

# Trainability Enhancement of Parameterized Quantum Circuits via Reduced-Domain Parameter Initialization

Yabo Wang and Bo Qi\*

*Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science,  
Chinese Academy of Sciences, Beijing 100190, P. R. China and  
University of Chinese Academy of Sciences, Beijing 100049, P. R. China*

Chris Ferrie

*Center for Quantum Software and Information, University of Technology Sydney, Ultimo NSW 2007, Australia*

Daoyi Dong

*CIICADA Lab, School of Engineering, Australian National University, Canberra ACT 2601, Australia  
(Dated: November 6, 2024)*

Parameterized quantum circuits (PQCs) have been widely used as a machine learning model to explore the potential of achieving quantum advantages for various tasks. However, training PQCs is notoriously challenging owing to the phenomenon of plateaus and/or the existence of (exponentially) many spurious local minima. To enhance trainability, in this work we propose an efficient parameter initialization strategy with theoretical guarantees. We prove that by reducing the initial domain of each parameter inversely proportional to the square root of circuit depth, the magnitude of the cost gradient decays at most polynomially with respect to qubit count and circuit depth. Our theoretical results are substantiated through numerical simulations of variational quantum eigensolver tasks. Moreover, we demonstrate that the reduced-domain initialization strategy can protect specific quantum neural networks from exponentially many spurious local minima. Our results highlight the significance of an appropriate parameter initialization strategy, offering insights to enhance the trainability and convergence of variational quantum algorithms.

## I. INTRODUCTION

Variational quantum algorithms (VQAs) with parameterized quantum circuits (PQCs) [1–5] have emerged as a leading strategy to explore the power of quantum computing in the current noisy intermediate-scale quantum era. VQAs are optimization-based quantum-classical hybrid algorithms. The core idea is exploiting PQCs to delegate classically difficult computation and updating the variational parameters in PQCs with a classical optimizer [6–9]. VQAs have been applied across various practical fields, including machine learning [10–14], error correction [15–17], dynamical simulations [18–22], combinatorial optimization [23–26], among others [27–29].

Despite the success in many fields, training VQAs still faces many challenges when scaling up the size of PQCs. One of the key bottlenecks is the magnitude of the cost gradient decaying exponentially with the number of qubits, also known as barren plateaus (BPs) [30]. Once BPs occur, training PQCs is prohibited unless exponentially many copies of quantum states are provided to identify the update direction through measurement. BPs have been found arising in many scenarios, such as randomly initialized PQCs [30, 31], deep and problem-agnostic PQCs [30, 32], global cost function [33, 34], large expressibility [30, 35], entanglement [36–38], and noise [39]. In addition to BPs, there

is another bottleneck for training. For a wide class of under-parameterized variational quantum models, the landscapes of the cost function are generally swamped with spurious local minima and traps [40]. These quantities may be exponentially many in relation to the number of parameters [41], making it extremely hard to find a global minimum [42–44].

To address the trainability issue, various strategies have been proposed. For instance, certain constructed PQCs, including tree tensor network [45, 46] and quantum convolutional neural network [47], can help avoid BPs [32, 48–50]. However, due to their specific architectures, where the circuit depth is logarithmic with respect to the number of qubits, the gradient usually decays exponentially with the circuit depth instead. Ansatzes constructed in a problem-inspired manner [37, 51–53] have been put forward to mitigate BPs by constraining the explored space of unitaries. Ansatzes can also be trained layer-by-layer [54, 55], or designed using classical reinforcement learning [56]. Since BPs are cost function-dependent, it was proposed to encode the problem into local cost function [33]. In addition, over-parametrization has been investigated to address the training issue caused by the presence of many spurious local minima or traps in the cost landscapes far away from the global minimum [57–59]. However, efficient over-parametrization, i.e., with polynomially many parameters, is achievable only under some conditions for quantum neural networks (QNNs) [60]. Generally, the number of parameters required for over-parametrization is comparable to the system dimension [58]. This

\* qibo@amss.ac.cn

exponential overhead may incur BPs, especially when considering noises on quantum hardware.

From a practical viewpoint, an effective parameter initialization strategy is crucial for the trainability of PQCs and the convergence of VQAs [30, 35, 42, 52, 61]. An identity block initialization strategy has been introduced with the purpose of limiting the effective depth of circuits [62]. It was demonstrated that ansatzes with correlated parameters can effectively restrict the optimized parameter space and result in large gradients of cost function [63]. A Bayesian learning initialization was employed to identify a promising region in the parameter space before performing local search [64]. Simulations conducted in [31] revealed an initial region of parameters that can improve the trainability while maintaining a high level of expressibility. A Gaussian initialization was proposed with theoretical guarantees to escape from BPs in general deep circuits [65], where the lower bound of the norm of cost gradient decays at most polynomially with the circuit depth. Besides, for correlated parameterized gates, Gaussian initialization guarantees the update direction when training global observables [65]. However, the theoretical or empirical results of the aforementioned strategies are confined to single Pauli strings. For complex objectives corresponding to a summation of local Pauli strings, a Floquet initialization strategy was proposed to mitigate BPs [66]. They demonstrated that by initializing PQCs within the many-body localization phase, the gradient does not vanish as the system size grows. Nevertheless, the phase transitions between the many-body localization and thermalization (where BPs occur) in Floquet systems need to be characterized through additional simulations.

In this work, we propose an efficient reduced-domain initialization strategy to enhance trainability of PQCs. By focusing on the variational quantum eigensolver (VQE), we theoretically prove that for an  $L$ -depth circuit, if the initial domain of each parameter is reduced in proportion to  $\frac{1}{\sqrt{L}}$ , the norm of the cost gradient decays at most polynomially with  $L$  as the number of qubits and the depth increase. Our result is quite different from the belief that BPs occur when the circuit depth is polynomial of the number of qubits. This is because the reduced-domain initialization can help maintain a balance between the trainability and expressibility of PQCs. Via simulations, we first validate our theoretical results. Then we show that our reduced-domain initialization can protect QNNs from the phenomenon of exponentially many spurious local minima in cost function landscapes, and help achieve improved performance.

The paper is organized as follows. In Section II, we describe the generic setup of VQAs. In Section III, we present the theoretical results of enhancing trainability of PQCs through the reduced-domain parameter initialization strategy. In Section IV, via simulations we demonstrate the enhanced trainability in VQE tasks, as well as the improved convergence in both VQE and QNN instances. Section V concludes the paper.

## II. SETUP of VQAS

VQAs provide a general framework for solving a wide range of problems. A typical VQA is comprised of basic elements: cost, ansatz, measurement, optimizer as well as parameter initialization.

**Cost:** The first task of VQAs is to identify an appropriate cost function which effectively encodes the problem to be solved. Without loss of generality, we consider the expectation value of an observable  $O$  as cost function in the form of

$$C(\theta) = \text{Tr} [OU(\theta)\rho U^\dagger(\theta)]. \quad (1)$$

Here,  $\rho$  is the input state, and  $U(\theta)$  denotes a parameterized quantum unitary with the parameter vector  $\theta$  to be optimized. In VQAs, an implicit assumption is that the cost function should be classically computationally intensive and the value can be obtained by performing quantum measurements on a quantum circuit with possible classical postprocessing.

In this work, our aim is finding the least eigenvalue of a given Hamiltonian  $H$ . In the following, given a positive integer  $n$ , denote by  $[n]$  the set  $\{1, 2, \dots, n\}$ . Let

$$\sigma_{\mathbf{i}} = \sigma_{(i_1, i_2, \dots, i_N)} = \sigma_{i_1} \otimes \sigma_{i_2} \otimes \dots \otimes \sigma_{i_N} \quad (2)$$

denote an  $N$ -qubit Pauli string, where  $i_j \in \{0, 1, 2, 3\}$  for  $j \in [N]$ , and  $\sigma_0 = I$ ,  $\sigma_1 = X$ ,  $\sigma_2 = Y$ ,  $\sigma_3 = Z$  are Pauli matrices. For instance, we denote

$$\begin{aligned} \sigma_{(3,0,0,\dots,0)} &= Z \otimes I \otimes I \otimes \dots \otimes I = Z_1, \\ \sigma_{(3,3,0,\dots,0)} &= Z \otimes Z \otimes I \otimes \dots \otimes I = Z_1 Z_2. \end{aligned}$$

Moreover,  $\sigma_{\mathbf{i}}$  is said to be  $S$ -local if there are  $S$  non-identity components in the tensor product of Eq. (2).

We choose  $\rho = |0\rangle\langle 0|$  and set  $O = H$  in Eq. (1). We consider  $H$  in the form as

$$H = \sum_{\mathbf{i} \in \mathcal{N}} \sigma_{\mathbf{i}} = \sum_{\mathbf{i} \in \mathcal{N}} \sigma_{(i_1, i_2, \dots, i_N)}, \quad (3)$$

where  $\mathcal{N}$  is a subset of  $\mathcal{N}_S = \{\mathbf{i} = (i_1, i_2, \dots, i_N) : 1 \leq \sum_{j=1}^N 1_{i_j \neq 0} \leq S\}$ , and  $1_\omega$  is the indicator function of the event  $\omega$ . That is,  $H$  is a sum of  $N$ -qubit Pauli strings with locality of at most  $S$ . This model encompasses ground state determination problems for various quantum many-body systems and chemical molecules [67–70]. In addition, the Hamiltonian Eq. (3) can be employed to tackle many NP-complete problems [71].

**Ansatz:** Another key element of VQAs is ansatz, i.e., PQC whose parameters are optimized to minimize the corresponding cost function. In this work, we consider a PQC similar to that in [65], called the hardware-efficient ansatz (HEA). As illustrated in Fig. 1(a), it is comprised of  $L$  blocks, each of which is composed of an entanglement layer  $\text{CZ}_l$  including CZ gates on arbitrarily many qubit pairs and two layers of  $R_X$  and  $R_Y$  gates on all qubits. There are two widely used structures of

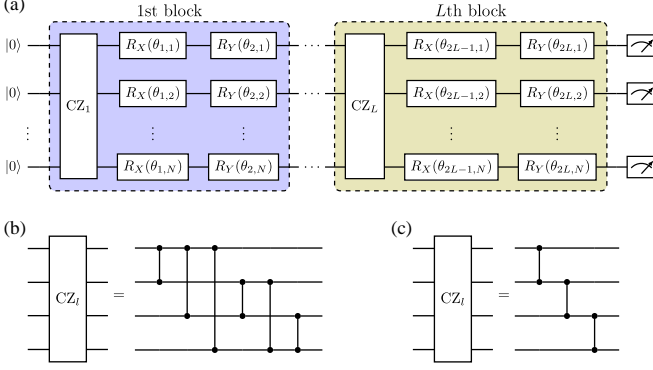


FIG. 1. Setup of HEA. (a) PQC for VQE. It consists of  $L$  blocks. In the  $l$ th block, we first perform the entanglement layer composed of  $CZ$  gates, and then successively perform  $R_X$  gate and  $R_Y$  gate on each qubit. (b)  $CZ_l$  layer with the fully connected topology. (c)  $CZ_l$  layer with the nearest-neighbor pairs topology and open boundary condition.

the entanglement layer: the fully connected topology as shown in Fig. 1(b), and the nearest-neighbor pairs topology with open boundary condition as shown in Fig. 1(c). Notably, our results do not rely on the specific topology of  $CZ_l$  and can also be generalized to other structures of PQC.

**Measurement:** In our VQE task, the Hamiltonian is formulated in terms of Pauli strings. The expectation values of the  $\sigma_3$  terms can be directly measured by reading out the probabilities of the computational basis, while to obtain the expectation values of the  $\sigma_1$  and  $\sigma_2$  terms, we should rotate them into the  $\sigma_3$  basis first. To minimize the number of measurements needed, we should group different Pauli strings that can be measured simultaneously. There is still an urgent need for more efficient compilation of measurement circuits.

**Optimizer:** The core task of the classical processor is updating the parameters to minimize the cost. In our VQE task, the cost gradient can be obtained by utilizing the parameter-shift rule [11], which also suffers from the measurement efficiency issue. In this work, we aim to analyze the norm of the gradient which greatly impacts the trainability of VQAs.

**Parameter initialization:** Due to the possibility of occurrence of BPs and/or the existence of exponentially many spurious local minima when training VQAs, it is critical to find an appropriate parameter initialization strategy to enhance trainability.

For the PQC illustrated in Fig. 1(a), there are  $2L$  layers of parameterized rotation gates with  $2NL$  variational parameters in total, denoted by

$$\theta = \theta_{[2L]} = (\theta_1, \theta_2, \dots, \theta_{2L}),$$

where  $\theta_q = (\theta_{q,1}, \theta_{q,2}, \dots, \theta_{q,N})$  denotes the vector of rotation angles in the  $q$ th layer for  $q \in [2L]$ . To be specific, for the  $l$ th block,  $l \in [L]$ , and the  $j$ th qubit,  $j \in [N]$ ,  $\theta_{2l-1,j}$  denotes the rotation angle around  $X$ , while  $\theta_{2l,j}$  denotes that around  $Y$ .

### III. TRAINABILITY ENHANCEMENT

#### A. Lower bounds of gradient

In previous works, the initial parameters are usually drawn independently and identically from a uniform distribution over an entire period  $[-\pi, \pi]$ , and the  $N$ -qubit local observable is always chosen as a simple Pauli string, for instance,  $Z_1 Z_2$ . Under a PQC similar to ours, Ref. [30] demonstrated that a BP occurs even when the depth  $L$  is a modest linear function, that is,  $L = 10N$ , and the exponential decay cannot be circumvented for any polynomial sum of these local operators. For the simple observable  $Z_1 Z_2$ , Ref. [31] numerically showed that the trainability can be improved if each parameter is randomly chosen from  $[-a\pi, a\pi]$ . The satisfying value of  $a$  was identified by searching from  $10^{-10}$  to 1 via simulations.

In this work, we consider a general Hamiltonian in the form of Eq. (3), where the number of individual terms in the summation can be arbitrary. For the complex objective function, we propose a reduced-domain parameter initialization strategy, in which each parameter is independently and uniformly chosen from a reduced interval  $[-a\pi, a\pi]$ , where the hyperparameter  $a$  is depth-dependent and will be given in the following theorems. Importantly, there is no constraint on the parameters' domain during the optimization process. That is, the variational parameters are free to leave their initial reduced-domains as guided by the negative gradient to minimize the objective function.

In view of the setup in Section II, the decrease of the cost function value after an update of the parameters can be quantified in terms of the magnitude of  $\|\nabla_{\theta} C\|_2^2$  (see Appendix A). Thus, we adopt its average value  $\mathbb{E}_{\theta} \|\nabla_{\theta} C\|_2^2$  as the trainability figure of merit.

Now we provide a theoretical guarantee for the reduced-domain initialization by establishing a lower bound on the norm of cost gradient at the first step. For the ease of notation, denote  $\alpha = \frac{1}{4} (2 + \frac{\sin 2a\pi}{a\pi})$ ,  $\beta = \frac{1}{4} (2 - \frac{\sin 2a\pi}{a\pi})$ , and  $\gamma = \frac{\sin a\pi}{a\pi}$ .

**Theorem 1.** Consider the VQE problem, where the cost function is Eq. (1) with the Hamiltonian being Eq. (3), the ansatz illustrated as in Fig. 1(a), the initial state chosen as  $\rho = |0\rangle\langle 0|$ , and each parameter in  $\theta$  chosen from  $[-a\pi, a\pi]$  independently and uniformly. Let  $\nabla_{\theta} C$  denote the gradient of the cost function with respect to  $\theta$ . Then the average of the squared norm of the gradient is lower bounded by

$$\mathbb{E}_{\theta} \|\nabla_{\theta} C\|_2^2 \geq 2|\mathcal{N}|(L-1)\alpha^{2SL-1}\beta^{S+1}, \quad (4)$$

where  $|\mathcal{N}|$  denotes the cardinality of the set  $\mathcal{N}$ .

