

# Exponential Hardness of Optimization from the Locality in Quantum Neural Networks

Hao-Kai Zhang<sup>1,3\*</sup>, Chengkai Zhu<sup>2,3\*</sup>, Geng Liu<sup>3</sup>, Xin Wang<sup>2,3†</sup>

<sup>1</sup>Institute for Advanced Study, Tsinghua University, Beijing 100084, China

<sup>2</sup>Thrust of Artificial Intelligence, Information Hub,

The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, China

<sup>3</sup>Institute for Quantum Computing, Baidu Research, Beijing 100193, China

zhk20@mails.tsinghua.edu.cn, czhu696@connect.hkust-gz.edu.cn, luke97624@gmail.com, felixxinwang@hkust-gz.edu.cn

## Abstract

Quantum neural networks (QNNs) have become a leading paradigm for establishing near-term quantum applications in recent years. The trainability issue of QNNs has garnered extensive attention, spurring demand for a comprehensive analysis of QNNs in order to identify viable solutions. In this work, we propose a perspective that characterizes the trainability of QNNs based on their locality. We prove that the entire variation range of the loss function via adjusting any local quantum gate vanishes exponentially in the number of qubits with a high probability for a broad class of QNNs. This result reveals extra harsh constraints independent of gradients and unifies the restrictions on gradient-based and gradient-free optimizations naturally. We showcase the validity of our results with numerical simulations of representative models and examples. Our findings, as a fundamental property of random quantum circuits, deepen the understanding of the role of locality in QNNs and serve as a guideline for assessing the effectiveness of diverse training strategies for quantum neural networks.

## Introduction

Quantum computing is a rapidly growing technology that exploits quantum mechanics to solve intricate problems that classical computers cannot solve. With enormous efforts having been made to develop noisy intermediate scale quantum (NISQ) devices (Preskill 2018), current quantum devices have demonstrated the ability to achieve near-term quantum advantage for practical applications in key areas including many-body physics (Wecker, Hastings, and Troyer 2015; Uvarov, Biamonte, and Yudin 2020), chemistry (McArdle et al. 2020), finance (Egger et al. 2020), and machine learning (Biamonte et al. 2017). Specifically, quantum machine learning (QML) represents an exciting, emerging interdisciplinary field that seeks to enhance machine learning algorithms by harnessing the inherent parallelism of quantum systems (LaRose and Coyle 2020; Cerezo et al. 2022; You et al. 2023; Gebhart et al. 2023; Li et al. 2021; Huang et al. 2022; Yu et al. 2022). Quantum neural networks (QNNs) stand at the forefront of QML, cap-

italizing on the unprecedented potential of quantum computing to revolutionize data analysis and pattern recognition. Inspired by classical neural networks, QNNs employ quantum gates and quantum states as fundamental building blocks within their computational framework. These networks can be trained using a diverse range of methods, including gradient-based optimization techniques akin to classical neural network training (McClellan et al. 2016; Cerezo et al. 2021a; Bharti et al. 2022; Endo et al. 2021).

With the aim to show quantum advantage on certain tasks, a critical issue is whether QNNs can be extended to solve large-scale systems, i.e., scalability. Unfortunately, many studies point out that training of QNNs requires exponential resources with the system size under certain conditions (McClellan et al. 2018; Arrasmith et al. 2020; Cerezo and Coles 2021; Arrasmith et al. 2022; Wang et al. 2021; Holmes et al. 2021; Bittel and Kliesch 2021; Ortiz Marrero, Kieferová, and Wiebe 2021; Stilck Franca and García-Patrón 2021; Uvarov and Biamonte 2021; Campos, Nasrallah, and Biamonte 2021; De Palma et al. 2023). Besides the practical limitations such as noises (Wang et al. 2021), even ideal quantum devices will suffer from the so-called *barren plateau* phenomenon (McClellan et al. 2018), which is the quantum counterpart of vanishing gradient problem in classical machine learning. It was shown that the gradient of the cost function vanishes exponentially in the number of qubits with a high probability for a random initialized QNN with sufficient depth, analogous to the vanishing gradient issue in classical neural networks. Consequently, exponentially vanishing gradients demand exponential precision in the cost function measurement on a quantum device (Knill, Ortiz, and Somma 2007) to make progress in the gradient-based optimization, and hence an exponential complexity in the number of qubits.

Several attempts have been made to avoid barren plateaus, such as higher order derivatives (Huembeli and Dauphin 2021), gradient-free optimizers including gate-by-gate optimization (Nakanishi, Fujii, and Todo 2019; Ostaszewski, Grant, and Benedetti 2019), proper initialization (Grant et al. 2019; Zhang et al. 2022), pre-training including adaptive methods (Verdon et al. 2019; Grimsley et al. 2019; Zhang et al. 2021; Skolik et al. 2021; Grimsley et al. 2023), QNN architectures (Pesah et al. 2021; Liu et al. 2022) and cost

\*These authors contributed equally.

†Corresponding author.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

function choices (Cerezo et al. 2021b; Kieferova, Carlos, and Wiebe 2021). More efforts are needed to study the general effectiveness of these attempts (Arrasmith et al. 2020; Cerezo and Coles 2021) and develop new strategies to improve the trainability and scalability of QNNs. As a guide for exploring effective training strategies, it is crucial to uncover the essential mechanisms behind the barren plateau phenomenon.

However, few rigorous scaling results are known for generic QNNs besides phenomenological calculations, i.e., gradient analyses and their descendent (Arrasmith et al. 2020; Cerezo and Coles 2021; Arrasmith et al. 2022). Instead of just the limited information of vicinity from gradient analyses, it would be quite helpful for designing efficient algorithms if we could gain information on the entire variation range of the cost function when adjusting a single (Nakanishi, Fujii, and Todo 2019; Ostaszewski, Grant, and Benedetti 2019) or several parameters. Combined with the fact that parameters usually enter the circuit independently through local quantum gates, all of which motivate our work where we are chiefly concerned with the variation range of the cost function via varying a local unitary within a quantum circuit.

In this work, we present a rigorous scaling theorem on the trainability of QNNs beyond gradients from the perspective of QNN locality. As summarized in Figure 1, we prove that when varying a *local unitary* within a sufficiently random circuit, the expectation and variance of the variation range of the cost function vanish exponentially in the number of qubits. Then through simple derivations, we show that this theorem implies exponentially vanishing gradients and cost function differences, and hence unifies the restrictions on gradient-based and gradient-free optimizations. Meanwhile, this theorem further delivers extra meaningful information about the training landscapes and optimization possibilities of QNNs. In this sense, we obtain a fundamental limitation on QNN training. Next, we illustrate the applications of our theorem on representative QNN models, where a tighter bound for the fidelity-type cost function is provided specifically even with shallow random circuits. At last, we perform numerical simulations on these representative models, where the scaling exponents coincide with our analytical results almost precisely.

