Symmetric Pruning in Quantum Neural Networks

Anonymous authors
Paper under double-blind review

Abstract

Many fundamental properties of a quantum system are captured by its Hamiltonian and ground state. Despite the significance, ground states preparation (GSP) is classically intractable for large-scale Hamiltonians. Quantum neural networks (QNNs), which exert the power of modern quantum machines, have emerged as a leading protocol to conquer this issue. As such, the performance enhancement of QNNs becomes the core in GSP. Empirical evidence showed that QNNs with handcraft symmetric ansätze generally experience better trainability than those with asymmetric ansätze, while theoretical explanations remain vague. To fill this knowledge gap, here we propose the effective quantum neural tangent kernel (EQNTK) and connect this concept with over-parameterization theory to quantify the convergence of QNNs towards the global optima. We uncover that the advance of symmetric ansätze attributes to their large EQNTK value with low effective dimension, which requests few parameters and quantum circuit depth to reach the over-parameterization regime permitting a benign loss landscape and fast convergence. Guided by EQNTK, we further devise a symmetric pruning (SP) scheme to automatically tailor a symmetric ansatz from an over-parameterized and asymmetric one to greatly improve the performance of QNNs when the explicit symmetry information of Hamiltonian is unavailable. Extensive numerical simulations are conducted to validate the analytical results of EQNTK and the effectiveness of SP.

1 Introduction

The law of quantum mechanics advocates that any quantum system can be described by a Hamiltonian, and many important physical properties are reflected by its ground state. For this reason, the ground state preparation (GSP) of Hamiltonians is the key to understanding and fabricating novel quantum matters. Due to the intrinsic hardness of GSP (Poulin & Wocjan, 2009; Carleo et al., 2019), the required computational resources of classical methods are unaffordable when the size of Hamiltonian becomes large. Quantum computers, whose operations can harness the strength of quantum mechanics, promise to tackle this problem with potential computational merits. In the noisy intermediate-scale quantum (NISQ) era (Preskill, 2018), quantum neural networks (QNNs) (Farhi & Neven, 2018; Cong et al., 2019; Cerezo et al., 2021a) are leading candidates toward this goal. The building blocks of QNNs, analogous to deep neural networks, consist of variational ansätze (or parameterized quantum circuits) and classical optimizers. In order to enhance the power of QNNs in GSP, great efforts have been made to design advanced ansätze with varied circuit structures (Peruzzo et al., 2014; Wecker et al., 2015; Kandala et al., 2017).

Despite the achievements aforementioned, recent progress has shown that QNNs may suffer from severe trainability issues when the circuit depth of ansätze is either shallow or deep. Namely, for the deep ansätze, the magnitude of the gradients exponentially decays with the increased system size (McClean et al., 2018; Cerezo et al., 2021b). This phenomenon, dubbed the barren plateau, hints at the difficulty of optimizing deep QNNs, where an exponential runtime is necessitated for convergence. The wisdom to alleviate barren plateaus is exploiting shallow ansätze to accomplish learning tasks (Grant et al., 2019; Skolik et al., 2021; Zhang et al., 2020; Pesah et al., 2021), while the price to pay is incurring another

serious trainability issue—convergence (Boyd & Vandenberghe, 2004; Du et al., 2021). The trainable parameters may get stuck into sub-optimal local minima or saddle points with high probability because of the unfavorable loss landscape (Anschuetz, 2021; Anschuetz & Kiani, 2022). Orthogonal to these negative results, several studies pointed out that when the depth of ansätze becomes overwhelmingly deep and surpasses a critical point, the over-parameterized QNNs embrace a benign landscape and permit fast convergence towards good local minima (Kiani et al., 2020; Wiersema et al., 2020; Larocca et al., 2021b). Nevertheless, the criteria to reach such a critical point is stringent, i.e., the number of parameterized gates or the circuit depth scales exponentially with the problem size, which hurdles the application of over-parameterized QNNs in practice.

Empirical evidence sheds new light on exploiting over-parameterized QNNs to tackle GSP. QNNs with symmetric ansätze only demand a polynomial number of trainable parameters and the circuit depth with the problem size to reach the over-parameterized region and achieve a fast convergence rate (Herasymenko & O'Brien, 2021; Gard et al., 2020; Zheng et al., 2021; 2022; Shaydulin & Wild, 2021; Mernyei et al., 2022; Marvian, 2022; Meyer et al., 2022; Larocca et al., 2022; Sauvage et al., 2022). A common feature of these symmetric ansätze is capitalizing on the symmetric properties underlying the problem Hamiltonian to shrink the solution space and facilitate seeking near-optimal solutions. Unfortunately, current symmetric ansätze are inapplicable to a broad class of Hamiltonians whose symmetry is implicit, since their constructions rely on the explicit information for the symmetry of Hamiltonians. Additionally, it remains largely obscure whether the symmetry contributes to lowering the critical point to reach the over-parameterization regime.

In this work, we fill the above knowledge gap from both theoretical and practical aspects. Concretely, we develop a novel notion—effective quantum neural tangent kernel (EQNTK) to capture the training dynamic of various ansätze via their effective dimension. In doing so, we expose that compared with the asymmetric ansätze, the symmetric ansätze possess dramatically lower effective dimensions and the required number of parameters and circuit depth to reach the over-parameterization may polynomially scale with the problem size. We next exploit EQNTK to prove that when the condition of over-parameterization is satisfied, the trainable parameters of QNNs with symmetric ansätze can exponentially converge to the global optima with the increased iterations. Taken together, our analysis recognizes that over-parameterized QNNs with symmetric ansätze is a possible solution toward large-scale GSP tasks. Envisioned by EQNTK and pruning in deep neural networks (Han et al., 2015; Blalock et al., 2020; Frankle et al., 2020; Wang et al., 2022), we further devise a symmetric pruning scheme (SP) to automatically tailor a symmetric ansatz from an over-parameterized and asymmetric ansatz with the enhanced trainability and applicability. Analogous to the pruning method, SP continuously eliminates the redundant quantum gates from the given asymmetric ansatz and correlates parameters to assign different types of symmetries on the slimmed ansatz. In this way, SP generates a symmetric ansatz with a modest depth with a fast convergence guarantee. Extensive numerical simulations on transverse-field Ising model and the maximum cut problem of random graphs validate the analytical property of EQNKT and the effectiveness of SP. These results deepen our understating about how to merge symmetry with over-parameterization theory and indicate the signification of designing symmetric ansätze.

We summarize our main contributions as follows:

1. EQNTK, easy to understand QNN in GSP xxx;
2. Theoretical bound (recap Thm 1,2), the dependence of factors;
3. SP scheme xxx;

2 Effective QNTK allows an improved convergence of QNNs

Here we establish foundations about why symmetric ansätze have the ability to enhance the trainability of QNNs in ground state preparation (GSP) tasks. To do so, we propose a novel concept—effective quantum neural tangent kernel (EQNTK), to reconcile the QNTK theory with the symmetry of the problem Hamiltonian. Attributed to EQNTK, we uncover that the

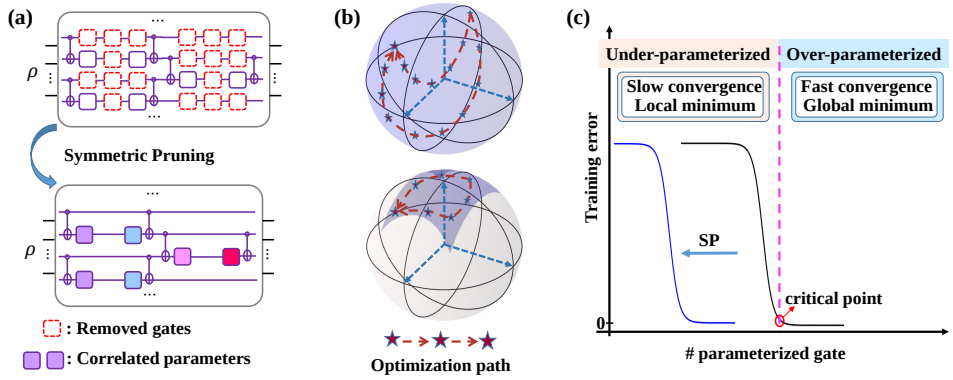

Figure 1: **(a) Schematic of symmetric pruning.** The proposed symmetric pruning (SP) distills the symmetric ansatz from a given asymmetric ansatz, completed by removing the redundant gates (highlighted by the red dashed boxes) and correlating the parameters in the gate respecting the spatial symmetry (highlighted by the solid boxes with the same color). **(b) Training dynamics of QNNs with symmetric ansätze.** The upper and lower panels illustrate the dynamic of variational states corresponding to the asymmetric and symmetric ansätze, respectively. The shadow region refers to the solution space, which is the whole Hilbert space for asymmetric ansatz, and the restricted invariant subspace for symmetric ansatz. **(c) The critical point of the over-parameterized regime.** When the number of parameters is beyond the critical point (highlighted by the red circle), the training error exponentially converges to a nearly global minimum. Compared with asymmetric ansätze, symmetric ansätze (highlighted by the blue curve) require few parameters to reach the critical point. (Decompose into small figures. (c) should be posited in Introduction for intuition.)

advance of symmetric ansätze originates from their ability to dramatically decrease the over-parameterization threshold. For elucidating, we first interpret the necessary backgrounds in Sec. 2.1. Then, we present our main theoretical results in Sec. 2.2.

2.1 Problem setup

Ground state preparation. Given an n-qubit Hamiltonian $H \in \mathbb{C}^{2^n \times 2^n}$, GSP aims to find the eigenvector $|\psi^*\rangle \in \mathbb{C}^{2^n}$ (i.e., the ground state) of H corresponding to its minimum eigenvalue. For any n-qubit state $|\psi\rangle$, we have $\langle\psi^*| H |\psi^*\rangle \leq \langle\psi| H |\psi\rangle$ and the equality is satisfied iff $|\psi\rangle = |\psi^*\rangle$. Since the dimension of $|\psi^*\rangle$ scales exponentially with n, GSP is classically intractable for a large n.

Quantum neural networks. A QNN can be described by a triplet $(|\psi_0\rangle, U(\boldsymbol{\theta}), H)$. When it is applied to solve GSP, an ansatz $U(\boldsymbol{\theta})$ (i.e., a parameterized unitary) prepares a variational state $|\psi(\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})|\psi_0\rangle$ with the fixed input state $|\psi_0\rangle$. The parameters $\boldsymbol{\theta}$ are optimized to minimize the loss function

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\left(\langle\psi_0| U(\boldsymbol{\theta})^\dagger H U(\boldsymbol{\theta}) |\psi_0\rangle - E_0\right)^2, \tag{1}$$

where $E_0 = \langle\psi^*| H |\psi^*\rangle$ refers to the ground state energy. As shown in Fig. 1(a), the optimization follows an iterative manner, i.e., the classical optimizer continuously leverages the output of the quantum circuits to update $\boldsymbol{\theta}$ and the update rule is $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta\partial\mathcal{L}(\boldsymbol{\theta}^{(t)})/\partial\boldsymbol{\theta}$, where η refers to the learning rate.