Furthermore, if the hyperparameter  $a$  is selected as the solution of the following equation

$$\frac{\sin 2a\pi}{2a\pi} = \frac{S(2L-1)-2}{S(2L+1)}, \quad (5)$$

where  $a \in \Theta\left(\frac{1}{\sqrt{L}}\right)$ , then the gradient norm can be lower bounded by

$$\mathbb{E}_{\theta} \|\nabla_{\theta} C\|_2^2 \geq 2|\mathcal{N}| \left(\frac{S+1}{eS}\right)^{S+1} \frac{L-1}{(2L+1)^{S+1}}. \quad (6)$$

From Theorem 1, for VQE problems with Hamiltonian being a summation of a series of local Pauli strings, the trainability can be greatly enhanced in the initial phase by appropriately reducing the initial domain of parameters. Specifically, the lower bound of the gradient norm decays at most polynomially with the circuit depth. In addition, the lower bound increases linearly with the number of individual Pauli strings in the Hamiltonian, which guarantees the efficiency of the reduced-domain strategy for complex objective functions.

An intuitive interpretation of the effectiveness of the reduced-domain initialization is as follows. It is well known that there is a conflict between the trainability and expressibility of PQC. Over-expressed PQCs may suffer from phenomena like BPs and/or the landscape being swamped with spurious local minima, making the training extremely difficult. Note that the expressibility of a PQC depends on the structure of the circuit as well as the size of the initial domains of the parameters. It generally grows fast as the circuit depth  $L$  increases and the range of the initial domains expands. Therefore, to balance the conflict between trainability and expressibility of PQCs, we should appropriately reduce the initial domains of the parameters. As we have stated, to enhance the trainability of a PQC with depth  $L$ , the domain of each parameter should be reduced in proportion to  $\frac{1}{\sqrt{L}}$ . We will provide more detailed sensitivity analysis on the hyperparameter  $a$  in Sec. IV; see Table I, Fig. 3(a), and Fig. 5(a).

To further illustrate the result and explain the idea of proof of Theorem 1, let us consider an example where the  $N$ -qubit Hamiltonian reads

$$H = \sum_{j=1}^{N-1} Z_j Z_{j+1}. \quad (7)$$

**Lemma 2.** Consider the VQE problem, where the cost function is Eq. (1) with the Hamiltonian being Eq. (7), the ansatz illustrated as in Fig. 1(a), the initial state chosen as  $\rho = |0\rangle\langle 0|$ , and each parameter in  $\theta$  chosen from  $[-a\pi, a\pi]$  independently and uniformly. Let  $\nabla_{\theta} C$  denote the gradient of the cost function with respect to  $\theta$ . Then the average of the squared norm of the gradient is lower bounded by

$$\mathbb{E}_{\theta} \|\nabla_{\theta} C\|_2^2 \geq 4(2N-3)L\gamma^{8L-2}\beta. \quad (8)$$

Furthermore, if the hyperparameter  $a$  is selected as

$$a = \frac{1}{4\pi} \sqrt{\frac{40L+7-\sqrt{1600L^2-400L+49}}{L}} \in \Theta\left(\frac{1}{\sqrt{L}}\right), \quad (9)$$

then the gradient norm can be lower bounded by

$$\mathbb{E}_{\theta} \|\nabla_{\theta} C\|_2^2 \geq e^{-1}(2N-3), \quad (10)$$

where  $e$  is the natural constant.

It is clear that for the Hamiltonian of Eq. (7), under the reduced-domain initialization, the trainability can be remarkably enhanced in the initial phase. Not only the lower bound of the gradient norm does not decrease when the circuit depth increases, but also it increases as the number of qubits increases. In particular, for the simple  $N$ -qubit observable  $Z_1 Z_2$ , the gradient norm can be lower bounded by a constant under the reduced-domain initialization. This is quite different from the result in Ref. [30] with the domain being  $[-\pi, \pi]$ , where the gradient vanishes exponentially with respect to both the qubit number and the circuit depth.

The proofs of Lemma 2 and Theorem 1 are given in Appendix C and Appendix D, respectively. The basic idea is inspired by the proof of Theorem 4.1 in [65], where the observable considered is restricted to a single local Pauli string. However, in our work, the observables are in the form of a summation of local Pauli strings, so that additionally we need to deal with the cross terms generated from different Pauli strings when analyzing the norm of gradient. Specifically, we prove that under the selected initial state  $\rho = |0\rangle\langle 0|$ , these cross terms are non-negative by analyzing the transformations of the involved Pauli strings after taking the expectations over the trainable parameters and applying the entanglement layers.

## B. Addressing BPs

We now prove that the reduced-domain initialization strategy can help escape from BPs by considering the variance of the partial derivative of the cost function with respect to specific variational parameters.

In general, an  $N$ -qubit cost function  $C$  is said to exhibit a BP in the  $\theta_v$  direction, if  $\mathbb{E}_{\theta} \partial_{\theta_v} C = 0$ , and the variance decreases exponentially with  $N$ , that is,

$$\mathbb{V}_{\theta} \partial_{\theta_v} C \in \mathcal{O}(b^N), \text{ for some } b \in (0, 1).$$

On the other hand, it is said to be trainable, if the variance decreases polynomially with  $N$ , that is,

$$\mathbb{V}_{\theta} \partial_{\theta_v} C \in \Omega\left(\frac{1}{\text{poly}(N)}\right).$$

Similarly, we can define the circuit-depth-dependent BP, as the depth is also a key figure of merit in VQAs. The number of qubits and the circuit depth are unified as the size of PQC.

Recall that once a BP occurs, the training of PQCs is prohibited. The following theorem states that the



reduced-domain initialization strategy with a selected hyperparameter in the scaling of  $\frac{1}{\sqrt{L}}$  can help PQCs avoid BPs. Remarkably, it holds for PQCs whose depths are polynomial functions of the number of qubits, while Ref. [30] demonstrated that for a wide class of randomly initialized PQCs of reasonably linear depth, BPs occur even with simple local observables.

**Theorem 3.** *Under the same assumptions as in Theorem 1, let  $\partial_{\theta_{q,n}} C$  denote the partial derivative of cost function with respect to  $\theta_{q,n}$ . For  $q \in [2L-2]$  and  $n \in [N]$ , the average of the squared norm of the partial derivative is lower bounded by*

$$\mathbb{E}_{\theta} (\partial_{\theta_{q,n}} C)^2 \geq \alpha^{2SL-1} \beta^{S+1}. \quad (11)$$

*In addition, if for each  $i \in \mathcal{N}$ , the number of its elements belonging to the set  $\{1, 2\}$  is not equal to one, then we have*

$$\mathbb{E}_{\theta} \partial_{\theta_{q,n}} C = 0, \quad (12)$$

*and the variance is lower bounded by*

$$\mathbb{V}_{\theta} \partial_{\theta_{q,n}} C \geq \alpha^{2SL-1} \beta^{S+1}. \quad (13)$$

*Furthermore, if the hyperparameter  $a$  is selected according to Eq. (5) in Theorem 1, where  $a \in \Theta\left(\frac{1}{\sqrt{L}}\right)$ , we have*

$$\mathbb{E}_{\theta} (\partial_{\theta_{q,n}} C)^2 \geq \left(\frac{S+1}{eS}\right)^{S+1} \frac{1}{(2L+1)^{S+1}}. \quad (14)$$

Similar to Lemma 2, to further illustrate the result and explain the idea of proof of Theorem 3, we present the following lemma on the basis of Hamiltonian Eq. (7).

**Lemma 4.** *Under the same assumptions as in Lemma 2, let  $\partial_{\theta_{q,n}} C$  denote the partial derivative of cost function with respect to  $\theta_{q,n}$ . For  $q \in [2L]$  and  $n \in [N]$ , we have*

$$\mathbb{E}_{\theta} \partial_{\theta_{q,n}} C = 0, \quad (15)$$

*and for  $q \in [2L]$ , the variance is lower bounded by*

$$\mathbb{V}_{\theta} \partial_{\theta_{q,n}} C \geq f_n(L, a). \quad (16)$$

(i) For  $n \in \{1, N\}$ ,

$$f_n(L, a) = \alpha^{4L-1} \beta. \quad (17)$$

*Furthermore, if the hyperparameter  $a$  is selected as the solution of the following equation*

$$\frac{\sin a\pi}{a\pi} = \frac{2L-1}{L}, \quad (18)$$

*where  $a \in \Theta\left(\frac{1}{\sqrt{L}}\right)$ , then we have*

$$\mathbb{V}_{\theta} \partial_{\theta_{q,n}} C \geq \frac{1}{4eL}. \quad (19)$$

(ii) For  $n \in \{2, 3, \dots, N-1\}$ ,

$$f_n(L, a) = 2\alpha^{4L-1} \beta + 2\gamma^{4L} \alpha^{2L-1} \beta. \quad (20)$$

*Furthermore, if the hyperparameter  $a$  is selected according to Eq. (9) in Lemma 2, where  $a \in \Theta\left(\frac{1}{\sqrt{L}}\right)$ , then we have*

$$\mathbb{V}_{\theta} \partial_{\theta_{q,n}} C \geq \frac{1}{eL}. \quad (21)$$

The proofs of Lemma 4 and Theorem 3 are provided in Appendix E and Appendix F, respectively.

## IV. NUMERICAL SIMULATIONS

In this section, we demonstrate the efficacy of the reduced-domain initialization strategy in enhancing trainability and convergence via VQE and QNN. Unless otherwise stated, we conduct classical simulations of demonstrations using the analytical form of the cost function and its gradient, simulated through tensor operations in TensorFlow [72] for VQE tasks and in PyTorch [73] for QNN tasks. Additionally, we utilize TensorCircuit [74], which supports just-in-time compilation to accelerate the simulation processes of VQE.

### A. VQE

For VQE tasks, we focus on the  $N$ -qubit Heisenberg model

$$H = \sum_{i=1}^{N-1} (X_i X_{i+1} + Y_i Y_{i+1} + Z_i Z_{i+1}), \quad (22)$$

where the locality  $S = 2$ . It becomes challenging to obtain the ground energy of  $H$  as  $N$  increases.

Unless otherwise stated, we adopt the HEA illustrated in Fig. 1(a) and fix the entanglement layer in each block arranged in the nearest-neighbor pairs topology with open boundary as shown in Fig. 1(c). For trainability and convergence, we compare five different initialization strategies, including zero-initialization by setting all parameters to be zero, uniform  $\mathcal{U}[-\pi, \pi]$ , Floquet  $\mathcal{U}[-W, W]$  with  $W = 0.4$  for intermediate-depth circuits and  $W = 0.2$  for deep circuits [66]<sup>1</sup>, Gaussian  $\mathcal{N}(0, \sigma^2)$

<sup>1</sup> In Ref. [66], the Floquet-initialized PQC incorporates  $X$ - and  $Y$ -type rotation gates with angle uniformly sampled from  $[-W, W]$ , and  $Z$ -type rotation gates with angle uniformly sampled from  $[-\pi, \pi]$ . In our work, the  $W$ s employed are consistent with those utilized in Ref. [66], given the similar settings. Specifically, in the HEA featuring only  $X$ - and  $Y$ -type rotation gates (see Fig. 1(a)), all the parameters are initialized from  $\mathcal{U}[-W, W]$ . In the Hamiltonian variational ansatz including all three types of gates (see Fig. 4), the parameters are initialized accordingly.

TABLE I. Gradient magnitudes of the 12-qubit Heisenberg model at the initial step across different initializations.

	reduced-domain-0.07	reduced-domain-0.007	reduced-domain-0.7	zero/ $\pi$ -initialization	uniform	Floquet	Gaussian
HEA	502.8	35.3	3.6	0.0/ N/A	3.5	236.1	266.9
HVA	4803.0	277.6	61.9	0.0/0.00037549	61.4	1322.1	1947.4

with the variance  $\sigma^2 = \frac{1}{8SL} = \frac{1}{16L}$  [65], and our reduced-domain  $\mathcal{U}[-a\pi, a\pi]$  with the hyperparameter  $a$  selected according to Eq. (5). Recall that after the initialization, the variational parameters are updated by the classical optimizer to minimize the cost function, and are not constrained to the initial domains.

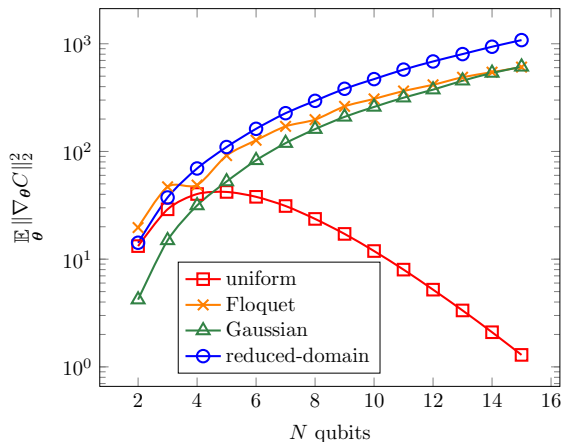


FIG. 2. Gradient magnitudes of the Heisenberg model at the initial step versus the number of qubits under five initialization strategies: zero-initialization (not displayed where the gradient values are consistently zero), uniform (square), Floquet (cross), Gaussian (triangle), and reduced-domain (circle). Here,  $L = 5N$ ,  $W = 0.2$ ,  $\sigma^2 = \frac{1}{8SL} = \frac{1}{80N}$ , and  $a$  is determined according to Eq. (5). There exists a BP under the uniform initialization, whereas the other three strategies help PQCs avoid BPs. Under the three advanced initialization strategies, the gradient norms increase exponentially with the qubit number, among which the reduced-domain exhibits the most favorable growth.

In Fig. 2, we compare the trainability of VQE under the five initialization strategies by evaluating the mean squared norm of the gradient at the first step. Here, we set  $L = 5N$  so that the HEA in Fig. 1(a) includes the same number of rotation layers  $10N$  as that in Figure 3 of Ref. [30]. We plot the average points over 1000 rounds of demonstrations on a semi-log plot. The gradient values under zero-initialization are always zero and we do not display them. From Fig. 2, under the uniform initialization, the magnitude of the cost gradient decays exponentially as the number of qubits  $N$  increases, which coincides with both the theoretical and empirical results concerning BPs in Ref. [30]. In contrast, under the other three advanced initializations, the magnitudes of the gradient increase exponentially with the number

of qubits. It is clear that the reduced-domain exhibits the most favorable growth among the five initialization strategies.

Next, we demonstrate the effect of initialization strategies on the convergence behavior of finding the ground state of the 12-qubit Heisenberg model Eq. (22). In our demonstrations, VQEs are optimized by Adam [75] with a learning rate multiplier schedule, which has been shown to substantially improve Adam’s performance [76]. Specifically, we utilize a linearly decreasing learning rate  $\eta$  ranging from 0.1 to 0.05 during the 500 training iterations. For each initialization strategy, we sample 1000 sets of initial parameters to evaluate the initial gradient, and sample 10 sets to track the optimization process, presenting the average performance. The filled area in the figures of the rest of this subsection represents the smallest area that encompasses the behaviors of these 10 samples. To ensure the expressibility of the HEA in Fig. 1(a), we set  $L = 3.5N = 42$ . Accordingly, the hyperparameter  $a^*$  selected according to Eq. (5) is 0.0736328.