**Comparison with Previous Works.** The advances of our results compared to previous works (McClean et al. 2018; Cerezo and Coles 2021; Arrasmith et al. 2020, 2022; Leone et al. 2022; Thanasilp et al. 2023; Garcia et al. 2023; Larocca et al. 2022) exist in two aspects. Firstly, the exponentially vanishing quantity we claim is the *entire* variation range of the cost function in the *whole parameter subspace* corresponding to the local unitary. This provides constraints on multiple parameters at finite intervals simultaneously, instead of an infinitesimal vicinity or two fixed-parameter points. In particular, Ref. (Arrasmith et al. 2022) considers the cost function difference between two points either both randomly chosen or one random, and the other has a deterministically chosen distance with it. Whereas we consider the difference between the maximum and minimum within

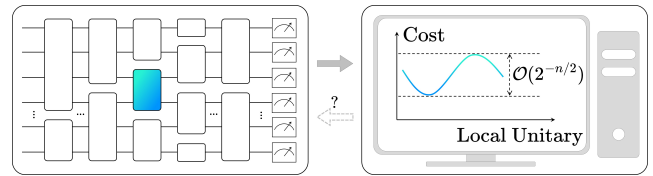


Figure 1: Training limitations from QNN locality. The left part depicts a PQC on  $n$  qubits composed of local unitaries. The right part symbolically depicts the cost function on a classical device vs. the local unitary highlighted in the left part. This work proves that the cost function will fluctuate in an exponentially small range in the number of qubits with a high probability when we vary an arbitrary local unitary within the QNN in certain cases.

the whole subspace w.r.t. a local unitary. The process of taking extreme values in our work is not involved there. Secondly, Our results have a broader scope as they are not dependent on the parameterization of the local unitary, such as  $e^{-i\theta\Omega}$ , used in previous studies. The only requirement for our results is the circuit locality, which opens up new possibilities for analyzing the trainability of QNNs.

## Preliminaries

**Quantum State.** We first introduce basic concepts and notations in quantum computing. A pure single-qubit quantum state is a linear combination of two computational basis states, represented as  $|\phi\rangle = \alpha|0\rangle + \beta|1\rangle$  in Dirac notation, where  $\alpha, \beta \in \mathbb{C}$ ,  $|\alpha|^2 + |\beta|^2 = 1$ . Here,  $|0\rangle$  and  $|1\rangle$  denote the basis states  $[1, 0]^T$  and  $[0, 1]^T$  in the single-qubit space  $\mathbb{C}^2$ , respectively. The  $n$ -qubit space  $\mathbb{C}^{2^n}$  is formed by the tensor product of  $n$  single-qubit spaces. Additionally, the quantum state can be represented by a positive semidefinite matrix, also known as a density matrix. The density matrix  $\rho$  of a pure state  $|\phi\rangle$  consisting of  $n$  qubits is expressed as  $\rho = |\phi\rangle\langle\phi|$ , where  $\langle\phi| = |\phi\rangle^\dagger$ . A general mixed quantum state is represented by  $\rho = \sum_k c_k |\phi_k\rangle\langle\phi_k|$ , where  $c_k \in \mathbb{R}$ ,  $\sum_k c_k = 1$ .

**Quantum Gate.** Quantum gates are mathematically described as unitary operators. Common single-qubit gates include the Pauli rotations  $\{R_P(\theta) = e^{-i\frac{\theta}{2}P} | P \in \{X, Y, Z\}\}$ , which are in the matrix exponential form of Pauli matrices

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

Common two-qubit gates include controlled-X gate CNOT =  $I \oplus X$  ( $\oplus$  is the direct sum) and controlled-Z gate CZ =  $I \oplus Z$ , which can generate quantum entanglement among qubits.

**Quantum Measurement.** Quantum measurement is a quantum operation to obtain information from the quantum system. For example, for a single-qubit state  $|\phi\rangle = \alpha|0\rangle + \beta|1\rangle$ , the outcome of a computational basis measurement is either  $|0\rangle$  with probability  $|\alpha|^2$  or  $|1\rangle$  with probability  $|\beta|^2$ . This measurement operation can be mathematically

referred to as the average of the observable  $O = Z$  under the state  $|\phi\rangle$ :  $\langle\phi|O|\phi\rangle = \text{tr}[Z|\phi\rangle\langle\phi|] = |\alpha|^2 - |\beta|^2$ . Generally, quantum observables  $O$  are Hermitian matrices and  $\mathcal{O}(1/\varepsilon^2)$  times of measurements could give an  $\varepsilon\|O\|_\infty$ -error estimation to the value  $\text{tr}[O\rho]$ , where  $\|\cdot\|_\infty$  is the spectral norm of the matrix.

**Quantum Neural Network.** While classical neural networks operate on classical bits and use classical logic gates, quantum neural networks (QNNs) use quantum bits, or qubits, and quantum gates to process and store information. QNNs are often described as parameterized quantum circuits (PQCs) that are composed of rotation gates with adjustable rotating angles. In general, a QNN takes the mathematical form  $\mathbf{U}(\boldsymbol{\theta}) = \prod_\mu U_\mu(\theta_\mu)W_\mu$ , where  $U_\mu(\theta_\mu) = e^{-i\theta_\mu\Omega_\mu}$  denotes a parameterized gate, such as a single-qubit rotation gate with  $\Omega_\mu$  representing a Hermitian operator, and  $W_\mu$  corresponds to fixed gates like the CNOT gate and SWAP gate. Commonly used templates of QNNs include the hardware efficient ansatz, the alternating-layered ansatz, and the tensor-network-based ansatz (Cerezo et al. 2021b; Ran 2020). Note that QNNs with intermediate classical controls such as QCNNs (Cong, Choi, and Lukin 2019) can also be included in this general form theoretically.

### Limitations of Local Unitary Optimization in QNNs

We start by introducing a general setting of a QNN model used throughout our analysis. A hybrid quantum-classical framework in QML usually uses a classical optimizer to train a QNN, denoted by  $\mathbf{U}$ , with an input state  $\rho$  by minimizing a task-dependent cost function  $C$ , which is typically chosen as the expectation value of some Hermitian operator  $H$ :

$$C_{H,\rho}(\mathbf{U}) = \text{tr}(H\mathbf{U}\rho\mathbf{U}^\dagger). \quad (1)$$

Note that other cost function forms can be regarded as compositions of observable expectations and some other classical post-processing functions. Here we focus on Eq. 1 for simplicity. Divide the whole qubit system into two parts  $A, B$  with  $m$  qubits and  $n - m$  qubits, respectively. Here  $m$  is a fixed constant not scaling with  $n$  so that we call  $A$  a local subsystem. The QNN  $\mathbf{U}$  is often composed of local unitaries on real devices, such as the single-qubit rotation gates and the CNOT gate. We focus on a local unitary  $U_A$  within  $\mathbf{U}$  acting on subsystem  $A$ . As shown in Figure 2, we denote the sub-circuit of  $\mathbf{U}$  before  $U_A$  as  $V_1$  and that behind  $U_A$  as  $V_2$ , such that  $\mathbf{U} = V_2(U_A \otimes I_B)V_1$  where  $I_B$  is the identity operator on  $B$ .  $V_1, V_2$  and  $U_A$  are independent of each other. We also remark that this circuit setting is sufficiently general to cover common representative QNN models, e.g., the variational quantum eigensolver, the quantum autoencoder, and the quantum state learning.

To characterize the training landscape beyond the limited information of the vicinity from gradient analyses, we introduce a central quantity throughout this work, i.e., the *variation range of the cost function* via varying a local unitary.

