A Parameter Initialization Method for Variational Quantum Algorithms to Mitigate Barren Plateaus Based on Transfer Learning

Huan-Yu Liu,1,2 Tai-Ping Sun,1,2 Yu-Chun Wu,1,2,3,∗ Yong-Jian Han,1,2,3 and Guo-Ping Guo1,2,3,4

1Key Laboratory of Quantum Information, Chinese Academy of Sciences, School of Physics, University of Science and Technology of China, Hefei, Anhui, 230026, P. R. China
2CAS Center For Excellence in Quantum Information and Quantum Physics, University of Science and Technology of China, Hefei, Anhui, 230026, P. R. China
3Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, Anhui, 230088, P. R. China
4Origin Quantum Computing, Hefei, Anhui, 230026, P. R. China

(Dated: December 22, 2021)

Variational quantum algorithms (VQAs) are widely applied in the noisy intermediate-scale quantum era and are expected to demonstrate quantum advantage. However, training VQAs faces difficulties, one of which is the so-called barren plateaus (BP) phenomenon, where gradients of the cost function vanish exponentially with the number of qubits. In this paper, based on the basic idea of transfer learning, where knowledge of pre-solved tasks could be further used in a different but related work with the training efficiency improved, we report a parameter initialization method to mitigate BP. In the method, the quantum neural network, as well as the optimal parameters for the task with a small size, are transferred to tasks with larger sizes. Numerical simulations show that this method outperforms random initializations and could mitigate BP as well. This work provides a reference for mitigating BP, and therefore, VQAs could be applied to more practical problems.

I. INTRODUCTION

While the realization of quantum computers that can carry out fault-tolerant quantum computations [1] will potentially take decades of research, designing algorithms executable with the noisy intermediate-scale quantum (NISQ) device [2, 3] to solve practical problems is imminent. Variational quantum algorithms (VQAs) [4–6] are promising in this aspect. The main feature of VQAs is the hybrid quantum-classical loop: A quantum processor prepares the ansatz with the quantum neural network (QNN) (It is also called parameterized quantum circuits,) and performs measurements on the corresponding Hamiltonian; then a classical processor evaluates the cost function based on the measured data and performs parameter optimizations. VQAs have been applied in various regions including solving molecular systems in quantum chemistry [6–9], scientific computation [10–12], machine learning [13–15], system-environment interaction simulation [16, 17], etc.

Compared to traditional quantum algorithms like quantum phase estimation [18], the requirements for the qubit number and quantum operation depth available for the quantum processor are weaker. However, an extra cost of training parameters in the QNN is included in VQAs. Recent works reported that such a training process can suffer from the barren plateaus (BP) phenomenon [19, 20], where gradients of the cost functions vanish exponentially with the number of qubits. Exponential resources would be required to obtain the gradient information as the sampling error in quantum measurements is $O(1/\sqrt{N})$ for $N$ shots. This exponential cost limits the application of VQAs to practical problems, where more qubits are needed.

An example of the cost landscape when BP establishes is shown in Fig. 1. As we can see, gradients of parameters in the BP region (red part) vanish and only parameters in the exponentially suppressed central part can be efficiently optimized. There are ways to address this difficulty. For instance, the QNN with a shallow depth will not suffer from BP if the Hamiltonian is local [20]. Also, BP is absent with the quantum convolution neural

FIG. 1. An example of graphic representation of the cost landscape when BP exists. Only parameters in the exponentially suppressed central part can be optimized efficiently. This plot is obtained from the task of optimizing parameters to maximum the fidelity between $|0\rangle^{\otimes n}$ and $(R_{\theta}(0)|0\rangle)^{\otimes n}$ with a global Hamiltonian $H = 1 - (|0\rangle\langle 0|)^{\otimes n}$ [20].
network and the tree tensor network [21]. The method to mitigate the noise-induced BP [22] was also proposed [23]. Transfer learning [24–26] is a machine learning method that tries to apply the information of pre-solved tasks to a different but related work, which are called the base and the target, respectively. It is more efficient with the application of that knowledge in some simulations [26, 27].