We first conduct a sensitivity analysis on the hyperparameter  $a$  by comparing the performance under three different values of  $a$ . They are in different orders of magnitudes from  $a^* = 0.0736328$ , namely,  $a = 0.007$ ,  $a = 0.07$ , and  $a = 0.7$ . We list the mean squared norm of the gradient at the first step for these different values of  $a$  in Table I. We find that  $a = 0.07$ , having the same order with  $a^*$ , provides the most favorable trainability. We illustrate the convergence behavior of cost values under different values of  $a$  in Fig. 3(a). We find that the soft reduced-domain with  $a = 0.7$  does not outperform the uniform initialization (see Fig. 3(b)), while the hard one with  $a = 0.007$  is less effective than the appropriate one with  $a = 0.07$ . Among the four reduced-domain initializations, the one with  $a^*$  has the best convergence behavior, which supports the validity of our theoretical results.

We then compare the reduced-domain strategy with  $a^*$  against the other four strategies. We list the mean squared norm of the gradient at the first step of the other four initializations in Table I. We find that the magnitude of the initial gradient under the reduced-domain with  $a = 0.07$  is significantly larger than those under the other initializations. We depict their convergence behaviors of cost values in Fig. 3(b). It is clear that the zero-initialization fails to optimize due to zero gradients. Under the uniform initialization, the optimization quickly gets stuck far away from the global minimum. In contrast, under the reduced-domain,

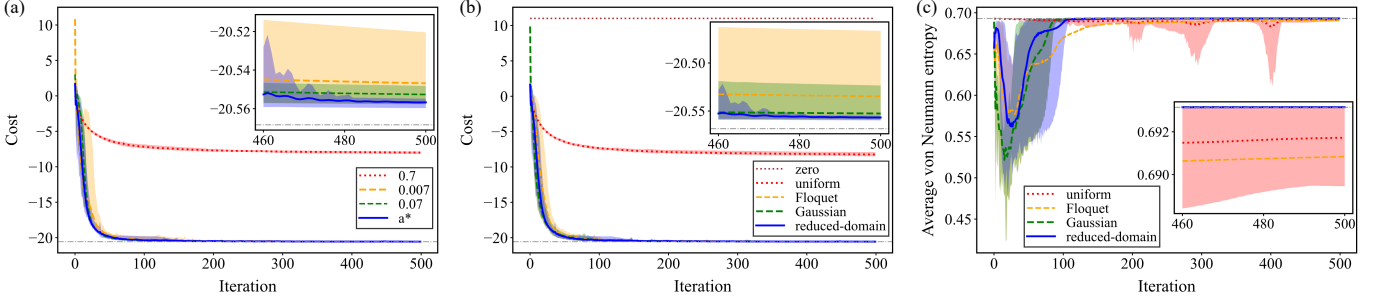


FIG. 3. Convergence behavior of the 12-qubit Heisenberg model under different initializations in the HEA context. The dash-dotted grey line corresponds to that of the global minimum. Here,  $L = 3.5N = 42$ . (a) Convergence of the cost values under different hyperparameters of the reduced-domain initialization:  $a^* \approx 0.0736328$  (solid blue),  $a = 0.7$  (dotted red),  $a = 0.07$  (thin dashed green), and  $a = 0.007$  (dashed orange). (b) Convergence of the cost values under different initializations. (c) Convergence of the average von Neumann entropy under different initializations. Here, zero-initialization: dotted dark red, uniform: dotted red, Floquet: thin dashed orange, Gaussian: dashed green, and reduced-domain: solid blue. The corresponding hyperparameters are  $W = 0.4$ ,  $\sigma^2 = \frac{1}{85L} = \frac{1}{672} \approx 0.001488$ , and  $a^* = 0.0736328$ , respectively.

Floquet, and Gaussian, large gradients at the initial stage can swiftly steer the cost to the vicinity of the global minimum. At the end (see the inset of Fig. 3(b)), the reduced-domain strategy is most robust across random parametrizations and achieves the most favorable convergence value, followed by the Gaussian, and then the Floquet.

We also track the dynamics of quantum correlations throughout the optimization process by calculating the average von Neumann entropy, defined as

$$S(\rho) \triangleq -\frac{1}{N} \sum_{i=1}^N \text{Tr} [\rho^i \log(\rho^i)],$$

where  $\rho^i = \text{Tr}_{\bar{i}}[\rho]$  denotes the reduced state of  $\rho$  on the  $i$ th qubit, obtained by tracing out all the other qubits  $\bar{i}$ . It helps to evaluate the efficiency of the PQC in capturing the entanglement of the target ground state. As depicted in Fig. 3(c), due to the imbalance between the expressibility and trainability, the primitive uniform initialization has only limited capability to adjust quantum correlations, and does not allow the optimizer to explore the cost landscape far away from the starting point. Under zero-initialization, the von Neumann entropy remains zero (not plotted), which is consistent with the cost values remaining unchanged in Fig. 3(b). In contrast, the other three advanced initialization strategies exhibit capability for adjusting entanglement, even in the hardware-efficient circuits. Although these advanced strategies restrict the initial domain, they effectively allow the optimizer to explore a broad range of the cost landscape and approach the global optimum. Notably, the reduced-domain and Gaussian initializations are more effective in achieving quantum correlations in the target ground state, as indicated by the metric of average von Neumann entropy. The efficiency of the Floquet one can be improved through a careful selection of initial trial states, whose entanglement structures closely align with that of the

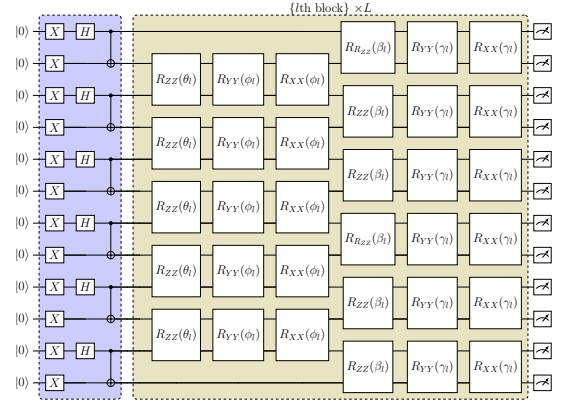


FIG. 4. Setup of  $L$ -block HVA for the 12-qubit Heisenberg model. Here,  $L = 2N = 24$ . The purple shaded consists of Pauli- $X$  gates, Hadamard gates and CX gates. It denotes the preparation of the initial reference state  $\otimes^6 |\Psi^-\rangle$  with  $|\Psi^-\rangle = (1/\sqrt{2})(01) - |10\rangle$ . The two-qubit parameterized rotation gates are  $R_{\Gamma\Gamma}(x) = \exp\{-i\frac{x}{2}\Gamma \otimes \Gamma\}$  with  $\Gamma \in \{XX, YY, ZZ\}$ , and  $x \in \{\theta_l, \phi_l, \beta_l, \gamma_l\}$ .

target ground state [66].

We further investigate the effectiveness of the advanced initialization strategies on a problem-specific ansatz known as the Hamiltonian variational ansatz (HVA) [52]. In contrast to the HEA, the PQC of HVA is designed on the basis of the Hamiltonian of interest. For the 12-qubit Heisenberg model, we illustrate the PQC in Fig. 4. Note that the initial reference state plays a crucial role in HVA. In Fig. 4, the initial reference state is  $\otimes^6 |\Psi^-\rangle$  with  $|\Psi^-\rangle = (1/\sqrt{2})(01) - |10\rangle$ . It is suggested in Ref. [52] that a suitable initialization strategy is crucial for circumventing BPs and successfully finding the ground state. In particular, they employed a simpler version of identity-block initialization than that in Ref. [62] by setting all variational parameters to be  $\pi$ , which we refer to as  $\pi$ -initialization in our paper. It

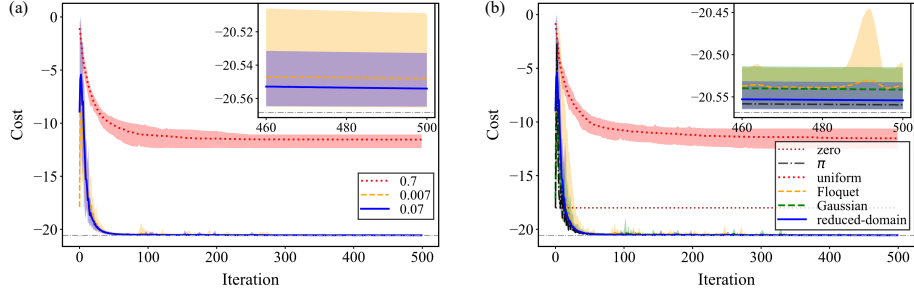


FIG. 5. Convergence behavior of the 12-qubit Heisenberg model under different initializations in the HVA context. The dash-dotted grey line corresponds to that of the global minimum. Here,  $L = 2N = 24$ . (a) Convergence of the cost values under different hyperparameters of the reduced-domain initialization  $\mathcal{U}[-a\pi, a\pi]$ :  $a = 0.07$  (solid blue),  $a = 0.7$  (dotted red), and  $a = 0.007$  (dashed orange). (b) Convergence of the cost values under different initializations. Here, zero-initialization: dotted dark red,  $\pi$ -initialization: dash dotted black, uniform: dotted red, Floquet: thin dashed orange, Gaussian: dashed green, and reduced-domain: solid blue. The corresponding hyperparameters are  $W = 0.4$ ,  $\sigma^2 = 0.001488$ , and  $a = 0.07$ , respectively.

enables accurate approximations to the ground state with rapid convergence.

In the context of HVA, we set  $L = 2N = 24$ , resulting in  $6N(N - 1)$  two-qubit parameterized rotation gates, approximately equal to the total number of  $7N^2$  single-qubit parameterized rotation gates in the previous HEA; see Fig. 1(a). In view of the comparable number of parameterized gates, we set the hyperparameter  $a$  of the reduced-domain initialization for HVA to be 0.07. This choice is numerically supported by the sensitivity analysis on  $a$  concerning the initial trainability and the final convergence behavior of cost values illustrated in Table I and Fig. 5(a), respectively. In addition, we assess the five initialization strategies as well as the  $\pi$ -initialization employed in Ref. [52], by examining their initial gradient and cost convergence, as shown in Table I and Fig. 5(b), respectively. Similar to the numerical findings in the HEA context, the zero-initialization, uniform and soft reduced-domain ( $a = 0.7$ ) strategies suffer from trainability issues, whereas the other initialization strategies can help approximate the ground state, among which the appropriate reduced-domain strategy ( $a = 0.07$ ) demonstrates superior performance. Note that the  $\pi$ -initialization strategy employed in Ref. [52] is a deterministic parameter initialization that exhibits a fixed minor gradient yet achieves better convergence performance than the average performance of the other random initializations. However, the performance of deterministic initializations is highly dependent on the Hamiltonian model and the choice of the reference state. Specifically, although the HVA reduces to the identity when all parameters are set to either 0 or  $\pi$ , their performances are significantly different due to the gradient differences in their landscapes. In practical VQE tasks, deterministic initializations may not be as effective as random ones. Random samples can potentially achieve lower cost values, as illustrated by the blue shaded area in the inset of Fig. 5(b).

Although the reduced-domain initialization is effective in addressing trainability issues, there is still a gap

between the cost values and the global minimum, which is around 0.005 at best in our demonstrations. The role of reduced-domain is to balance the conflict between trainability and expressibility of PQC. Recall that for VQAs, a key factor is designing an appropriate ansatz that is either hardware-efficient, problem-inspired, or in a learning way. Thus, to further improve the performance of VQE, we may integrate the reduced-domain strategy with an effective ansatz for PQC, e.g., the entanglement-variational hardware-efficient ansatz [77].

To take into account of the impact of statistical noise owing to limited measurement shots, for the 10-qubit Heisenberg model, we consider four cases where the numbers of measurement counts per evaluation are 50, 500, 5000, and infinitely many. Here, the circuit depth is set to be  $L = 3.5N = 35$  and the VQEs are optimized using Adam with a fixed learning rate of 0.05 during the 200 training iterations. These numerical simulations are conducted using the PennyLane's built-in "default.qubit" device [78] with a finite number of shots to measure observables.

As illustrated in Fig. 6, regardless of the number of measurement shots, compared to the uniform initialization, the performances under our reduced-domain and Gaussian can be greatly enhanced. Moreover, from the insets of Fig. 6, the performance of our reduced-domain is always more robust with respect to different realizations as compared with Gaussian. From Fig. 6(c) and Fig. 6(d), we can see that for Gaussian initialization, the variance of the cost in the ideal case is larger than that with 5000-shots. This is because the statistical noise may play a positive role. As verified in Section 9.5 in [79], gradient descent with small random noise can implicitly compute the second-order information and exploit the direction of the negative curvature to achieve adequate local descent.



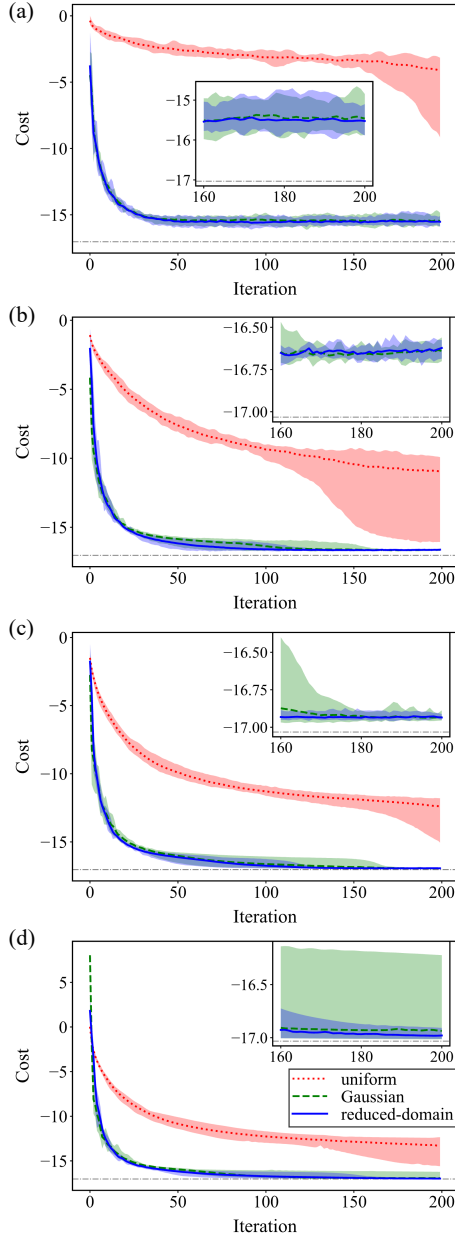


FIG. 6. Convergence behavior of the 10-qubit Heisenberg model under uniform (dotted red), Gaussian (dashed green), and reduced-domain (solid blue) initialization, with different numbers of measurement shots per evaluation: (a) 50 shots; (b) 500 shots; (c) 5000 shots; (d) infinitely many. Here,  $S = 2$ ,  $L = 35$ , the variance of Gaussian initialization is chosen as  $\frac{1}{8SL} = \frac{1}{560} \approx 0.0017857$ , while the hyperparameter  $a$  of the reduced-domain is selected to be 0.0806522 according to Eq. (5).

## B. QNN

We now extend the reduced-domain initialization strategy to QNN. In Proposition 5 of Ref. [41], it was demonstrated that in the under-parameterized region, there exists such a dataset that can induce a

loss function with exponentially many spurious local minima. Subsequently, several simple yet extremely hard QNN instances for training were presented, where the initial parameters are uniformly drawn from the entire period. Here we focus on Example 2 in Ref. [41] and demonstrate that the reduced-domain parameter initialization strategy can help achieve global convergence efficiently.