**Definition 1** For a generic cost function  $C_{H,\rho}(\mathbf{U})$  with a QNN  $\mathbf{U}$  in Eq. 1, we define its variation range with given

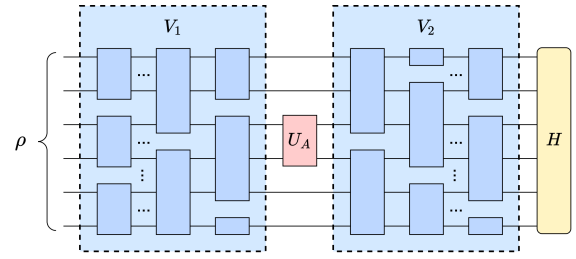


Figure 2: Partition of the QNN in our analysis. The QNN is decomposed as  $\mathbf{U} = V_2(U_A \otimes I_B)V_1$  with an input state  $\rho$  and an observable  $H$ . A tunable local unitary  $U_A$  is implemented by some local quantum gates with the left and right parts assembled as  $V_1$  and  $V_2$ .

$V_1, V_2$  as

$$\Delta_{H,\rho}(V_1, V_2) := \max_{U_A} C_{H,\rho}(\mathbf{U}) - \min_{U_A} C_{H,\rho}(\mathbf{U}), \quad (2)$$

where the maximum and minimum with respect to  $U_A$  are taken over the unitary group  $\mathcal{U}(2^m)$  of degree  $2^m$ .

The quantity  $\Delta_{H,\rho}(V_1, V_2)$  intuitively reflects the maximal possible influence that the local unitary  $U_A$  can have on the cost function. We establish an upper bound on  $\Delta_{H,\rho}(V_1, V_2)$  in the sense of probability by Theorem 1, which thus delivers a limitation on optimizing an arbitrary local unitary. To be specific, we prove that if either  $V_1, V_2$ , or both match the Haar distribution up to the second moment, i.e., are sampled from unitary 2-designs (Dankert et al. 2009), the expectation of  $\Delta_{H,\rho}(V_1, V_2)$  vanishes exponentially in the number of qubits. See Appendix A for preliminaries on unitary designs.

**Theorem 1** Suppose  $\mathbb{V}_1, \mathbb{V}_2$  are ensembles from which  $V_1, V_2$  are sampled, respectively. If either  $\mathbb{V}_1$  or  $\mathbb{V}_2$ , or both form unitary 2-designs, then for arbitrary  $H$  and  $\rho$ , the following inequality holds

$$\mathbb{E}_{\mathbb{V}_1, \mathbb{V}_2}[\Delta_{H,\rho}(V_1, V_2)] \leq \frac{w(H)}{2^{n/2-3m-2}}, \quad (3)$$

where  $\mathbb{E}_{\mathbb{V}_1, \mathbb{V}_2}$  denotes the expectation over  $\mathbb{V}_1, \mathbb{V}_2$  independently.  $w(H) = \lambda_{\max}(H) - \lambda_{\min}(H)$  denotes the spectral width of  $H$ , where  $\lambda_{\max}(H)$  is the maximum eigenvalue of  $H$  and  $\lambda_{\min}(H)$  is the minimum.

Theorem 1 demonstrates that the maximal influence of a local unitary within a random QNN on the cost function diminishes exponentially in the number of qubits, with a high probability. This inherent locality of QNN poses an exponential hardness of optimization in QNN training and we would like to make several remarks to better reveal the underlying implications of the theorem below. The main proof idea of Theorem 1 is to calculate the expectation value over  $\mathbb{V}_1, \mathbb{V}_2$  separately. To tackle the maximization over  $U_A$ , the main technique is to employ Hölder’s inequality to extract  $U_A$  out and bound the remaining part with specific calculations of 2-design element-wise integrals. For the detailed proof, we defer to Appendix B.

**Remark 1** Firstly, due to the non-negativity and boundedness of the variation range, i.e.,  $\Delta_{H,\rho} \in [0, w(H)]$ , the variance of  $\Delta_{H,\rho}$  can be bounded by its expectation times  $w(H)$ . Thus from Theorem 1 we know that the variance also vanishes exponentially:

$$\text{Var}_{V_1, V_2}[\Delta_{H,\rho}(V_1, V_2)] \leq \frac{w^2(H)}{2^{n/2-3m-2}}. \quad (4)$$

Note that  $w(H) \in \mathcal{O}(\text{poly}(n))$  holds for common VQAs. Moreover, Theorem 1 together with Markov’s inequality provides an exponentially small upper bound of the probability that  $\Delta_{H,\rho}(V_1, V_2)$  deviates from zero, i.e.,

$$\Pr[\Delta_{H,\rho}(V_1, V_2) \geq \epsilon] \leq \frac{1}{\epsilon} \cdot \frac{w(H)}{2^{n/2-3m-2}}, \forall \epsilon > 0. \quad (5)$$

That is to say, the probability that  $\Delta_{H,\rho}$  is non-zero to some fixed precision is exponentially small.

**Remark 2** Secondly, we can even establish an exponentially small bound using Theorem 1 for the case where  $U_A$  is a *global unitary* satisfying the parameter-shift rule (Guerreschi and Smelyanskiy 2017; Mitarai et al. 2018; Schuld et al. 2018; Crooks 2019; Mari, Bromley, and Killoran 2021). Suppose  $U_A = e^{-i\theta\Omega}$  with the Hermitian generator  $\Omega$  satisfying  $\Omega^2 = I$  and  $\text{tr} \Omega = 0$ . Then we have  $e^{-i\theta\Omega} = I \cos \theta - i\Omega \sin \theta$  and hence the expectation value w.r.t. a single parameter must be some triangle function, the derivative of which can be exactly expressed as finite difference, satisfying the parameter-shift rule. The conditions  $\Omega^2 = I$  and  $\text{tr} \Omega = 0$  imply that half of the eigenvalues of  $\Omega$  is 1, and the other half  $-1$ . Thus  $\Omega$  has the same eigen spectrum with the local operator  $Z \otimes I \otimes \dots \otimes I$ . there exists a unitary  $W$  such that  $We^{-i\theta\Omega}W^\dagger$  becomes a local unitary acting on a single qubit non-trivially.  $W$  and  $W^\dagger$  could be absorbed into the rest of the circuit with  $W^\dagger \mathbb{V}_1$  or  $\mathbb{V}_2 W$  still forming 2-designs. Therefore, the proof for global unitaries satisfying the parameter-shift rule can be reduced back to the case of local unitaries. However, for the global unitary that does not satisfy the parameter-shift rule, such as the controlled Pauli rotation gates, this argument does not generally hold.

**Remark 3** Thirdly, it’s important to underscore that our theorem involves a scaling relationship of the gate locality  $m$  with respect to the qubit count  $n$ . Our key emphasis lies in the situation of the local gate where  $m$  does not scale with  $n$  so that our theorem can give a non-trivial exponential upper bound. This is indeed a situation frequently encountered in practice like in variational quantum eigensolver and quantum state learning, considering the fact that common elementary quantum gates available on digital quantum computers (e.g., Pauli rotation gates and CNOT gates) are inherently local. This aligns with the hardware constraints and technological advancements shaping current quantum devices. For global unitaries within an ansatz, e.g., QAOA, which is beyond the scope of our theorem, we usually realize it by correlating parameters among many local parameterized gates. This inspires us that one possible strategy to

escape from the vanishing variation range of the cost function is correlating parameters in multiple local gates.

**Remark 4** Moreover, it is worth noticing that the compact bound in Eq. 3 only involves the spectral width  $w(H)$  and does not depend on any detail of the Hermitian operator  $H$ . But if some specific structures about  $H$  are known, e.g., the Pauli decomposition of  $H$ , a tighter bound could be derived in Appendix B which depends on the coupling complexity of  $H$ . In addition, if the cost function reduces to the form of the fidelity between pure states, we could have a tighter bound with scaling  $\mathcal{O}(2^{-n})$  in Proposition 2. Theorem 1 can be generalized to arbitrary dimensions besides qubit systems of dimension  $2^n$ , e.g., qutrit and qudit systems. The detailed proof is provided in Appendix B.