Note that we adopt E_0 to facilitate the convergence analysis. Our results cover general loss functions where E_0 is replaced by $C \in \mathbb{R}$ with $C \leq E_0$. See Appendix A for details.

Constructions of (symmetric) ansätze. The power of QNNs depends on the employed ansatz $U(\boldsymbol{\theta})$. A general form of $U(\boldsymbol{\theta})$ covering many typical ansätze such as Hamiltonian variational ansatz (HVA) and hardware efficient ansatz (HEA) (Bharti et al., 2022; Qian

et al., 2021) yields

$$U(\boldsymbol{\theta}) = \prod_{l=1}^{L} U_l(\boldsymbol{\theta}_l), \quad U_l(\boldsymbol{\theta}_l) = \prod_{k=1}^{K} e^{-iG_k \boldsymbol{\theta}_{lk}}, \tag{2}$$

where L refers to the layer number, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_L) \in \Theta \subseteq \mathbb{R}^{LK}$ is trainable parameters living in the parameter space Θ, $\boldsymbol{\theta}_l = (\boldsymbol{\theta}_{l1}, \cdots, \boldsymbol{\theta}_{lK})$ is trainable parameters at the l-th layer, and $\mathcal{A} = \{G_1, \cdots, G_K\}$ is a set of Hermitian traceless operators called an *ansatz design*. Given Θ and $\mathcal{A}$, a set of trainable unitary matrices of QNNs forms a subgroup of $SU(2^n)$ with $\mathcal{U}_{\mathcal{A}} = \cup_{L=0}^{\infty}\{U(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$.

The difference of ansätze originates from the varied Θ and $\mathcal{A}$. To construct a *symmetric* ansatz $U(\boldsymbol{\theta})$, $\mathcal{A}$ should be restricted such that each element in the unitary group $\mathcal{U}_{\mathcal{A}}$ is expected to commute with a Hermitian matrix Σ. This leads to a direct sum decomposition $\mathcal{H} = \oplus_{j=1}^{p} V_j$ of the quantum state space, where V_j is the invariant subspace spanned by the eigenvectors corresponding to distinct eigenvalues of Σ. More precisely, denote $\Sigma = \sum_{j=1}^{p}\sum_{k=1}^{s_j} \lambda_j \boldsymbol{v}_{jk}$ where λ_j is the eigenvalue with $\lambda_i \neq \lambda_j$ for $i \neq j$, $\boldsymbol{v}_{jk}$ is the corresponding eigenvector, and $\sum_{j=1}^{p} s_j = 2^n$, then V_j is spanned by the eigenvectors $\{\boldsymbol{v}_{j1}, \cdots, \boldsymbol{v}_{js_j}\}$.

Convergence of QNNs. A crucial metric to assess the performance of different QNNs is the ϵ-convergence rate towards the global minimum $\mathcal{L}(\boldsymbol{\theta}^*)$ with $\boldsymbol{\theta}^* = \min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta})$.

Definition 1 (ϵ-convergence)**.** *A QNN instance* $(|\psi_0\rangle, U(\boldsymbol{\theta}), H)$ *achieves an* ϵ*-convergence if the trained parameters after* T *iterations* $\boldsymbol{\theta}^{(T)}$ *satisfy* $\mathcal{L}(\boldsymbol{\theta}^{(T)}) \leq \epsilon$ *with* $\epsilon \in \mathbb{R}$.

This quantity measures the distance between the estimated and the optimal loss values, which can be derived via quantum neural tangent kernel (QNTK), i.e.,

$$Q^{(t)} = \nabla\varepsilon^{(t)\top}\nabla\varepsilon^{(t)}, \tag{3}$$

where $\varepsilon^{(t)} = \langle\psi_0| U(\boldsymbol{\theta}^{(t)})^\dagger H U(\boldsymbol{\theta}^{(t)}) |\psi_0\rangle - E_0$ denotes the residual training error and $\nabla\varepsilon^{(t)}$ is the gradients of $\varepsilon^{(t)}$ with respect to $\boldsymbol{\theta}$. Liu et al. (2022b) proved the ϵ-convergence of the over-parameterized QNN with an arbitrary ansatz.

Theorem 1 (Liu et al. (2022b))**.** *Following notations in Eqns. (1)-(3), when* $U(\boldsymbol{\theta})$ *matches the Haar distribution up to the fourth moment, the number of parameters satisfies* $LK \gg 1$*, and the learning rate* $\eta \ll 1$*, the training dynamics of a QNN instance* $(|\psi_0\rangle, U(\boldsymbol{\theta}), H)$ *yields*

$$\varepsilon_t \approx (1 - \eta\bar{Q})^t \varepsilon_0 \approx e^{-\gamma t}\varepsilon_0. \tag{4}$$

where $\gamma = \eta\bar{Q}$ *is the indicator of the decay rate and* $\bar{Q} = \mathcal{O}(LK\,\mathrm{Tr}(H^2)/4^n)$ *refers to the expectation of* Q *on Haar average.*

It indicates that the critical point to reach the over-parameterization region is $|\boldsymbol{\theta}| \sim \mathcal{O}(4^n/(\eta\,\mathrm{Tr}(H^2)))$. In this setting, the exponent meets $\gamma \sim \mathcal{O}(1)$ and promises an exponential convergence. Besides, Eqn. (4) hints that the convergence rate of QNNs is continuously enhanced by increasing the value of QNTK, which can be achieved by growing the number of parameters or decreasing the system size.

2.2 Effective Quantum Neural Tangent Kernel

The exponential number of parameters with respect to the number of qubits causes the realization of the over-parameterized QNNs to be impractical for large quantum systems. Moreover, the convergence rate established by QNTK is independent of the refined structure information of ansätze. This contradicts the empirical evidence such that symmetric ansätze outperform asymmetric ansätze with fast convergence in training QNNs. It highly demands carrying out new theories to stress these issues.

Here we propose a novel concept—effective QNTK (EQNTK) to resolve the above dilemma and exhibit how symmetry improves the trainability of QNNs. An intuition is depicted in Fig. 1(b). Specifically, given a QNN $(|\psi_0\rangle, U(\boldsymbol{\theta}), H)$ whose input state $|\psi_0\rangle$ and the ground state $|\psi^*\rangle$ live in the same subspace $V^* \subset \mathbb{C}^{2^n}$ and V^* is invariant with the employed symmetric ansatz $U(\boldsymbol{\theta})$, the training dynamics of $|\psi(\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})|\psi_0\rangle$ can be captured by

V^*, a much smaller space than the whole state space. Suppose that the state space $\mathcal{H}$ under the symmetric ansatz design $\mathcal{A}$ can be decomposed into $\mathcal{H} = \oplus_{j=1}^{p} V_j$ and there exists $j^* \in [p]$ such that $V_{j^*} = V^*$ includes the input state $|\psi_0\rangle$, the ground state $|\psi^*\rangle$, and all possible variational states $\{|\psi(\boldsymbol{\theta})\rangle \,|\boldsymbol{\theta} \in \Theta\}$ by varying $\boldsymbol{\theta}$. In this regard, the behavior of $|\psi(\boldsymbol{\theta})\rangle$ can be quantitatively characterized by the dimension of the subspace $d_{\text{eff}} = |V^*|$, dubbed the *effective dimension* of the QNN.

Definition 2 (Effective dimension)**.** *Consider a QNN instance $(|\psi_0\rangle, U(\boldsymbol{\theta}), H)$ with symmetric ansatz design $\mathcal{A}$. Let V^* denote the invariant subspace of interest. The effective dimension d_{eff} is defined as the dimension of V^*. And we denote the projection on the subspace V^* as $\mathbf{\Pi} = PP^\dagger$, where $P \in \mathbb{C}^{d \times d_{\text{eff}}}$ is an arbitrary set of orthonormal basis.*

As a result, for symmetric ansätze, the value of $Q^{(t)}$ and $\bar{Q}$ in Theorem 1 is controlled by d_{eff} instead of 2^n, which reduces the threshold to reach over-parameterization and accelerates the convergence. Under these elaborations, the following theorem establishes the convergence of QNNs with symmetric ansätze under EQNTK, whose proof is deferred to Appendix B.

Theorem 2. *Consider the QNN instance $(|\psi_0\rangle, U(\boldsymbol{\theta}), H)$ with the effective dimension d_{eff}. Following notations in Eqns. (1)-(3), when the distribution of $U(\boldsymbol{\theta})$ constrained to the invariant subspace matches the Haar distribution up to the fourth moment, the number of parameters satisfies $LK \gg 1$, and the learning rate $\eta \ll 1$, the training dynamics of a QNN instance $(|\psi_0\rangle, U(\boldsymbol{\theta}), H)$ yields*

$$\varepsilon_t \approx (1 - \eta \bar{Q}_S)^t \varepsilon_0 \approx e^{-\gamma t} \varepsilon_0. \tag{5}$$

where $\bar{Q}_S = \mathcal{O}(LK \operatorname{Tr}(H^2)/d_{\text{eff}}^2)$ refers to the expectation of EQNTK Q_S on Haar average.

The above results indicate that when the number of trainable parameters LK scales with $\mathcal{O}(d_{\text{eff}}^2/(\eta \operatorname{Tr}(H^2)))$, the adopted symmetric ansatz can reach the over-parameterization regime. Compared with QNTK, the reduction in the order of $(2^n/d_{\text{eff}})^2$ not only ensures the practical value of using symmetric ansätze to solve GSP for large-scale quantum systems, but also explains the empirical observations that symmetric ansätze require fewer parameters to reach the critical point than that of the asymmetric ansätze. Moreover, in contrast with prior results that the trainability can always be improved by over-parameterization, our bound suggests that involving more parameterized gates over the critical point could degrade the convergence of QNNs as the underlying symmetry may be broken and the effective dimension d_{eff} could be large.

Remark. (i) The derived EQNTK can also be used to diagnose the barren plateau phenomenon of QNNs. Particularly, the quantity $Q_S/(LK)$ amounts to the variance of gradient whose average is zero under the 4-design assumption. In other words, when the number of parameters LK is fixed, the EQNTK should be as large as possible to avoid barren plateaus. (ii) Besides invariant subspace, the EQNTK, especially for the effective dimension, can be quantified by other metrics. An alternative is the dynamical Lie algebra (DLA) (Larocca et al., 2021a;b), which measures the controllability of the quantum system. The following lemma analyzes the convergence rate of QNNs with symmetric ansätze under the DLA-based EQNTK, whose proof is given in Appendix C.

