathcal{R}_\rho = \{r_1, r_2, \ldots, r_{n-1}, r_n\}, \tag{12}$$

where r_k indicates the Schmidt rank of the kth partition reduced states ρ_k in Fig. 2 and $r_n = 1$.

We then have the following sufficient and necessary conditions for the perfect learning of the QSSM,

Proposition 1. For a given n-qubit pure target state ρ represented by certainly ordered qubits, if the rank sequence is

$$\mathcal{R}_\rho = \{r_1, r_2, \ldots, r_{n-1}, r_n\},$$

then the QSSM can achieve perfect learning for ρ if and only if the minimum width of the kth scattering layer is $\lceil \log_2 r_k \rceil + 1$.

The proof of Proposition 1 directly follows by Lemma 1 and Schmidt decomposition. The proposition identifies a group of quantum states that can be learned more efficiently using the QSSM. One notable exemplar within this proposition is the n-qubit Greenberger-Horne-Zeilinger (GHZ) state.

Remark 2. An n-qubit GHZ state [45] has constant rank $r_k = 2$ for $1 \leqslant k < n$. Hence setting $w_k = 2\,\forall\, k$ is sufficient to obtain perfect learning of QSSM state learning on the GHZ state.

Direct intuition from the above phenomenon suggests a connection between the amount of entanglement within a target state and the sufficient widths w_k to achieve perfect learning. The higher the rank, the harder the target state that could be learned via the QSSM.

B. Avoiding barren plateaus

Trainability is a critical challenge for the practical usage of QNNs. Using a deep QNN for a global system significantly increases the randomness of an initial guess, which nevertheless leads to the emergence of the barren plateau (BP) issue [26] for sizable state learning. In this case, the partial derivative of the cost function would have a zero mean and an exponentially small variance with respect to the number of qubits, thereby making it challenging to identify the correct direction to decrease the cost function value.

As shown in previous sections, the QSSM has illustrated a potential capability of addressing trainability issues because of its nature of focusing on subsystems instead of the whole state. In this section we show that the QSSM has explicit advantages in trainability by directly computing the variance of the cost gradient, thus enabling it to avoid a BP in many cases. For the gradient of the kth step $\partial_\mu C_k$ [see Eq. (5)], the kth scattering layer can be expressed as $U_k(\boldsymbol{\theta}) = U_+^{(k)}(\boldsymbol{\theta}_+)e^{-i\theta_\mu H_\mu}U_-^{(k)}(\boldsymbol{\theta}_-)$, where the total parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_+, \theta_\mu, \boldsymbol{\theta}_-)$. For simplicity, we have omitted the superscript k in the parameter. The results are summarized as follows.

Definition 2. A unitary t-design of dimension d [46] with respect to the Haar measure is defined as a finite set of unitaries $\{U_k\}_{k=1}^M$ on a d-dimensional Hilbert space such that

$$\frac{1}{M}\sum_{k=1}^{M} P_{(t,t)}(U_k) = \int_{\mathcal{U}(d)} d\mu_{\mathrm{Haar}}(U) P_{(t,t)}(U), \tag{13}$$

where $P_{(t,t)}(U)$ denotes a homogeneous polynomial of degree at most t on the elements of U and $U^\dagger$.

Proposition 2. For QSSM learning of an n-qubit target state, ρ has a fixed-order representation and a rank sequence

$$\mathcal{R}_\rho = \{r_1, r_2, \ldots, r_{n-1}, r_n\}.$$

If $U_\pm^{(k)}$ of the kth step form at least a local unitary 4-design, the expectation and the variance of the analytic gradient for the kth learning step with respect to θ_μ are evaluated as

$$\mathbb{E}(\partial_\mu C_k) = 0, \quad \mathrm{Var}(\partial_\mu C_k) \in O\left(\frac{1}{r_k}\right). \tag{14}$$

The proof of this proposition is presented in Appendix E. This proposition notably implies that the gradient magnitude is greatly determined by the largest Schmidt rank in the rank sequence of the target state rather than the total number of qubits n. In other words, the gradient magnitude scales with the width of each scattering layer as $O(2^{-w_k})$ and the QSSM could escape from the barren plateaus by carefully setting the layer widths. A typical example is given in the following remark.

Remark 3. The variance of the gradient magnitude for learning an n-qubit GHZ state scales as

$$\mathrm{Var}(\partial_\mu C_k) \in O(1) \tag{15}$$

for any integer $1 \leqslant k \leqslant n$, which means there is no barren plateau.

Proposition 2 implies that the QSSM can efficiently facilitate the learning of any pure states with $r_k \sim O(\mathrm{Poly}(n))$. This encompasses a broad spectrum of quantum states, including

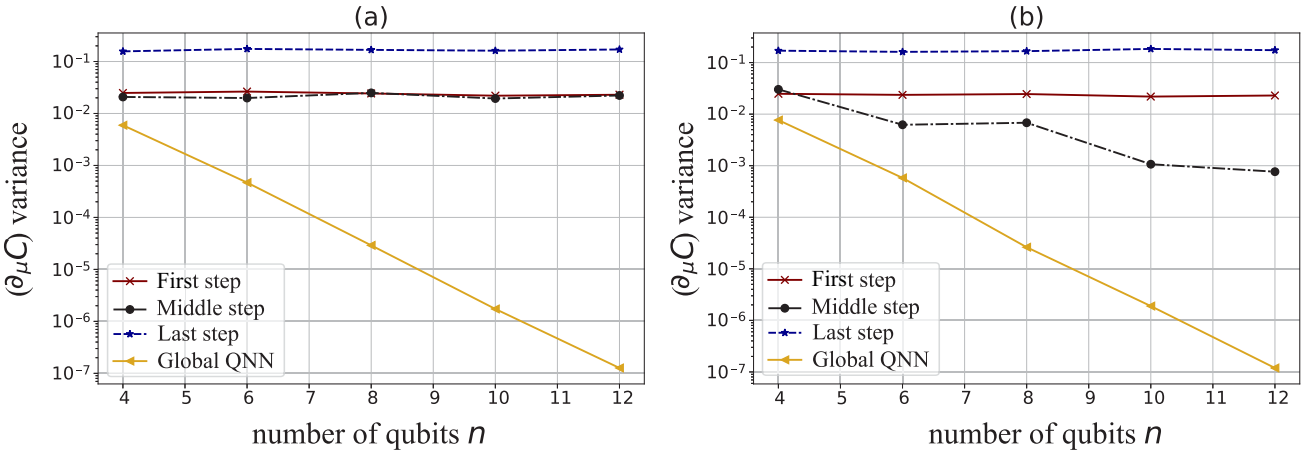


FIG. 3. Comparison of the gradient magnitude between different steps in the QSSM and global QNN. The data points represent the sample variance of gradients of the cost function as a function of the number of qubits on a semilogarithmic plot for the learning of (a) the GHZ state and (b) the ground state of the Heisenberg model. The red solid, black dash-dotted, and blue dashed lines represent the gradient magnitude of the first step, the $\frac{n}{2}$th step, and the last step, respectively. The yellow diagonal line represents the gradient magnitude of using the randomly initialized global QNN. Our method apparently outperforms the conventional global QNN in terms of gradient variance scaling, indicating the absence of barren plateaus.

cases like slightly entangled states [7] and matrix product states [47], which can be classically efficiently characterized and simulated [48]. However, while the former necessitates a maximal Schmidt rank over all possible bipartite splits of polynomial scaling, our QSSM applies more than those merely requiring a polynomial scaling of the Schmidt rank over bipartite splits from all possible permutations of those representing qubits. Moreover, although our QSSM illustrates similar favor of tensor network techniques, in contrast to existing methods [49–51] tailored for learning matrix product states, primarily applicable to one-dimensional quantum spin model ground states, the QSSM has broader applications and could work more generally for learning different quantum states. Therefore, our QSSM exhibits broader utility and potential applicability across a wider spectrum of quantum states compared to these methods. Even in the worst case, wherein the rank sequence of the target state is $\mathcal{R}_\rho = \{2^1, 2^2, \ldots, 2^{\lfloor n/2 \rfloor}, \ldots, 2, 1\}$, the gradient magnitude scales as $O(2^{-\lfloor n/2 \rfloor})$. Our QSSM state learning still gains a square root advantage compared with the conventional global QNN having a gradient variance scaling as $O(2^{-n})$.

From another point of view, random pure states ρ, despite having maximum rank in $\mathcal{R}_\rho$, scale as $2^{\lfloor n/2 \rfloor}$ and can be useless as practical computational resources [52]. In addition, the few smallest eigenvalues of the middle partition $\rho_{\lfloor n/2 \rfloor}$ can be negligible. Learning their low-rank approximation predetermined by the quantum principal component analysis [53] can be treated as a quantum compressing of unknown states, which still captures the main statistical behaviors of target states. By allowing a certain error tolerance of learning ρ instead of perfect learning, one can omit the influence of those tail eigenvalues and the efficient training condition of the QSSM still applies by fixing the maximum layer width.

Compared to the n-qubit universal QNN model state learning, the QSSM demands significantly fewer parametric degrees of freedom (DOFs) to reach the same approximating error. The generating Lie algebra of an n-qubit universal QNN model has to span $\mathrm{SU}(2^n)$, resulting in a model DOF of $O(4^n)$. In contrast, since the kth scattering layer involves at most $\lfloor \frac{n}{2} \rfloor + 1$ quantum registers, the total DOF of the QSSM experiences a quadratic reduction to at most $O(4^{\lfloor n/2 \rfloor})$. Also, to learn the polynomial rank-bounded target state ρ, $r_{\max} = \max \mathcal{R}_\rho \sim O(\mathrm{Poly}(n))$. The DOF required for each scattering layer in the QSSM scales as $O(\mathrm{Poly}(n))$. Therefore, the entire model comprises fewer quantum gates, rendering this approach considerably more hardware efficient.

To illustrate the result shown in Proposition 2, we compare the gradient variances of cost (2) as a function of the number of qubits for the QSSM and global QNN state learning. For the QSSM, we particularly investigate the values in the first step, the middle step ($\frac{n}{2}$th step), and the last step of the learning procedure. We look into a single-parameter R_Z gate in the middle of a circuit forming a local 4-design. The two parts split by the gate are also local 4-designs and are represented as Haar random unitaries in our experiment. For the global QNN, we calculate the gradient variances of a single-parameter R_Z gate which is sandwiched between two global Haar unitaries. We learn the GHZ state and ground state of the Heisenberg model [54] with maximum width $w_{\max}$ being 2 and 4, respectively. The variance values are computed from sampling 500 Haar unitary pairs for both cases.

As can be seen in Fig. 3, the variance of the gradient vanishes exponentially with the number of qubits when using the randomly initialized global QNNs. In contrast, the QSSM demonstrates a constant scaling of variance magnitude. We note that there is a decay of the gradient variance of the middle step in Fig. 3(b). Nevertheless, this decay is caused by a constant factor that originates from the nature of the ground state (GS). It does not exhibit the exponentially decreasing behavior appearing in the global QNN cases.

It is worth noting that the recent work by Cerezo *et al.* [27] claims that strategies aimed at avoiding barren plateaus will result in a polynomially sized subspace constraining the evolved observable, rendering them classically simulable.

TABLE I. Effectiveness (noise-free) validation of the QSSM in learning diverse 12-qubit quantum states regarding the final-state fidelities. The QSSM consistently surpasses the conventional global QNN in terms of achieving higher final-state fidelity across a variety of quantum states, showing its remarkable efficiency in tackling quantum state learning tasks.

Physical states	Global QNN	QSSM
XXX model GS	0.533	0.926
XXZ model GS	0.523	0.948
LiH molecular GS	0.531	0.973
Algorithmic states	**Global QNN**	**QSSM**
GHZ state	0.535	0.971
W state	0.527	0.958
Gaussian distribution encoding	0.561	0.978
MNIST data encoding	0.330	0.867
random state	0.317	0.834

However, we note that the enhanced trainability of our algorithm primarily stems from its layerwise learning scheme, which can be regarded as applying a warm start for each learning step. In addition, the loss function employed in the QSSM diverges from the assumption considered in that work [27]. Therefore, the QSSM does not align with the classically simulable algorithms in the same manner as the strategies considered in [27]. To the best of our knowledge, it remains uncertain if the QSSM is classically simulable. We believe it would be a valuable research direction to explore in the future.