In fact, Theorem 1 has a natural physical interpretation: the effect of a local operation on a physical observable will vanish exponentially after a chaotic evolution. Remarkably, the concept of local operations yielding minor global influences is a physically intuitive yet mathematically intricate notion. For instance, even a single-qubit unitary is enough to rotate an arbitrary  $n$ -qubit pure state to a new state with zero fidelity with the original one, showcasing local operations do make a great global influence. Hence, Theorem 1 may be invaluable as a rigorous formulation of the aforementioned argument within the domain of QNN training, elucidating the locality of QNNs.

### Unifying the Limitations on Training QNNs

Here we briefly demonstrate how Theorem 1 unifies the restrictions on gradient-based (McClellan et al. 2018; Cerezo and Coles 2021) and gradient-free optimizations (Arrasmith et al. 2020, 2022) in a more natural manner, and indicates the extra restrictions besides the PQC on QNN training. In the following, we focus on a PQC applicable for Theorem 1 with  $M$  trainable parameters  $\{\theta_\mu\}_{\mu=1}^M$  and denote the variation range of the cost function via varying  $\theta_\mu$  as  $\Delta_\mu$ .

Consider the gradient-based optimization first. On the one hand, in the case where the parameter-shift rule is valid (Guerreschi and Smelyanskiy 2017; Mitarai et al. 2018; Schuld et al. 2018; Crooks 2019; Mari, Bromley, and Killoran 2021), Theorem 1 can strictly deduce vanishing gradients. Suppose  $\{\theta_\mu\}_{\mu=1}^M$  are applicable for the parameter-shift rule (e.g., hardware-efficient ansatzes). Namely,  $\theta_\mu$  enters the unitary  $e^{-i\theta_\mu\Omega_\mu}$  within the circuit where  $\Omega_\mu$  is a Hermitian generator satisfying  $\Omega_\mu^2 = I$ . From Theorem 1 we know that the expectation of  $\Delta_\mu$  vanishes exponentially. Therefore, the derivative  $\partial_\mu C := \frac{\partial C}{\partial \theta_\mu}$  with respect to  $\theta_\mu$  satisfies

$$\begin{aligned} \mathbb{E}[|\partial_\mu C|] &= \mathbb{E} \left[ \left| C \left( \boldsymbol{\theta} + \frac{\pi}{4} \mathbf{e}_\mu \right) - C \left( \boldsymbol{\theta} - \frac{\pi}{4} \mathbf{e}_\mu \right) \right| \right] \\ &\leq \mathbb{E}[\Delta_\mu] \in \mathcal{O}(2^{-n/2}), \end{aligned} \quad (6)$$

where  $\mathbf{e}_\mu$  is the unit vector in the parameter space corresponding to  $\theta_\mu$ . From Markov’s inequality as in Eq. 5, we know that the probability that the derivative  $\partial_\mu C$  deviates from zero by a small constant is exponentially small.

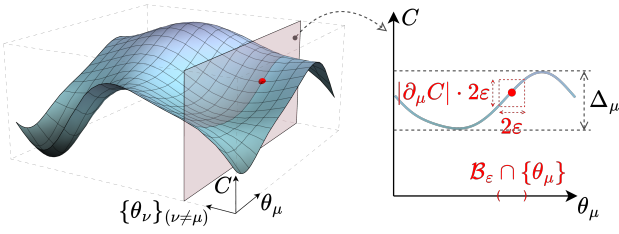


Figure 3: Sketch of our results implying vanishing gradients. The left panel sketches the whole training landscape with one of the parameters  $\theta_\mu$  as the  $x$ -axis, all of the other parameters  $\{\theta_\nu\}_{\nu \neq \mu}$  as the  $y$ -axis symbolically and the cost function value  $C$  as the  $z$ -axis. The right panel depicts a typical sample of the  $z$ - $x$  cross-section from the landscape on the left with variation range  $\Delta_\mu$ . Up to the linear approximation error,  $\Delta_\mu$  serves as an upper bound for the absolute derivative  $|\partial_\mu C|$  times the vicinity size  $2\varepsilon$ .

On the other hand, even in the absence of the parameter-shift rule, vanishing gradients could still be obtained approximately by the following arguments. Consider the vicinity of a random initialized parameter point where the linear approximation error is negligible, denoted as an  $\varepsilon$ -ball  $\mathcal{B}_\varepsilon$  of radius  $\varepsilon$  (here  $\varepsilon$  plays the same role as the learning rate). As shown in Figure 3, the linearity in  $\mathcal{B}_\varepsilon$  together with Theorem 1 leads to

$$\mathbb{E}[|\partial_\mu C|] \leq \mathbb{E}\left[\frac{\Delta_\mu}{2\varepsilon}\right] \in \mathcal{O}(2^{-n/2} \frac{1}{\varepsilon}), \quad (7)$$

up to the linear approximation error, where  $1/\varepsilon$  is not an essential factor since it reflects the frequencies of the landscape fluctuation rather than magnitudes, similar to the role of the factor  $\text{tr}(V^2)$  in the expression of  $\text{Var}[\partial_\mu C]$  (McClean et al. 2018).

For the gradient-free optimization based on the cost function difference between any two *fixed* parameter points  $\theta'$  and  $\theta$ , Theorem 1 leads to

$$\begin{aligned} \mathbb{E}[|C(\theta') - C(\theta)|] &\leq \mathbb{E}\left[\sum_{\mu=1}^M |C(\theta^{(\mu)}) - C(\theta^{(\mu-1)})|\right] \\ &\leq \sum_{\mu=1}^M \mathbb{E}[|\Delta_\mu|] \in \mathcal{O}(M2^{-n/2}), \end{aligned}$$

where  $\theta^{(\mu)} = \theta + \sum_{\nu=1}^{\mu} (\theta'_\nu - \theta_\nu) \mathbf{e}_\nu$  for  $\mu = 1, \dots, M$  and  $\theta^{(\mu)} = \theta$  for  $\mu = 0$ . Thus, as long as the number of parameters satisfies  $M \in \mathcal{O}(\text{poly}(n))$ , the cost function difference between any two points vanishes exponentially with a high probability, demanding an exponential precision to make progress in the gradient-free optimization.

Furthermore, Theorem 1 goes beyond vanishing gradients and vanishing differences between two fixed points. The exponentially vanishing quantity claimed by Theorem 1 is the variation range of the cost function in the *whole parameter subspace* corresponding to a local unitary, e.g., the subspace of the 3 Euler angles in a single-qubit rotation gate from

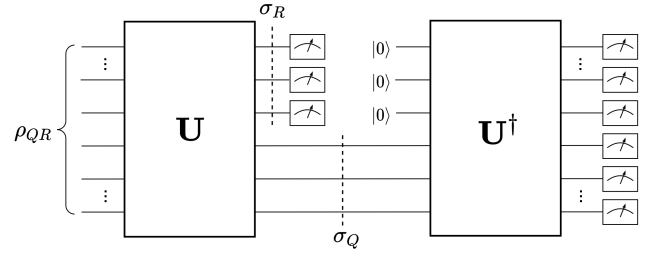


Figure 4: Circuit setting of the quantum autoencoder.  $\rho_{QR}$  is the given state to be compressed and  $\sigma_Q$  is the compressed state through the encoder  $\mathbf{U}$ . The quantum autoencoder aims to train  $\mathbf{U}$  such that  $\rho_{QR}$  can be reconstructed from  $\sigma_Q$  with high fidelity through the decoder  $\mathbf{U}^\dagger$  combined with an ancilla zero state  $|0\rangle\langle 0|_R$ .  $\sigma_R$  denotes the state of the discarded part after compression.