In this paper, based on transfer learning, we propose a parameter initialization method to mitigate BP. The task with a small size is trained as the base task. Then the QNN and its optimal parameters are transferred to target tasks with larger sizes. Numerical simulations show that this method can save resources, and more importantly, mitigate BP.

The rest of this paper is organized as follows: VQAs, the ansatz, and BP will be introduced in Sec. II. In Sec. III, we will analyze how a trainable set of initial parameters can address BP and introduce methods for the transfer of the QNN and their optimal parameters. Numerical simulations to show the performance will be in Sec. IV and a conclusion and discussion will be in Sec. V.

II. BACKGROUND

A. Variational quantum algorithms

The VQA is a hybrid quantum-classical algorithm that employs a quantum processor for ansatz preparation and expectation value measurement and a classical optimizer for parameter optimization. Generally, the ansatz is generated as $|\psi(\theta)\rangle = U(\theta)|0\rangle$ with the QNN $U(\theta)$ having the form:

$$U(\theta) = \prod_{l=1}^{L} U_l(\theta_l),$$

(1)

where $U_l(\theta_l) = \exp(-i\theta_l V_l/2) W_l$ with $V_l^2 = I$, which is usually a tensor product of Pauli operators. $W_l$ is the un-parameterized quantum gate and $\theta = \{\theta_j\}_{j=1}^{L}$ is the set of parameters to be optimized. Various types of ansatz have been proposed and we will apply the hardware-efficient ansatz [7, 28] and the Hamiltonian variational ansatz [29, 30] to the simulations in this work, which will be introduced later in this section. The cost function is usually the expectation value of the Hamiltonian $H$ corresponding to the problem:

$$C(\theta) = Tr[H U(\theta) \rho_0 U(\theta)^\dagger],$$

(2)

with $\rho_0$ the initial state. Generalized cost function can be functions of the measured expectation values, which can be efficiently evaluated with a classical computer once the measurement outcome is obtained.

The parameter optimization process is done via a classical optimizer. The optimization method can be either gradient-based or gradient-free, where the gradient of the cost function and cost difference are used to update the parameters, respectively. The desired information can be acquired once the optimization finishes. The ansatz and the optimization method will highly determine the performance of VQAs, which need to be well-designed.

B. Ansatz

The selection of the ansatz is the core part of VQAs, which will highly determine the performance. Generally, there are two types of ansatz: The problem-agnostic ansatz and the problem-inspired ansatz. The problem-agnostic ansatz is independent on the problem to be solved. Strong expressibility [31–34] is always needed as it is a direct search within the representation region of the ansatz. This large searching area could bring difficulties with the parameter optimization process. Therefore, the application of the problem-agnostic ansatz should be well-concerned. One widely used example is the hardware-efficient ansatz.

In contrast, the problem-inspired ansatz tries to include the known information of the problem. It encodes the information into the ansatz and constrains the searching area into a smaller region. Examples of such ansatz include the unitary coupled-cluster ansatz [35, 36] in quantum chemistry and the Hamiltonian variational ansatz. Initial states for these ansatzes should be designed, which may not be all-zero-initializations. Below we will introduce the ansatz used in this paper.

a. Hardware-efficient ansatz (HEA) As decomposing multi-qubit unitary operations into elementary gates will lead to exponentially circuit depth, which is not achievable with the NISQ device, ansatz that directly uses the easy-implementable gates was proposed. Generally, the HEA is comprised of many blocks, each of which consists of the single- and two-qubit gates easy-implementable with the hardware. The circuit of the HEA used in this work is:

$$U(\theta) = \prod_{k} R(\theta_k) U_{ent},$$

(3)

where $R(\theta_k)$ and $U_{ent}$ are the single-qubit rotation and two-qubit entangling operation parts, respectively. The gate sequences of these operations are:

$$R(\theta_k) = \prod_{i=1}^{n} R^a_x(\theta_{ki1}) R^a_z(\theta_{ki2}) R^x_z(\theta_{ki3}),$$

$$U_{ent} = \prod_{i=1}^{n-1} CZ_{i,i+1}.$$  

(4)