Lemma 1. *Consider the QNN instance $(|\psi_0\rangle, U(\boldsymbol{\theta}), H)$ with the corresponding dynamical Lie algebra denoted as $\mathfrak{g}$. If there exists an invariant subspace $V_{\mathfrak{g}}$ with dimension $d_{\mathfrak{g}}$ under the DLA $\mathfrak{g}$ including the input state $|\psi_0\rangle$ and the ground state $|\psi^*\rangle$, Following notations in Eqns. (1)-(3), when the distribution of $U(\boldsymbol{\theta})$ constrained to the invariant subspace matches the Haar distribution up to the fourth moment, the number of parameters satisfies $LK \gg 1$, and the learning rate $\eta \ll 1$, the training dynamics of a QNN instance $(|\psi_0\rangle, U(\boldsymbol{\theta}), H)$ yields*

$$\varepsilon_t \approx (1 - \eta \bar{Q}_D)^t \varepsilon_0 \approx e^{-\gamma t} \varepsilon_0. \tag{6}$$

where $\bar{Q}_D = \mathcal{O}(LK \operatorname{Tr}(H^2)/d_{\mathfrak{g}}^2)$ refers to the expectation of DLA-based EQNTK Q_D on Haar average.

3 Symmetrical pruning with effective QNTK

Beyond analyzing the convergence rate, another ad-hoc topic in GSP is designing advanced ansätze to improve the trainability of QNNs. Although over-parameterization and contemporary symmetric ansätze partially address this problem, both of them have evident caveats. The former may request exponential trainable parameters to satisfy the over-parameterized condition, and the latter requires explicit information for the symmetry of the problem Hamiltonian. To compensate for these deficiencies, here we devise symmetrical pruning (SP), an automatic scheme to design symmetric ansätze with the enhanced trainability of QNNs. Conceptually, SP distills a symmetric over-parameterized ansatz from an asymmetric over-parameterized ansatz. Supported by the EQNTK theory, the extracted ansatz is resource-friendly in implementation since it holds a small effective dimension and only needs a few trainable parameters to compass the over-parameterization.

Algorithm 1: Symmetric pruning (SP)

Input : Problem Hamiltonian $H = (\sum_{j=1}^{q} \alpha_j H_j) \otimes \mathbb{I}^m$, the ansatz design $\mathcal{A}$ and the parameter space Θ in Eqn. (2).

Step 1. Initialize an over-parameterized and asymmetric ansatz via $\mathcal{A}$ and Θ;

Step 2. Symmetry identification.

2-1. Remove the gates on wires corresponding to the redundant part of H in $\mathcal{A}$, i.e., $\mathbb{I}^m$.

2-2. Remove the gates such that the pruned ansatz design $\mathcal{A}_{\text{pr}} = \{H_1, \cdots, H_q\}$.

2-3. Assign the spatial symmetry of $\mathcal{A}_{\text{pr}}$ by correlating the parameterized gates with one individual parameter and obtain $\Theta_{\text{pr}} \subseteq \Theta$.

Output: Pruned ansatz design $\mathcal{A}_{\text{pr}}$ and parameter space Θ_{pr}.

The Pseudo code of SP is summarized in Alg. 1 and its schematic illustration is shown in Fig. 1(a). Suppose the problem Hamiltonian is $H = (\sum_{j=1}^{q} \alpha_j H_j) \otimes \mathbb{I}^{\otimes m}$, where α_j is the real coefficients and H_j is the tensor product of Pauli matrices on n qubits, SP builds the symmetric ansatz of H with two primary steps, i.e., initialization and symmetry identification. The initialization step is choosing an initial over-parameterized QNN, which includes setting down the ansatz design $\mathcal{A}$ and the parameter space Θ. Note that $\mathcal{A}$ should contain all Pauli terms in H and Θ enabling an ϵ-convergence of QNNs. This can be achieved by adopting a sufficient deep hardware efficient ansatz. After initialization, the symmetry identification step iteratively discovers the system symmetry, structure symmetry, and spatial symmetry, which is completed by three sub-steps. The first sub-step is symmetrically pruning the qubit wires. That is, all qubit gates interact with the redundant part of H, i.e., the identity term $\mathbb{I}^{\otimes m}$, are removed. The second sub-step is symmetrically pruning the structure. This step drops the parameterized single-qubit gates and the two-qubit gates so that the pruned ansatz design $\mathcal{A}_{\text{pr}}$ can be block diagonalized under the projection on the eigenspace of $H = \sum_{j=1}^{q} \alpha_j H_j$. Mathematically, this can be achieved by setting $\mathcal{A}_{\text{pr}} = \{H_1, \cdots, H_q\}$ and thus the pruned ansatz $U_{\text{pr}}(\boldsymbol{\theta})$ takes the form of Eqn. (2) with $U_l(\boldsymbol{\theta}_l) = \Pi_{k=1}^{q} e^{-iH_k \boldsymbol{\theta}_{lk}}$. The last sub-step is correlating symmetric parameters to demystify the spatial symmetry of H and is accomplished by a heuristic algorithm used in determining the graph automorphism group (Stoichev, 2019). See Appendix E for more algorithmic details.

Remark. (i) We emphasize that although both SP and the pruning techniques used in deep neural networks orient to remove redundant parameters and (quantum) neurons, they are fundamentally different. This is because classical pruning methods generally leverage the magnitude of weights or the gradient information to recognize such redundancy, which is impermissible in QNNs (Wang et.al, 2022) (refer to Appendix F for elaborations). (ii) SP is a flexible framework. Besides three symmetric properties in Alg. 1, SP can effectively integrate other symmetry identification methods in the second step.

4 Experiments

In this section, we carry out numerical simulations to validate the theoretical properties of EQNTK and the effectiveness of the SP scheme in GSP. Two typical problem Hamiltonians are considered, i.e., Transverse-field Ising model and Maximum Cut with random graphs. The organization of this section is as follows. We introduce the experimental setup in Sec. 4.1

and analyze the obtained numerical results in Sec. 4.2. Refer to Appendix G for the omitted details.

4.1 Setup and hyper-parameters

Problem Hamiltonian. Let us first recap the two problem Hamiltonians.

1) Transverse-field Ising model. A central problem in quantum many-body physics is predicting the properties of these quantum systems from the first principles of quantum mechanics. Transverse-field Ising model (TFIM) has been employed to explore many interesting quantum systems. An n-qubit Hamiltonian of 1D TFIM with an open boundary condition is defined as $H_{\text{TFIM}} = -\sum_{j=1}^{n-1} \sigma_j^z \sigma_{j+1}^z - h \sum_{j=1}^{n} \sigma_j^x$, where σ_j^μ denotes the μ-Pauli matrix (with $\mu = x, z$) acting on the j-th qubit, and h is the strength of the transverse field. For simplicity, we set $h = 1$ in the following simulations. The Hamiltonian is graphically depicted in Fig. 2(a).

2) **Maximum cut.** Maximum cut (MaxCut) problem aims to partition the set of nodes V in a graph $G = (V, E)$ into two parts such that the number of edges spanning multiple parts is maximized. Although many important problems in statistical physics and operation research (Wheeler, 2004) can be formulated as MaxCut, finding the optimal solution of MaxCut has been proven to be NP-hard (Karp, 1972) and quantum computers are expected to attain better approximated solutions than those of classical computers (Farhi et al., 2014; Zhou et al., 2022). Notably, when quantum machines are applied, the MaxCut problem can be recast to GSP. Namely, the objective of an n-node graph is encoded by an n-qubit Hamiltonian $H_{\text{MC}} = \frac{1}{2} \sum_{(u,v)\in E} (I - \sigma_u^z \sigma_v^z)$ and the optimal solution corresponds to the ground state of H_{MC} as formulated in Eqn. (1). In all numerical simulations, the Erdos-Renyi graphs are adopted, which are generated by randomly connecting any pair nodes among n nodes with probability $p = 0.6$. Fig. 3(a) shows an instance of the Erdos-Renyi graph.

To verify the effectiveness of SP, the above problem Hamiltonians are modified as an $(n+m)$-qubit Hamiltonian $H = H_M \otimes \mathbb{I}^m (H_M = H_{\text{TFIM}}, H_{\text{MC}})$. The number of qubits n and m is set as 6 and 2, respectively.

Initialization of QNNs. We take the hardware efficient ansatz (HEA) with the form of Eqn. (2) as the initial ansatz, which is over-parameterized and asymmetric. HEA yields the layer-stacking structure following Eqn. (2), where each layer consists of single-qubit Pauli rotation gates and fixed two-qubit CNOT gates. The number of layers is set as $L \in \{4, 6, 8 \cdots, 28\}$ for TFIM and $L \in \{4, 6, 10, \cdots, 36\}$ for the MaxCut problem, respectively.

For each problem Hamiltonian, the input state is set as $|00\cdots0\rangle$. The parameters $\boldsymbol{\theta}$ are uniformly sampled from the uniform distribution $[-\pi, \pi]$. The variational ansatz is trained by the Adam optimizer where the learning rate is 0.001 and the rest hyper-parameters follow the default settings. The training of QNNs stops when the loss value is less than 10^{-8} or when the change in the loss function is less than 10^{-8} three times in a row. The maximum number of iteration steps is set as $T = 10000$. The ϵ value of ϵ-convergence is set as 10^{-5} for both TFIM and MaxCut. Each setting is repeated with 5 times to collect the statistical results.

Evaluation metrics. We utilize three metrics to assess the convergence rate of QNNs, i.e., (1) the loss value at convergence $\mathcal{L}(\boldsymbol{\theta}^{(T)})$; (2) the number of iteration steps $T(\epsilon) \leq T$ required to achieve ϵ-convergence; (3) the minimum number of parameterized gates required to achieve ϵ-convergence, which can also be interpreted as the threshold to achieve the over-parameterization regime. Additionally, we record the norm of the gradient at the initialization to verify the effectiveness of EQNTK for predicting convergence.

Code availability. The source code will be released to Github repository.

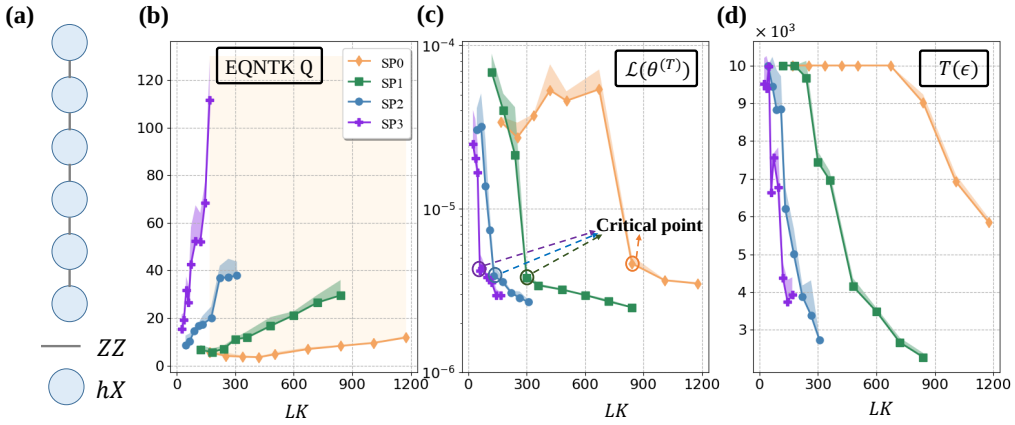

Figure 2: **Results for TFIM model under symmetric pruning.** Panel (a) depicts the graph representation of the TFIM model. The panel (b)—(d) plot the value of EQNTK at initialization Q, mean square loss at convergence $\mathcal{L}(\boldsymbol{\theta}^{(T)})$, and the number of iterations required for achieving ϵ-convergence $T(\varepsilon)$ versus the number of parameters LK, respectively. The labels 'SP0'—'SP3' respectively refer to the initial ansatz, the pruned ansatz after system symmetric pruning, structure symmetric pruning, and spatial symmetric pruning. The light orange shadow refers to the area where the trainability of QNN can be improved by increasing EQNTK.