$SU(2)$ , or the subspace of the 15 parameters in a two-qubit rotation gate from  $SU(4)$ , etc. This gives constraints on multiple parameters at finite intervals simultaneously, instead of a vicinity or two fixed parameter points.

### Application on Representative QNN Models

To better illustrate the meaning of our findings in practice, we investigate the applications of Theorem 1 on three representative QNN models, including the variational quantum eigensolver (VQE), quantum autoencoder, and quantum state learning. The corresponding numerical simulation results are summarized in Figure 5.

**Application on VQE.** The variational quantum eigensolver is the most famous implementation of a hybrid quantum-classical algorithm with the goal of preparing the ground state of a given Hamiltonian  $\hat{H}$  of a physical system (Peruzzo et al. 2014). The cost function is the energy expectation with respect to an ansatz state  $\mathbf{U}|0\rangle$ , i.e.,

$$C_{\text{VQE}}(\mathbf{U}) = \langle 0|\mathbf{U}^\dagger \hat{H} \mathbf{U}|0\rangle. \quad (8)$$

For most physical models with local interactions, the spectral width is proportional to the system size, i.e.,  $w(\hat{H}) \in \mathcal{O}(n)$ . For common repeated-layer-type ansatzes, e.g., the hardware-efficient ansatzes (Kandala et al. 2017), linear depth  $\mathcal{O}(n)$  is enough to make a randomly initialized circuit to be a sample from an approximate 2-design ensemble (McClean et al. 2018; Harrow, Hassidim, and Lloyd 2009; Brandão, Harrow, and Horodecki 2016). Hence from Theorem 1 we know that  $\Delta_{\text{VQE}}(V_1, V_2)$  vanishes exponentially with a high probability for random circuits forming 2-designs. We conduct numerical simulations for the variation range of the VQE cost function  $\Delta_{\text{VQE}}$  using the 1-dimensional spin-1/2 antiferromagnetic Heisenberg model:

$$\hat{H} = \sum_{i=1}^n (X_i X_{i+1} + Y_i Y_{i+1} + Z_i Z_{i+1}), \quad (9)$$

with periodic boundary condition, as shown in Figure 5(a).

**Application on Quantum Autoencoder.** The quantum autoencoder (QAE) is an approach for quantum data compression (Romero, Olson, and Aspuru-Guzik 2017; Cao and Wang 2021). As shown in Figure 4, a QNN  $\mathbf{U}$  is trained as an encoder to compress a given state  $\rho_{QR}$  on a bipartite system  $QR$  into a reduced state  $\sigma_Q = \text{tr}_R(\mathbf{U}\rho_{QR}\mathbf{U}^\dagger)$  on subsystem  $Q$ , such that  $\rho_{QR}$  can be reproduced from  $\sigma_Q$  by the decoder isometry  $\langle 0|_R\mathbf{U}^\dagger$  with a high fidelity. According to the monotonicity of the fidelity under partial trace, an easy-to-measure cost function could be reduced from the fidelity between  $\rho_{QR}$  and the reconstructed state as

$$C_{\text{QAE}}(\mathbf{U}) := 1 - \text{tr}(|0\rangle\langle 0|_R \otimes I_Q \mathbf{U}\rho_{QR}\mathbf{U}^\dagger), \quad (10)$$

where the second term is exactly the fidelity between the state of the discarded part  $\sigma_R = \text{tr}_Q(\mathbf{U}\rho_{QR}\mathbf{U}^\dagger)$  and the zero state  $|0\rangle_R$  on subsystem  $R$ . The spectral width for the QAE cost function in Eq. 10 is  $w(H_{\text{QAE}}) = 1$  with  $H_{\text{QAE}} = I_{QR} - |0\rangle\langle 0|_R \otimes I_Q$ . Thus again from Theorem 1 we know that  $\Delta_{\text{QAE}}(V_1, V_2)$  vanishes exponentially in the number of qubits, specifically with the scaling  $\mathcal{O}(2^{-n/2})$  as shown in Figure 5(b).

**Application on Quantum State Learning.** The fidelity between pure states is a special case of the cost function in Eq. 1 with a low-rank observable. Many QML applications make use of fidelity as their cost functions (Lee, Lee, and Bang 2018; Shirakawa, Ueda, and Yunoki 2021; Bravo-Prieto et al. 2023). Here we uniformly call them quantum state learning (QSL) tasks. Denote the input state as  $|\psi\rangle$  and the target state as  $|\phi\rangle$ . The QSL cost function can be written as

$$C_{\text{QSL}}(\mathbf{U}) = 1 - |\langle \phi | \mathbf{U} | \psi \rangle|^2. \quad (11)$$

Theorem 1 can be applied here with  $H_{\text{QSL}} = I - |\phi\rangle\langle \phi|$  and  $w(H_{\text{QSL}}) = 1$ . Here a tighter bound for  $\Delta_{\text{QSL}}$  is provided in Proposition 2, which generally holds for the Bures fidelity. The proof of Proposition 2 is detailed in Appendix C.

**Proposition 2** *If either  $\mathbb{V}_1$  or  $\mathbb{V}_2$ , or both form unitary 1-designs, then for the variation range of the fidelity-type cost function  $\Delta_{\text{QSL}}$ , the following inequality holds*

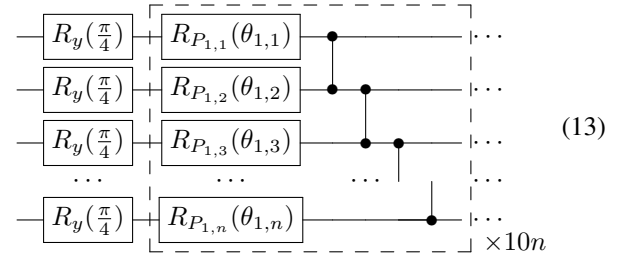
$$\mathbb{E}_{V_1, V_2} [\Delta_{\text{QSL}}(V_1, V_2)] \leq \frac{1}{2^{n-2m}}. \quad (12)$$

Compared with Theorem 1, the bound  $\mathcal{O}(2^{-n})$  becomes tighter and the demanded randomness becomes weaker in this special case. Notably, even a random circuit of constant depth is enough to form a 1-design, which is much shallower than 2-designs. Like in Eq. 4 and Eq. 5, the variance and the probability that  $\Delta_{\text{QSL}}$  deviates from zero also vanish exponentially, but only require random circuits forming unitary 1-designs. Moreover, still with 1-designs, Proposition 2 implies exponentially vanishing cost gradients and cost differences in the same way as Theorem 1, which may be considered as the underlying mechanism behind the severe barren plateaus for global cost functions even with shallow quantum circuits (Cerezo et al. 2021b).

## Numerical Simulations of Experiments

Previously, we have theoretically shown that with a high probability, the maximal influence of a local unitary within a random QNN on the cost function will vanish exponentially in the number of qubits. We further demonstrate the validity of our results with numerical simulations of experiments on the three representative QNN models. All of these experimental results show the exponentially vanishing variation range in the number of qubits, which is consistent with Theorem 1 and Proposition 2.