IV. LEARNING PHYSICAL AND ALGORITHMIC QUANTUM STATES

To showcase the effectiveness of the QSSM on state learning tasks, we conduct numerical simulations of learning both physical and algorithmic 12-qubit quantum states using the QSSM and the traditional global QNN, shown in Table I. The XXX and XXZ models represent the Heisenberg spin-1/2 chains ($J_x = J_y = J_z = 1$, and $J_x = J_y = 1$ and $J_z = 2$, respectively) with zero external magnetic fields, satisfying periodic boundary conditions. The ground state of the LiH molecule is provided by OpenFermion tools. For the Gaussian and MNIST experiments, the distribution and image data are encoded to the unit quantum state vectors of dimension 2^n via amplitude encoding [55] with automatic padding of 0's filling out the extra grayscale pixels. A consistent experimental configuration is employed across all the following simulations.

In our numerical simulations involving the global QNN and the QSSM, we employ a general hardware-efficient ansatz [23] of depth $d = 20$ with random initialized parameters. This choice of a random circuit can be considered as an approximate 2-design, thus providing robust expressibility. The optimization uses the Adam optimizer with a learning rate of 0.1, spanning 200 iterations. Comparing the outcomes with those of the global QNN, we discern clear advantages exhibited by the QSSM, which consistently attains notably high fidelity in the learning of diverse quantum states. Conversely, conventional methods do not perform well, primarily due to the challenges encountered during the training process when dealing with a large number of qubits, where the convergence speed significantly decreases.

More specifically, we present the trend line depicting the outcome performance in terms of fidelity between the network-produced results and the target states, as showcased in Fig. 4. This figure offers a comparative view of the performance of the two models as learning the ground states of the Heisenberg spin-1/2 chain model ($J_{x,y,z} = 1$ and $h_z = 0$) across an increasing number of qubits. Each data point is an average from five independent numerical experiments, and the error bars indicate the range of trained outcome fidelities. The performance of the global QNN, represented by the red curve, exhibits a marked decline when dealing with higher-dimensional states. On the other hand, the QSSM effectively sustains its convergence speed under the same computational resource allocation, as indicated by the slight decrease in the blue curve.

We also perform noisy quantum simulations in which the QSSM is used to learn a four-qubit GHZ state on the IBMQ Qiskit simulator [56]. We build our noise model from single-qubit and multiqubit depolarizing channels (DCs) and thermal relaxation channels (TRCs) [57]. The error rate of the DCs is set to 10^{-3} and the T_1, T_2, and gate time of the TRCs are set to 1000 μs, 100 μs, and 1 ns, respectively. In Fig. 5(a) we show the behavior of the cost function in every step of the QSSM's optimization. At each step, we run the optimization of the QSSM circuit 20 times and use the parameters that correspond to the lowest cost to update the circuit before going to the next step. We find this trick can significantly alleviate the randomness arising from sampling of bit strings in the measurement of quantum circuits and therefore improves the stability of our state learning task. We also show the distribution of bit strings generated from the learned QSSM circuit after the measurement in Fig. 5(b). The fidelity between the quantum state generated from the QSSM and the GHZ state could reach 91%. The figure validates the efficacy and efficiency of the QSSM even in the presence of noisy environments, consequently reinforcing the practical applicability and prospects of our method.

From the analytical description and numerical demonstration, we see that the QSSM has the ability to learn arbitrary quantum states with high fidelity compared to the conventional variational methods. It is worth noting that, based on the result shown in Proposition 1, the QSSM will only require narrow and shallow circuits in learning quantum states that are weakly entangled in each partition, thus being extremely efficient in learning such a class of quantum states.

A. Truncation performance

The original QSSM assumes perfect learning for arbitrary pure quantum states, necessitating $w_k = k + 1$ or $n - k + 1$ to cover any state ρ of $r_k = 2^{\min\{k,n-k\}}$ in $\mathcal{R}_\rho$, based on Proposition 1. However, as highlighted in Remark 2 for the GHZ state, the flexibility of confining the maximum width $w_{\max}$ of the scattering layers can inspire a truncated version of the QSSM (TQSSM). With prior knowledge of the rank sequence of the target state, we could use narrower layers to accomplish the kth learning step.

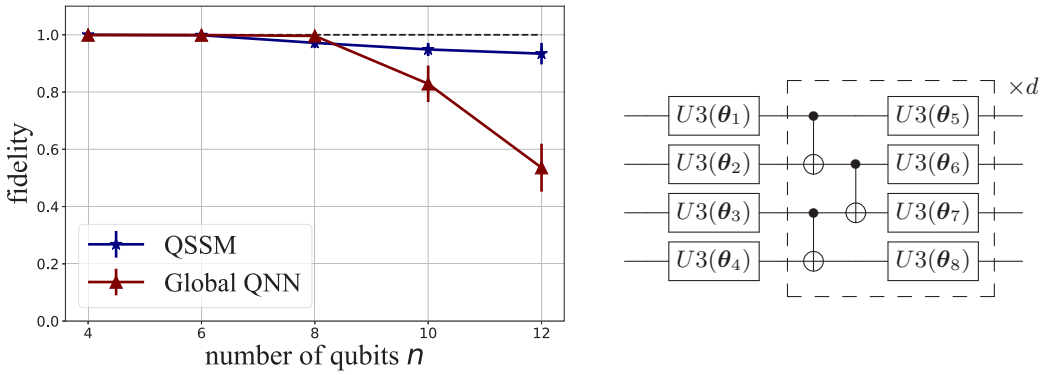


FIG. 4. Comparison of effectiveness (noise-free) trends between the QSSM and global QNN in learning the ground state of the periodic Heisenberg model of different system sizes. The blue and red curves illustrate the final learned fidelity from the QSSM and the global QNN model, respectively. On the right we show the architecture of the circuit model used in the experiments. The circuit is composed of the repeated block consisting of CNOT gates and $U3$ gates. The dashed block circuit repeats d times.

In addition, a state with exponentially large ranks in the rank sequence might not necessarily be hard to learn. Those with concentrated eigenvalues contain large ranks but can be learned up to a high fidelity [59] with limited resources. Intuitively, only highly entangled states have insignificant tail eigenvalues, which occupy a limited region throughout the entire Hilbert space. In the worst case, learning a state that is maximally entangled [60] is the most challenging task for the QSSM since every kth partition reduced state of the maximally entangled states has the maximum attainable rank r_k with eigenvalues uniformly spread.

In most cases, we could omit the insignificant impactful tail Schmidt coefficients of the target state and concentrate on the dominant terms. In essence, a truncated version of the target state can be learned, encapsulating the majority of stochastic characteristics yet requiring fewer computational resources. As a result, during practical QSSM implementations, it becomes viable to truncate the maximum layer width $w_{\max}$, thereby enhancing efficiency without significantly compromising the quality of state learning.

We determine a width constraint to the QSSM so that $w_k = \min\{k+1, w_{\max}, n-k+1\}$. To clarify the performance of the truncated QSSM, we conduct numerical simulations on the model, as before, by setting different values of $w_{\max}$ and learning the same quantum states used in Table I. All the other hyperparameters stay the same. Results are shown in Table II.

In most situations, reasonably constrained widths for scattering layers would counterintuitively yield superior performance compared to the original strategy used in the QSSM. As evident from Table II, it is apparent that attaining an acceptable fidelity level (approximately 0.9) for learning these quantum states only necessitates $w_{\max} = 4$ or 5. Larger values of $w_{\max}$ could even lead to a decrease in the model performances of state learning. For instance, in the case of learning the GHZ state, opting for $w_{\max} = 2$ yields an optimal fidelity of 0.994, whereas a value of only 0.971 is achieved with

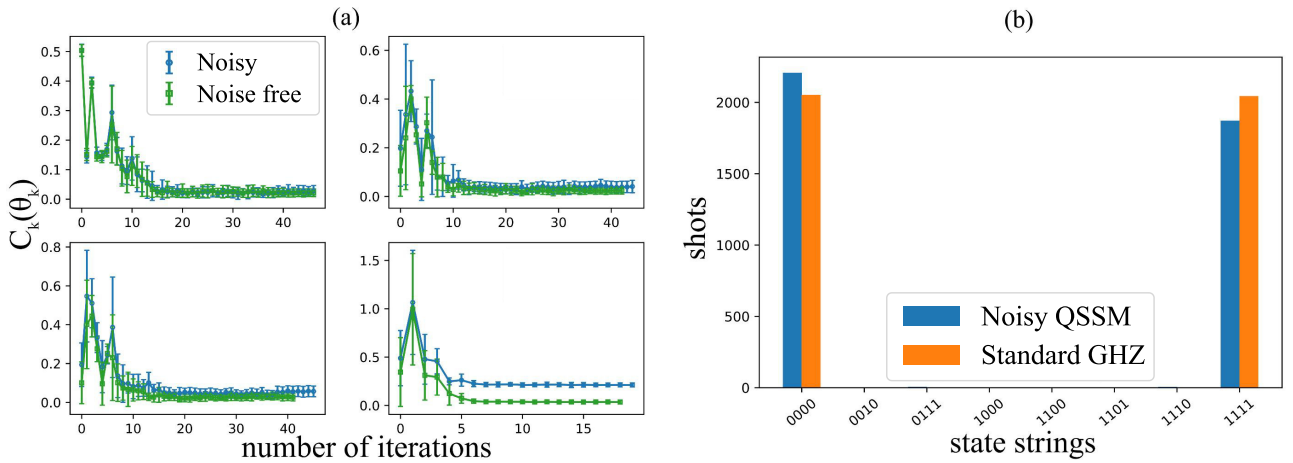


FIG. 5. Noisy quantum simulation of the QSSM for learning a four-qubit GHZ state. (a) Comparison of the variation of the cost function with the noisy quantum simulation and noise-free simulation. For both cases, the optimization was processed via the COBYLA optimizer [58] on SWAP-test estimated cost values. (b) Distribution of measurement outcomes which are generated noise-free from the state that was obtained by the noisy trained QSSM. The blue (left) bar shows the result of the noisy simulation and the orange (right) bar shows the result of the standard GHZ state.

TABLE II. Flexibility of TQSSM (noise-free) numerical simulations on learning both physical and algorithmic quantum states with increasing maximum layer widths $w_{\max}$. All layers with different $w_{\max}$ have the same ansatz setup as in Table I. For $n = 12$ states, $w_{\max} = 7$ is equivalent to the worst case of doubling the dimensionality.

	Maximum layer widths					
Physical states	2	3	4	5	6	7
XXX model GS	0.523	0.883	0.915	0.956	0.950	0.926
XXZ model GS	0.750	0.887	0.954	0.952	0.952	0.948
LiH molecule GS	0.978	0.973	0.967	0.978	0.982	0.973
Algorithmic states	2	3	4	5	6	7
GHZ state	0.994	0.993	0.990	0.989	0.975	0.971
W state	0.990	0.992	0.982	0.985	0.982	0.958
Gaussian distribution	0.969	0.985	0.976	0.981	0.986	0.978
MNIST data encoding	0.517	0.759	0.891	0.903	0.887	0.867
random state	0.318	0.768	0.856	0.871	0.879	0.834

$w_{\max} = 7$. A plausible explanation for this phenomenon could be the overparametrization and the mild BP effect during the training of the half-dimensional scattering layers. Notably, learning random state undoubtedly obtains the worst learning results.

V. DISCUSSION

We have proposed a QSSM state learning framework combining purification theory and QNN training. The main feature of our QSSM is that it avoids barren plateaus in learning a large class of quantum states containing a medium amount of entanglement. By theoretical and numerical demonstrations, we have shown that the QSSM can outperform the conventional QNN architecture on quantum state learning tasks with higher fidelity and convergence speed. Our model only requires QNN layers crossing adjacent systems and fewer network parameters, showing convenient topological connectivity and robustness on circuit noise from simulations. From the perspective of trainability, we briefly introduced the issue of barren plateaus in training QNN-based algorithms and reviewed some recent solutions and schemes resolving it. In particular, we rigorously proved that the QSSM would avoid barren plateaus for learning a large class of quantum states. We also performed numerical sampling to compute the gradient variances with respect to some fixed network parameters. Our QSSM shows constant and polynomial scaling of gradient variance for those states, which partially solves the challenging trainability issues in QML.