Let us first briefly introduce the hard instance for training in Ref. [41]. In the under-parameterized region with depth  $L = 1$ , constructing a  $p$ -parameter hard instance involves  $N = p$  qubits with the loss function

$$\frac{1}{2p} \sum_{l=1}^p \left[ \left( \sin 2\theta_l - \sin \frac{\pi}{50} \right)^2 + \frac{1}{4} \left( \cos 2\theta_l - \cos \frac{\pi}{50} \right)^2 \right]. \quad (23)$$

Here, the  $p$ -dimensional parameter vector is denoted by  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)$ . It was proved that within each period, for instance,  $[-\frac{\pi}{2}, \frac{\pi}{2}]$ , there are  $2^p$  minima. Among them, the global minimum is attained at  $\boldsymbol{\theta}^* = (\frac{\pi}{100}, \dots, \frac{\pi}{100})$ , while the other  $2^p - 1$  local minima are spurious with non-negligible suboptimality gaps.

We follow the same routine as outlined in Ref. [41], where QNNs are trained using back-propagating gradients with RMSProp. The constant learning rate is set to be 0.01, and the smoothing constant for mean-squared estimation is 0.99. We perform 5000 rounds of demonstrations with 200 iterations for each round. We consider the convergence of the global minimum within a margin of error of  $10^{-4}$ .

We compare the relative frequencies of achieving global convergence under five initialization strategies: zero-initialization by setting all parameters to be zero, uniform  $\mathcal{U}[-\frac{\pi}{2}, \frac{\pi}{2}]$  as in Ref. [41], Floquet  $\mathcal{U}[-W, W]$  with  $W = 0.2$ , reduced-domain  $\mathcal{U}[-a\frac{\pi}{2}, a\frac{\pi}{2}]$  with  $a = 0.1$ , and Gaussian  $\mathcal{N}(0, 0.008225)$ , whose variance is set to be the same as the reduced-domain strategy. From the semi-log plot in Fig. 7(a), it is clear that under uniform initialization, the probability of success decays exponentially with the number of qubits (or parameters). In particular, in the 18-qubit case, there is no round achieving the global minimum within 5000 rounds of demonstrations. In contrast, all the other advanced strategies consistently achieve global convergence.

To further compare the efficiency of these advanced initialization strategies, we shift the parameter initialization domain away from the global minimum. Specifically, we adjust the center of the above five initialization strategies to  $\frac{\pi}{4}$ , instead of 0. From the semi-log plot in Fig. 7(b), both the Gaussian and Floquet initializations exhibit exponential growth of failure in achieving global convergence as the system size increases, although the rates of their exponential growth are significantly lower than that of uniform initialization. Notably, the reduced-domain and  $\frac{\pi}{4}$ -initialization strategies consistently succeed in achieving global convergence in this context of shallow circuits. Especially, for the 16-qubit instance, the cumulative frequency of global convergence is only 2 under

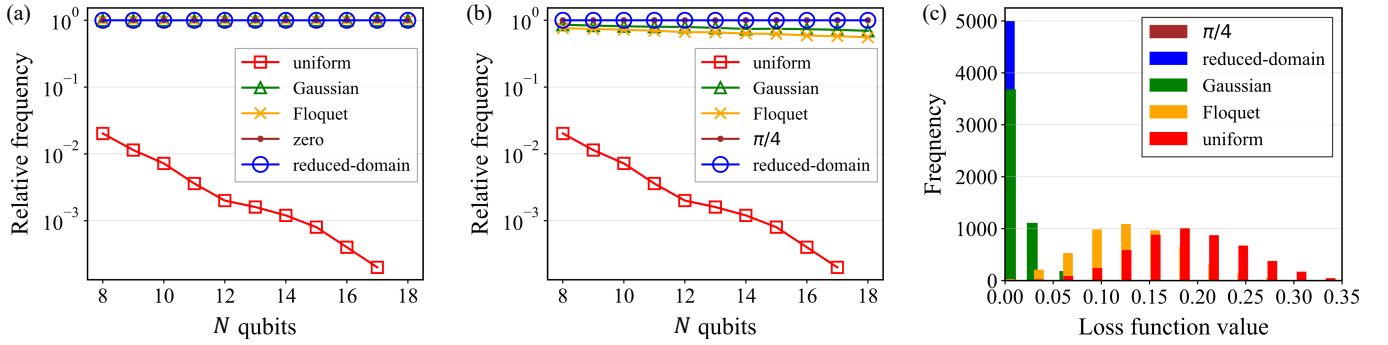


FIG. 7. Relative frequencies of achieving global convergence in 5000 rounds of demonstrations under five initializations: uniform, Gaussian, Floquet, zero/ $\pi/4$ -initialization (parameters are set to be zero/ $\pi/4$ ), and reduced-domain. Here,  $\sigma^2 = 0.008225$ ,  $W = 0.2$ , and  $a = 0.1$ . (a) The center of the parameters' domain is zero for the five initializations. The probability of success decays exponentially with the number of qubits under uniform initialization, while it is always 1 under the other four strategies. (b) The center of the parameters' domain is  $\pi/4$  for the five initializations. Both the Floquet and Gaussian initializations exhibit a slightly exponentially decreasing probability of success, while the reduced-domain and  $\pi/4$ -initialization strategies consistently succeed. (c) Detailed distribution of the loss function value at convergence for the 16-qubit instance. The cumulative frequency of global convergence is 5000 for both  $\pi/4$ -initialization and reduced-domain ( $\mathcal{U}[\frac{\pi}{4} - a\frac{\pi}{2}, \frac{\pi}{4} + a\frac{\pi}{2}]$  with  $a = 0.1$ ), 3602 for Gaussian ( $\mathcal{N}(\frac{\pi}{4}, 0.008225)$ ), 2887 for Floquet ( $\mathcal{U}[\frac{\pi}{4} - 0.2, \frac{\pi}{4} + 0.2]$ ), and only 2 for uniform ( $\mathcal{U}[\frac{\pi}{4} - \frac{\pi}{2}, \frac{\pi}{4} + \frac{\pi}{2}]$ ).

uniform initialization, while it is 2887, 3602, 5000, and 5000 under the Floquet, Gaussian,  $\pi/4$ -initialization, and reduced-domain strategies, respectively. Furthermore, we illustrate a detailed distribution of the loss function's convergence for the 16-qubit case in Fig. 7(c).

From the above demonstrations, we conclude that when the landscape of the cost function has abundance of local minima, it is a smart choice to properly reduce the initial domain to improve the trainability by limiting the expressibility of initial PQC.

## V. CONCLUSION and DISCUSSION

In this paper, we have presented an efficient reduced-domain parameter initialization, enhancing trainability of VQAs both theoretically and empirically. For Hamiltonian represented as a sum of local Pauli strings, we have proved that with the reduced-domain strategy, the average of the squared norm of the cost gradient decays at most polynomially with the depth of circuit and

increases linearly with the number of individual terms in the summation of Hamiltonian. Thus, it clearly implies that reduced-domain initialization helps PQC avoid the phenomenon of BPs. Our empirical results have validated the theoretical results in VQE tasks. Moreover, via simulations, we have demonstrated that the reduced-domain strategy can also help address the poor convergence in presence of BPs in the VQE context of both HEA and HVA, and exponentially many spurious local minima in under-parameterized QNNs. The reduced-domain initialization strategy has the potential to significantly enhance trainability and convergence of PQCs and generate wide impact on developing algorithms with quantum advantages for various practical applications.

## ACKNOWLEDGMENTS

B. Q. acknowledges the support of the National Natural Science Foundation of China (No. 61773370 and No. 61833010), and D. D. acknowledges the support of the Australian Research Council's Future Fellowship funding scheme under project FT220100656.

- 
- [1] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, The theory of variational hybrid quantum-classical algorithms, *New J. Phys.* **18**, 023023 (2016).
  - [2] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, *et al.*, Variational quantum algorithms, *Nat. Rev. Phys.* **3**, 625 (2021).
  - [3] S. Jerbi, L. J. Fiderer, H. Poulsen Nautrup, J. M. Kübler, H. J. Briegel, and V. Dunjko, Quantum machine learning beyond kernel methods, *Nat. Commun.* **14**, 517 (2023).
  - [4] Y. Liu, S. Arunachalam, and K. Temme, A rigorous and robust quantum speed-up in supervised machine learning, *Nat. Phys.* **17**, 1013 (2021).
  - [5] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, Parameterized quantum circuits as machine learning models, *Quantum Sci. Technol.* **4**, 043001 (2019).
  - [6] R. Sweke, F. Wilde, J. Meyer, M. Schuld, P. K. Fährmann, B. Meynard-Piganeau, and J. Eisert, Stochastic gradient descent for hybrid quantum-classical optimization, *Quantum* **4**, 314 (2020).
  - [7] D. Dong and I. R. Petersen, Quantum estimation, control and learning: Opportunities and challenges, *Annual Reviews in Control* **54**, 243 (2022).

- [8] A. Basheer, Y. Feng, C. Ferrie, and S. Li, Alternating layered variational quantum circuits can be classically optimized efficiently using classical shadows, [arXiv preprint arXiv:2208.11623](#) (2022).
- [9] W. Lavrijsen, A. Tudor, J. Müller, C. Iancu, and W. De Jong, Classical optimizers for noisy intermediate-scale quantum devices, in *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)* (2020) pp. 267–277.
- [10] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Quantum machine learning, *Nature* **549**, 195 (2017).
- [11] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, *Phys. Rev. A* **98**, 032309 (2018).
- [12] E. Farhi and H. Neven, Classification with quantum neural networks on near term processors, [arXiv preprint arXiv:1802.06002](#) (2018).
- [13] R.-B. Wu, X. Cao, P. Xie, and Y.-x. Liu, End-to-end quantum machine learning implemented with controlled quantum dynamics, *Phys. Rev. Appl.* **14**, 064020 (2020).
- [14] E. Perrier, D. Tao, and C. Ferrie, Quantum geometric machine learning for quantum circuits and control, *New J. Phys.* **22**, 103056 (2020).
- [15] X. Xu, S. C. Benjamin, and X. Yuan, Variational circuit compiler for quantum error correction, *Phys. Rev. Appl.* **15**, 034068 (2021).
- [16] C. Cao, C. Zhang, Z. Wu, M. Grassl, and B. Zeng, Quantum variational learning for quantum error-correcting codes, *Quantum* **6**, 828 (2022).
- [17] P. D. Johnson, J. Romero, J. Olson, Y. Cao, and A. Aspuru-Guzik, QVECTOR: an algorithm for device-tailored quantum error correction, [arXiv preprint arXiv:1711.02249](#) (2017).
- [18] X. Yuan, S. Endo, Q. Zhao, Y. Li, and S. C. Benjamin, Theory of variational quantum simulation, *Quantum* **3**, 191 (2019).
- [19] E. Altman, K. R. Brown, G. Carleo, L. D. Carr, E. Demler, C. Chin, B. DeMarco, S. E. Economou, M. A. Eriksson, K.-M. C. Fu, M. Greiner, K. R. A. Hazzard, R. G. Hulet, A. J. Kollár, B. L. Lev, M. D. Lukin, R. Ma, X. Mi, S. Misra, C. Monroe, K. Murch, Z. Nazario, K.-K. Ni, A. C. Potter, P. Roushan, M. Saffman, M. Schleier-Smith, I. Siddiqi, R. Simmonds, M. Singh, I. B. Spielman, K. Temme, D. S. Weiss, J. Vučković, V. Vuletić, J. Ye, and M. Zwierlein, Quantum simulators: Architectures and opportunities, *PRX Quantum* **2**, 017003 (2021).
- [20] K. Bharti and T. Haug, Quantum-assisted simulator, *Phys. Rev. A* **104**, 042418 (2021).
- [21] W. W. Ho and T. H. Hsieh, Efficient variational simulation of non-trivial quantum states, *SciPost Phys.* **6**, 029 (2019).
- [22] C. Tabares, A. Muñoz de las Heras, L. Tagliacozzo, D. Porras, and A. González-Tudela, Variational quantum simulators based on waveguide QED, *Phys. Rev. Lett.* **131**, 073602 (2023).
- [23] M. Svensson, M. Andersson, M. Grönkvist, P. Vikstål, D. Dubhashi, G. Ferrini, and G. Johansson, Hybrid quantum-classical heuristic to solve large-scale integer linear programs, *Phys. Rev. Appl.* **20**, 034062 (2023).
- [24] E. Farhi, J. Goldstone, and S. Gutmann, A quantum approximate optimization algorithm, [arXiv preprint arXiv:1411.4028](#) (2014).
- [25] G. E. Crooks, Performance of the quantum approximate optimization algorithm on the maximum cut problem, [arXiv preprint arXiv:1811.08419](#) (2018).
- [26] P. Vikstål, M. Grönkvist, M. Svensson, M. Andersson, G. Johansson, and G. Ferrini, Applying the quantum approximate optimization algorithm to the tail-assignment problem, *Phys. Rev. Appl.* **14**, 034009 (2020).
- [27] Y. Wang, G. Li, and X. Wang, Variational quantum Gibbs state preparation with a truncated Taylor series, *Phys. Rev. Appl.* **16**, 054035 (2021).
- [28] Z. Jiang, K. J. Sung, K. Kechedzhi, V. N. Smelyanskiy, and S. Boixo, Quantum algorithms to simulate many-body physics of correlated fermions, *Phys. Rev. Appl.* **9**, 044036 (2018).
- [29] Z. Liang, H. Wang, J. Cheng, Y. Ding, H. Ren, Z. Gao, Y. Hu, D. S. Boning, X. Qian, S. Han, W. Jiang, and Y. Shi, Variational quantum pulse learning, in *2022 IEEE International Conference on Quantum Computing and Engineering (QCE)* (2022) pp. 556–565.
- [30] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, *Nat. Commun.* **9**, 1 (2018).
- [31] T. Haug, K. Bharti, and M. S. Kim, Capacity and quantum geometry of parametrized quantum circuits, *PRX Quantum* **2**, 040309 (2021).
- [32] C. Zhao and X.-S. Gao, Analyzing the barren plateau phenomenon in training quantum neural networks with the ZX-calculus, *Quantum* **5**, 466 (2021).
- [33] 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**, 1 (2021).
- [34] Z. Liu, L.-W. Yu, L.-M. Duan, and D.-L. Deng, Presence and absence of barren plateaus in tensor-network based machine learning, *Phys. Rev. Lett.* **129**, 270501 (2022).
- [35] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, *PRX Quantum* **3**, 010313 (2022).
- [36] C. Ortiz Marrero, M. Kieferová, and N. Wiebe, Entanglement-induced barren plateaus, *PRX Quantum* **2**, 040316 (2021).
- [37] K. Sharma, M. Cerezo, L. Cincio, and P. J. Coles, Trainability of dissipative perceptron-based quantum neural networks, *Phys. Rev. Lett.* **128**, 180505 (2022).
- [38] T. L. Patti, K. Najafi, X. Gao, and S. F. Yelin, Entanglement devised barren plateau mitigation, *Phys. Rev. Res.* **3**, 033090 (2021).
- [39] S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, and P. J. Coles, Noise-induced barren plateaus in variational quantum algorithms, *Nat. Commun.* **12**, 1 (2021).
- [40] E. R. Anschuetz and B. T. Kiani, Quantum variational algorithms are swamped with traps, *Nat. Commun.* **13**, 7760 (2022).
- [41] X. You and X. Wu, Exponentially many local minima in quantum neural networks, in *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)* (2021) pp. 12144–12155.
- [42] L. Bittel and M. Kliesch, Training variational quantum algorithms is NP-hard, *Phys. Rev. Lett.* **127**, 120502 (2021).
- [43] D. Wierichs, C. Gogolin, and M. Kastoryano, Avoiding local minima in variational quantum eigensolvers with the natural gradient optimizer, *Phys. Rev. Res.* **2**, 043246 (2020).