**Circuit Setting.** Consider subsystem  $A$  only containing a single qubit, namely  $m = 1$ , and parameterize the local unitary  $U_A \in \mathcal{U}(2)$  with 3 Euler angles up to a global phase, i.e.,  $U_A(\phi, \theta, \alpha) = R_z(\phi)R_y(\theta)R_z(\alpha)$ , where  $R_y$  and  $R_z$  are single-qubit rotation gates with generators being  $Y$  and  $Z$  Pauli matrices. To construct random circuits forming 2-designs as  $V_1$  or  $V_2$  used in the VQE and QAE examples, we employ the following hardware-efficient ansatz as in (McClean et al. 2018) for comparison.



A single layer of  $R_y(\pi/4) = \exp(-iY\pi/8)$  gates are laid at the very beginning of the circuit to make the three rotation axes have equal status, then followed by  $10 \times n$  repeated layers. Each layer consists of  $n$  single-qubit rotation gates  $R_P(\theta)$  on each qubit together with  $n - 1$  controlled phase gates between nearest neighboring qubits aligned as a 1-dimensional array, where the rotation axes  $P \in \{x, y, z\}$  is chosen with uniform probability and  $\theta \in [0, 2\pi)$  is also chosen uniformly. A such random circuit with  $\mathcal{O}(n)$  repeated layers could be considered as an approximate 2-design (here we employ  $10 \times n$ ) (McClean et al. 2018; Harrow, Hasidim, and Lloyd 2009; Brandão, Harrow, and Horodecki 2016). Experimental results with different numbers of layers are also presented in Appendix D to show how the expectation of the cost variation range  $\Delta_{H,\rho}$  vanishes with the circuit depth. To construct random circuits forming 1-designs used in the QSL example, we just replace the repeated layers above with a single layer of  $S\mathcal{U}(2)$  elements  $R_z(\phi)R_y(\theta)R_z(\alpha)$  on each qubit with  $\phi, \theta, \alpha \in [0, 2\pi)$  are chosen with uniform probability.

**Implementation Details.** To compute  $\max_{U_A} C$  and  $\min_{U_A} C$  in the definition of  $\Delta_{H,\rho}(V_1, V_2)$  with respect to  $U_A$ , we employ the Adam optimizer to update  $U_A$  iteratively until convergence for each of the 100 samples of  $V_1, V_2$ . We consider the converged value as a good estimation with a tolerable error at least for circuits with a small number of qubits ( $\leq 10$ ) and a modest depth ( $\leq 10 \times n$ ). We repeat this procedure for different numbers of qubits and different statistics of  $\mathbb{V}_1$  and  $\mathbb{V}_2$ , i.e.,  $\mathbb{V}_1$  or  $\mathbb{V}_2$  being a 2-design (1-design) while the other being identity.

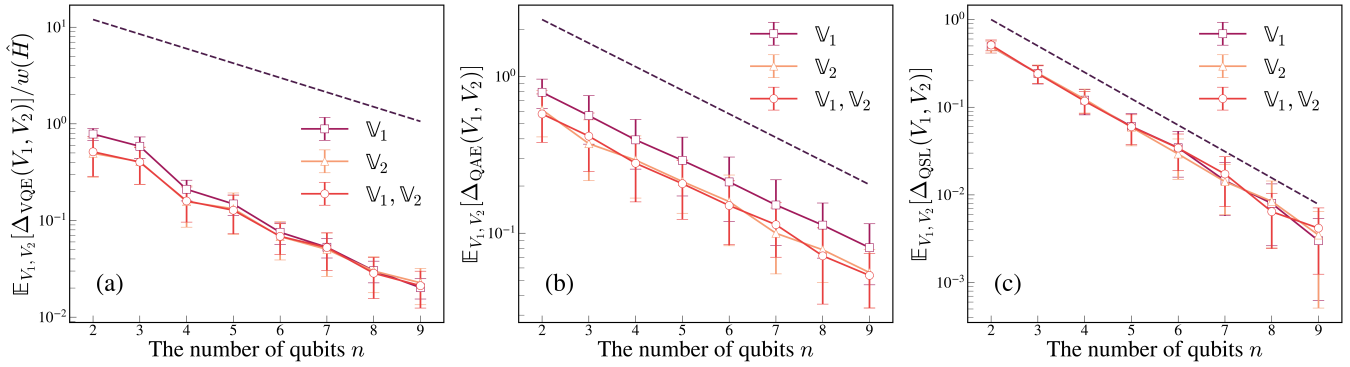


Figure 5: Exponentially vanishing variation range of the cost function via varying a local unitary. The data points represent the sample averages of the cost variation range  $\Delta_{H,\rho}$  via varying a single-qubit unitary over the spectral width  $w(H)$  as a function of the number of qubits on semi-log plots. Panel (a) and (b) correspond to the VQE with the 1-dimensional Heisenberg model and the quantum autoencoder with one qubit discarded, respectively, where the error bars represent the standard deviations over samples. Panel (c) corresponds to the quantum state learning with the cost function being the fidelity with the zero state. Different legends stand for  $\mathbb{V}_1$ ,  $\mathbb{V}_2$  or both being approximate 2-designs in (a), (b) and 1-designs in (c). The dashed lines depict our theoretical upper bounds for the three tasks where the scaling exponents show a good coincidence with the experimental results.

**Numerical Results.** We summarize the simulation results of the three examples in Figure 5. The slopes of the lines imply the rates of exponential decay. The data points represent the sample averages of the cost variation range  $\Delta_{H,\rho}$  via varying  $U_A$  over  $w(H)$ , and the error bars represent the standard deviations over samples. We specially rescale the error bar in the QSL example as a quarter of the standard deviation for better presentation on semi-log plots. One can see that in all the cases, the expectations of  $\Delta_{H,\rho}(V_1, V_2)$  vanish exponentially in the number of qubits. The data lines are almost parallel to the dashed lines depicting the theoretical upper bounds. That is to say, the scaling behaviors almost coincide with the predictions from Theorem 1 and Proposition 2. These results suggest that while optimizing a local unitary within a random QNN, the cost function exhibits fluctuations within an exponentially small range relative to the number of qubits. It is this phenomenon that elucidates the vanishing gradient issue and contributes to the exponential difficulty of training as the QNN scales up. A detailed derivation can be found in Appendix B for the tighter task-dependent upper bounds used in Figure 5(a) and (b).

## Conclusion

We have shown that the maximal possible influence of a local unitary within a QNN on the cost function vanishes exponentially in the number of qubits with a high probability. This finding unveils the exponential hardness associated with training QNNs as they scale up. The randomness required is just a 2-design for the generic cost function and a 1-design for the fidelity-type cost function, in spite that the integrand  $\Delta_{H,\rho}(V_1, V_2)$  is not necessarily a polynomial of degree at most 2 or 1 in the entries of  $V_1$  and  $V_2$ . We remark that a 2-design circuit can be achieved approximately by only  $\mathcal{O}(n)$  depth (McClean et al. 2018; Harrow, Hasidim, and Lloyd 2009; Brandão, Harrow, and Horodecki

2016) for common repeated-layer-type ansatzes, e.g., the hardware-efficient ansatzes (Kandala et al. 2017), and a 1-design circuit can be achieved more easily by only  $\mathcal{O}(1)$  depth.