While barren plateaus have received considerable attention in the literature, recent investigations have revealed a broader challenge posed by the concentration of local minima [28,29] near the global minimum due to the nonconvex nature of the cost landscape. This phenomenon can impede training progress, leading to unreliable results. Given that our model, the QSSM, adopts a layerwise learning strategy wherein the outcome of each step depends on the preceding one, it is susceptible to this issue. However, some promising methodologies [61,62] have emerged to mitigate such challenges, demonstrating efficiency in escaping local minima. Therefore, by integrating these techniques, the QSSM can potentially overcome these hurdles and achieve better performance. Another challenge for the sequential training procedure is the phenomenon known as abrupt phase transition [31,32], which could also be problematic in the training of our model. This phenomenon means that the loss value cannot be improved by adding subsequent layers and the trained layers will be close to an identity, thus rendering the layerwise training strategy ineffective. However, it is essential to distinguish the QSSM from the standard layerwise training methods, wherein a new layer is trained at each step while maintaining the same loss function. In the QSSM, the training at each step can be regarded as an independent global optimization task, with a distinct loss function tailored to the different target states. Consequently, the QSSM circumvents the limitations associated with traditional layerwise training approaches.

There are other remaining issues of the QSSM for future discussion. Different choices of scattering layers would influence the learning performance, which has to be exemplified. How to further improve the state fidelity provided the high-fidelity state from the QSSM could become a significant open question. Understanding and resolving the effect of overparametrization from the QSSM should be explained. A theoretical performance guarantee and the connection between scattering layer dilation and information flow of QSSM state learning should be established for a complete story of truncated state learning. Finally, we also expect some extended applications of the QSSM, for instance, learning special probability distributions that have been encoded as quantum states instead of only state learning on near-term quantum devices.

ACKNOWLEDGMENTS

Part of this work was done when M.J. and G.L. were at Baidu Research. The authors would like to thank Benchi Zhao for valuable comments. The authors would also like to thank the anonymous reviewers for their valuable suggestions to improve the manuscript. M.J. and X.W. were partly supported by the Start-up Fund (Grant No. G0101000151) from The Hong Kong University of Science and Technology (Guangzhou), the Guangdong Provincial Quantum Science Strategic Initiative (Grant No. GDZX2303007), and the Education Bureau of Guangzhou Municipality.

APPENDIX A: FOUNDATION OF QUANTUM COMPUTING

We first briefly introduce some basic concepts of quantum computation necessary for a self-contained reading of the paper. Our notation follows the conventional textbook by Nielsen and Chuang [35]. Quantum information is encoded and processed via the fundamental cells, namely, qubits, and described by quantum states. An n-qubit state can be mathematically represented by a $2^n \times 2^n$ positive-semidefinite density matrix ρ over the complex field where $\rho \succeq 0$ and $\mathrm{Tr}(\rho) = 1$. A pure state, in this formulation, satisfies $\mathrm{rank}\,(\rho) = 1$ and can be expressed in Dirac bra-ket notation as $\rho = |\psi\rangle\langle\psi|$, where $|\psi\rangle \in \mathbb{C}^{2^n}$ denotes a Hilbert space unit column vector with the corresponding dual vector $\langle\psi|^\dagger = |\psi\rangle$, with the dagger denoting the complex conjugate transpose operation. In general, we also use $|\psi\rangle$ to represent a pure state. A mixed state satisfies $\mathrm{rank}\,(\rho) > 1$, and based on the spectral theorem, it has a decomposition form $\rho = \sum_j p_j|\psi_j\rangle\langle\psi_j|$, where $p_j > 0$ denotes the probability of observing $|\psi_j\rangle\langle\psi_j|$ in ρ and $\sum_j p_j = 1$. Based on Uhlmann's theorem [35], for every mixed state ρ acting as a linear operator on a Hilbert space A there exists a purified state $|AR\rangle$ (i.e., pure state) in the composite system AR such that $\mathrm{Tr}_R(|AR\rangle\langle AR|) = \rho$, where $\mathrm{Tr}_R(\cdot)$ denotes the partial trace operation tracing out the ancillary system R. The purification $|AR\rangle$ has a Schmidt decomposition form $|AR\rangle = \sum_j \sqrt{p_j}|\psi_j\rangle \otimes |j_R\rangle$ for some orthonormal set $|j_R\rangle$ in R.

The evolution of a quantum state ρ is realized by applying a series of quantum gates, which are mathematically described as unitary operators. The state ρ' that undergoes transformation via a quantum gate U can be obtained through direct matrix multiplication, expressed as $\rho' = U\rho U^\dagger$. Common single-qubit gates include the Pauli rotations $\{R_P(\theta) = e^{-i(\theta/2)P}|P \in \{X, Y, Z\}\}$, which are in the matrix exponential form of Pauli matrices

$$X \coloneqq \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad Y \coloneqq \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad Z \coloneqq \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{A1}$$

In QML, QNNs are usually represented as parametrized unitaries consisting of a bunch of single-qubit gates and several two-qubit gates, including a controlled-X [or controlled-NOT (CNOT)] gate equal to $I \oplus X$ and a controlled-Z (CZ) gate equal to $I \oplus Z$, where $\oplus$ denotes the direct sum operation. An n-qubit operator generally exists in the linear operator space $\mathcal{L}(\mathbb{C}^{2^n})$ over the complex field. Quantum measurements are then applied at the end of the quantum circuits, which extract classical information by projecting the quantum states onto its classical shadow.

APPENDIX B: LITERATURE ON QUANTUM NEURAL NETWORKS

In quantum machine learning, QNNs are usually represented as parametrized unitaries consisting of a bunch of single-qubit rotation gates and several two-qubit gates, denoted by $\mathbf{U}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the trainable parameters. The model is trained using a classical optimizer according to a minimization process on some cost function $C(\boldsymbol{\theta})$ based on the quantum measurement results.

Quantum neural networks can be used to handle a variety of computational tasks, which is usually seen as a quantum version of classical neural networks. In the most general form, a QNN model can be expressed as $\mathbf{U}(\boldsymbol{\theta}) = \prod_{k=1}^{M} U_k(\boldsymbol{\theta}_k)$ for some subnetwork layers $U_k(\boldsymbol{\theta}_k)$, where each layer can also be seen as a combination of parametrized circuits as $U_k(\boldsymbol{\theta}_k) = \prod_{j=1}^{d} U_j(\theta_j^{(k)})W_j$, where $U_j(\theta_j^{(k)}) = e^{-ig_j\theta_j^{(k)}}$ is a parametrized gate with a Hermitian generator g_j. The W_j is usually non-parametrized, like the networks of CNOT and CZ gates. The product $\prod_k$ here is, by default, in increasing order from right to left in the above representations.

The idea of quantum neural networks has obtained massive attention since its introduction [63]. Various QNN architectures have been introduced to address a diverse range of computational challenges, spanning both classical and quantum problem domains [64–68], thereby pioneering an entirely novel realm of machine learning models. Recent literature focusing on the trainability theory of QNNs indicates a prospective direction for coping with barren plateaus by reducing the expressibility of QNN architectures [69,70]. Beyond that, some strategies have been proposed under certain conditions, for example, adopting a clever initialization strategy [71,72], using adaptive algorithms [34,73–75], making a parametrization generalization [76,77], and choosing different cost forms and circuit architectures [59,69,78].

APPENDIX C: ANALYTIC EVALUATION OF THE COST FUNCTION AND GRADIENT

In this Appendix we provide a detailed analysis of the analytic gradient of our cost function C_k (2). We take the 2-norm squared cost function as our objective. At the kth learning step, analyzing the exact form of $\partial_\mu C_k$ is necessary for further designing the training strategy of the QSSM. Recalling the expression of C_k, we could derive the derivative form with respect to the parameter $\theta_\mu = \theta_k^\mu$. From here we concentrate on the kth step and for convenience we will omit the subscript k of the parameter in the following. The partial derivative of C_k with respect to θ_μ is then expressed as

$$\partial_\mu C_k = 2\,\mathrm{Tr}[2\sigma_k\partial_\mu(\sigma_k)] - 2\,\mathrm{Tr}[\rho_k\partial_\mu(\sigma_k)], \tag{C1}$$

where $\sigma_k = \sigma_k(\boldsymbol{\theta})$, which is constructed via the parametrized circuit $U_k(\boldsymbol{\theta})$, and ρ_k is the kth step reduced target. In a practical sense, our U_k is composed of the quantum gates satisfying the parameter-shift rule and $U_k = U_l e^{-i(\theta_\mu/2)\Omega_\mu}U_r =$

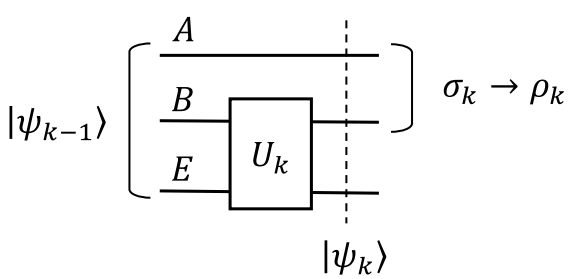


FIG. 6. The kth learning step layer. Based on adaptive learning processes, the previously learned state $|\psi_{k-1}\rangle$ on system ABE must be pure, where E is the additional system acted upon by the kth step layer U_k. Under perfect learning defined in Sec. III A, we have $\sigma_{k-1} = \mathrm{Tr}_{BE}(\psi_{k-1}) = \rho_{k-1}$.

$\tilde{U}_l U_r$, where $\Omega_\mu^2 = I$. The kth scattering layer is shown in Fig. 6. Then the following lemma holds.

Lemma 2. The kth step cost function C_k has the partial derivative form (with respect to θ_μ and evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$)

$$\partial_\mu C_k^* = \left\langle \Delta_k^* \otimes \frac{I_E}{d_E} \right\rangle_{\theta_\mu + \pi/2} - \left\langle \Delta_k^* \otimes \frac{I_E}{d_E} \right\rangle_{\theta_\mu - \pi/2}, \tag{C2}$$

where $\Delta_k = \sigma_k - \rho_k$ and the asterisk indicates the state difference evaluated at $\boldsymbol{\theta}^*$. The other symbols all match the settings in Fig. 6.

Proof. By observing $\sigma_k = \mathrm{Tr}_E[(I_A \otimes U_k) P_{\psi_{k-1}} (I_A \otimes U_k^\dagger)]$, where $P_{\psi_{k-1}} = |\psi_{k-1}\rangle\langle\psi_{k-1}|$, we could compute the expression of $\partial_\mu \sigma_k$ based on the linearity of the derivative operation,

$$\begin{aligned} \partial_\mu \sigma_k = {} & \mathrm{Tr}_E\{[I_A \otimes \partial_\mu(U_k)] P_{\psi_{k-1}} (I_A \otimes U_k^\dagger)\} \\ & + \mathrm{Tr}_E\{(I_A \otimes U_k) P_{\psi_{k-1}} [I_A \otimes \partial_\mu(U_k^\dagger)]\}. \end{aligned} \tag{C3}$$

Recalling the expression of $\partial_\mu(U_k)$ and $\partial_\mu(U_k^\dagger)$, we have

$$\partial_\mu \sigma_k = -\frac{i}{2} \mathrm{Tr}_E\{(I_A \otimes \tilde{U}_l)[(I_A \otimes \Omega_\mu), (I_A \otimes U_r) P_{\psi_{k-1}} (I_A \otimes U_r^\dagger)](I_A \otimes \tilde{U}_l^\dagger)\} = -\frac{i}{2} \mathrm{Tr}_E[\tilde{U}_l (\Omega_\mu, U_r P_{\psi_{k-1}} U_r^\dagger) \tilde{U}_l^\dagger], \tag{C4}$$

where we have abbreviated the $I_A\otimes$ correspondence for simplicity, meaning the subsystem A would never join the optimization during the kth step. Since $U_\mu(\theta_\mu) = e^{-i(\theta_\mu/2)\Omega_\mu}$ satisfies the parameter-shift rule, we could use the gate identity

$$i(\Omega_\mu, M) = U_\mu\left(-\frac{\pi}{2}\right) M U_\mu^\dagger\left(-\frac{\pi}{2}\right) - U_\mu\left(\frac{\pi}{2}\right) M U_\mu^\dagger\left(\frac{\pi}{2}\right) \tag{C5}$$

for any linear operator M and then derive the exact value of $\partial_\mu \sigma_k^*$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ as

$$\begin{aligned} \partial_\mu(\sigma_k^*) = \frac{1}{2} \mathrm{Tr}_E \Bigg[& U_k\left(\theta_\mu^* + \frac{\pi}{2}\right) P_{\psi_{k-1}} U_k^\dagger\left(\theta_\mu^* + \frac{\pi}{2}\right) \\ & - U_k\left(\theta_\mu^* - \frac{\pi}{2}\right) P_{\psi_{k-1}} U_k^\dagger\left(\theta_\mu^* - \frac{\pi}{2}\right) \Bigg]. \end{aligned} \tag{C6}$$