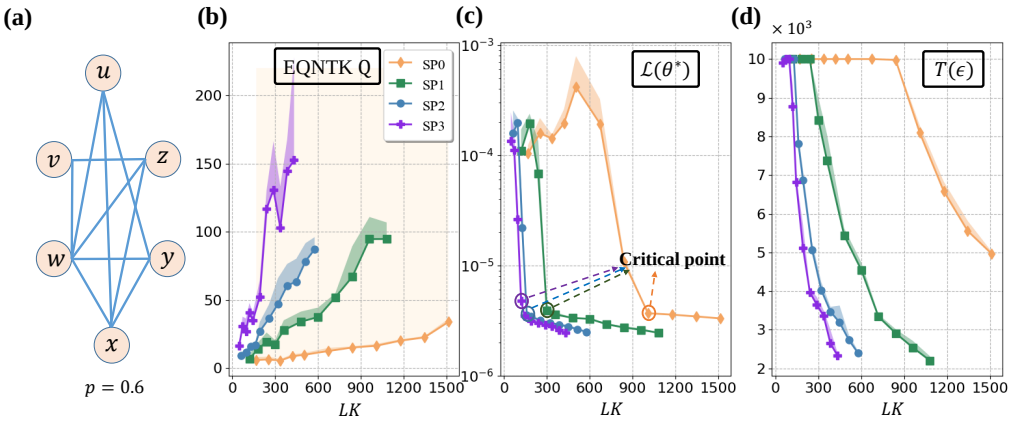

Figure 3: **Results for MaxCut problem under symmetric pruning.** Panel (a) depicts the graph representation of the MaxCut. The panel (b)—(d) plot the value of EQNTK at initialization Q, mean square loss at convergence $\mathcal{L}(\boldsymbol{\theta}^{(T)})$, and the number of iterations required for achieving ϵ-convergence $T(\varepsilon)$ versus the number of parameters LK, respectively. The labels 'SP0'—'SP3' respectively refer to the initial ansatz, the pruned ansatz after system symmetric pruning, structure symmetric pruning, and spatial symmetric pruning. The light orange shadow refers to the area where the trainability of QNN can be improved by increasing EQNTK.

4.2 Simulation results

EQNTK at initilization. Fig. 2(b) and Fig. 3(b) show that each symmetric pruning step in Alg. 1 (i.e., the sub-steps 2-1, 2-2, and 2-3 refer to 'SP1', 'SP2', 'SP3', respectively) constantly improves the squared norm of gradients (i.e., the EQNTK value Q) for both TFIM and MaxCut. This implicates that SP is capable of accelerating convergence of QNNs and alleviating barren plateaus based on Theorem 2. Notably, the EQNTK Q of the pruned ansatz (labeled by 'SP3') is improved 20 (14) times compared to the initial over-parameterized ansatz (labeled by 'SP0') in the problem of TFIM (MaxCut).

Critical point of QNNs. We see from Fig. 2(c) and Fig. 3(c) that when increasing the number of parameters beyond some threshold, the convergence of the optimization presents an overparameterization phenomenon through a computational phase transition where the loss value at convergence $\mathcal{L}(\boldsymbol{\theta}^{(T)})$ dramatically sharply drops by about an order of magnitude. Moreover, the minimum number of parameters required to reach the over-parameterization regime, highlighted by 'critical point', can be dramatically reduced by SP. Specifically, in the problem of TFIM (MaxCut), the number of parameters at the critical point of the variational ansatz corresponding to the label 'SP0' is at most and scales exponentially with the system size. It is gradually reduced from 800 (1000) to 300 (300) after SP1, then to 120 (150) after SP2, and finally to 50 (100) after SP3, which scales polynomially with the system size and is available on the NISQ device.

Convergence of QNNs. Fig. 2(d) and Fig. 3(d) show that SP algorithm can dramatically improve the convergence rate of QNNs. In the common over-parameterization regime of 'SP1'–'SP3' with the number of parameters $LK \geq 300$, the iteration step required for achieving ϵ-convergence can be reduced by up to 6000 steps for TFIM and 5000 steps for MaxCut. In addition, for the same ansatz design, increasing the number of parameters can linearly improve the convergence rate, which accords with the achieved results in Theorem 2.

EQNTK echos with the trainability of QNNs. Both Fig. 2 and Fig. 3 show the correlation between the value of EQNTK and the trainability of QNNs, where the convergence rate and the number of parameters in the critical point decrease with the increase of EQNTK. We see that in the common over-parameterization regime for 'SP1'–'SP3' (i.e., the number of parameters is more than $LK = 300$ in the MaxCut problem), the quantitative relation of the iteration steps $T(\epsilon)$ between 'SP1'–'SP3' is consistent with that of the EQNTK between 'SP1'–'SP3'. Specifically, in the task of MaxCut with $LK \approx 450$, $T_0 : T_1 : T_2 \approx Q_0 : Q_1 : Q_2 \approx 1 : 2 : 4$ with T_j and Q_j corresponding to $T(\epsilon)$ and Q in 'SPj' with $j = 1, 2, 3$. A similar quantitative correlation occurs in the task of TFIM. These observations accord with the achieved results in Theorem 2, justifying the effectiveness of EQNTK as guidance of trainability.

5 Related work

Prior literature related to our work can be cast into two categories: theories about the trainability of QNNs and the design of symmetric ansätze.

Trainability of QNNs. McClean et al. (2018) first discovered the barren plateau of QNNs. Since then, a line of research is uncovering intrinsic reasons leading to this phenomenon. Current progress has revealed that these reasons include high entanglement of QNNs (Marrero et al., 2021), the employed global measurements (Cerezo et al., 2021b), and the presence of noise (Wang et al., 2021). To mitigate barren plateaus, two popular methods are adopting shallow QNNs with local measurements (Uvarov & Biamonte, 2021; Pesah et al., 2021; Du et al., 2022a; Zhang et al., 2020) and correlating trainable parameters (Volkoff & Coles, 2021).

Another crucial line of research is investigating the convergence rate of QNNs. Several empirical studies have observed that over-parameterized QNNs promise faster convergence, and the trained parameters are near-optimal (Kiani et al., 2020; Wiersema et al., 2020; Zhang & Cui, 2020). Afterward, initial attempts have been made to theoretically explain the superiority of over-parameterized QNNs. Specifically, Larocca et al. (2021b) and Anschuetz (2021) separately leveraged the tools of dynamical Lie algebra and random matrix theory

to quantify the critical point of over-parameterized QNNs; You et al. (2022) extended the result of Xu et al. (2018) to the quantum regime and proved the exponential convergence rate of over-parameterized QNNs; Liu et al. (2022c;b;d;a) proposed the quantum neural tangent kernel to exhibit an exponential convergence rate of over-parameterized QNNs.

Ansätze with symmetric properties. Previous studies focus on unearthing inherent symmetry behind the problem Hamiltonian to design problem-specific ansätze. The mainstream approaches contain arranging the layout of ansätze (Liu et al., 2019; Seki et al., 2020; Gard et al., 2020; Zheng et al., 2021; 2022), correlating trainable parameters (Shaydulin et al., 2021; Shaydulin & Wild, 2021; Sauvage et al., 2022), and utilizing results from the geometric deep learning (Shaydulin et al., 2021; Shaydulin & Wild, 2021; Sauvage et al., 2022; Meyer et al., 2022). Compared with asymmetric ansätze, these ansätze enable better trainability and generalization ability in GSP.

Our results differ from the above literature in both theoretical and practical aspects. On the theoretical side, EQNTK greatly reconciles the harsh requirement exhibited in previous literature to reach over-parameterization. Moreover, although there is numerical evidence that symmetric ansätze can accelerate the convergence, theoretical analysis is still lacking. Our results readily explain this open question. On the practical side, EQNTK provides an efficient measure to compare the trainability of various ansätze and allows us devising two automatic methods to design symmetrical ansatz with the convergence guarantee.

6 Conclusions

In this study, we investigate the training performance of QNNs for the GSP problem by developing a novel tool—EQTNK, which is capable of capturing the training dynamics of various ansätze via their effective dimension. We theoretically exhibit that a symmetric ansatz design with a small effective dimension contributes to improving the trainability of QNNs, including alleviating the barren plateau phenomenon and reducing the number of parameters and circuit depth required to reach the over-parameterization regime. Besides, we propose a novel symmetric pruning algorithm to automatically extract the symmetric ansatz from an over-parameterized and asymmetric ansatz. We empirically demonstrate that SP has the capability of improving the trainability of QNNs. A future research direction is extending the results of EQNTK from GSP to the regime of machine learning and exploring whether over-parameterized QNNs can simultaneously attain good trainability and generalization (Abbas et al., 2021; Du et al., 2022b; Caro et al., 2022).

References

Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *Nature Computational Science*, 1(6): 403–409, 2021.

Eric R Anschuetz. Critical points in hamiltonian agnostic variational quantum algorithms. *arXiv preprint arXiv:2109.06957*, 2021.

Eric R Anschuetz and Bobak T Kiani. Beyond barren plateaus: Quantum variational algorithms are swamped with traps. *arXiv preprint arXiv:2205.05786*, 2022.

Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S Kottmann, Tim Menke, et al. Noisy intermediate-scale quantum algorithms. *Reviews of Modern Physics*, 94(1): 015004, 2022.

Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? *arXiv preprint arXiv:2003.03033*, 2020.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.

Matthias C Caro, Hsin-Yuan Huang, Marco Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J Coles. Generalization in quantum machine learning from few training data. *Nature Communications*, 13(1):1–11, 2022.

Marco Cerezo and Patrick J Coles. Higher order derivatives of quantum neural networks with barren plateaus. *Quantum Science and Technology*, 6(3):035006, 2021.

Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. Variational quantum algorithms. *Nature Reviews Physics*, pp. 1–20, 2021a. doi: https://doi.org/10.1038/s42254-021-00348-9.