- (2020).
- [44] X. Ge, R.-B. Wu, and H. Rabitz, The optimization landscape of hybrid quantum–classical algorithms: From quantum control to NISQ applications, *Annual Reviews in Control* **54**, 314 (2022).
  - [45] Y.-Y. Shi, L.-M. Duan, and G. Vidal, Classical simulation of quantum many-body systems with a tree tensor network, *Phys. Rev. A* **74**, 022320 (2006).
  - [46] E. Grant, M. Benedetti, S. Cao, A. Hallam, J. Lockhart, V. Stojevic, A. G. Green, and S. Severini, Hierarchical quantum classifiers, *npj Quantum Inf.* **4**, 1 (2018).
  - [47] I. Cong, S. Choi, and M. D. Lukin, Quantum convolutional neural networks, *Nat. Phys.* **15**, 1273 (2019).
  - [48] K. Zhang, M.-H. Hsieh, L. Liu, and D. Tao, Toward trainability of quantum neural networks, *arXiv preprint arXiv:2011.06258* (2020).
  - [49] A. Pesah, M. Cerezo, S. Wang, T. Volkoff, A. T. Sornborger, and P. J. Coles, Absence of barren plateaus in quantum convolutional neural networks, *Phys. Rev. X* **11**, 041011 (2021).
  - [50] T. Barthel and Q. Miao, Absence of barren plateaus and scaling of gradients in the energy optimization of isometric tensor network states, *arXiv preprint arXiv:2304.00161* (2023).
  - [51] S. Hadfield, Z. Wang, B. O’gorman, E. G. Rieffel, D. Venturelli, and R. Biswas, From the quantum approximate optimization algorithm to a quantum alternating operator ansatz, *Algorithms* **12**, 34 (2019).
  - [52] R. Wiersema, C. Zhou, Y. de Sereville, J. F. Carrasquilla, Y. B. Kim, and H. Yuen, Exploring entanglement and optimization within the Hamiltonian variational ansatz, *PRX Quantum* **1**, 020319 (2020).
  - [53] C. Lyu, X. Xu, M.-H. Yung, and A. Bayat, Symmetry enhanced variational quantum spin eigensolver, *Quantum* **7**, 899 (2023).
  - [54] A. Skolik, J. R. McClean, M. Mohseni, P. van der Smagt, and M. Leib, Layerwise learning for quantum neural networks, *Quantum Mach. Intell.* **3**, 1 (2021).
  - [55] E. Campos, D. Rabinovich, V. Akshay, and J. Biamonte, Training saturation in layerwise quantum approximate optimization, *Phys. Rev. A* **104**, L030401 (2021).
  - [56] O. Lockwood, Optimizing quantum variational circuits with deep reinforcement learning, *arXiv preprint arXiv:2109.03188* (2021).
  - [57] J. Liu, K. Najafi, K. Sharma, F. Tacchino, L. Jiang, and A. Mezzacapo, Analytic theory for the dynamics of wide quantum neural networks, *Phys. Rev. Lett.* **130**, 150601 (2023).
  - [58] X. You, S. Chakrabarti, and X. Wu, A convergence theory for over-parameterized variational quantum eigensolvers, *arXiv preprint arXiv:2205.12481* (2022).
  - [59] E. R. Anschuetz, Critical points in quantum generative models, in *International Conference on Learning Representations* (2022).
  - [60] M. Larocca, N. Ju, D. García-Martín, P. J. Coles, and M. Cerezo, Theory of overparametrization in quantum neural networks, *Nature Computational Science* **3**, 542 (2023).
  - [61] L. Zhou, S.-T. Wang, S. Choi, H. Pichler, and M. D. Lukin, Quantum approximate optimization algorithm: Performance, mechanism, and implementation on near-term devices, *Phys. Rev. X* **10**, 021067 (2020).
  - [62] E. Grant, L. Wossnig, M. Ostaszewski, and M. Benedetti, An initialization strategy for addressing barren plateaus in parametrized quantum circuits, *Quantum* **3**, 214 (2019).
  - [63] T. Volkoff and P. J. Coles, Large gradients via correlation in random parameterized quantum circuits, *Quantum Sci. Technol.* **6**, 025008 (2021).
  - [64] A. Rad, A. Seif, and N. M. Linke, Surviving the barren plateau in variational quantum circuits with Bayesian learning initialization, *arXiv preprint arXiv:2203.02464* (2022).
  - [65] K. Zhang, L. Liu, M.-H. Hsieh, and D. Tao, Escaping from the barren plateau via gaussian initializations in deep variational quantum circuits, in *Advances in Neural Information Processing Systems*, edited by A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho (2022).
  - [66] C. Cao, Y. Zhou, S. Tannu, N. Shannon, and R. Joynt, Exploiting many-body localization for scalable variational quantum simulation, *arXiv preprint arXiv:2404.17560* (2024).
  - [67] M. S. Alam, F. A. Wudarski, M. J. Reagor, J. Sud, S. Grabbe, Z. Wang, M. Hodson, P. A. Lott, E. G. Rieffel, and D. Venturelli, Practical verification of quantum properties in quantum-approximate-optimization runs, *Phys. Rev. Appl.* **17**, 024026 (2022).
  - [68] L. Lewis, H.-Y. Huang, V. T. Tran, S. Lehner, R. Kueng, and J. Preskill, Improved machine learning algorithm for predicting ground state properties, *Nat. Commun.* **15**, 895 (2024).
  - [69] S. McArdle, S. Endo, A. Aspuru-Guzik, S. C. Benjamin, and X. Yuan, Quantum computational chemistry, *Rev. Mod. Phys.* **92**, 015003 (2020).
  - [70] Y. Cao, J. Romero, J. P. Olson, M. Degroote, P. D. Johnson, M. Kieferová, I. D. Kivlichan, T. Menke, B. Peropadre, N. P. Sawaya, *et al.*, Quantum chemistry in the age of quantum computing, *Chemical reviews* **119**, 10856 (2019).
  - [71] A. Lucas, Ising formulations of many NP problems, *Frontiers in Physics* **2**, 10.3389/fphy.2014.00005 (2014).
  - [72] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, TensorFlow: a system for large-scale machine learning, in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI’16 (USENIX Association, USA, 2016) p. 265–283.
  - [73] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, PyTorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems*, Vol. 32, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019).
  - [74] S.-X. Zhang, J. Allcock, Z.-Q. Wan, S. Liu, J. Sun, H. Yu, X.-H. Yang, J. Qiu, Z. Ye, Y.-Q. Chen, C.-K. Lee, Y.-C. Zheng, S.-K. Jian, H. Yao, C.-Y. Hsieh, and S. Zhang, TensorCircuit: a quantum software framework for the NISQ era, *Quantum* **7**, 912 (2023).



- [75] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, [arXiv preprint arXiv:1412.6980](#) (2014).
- [76] I. Loshchilov and F. Hutter, Decoupled weight decay regularization, in *International Conference on Learning Representations* (2019).
- [77] X. Wang, B. Qi, Y. Wang, and D. Dong, Entanglement-variational hardware-efficient ansatz for eigensolvers, *Phys. Rev. Appl.* **21**, 034059 (2024).
- [78] V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, S. Ahmed, V. Ajith, M. S. Alam, G. Alonso-Linaje, B. AkashNarayanan, A. Asadi, *et al.*, PennyLane: Automatic differentiation of hybrid quantum-classical computations, [arXiv preprint arXiv:1811.04968](#) (2018).
- [79] J. Wright and Y. Ma, *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications* (Cambridge University Press, 2022).

## Appendix A: The norm of gradient

In this section, we explain the meaning of the figure of merit  $\|\nabla_{\theta} C\|_2^2$ . In gradient-based optimizers, the parameter vector  $\theta$  updates as

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta} C(\theta_t),$$

where  $\eta$  denotes the learning rate, which is usually small. With the update, the value of cost function decreases as

$$C(\theta_{t+1}) = C(\theta_t - \eta \nabla_{\theta} C(\theta_t)) = C(\theta_t) - \eta \|\nabla_{\theta} C(\theta_t)\|_2^2 + o(\eta).$$

It is clear to see that the decrease with parameters updating is quantified in terms of the magnitude of  $\|\nabla_{\theta} C\|_2^2$ . Thus, we adopt the average  $\mathbb{E}_{\theta} \|\nabla_{\theta} C\|_2^2$  as the trainability figure of merit.

## Appendix B: Technical Lemmas

In this section, we present several technical lemmas, which are all based on the following assumption.

**Assumption 5.** A random variable  $\theta$  is drawn from uniform distribution over  $[-a\pi, a\pi]$ , where  $a \in (0, 1)$  is a hyperparameter to be further determined.

Under the assumption, let  $\mathbb{E}_{\theta}$  denote the expectation over  $\theta$  with respect to the uniform distribution  $\mathcal{U}[-a\pi, a\pi]$ . Recall that

$$\alpha = \frac{1}{4} \left( 2 + \frac{\sin 2a\pi}{a\pi} \right), \beta = \frac{1}{4} \left( 2 - \frac{\sin 2a\pi}{a\pi} \right), \gamma = \frac{\sin a\pi}{a\pi}.$$

**Lemma 6.** Let  $A$  be an arbitrary linear operator,  $G$  be a Hermitian unitary and  $V = e^{-i\frac{\theta}{2}G}$ . Let  $O$  be an arbitrary Hermitian operator that anti-commutes with  $G$ , and  $\tilde{O}$  be an arbitrary Hermitian operator that commutes with  $G$ . Then,

$$\mathbb{E}_{\theta} \text{Tr} [O V A V^{\dagger}] = \gamma \text{Tr} [O A], \quad (\text{B1})$$

$$\mathbb{E}_{\theta} \frac{\partial}{\partial \theta} \text{Tr} [O V A V^{\dagger}] = \gamma \text{Tr} [i G O A], \quad (\text{B2})$$

$$\mathbb{E}_{\theta} \text{Tr} [\tilde{O} V A V^{\dagger}] = \text{Tr} [\tilde{O} A], \quad (\text{B3})$$

$$\mathbb{E}_{\theta} \frac{\partial}{\partial \theta} \text{Tr} [\tilde{O} V A V^{\dagger}] = 0. \quad (\text{B4})$$

*Proof.* From  $V = e^{-i\frac{\theta}{2}G} = I \cos \frac{\theta}{2} - iG \sin \frac{\theta}{2}$ , for any operator  $B$ , we have

$$\begin{aligned} \text{Tr} [B V A V^{\dagger}] &= \text{Tr} \left[ B \left( I \cos \frac{\theta}{2} - iG \sin \frac{\theta}{2} \right) A \left( I \cos \frac{\theta}{2} + iG \sin \frac{\theta}{2} \right) \right] \\ &= \frac{1 + \cos \theta}{2} \text{Tr} [B A] + \frac{1 - \cos \theta}{2} \text{Tr} [B G A G] + \frac{\sin \theta}{2} (\text{Tr} [i B A G] - \text{Tr} [i B G A]). \end{aligned} \quad (\text{B5})$$

If  $O$  anti-commutes with  $G$ , then  $iGO$  is Hermitian. In addition, with  $G$  being unitary and  $\{O, G\} = 0$ , from Eq. (B5) we have

$$\text{Tr} [O V A V^{\dagger}] = \cos \theta \text{Tr} [O A] + \sin \theta \text{Tr} [i G O A], \quad (\text{B6})$$

and accordingly,

$$\frac{\partial}{\partial \theta} \text{Tr} [O V A V^\dagger] = -\sin \theta \text{Tr} [O A] + \cos \theta \text{Tr} [i G O A]. \quad (\text{B7})$$

Thus, Eqs. (B1) and (B2) can be obtained by taking the expectations of Eqs. (B6) and (B7) over  $\theta$ , respectively.

Moreover, when  $\tilde{O}$  commutes with the unitary  $G$ , following similar calculations, it is straightforward to have Eqs. (B3) and (B4).  $\square$

From Lemma 6, we have the following conclusion.

**Corollary 7.** *Let  $A$  be an arbitrary linear operator,  $G$  be a Hermitian unitary and  $V = e^{-i\frac{\theta}{2}G}$ . Let  $O$  denote an arbitrary Hermitian operator that anti-commutes with  $G$ , and  $\tilde{O}_1$  and  $\tilde{O}_2$  be arbitrary Hermitian operators that commute with  $G$ . Then,*

$$\mathbb{E}_\theta \text{Tr} [O V A V^\dagger] \text{Tr} [\tilde{O}_1 V A V^\dagger] = \gamma \text{Tr} [O A] \text{Tr} [\tilde{O}_1 A], \quad (\text{B8})$$

$$\mathbb{E}_\theta \frac{\partial}{\partial \theta} \text{Tr} [O V A V^\dagger] \cdot \frac{\partial}{\partial \theta} \text{Tr} [\tilde{O}_1 V A V^\dagger] = 0, \quad (\text{B9})$$

$$\mathbb{E}_\theta \text{Tr} [\tilde{O}_1 V A V^\dagger] \text{Tr} [\tilde{O}_2 V A V^\dagger] = \text{Tr} [\tilde{O}_1 A] \text{Tr} [\tilde{O}_2 A], \quad (\text{B10})$$

$$\mathbb{E}_\theta \frac{\partial}{\partial \theta} \text{Tr} [\tilde{O}_1 V A V^\dagger] \cdot \frac{\partial}{\partial \theta} \text{Tr} [\tilde{O}_2 V A V^\dagger] = 0. \quad (\text{B11})$$

**Lemma 8.** *Let  $A$  be an arbitrary linear operator,  $G$  be a Hermitian unitary and  $V = e^{-i\frac{\theta}{2}G}$ . Let  $O_1$  and  $O_2$  be arbitrary Hermitian operators that anti-commute with  $G$ . Then,*

$$\mathbb{E}_\theta \text{Tr} [O_1 V A V^\dagger] \text{Tr} [O_2 V A V^\dagger] = \alpha \text{Tr} [O_1 A] \text{Tr} [O_2 A] + \beta \text{Tr} [i G O_1 A] \text{Tr} [i G O_2 A], \quad (\text{B12})$$

$$\mathbb{E}_\theta \frac{\partial}{\partial \theta} \text{Tr} [O_1 V A V^\dagger] \cdot \frac{\partial}{\partial \theta} \text{Tr} [O_2 V A V^\dagger] = \beta \text{Tr} [O_1 A] \text{Tr} [O_2 A] + \alpha \text{Tr} [i G O_1 A] \text{Tr} [i G O_2 A]. \quad (\text{B13})$$

*Proof.* According to Eqs. (B6) and (B7), we have

$$\begin{aligned} \text{Tr} [O_1 V A V^\dagger] \cdot \text{Tr} [O_2 V A V^\dagger] &= \cos^2 \theta \text{Tr} [O_1 A] \text{Tr} [O_2 A] + \sin^2 \theta \text{Tr} [i G O_1 A] \text{Tr} [i G O_2 A] \\ &\quad + \sin \theta \cos \theta \{ \text{Tr} [O_1 A] \text{Tr} [i G O_2 A] + \text{Tr} [O_2 A] \text{Tr} [i G O_1 A] \}, \end{aligned} \quad (\text{B14})$$

$$\begin{aligned} \frac{\partial}{\partial \theta} \text{Tr} [O_1 V A V^\dagger] \cdot \frac{\partial}{\partial \theta} \text{Tr} [O_2 V A V^\dagger] &= \sin^2 \theta \text{Tr} [O_1 A] \text{Tr} [O_2 A] + \cos^2 \theta \text{Tr} [i G O_1 A] \text{Tr} [i G O_2 A] \\ &\quad - \sin \theta \cos \theta \{ \text{Tr} [O_1 A] \text{Tr} [i G O_2 A] + \text{Tr} [O_2 A] \text{Tr} [i G O_1 A] \}. \end{aligned} \quad (\text{B15})$$

Then, Eqs. (B12) and (B13) can be derived by taking the expectations of Eqs. (B14) and (B15) and employing

$$\mathbb{E}_\theta \cos^2 \theta = \alpha, \quad \mathbb{E}_\theta \sin^2 \theta = \beta, \quad \mathbb{E}_\theta \sin \theta \cos \theta = 0.$$

$\square$

It follows that the following conclusion holds.