From the perspective of quantum information theory, our results can be regarded as a basic property of random quantum circuits. That is, a local unitary within a random circuit of polynomial depth has an exponentially small impact on the expectation of physical observables, which is expected to have potential applications in other areas involving random quantum circuits.

For the training of QNN, our results unify the restrictions on gradient-based and gradient-free optimizations in a natural way and hence can be regarded as the underlying mechanism behind the barren plateau phenomenon. Therefore, a fundamental limitation is unraveled in training QNNs, which can serve as a guide for designing better training strategies to improve the scalability of QNNs. A direct consequence is that the gate-by-gate optimization strategy (Nakanishi, Fujii, and Todo 2019; Ostaszewski, Grant, and Benedetti 2019) is ineffective no matter what optimizers are utilized. Reparameterization within local unitaries is also unhelpful. For future research, it will be of great interest to explore potential solutions via proper initialization (Grant et al. 2019), pre-training including adaptive methods (Verdon et al. 2019; Grimsley et al. 2019; Zhang et al. 2021; Skolik et al. 2021; Grimsley et al. 2023), circuit architectures (Pesah et al. 2021; Liu et al. 2022) and cost function choices (Cerezo et al. 2021b; Kieferova, Carlos, and Wiebe 2021).

## Acknowledgements

This work was supported by the Start-up Fund from The Hong Kong University of Science and Technology (Guangzhou), the Guangdong Quantum Science and Technology Strategic Fund (Grant No. GDZX2303007), and the

Education Bureau of Guangzhou Municipality.

## References

- Arrasmith, A.; Cerezo, M.; Czarnik, P.; Cincio, L.; and Coles, P. J. 2020. Effect of barren plateaus on gradient-free optimization. *Quantum*, 5: 1–9.
- Arrasmith, A.; Holmes, Z.; Cerezo, M.; and Coles, P. J. 2022. Equivalence of quantum barren plateaus to cost concentration and narrow gorges. *Quantum Science and Technology*, 7(4): 045015.
- Bharti, K.; Cervera-Lierta, A.; Kyaw, T. H.; Haug, T.; Alperin-Lea, S.; Anand, A.; Degroote, M.; Heimonen, H.; Kottmann, J. S.; Menke, T.; Mok, W.-K.; Sim, S.; Kwek, L.-C.; and Aspuru-Guzik, A. 2022. Noisy intermediate-scale quantum algorithms. *Reviews of Modern Physics*, 94(1).
- Biamonte, J.; Wittek, P.; Pancotti, N.; Rebentrost, P.; Wiebe, N.; and Lloyd, S. 2017. Quantum machine learning. *Nature*, 549(7671): 195–202.
- Bittel, L.; and Kliesch, M. 2021. Training Variational Quantum Algorithms Is NP-Hard. *Physical Review Letters*, 127(12): 120502.
- Brandão, F. G. S. L.; Harrow, A. W.; and Horodecki, M. 2016. Local Random Quantum Circuits are Approximate Polynomial-Designs. *Communications in Mathematical Physics*, 346(2): 397–434.
- Bravo-Prieto, C.; LaRose, R.; Cerezo, M.; Subasi, Y.; Cincio, L.; and Coles, P. J. 2023. Variational Quantum Linear Solver. *Quantum*, 7: 1188.
- Campos, E.; Nasrallah, A.; and Biamonte, J. 2021. Abrupt transitions in variational quantum circuit training. *Physical Review A*, 103(3): 032607.
- Cao, C.; and Wang, X. 2021. Noise-Assisted Quantum Autoencoder. *Physical Review Applied*, 15(5): 054012.
- Cerezo, M.; Arrasmith, A.; Babbush, R.; Benjamin, S. C.; Endo, S.; Fujii, K.; McClean, J. R.; Mitarai, K.; Yuan, X.; Cincio, L.; and Coles, P. J. 2021a. Variational quantum algorithms. *Nature Reviews Physics*, 3(9): 625–644.
- Cerezo, M.; and Coles, P. J. 2021. Higher order derivatives of quantum neural networks with barren plateaus. *Quantum Science and Technology*, 6(3): 035006.
- Cerezo, M.; Sone, A.; Volkoff, T.; Cincio, L.; and Coles, P. J. 2021b. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications*, 12(1): 1791.
- Cerezo, M.; Verdon, G.; Huang, H.-Y.; Cincio, L.; and Coles, P. J. 2022. Challenges and opportunities in quantum machine learning. *Nature Computational Science*, 2(9): 567–576.
- Cong, I.; Choi, S.; and Lukin, M. D. 2019. Quantum convolutional neural networks. *Nature Physics*, 15(12): 1273–1278.
- Crooks, G. E. 2019. Gradients of parameterized quantum gates using the parameter-shift rule and gate decomposition. arXiv:1905.13311.
- Dankert, C.; Cleve, R.; Emerson, J.; and Livine, E. 2009. Exact and approximate unitary 2-designs and their application to fidelity estimation. *Physical Review A*, 80(1): 012304.
- De Palma, G.; Marvian, M.; Rouzé, C.; and França, D. S. 2023. Limitations of Variational Quantum Algorithms: A Quantum Optimal Transport Approach. *PRX Quantum*, 4(1).
- Egger, D. J.; Gambella, C.; Marecek, J.; McFaddin, S.; Mevissen, M.; Raymond, R.; Simonetto, A.; Woerner, S.; and Yndurain, E. 2020. Quantum Computing for Finance: State-of-the-Art and Future Prospects. *IEEE Transactions on Quantum Engineering*, 1: 1–24.
- Endo, S.; Cai, Z.; Benjamin, S. C.; and Yuan, X. 2021. Hybrid Quantum-Classical Algorithms and Quantum Error Mitigation. *Journal of the Physical Society of Japan*, 90(3): 032001.
- Garcia, R. J.; Zhao, C.; Bu, K.; and Jaffe, A. 2023. Barren plateaus from learning scramblers with local cost functions. *Journal of High Energy Physics*, 2023(1): 1–79.
- Gebhart, V.; Santagati, R.; Gentile, A. A.; Gauger, E. M.; Craig, D.; Ares, N.; Banchi, L.; Marquardt, F.; Pezzè, L.; and Bonato, C. 2023. Learning quantum systems. *Nature Reviews Physics*, 5(3): 141–156.
- Grant, E.; Wossnig, L.; Ostaszewski, M.; and Benedetti, M. 2019. An initialization strategy for addressing barren plateaus in parametrized quantum circuits. *Quantum*, 3.
- Grimsley, H. R.; Barron, G. S.; Barnes, E.; Economou, S. E.; and Mayhall, N. J. 2023. Adaptive, problem-tailored variational quantum eigensolver mitigates rough parameter landscapes and barren plateaus. *npj Quantum Information*, 9(1).
- Grimsley, H. R.; Economou, S. E.; Barnes, E.; and Mayhall, N. J. 2019. An adaptive variational algorithm for exact molecular simulations on a quantum computer. *Nature Communications*, 10(1): 3007.
- Guerreschi, G. G.; and Smelyanskiy, M. 2017. Practical optimization for hybrid quantum-classical algorithms. arXiv:1701.01450.
- Harrow, A. W.; Hassidim, A.; and Lloyd, S. 2009. Quantum Algorithm for Linear Systems of Equations. *Physical Review Letters*, 103(15): 150502.
- Holmes, Z.; Sharma, K.; Cerezo, M.; and Coles, P. J. 2021. Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *PRX Quantum*, 3(1): 1–20.
- Huang, H.-Y.; Kueng, R.; Torlai, G.; Albert, V. V.; and Preskill, J. 2022. Provably efficient machine learning for quantum many-body problems. *Science*, 377(6613): eabk3333.
- Huembeli, P.; and Dauphin, A. 2021. Characterizing the loss landscape of variational quantum circuits. *Quantum Science and Technology*, 6(2): 025011.
- Kandala, A.; Mezzacapo, A.; Temme, K.; Takita, M.; Brink, M.; Chow, J. M.; and Gambetta, J. M. 2017. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, 549(7671): 242–246.