Here $\partial_\mu(\sigma_k^*) = \partial_\mu(\sigma_k)|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ and the circuit $U_k(\theta_\mu^* + \alpha)$ takes in $\boldsymbol{\theta}^*$ and modifies the parameter θ_μ^* to $\theta_\mu^* + \frac{\pi}{2}$. Now recalling the fact that

$$\mathrm{Tr}[\mathrm{Tr}_B(\rho_{AB})\sigma_A] = \mathrm{Tr}\left[\rho_{AB}\left(\sigma_A \otimes \frac{I_B}{d_B}\right)\right], \tag{C7}$$

we have

$$\begin{aligned} \mathrm{Tr}[\rho_k \partial_\mu(\sigma_k^*)] &= \left\langle \rho_k \otimes \frac{I_E}{d_E} \right\rangle_{\theta_\mu^* + \pi/2} - \left\langle \rho_k \otimes \frac{I_E}{d_E} \right\rangle_{\theta_\mu^* - \pi/2}, \\ \mathrm{Tr}[\sigma_k^* \partial_\mu(\sigma_k^*)] &= \left\langle \sigma_k^* \otimes \frac{I_E}{d_E} \right\rangle_{\theta_\mu^* + \pi/2} - \left\langle \sigma_k^* \otimes \frac{I_E}{d_E} \right\rangle_{\theta_\mu^* - \pi/2}, \end{aligned} \tag{C8}$$

where $\langle M \rangle_\theta = \langle \psi_k(\theta)|M|\psi_k(\theta)\rangle$ and $|\psi_k(\theta)\rangle$ is derived by applying $U_k(\theta)$ on $|\psi_{k-1}\rangle$. We combine the above calculations to obtain the desired result in Lemma 2, taking $\Delta^* = \sigma_k(\boldsymbol{\theta}^*) - \rho_k$. Finally, by taking in the actual dimensional factors, we could derive the analytic form of the partial derivative as shown in Sec. II C. ■

APPENDIX D: PROOF OF THE EFFECTIVENESS OF THE QSSM

In this Appendix we prove the effectiveness of the QSSM based on Schmidt decomposition and the properties of purification. Purification is a commonly used mathematical procedure in quantum computing. For an arbitrary quantum state, its purification is not unique. However, we could bridge these purification states via unitary transformations, which we called the freedom in purification.

Lemma 3. Let $|\psi\rangle$ and $|\phi\rangle$ be two purifications of a state ρ acting on a composite system AE. Then there exists a unitary U_E locally acting on E such that

$$|\psi\rangle = (I_A \otimes U_E)|\phi\rangle. \tag{D1}$$

Proof. The proof is simply inspired by the Schmidt decomposition. Let $|\psi\rangle$ and $|\phi\rangle$ be the purifications of ρ acting on AE. Write the Schmidt decomposition of these two states as

$$|\psi\rangle = \sum_j \sqrt{\lambda_j} |j^A\rangle|j^E\rangle, \quad |\phi\rangle = \sum_k \sqrt{\eta_k} |k^A\rangle|k^E\rangle. \tag{D2}$$

Note that $\mathrm{Tr}_E(\psi) = \rho = \mathrm{Tr}_E(\phi)$, which then induces

$$\sum_j \lambda_j |j^A\rangle\langle j^A| = \sum_k \eta_k |k^A\rangle\langle k^A|. \tag{D3}$$

By linear algebra, we could easily extend both $\{|j^A\rangle\}_j$ and $\{|k^E\rangle\}_k$ to the basis set of $\mathcal{H}^E$, via the Gram-Schmidt method, and hence prove the existence of a unitary U_E such that

$$U_E|k^E\rangle = |j^E\rangle, \tag{D4}$$

which is then substituted into the above equations to prove the lemma. ■

Based on the freedom in purification, we could prove Lemma 1 and therefore prove the effectiveness of our QSSM.

Proof. From the definition, $|\psi\rangle\langle\psi|$ and ρ have the same reduced state acting on A. Suppose the state $|\phi\rangle$ is the purification of ρ on system ABE. Thus it is also a purification of $\rho^A = \mathrm{Tr}_B(\rho)$. As we have that $|\phi\rangle$ and $|\psi\rangle$ are both acting on the composite system ABE, by Lemma 3 there exists a U_{BE} such that

$$|\phi\rangle\langle\phi| = (I_A \otimes U_{BE})|\psi\rangle\langle\psi|(I_A \otimes U_{BE}^\dagger). \tag{D5}$$

Since $|\phi\rangle$ is the purification of ρ, we have

$$\mathrm{tr}_E[|\phi\rangle\langle\phi|] = \rho, \tag{D6}$$

as required. ■

Now we are ready to prove the effectiveness of the QSSM. The proof assumes a sufficient number of computational

resources, which then ensures ideal learning for each step's reduced target.

Proposition 3. Given any target n-qubit pure state ψ, we claim that our QSSM could ideally learn the state.

Proof. Based on the setup of the algorithm, we divide the entire learning task into three main stages. In the beginning, a state $|0\rangle$ is initialized for the model. We denote the step by $k = 1$ for learning the reduced state acting on $\mathcal{A}_1$ of a single qubit. Note that for any one-qubit state $\rho_{\mathcal{A}_1}$ with a eigendecomposition,

$$\rho_{\mathcal{A}_1} = \lambda_1^{(1)}|0^{(1)}\rangle\langle 0^{(1)}| + \lambda_2^{(1)}|1^{(1)}\rangle\langle 1^{(1)}|, \tag{D7}$$

where the states $|0^{(1)}\rangle$ and $|1^{(1)}\rangle$ are not necessarily the computational basis elements. There exists a purification unitary $U_{\mathcal{A}_1\mathcal{A}_2}$ such that

$$U_{\mathcal{A}_1\mathcal{A}_2}|00\rangle = \sqrt{\lambda_1^{(1)}}|0^{(1)}\rangle|0^{(2)}\rangle + \sqrt{\lambda_2^{(1)}}|1^{(1)}\rangle|1^{(2)}\rangle. \tag{D8}$$

Such a unitary should have the following components. The rest of the matrix can be extended using the Gram-Schmidt process. We could write the computational basis representation of $U_{\mathcal{A}_1\mathcal{A}_2}$ as

$$(U_{\mathcal{A}_1\mathcal{A}_2})_{mn} = \begin{pmatrix} \sqrt{\lambda_1^{(1)}}\langle 00|0^{(1)}0^{(2)}\rangle + \sqrt{\lambda_2^{(1)}}\langle 00|1^{(1)}1^{(2)}\rangle & \cdots \\ \sqrt{\lambda_1^{(1)}}\langle 01|0^{(1)}0^{(2)}\rangle + \sqrt{\lambda_2^{(1)}}\langle 01|1^{(1)}1^{(2)}\rangle & \cdots \\ \sqrt{\lambda_1^{(1)}}\langle 10|0^{(1)}0^{(2)}\rangle + \sqrt{\lambda_2^{(1)}}\langle 10|1^{(1)}1^{(2)}\rangle & \cdots \\ \sqrt{\lambda_1^{(1)}}\langle 11|0^{(1)}0^{(2)}\rangle + \sqrt{\lambda_2^{(1)}}\langle 11|1^{(1)}1^{(2)}\rangle & \cdots \end{pmatrix}. \tag{D9}$$

For $1 < k \leqslant \lceil n/2 \rceil$, by the assumption of ideal learning of the state $\rho_{\mathcal{A}_{k-1}}$, a purification (denoted by $|\psi_{k-1}\rangle$) of it would be imported from the $(k-1)$th step. The reduced state $\rho_{\mathcal{A}_k}$ would in general require at least k extra ancillary qubits to be purified, which is the reason a width control $w_k = k+1$ or $n-k+1$ was settled on in the original QSSM.

Now suppose a purification $|\phi_k\rangle$ of $\rho_{\mathcal{A}_k}$. Since $\dim(|\psi_{k-1}\rangle) \leqslant \dim(|\phi_k\rangle)$, we could always extend $|\psi_k\rangle$ to $|\tilde{\psi}_k\rangle = |\psi_k\rangle|0\rangle$ so that the resulting pure state exists in the same dimensional Hilbert space as $|\phi_k\rangle$. Now we could observe $|\tilde{\psi}_k\rangle$ and $|\phi_k\rangle$ as two purifications of $\rho_{\mathcal{A}_{k-1}}$. Based on Lemma 1, there exists a U_k acting on the qubit index from $k+1$ to ζ_k such that

$$\mathrm{Tr}_{\bar{\mathcal{A}}_{k+1}}[(I_{\mathcal{A}_{k-1}} \otimes U_k)\tilde{\psi}_k(I_{\mathcal{A}_{k-1}} \otimes U_k^\dagger)] = \mathrm{Tr}_{\bar{\mathcal{A}}_{k+1}}(\phi_k) = \rho_{\mathcal{A}_k}. \tag{D10}$$

Finally, for $\lceil n/2 \rceil < k \leqslant n$, $|\phi_k\rangle$ becomes the pure state acting on the entire system of n qubits. The imported purification $|\psi_{k-1}\rangle$ of $\rho_{\mathcal{A}_{k-1}}$ is also a pure state of n qubits. The result follows by applying Lemma 1 again. Above all, we have proven the effectiveness of the QSSM. The proof also encourages us to study situations when the sequential scattering unitary U_k is not determined via the doubling strategy and better understand the truncated version of the QSSM. ■

Based on Proposition 3 and freedom in purification, we can easily arrive at Proposition 1.

APPENDIX E: PROOF OF THE TRAINABILITY OF THE QSSM

In this Appendix we give the proof for Proposition 2 stated about the trainability of the QSSM in this paper. To make the proof easy to read and to emphasize important intermediate results, we first recall some useful lemmas. The following lemmas were derived from the studies of unitary t-design. These were originally computed in [69].

Lemma 4. Suppose $X \subset \mathcal{U}(d)$ is a unitary t-design and A, B, C, D are arbitrary linear operators. If $t \geqslant 1$, then we have

$$\frac{1}{|X|}\sum_{U\in X}\mathrm{Tr}(U^\dagger AUB) = \int_{\mathcal{U}(d)}\mathrm{Tr}(U^\dagger AUB)d\eta(U) = \frac{\mathrm{Tr}(A)\,\mathrm{Tr}(B)}{d}. \tag{E1}$$

If $t \geqslant 2$, then we have

$$\frac{1}{|X|}\sum_{U\in X}\mathrm{Tr}(U^\dagger AUBU^\dagger CUD) = \int_{\mathcal{U}(d)}\mathrm{Tr}(U^\dagger AUBU^\dagger CUD)d\eta(U) \tag{E2}$$

$$= \frac{\mathrm{Tr}(A)\,\mathrm{Tr}(C)\,\mathrm{Tr}(BD) + \mathrm{Tr}(AC)\,\mathrm{Tr}(B)\,\mathrm{Tr}(D)}{d^2-1} - \frac{\mathrm{Tr}(AC)\,\mathrm{Tr}(BD) + \mathrm{Tr}(A)\,\mathrm{Tr}(B)\,\mathrm{Tr}(C)\,\mathrm{Tr}(D)}{d(d^2-1)}. \tag{E3}$$

Lemma 5. Suppose A, B, C, D are arbitrary linear operators. Then

$$\int_{\mathcal{U}(d)}\mathrm{Tr}(UAU^\dagger B)\,\mathrm{Tr}(UCU^\dagger D)d\eta(U) = \frac{1}{d^2-1}[\mathrm{Tr}(A)\,\mathrm{Tr}(B)\,\mathrm{Tr}(C)\,\mathrm{Tr}(D) + \mathrm{Tr}(AC)\,\mathrm{Tr}(BD)] - \frac{1}{d(d^2-1)}[\mathrm{Tr}(AC)\,\mathrm{Tr}(B)\,\mathrm{Tr}(D) + \mathrm{Tr}(A)\,\mathrm{Tr}(C)\,\mathrm{Tr}(BD)]. \tag{E4}$$