Here, $n$ refers to the number of qubits. $R^a_x(\beta), (a = x, z)$ is the rotation gate with a angle $\beta$ in the $a$-axis acting on the $i$-th qubit and $CZ = diag\{1,1,1,-1\}$ is the Control-Z gate. This ansatz is independent on any physical information of the problem and its expressibility has
been shown to overpass the Boltzmann machine and tensor networks [31]. However, as this can lead to BP [21], parameters optimization should be concerned. The quantum circuit for the 4-qubit HEA is shown in Fig. 2(a).

b. Hamiltonian variational ansatz (HVA) The HVA originates from quantum adiabatic computing, which uses the operators in the Hamiltonian to evolve the ansatz. Specifically, for the Hamiltonian represented as a linear combination of unitary operators:

$$H = \sum_l H_l,$$

the HVA has the form:

$$U(\theta) = \prod_k \prod_l R(H_l) = \prod_k \left( \prod_l e^{-i\theta_{k,l} H_l} \right). \quad (5)$$

The initial state should have a non-zero overlap with the ground state (we can prepare it as the ground state of some $H_l$). Similar to the quantum alternating operator ansatz (QAOA), the performance depends on its number of layers [37]. The performance is also related to the interaction distance to the target state [38]. The quantum circuit for the 4-qubit HVA is shown in Fig. 2(b).

c. Barren plateaus and cost concentration

The gradient $\nabla C$ is always needed in the parameter optimization process via the gradient-based method, which is a vector of partial derivatives of the cost function at the parameters:

$$\nabla C = \{ \partial_j C \}_{j=1}^L, \quad \partial_j C = \frac{\partial C}{\partial \theta_j}. \quad (6)$$

Barren plateaus refers to the situation that the gradient vanishes exponentially with the number of qubits. A formal statement is:

**Lemma 1 (Barren plateaus).** Consider the ansatz and cost function introduced in Eq. (1) and Eq. (2). Denote:

$$U = U_L U_R,$$

$$U_L = \prod_{l<j} U_l, \quad (7)$$

$$U_R = \prod_{l>j} U_l. \quad (8)$$

Then we have:

- The average of the partial derivative at any parameter over the ansatz is 0, i.e., $\langle \partial_j C \rangle = 0$.
- The variance of the above partial derivative vanishes exponentially with the number of qubits, i.e., $\text{Var} [\partial_j C] \in O(p^{-n}), p > 1$, when either $U_L$ or $U_R$ is unitary 2-design.

According to Chebyshev's inequality:

$$P(|\partial_j C| \geq c) \leq \frac{\text{Var}[\partial_j C]}{c^2}. \quad (10)$$
When the variance of the partial derivative establishes an exponential decay, the probability of finding a $|\partial C|$ that is bigger than $c$ also decreases exponentially. This indicates that one needs exponential resources to sample the gradient. Therefore, optimizing parameters in the BP region with the gradient-based method is difficult.

A circuit forms unitary 2-design [39, 40] means that average of the function of $f(U)$ up to the second-order (like $Tr[UAU^\dagger B]Tr[UCU^\dagger D]$, which will be used in the analysis of the variance of $\partial C$) over such circuit is indistinguishable from average of that function over the unitary group via the Haar measure. While it is shown that the random circuit forms an approximate 2-design [41], it indicates that the HEA, which is a non-structured circuit, will establish BP. The expressibility [32–34] of quantum circuits is also defined based on the "distance" from the circuit to unitary 2-design. This shows that strong expressibility can establish BP. And there are tests [42] showing that methods to reduce the expressibility of the circuit, like restricting the freedom of parameters and quantum operations, can reduce the speed of the gradient decay.

From the above analysis with Chebyshev’s inequality, BP also establishes a cost concentration [43], where the cost function is exponentially concentrated around the mean value. This can be derived using the variance of the cost difference similar to the gradients in Lemma 1:

$$Var[C(\theta) - \langle C \rangle] \in O(b^{-n}), b > 1. \quad (11)$$

This means that when there establishes BP, the probability of finding parameters whose cost function is lower than the average value at a constant $c$ is also exponentially suppressed. This exponential suppression of the cost difference makes the parameter optimization also difficult with gradient-free methods [44], where the parameter update is based on the sampled cost difference. It has been shown that BP is equivalent to the cost concentration.