Marco Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature communications*, 12(1):1–12, 2021b.

Iris Cong, Soonwon Choi, and Mikhail D Lukin. Quantum convolutional neural networks. *Nature Physics*, 15(12):1273–1278, 2019.

Paul T Darga, Mark H Liffiton, Karem A Sakallah, and Igor L Markov. Exploiting structure in symmetry detection for cnf. In *Proceedings of the 41st Annual Design Automation Conference*, pp. 530–534, 2004.

Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, Shan You, and Dacheng Tao. Learnability of quantum neural networks. *PRX Quantum*, 2(4):040337, 2021.

Yuxuan Du, Tao Huang, Shan You, Min-Hsiu Hsieh, and Dacheng Tao. Quantum circuit architecture search for variational quantum algorithms. *npj Quantum Information*, 8(1): 1–8, 2022a.

Yuxuan Du, Zhuozhuo Tu, Xiao Yuan, and Dacheng Tao. Efficient measure for the expressivity of variational quantum algorithms. *Physical Review Letters*, 128(8):080506, 2022b.

Edward Farhi and Hartmut Neven. Classification with quantum neural networks on near term processors. *arXiv preprint arXiv:1802.06002*, 2018. URL `https://arxiv.org/abs/1802.06002`.

Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm. *arXiv preprint arXiv:1411.4028*, 2014.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. Pruning neural networks at initialization: Why are we missing the mark? *arXiv preprint arXiv:2009.08576*, 2020.

Motohisa Fukuda, Robert König, and Ion Nechita. Rtni—a symbolic integrator for haar-random tensor networks. *Journal of Physics A: Mathematical and Theoretical*, 52(42): 425303, 2019.

Bryan T Gard, Linghua Zhu, George S Barron, Nicholas J Mayhall, Sophia E Economou, and Edwin Barnes. Efficient symmetry-preserving state preparation circuits for the variational quantum eigensolver algorithm. *npj Quantum Information*, 6(1):1–9, 2020.

Edward Grant, Leonard Wossnig, Mateusz Ostaszewski, and Marcello Benedetti. An initialization strategy for addressing barren plateaus in parametrized quantum circuits. *Quantum*, 3:214, 2019.

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

Babak Hassibi and David Stork. Second order derivatives for network pruning: Optimal brain surgeon. *Advances in neural information processing systems*, 5, 1992.

Y Herasymenko and TE O'Brien. A diagrammatic approach to variational quantum ansatz construction. *Quantum*, 5:596, 2021.

Tommi Junttila and Petteri Kaski. Engineering an efficient canonical labeling tool for large and sparse graphs. In *2007 Proceedings of the Ninth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pp. 135–149. SIAM, 2007.

Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M Chow, and Jay M Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, 549(7671):242–246, 2017.

Richard M Karp. Reducibility among combinatorial problems. In *Complexity of computer computations*, pp. 85–103. Springer, 1972.

Bobak Toussi Kiani, Seth Lloyd, and Reevu Maity. Learning unitaries by gradient descent. *arXiv preprint arXiv:2001.11897*, 2020.

Martin Larocca, Piotr Czarnik, Kunal Sharma, Gopikrishnan Muraleedharan, Patrick J Coles, and M Cerezo. Diagnosing barren plateaus with tools from quantum optimal control. *arXiv preprint arXiv:2105.14377*, 2021a.

Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J Coles, and M Cerezo. Theory of overparametrization in quantum neural networks. *arXiv preprint arXiv:2109.11676*, 2021b.

Martin Larocca, Frederic Sauvage, Faris M Sbahi, Guillaume Verdon, Patrick J Coles, and Marco Cerezo. Group-invariant quantum machine learning. *arXiv preprint arXiv:2205.02261*, 2022.

Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. *Advances in neural information processing systems*, 2, 1989.

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip HS Torr. Snip: Single-shot network pruning based on connection sensitivity. *arXiv preprint arXiv:1810.02340*, 2018.

Jin-Guo Liu, Yi-Hong Zhang, Yuan Wan, and Lei Wang. Variational quantum eigensolver with fewer qubits. *Physical Review Research*, 1(2):023025, 2019.

Junyu Liu, Zexi Lin, and Liang Jiang. Laziness, barren plateau, and noise in machine learning. *arXiv preprint arXiv:2206.09313*, 2022a.

Junyu Liu, Khadijeh Najafi, Kunal Sharma, Francesco Tacchino, Liang Jiang, and Antonio Mezzacapo. An analytic theory for the dynamics of wide quantum neural networks. *arXiv preprint arXiv:2203.16711*, 2022b.

Junyu Liu, Francesco Tacchino, Jennifer R. Glick, Liang Jiang, and Antonio Mezzacapo. Representation Learning via Quantum Neural Tangent Kernels. *PRX Quantum*, 3(3): 030323, 2022c. doi: 10.1103/PRXQuantum.3.030323.

Junyu Liu, Changchun Zhong, Matthew Otten, Cristian L. Cortes, Chaoyang Ti, Stephen K. Gray, and Xu Han. Quantum Kerr Learning. 5 2022d.

José Luis López-Presa, Luis F Chiroque, and Antonio Fernández Anta. Novel techniques to speed up the computation of the automorphism group of a graph. *Journal of Applied Mathematics*, 2014, 2014.

Carlos Ortiz Marrero, Mária Kieferová, and Nathan Wiebe. Entanglement-induced barren plateaus. *PRX Quantum*, 2(4):040316, 2021.

Iman Marvian. Restrictions on realizable unitary operations imposed by symmetry and locality. *Nature Physics*, 18(3):283–289, 2022.

Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature communications*, 9(1):1–6, 2018.

Brendan D McKay and Adolfo Piperno. Practical graph isomorphism, ii. *Journal of symbolic computation*, 60:94–112, 2014.

Brendan D McKay et al. Practical graph isomorphism. 1981.

Péter Mernyei, Konstantinos Meichanetzidis, and Ismail Ilkan Ceylan. Equivariant quantum graph circuits. In *International Conference on Machine Learning*, pp. 15401–15420. PMLR, 2022.

Johannes Jakob Meyer, Marian Mularski, Elies Gil-Fuster, Antonio Anna Mele, Francesco Arzani, Alissa Wilms, and Jens Eisert. Exploiting symmetry in variational quantum machine learning. *arXiv preprint arXiv:2205.06217*, 2022.

Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J Love, Alán Aspuru-Guzik, and Jeremy L O'brien. A variational eigenvalue solver on a photonic quantum processor. *Nature communications*, 5:4213, 2014.

Arthur Pesah, M Cerezo, Samson Wang, Tyler Volkoff, Andrew T Sornborger, and Patrick J Coles. Absence of barren plateaus in quantum convolutional neural networks. *Physical Review X*, 11(4):041011, 2021.

David Poulin and Pawel Wocjan. Preparing ground states of quantum many-body systems on a quantum computer. *Physical review letters*, 102(13):130503, 2009.

John Preskill. Quantum computing in the nisq era and beyond. *Quantum*, 2:79, 2018. doi: https://doi.org/10.22331/q-2018-08-06-79.

Yang Qian, Xinbiao Wang, Yuxuan Du, Xingyao Wu, and Dacheng Tao. The dilemma of quantum neural networks. *arXiv preprint arXiv:2106.04975*, 2021.

Frederic Sauvage, Martin Larocca, Patrick J Coles, and M Cerezo. Building spatial symmetries into parameterized quantum circuits for faster training. *arXiv preprint arXiv:2207.14413*, 2022.

Kazuhiro Seki, Tomonori Shirakawa, and Seiji Yunoki. Symmetry-adapted variational quantum eigensolver. *Physical Review A*, 101(5):052340, 2020.

Ruslan Shaydulin and Stefan M Wild. Exploiting symmetry reduces the cost of training qaoa. *IEEE Transactions on Quantum Engineering*, 2:1–9, 2021.

Ruslan Shaydulin, Stuart Hadfield, Tad Hogg, and Ilya Safro. Classical symmetries and the quantum approximate optimization algorithm. *Quantum Information Processing*, 20(11): 1–28, 2021.

Andrea Skolik, Jarrod R McClean, Masoud Mohseni, Patrick van der Smagt, and Martin Leib. Layerwise learning for quantum neural networks. *Quantum Machine Intelligence*, 3(1):1–11, 2021.

Stoicho D Stoichev. New exact and heuristic algorithms for graph automorphism group and graph isomorphism. *Journal of Experimental Algorithmics (JEA)*, 24:1–27, 2019.

AV Uvarov and Jacob D Biamonte. On barren plateaus and cost function locality in variational quantum algorithms. *Journal of Physics A: Mathematical and Theoretical*, 54(24): 245301, 2021.

Tyler Volkoff and Patrick J Coles. Large gradients via correlation in random parameterized quantum circuits. *Quantum Science and Technology*, 6(2):025008, 2021.

Chaoqi Wang, Guodong Zhang, and Roger Grosse. Picking winning tickets before training by preserving gradient flow. *arXiv preprint arXiv:2002.07376*, 2020.

Hanrui Wang, Yongshan Ding, Jiaqi Gu, Yujun Lin, David Z Pan, Frederic T Chong, and Song Han. Quantumnas: Noise-adaptive search for robust quantum circuits. In *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pp. 692–708. IEEE, 2022.

Samson Wang, Enrico Fontana, Marco Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J Coles. Noise-induced barren plateaus in variational quantum algorithms. *Nature communications*, 12(1):1–11, 2021.

Xinbiao Wang et.al. In preparation. 2022.

Dave Wecker, Matthew B Hastings, and Matthias Troyer. Progress towards practical quantum variational algorithms. *Physical Review A*, 92(4):042303, 2015.

John W Wheeler. An investigation of the max-cut problem. *University of Iowa*, 2004.

Roeland Wiersema, Cunlu Zhou, Yvette de Sereville, Juan Felipe Carrasquilla, Yong Baek Kim, and Henry Yuen. Exploring entanglement and optimization within the hamiltonian variational ansatz. *PRX Quantum*, 1(2):020319, 2020.

Zhiqiang Xu, Xin Cao, and Xin Gao. Convergence analysis of gradient descent for eigenvector computation. International Joint Conferences on Artificial Intelligence, 2018.

Xuchen You, Shouvanik Chakrabarti, and Xiaodi Wu. A convergence theory for overparameterized variational quantum eigensolvers. *arXiv preprint arXiv:2205.12481*, 2022.

Kaining Zhang, Min-Hsiu Hsieh, Liu Liu, and Dacheng Tao. Toward trainability of quantum neural networks. *arXiv preprint arXiv:2011.06258*, 2020.

Stephen Zhang and Wentao Cui. Overparametrization in qaoa. *Written Report*, 2020.

Han Zheng, Zimu Li, Junyu Liu, Sergii Strelchuk, and Risi Kondor. Speeding up learning quantum states through group equivariant convolutional quantum ans {\" a} tze. *arXiv preprint arXiv:2112.07611*, 2021.