Lemma 6. Let $\mathcal{H} = \mathcal{H}_\mathcal{A} \otimes \mathcal{H}_\mathcal{B}$ be a bipartite Hilbert space of dimension $d = d_A d_B$, and for arbitrary linear operators $M, N : \mathcal{H} \to \mathcal{H}$ we have

$$\int_{\mathcal{U}(d_B)} d\eta(U)(I_A \otimes U)M(I_A \otimes U^\dagger)N = \frac{\mathrm{Tr}_B(M) \otimes I_B}{d_B}N \tag{E5}$$

and

$$\int_{\mathcal{U}(d_B)} d\eta(U)\,\mathrm{Tr}[(I_A \otimes U)M(I_A \otimes U^\dagger)N] = \frac{\mathrm{Tr}[\mathrm{Tr}_B(M)\,\mathrm{Tr}_B(N)]}{d_B}. \tag{E6}$$

Lemma 7. Let $\mathcal{H} = \mathcal{H}_\mathcal{A} \otimes \mathcal{H}_\mathcal{B}$ be a bipartite Hilbert space of dimension $d = d_A d_B$ ($d = 2^n$ and $d_A = 2^{n'}$), and for arbitrary linear operators $M, N, U : \mathcal{H} \to \mathcal{H}$ we have

$$\mathrm{Tr}[(I_A \otimes U)M(I_A \otimes U^\dagger)N] = \sum_{p,q}\mathrm{Tr}(UM_{qp}U^\dagger N_{pq}), \tag{E7}$$

where the summation runs over all bit strings of length n' and

$$M_{qp} = \mathrm{Tr}_A[(|p\rangle\langle q| \otimes I)M], \tag{E8}$$

$$N_{pq} = \mathrm{Tr}_A[(|q\rangle\langle p| \otimes I)N]. \tag{E9}$$

With these lemmas, now we can start our proof by directly calculating the variance of gradients. The whole proof includes three parts indicating the gradient magnitude of different stages in the algorithm.

Proposition 4. For an n-qubit target state ρ, assume we start from the $\hat{\sigma}$ such that $\mathrm{Tr}_n(\rho) = \mathrm{Tr}_n(\hat{\sigma})$, where $\mathrm{Tr}_n(\rho)$ denotes the partial trace over the last qubit of the state. In addition, if the circuit is only acting on the last qubit and forms a 2-design, then $\mathbb{E}(\partial_\mu C_n) = 0$ and the variance $\mathrm{Var}(\partial_\mu C_n) \in [\frac{16}{27}, \frac{8}{9}]$.

Proof. Suppose the output state is σ. Then the cost function is

$$C_n(\boldsymbol{\theta}) = \mathrm{Tr}\{[\rho - \sigma(\boldsymbol{\theta})][\rho - \sigma(\boldsymbol{\theta})]^\dagger\}. \tag{E10}$$

With notation similar to that used by McClean *et al.* in [26], we can use U to denote the unitary representation of circuits. We can write it as $U = U_+ e^{-i\theta_\mu H} U_-$, where H denotes the Hermitian operator and in most cases it will be the Pauli matrices, which are traceless. Since $\mathrm{Tr}_n(\rho) = \mathrm{Tr}_n(\hat{\sigma})$, we have

$$\hat{\sigma} = (I_A \otimes V_B)\rho(I_A \otimes V_B^\dagger), \tag{E11}$$

where V is a fixed unitary and system A denotes the first $n-1$ qubits and system B denotes the last qubit. So $d_A = 2^{n-1}$ and $d_B = 2$. For simplicity, we will omit the subscript in the following proof. We then arrive at

$$\sigma = (I \otimes UV)\rho(I \otimes V^\dagger U^\dagger). \tag{E12}$$

Next we compute the partial derivative of C_n with respect to the kth parameter. Note that the trace is linear; the derivative operation could pass through the trace and hence we obtain

$$\partial_\mu C_n = \partial_\mu \mathrm{Tr}[\rho^2 + \sigma^2 - 2(\rho\sigma)] = -2\,\mathrm{Tr}[\rho\,\partial_\mu(\sigma)].$$

Now we start by calculating the mean of gradients. Expanding the expression for σ, we could find

$$\partial_\mu C_n = -2\,\mathrm{Tr}(\rho\{[I \otimes (\partial_\mu U)V]\rho(I \otimes V^\dagger U^\dagger) + (I \otimes UV)\rho[I \otimes V^\dagger(\partial_\mu U^\dagger)]\}),$$

by the chain rule of derivative. Since $U = U_+ e^{-i\theta_\mu H} U_-$, we could compute the derivatives as

$$\partial_\mu U = -iU_+ e^{-i\theta_\mu H} H U_-, \quad \partial_\mu U^\dagger = iU_-^\dagger H e^{i\theta_\mu H} U_+^\dagger.$$

For convenience, we define $\tilde{U}_+ = U_+ e^{-i\theta_\mu H}$. We substitute the above into the expression of the cost derivative to achieve

$$\partial_\mu C_n = 2i\,\mathrm{Tr}\{\rho[(I \otimes \tilde{U}_+ H U_- V)\rho(I \otimes V^\dagger U^\dagger) - (I \otimes UV)\rho(I \otimes V^\dagger U_-^\dagger H \tilde{U}_+^\dagger)]\}.$$

Now if we expand $U = \tilde{U}_+ U_-$ and assume the $\tilde{U}_- = U_- V$, we obtain

$$\begin{aligned}\partial_\mu C_n &= 2i\,\mathrm{Tr}\{\rho[(I \otimes \tilde{U}_+ H U_- V)\rho(I \otimes V^\dagger U_-^\dagger \tilde{U}_+^\dagger) - (I \otimes \tilde{U}_+ U_- V)\rho(I \otimes V^\dagger U_-^\dagger H \tilde{U}_+^\dagger)]\} \\ &= 2i\,\mathrm{Tr}\{\rho[(I \otimes \tilde{U}_+ H \tilde{U}_-)\rho(I \otimes \tilde{U}_-^\dagger \tilde{U}_+^\dagger) - (I \otimes \tilde{U}_+ \tilde{U}_-)\rho(I \otimes \tilde{U}_-^\dagger H \tilde{U}_+^\dagger)]\} \\ &= 2i\,\mathrm{Tr}\{(I \otimes \tilde{U}_+^\dagger)\rho(I \otimes \tilde{U}_+) \times [I \otimes H, (I \otimes \tilde{U}_-)\rho(I \otimes \tilde{U}_-^\dagger)]\},\end{aligned}$$

where $[A, B] = AB - BA$ denotes the commutator notation. We denote the commutator $[I \otimes H, (I \otimes \tilde{U}_-)\rho(I \otimes \tilde{U}_-^\dagger)]$ by T_-; thus we have

$$\partial_\mu C_n = 2i\,\mathrm{Tr}[(I \otimes \tilde{U}_+^\dagger)\rho(I \otimes \tilde{U}_+)T_-]. \tag{E13}$$

Then we integrate over $\tilde{U}_+$ by using Lemma 6,

$$\begin{aligned}\mathbb{E}(\partial_\mu C_n) &= 2i\frac{\mathrm{Tr}[\mathrm{Tr}_B(\rho)\,\mathrm{Tr}_B(T_-)]}{d_B} \\ &= i\,\mathrm{Tr}[\mathrm{Tr}_B(\rho)\,\mathrm{Tr}_B(T_-)].\end{aligned}$$

We can write ρ as

$$\rho = \sum_{i,j} |i\rangle\langle j|_A \otimes X_{i,j}, \tag{E14}$$

which thus leads to

$$\begin{aligned}&\mathrm{Tr}_B(T_-) \\ &= \mathrm{Tr}_B\{[I \otimes H, (I \otimes \tilde{U}_-)\rho(I \otimes \tilde{U}_-^\dagger)]\} \\ &= \sum_{i,j} \mathrm{Tr}_B\{[I \otimes H, (I \otimes \tilde{U}_-)(|i\rangle\langle j|_A \otimes X_{i,j})(I \otimes \tilde{U}_-^\dagger)]\} \\ &= \sum_{i,j} \mathrm{Tr}_B(|i\rangle\langle j| \otimes H\tilde{U}_- X_{i,j}\tilde{U}_-^\dagger - |i\rangle\langle j| \otimes \tilde{U}_- X_{i,j}\tilde{U}_-^\dagger H) \\ &= \sum_{i,j} |i\rangle\langle j|[\mathrm{Tr}(H\tilde{U}_- X_{i,j}\tilde{U}_-^\dagger) - \mathrm{Tr}(\tilde{U}_- X_{i,j}\tilde{U}_-^\dagger H)] \\ &= 0.\end{aligned} \tag{E15}$$

Therefore, we have

$$\mathbb{E}(\partial_\mu C_n) = 0. \tag{E16}$$

The mean of gradients is 0. Based on the fact that the mean of gradients is 0, we then only need to consider the $\mathbb{E}[(\partial_\mu C_n)^2]$ in order to determine the variance

$$\begin{aligned}\mathrm{Var}(\partial_\mu C_n) &= \mathbb{E}[(\partial_\mu C_n)^2] \\ &= -4\mathbb{E}_{\tilde{U}_+,\tilde{U}_-}(\{\mathrm{Tr}[(I \otimes \tilde{U}_+^\dagger)\rho(I \otimes \tilde{U}_+)T_-]\}^2).\end{aligned} \tag{E17}$$

Using Lemma 7, we have

$$\begin{aligned}&\mathbb{E}_{\tilde{U}_+,\tilde{U}_-}(\{\mathrm{Tr}[(I \otimes \tilde{U}_+^\dagger)\rho(I \otimes \tilde{U}_+)T_-]\}^2) \\ &= \mathbb{E}_{\tilde{U}_+,\tilde{U}_-}\left[\left(\sum_{p,q} \mathrm{Tr}(\tilde{U}_+ \rho_{qp} \tilde{U}_+^\dagger T_{-pq})\right) \times \left(\sum_{m,n} \mathrm{Tr}(\tilde{U}_+ \rho_{nm} \tilde{U}_+^\dagger T_{-mn})\right)\right] \\ &= \mathbb{E}_{\tilde{U}_+,\tilde{U}_-}\left(\sum_{p,q,m,n} \mathrm{Tr}(\tilde{U}_+ \rho_{qp} \tilde{U}_+^\dagger T_{-pq})\,\mathrm{Tr}(\tilde{U}_+ \rho_{nm} \tilde{U}_+^\dagger T_{-mn})\right) \\ &= \sum_{p,q,m,n} \mathbb{E}_{\tilde{U}_+,\tilde{U}_-}[\mathrm{Tr}(\tilde{U}_+ \rho_{qp} \tilde{U}_+^\dagger T_{-pq})\,\mathrm{Tr}(\tilde{U}_+ \rho_{nm} \tilde{U}_+^\dagger T_{-mn})].\end{aligned}$$

Then, according to Lemma 5,

$$\begin{aligned}
&\sum_{p,q,m,n} \mathbb{E}_{\tilde{U}_+,\tilde{U}_-}[\mathrm{Tr}(\tilde{U}_+\rho_{qp}\tilde{U}_+^\dagger T_{-pq})\,\mathrm{Tr}(\tilde{U}_+\rho_{nm}\tilde{U}_+^\dagger T_{-mn})] \\
&= \sum_{p,q,m,n} \mathbb{E}_{\tilde{U}_-}\bigg(\frac{1}{d_B^2-1}[\mathrm{Tr}(\rho_{qp})\,\mathrm{Tr}(T_{-pq})\,\mathrm{Tr}(\rho_{nm})\,\mathrm{Tr}(T_{-mn}) + \mathrm{Tr}(\rho_{qp}\rho_{nm})\,\mathrm{Tr}(T_{-pq}T_{-mn})] \\
&\qquad - \frac{1}{d_B(d_B^2-1)}[\mathrm{Tr}(\rho_{qp}\rho_{nm})\,\mathrm{Tr}(T_{-pq})\,\mathrm{Tr}(T_{-mn}) + \mathrm{Tr}(\rho_{qp})\,\mathrm{Tr}(\rho_{nm})\,\mathrm{Tr}(T_{-pq}T_{-mn})]\bigg).
\end{aligned} \tag{E18}$$