- Kieferova, M.; Carlos, O. M.; and Wiebe, N. 2021. Quantum Generative Training Using Rényi Divergences. arXiv:2106.09567.
- Knill, E.; Ortiz, G.; and Somma, R. D. 2007. Optimal quantum measurements of expectation values of observables. *Phys. Rev. A*, 75: 012328.
- Larocca, M.; Czarnik, P.; Sharma, K.; Muraleedharan, G.; Coles, P. J.; and Cerezo, M. 2022. Diagnosing barren plateaus with tools from quantum optimal control. *Quantum*, 6: 824.
- LaRose, R.; and Coyle, B. 2020. Robust data encodings for quantum classifiers. *Physical Review A*, 102(3): 032420.
- Lee, S. M.; Lee, J.; and Bang, J. 2018. Learning unknown pure quantum states. *Physical Review A*, 98(5): 052302.
- Leone, L.; Oliviero, S. F. E.; Cincio, L.; and Cerezo, M. 2022. On the practical usefulness of the Hardware Efficient Ansatz. arXiv:2211.01477.
- Li, T.; Wang, C.; Chakrabarti, S.; and Wu, X. 2021. Sub-linear classical and quantum algorithms for general matrix games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8465–8473. ISBN 2374-3468.
- Liu, X.; Liu, G.; Huang, J.; Zhang, H.-K.; and Wang, X. 2022. Mitigating barren plateaus of variational quantum eigensolvers. arXiv:2205.13539.
- Mari, A.; Bromley, T. R.; and Killoran, N. 2021. Estimating the gradient and higher-order derivatives on quantum hardware. *Physical Review A*, 103(1): 012405.
- McArdle, S.; Endo, S.; Aspuru-Guzik, A.; Benjamin, S. C.; and Yuan, X. 2020. Quantum computational chemistry. *Reviews of Modern Physics*, 92(1): 015003.
- McClean, J. R.; Boixo, S.; Smelyanskiy, V. N.; Babbush, R.; and Neven, H. 2018. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9(1): 1–7.
- McClean, J. R.; Romero, J.; Babbush, R.; and Aspuru-Guzik, A. 2016. The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics*, 18(2): 023023.
- Mitarai, K.; Negoro, M.; Kitagawa, M.; and Fujii, K. 2018. Quantum circuit learning. *Physical Review A*, 98(3).
- Nakanishi, K. M.; Fujii, K.; and Todo, S. 2019. Sequential minimal optimization for quantum-classical hybrid algorithms. *Physical Review Research*, 2(4): 1–11.
- Ortiz Marrero, C.; Kieferová, M.; and Wiebe, N. 2021. Entanglement-Induced Barren Plateaus. *PRX Quantum*, 2(4): 040316.
- Ostaszewski, M.; Grant, E.; and Benedetti, M. 2019. Structure optimization for parameterized quantum circuits. *Quantum*, 5: 1–13.
- Peruzzo, A.; McClean, J.; Shadbolt, P.; Yung, M.-H.; Zhou, X.-Q.; Love, P. J.; Aspuru-Guzik, A.; and O’Brien, J. L. 2014. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*, 5(1): 4213.
- Pesah, A.; Cerezo, M.; Wang, S.; Volkoff, T.; Sornborger, A. T.; and Coles, P. J. 2021. Absence of Barren Plateaus in Quantum Convolutional Neural Networks. *Physical Review X*, 11(4): 041011.
- Preskill, J. 2018. Quantum Computing in the NISQ era and beyond. *Quantum*, 2: 79.
- Ran, S.-J. 2020. Encoding of matrix product states into quantum circuits of one- and two-qubit gates. *Physical Review A*, 101(3): 032310.
- Romero, J.; Olson, J. P.; and Aspuru-Guzik, A. 2017. Quantum autoencoders for efficient compression of quantum data. *Quantum Science and Technology*, 2(4): 1–10.
- Schuld, M.; Bergholm, V.; Gogolin, C.; Izaac, J.; and Killoran, N. 2018. Evaluating analytic gradients on quantum hardware. *Physical Review A*, 99(3): 032331.
- Shirakawa, T.; Ueda, H.; and Yunoki, S. 2021. Automatic quantum circuit encoding of a given arbitrary quantum state. arXiv:2112.14524.
- Skolik, A.; McClean, J. R.; Mohseni, M.; van der Smagt, P.; and Leib, M. 2021. Layerwise learning for quantum neural networks. *Quantum Machine Intelligence*, 3(1): 5.
- Stilck Franca, D.; and García-Patrón, R. 2021. Limitations of optimization algorithms on noisy quantum devices. *Nature Physics*, 17(11): 1221–1227.
- Thanasilp, S.; Wang, S.; Nghiem, N. A.; Coles, P.; and Cerezo, M. 2023. Subtleties in the trainability of quantum machine learning models. *Quantum Machine Intelligence*, 5(1): 21.
- Uvarov, A.; Biamonte, J.; and Yudin, D. 2020. Variational Quantum Eigensolver for Frustrated Quantum Systems. *Physical Review B*, 102(7): 075104.
- Uvarov, A. V.; and Biamonte, J. D. 2021. On barren plateaus and cost function locality in variational quantum algorithms. *Journal of Physics A: Mathematical and Theoretical*, 54(24): 245301.
- Verdon, G.; Broughton, M.; McClean, J. R.; Sung, K. J.; Babbush, R.; Jiang, Z.; Neven, H.; and Mohseni, M. 2019. Learning to learn with quantum neural networks via classical neural networks. arXiv:1907.05415.
- Wang, S.; Fontana, E.; Cerezo, M.; Sharma, K.; Sone, A.; Cincio, L.; and Coles, P. J. 2021. Noise-induced barren plateaus in variational quantum algorithms. *Nature Communications*, 12(1): 6961.
- Wecker, D.; Hastings, M. B.; and Troyer, M. 2015. Progress towards practical quantum variational algorithms. *Physical Review A*, 92(4): 042303.
- You, X.; Chakrabarti, S.; Chen, B.; and Wu, X. 2023. Analyzing Convergence in Quantum Neural Networks: Deviations from Neural Tangent Kernels. arXiv preprint arXiv:2303.14844.
- Yu, Z.; Yao, H.; Li, M.; and Wang, X. 2022. Power and limitations of single-qubit native quantum neural networks. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*.
- Zhang, F.; Gomes, N.; Yao, Y.; Orth, P. P.; and Iadecola, T. 2021. Adaptive variational quantum eigensolvers for highly excited states. *Physical Review B*, 104(7): 1–10.
- Zhang, K.; Hsieh, M.-H.; Liu, L.; and Tao, D. 2022. Gaussian initializations help deep variational quantum circuits escape from the barren plateau. arXiv preprint arXiv:2203.09376.