Han Zheng, Zimu Li, Junyu Liu, Sergii Strelchuk, and Risi Kondor. On the Super-exponential Quantum Speedup of Equivariant Quantum Machine Learning Algorithms with SU(d) Symmetry. 7 2022.

Zeqiao Zhou, Yuxuan Du, Xinmei Tian, and Dacheng Tao. Qaoa-in-qaoa: solving large-scale maxcut problems on small quantum machines. *arXiv preprint arXiv:2205.11762*, 2022.

A Equivalent training dynamics under the mean square error loss

In this section, we exhibit the training dynamics of the parameterized quantum state $|\psi(\boldsymbol{\theta})\rangle$ under a general mean square error loss function

$$\mathcal{L}(\boldsymbol{\theta}, C) = \frac{1}{2}\left(\langle\psi_0| U(\boldsymbol{\theta})^\dagger H U(\boldsymbol{\theta}) |\psi_0\rangle - C\right)^2 \equiv \frac{1}{2}\varepsilon(\boldsymbol{\theta}, C)^2, \tag{7}$$

where $C \leq E_0$ is a constant and

$$\varepsilon(\boldsymbol{\theta}, C) = \langle\psi_0| U(\boldsymbol{\theta})^\dagger H U(\boldsymbol{\theta}) |\psi_0\rangle - C$$

is the training error associated with C.

The following lemma exhibits that given any C and C', the corresponding training dynamic of $|\psi(\boldsymbol{\theta})\rangle$ with respect to $\mathcal{L}(\boldsymbol{\theta}, C)$ and $\mathcal{L}(\boldsymbol{\theta}, C')$ is equivalent.

Lemma 2. *Under the framework of quantum neural tangent kernel, given any $C \leq E_0$, the loss function in Eqn. (7) with the varied constant $C \leq E_0$ has the same convergence rate and yields the same variational quantum state $|\psi(\boldsymbol{\theta}^*)\rangle$ minimized the loss function.*

Proof. Following the results of Theorem 2, the training error $\varepsilon(\boldsymbol{\theta}, C)$ decays as follows:

$$\varepsilon_t(\boldsymbol{\theta}, C) + (C - E^*) \approx (1 - \eta Q_c)^t \left(\varepsilon_0(\boldsymbol{\theta}, C) + (C - E^*)\right) \tag{8}$$

where Q_c is the quantum neural tangent kernel under the loss function in Eqn. (7), $E^* - C := \min_{\boldsymbol{\theta}} \varepsilon(\boldsymbol{\theta}, C)$, and $E^* = \min_{\boldsymbol{\theta}} \langle\psi_0| U(\boldsymbol{\theta})^\dagger H U(\boldsymbol{\theta}) |\psi_0\rangle$. Note that the convergence rate of QNN is determined by $Q_c = \nabla\varepsilon_c^\top \nabla\varepsilon_c$. This quantity is independent with the constant C because $\nabla\varepsilon_c^\top$ is independent with C. Meanwhile, following Eqn. (8), the output of QNN exponentially converges to the minimum value E^*. When $E^* = E_0$, the optimized quantum state $|\psi(\boldsymbol{\theta^*})\rangle$ is the ground state of H for any constant C. □

B Proof of Theorem 2

The proof of Theorem 2 employs the following two lemmas whose proofs are given in the subsequent two subsections.

Lemma 3 (You et al. (2022), Lemma D.1)**.** *Let $\mathcal{U}_\mathcal{A}$ be a matrix subgroup of $SU(d)$ where each element in $\mathcal{U}_\mathcal{A}$ commutes with a Hermitian matrix Σ. The corresponding direct decomposition is denoted as $V = \sum_{j=1}^p V_j$ with projection Π_j. Let V^* be the subspace of interest which includes the input state $|\psi_0\rangle$ and ground state $|\psi^*\rangle$ and $\Pi^* = P^\dagger P$ the corresponding projection. Then for any Hermitian W and unitary matrix U in the group $\mathcal{U}_\mathcal{A}$:*

$$\Pi^* U W U^\dagger \Pi^* = \Pi^* U \Pi \; \Pi^* W \Pi^* \; \Pi^* U^\dagger \Pi^*. \tag{9}$$

Lemma 4. *The quantum neural tangent kernel has the following form*

$$Q = -\sum_{\ell=1}^{L}\sum_{k=1}^{K} \left\langle \psi_0 \middle| U_{+,\ell k}^\dagger \left[G_k, U_{-,\ell k}^\dagger H U_{-,\ell k}\right] U_{+,\ell k} \middle| \psi_0 \right\rangle^2, \tag{10}$$

where

$$U_{-,\ell k} \equiv \prod_{\ell'=1}^{\ell-1} U_{\ell'}(\boldsymbol{\theta}_{\ell'}) \prod_{k'=1}^{k-1} e^{-i\boldsymbol{\theta}_{\ell k'} G_{k'}}, \quad U_{+,\ell k} \equiv \prod_{k'=k}^{K} e^{-i\boldsymbol{\theta}_{\ell k'} G_{k'}} \prod_{\ell'=\ell+1}^{L} U_{\ell'}(\boldsymbol{\theta}_{\ell'}). \tag{11}$$

When the conditions in Lemma 3 hold, EQNTK has the form of

$$Q_S = -\sum_{\ell=1}^{L}\sum_{k=1}^{K} \left\langle \psi_0^* \middle| (U_{+,\ell k}^*)^\dagger \left[G_k^*, (U_{-,\ell k}^*)^\dagger H^* U_{-,\ell k}^*\right] U_{+,\ell k}^* \middle| \psi_0^* \right\rangle^2, \tag{12}$$

where $|\psi_0^*\rangle = P|\psi^*\rangle$ and $A^* = PAP^\dagger$ (with $A = U_{+,\ell k}, G_k, U_{-,\ell k}, H$).

Proof of Theorem 2. Following the gradient descent optimizer with learning rate $\eta \leq 1$, the change in the training error of QNN can be expressed as

$$đ\varepsilon = \sum_{\ell,k} \frac{\partial\varepsilon}{\partial\boldsymbol{\theta}_{\ell k}} đ\boldsymbol{\theta}_{\ell k} = -\eta \sum_{\ell,k} \frac{\partial\varepsilon}{\partial\boldsymbol{\theta}_{\ell k}} \frac{\partial\varepsilon}{\partial\boldsymbol{\theta}_{\ell k}} \varepsilon = -\eta Q_S \varepsilon. \tag{13}$$

where $đ\mu = \mu(t+1) - \mu(t)$ (i.e., μ can either be ε or $\boldsymbol{\theta}_{\ell k}$), the second equality comes from the update rule of gradient descent $đ\boldsymbol{\theta}_{\ell k} = \eta\varepsilon\partial\varepsilon/\partial\boldsymbol{\theta}_{\ell k}$, and the third equality uses the representation of QNTK in Eqn. (3).

Following from the results in Liu et al. (2022c, Theorem 1), when the EQNTK is a constant, the training error decays with

$$\varepsilon_t \approx (1 - \eta Q_S)^t \varepsilon_0 \approx e^{-\eta Q_S t} \varepsilon_0, \tag{14}$$

which guarantees an exponential convergence towards the global optima. To this end, the proof of Theorem 2 amounts to proving that when the number of parameters satisfies $|\boldsymbol{\theta}| =$

$LK \gg 1$, the EQNTK can be regarded as a constant, which can be achieved by deriving an analytical solution of Q_S on average as well as the fluctuations around the average.

We next analyze the average of Q_S. Following Lemma 3 and the assumption that $U_{-,\ell k}$ and $U_{+,\ell k}$ are independent and match the Haar distribution up to the second moment, the Haar average of EQNTK yields

$$
\begin{aligned}
Q_S &= -\sum_{\ell=1}^{L}\sum_{k=1}^{K}\int dU_{+,\ell k}dU_{-,\ell k}\left\langle\psi_0^*\middle|(U_{+,\ell k}^*)^\dagger\left[G_k^*,(U_{-,\ell k}^*)^\dagger H^* U_{-,\ell k}^*\right]U_{+,\ell k}^*\middle|\psi_0^*\right\rangle^2, \\
&= -\sum_{\ell=1}^{L}\sum_{k=1}^{K}\int dU_{+,\ell k}^* dU_{-,\ell k}^* \operatorname{Tr}\left(\rho_0^*(U_{+,\ell k}^*)^\dagger M_{-,\ell k}U_{+,\ell k}^*\rho_0^*(U_{+,\ell k}^*)^\dagger M_{-,\ell k}U_{+,\ell k}^*\right), \\
&= -\sum_{\ell=1}^{L}\sum_{k=1}^{K}\int dU_{+,\ell k}^*\left(\frac{\operatorname{Tr}^2\left(M_{-,\ell k}\right)\operatorname{Tr}\left((\rho_0^*)^2\right)}{d_{\text{eff}}^2-1}+\frac{\operatorname{Tr}\left((M_{-,\ell k}^*)^2\right)\operatorname{Tr}\left(\rho_0^2\right)}{d_{\text{eff}}^2-1}\right. \\
&\left.+\frac{\operatorname{Tr}^2\left(M_{-,\ell k}\right)\operatorname{Tr}^2\left(\rho_0\right)}{d_{\text{eff}}-d_{\text{eff}}^3}+\frac{\operatorname{Tr}\left((M_{-,\ell k}^*)^2\right)\operatorname{Tr}\left((\rho_0^*)^2\right)}{d_{\text{eff}}-d_{\text{eff}}^3}\right) \\
&= -\sum_{\ell=1}^{L}\sum_{k=1}^{K}\int dU_{+,\ell k}^*\frac{\operatorname{Tr}\left((M_{-,\ell k}^*)^2\right)}{d_{\text{eff}}^2+d_{\text{eff}}} \\
&= -\sum_{\ell=1}^{L}\sum_{k=1}^{K}\int dU_{+,\ell k}^*\frac{2\operatorname{Tr}\left(\left(G_k^*(U_{-,\ell k}^*)^\dagger H^* U_{-,\ell k}^*\right)^2\right)-2\operatorname{Tr}\left((G_k^*)^2((U_{-,\ell k}^*)^\dagger H^* U_{-,\ell k}^*)^2\right)}{d_{\text{eff}}^2+d_{\text{eff}}} \\
&= \frac{2}{d_{\text{eff}}^2+d_{\text{eff}}}\left(\frac{d_{\text{eff}}\operatorname{Tr}((H^*)^2-\operatorname{Tr}^2(H^*))}{d_{\text{eff}}^2-1}\right)\operatorname{Tr}\left(L\sum_{k=1}^{K}(G_k^*)^2\right) \\
&\approx \frac{LK\operatorname{Tr}\left((H^*)^2\right)}{d_{\text{eff}}^2}.
\end{aligned}
\tag{15}
$$

where the second equality follows from the fact that $U_{-,\ell k}^*$ and $U_{+,\ell k}^*$ match the Haar distibution on the group $SU(d_{\text{eff}})$ to the second moment, $\rho_0^* = |\psi_0^*\rangle\langle\psi_0^*|$, and $M_{-,\ell k} = [G_k^*,(U_{-,\ell k}^*)^\dagger H^* U_{-,\ell k}^*]$, the calculation of integration with respect to the Haar measure in the third and the sixth equality exploits the **RTNI** package (Fukuda et al., 2019), the fifth equality employs the equation $\operatorname{Tr}([A,B]^2) = 2\operatorname{Tr}(ABAB) - 2\operatorname{Tr}(A^2B^2)$, the last equality utilizes the fact

$$
\operatorname{Tr}((G_k^*)^2) = \frac{\operatorname{Tr}(G_k\Pi^* G_k\Pi^*)}{\sum_{j=1}^{p}\operatorname{Tr}(G_k\Pi_j G_k\Pi_j)}\operatorname{Tr}(G_k^2) \approx \frac{d_{\text{eff}}}{2^n}\cdot 2^n = d_{\text{eff}}. \tag{16}
$$

where $\operatorname{Tr}(G_k^2) = \operatorname{Tr}(\mathbb{I}) = 2^n$.