Since

$$\mathrm{Tr}(\rho_{qp}) = \mathrm{Tr}\{\mathrm{Tr}_A[(|p\rangle\langle q|\otimes I)\rho]\} = \mathrm{Tr}[(|p\rangle\langle q|\otimes I)\rho] = \mathrm{Tr}[|p\rangle\langle q|\,\mathrm{Tr}_B(\rho)] = \langle q|\,\mathrm{Tr}_B(\rho)|p\rangle \tag{E19}$$

and

$$\mathrm{Tr}(T_{-pq}) = \mathrm{Tr}\{\mathrm{Tr}_A[(|q\rangle\langle p|\otimes I)T_-]\} = \mathrm{Tr}[(|q\rangle\langle p|\otimes I)T_-] = \mathrm{Tr}[|q\rangle\langle p|\,\mathrm{Tr}_B(T_-)] = 0, \tag{E20}$$

Eq. (E20) holds because of Eq. (E15). Thus Eq. (E18) can be simplified as

$$\begin{aligned}
&\sum_{p,q,m,n} \mathbb{E}_{\tilde{U}_+,\tilde{U}_-}[\mathrm{Tr}(\tilde{U}_+\rho_{qp}\tilde{U}_+^\dagger T_{-pq})\,\mathrm{Tr}(\tilde{U}_+\rho_{nm}\tilde{U}_+^\dagger T_{-mn})] \\
&= \sum_{p,q,m,n} \mathbb{E}_{\tilde{U}_-}\bigg(\frac{1}{d_B^2-1}\,\mathrm{Tr}(\rho_{qp}\rho_{nm})\,\mathrm{Tr}(T_{-pq}T_{-mn}) - \frac{1}{d_B(d_B^2-1)}\,\mathrm{Tr}(\rho_{qp})\,\mathrm{Tr}(\rho_{nm})\,\mathrm{Tr}(T_{-pq}T_{-mn})\bigg) \\
&= \sum_{p,q,m,n} \mathbb{E}_{\tilde{U}_-}\bigg(\frac{1}{d_B(d_B^2-1)}\,\mathrm{Tr}(T_{-pq}T_{-mn})[d_B\,\mathrm{Tr}(\rho_{qp}\rho_{nm}) - \mathrm{Tr}(\rho_{qp})\,\mathrm{Tr}(\rho_{nm})]\bigg) \\
&= \sum_{p,q,m,n} \frac{1}{d_B(d_B^2-1)}[d_B\,\mathrm{Tr}(\rho_{qp}\rho_{nm}) - \mathrm{Tr}(\rho_{qp})\,\mathrm{Tr}(\rho_{nm})]\mathbb{E}_{\tilde{U}_-}[\mathrm{Tr}(T_{-pq}T_{-mn})].
\end{aligned}$$

We now need to evaluate the other integral with respect to $\tilde{U}_-$. A simplification can be made by noting that

$$\begin{aligned}
T_{-pq} &= \mathrm{Tr}_A[(|q\rangle\langle p|\otimes I)T_-] = \mathrm{Tr}_A[I\otimes H, (I\otimes\tilde{U}_-)(|q\rangle\langle p|\otimes I)\rho(I\otimes\tilde{U}_-^\dagger)] \\
&= \mathrm{tr}_A([I\otimes H, (I\otimes\tilde{U}_-)(|q\rangle\langle p|\otimes I)\rho(I\otimes\tilde{U}_-^\dagger)]) = [H, \tilde{U}_-\rho_{pq}\tilde{U}_-^\dagger],
\end{aligned}$$

since $|p\rangle\langle q|\otimes I$ commutes with other operators. Therefore,

$$\begin{aligned}
\mathrm{Tr}(T_{-pq}T_{-mn}) &= \mathrm{Tr}([H,\tilde{U}_-\rho_{pq}\tilde{U}_-^\dagger][H,\tilde{U}_-\rho_{mn}\tilde{U}_-^\dagger]) \\
&= 2\,\mathrm{Tr}(H\tilde{U}_-\rho_{pq}\tilde{U}_-^\dagger H\tilde{U}_-\rho_{mn}\tilde{U}_-^\dagger) - \mathrm{Tr}(\tilde{U}_-\rho_{pq}\rho_{mn}\tilde{U}_-^\dagger H^2) - \mathrm{Tr}(\tilde{U}_-\rho_{mn}\rho_{pq}\tilde{U}_-^\dagger H^2).
\end{aligned}$$

So according to Lemma 4,

$$\begin{aligned}
\mathbb{E}_{\tilde{U}_-}[\mathrm{Tr}(T_{-pq}T_{-mn})] &= \frac{2}{d_B^2-1}[\mathrm{Tr}(\rho_{pq})\,\mathrm{Tr}(\rho_{mn})\,\mathrm{Tr}(H^2) + \mathrm{Tr}(\rho_{pq}\rho_{mn})\,\mathrm{Tr}^2(H)] \\
&\quad - \frac{2}{d_B(d_B^2-1)}[\mathrm{Tr}(\rho_{pq}\rho_{mn})\,\mathrm{Tr}(H^2) + \mathrm{Tr}(\rho_{pq})\,\mathrm{Tr}(\rho_{mn})\,\mathrm{Tr}^2(H)] - \frac{2}{d_B}\,\mathrm{Tr}(\rho_{pq}\rho_{mn})\,\mathrm{Tr}(H^2) \\
&= \frac{-2}{d_B(d_B^2-1)}[d_B\,\mathrm{Tr}(\rho_{pq}\rho_{mn}) - \mathrm{Tr}(\rho_{pq})\,\mathrm{Tr}(\rho_{mn})][d_B\,\mathrm{Tr}(H^2) - \mathrm{Tr}^2(H)] \\
&= \frac{-2}{d_B^2-1}\,\mathrm{Tr}(H^2)[d_B\,\mathrm{Tr}(\rho_{pq}\rho_{mn}) - \mathrm{Tr}(\rho_{pq})\,\mathrm{Tr}(\rho_{mn})].
\end{aligned}$$

Subsitituting the above into Eq. (E18) to obtain,

$$\begin{aligned}
&\sum_{p,q,m,n} \mathbb{E}_{\tilde{U}_+,\tilde{U}_-}[\mathrm{Tr}(\tilde{U}_+\rho_{qp}\tilde{U}_+^\dagger T_{-pq})\,\mathrm{Tr}(\tilde{U}_+\rho_{nm}\tilde{U}_+^\dagger T_{-mn})] \\
&= \sum_{p,q,m,n} \frac{-2}{d_B(d_B^2-1)^2}\,\mathrm{Tr}(H^2)[d_B\,\mathrm{Tr}(\rho_{qp}\rho_{nm}) - \mathrm{Tr}(\rho_{qp})\,\mathrm{Tr}(\rho_{nm})][d_B\,\mathrm{Tr}(\rho_{pq}\rho_{mn}) - \mathrm{Tr}(\rho_{pq})\,\mathrm{Tr}(\rho_{mn})].
\end{aligned}$$

First, we look at $\mathrm{Tr}(\rho_{qp}\rho_{nm})$,

$$\begin{aligned}\mathrm{Tr}(\rho_{qp}\rho_{nm}) &= \mathrm{Tr}\{\mathrm{Tr}_A[(|p\rangle\langle q|\otimes I)\rho]\,\mathrm{Tr}_A[(|m\rangle\langle n|\otimes I)\rho]\}\\ &= \mathrm{Tr}\left(\sum_i\{\langle i|\otimes I[(|p\rangle\langle q|\otimes I)\rho]|i\rangle\otimes I\}\sum_j[\langle j|\otimes I(|p\rangle\langle q|\otimes I)\rho|j\rangle\otimes I]\right)\\ &= \mathrm{Tr}[(\langle q|\otimes I)\rho(|p\rangle\langle n|\otimes I)\rho(|m\rangle\otimes I)] = \mathrm{Tr}\{\langle q|\,\mathrm{Tr}_B[\rho(|p\rangle\langle n|\otimes I)\rho]|m\rangle\}\\ &= \langle q|\,\mathrm{Tr}_B[\rho(|p\rangle\langle n|\otimes I)\rho]|m\rangle.\end{aligned}$$

Then

$$\begin{aligned}\sum_{p,q,m,n}\mathrm{Tr}(\rho_{qp}\rho_{nm})\,\mathrm{Tr}(\rho_{pq}\rho_{mn}) &= \sum_{p,q,m,n}\langle q|\,\mathrm{Tr}_B[\rho(|p\rangle\langle n|\otimes I)\rho]|m\rangle\langle m|\,\mathrm{Tr}_B[\rho(|n\rangle\langle p|\otimes I)\rho]|q\rangle\\ &= \sum_{p,n}\mathrm{Tr}\{\mathrm{Tr}_B[\rho(|p\rangle\langle n|\otimes I)\rho]\,\mathrm{Tr}_B[\rho(|n\rangle\langle p|\otimes I)\rho]\}.\end{aligned}$$

Suppose the Schmidt decomposition of $|\phi\rangle$ is

$$|\phi\rangle = \sum_k \lambda_k|u_k\rangle_A|v_k\rangle_B, \tag{E21}$$

where $\{|u_k\rangle\}$ is the orthogonal basis on the system A and $\{|v_k\rangle\}$ is the orthogonal basis on the system B. Therefore, we can write ρ as

$$\rho = \sum_{i,j}\lambda_i\lambda_j|u_i\rangle\langle u_j|\otimes|v_i\rangle\langle v_j|. \tag{E22}$$

We can expand ρ in $\mathrm{Tr}_B[\rho(|p\rangle\langle n|\otimes I)\rho]$,

$$\begin{aligned}\mathrm{Tr}_B[\rho(|p\rangle\langle n|\otimes I)\rho] &= \mathrm{Tr}_B\left[\left(\sum_{i,j}\lambda_i\lambda_j|u_i\rangle\langle u_j|\otimes|v_i\rangle\langle v_j|\right)(|p\rangle\langle n|\otimes I\right)\left(\sum_{k,l}\lambda_k\lambda_l|u_k\rangle\langle u_l|\otimes|v_k\rangle\langle v_l|\right)\right]\\ &= \sum_{i,j,k,l}\lambda_i\lambda_j\lambda_k\lambda_l\,\mathrm{Tr}_B(|u_i\rangle\langle u_j||p\rangle\langle n||u_k\rangle\langle u_l|\otimes|v_i\rangle\langle v_j||v_k\rangle\langle v_l|)\\ &= \sum_{i,j}\lambda_i^2\lambda_j^2|u_i\rangle\langle u_j||p\rangle\langle n||u_j\rangle\langle u_i|.\end{aligned}$$

Thus, we arrive at

$$\begin{aligned}&\sum_{p,n}\mathrm{Tr}\{\mathrm{Tr}_B[\rho(|p\rangle\langle n|\otimes I)\rho]\,\mathrm{Tr}_B[\rho(|n\rangle\langle p|\otimes I)\rho]\}\\ &= \sum_{p,n}\mathrm{Tr}\left[\left(\sum_{i,j}\lambda_i^2\lambda_j^2|u_i\rangle\langle u_j||p\rangle\langle n||u_j\rangle\langle u_i|\right)\left(\sum_{k,l}\lambda_k^2\lambda_l^2|u_k\rangle\langle u_l||p\rangle\langle n||u_k\rangle\langle u_l|\right)\right]\\ &= \sum_{p,n}\mathrm{Tr}\left(\sum_{i,j,k,l}\right)\lambda_i^2\lambda_j^2\lambda_k^2\lambda_l^2|u_i\rangle\langle u_j||p\rangle\langle n||u_j\rangle\langle u_i||u_k\rangle\langle u_l||n\rangle\langle p||u_l\rangle\langle u_k|\\ &= \sum_{p,n}\sum_{i,j,l}\lambda_i^4\lambda_j^2\lambda_l^2\,\mathrm{Tr}(\langle u_j||p\rangle\langle n||u_j\rangle\langle u_l||n\rangle\langle p||u_l\rangle)\\ &= \sum_{i,j,l}\lambda_i^4\lambda_j^2\lambda_l^2\,\mathrm{Tr}[\mathrm{Tr}(|u_l\rangle\langle u_j|)\,\mathrm{Tr}(|u_j\rangle\langle u_l|)] = \sum_{i,j}\lambda_i^4\lambda_j^4 = \left(\sum_i\lambda_i^4\right)^2.\end{aligned}$$