**Corollary 9.** *Let  $A$  be an arbitrary linear operator,  $G$  be a Hermitian unitary and  $V = e^{-i\frac{\theta}{2}G}$ . Let  $O$  be an arbitrary Hermitian operator that anti-commutes with  $G$ . Then,*

$$\mathbb{E}_\theta \text{Tr}^2 [O V A V^\dagger] = \alpha \text{Tr}^2 [O A] + \beta \text{Tr}^2 [i G O A], \quad (\text{B16})$$

$$\mathbb{E}_\theta \left( \frac{\partial}{\partial \theta} \text{Tr} [O V A V^\dagger] \right)^2 = \beta \text{Tr}^2 [O A] + \alpha \text{Tr}^2 [i G O A], \quad (\text{B17})$$

where  $\text{Tr}^2[\cdot] = (\text{Tr}[\cdot])^2$ .

The following lemma concerns the transformations of two-qubit Pauli tensor products after a controlled- $Z$  gate.

**Lemma 10.** Let  $CZ$  denote a controlled- $Z$  gate, and  $\sigma_i \otimes \sigma_j$  be a two-qubit Pauli tensor product, where  $\sigma_i$  and  $\sigma_j$  are Pauli matrices. If  $\sigma_{i'} \otimes \sigma_{j'} = CZ^\dagger(\sigma_i \otimes \sigma_j)CZ$ , we denote the transformation by  $\sigma_i \otimes \sigma_j \rightarrow \sigma_{i'} \otimes \sigma_{j'}$ . All of the specific transformations can be summarized as follows:

$$\begin{aligned} Z \otimes I &\leftrightarrow Z \otimes I, & Z \otimes Z &\leftrightarrow Z \otimes Z, \\ Y \otimes I &\leftrightarrow Y \otimes Z, & Y \otimes Y &\leftrightarrow X \otimes X, \\ X \otimes I &\leftrightarrow X \otimes Z, & X \otimes Y &\leftrightarrow -Y \otimes X. \end{aligned}$$

### Appendix C: Proof of Lemma 2

*Proof.* For the Hamiltonian of Eq. (7), the average of the squared norm of the gradient can be described as

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}} \|\nabla_{\boldsymbol{\theta}} C\|_2^2 &= \sum_{q=1}^{2L} \sum_{n=1}^N \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C}{\partial \theta_{q,n}} \right)^2 \\ &= \sum_{q=1}^{2L} \sum_{n=1}^N \mathbb{E}_{\boldsymbol{\theta}} \left( \sum_{i=1}^{N-1} \frac{\partial C_i}{\partial \theta_{q,n}} \right)^2 \\ &= \sum_{i=1}^{N-1} \sum_{q=1}^{2L} \sum_{n=1}^N \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \right)^2 + \sum_{i \neq j=1}^{N-1} \sum_{q=1}^{2L} \sum_{n=1}^N \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \frac{\partial C_j}{\partial \theta_{q,n}} \right), \end{aligned} \quad (\text{C1})$$

where  $C_i = \text{Tr} [Z_i Z_{i+1} U(\boldsymbol{\theta}) \rho U^\dagger(\boldsymbol{\theta})]$ . Thus, we turn to analyze the lower bound of Eq. (C1).

For clarity, we first introduce notation. As illustrated in Fig. 1(a), the  $L$ -layer parameterized unitary  $U(\boldsymbol{\theta})$  can be expressed as the product of  $L$  sequential unitaries

$$U(\boldsymbol{\theta}) = U_L(\boldsymbol{\theta}_{2L}, \boldsymbol{\theta}_{2L-1}) \cdots U_1(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1),$$

with

$$\begin{aligned} U_l(\boldsymbol{\theta}_{2l}, \boldsymbol{\theta}_{2l-1}) &= R_{2l}(\boldsymbol{\theta}_{2l}) R_{2l-1}(\boldsymbol{\theta}_{2l-1}) CZ_l, \\ R_{2l}(\boldsymbol{\theta}_{2l}) &= e^{-\frac{i}{2} \theta_{2l,1} Y} \otimes \cdots \otimes e^{-\frac{i}{2} \theta_{2l,N} Y}, \\ R_{2l-1}(\boldsymbol{\theta}_{2l-1}) &= e^{-\frac{i}{2} \theta_{2l-1,1} X} \otimes \cdots \otimes e^{-\frac{i}{2} \theta_{2l-1,N} X}. \end{aligned}$$

Here, for  $l \in [L]$ ,  $CZ_l$  denotes the  $l$ th entanglement layer composed of arbitrarily many  $CZ$  gates, and  $R_{2l}(\boldsymbol{\theta}_{2l})$  and  $R_{2l-1}(\boldsymbol{\theta}_{2l-1})$  denote the  $l$ th  $R_Y$  and  $R_X$  rotation layer, respectively. Denote  $\rho_0 = |0\rangle\langle 0|$ , and for  $l \in [L]$ , let

$$\begin{cases} \rho_{2l} = R_{2l}(\boldsymbol{\theta}_{2l}) \rho_{2l-1} R_{2l}^\dagger(\boldsymbol{\theta}_{2l}), \\ \rho_{2l-1} = R_{2l-1}(\boldsymbol{\theta}_{2l-1}) CZ_l \rho_{2l-2} CZ_l^\dagger R_{2l-1}^\dagger(\boldsymbol{\theta}_{2l-1}). \end{cases}$$

Now we evaluate the lower bound of the first summation in Eq. (C1). Note that

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \right)^2 &= \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_i Z_{i+1} U(\boldsymbol{\theta}) \rho_0 U^\dagger(\boldsymbol{\theta})] \right)^2 \\ &= \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_i Z_{i+1} \rho_{2L}] \right)^2 \\ &= \mathbb{E}_{\boldsymbol{\theta}_{[2L-1]}} \mathbb{E}_{\boldsymbol{\theta}_{2L}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_i Z_{i+1} R_{2L}(\boldsymbol{\theta}_{2L}) \rho_{2L-1} R_{2L}^\dagger(\boldsymbol{\theta}_{2L})] \right)^2 \end{aligned} \quad (\text{C2})$$

$$= \mathbb{E}_{\boldsymbol{\theta}_{[2L-1]}} \mathbb{E}_{\boldsymbol{\theta}_{2L}} \text{Tr}^2 \left[ Z_i Z_{i+1} R_{2L}(\boldsymbol{\theta}_{2L}) \frac{\partial \rho_{2L-1}}{\partial \theta_{q,n}} R_{2L}^\dagger(\boldsymbol{\theta}_{2L}) \right] \quad (\text{C3})$$

$$\geq \alpha^2 \mathbb{E}_{\boldsymbol{\theta}_{[2L-1]}} \text{Tr}^2 \left[ Z_i Z_{i+1} \frac{\partial \rho_{2L-1}}{\partial \theta_{q,n}} \right] \quad (\text{C4})$$

$$= \alpha^2 \mathbb{E}_{\boldsymbol{\theta}_{[2L-2]}} \mathbb{E}_{\boldsymbol{\theta}_{2L-1}} \text{Tr}^2 \left[ Z_i Z_{i+1} R_{2L-1}(\boldsymbol{\theta}_{2L-1}) \text{CZ}_L \frac{\partial \rho_{2L-2}}{\partial \theta_{q,n}} \text{CZ}_L^\dagger R_{2L-1}^\dagger(\boldsymbol{\theta}_{2L-1}) \right] \quad (\text{C5})$$

$$\geq \alpha^4 \mathbb{E}_{\boldsymbol{\theta}_{[2L-2]}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [\text{CZ}_L^\dagger Z_i Z_{i+1} \text{CZ}_L \rho_{2L-2}] \right)^2 \quad (\text{C6})$$

Here, Eq. (C3) is due to the linearity of the trace operation, which is frequently utilized in the proof. Eq. (C4) is obtained by first employing Eq. (B16) to take the expectations over  $\boldsymbol{\theta}_{2L,i}$  and  $\boldsymbol{\theta}_{2L,i+1}$  producing the coefficient  $\alpha^2$ , and then leveraging Eq. (B10) to take the expectations over the other  $N-2$  components in  $\boldsymbol{\theta}_{2L}$ . Following a similar analysis, Eq. (C5) can be derived. Eq. (C6) is due to Lemma 10, that is, the tensor product  $Z_i Z_{i+1}$  remains unchanged after applying arbitrarily many CZ gates.

We repeat the techniques for deriving Eqs. (C2)-(C6) to calculate the expectations over  $\boldsymbol{\theta}_{2L-1}, \dots, \boldsymbol{\theta}_{q+1}$  and obtain

$$\mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \right)^2 \geq \alpha^{2(2L-q)} \mathbb{E}_{\boldsymbol{\theta}_{[q]}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_i Z_{i+1} \rho_q] \right)^2. \quad (\text{C7})$$

Next, we proceed upon the index  $n$ , the ordinal number of the qubit where the trainable parameter  $\theta_{q,n}$  lies. In the case where  $n \in \{i, i+1\}$ , we have

$$\mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \right)^2 \geq \alpha^{2(2L-q)} \cdot \alpha \beta \mathbb{E}_{\boldsymbol{\theta}_{[q-1]}} \text{Tr}^2 [Z_i Z_{i+1} \rho_{q-1}] \quad (\text{C8})$$

$$\geq \alpha^{2(2L-q)+1} \beta \cdot \alpha^{2(q-1)} \text{Tr}^2 [Z_i Z_{i+1} \rho_0], \quad (\text{C9})$$

where  $\beta$  in Eq. (C8) is obtained by employing Eq. (B17) to take the expectation over  $\theta_{q,n}$ . It is worth pointing out that to persist the observable  $Z_i Z_{i+1}$  in the following analysis, we keep the item with coefficient  $\beta$  rather than  $\alpha$  in the inequality. As for the other  $N-1$  components in  $\boldsymbol{\theta}_q$ , the analysis is the same as that in Eq. (C4) or (C5). Following a similar derivation to Eq. (C7), Eq. (C9) can be obtained.

In the case where  $n \notin \{i, i+1\}$ , when taking the expectation over  $\theta_{q,n}$ , by Eq. (B11) we have

$$\mathbb{E}_{\boldsymbol{\theta}_{[q]}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_i Z_{i+1} \rho_q] \right)^2 = 0. \quad (\text{C10})$$

Therefore, combining Eq. (C9) and Eq. (C10), it can be seen that

$$\mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \right)^2 \geq \begin{cases} \alpha^{4L-1} \beta, & n \in \{i, i+1\}, \\ 0, & \text{otherwise.} \end{cases} \quad (\text{C11})$$

and the first summation in Eq. (C1) is lower bounded by

$$\sum_{i=1}^{N-1} \sum_{q=1}^{2L} \sum_{n=1}^N \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \right)^2 \geq (N-1) \times 2L \times 2\alpha^{4L-1} \beta. \quad (\text{C12})$$



Then, we analyze the second summation in Eq. (C1). Note that

$$\mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \frac{\partial C_j}{\partial \theta_{q,n}} \right) = \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_i Z_{i+1} \rho_{2L}] \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_j Z_{j+1} \rho_{2L}] \right).$$

We first consider the case where  $\{i, i+1\} \cap \{j, j+1\} = \emptyset$ .

$$\mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \frac{\partial C_j}{\partial \theta_{q,n}} \right) = \gamma^{4(2L-q)} \mathbb{E}_{\boldsymbol{\theta}_{[q]}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_i Z_{i+1} \rho_q] \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_j Z_{j+1} \rho_q] \right) \quad (\text{C13})$$

$$= 0. \quad (\text{C14})$$

In regard to the expectation over  $\boldsymbol{\theta}_{2L}$  in Eq. (C13), the coefficient  $\gamma$  appears each time when we take the expectation over  $\theta_{2L,i}$ ,  $\theta_{2L,i+1}$ ,  $\theta_{2L,j}$  or  $\theta_{2L,j+1}$  by employing Eq. (B8). In addition, Eq. (B10) is utilized to calculate the expectations over the other components in  $\boldsymbol{\theta}_{2L}$ . Similarly, we can take the expectations over  $\boldsymbol{\theta}_{2L-1}, \dots, \boldsymbol{\theta}_{q+1}$ , during which the calculation of each CZ layer is the same as that in Eq. (C6). When taking the expectation over  $\theta_{q,n}$ , Eq. (C14) can be obtained by employing either Eq. (B9) for  $n \in \{i, i+1, j, j+1\}$  or Eq. (B11), otherwise.

Next, we consider the case where  $i = j+1$  or  $i+1 = j$ . Let us consider the case  $i+1 = j$  in detail, and the other case can be followed in a similar way. Note that

$$\mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \frac{\partial C_{i+1}}{\partial \theta_{q,n}} \right) = \mathbb{E}_{\boldsymbol{\theta}_{[2L]}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_i Z_{i+1} \rho_{2L}] \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_{i+1} Z_{i+2} \rho_{2L}] \right) \quad (\text{C15})$$

$$= \gamma^2 \alpha \mathbb{E}_{\boldsymbol{\theta}_{[2L-1]}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_i Z_{i+1} \rho_{2L-1}] \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_{i+1} Z_{i+2} \rho_{2L-1}] \right) \quad (\text{C16})$$

$$+ \gamma^2 \beta \mathbb{E}_{\boldsymbol{\theta}_{[2L-1]}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_i X_{i+1} \rho_{2L-1}] \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [X_{i+1} Z_{i+2} \rho_{2L-1}] \right) \\ = \gamma^4 \alpha^2 \mathbb{E}_{\boldsymbol{\theta}_{[2L-2]}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [CZ_L^\dagger Z_i Z_{i+1} CZ_L \rho_{2L-2}] \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [CZ_L^\dagger Z_{i+1} Z_{i+2} CZ_L \rho_{2L-2}] \right) \\ + \gamma^4 \alpha \beta \mathbb{E}_{\boldsymbol{\theta}_{[2L-2]}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [CZ_L^\dagger Z_i X_{i+1} CZ_L \rho_{2L-2}] \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [CZ_L^\dagger X_{i+1} Z_{i+2} CZ_L \rho_{2L-2}] \right) \quad (\text{C17}) \\ + \gamma^4 \beta \mathbb{E}_{\boldsymbol{\theta}_{[2L-2]}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [CZ_L^\dagger Z_i X_{i+1} CZ_L \rho_{2L-2}] \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [CZ_L^\dagger X_{i+1} Z_{i+2} CZ_L \rho_{2L-2}] \right).$$

Here, Eq. (C16) is obtained by taking the expectations over  $\theta_{2L,i+1}$  by Eq. (B12),  $\theta_{2L,i}$  and  $\theta_{2L,i+2}$  by Eq. (B8), and the other  $N-3$  components by Eq. (B10). Similarly, Eq. (C17) can be derived.