Last, the fluctuations of EQNTK can be expressed as the form of $\Delta Q_S^2 = E(Q_S^2) - \bar{Q}_S^2$:

$$\begin{aligned}
\Delta Q_S^2 =& 2 \sum_{\ell_1,k_1<\ell_2,k_2} \int dU^*_{+,\ell_1k_1} dU^*_{+,\ell_2k_2} dU^*_{-,\ell_1k_1} dU^*_{-,\ell_2k_2} \times \\
& \begin{pmatrix} \mathrm{Tr}\left(\rho_0^*(U^*_{+,\ell_1k_1})^\dagger M_{-,\ell_1k_1}(U^*_{+,\ell_1k_1})^\dagger \rho_0^* U^*_{+,\ell_1k_1} M_{-,\ell_1k_1} U^*_{+,\ell_1k_1}\right) \\ \mathrm{Tr}\left(\rho_0^*(U^*_{+,\ell_2k_2})^\dagger M_{-,\ell_2k_2}(U^*_{+,\ell_2k_2})^\dagger \rho_0^* U^*_{+,\ell_2k_2} M_{-,\ell_2k_2} U^*_{+,\ell_2k_2}\right) \end{pmatrix} \\
& + \sum_{\ell,k} \int dU^*_{+,\ell k} dU^*_{-,\ell k} \begin{pmatrix} \mathrm{Tr}\left(\rho_0^*(U^*_{+,\ell k})^\dagger M_{-,\ell k}(U^*_{+,\ell k})^\dagger \rho_0^* U^*_{+,\ell k} M_{-,\ell k} U^*_{+,\ell k}\right) \\ \mathrm{Tr}\left(\rho_0^*(U^*_{+,\ell k})^\dagger M_{-,\ell k}(U^*_{+,\ell k})^\dagger \rho_0^* U^*_{+,\ell k} M_{-,\ell k} U^*_{+,\ell k}\right) \end{pmatrix} \\
& - \bar{Q}_S^2 \\
=& \frac{LK}{d_{\mathrm{eff}}^4} \left(8\,\mathrm{Tr}^2((H^*)^2) + 12\,\mathrm{Tr}((H^*)^4)\right) + \\
& \frac{LK}{d_{\mathrm{eff}}^5} \Big(16\,\mathrm{Tr}((H^*)^2)\,\mathrm{Tr}^2(H^*) + 48\,\mathrm{Tr}((H^*)^3)\,\mathrm{Tr}(H^*) + 40\,\mathrm{Tr}^2((H^*)^2) + \cdots \\
\approx& \frac{LK}{d_{\mathrm{eff}}^4} \left(8\,\mathrm{Tr}^2((H^*)^2) + 12\,\mathrm{Tr}((H^*)^4)\right), \qquad (17)
\end{aligned}$$

where $\ell_1, k_1 < \ell_2, k_2$ refers to $\ell_1 K + k_1 < \ell_2 K + k_2$, the last equality adopts the leading order term as an approximation, the derivation of the second equality, which is omitted here, is mainly concerned with calculating the Haar integral using **RTNI** package following the results of Liu et al. (2022b, Appendix C).

Taken together, when the number of parameters $LK \gg 1$ such that $Q_S/\Delta Q_S \approx \frac{1}{\sqrt{LK}} \ll 1$, the EQNTK can be viewed as a constant and Eqn. (14) is satisfied.

□

B.1 Proof of Lemma 3

Proof of Lemma 3. The decomposition of invariant subspaces dictates that for any $U \in \mathcal{U}_\mathcal{A}$ is block-diagonal under $\{\Pi_j\}_{j=1}^p$, i.e., $\forall U \in U \in \mathcal{U}_\mathcal{A}, \forall j \neq j', \Pi_{j'} \mathbf{U} \Pi_j = 0$.

$$\begin{aligned}
& \Pi^* U W U^\dagger \Pi^* \\
=& \Pi^* U \sum_{j=1}^p \Pi_j W \sum_{j'=1}^p \Pi_{j'} U^\dagger \Pi^* \\
=& \sum_{j,j' \in [p]} (\Pi^* U \Pi_j)\, W \left(\Pi_{j'} U^\dagger \Pi^*\right) \\
=& \Pi^* U \Pi^* W \Pi^*\ U^\dagger \Pi^* \\
=& \Pi^* U \Pi^* \Pi^* W \Pi^*\ \Pi^* U^\dagger \Pi^*,
\end{aligned}$$

where the first equality employs the fact of $\mathbb{I} = \sum_{j=1}^p \Pi_j$, and the last equality uses the property of projections $\Pi_j^2 = \Pi_j$. □

B.2 Proof of Lemma 4

Proof of Lemma 4. Utilizing the form of QNTK with respect to the gradient vector of training error $\varepsilon(\boldsymbol{\theta}) = \langle \psi_0 | U(\boldsymbol{\theta})^\dagger H U(\boldsymbol{\theta}) | \psi_0 \rangle - E_0$ directly yields

$$\begin{aligned}
Q &= (\nabla \varepsilon(\boldsymbol{\theta}))^T \nabla \varepsilon(\boldsymbol{\theta}) \\
&= -\sum_{\ell=1}^L \sum_{k=1}^K \left\langle \psi_0 \middle| U^\dagger_{+,\ell k} \left[G_k, U^\dagger_{-,\ell k} H U_{-,\ell k}\right] U_{+,\ell k} \middle| \psi_0 \right\rangle^2. \qquad (18)
\end{aligned}$$

Moreover, for the symmetric ansatz $U(\boldsymbol{\theta})$ with projection $\Pi^* = PP^\dagger$, the EQNTK yields

$$\begin{aligned}
Q_S &= -\sum_{\ell=1}^{L}\sum_{k=1}^{K} \mathrm{Tr}\left(\left|\psi_0\right\rangle\left\langle\psi_0\right| U_{+,\ell k}^\dagger \left[G_k, U_{-,\ell k}^\dagger H U_{-,\ell k}\right] U_{+,\ell k}\right)^2 \\
&= -\sum_{\ell=1}^{L}\sum_{k=1}^{K} \mathrm{Tr}\left(\Pi^*\rho_0\Pi^* U_{+,\ell k}^\dagger \left[G_k, U_{-,\ell k}^\dagger H U_{-,\ell k}\right] U_{+,\ell k}\right)^2 \\
&= -\sum_{\ell=1}^{L}\sum_{k=1}^{K} \mathrm{Tr}\left(\Pi^*\rho_0\Pi^* U_{+,\ell k}^\dagger \left[G_k, U_{-,\ell k}^\dagger O U_{-,\ell k}\right] U_{+,\ell k}\Pi^*\right)^2 \\
&= -\sum_{\ell=1}^{L}\sum_{k=1}^{K} \mathrm{Tr}\left(\Pi^*\rho_0\Pi^* U_{+,\ell k}^\dagger \Pi^* \left[G_k, U_{-,\ell k}^\dagger O U_{-,\ell k}\right] \Pi^* U_{+,\ell k}\Pi^*\right)^2 \\
&= -\sum_{\ell=1}^{L}\sum_{k=1}^{K} \mathrm{Tr}\left(\Pi^*\rho_0\Pi^* U_{+,\ell k}^\dagger \Pi^* \left[\Pi^* G_k \Pi^*, \Pi^* U_{-,\ell k}^\dagger \Pi^* H \Pi^* U_{-,\ell k}\Pi^*\right] \Pi^* U_{+,\ell k}\Pi^*\right)^2 \\
&= -\sum_{\ell=1}^{L}\sum_{k=1}^{K} \mathrm{Tr}\left(P^\dagger\rho_0 P P^\dagger U_{+,\ell k}^\dagger P \left[P^\dagger G_k P, P^\dagger U_{-,\ell k}^\dagger P P^\dagger H P P^\dagger U_{-,\ell k} P\right] P^\dagger U_{+,\ell k} P\right)^2 \\
&= -\sum_{\ell=1}^{L}\sum_{k=1}^{K} \mathrm{Tr}\left(\rho_0^* (U_{+,\ell k}^*)^\dagger \left[G_k^*, (U_{-,\ell k}^*)^\dagger H^* U_{-,\ell k}^*\right] U_{+,\ell k}^*\right)^2 \\
&= -\sum_{\ell=1}^{L}\sum_{k=1}^{K} \left\langle\psi_0^*\right| (U_{+,\ell k}^*)^\dagger \left[G_k^*, (U_{-,\ell k}^*)^\dagger H^* U_{-,\ell k}^*\right] U_{+,\ell k}^* \left|\psi_0^*\right\rangle^2,
\end{aligned} \tag{19}$$

where $\rho_0 = \left|\psi_0\right\rangle\left\langle\psi_0\right|$, the second equality utilizes the fact that $\left|\psi_0\right\rangle$ lies in V^*, the third to the fifth equality employ the property of projection operator $(\Pi^*)^2 = \Pi^*$ and Lemma 3, the final equality follows from the definitions with $\left|\psi_0^*\right\rangle = P^\dagger\left|\psi_0\right\rangle$ and $A^* = P^\dagger A P$ (with $A = U_{-,\ell k}, U_{+,\ell k}, G_k, H, \rho_0$) . □

C Proof of Lemma 1

Before moving on to elaborate on the proof of Lemma 1, we first briefly review the definition of dynamical Lie algebra (DLA).