Then we look at $\mathrm{Tr}(\rho_{qp})\,\mathrm{Tr}(\rho_{pq})\,\mathrm{Tr}(\rho_{mn})\,\mathrm{Tr}(\rho_{nm})$,

$$\begin{aligned}\sum_{p,q,m,n}\mathrm{Tr}(\rho_{qp})\,\mathrm{Tr}(\rho_{pq})\,\mathrm{Tr}(\rho_{mn})\,\mathrm{Tr}(\rho_{nm}) &= \sum_{p,q,m,n}\langle q|\,\mathrm{Tr}_B(\rho)|p\rangle\langle p|\,\mathrm{Tr}_B(\rho)|q\rangle\langle m|\,\mathrm{Tr}_B(\rho)|n\rangle\langle n|\,\mathrm{Tr}_B(\rho)|m\rangle\\ &= \mathrm{Tr}[\mathrm{Tr}_B(\rho)\,\mathrm{Tr}_B(\rho)]\,\mathrm{Tr}[\mathrm{Tr}_B(\rho)\,\mathrm{Tr}_B(\rho)] = \{\mathrm{Tr}[\mathrm{Tr}_B(\rho)\,\mathrm{Tr}_B(\rho)]\}^2 = \left(\sum_i\lambda_i^4\right)^2.\end{aligned}$$

Now we look at $\mathrm{Tr}(\rho_{qp}\rho_{nm})\,\mathrm{Tr}(\rho_{pq})\,\mathrm{Tr}(\rho_{mn})$:

$$\mathrm{Tr}(\rho_{qp}\rho_{nm}) = \langle n|\,\mathrm{Tr}_B[\rho(|m\rangle\langle q|\otimes I)\rho]|p\rangle \tag{E23}$$

$$= \sum_{i,j}\lambda_i^2\lambda_j^2\langle n||u_i\rangle\langle u_j||m\rangle\langle q||u_j\rangle\langle u_i||p\rangle \tag{E24}$$

and

$$\mathrm{Tr}(\rho_{pq})\,\mathrm{Tr}(\rho_{mn}) = \langle p|\,\mathrm{Tr}_B(\rho)|q\rangle\langle m|\,\mathrm{Tr}_B(\rho)|n\rangle = \sum_{i,j}\lambda_i^2\lambda_j^2\langle p||u_i\rangle\langle u_i||q\rangle\langle m||u_j\rangle\langle u_j||n\rangle.$$

Thus,

$$\begin{aligned}\sum_{p,q,m,n}\mathrm{Tr}(\rho_{qp}\rho_{nm})\,\mathrm{Tr}(\rho_{pq})\,\mathrm{Tr}(\rho_{mn}) &= \sum_{p,q,m,n}\left(\sum_{k,l}\lambda_k^2\lambda_l^2\langle n||u_k\rangle\langle u_l||m\rangle\langle q||u_k\rangle\langle u_l||p\rangle\right)\left(\sum_{i,j}\lambda_i^2\lambda_j^2\langle p||u_i\rangle\langle u_i||q\rangle\langle m||u_j\rangle\langle u_j||n\rangle\right)\\ &= \sum_{p,q,m,n}\sum_{i,j,k,l}\lambda_i^2\lambda_j^2\lambda_k^2\lambda_l^2(\langle n||u_k\rangle\langle u_l||m\rangle\langle q||u_k\rangle\langle u_l||p\rangle\langle p||u_i\rangle\langle u_i||q\rangle\langle m||u_j\rangle\langle u_j||n\rangle)\\ &= \sum_{q,m}\sum_{i,j,k,l}\lambda_i^2\lambda_j^2\lambda_k^2\lambda_l^2\,\mathrm{Tr}(|u_k\rangle\langle u_l||m\rangle\langle q||u_k\rangle\langle u_l||u_i\rangle\langle u_i||q\rangle\langle m||u_j\rangle\langle u_j|)\\ &= \sum_{i,j,k,l}\lambda_i^2\lambda_j^2\lambda_k^2\lambda_l^2\,\mathrm{Tr}(|u_k\rangle\langle u_l||u_i\rangle\langle u_i|)\,\mathrm{Tr}(|u_j\rangle\langle u_j||u_k\rangle\langle u_l|) = \sum_i\lambda_i^8.\end{aligned}$$

Therefore, we have

$$\sum_{p,q,m,n}[d_B\,\mathrm{Tr}(\rho_{pq}\rho_{mn}) - \mathrm{Tr}(\rho_{pq})\,\mathrm{Tr}(\rho_{mn})]$$

$$= (d_B^2+1)\left(\sum_i\lambda_i^4\right)^2 - 2d_B\left(\sum_i\lambda_i^8\right).$$

So

$$\mathrm{Var}(\partial_\mu C_n) = \frac{8}{d_B(d_B^2-1)^2}\,\mathrm{Tr}(H^2)\left[(d_B^2+1)\left(\sum_i\lambda_i^4\right)^2 - 2d_B\left(\sum_i\lambda_i^8\right)\right]. \tag{E25}$$

Since d_B is 2, we can simplify the equation above as

$$\mathrm{Var}(\partial_\mu C_n) = \tfrac{4}{9}\,\mathrm{Tr}(H^2)\big(\lambda_1^8+\lambda_2^8+10\lambda_1^4\lambda_2^4\big) \tag{E26}$$

$$= \tfrac{8}{9}\big(c_1^4+c_2^4+10c_1^2c_2^2\big), \tag{E27}$$

where $c_1=\lambda_1^2$ and $c_2=\lambda_2^2$ such that $c_1+c_2=1$, and $\mathrm{Tr}(H^2)=d_B=2$. Therefore, we can simply to get the range of the variance

$$\tfrac{16}{27} \leqslant \mathrm{Var}(\partial_\mu C_n) \leqslant \tfrac{8}{9}. \tag{E28}$$

■

Lemma 8. For the target pure state ρ_{ABC} on system ABC, suppose we start from an initial state $\hat{\sigma}$ such that $\mathrm{Tr}_{BC}(\rho) = \mathrm{Tr}_{BC}(\hat{\sigma})$ and the output state is σ. If the cost function is

$$C = \mathrm{Tr}\{[\mathrm{Tr}_C(\rho)-\mathrm{Tr}_C(\sigma)][\mathrm{Tr}_C(\rho)-\mathrm{Tr}_C(\sigma)]\} \tag{E29}$$

and the circuit is acting on system BC while forming a local 4-design, then $\mathbb{E}(\partial_\mu C)=0$ and the variance of the cost gradient scales as $\mathrm{Var}(\partial_\mu C)\in O(\frac{1}{d_B^3 d_C})$, where d_B and d_C denote the dimensions of systems B and C, respectively.

Proof. Since $\mathrm{Tr}_{BC}(\rho)=\mathrm{Tr}_{BC}(\hat{\sigma})$, there exist a fixed unitary V such that

$$\hat{\sigma} = (I_A\otimes V_{BC})\rho(I_A\otimes V_{BC}^\dagger). \tag{E30}$$

Then

$$\sigma = (I\otimes UV)\rho(I\otimes V^\dagger U^\dagger). \tag{E31}$$

Then the cost gradient becomes

$$\begin{aligned}\partial_\mu C &= 2\,\mathrm{Tr}[\mathrm{Tr}_C(\sigma)\partial_\mu\,\mathrm{Tr}_C(\sigma) - 2\,\mathrm{Tr}[\mathrm{Tr}_C(\rho)\partial_\mu\,\mathrm{Tr}_C(\sigma)]]\\ &= 2i\,\mathrm{Tr}[\mathrm{Tr}_C[(I\otimes U_+U_-V)\rho(I\otimes V^\dagger U_-^\dagger U_+^\dagger)-\rho]\\ &\quad\times\mathrm{Tr}_C[(I\otimes U_+U_-V)\rho(I\otimes V^\dagger U_-^\dagger H U_+^\dagger)\\ &\quad-(I\otimes U_+HU_-V)\rho(I\otimes V^\dagger U_-^\dagger U_+^\dagger)]].\end{aligned}$$

We exploit the RTNI package [79] to calculate the mean of the cost gradient. It turns out that the mean of the cost gradient is zero,

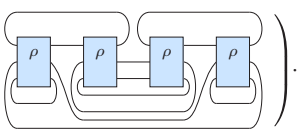
$$\mathbb{E}(\partial_\mu C) = 0. \tag{E32}$$

Then we consider the variance

$$\mathrm{Var}(\partial_\mu \mathrm{C}) = -\mathbb{E}[(\partial_\mu \mathrm{C})^2]. \tag{E33}$$

With the RTNI package, it turns out that the exact expression of the variance is dominant by

$$\mathrm{Var}[\partial_\mu C] \xrightarrow{d\to\infty} \frac{\mathrm{Tr}[H^2]}{d_B^2(d_B^2d_C^2-1)}$$



(E34)

We know that $\mathrm{Tr}(H^2)=d_Bd_C$; thus we have

$$\mathrm{Var}(\partial_\mu C) \in O\left(\frac{1}{d_B^3 d_C}\right). \tag{E35}$$

■

Proposition 5. For the kth learning step ($k \leqslant n$) in the QSSM, the mean of the cost gradient is 0 and the variance of the cost gradient scales as $\mathrm{Var}(\partial_\mu C_k) \in O(2^{-n_k})$, where n_k is the circuit width of the kth learning step.

Proof. Suppose the target state is ρ and the input state for the kth learning step is $\hat{\sigma}$. We assume system A denotes the first $k-1$ qubits, system B denotes the kth qubit, and system C denotes the $(k+1)$th qubit to the $(k+n_k-1)$th qubit. With the definition of n_k claimed in the main text, there exists a purification $\hat{\rho}_{ABC}$ of ρ_A on system ABC. According to Lemma 8, we can easily know that

$$\mathrm{Var}(\partial_\mu C_k) \in O\left(\frac{1}{2^{n_k+2}}\right) = O(2^{-n_k}). \tag{E36}$$

■

Proposition 3 (main proposition restated). For an n-qubit target state ρ with fixed-order representation, we suppose its rank sequence is $\mathcal{R}_\rho = \{r_1, r_2, \ldots, r_{n-1}, r_n\}$. Then for learning the target state ρ with the QSSM, if the circuit used for each step is sufficiently random in forming a local 4-design, the expectation gradient for the kth step $\mathbb{E}(\partial_\mu C_k) = 0$ and the variance of the cost gradient scales with r_k as

$$\mathrm{Var}(\partial_\mu C_k) \in O\left(\frac{1}{r_k}\right). \tag{E37}$$

Proof. Since we know that $2^{n_k-1} \leqslant r_k \leqslant 2^{n_k}$, according to Propositions 4 and 5, we can get the proof. ■

[1] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature (London) **521**, 436 (2015).

[2] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Quantum machine learning, Nature (London) **549**, 195 (2017).

[3] M. Schuld, I. Sinayskiy, and F. Petruccione, An introduction to quantum machine learning, Contemp. Phys. **56**, 172 (2015).

[4] S. Lloyd, M. Mohseni, and P. Rebentrost, Quantum algorithms for supervised and unsupervised machine learning, arXiv:1307.0411.

[5] M. Schuld, I. Sinayskiy, and F. Petruccione, The quest for a quantum neural network, Quantum Inf. Process. **13**, 2567 (2014).

[6] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. E. Schapire, and L. Sellie, On the learnability of discrete distributions, in *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing* (ACM Press, New York, NY, 1994), pp. 273–282.

[7] G. Vidal, Efficient classical simulation of slightly entangled quantum computations, Phys. Rev. Lett. **91**, 147902 (2003).

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

[9] G. Li, R. Ye, X. Zhao, and X. Wang, Concentration of data encoding in parameterized quantum circuits, Adv. Neural Inf. Process. Syst. **35**, 19456 (2022).

[10] M. Cerezo, G. Verdon, H.-Y. Huang, L. Cincio, and P. J. Coles, Challenges and opportunities in quantum machine learning, Nat. Comput. Sci. **2**, 567 (2022).

[11] N. Hansen, D. V. Arnold, and A. Auger, Evolution strategies, in *Springer Handbook of Computational Intelligence*, edited by J. Kacprzyk and W. Pedrycz (Springer, Berlin, 2015), pp. 871–898.