It is worth pointing out that after employing Lemmas 6 and 8 to calculate the expectations over the trainable parameters in rotation layers, except for the terms related to the original observables  $Z_i Z_{i+1}$  and  $Z_{i+1} Z_{i+2}$ , other terms related to new observables may be generated. For example, after taking the expectation over  $\boldsymbol{\theta}_{2L}$ , the original observable  $Z_i Z_{i+1}$  in Eq. (C15) generates two terms related to the observable  $Z_i Z_{i+1}$  and  $Z_i X_{i+1}$ , respectively, in Eq. (C16). It is clear that the generated new observables are always  $N$ -qubit Pauli strings as defined in Eq. (2), and the corresponding coefficients are monomials of  $\gamma$ ,  $\alpha$  and  $\beta$ , which are always positive. Moreover, according to Lemma 10, after applying an entanglement layer composed of CZ gates, the generated observables are still  $N$ -qubit Pauli strings, where, for example, the observable  $Z_i Z_{i+1}$  remains unchanged. Thus, whenever we finish the calculations with respect to a block, there are always terms related to the original observables  $Z_i Z_{i+1}$  and  $Z_{i+1} Z_{i+2}$ , and some other generated terms related to observables in the form of  $N$ -qubit Pauli strings. We point out that all the generated terms are non-negative, and we come back to its proof later. In the following, the terms with the original observables  $Z_i Z_{i+1}$  and  $Z_{i+1} Z_{i+2}$  are kept only for further analysis. Thus, we have

$$\mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \frac{\partial C_{i+1}}{\partial \theta_{q,n}} \right) \geq \gamma^4 \alpha^2 \mathbb{E}_{\boldsymbol{\theta}_{[2L-2]}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_i Z_{i+1} \rho_{2L-2}] \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_{i+1} Z_{i+2} \rho_{2L-2}] \right) \quad (\text{C18})$$

$$\geq \gamma^{2(2L-q)} \alpha^{2L-q} \mathbb{E}_{\boldsymbol{\theta}_{[q]}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_i Z_{i+1} \rho_q] \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_{i+1} Z_{i+2} \rho_q] \right), \quad (\text{C19})$$

where Eq. (C19) is obtained by repeating the same analysis as that in Eq. (C18) for the expectations over  $\boldsymbol{\theta}_{2L-2}, \dots, \boldsymbol{\theta}_{q+1}$ .

Then, we proceed upon the index  $n$ . When  $n = i + 1$ , we have

$$\begin{aligned} & \mathbb{E}_{\theta_{[q]}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_i Z_{i+1} \rho_q] \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_{i+1} Z_{i+2} \rho_q] \right) \\ & \geq \gamma^2 \beta \mathbb{E}_{\theta_{[q-1]}} (\text{Tr} [Z_i Z_{i+1} \rho_{q-1}] \text{Tr} [Z_{i+1} Z_{i+2} \rho_{q-1}]) \end{aligned} \quad (\text{C20})$$

$$\geq \gamma^2 \beta \cdot \gamma^{2(q-1)} \alpha^{q-1} \text{Tr} [Z_i Z_{i+1} \rho_0] \text{Tr} [Z_{i+1} Z_{i+2} \rho_0], \quad (\text{C21})$$

where in Eq. (C20), the coefficient  $\beta$  comes from Eq. (B13) for the expectation over  $\theta_{q,n}$ , and the coefficient  $\gamma^2$  comes from Eq. (B8) for  $\theta_{q,n-1}$  and  $\theta_{q,n+1}$ . Eq. (C21) is obtained in the same way as Eq. (C19).

When  $n \neq i + 1$ , by employing either Eq. (B9) for  $n \in \{i, i + 2\}$  or Eq. (B11), otherwise, to take the expectation over  $\theta_{q,n}$ , we have

$$\mathbb{E}_{\theta_{[q]}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_i Z_{i+1} \rho_q] \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [Z_{i+1} Z_{i+2} \rho_q] \right) = 0. \quad (\text{C22})$$

Now we prove the non-negativity of the generated terms. Without loss of generality, we take the inequality in Eq. (C21) as an example. The gap between the interested term (C15) and the final term (C21) can be expressed exactly as a linear combination of all generated terms

$$\sum_k c_k \text{Tr} [\sigma_{g_i^k} \rho_0] \text{Tr} [\sigma_{g_{i+1}^k} \rho_0], \quad (\text{C23})$$

where  $c_k$  is the positive coefficient,  $\sigma_{g_i^k}$  and  $\sigma_{g_{i+1}^k}$  denote the  $N$ -qubit Pauli strings generated from  $Z_i Z_{i+1}$  and  $Z_{i+1} Z_{i+2}$ , but not themselves, respectively. With the choice of the initial state  $\rho_0 = |0\rangle\langle 0|$ , other terms like Eq. (C23) are always non-negative.

Therefore, from Eqs. (C19), (C21), and (C22), we have

$$\mathbb{E}_{\theta} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \frac{\partial C_j}{\partial \theta_{q,n}} \right) \geq \begin{cases} \gamma^{4L} \alpha^{2L-1} \beta, & n = i = j + 1, \\ \gamma^{4L} \alpha^{2L-1} \beta, & n = i + 1 = j, \\ 0, & \text{otherwise,} \end{cases}$$

and the second summation in Eq. (C1) is lower bounded by

$$\sum_{i \neq j=1}^{N-1} \sum_{q=1}^{2L} \sum_{n=1}^N \mathbb{E}_{\theta} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \frac{\partial C_j}{\partial \theta_{q,n}} \right) \geq 2(N-2) \times 2L \gamma^{4L} \alpha^{2L-1} \beta. \quad (\text{C24})$$

Combining Eq. (C12) and Eq. (C24), we have

$$\mathbb{E}_{\theta} \|\nabla_{\theta} C\|_2^2 \geq 4(N-1)L \alpha^{4L-1} \beta + 4(N-2)L \gamma^{4L} \alpha^{2L-1} \beta \geq 4(2N-3)L \gamma^{8L-2} \beta, \quad (\text{C25})$$

where the second inequality is derived by using the inequality

$$\alpha = \frac{1}{4} \left( 2 + \frac{\sin 2a\pi}{a\pi} \right) \geq \gamma^2 = \left( \frac{\sin a\pi}{a\pi} \right)^2, \quad a \in (0, 1).$$

Finally, we determine the hyperparameter  $a$ . Let  $x$  denote  $(a\pi)^2$ . It can be verified that for  $a \in (0, 1)$ ,

$$\begin{aligned} \gamma & \geq 1 - \frac{1}{6} a^2 \pi^2 = 1 - \frac{1}{6} x, \\ \beta & \geq \frac{1}{3} a^2 \pi^2 - \frac{1}{15} a^4 \pi^4 = \frac{1}{3} x - \frac{1}{15} x^2. \end{aligned}$$

In order to ensure the positivity of the right sides of the two inequalities above, we further set  $a \in \left(0, \frac{\sqrt{5}}{\pi}\right)$ , i.e.,  $x \in (0, 5)$ . Then, the lower bound in Eq. (8), that is, Eq. (C25) can be further bounded by

$$4(2N-3)L \left( \frac{1}{3} x - \frac{1}{15} x^2 \right) \left( 1 - \frac{1}{6} x \right)^{8L-2} \triangleq g(N, L, x).$$

It can be verified that  $g(N, L, x)$  attains its maximum when

$$x = \frac{40L + 7 - \sqrt{1600L^2 - 400L + 49}}{16L}.$$

Thus, we choose the hyperparameter as

$$a = \frac{1}{4\pi} \sqrt{\frac{40L + 7 - \sqrt{1600L^2 - 400L + 49}}{L}}, \quad (\text{C26})$$

which has the same scaling as  $\frac{1}{\sqrt{L}}$ , that is,  $a \in \Theta\left(\frac{1}{\sqrt{L}}\right)$ . Moreover, with the selected  $a$  in Eq. (C26),  $g(N, L, a)$  can be lower bounded by

$$g(N, L, a) \geq e^{-1}(2N - 3),$$

which is obtained by the fact that the function  $\frac{1}{2N-3}g(N, L, a)$  decreases monotonically with  $L$  and converges to  $e^{-1}$ .  $\square$

#### Appendix D: Proof of Theorem 1

We continue to utilize the notation in Appendix C. In addition, for each term  $\sigma_{\mathbf{i}}$  in Eq. (3), denote  $S^{\mathbf{i}}$  as the number of nonzero components in the vector  $\mathbf{i} = (i_1, i_2, \dots, i_N)$ . According to the definition of  $S$ -local cost function in the form of Eq. (3), we have  $S^{\mathbf{i}} \leq S$ . The number of elements being  $j$  in vector  $\mathbf{i}$  is denoted by  $S_j^{\mathbf{i}}$ , for  $j \in \{1, 2, 3\}$ . Thus, we have

$$S^{\mathbf{i}} = S_1^{\mathbf{i}} + S_2^{\mathbf{i}} + S_3^{\mathbf{i}} \leq S. \quad (\text{D1})$$

We introduce  $I_{S^{\mathbf{i}}} \triangleq \{n | i_n \neq 0, n \in [N]\}$  to denote the set of qubits upon which the term  $\sigma_{\mathbf{i}}$  acts non-trivially. Similarly, we define  $I_{S_j^{\mathbf{i}}} \triangleq \{n | i_n = j, n \in [N]\}$  for  $j \in \{1, 2, 3\}$ .

Moreover, for  $\mathbf{i} = (i_1, i_2, \dots, i_N)$ ,  $i_j \in \{0, 1, 2, 3\}$ ,  $j \in [N]$ , let  $\mathbf{3} : \mathbf{i}$  denote the  $N$ -dimensional vector after replacing all nonzero components in  $\mathbf{i}$  with 3, and  $\mathbf{3} : \mathbf{i}; \mathbf{1}$  denote the  $N$ -dimensional vector after replacing all 1s in  $\mathbf{i}$  with 3.

*Proof.* For the general Hamiltonian of Eq. (3), the expectation of the squared norm of the gradient of the cost function can be described as

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}} \|\nabla_{\boldsymbol{\theta}} C\|_2^2 &= \sum_{q=1}^{2L} \sum_{n=1}^N \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C}{\partial \theta_{q,n}} \right)^2 \\ &= \sum_{q=1}^{2L} \sum_{n=1}^N \mathbb{E}_{\boldsymbol{\theta}} \left( \sum_{\mathbf{i} \in \mathcal{N}} \frac{\partial C_{\mathbf{i}}}{\partial \theta_{q,n}} \right)^2 \\ &= \sum_{\mathbf{i} \in \mathcal{N}} \sum_{q=1}^{2L} \sum_{n=1}^N \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_{\mathbf{i}}}{\partial \theta_{q,n}} \right)^2 + \sum_{\mathbf{i} \neq \mathbf{j} \in \mathcal{N}} \sum_{q=1}^{2L} \sum_{n=1}^N \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_{\mathbf{i}}}{\partial \theta_{q,n}} \frac{\partial C_{\mathbf{j}}}{\partial \theta_{q,n}} \right), \end{aligned} \quad (\text{D2})$$

where  $C_{\mathbf{i}} = \text{Tr} [\sigma_{\mathbf{i}} U(\boldsymbol{\theta}) \rho U^\dagger(\boldsymbol{\theta})]$ . Thus, we turn to analyze the lower bound of Eq. (D2).

For the first summation in Eq. (D2), we have

$$\sum_{\mathbf{i} \in \mathcal{N}} \sum_{q=1}^{2L} \sum_{n=1}^N \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_{\mathbf{i}}}{\partial \theta_{q,n}} \right)^2 \geq \sum_{\mathbf{i} \in \mathcal{N}} \sum_{q=1}^{2L-2} \sum_{n=1}^N \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_{\mathbf{i}}}{\partial \theta_{q,n}} \right)^2.$$

Note that, for  $q \in [2L - 2]$ ,

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \right)^2 &= \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [\boldsymbol{\sigma}_i \rho_{2L}] \right)^2 \\ &= \mathbb{E}_{\boldsymbol{\theta}_{[2L-1]}} \mathbb{E}_{\boldsymbol{\theta}_{2L}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [\boldsymbol{\sigma}_i R_{2L}(\boldsymbol{\theta}_{2L}) \rho_{2L-1} R_{2L}^\dagger(\boldsymbol{\theta}_{2L})] \right)^2 \\ &\geq \alpha^{S_3^i} \beta^{S_1^i} \mathbb{E}_{\boldsymbol{\theta}_{[2L-1]}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [\boldsymbol{\sigma}_{\mathbf{3};i;1} \rho_{2L-1}] \right)^2 \end{aligned} \quad (\text{D3})$$

$$\geq \alpha^{S_1^i + 2S_3^i} \beta^{S_1^i + S_2^i} \mathbb{E}_{\boldsymbol{\theta}_{[2L-2]}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [\text{CZ}_L^\dagger \boldsymbol{\sigma}_{\mathbf{3};i} \text{CZ}_L \rho_{2L-2}] \right)^2 \quad (\text{D4})$$

$$= \alpha^{S_1^i + 2S_3^i} \beta^{S_1^i + S_2^i} \mathbb{E}_{\boldsymbol{\theta}_{[2L-2]}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [\boldsymbol{\sigma}_{\mathbf{3};i} \rho_{2L-2}] \right)^2, \quad (\text{D5})$$

where Eq. (D3) is obtained by using Eq. (B16) up to  $S_3^i + S_1^i$  times with respect to  $\theta_{2L,n}, n \in I_{S_3^i} \cup I_{S_1^i}$ , and using Eq. (B10) up to  $N - S_3^i - S_1^i$  times with respect to the other components in  $\boldsymbol{\theta}_{2L}$ . The appearance of the coefficient  $\alpha^{S_3^i} \beta^{S_1^i}$  is due to the fact that each time when Eq. (B16) is employed we only keep the item with more Pauli  $Z$  tensors in the corresponding  $N$ -qubit Pauli string observable. Eq. (D4) is derived by using Eq. (B16) up to  $S^i$  times with respect to  $\theta_{2L-1,n}, n \in I_{S^i}$  with the appearance of the coefficient  $\alpha^{S_1^i + S_3^i} \beta^{S_2^i}$ , and using Eq. (B10) up to  $N - S^i$  times with respect to the remain components in  $\boldsymbol{\theta}_{2L-1}$ . Eq. (D5) follows the same reasoning behind Eq. (C6).

After taking the expectations over  $\boldsymbol{\theta}_{2L}$  and  $\boldsymbol{\theta}_{2L-1}$ , the observable remained in Eq. (D5) is  $\boldsymbol{\sigma}_{\mathbf{3};i}$ , which is an  $N$ -qubit tensor product composed of identity and at most  $S$  Pauli  $Z$  operators. Thus, to bound the expectation term in Eq. (D5), we can employ the same technique used for the interested term (C11) related to the original observable  $Z_i Z_{i+1}$ , specifically,

$$\mathbb{E}_{\boldsymbol{\theta}_{[2L-2]}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [\boldsymbol{\sigma}_{\mathbf{3};i} \rho_{2L-2}] \right)^2 \geq \begin{cases} \alpha^{2S^i(L-1)-1} \beta, & n \in I_{S^i}, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the first summation in Eq. (D2) is lower bounded by

$$\begin{aligned} \sum_{i \in \mathcal{N}} \sum_{q=1}^{2L} \sum_{n=1}^N \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \right)^2 &\geq (2L-2) \sum_{i \in \mathcal{N}} S^i \alpha^{2S^i(L-1) + S_1^i + 2S_3^i - 1} \beta^{S_1^i + S_2^i + 1} \\ &\geq (2L-2) \sum_{i \in \mathcal{N}} S^i \alpha^{2SL-1} \beta^{S+1} \end{aligned} \quad (\text{D6})$$

$$\geq |\mathcal{N}|(2L-2) \alpha^{2SL-1} \beta^{S+1}. \quad (\text{D7})$$

Here, Eq. (D6) is due to Eq. (D1), and Eq. (D7) is due to the non-triviality of each term in the general Hamiltonian of Eq. (3), that is,  $S^i \geq 1$ , for all  $i \in \mathcal{N}$ .