Definition 3 (Definition 3, Larocca et al. (2021a))**.** *Given an ansatz design $\mathcal{A}$, the dynamical Lie algebra $\mathfrak{g}$ is generated by repeated nested commutators of the operators in $\mathcal{A}$. That is*

$$\mathfrak{g} = \mathrm{span}\left\langle iG_1, \cdots, iG_K\right\rangle_{Lie} \tag{20}$$

where $\mathrm{span}\left\langle S\right\rangle_{Lie}$ *denotes the Lie closure, i.e., the set obtained by repeatedly taking the commutator of the elements in S.*

The proof of Lemma 1 employs the following Lemma.

Lemma 5. *Consider the QNN instance $\left(\left|\psi_0\right\rangle, U(\boldsymbol{\theta}), H\right)$ with the corresponding dynamical Lie algebra denoted as $\mathfrak{g}$. If there exists an invariant subspace $V_{\mathfrak{g}}$ with dimension $d_{\mathfrak{g}}$ under the DLA $\mathfrak{g}$ including the input state $\left|\psi_0\right\rangle$ and the ground state $\left|\psi^*\right\rangle$, then the effective dimension of the ansatz design $\mathcal{A}$ yields $d = d_{\mathfrak{g}}$.*

Proof of Lemma 5. From the Baker-Campbell-Hausdorff formula, the unitary group $\mathcal{U}_{\mathcal{A}}$ induced by the ansatz design $\mathcal{A}$ can be represented as

$$\mathcal{U}_{\mathcal{A}} = \cup_{L=0}^{\infty}\{U(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^{LK}\} = \{e^V, V \in \mathfrak{g}\}, \tag{21}$$

where the detailed derivation of the second equality, which is omitted here, can be found in section IV of Larocca et al. (2021b).

For any element $V \in \mathfrak{g}$, if there exists a Hermitian matrix Σ such that V commutes with Σ, then $e^V \in \mathcal{U}_\mathcal{A}$ commutes with Σ. This implies that the induced invariant subspace by $\mathcal{U}_\mathcal{A}$ is the same with that induced by $\mathfrak{g}$ and thus $d = d_\mathfrak{g}$.

□

Proof of Lemma 1. The Lemma 1 can be derived directly from Theorem 2 and Lemma 5. □

D Various symmetry in symmetric pruning algorithm

In this section, we elucidate the system symmetry, structure symmetry, and spatial symmetry discussed in the symmetric pruning algorithm. Recall that a symmetry S of a Hamiltonian H is a unitary operator leaving H invariant, such that $SHS^\dagger = H$. All the symmetries form a symmetry group $\mathcal{S}$ where for any two symmetries $S_1, S_2 \in \mathcal{S}$, their compositions $S_1 \circ S_2$ or $S_2 \circ S_1$ and their inverses S_1^{-1} and S_2^{-1} are also symmetries and in the symmetry group $\mathcal{S}$. We classify these symmetries into three categories, namely, system symmetry, structure symmetry, and spatial symmetry.

System symmetry. System symmetry consider the symmetry on qubit wires. Specifically, given a problem Hamiltonian with the form of $H \otimes \mathbb{I}^{\otimes m}$ acting on $n+m$ qubits, the system symmetry is $S_{sys} = \mathbb{I}^{\otimes n} \otimes U$ where $U \in SU(2^m)$ is an arbitrary unitary and thus S_{sys} satisfies $S_{sys}(H \otimes \mathbb{I}^{\otimes m})S_{sys}^\dagger = H \otimes \mathbb{I}^{\otimes m}$. Such symmetry motivates us to prune the redundant parameterized gates on the last m qubit wires through choosing an appropriate unitary U in the symmetry S_{sys}.

Structure symmetry. Based on the system symmetry

Spatial symmetry. Spatial symmetry is a discrete symmetry considering the permutation invariance on qubit wires. Specifically, a spatial symmetry of a problem Hamiltonian H is a permutation π over the sites leaving H invariant, i.e., $\pi H \pi^{-1} = H$. Accordingly, the sub-step 2-3 in Alg. 1 requires the varitional unitary $U(\boldsymbol{\theta})$ satisfying $\pi U(\boldsymbol{\theta})\pi^{-1} = U(\boldsymbol{\theta})$. A detailed eloboration is given in the next section.

E Automatically identifying the spatial symmetry

The sub-step 2-3 in Alg. 1 leverages the algorithms developed in graph theory to automatically identify the spatial symmetry of problem Hamiltonians. Here we elucidate the implementations of this sub-step.

From the graphical view, an n-qubit Hamiltonian H refers to a graph $G = (V, E)$ with n vertices, where the j-th nodes $v_j \in V$ represents the j-th quantum particle (site) of H and the edge $E_{i,j} \in E$ characterizes the interaction strength of the i-th and the j-th quantum particles (sites). This graph can further be described by an adjacency matrix D.

Recall that a spatial symmetry π of a Hamiltonian H is a permutation over the sites leaving H invariant, i.e., $\pi H \pi^{-1} = H$ (or equivalently $[\pi, H] = 0$). In other words, the spatial symmetry π preserves the topology invariance of G such that for any $(u, v) \in E$, we have

$$(\pi(v), \pi(u)) \in E, \text{ and } \pi D \pi^{-1} = D.$$

In GSP, the action of π on an n-qubit state $|\psi\rangle \to \pi |\psi\rangle$ means permuting the indices of qubits. For instance, a permutation π with $\pi(1) = 3, \pi(2) = 1, \pi(3) = 2$ acting on the state $|\psi_1\rangle |\psi_2\rangle |\psi_3\rangle$ yields $\pi(|\psi_1\rangle |\psi_2\rangle |\psi_3\rangle) = |\psi_3\rangle |\psi_1\rangle |\psi_2\rangle$. All these permutations form a discrete group of symmetries S_n with the cardinality $O(n!)$. Particularly, the spatial symmetries of the Hamiltonian is the automorphism group of its corresponding graph, defined as

$$Aut(H) = \{\pi_a \in S_n | \pi_a H \pi_a^{-1} = H\},$$

or equivalently $Aut(H) = \{\pi_a \in S_n | \pi_a D \pi_a^{-1} = D\}$. The qubits (qubit-pair) in the ansatz corresponding to the node (edge) which can be swapped refer to the equivalent qubits (qubit-pair). More precisely, for any node (edge) $u \in V$ ($(u, v) \in E$), if there exists $\pi \in Aut(H)$ such that $\pi(u) = x$ ($\pi(u, v) = (x, y)$), then the qubits (qubit-pair) corresponding to the

node (edge) u ((u, v)) and x ((x, y)) are called equivalent qubits (qubit-pairs). In sub-step 2-3, spatial symmetric pruning correlates the single-qubit parameterized gates on the equivalent qubits or the two-qubit parameterized gates on the equivalent qubit-pairs with one parameter per layer.

In practice, the automorphism group of graph corresponding to the Hamiltonian with complicated topological structure is hard to compute manually. There are a lot of heuristic algorithm to compute the automorphism group, including *nauty* (McKay et al., 1981), *Traces* (McKay & Piperno, 2014), *saucy* (Darga et al., 2004), *Bliss* (Junttila & Kaski, 2007) and *canauto* (López-Presa et al., 2014). The corresponding software packages of these algorithms are capable of solving most graphs for up to tens of thousands of nodes in less than a second (McKay & Piperno, 2014). In this work, we employ *nauty* to automatically recognize the automorphism group of graph corresponding to the Hamiltonian.

F The limitations of classical pruning

Although both symmetric pruning and classical pruning distill a smaller network (or ansatz) from an over-parameterized one in the view of algorithmic implementation, classical pruning method can not be directly employed to enhance the power of QNNs.

Recall that a common strategy of classical pruning methods is scoring each parameter or network element, and then removing those accompanied with low scores. Notably, such scores correspond to the magnitude of parameters (Frankle & Carbin, 2018), the gradient of parameters (Lee et al., 2018; Wang et al., 2020), or the Hessian matrix (LeCun et al., 1989; Hassibi & Stork, 1992) at the initialization stage or convergence. Unfortunately, Cerezo & Coles (2021) proved that the gradient information in QNNs with random deep ansatz exponentially vanishes with the increased number of qubits. In other words, the gradient information fails to provide any useful information to guide pruning. Therefore, the classical pruning methods are incapable of extracting the symmetric ansatz from an over-parameterized and asymmetric ansatz with the enhanced trainability.

In contrast with classical pruning methods, symmetric pruning does not require any gradient information to construct the symmetric ansätze. Instead, it removes the redundant gates and shrinks the solution space according to the information of the problem Hamiltonian. A deep comprehension about this topic is given in our companion study (Wang et.al, 2022).

G More numerical simulation details

In this section, we present the evolution of the ansatz structure for the transverse-field Ising model during symmetric pruning in Fig. 4, which serves as an example for better understanding the SP algorithm. Specifically, we adopt the Hardware efficient ansatz as the initial over-parameterized ansatz, shown on the left side of Fig. 4. The gates on the last two wires corresponding to $\mathbb{I}^{\otimes 2}$ are first removed through the sub-step 2-1 (referred to 'SP1') in Alg. 1 to ensure the system symmetry. Subsequently, SP employs the information of problem Hamiltonian H_{TFIM} to remove the parameterized single qubit gates and two-qubit gates on the first six wires such that the pruned ansatz design $\mathcal{A}_{pr} = \{\sigma_1^x, \cdots, \sigma_6^x, \sigma_1^z\sigma_2^z, \cdots \sigma_5^z\sigma_6^z\}$. Finally, spatial symmetry pruning correlates the parameterized gates on the equivalent qubits and qubit-pairs through the returned automorphism by the package *nauty*. In the case of TIFM, the package *nauty* returns a non-trivial automorphism $\pi(j) = n + 1 - j$, that permutes qubits from each side of the chain.

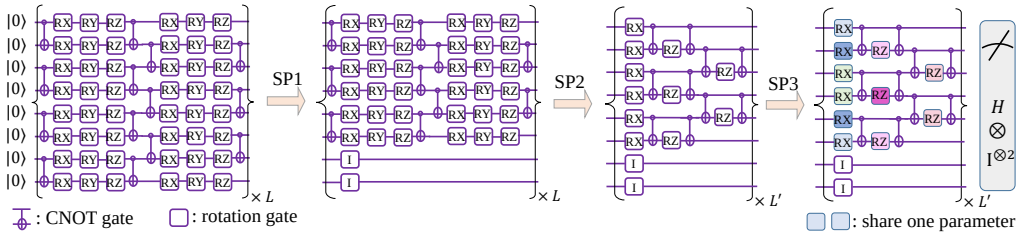

Figure 4: **Evolution of ansatz structure during symmetric pruning** From left to right shows the initial hardware efficient ansatz and ansatz structure at different stages of symmetric pruning, where 'SP1', 'SP2', 'SP3' refer to the sub-steps 2-1, 2-2, and 2-3 in Alg. 1 respectively and $L' < L$. The symbol 'RX' ('RY', 'RZ') refers to the single qubit rotation around the x (y, z)-axis and I refers to the identity gate. The rotation gates with the same color of the pruned ansatz are correlated by one individual parameter per layer.