[12] S. Kern, S. D. Müller, N. Hansen, D. Büche, J. Ocenasek, and P. Koumoutsakos, Learning probability distributions in continuous evolutionary algorithms—A comparative review, Nat. Comput. **3**, 77 (2004).

[13] E. Baum and F. Wilczek, Supervised learning of probability distributions by neural networks, in *Neural Information Processing Systems*, edited by D. Anderson (AIP Press, College Park, Maryland, 1987).

[14] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, A learning algorithm for boltzmann machines, Cognit. Sci. **9**, 147 (1985).

[15] L. Franceschi, M. Niepert, M. Pontil, and X. He, Learning discrete structures for graph neural networks, in *Proceedings of the 36th International Conference on Machine Learning* (PMLR, 2019), Vol. 97, pp. 1972–1982.

[16] E. Hoogeboom, D. Nielsen, P. Jaini, P. Forré, and M. Welling, Argmax flows and multinomial diffusion: Learning categorical distributions, Adv. Neural Inf. Process. Syst. **34**, 12454 (2021).

[17] J. Preskill, Quantum computing in the NISQ era and beyond, Quantum **2**, 79 (2018).

[18] M. Cerezo, K. Sharma, A. Arrasmith, and P. J. Coles, Variational quantum state eigensolver, npj Quantum Inf. **8**, 113 (2022).

[19] A. N. Chowdhury, G. H. Low, and N. Wiebe, A variational quantum algorithm for preparing quantum Gibbs states, arXiv:2002.00055.

[20] S. Ghosh, T. Paterek, and T. C. H. Liew, Quantum neuromorphic platform for quantum state preparation, Phys. Rev. Lett. **123**, 260404 (2019).

[21] Y. Wang, G. Li, and X. Wang, Variational quantum Gibbs state preparation with a truncated Taylor series, Phys. Rev. Appl. **16**, 054035 (2021).

[22] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, A variational eigenvalue solver on a photonic quantum processor, Nat. Commun. **5**, 4213 (2014).

[23] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets, Nature (London) **549**, 242 (2017).

[24] S. Aaronson, Shadow tomography of quantum states, in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing* (ACM Press, New York, NY, 2018), pp. 325–338.

[25] H.-Y. Huang, Learning quantum states from their classical shadows, Nat. Rev. Phys. **4**, 81 (2022).

[26] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. **9**, 4812 (2018).

[27] M. Cerezo, M. Larocca, D. García-Martín, N. L. Diaz, P. Braccia, E. Fontana, M. S. Rudolph, P. Bermejo, A. Ijaz,

S. Thanasilp *et al.*, Does provable absence of barren plateaus imply classical simulability? Or, why we need to rethink variational quantum computing, arXiv:2312.09121.

[28] E. R. Anschuetz, Critical points in quantum generative models, in *International Conference on Learning Representations* (ICLR, 2022).

[29] E. R. Anschuetz and B. T. Kiani, Quantum variational algorithms are swamped with traps, Nat. Commun. **13**, 7760 (2022).

[30] H.-k. Zhang, C. Zhu, M. Jing, and X. Wang, Statistical analysis of quantum state learning process in quantum neural networks, in *Thirty-Seventh Conference on Neural Information Processing Systems* (NeurIPS, 2023).

[31] E. Campos, A. Nasrallah, and J. Biamonte, Abrupt transitions in variational quantum circuit training, Phys. Rev. A **103**, 032607 (2021).

[32] E. Campos, D. Rabinovich, V. Akshay, and J. Biamonte, Training saturation in layerwise quantum approximate optimization, Phys. Rev. A **104**, L030401 (2021).

[33] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M.-H. Yang, Diffusion models: A comprehensive survey of methods and applications, ACM Comput. Surv. **56**, 1 (2023).

[34] A. Skolik, J. R. McClean, M. Mohseni, P. van der Smagt, and M. Leib, Layerwise learning for quantum neural networks, Quantum Mach. Intell. **3**, 5 (2021).

[35] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2010).

[36] A. Barenco, A. Berthiaume, D. Deutsch, A. Ekert, R. Jozsa, and C. Macchiavello, Stabilization of quantum computations by symmetrization, SIAM J. Comput. **26**, 1541 (1997).

[37] R. Islam, R. Ma, P. M. Preiss, M. E. Tai, A. Lukin, M. Rispoli, and M. Greiner, Measuring entanglement entropy in a quantum many-body system, Nature (London) **528**, 77 (2015).

[38] N. M. Linke, S. Johri, C. Figgatt, K. A. Landsman, A. Y. Matsuura, and C. Monroe, Measuring the Rényi entropy of a two-site Fermi-Hubbard model on a trapped ion quantum computer, Phys. Rev. A **98**, 052334 (2018).

[39] M. Fanizza, M. Rosati, M. Skotiniotis, J. Calsamiglia, and V. Giovannetti, Beyond the swap test, optimal estimation of quantum state overlap, Phys. Rev. Lett. **124**, 060503 (2020).

[40] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, Evaluating analytic gradients on quantum hardware, Phys. Rev. A **99**, 032331 (2019).

[41] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, Phys. Rev. A **98**, 032309 (2018).

[42] M. Ostaszewski, E. Grant, and M. Benedetti, Structure optimization for parameterized quantum circuits, Quantum **5**, 391 (2021).

[43] D. Kingma and J. Ba, Adam: A method for stochastic optimization, in *International Conference on Learning Representations (ICLR)* (ICLR, 2015).

[44] M. J. D. Powell, A direct search optimization method that models the objective and constraint functions by linear interpolation, in *Advances in Optimization and Numerical Analysis*, edited by S. Gomez and J.-P. Hennart (Springer Netherlands, Dordrecht, 1994), pp. 51–67.

[45] D. M. Greenberger, M. A. Horne, and A. Zeilinger, Going beyond Bell's theorem, in *Bell's Theorem, Quantum Theory and Conceptions of the Universe*, edited by M. Kafatos (Springer Netherlands, Dordrecht, 1989), pp. 69–72.

[46] C. Dankert, R. Cleve, J. Emerson, and E. Livine, Exact and approximate unitary 2-designs and their application to fidelity estimation, Phys. Rev. A **80**, 012304 (2009).

[47] D. Perez-Garcia, F. Verstraete, M. M. Wolf, and J. I. Cirac, Matrix product state representations, Quantum Inf. Comput. **7**, 401 (2007).

[48] M. Yoganathan, A condition under which classical simulability implies efficient state learnability, arXiv:1907.08163.

[49] O. Landon-Cardinal, Y.-K. Liu, and D. Poulin, Efficient direct tomography for matrix product states, arXiv:1002.4632.

[50] A. Anshu and S. Arunachalam, A survey on the complexity of learning quantum states, Nat. Rev. Phys. **6**, 59 (2024).

[51] M. Cramer, M. B. Plenio, S. T. Flammia, R. Somma, D. Gross, S. D. Bartlett, O. Landon-Cardinal, D. Poulin, and Y.-K. Liu, Efficient quantum state tomography, Nat. Commun. **1**, 149 (2010).

[52] D. Gross, S. T. Flammia, and J. Eisert, Most quantum states are too entangled to be useful as computational resources, Phys. Rev. Lett. **102**, 190501 (2009).

[53] S. Lloyd, M. Mohseni, and P. Rebentrost, Quantum principal component analysis, Nat. Phys. **10**, 631 (2014).

[54] M. Takahashi, One-dimensional Heisenberg model at finite temperature, Prog. Theor. Phys. **46**, 401 (1971).

[55] M. Schuld, Supervised quantum machine learning models are kernel methods, arXiv:2101.11020.

[56] Qiskit contributors, *Qiskit: An Open-Source Framework for Quantum Computing* (Zenodo, Geneva, 2023).

[57] K. Georgopoulos, C. Emary, and P. Zuliani, Modeling and simulating the noisy behavior of near-term quantum computers, Phys. Rev. A **104**, 062432 (2021).

[58] *Advances in Optimization and Numerical Analysis*, edited by S. Gomez and J.-P. Hennart, Mathematics and Its Applications Vol. 275 (Springer, Dordrecht, 1994).

[59] X. Liu, G. Liu, H.-K. Zhang, J. Huang, and X. Wang, Mitigating barren plateaus of variational quantum eigensolvers, IEEE Trans. Quantum Eng. (2024).

[60] N. Gisin and H. Bechmann-Pasquinucci, Bell inequality, Bell states and maximally entangled states for n qubits, Phys. Lett. A **246**, 1 (1998).

[61] J. Rivera-Dean, P. Huembeli, A. Acín, and J. Bowles, Avoiding local minima in variational quantum algorithms with neural networks, arXiv:2104.02955.

[62] D. Faílde, J. D. Viqueira, M. Mussa Juane, and A. Gómez, Using differential evolution to avoid local minima in variational quantum algorithms, Sci. Rep. **13**, 16230 (2023).

[63] G. Tóth, C. S. Lent, P. D. Tougaw, Y. Brazhnik, W. Weng, W. Porod, R.-W. Liu, and Y.-F. Huang, Quantum cellular neural networks, Superlattices Microstruct. **20**, 473 (1996).

[64] P. Rebentrost, T. R. Bromley, C. Weedbrook, and S. Lloyd, Quantum Hopfield neural network, Phys. Rev. A **98**, 042308 (2018).

[65] J. Zhao, Y.-H. Zhang, C.-P. Shao, Y.-C. Wu, G.-C. Guo, and G.-P. Guo, Building quantum neural networks based on a swap test, Phys. Rev. A **100**, 012334 (2019).

[66] C.-Y. Liu, C. Chen, C.-T. Chang, and L.-M. Shih, Single-hidden-layer feed-forward quantum neural network based on grover learning, Neural Networks **45**, 144 (2013).

[67] I. Cong, S. Choi, and M. D. Lukin, Quantum convolutional neural networks, Nat. Phys. **15**, 1273 (2019).

[68] N. Killoran, T. R. Bromley, J. M. Arrazola, M. Schuld, N. Quesada, and S. Lloyd, Continuous-variable quantum neural networks, Phys. Rev. Res. **1**, 033063 (2019).

[69] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Cost function dependent barren plateaus in shallow parametrized quantum circuits, Nat. Commun. **12**, 1791 (2021).

[70] J. Liu, K. Najafi, K. Sharma, F. Tacchino, L. Jiang, and A. Mezzacapo, An analytic theory for the dynamics of wide quantum neural networks, Phys. Rev. Lett. **130**, 150601 (2023).

[71] E. Grant, L. Wossnig, M. Ostaszewski, and M. Benedetti, An initialization strategy for addressing barren plateaus in parametrized quantum circuits, Quantum **3**, 214 (2019).

[72] A. Kulshrestha and I. Safro, Beinit: Avoiding barren plateaus in variational quantum algorithms, in *Proceedings of the 2022 IEEE International Conference on Quantum Computing and Engineering (QCE)* (IEEE, Piscataway, NJ, 2022), pp. 197–203.

[73] H. R. Grimsley, S. E. Economou, E. Barnes, and N. J. Mayhall, An adaptive variational algorithm for exact molecular simulations on a quantum computer, Nat. Commun. **10**, 3007 (2019).

[74] F. Zhang, N. Gomes, Y. Yao, P. P. Orth, and T. Iadecola, Adaptive variational quantum eigensolvers for highly excited states, Phys. Rev. B **104**, 075159 (2021).

[75] H. R. Grimsley, G. S. Barron, E. Barnes, S. E. Economou, and N. J. Mayhall, Adaptive, problem-tailored variational quantum eigensolver mitigates rough parameter landscapes and barren plateaus, npj Quantum Inf. **9**, 19 (2023).

[76] T. Volkoff and P. J. Coles, Large gradients via correlation in random parameterized quantum circuits, Quantum Sci. Technol. **6**, 025008 (2021).

[77] L. Friedrich and J. Maziero, Avoiding barren plateaus with classical deep neural networks, Phys. Rev. A **106**, 042433 (2022).

[78] M. Kieferova, O. M. Carlos, and N. Wiebe, Quantum generative training using Rényi divergences, arXiv:2106.09567.

[79] M. Fukuda, R. König, and I. Nechita, RTNI—A symbolic integrator for Haar-random tensor networks, J. Phys. A: Math. Theor. **52**, 425303 (2019).