Next, we calculate the second summation in Eq. (D2). Note that

$$\mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \frac{\partial C_j}{\partial \theta_{q,n}} \right) = \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [\boldsymbol{\sigma}_i \rho_{2L}] \frac{\partial}{\partial \theta_{q,n}} \text{Tr} [\boldsymbol{\sigma}_j \rho_{2L}] \right). \quad (\text{D8})$$

We first consider the case where  $I_{S^i} \cap I_{S^j} = \emptyset$ . Similar to obtaining Eqs. (C13) and (C14), we have

$$\mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \frac{\partial C_j}{\partial \theta_{q,n}} \right) = 0. \quad (\text{D9})$$

Then we consider the case where  $I_{S^i} \cap I_{S^j} \neq \emptyset$ . Following the technique used for the case  $i+1=j$  and the form of Eq. (C23) in Appendix C, Eq. (D8) can be rewritten as a sum of non-negative terms

$$\mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \frac{\partial C_j}{\partial \theta_{q,n}} \right) = \sum_k c_k \text{Tr} [\boldsymbol{\sigma}_{g_i^k} \rho_0] \text{Tr} [\boldsymbol{\sigma}_{g_j^k} \rho_0] \geq 0, \quad (\text{D10})$$

where  $c_k$  denotes the positive coefficient,  $\boldsymbol{\sigma}_{g_i^k}$  and  $\boldsymbol{\sigma}_{g_j^k}$  denote the  $N$ -qubit Pauli strings generated from  $\boldsymbol{\sigma}_i$  and  $\boldsymbol{\sigma}_j$ , respectively. The non-negativity is owing to the chosen initial state  $\rho_0 = |0\rangle\langle 0|$ .



Therefore, the second summation in Eq. (D2) is non-negative, i.e.,

$$\sum_{i \neq j \in \mathcal{N}} \sum_{q=1}^{2L} \sum_{n=1}^N \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{\partial C_i}{\partial \theta_{q,n}} \frac{\partial C_j}{\partial \theta_{q,n}} \right) \geq 0. \quad (\text{D11})$$

Combining Eq. (D7) and Eq. (D11), we have Eq. (4).

Finally, we determine the hyperparameter  $a$  by maximizing the lower bound in Eq. (4) denoted by

$$G(S, L, a) \triangleq 2|\mathcal{N}|(L-1)\alpha^{2SL-1}\beta^{S+1}.$$

It can be verified that  $G(S, L, a)$  attains its maximum when  $a$  is chosen as the solution of the following equation

$$\frac{\sin 2a\pi}{2a\pi} = \frac{S(2L-1)-2}{S(2L+1)},$$

and the solution  $a$  has the same scaling as  $\frac{1}{\sqrt{L}}$ , that is,  $a \in \Theta(\frac{1}{\sqrt{L}})$ . Moreover, with the selected  $a$ , we have

$$G(S, L, a) \geq 2|\mathcal{N}| \left( \frac{S+1}{eS} \right)^{S+1} \frac{L-1}{(2L+1)^{S+1}},$$

which is obtained by the fact that the function  $\left[ \frac{2SL-1}{S(2L+1)} \right]^{2SL-1}$  decreases monotonically with  $L$  and converges to  $\frac{1}{e^{S+1}}$ , when  $S \geq 1$ .  $\square$

## Appendix E: Proof of Lemma 4

*Proof.* Let us first prove Eq. (15) concerning the average of the partial derivative of the cost function with respect to  $\theta_{q,n}$ .

For the Hamiltonian of Eq. (7),

$$\mathbb{E}_{\boldsymbol{\theta}} \partial_{\theta_{q,n}} C = \sum_{i=1}^{N-1} \mathbb{E}_{\boldsymbol{\theta}} \partial_{\theta_{q,n}} C_i,$$

where  $C_i = \text{Tr} [Z_i Z_{i+1} U(\boldsymbol{\theta}) \rho U^\dagger(\boldsymbol{\theta})]$ . To proceed, assume that  $\theta_{q,n}$  lies in the  $l$ th block. Then, we have

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}} \partial_{\theta_{q,n}} C_i &= \mathbb{E}_{\boldsymbol{\theta}_{[2L]}} \partial_{\theta_{q,n}} \text{Tr} [Z_i Z_{i+1} \rho_{2L}] \\ &= \gamma^2 \mathbb{E}_{\boldsymbol{\theta}_{[2L-1]}} \partial_{\theta_{q,n}} \text{Tr} [Z_i Z_{i+1} \rho_{2L-1}] \end{aligned} \quad (\text{E1})$$

$$= \gamma^4 \mathbb{E}_{\boldsymbol{\theta}_{[2L-2]}} \partial_{\theta_{q,n}} \text{Tr} [C Z_L^\dagger Z_i Z_{i+1} C Z_L \rho_{2L-2}] \quad (\text{E2})$$

$$= \gamma^4 \mathbb{E}_{\boldsymbol{\theta}_{[2L-2]}} \partial_{\theta_{q,n}} \text{Tr} [Z_i Z_{i+1} \rho_{2L-2}] \quad (\text{E3})$$

$$= \gamma^{4(L-l)} \mathbb{E}_{\boldsymbol{\theta}_{[2l]}} \partial_{\theta_{q,n}} \text{Tr} [Z_i Z_{i+1} \rho_{2l}]. \quad (\text{E4})$$

Here, the appearance of  $\gamma^2$  in Eq. (E1) is due to taking the expectations over  $\theta_{2L,i}$  and  $\theta_{2L,i+1}$  by employing Eq. (B1) and taking the expectations over the other  $N-2$  parameters in  $\boldsymbol{\theta}_{2L}$  by utilizing Eq. (B3). Following the same reasoning, Eq. (E2) is derived. In addition, Eq. (E3) is due to Lemma 10. Eqs. (E1)-(E3) comprise a complete calculation for the  $L$ th block in Fig. 1(a). For the following  $L-l-1$  blocks, we repeat the same procedures and obtain Eq. (E4).

Now we proceed upon the index  $n$ . For  $n \notin \{i, i+1\}$ , from Eq. (B4), we have

$$\mathbb{E}_{\theta_{q,n}} \partial_{\theta_{q,n}} \text{Tr} [Z_i Z_{i+1} \rho_q] = 0,$$

which results in  $\mathbb{E}_{\boldsymbol{\theta}} \partial_{\theta_{q,n}} C = 0$ .

For  $n \in \{i, i+1\}$ , we first consider the case where  $q = 2l$ . Note that by employing Eq. (B2) for  $\theta_{q,n}$ , we have

$$\mathbb{E}_{\theta_{2l}} \partial_{\theta_{2l,n}} \text{Tr} [Z_i Z_{i+1} \rho_{2l}] = \begin{cases} -\gamma^2 \text{Tr} [X_i Z_{i+1} \rho_{2l-1}], & n = i \\ -\gamma^2 \text{Tr} [Z_i X_{i+1} \rho_{2l-1}], & n = i+1. \end{cases} \quad (\text{E5})$$

Moreover, by employing Eq. (B1), we have

$$\begin{cases} \mathbb{E}_{\theta_{2l-1}} \text{Tr} [X_i Z_{i+1} \rho_{2l-1}] = \gamma \text{Tr} [CZ_l^\dagger X_i Z_{i+1} CZ_l \rho_{2l-2}], & n = i \\ \mathbb{E}_{\theta_{2l-1}} \text{Tr} [Z_i X_{i+1} \rho_{2l-1}] = \gamma \text{Tr} [CZ_l^\dagger Z_i X_{i+1} CZ_l \rho_{2l-2}], & n = i+1. \end{cases} \quad (\text{E6})$$

Combining Eq. (E5) and Eq. (E6), we obtain

$$\mathbb{E}_{\theta_{2l-1}} \mathbb{E}_{\theta_{2l}} \partial_{\theta_{2l,n}} \text{Tr} [Z_i Z_{i+1} \rho_{2l}] = \begin{cases} -\gamma^3 \text{Tr} [CZ_l^\dagger X_i Z_{i+1} CZ_l \rho_{2l-2}], & n = i \\ -\gamma^3 \text{Tr} [CZ_l^\dagger Z_i X_{i+1} CZ_l \rho_{2l-2}], & n = i+1. \end{cases} \quad (\text{E7})$$

Similarly, for  $n \in \{i, i+1\}$  and  $q = 2l-1$ , we have

$$\begin{aligned} \mathbb{E}_{\theta_{2l-1}} \mathbb{E}_{\theta_{2l}} \partial_{\theta_{2l-1,n}} \text{Tr} [Z_i Z_{i+1} \rho_{2l}] &= \gamma^2 \mathbb{E}_{\theta_{2l-1}} \partial_{\theta_{2l-1,n}} \text{Tr} [Z_i Z_{i+1} \rho_{2l-1}] \\ &= \begin{cases} \gamma^4 \text{Tr} [CZ_l^\dagger Y_i Z_{i+1} CZ_l \rho_{2l-2}], & n = i \\ \gamma^4 \text{Tr} [CZ_l^\dagger Z_i Y_{i+1} CZ_l \rho_{2l-2}], & n = i+1. \end{cases} \end{aligned} \quad (\text{E8})$$

According to Lemma 10, after applying the  $l$ th entanglement layer, the generated observables in Eqs. (E7) and (E8) are all  $N$ -qubit Pauli strings with the  $n$ th tensor component unchanged and being either Pauli  $X$  or Pauli  $Y$ .

Subsequently, from Eqs. (B1) and (B3), taking the expectations over  $\theta_{q-1}, \dots, \theta_1$  does not change the observables any longer. Moreover, applying the entanglement layers does not change the  $n$ th tensor component of the observables in Eqs. (E7) and (E8), which is either Pauli  $X$  or Pauli  $Y$ . Thus, for  $n \in \{i, i+1\}$ , we have

$$\mathbb{E}_{\theta_{[2l]}} \partial_{\theta_{q,n}} \text{Tr} [Z_i Z_{i+1} \rho_{2l}] = c \text{Tr} [\sigma_{g_i} \rho_0], \quad (\text{E9})$$

where  $c$  is a positive coefficient,  $\sigma_{g_i}$  denotes the final Pauli string generated from  $Z_i Z_{i+1}$ , whose  $n$ th tensor component is Pauli  $X$  for  $q = 2l$  and Pauli  $Y$  for  $q = 2l-1$ . Moreover, with the choice of  $\rho_0 = |0\rangle\langle 0|$ , Eq. (E9) is equal to zero, which results in  $\mathbb{E}_{\theta} \partial_{\theta_{q,n}} C = 0$ .

Therefore, we have proved Eq. (15). The proof of Eqs. (16)-(21) can be straightforwardly derived from the proof in Appendix C.  $\square$

## Appendix F: Proof of Theorem 3

*Proof.* The proof of Eqs. (11) and (14) can be straightforwardly derived from the proof in Appendix D. We now focus on proving Eq. (12) when the number of the elements in vector  $\mathbf{i}$  belonging to  $\{1, 2\}$  is not equal to 1 for all  $\mathbf{i} \in \mathcal{N}$ .

Following the notation in Appendix D, we restate Eq. (12). Specifically, if  $S_1^i + S_2^i \neq 1$  for all  $\mathbf{i} \in \mathcal{N}$ , for an arbitrary trainable parameter  $\theta_{q,n}$ ,

$$\mathbb{E}_{\theta} \partial_{\theta_{q,n}} C = \sum_{\mathbf{i} \in \mathcal{N}} \mathbb{E}_{\theta} \partial_{\theta_{q,n}} C_{\mathbf{i}} = 0, \quad (\text{F1})$$

where  $C_{\mathbf{i}} = \text{Tr} [\sigma_{\mathbf{i}} U(\theta) \rho U^\dagger(\theta)]$ . To prove it, we assume that  $\theta_{q,n}$  lies in the  $l$ th block.

First, if  $S_1^i + S_2^i = 0$ , we can directly generalize the proof in Appendix E to obtain

$$\mathbb{E}_{\theta} \partial_{\theta_{q,n}} C_{\mathbf{i}} = 0. \quad (\text{F2})$$

Next, if  $S_1^i + S_2^i \geq 2$ , note that

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}} \partial_{\theta_{q,n}} C_i &= \mathbb{E}_{\boldsymbol{\theta}_{[2L]}} \partial_{\theta_{q,n}} \text{Tr} [\boldsymbol{\sigma}_i \rho_{2L}] \\ &= \gamma^{S_1^i + S_3^i} \mathbb{E}_{\boldsymbol{\theta}_{[2L-1]}} \partial_{\theta_{q,n}} \text{Tr} [\boldsymbol{\sigma}_i \rho_{2L-1}] \end{aligned} \quad (\text{F3})$$

$$= \gamma^{S_1^i + S_3^i} \mathbb{E}_{\boldsymbol{\theta}_{[2L-2]}} \partial_{\theta_{q,n}} \text{Tr} [\text{CZ}_L^\dagger \boldsymbol{\sigma}_i \text{CZ}_L \rho_{2L-2}] \quad (\text{F4})$$

$$= \gamma^{(S_1^i + S_3^i)(L-l)} \mathbb{E}_{\boldsymbol{\theta}_{[2l]}} \partial_{\theta_{q,n}} \text{Tr} [\boldsymbol{\sigma}_{g_{i,2l}} \rho_{2l}]. \quad (\text{F5})$$

Here, Eqs. (F3) and (F4) are derived from the same reasoning as Eqs. (E1) and (E2), respectively. According to Lemma 10, after applying the entanglement layer in Eq. (F4), the position set of the tensor components being  $X$  or  $Y$  does not change, being equal to  $I_{S_1^i} \cup I_{S_2^i}$ . Then we repeat the above procedures for the following  $L-l-1$  blocks and obtain Eq. (F5), where  $\boldsymbol{\sigma}_{g_{i,2l}}$  denotes the generated observable after calculating the last  $L-l$  blocks with the invariant set  $I_{S_1^{g_{i,2l}}} \cup I_{S_2^{g_{i,2l}}} = I_{S_1^i} \cup I_{S_2^i}$ . By now,  $S_1^{g_{i,2l}} + S_2^{g_{i,2l}} = S_1^i + S_2^i \geq 2$ .

For the  $l$ th block where  $\theta_{q,n}$  lies, from Lemma 6, when taking the expectations over  $\boldsymbol{\theta}_{2l}$  and  $\boldsymbol{\theta}_{2l-1}$ , the number of the tensor components being  $X$  or  $Y$  changes, which either increases or decreases by at most 1. In fact, the possible change only happens when employing Eq. (B2). Moreover, according to Lemma 10, after calculating the  $l$ th block, we have

$$S_1^{g_{i,2(l-1)}} + S_2^{g_{i,2(l-1)}} \geq S_1^i + S_2^i - 1 \geq 1.$$

During the calculations of the expectations over  $\boldsymbol{\theta}_{[2(l-1)]}$ , from Lemmas 6 and 10, the number of the tensor components being  $X$  or  $Y$  does not change any longer. Since

$$\mathbb{E}_{\boldsymbol{\theta}} \partial_{\theta_{q,n}} C_i \propto \text{Tr} [\boldsymbol{\sigma}_{g_{i,0}} \rho_0],$$

and at least one of the tensor components in  $\boldsymbol{\sigma}_{g_{i,0}}$  is  $X$  or  $Y$ , with the selected  $\rho_0 = |0\rangle\langle 0|$ , we have

$$\mathbb{E}_{\boldsymbol{\theta}} \partial_{\theta_{q,n}} C_i = 0. \quad (\text{F6})$$

Thus, Eq. (12) is proved and so is Eq. (13).  $\square$