### III. METHODS

#### A. Transfer-learning-based parameter initialization

While the cost concentration describes a flattened cost landscape where parameters in exponentially suppressed areas can be efficiently optimized, it also indicates that we can get away from the BP region with a high probability if we were able to start training, making the cost function decrease about $O(1)$. Therefore, a good set of initial parameters, whose gradients are larger than the suppressed ones, would be significant to mitigate BP.

The work [45] provides an initial strategy for the QNN. They randomly choose some of the parameters and adjust the others such that the circuit becomes a sequence of shallow blocks, each of which compiles to an Identity. This will reduce the number of effective layers and break the unitary 2-design structure, which makes it able to start that training.

Based on the definition of transfer learning, we propose a parameter initialization method to mitigate BP. The task with a small size is trained as the base network. Then we will try to apply the information from the base task to target tasks with larger sizes. Simulations with the transfer-based initial parameters and randomly initial parameters are called hot start and cold start, respectively. It is expected that hot start parameters can save resources compared to cold start ones, and more importantly, mitigate BP.

#### B. Transfer of the information

In this part, we will introduce how the information from the base task is transferred to target tasks. Two methods for the transfer of the QNN are given. The first is the network transfer method, which directly copies the base network. The second is the structure transfer method, which keeps the structure of the QNN unchanged from the base to the target. In all cases, optimal parameters in the base network will be a whole for transfer.

##### a. Network transfer

The network transfer method generates the ansatz by simply copying the base network. Specifically, suppose an $n$-qubit base network $U(\theta)_{i,n}$ is trained with optimal parameters $\theta^*$. Where subscripts $p, q$ in $U(\theta)_{p,q}$ means that the circuit acts on the $p$-th qubit to the $q$-th one. We can obtain the target network $U(\theta)$ by arranging the base network in an alternating layer if we were to solve the target task with $m(m > n)$ qubits:

$$U(\theta) = \prod_{i=1}^{m-n+1} U(\theta)_{i,i+n-1}, \quad (12)$$

where each $\theta_i$ can be either $\theta^*$ or randomly initialized $\theta_i$, which is represented with T(transfer) and R(random). Finally, the parameter initialization will be described by an $(m - n + 1)$-bit binary-valued string like "TTR1...". One can design different arrangements of the base networks for different tasks and we just provide the simple form here. Note that this method is only suitable for the problem-agnostic ansatz. While for the problem-inspired ansatz, this alternating arrangement does not have any physical meanings at all. The quantum circuit for the network transfer is shown in Fig. 2(c).

##### b. Structure transfer

Different from the network transfer method, the structure transfer method keeps the structure unchanged in the transfer process. The whole parameters are divided into several parts to be transferred, which will be described by a binary-valued string comprised of T and R similar to the network transfer method. An example of structure transfer of HEA is shown in Fig. 2(d). As it preserves the structure of
the ansatz, both problem-inspired and problem-agnostic ansatz can be transferred with this method.

IV. NUMERICAL SIMULATIONS

To specify whether hot start parameters can help to save resources compared to cold start ones, and more importantly, mitigate BP, numerical simulations will be performed in this section. We consider the task of solving the ground state of some models under the VQE framework. The introduced ansatz and transfer methods will be applied. The parameter optimization is done via the BFGS [46] algorithm, which is a gradient-based method.

Both the base and target networks will be trained to obtain a totally of 100 successful runs, where a successful run means that the optimal ground state energy is within chemical accuracy \((1.6 \times 10^{-3} \text{ Hartree})\) compared to the exact value. In all models, the simulation starts with a training of the base network with randomly initialized parameters. meanwhile recording the optimal parameters in each successful run. These optimal parameters will form a pool and will be randomly picked out for simulations of target tasks. The target network with different parameter initialization methods will be simulated. Several indicators are evaluated to compare their performance:

• Average number of iterations in the parameter optimization process with the gradient-based method over the 100 successful runs. This is a direct comparison between the hot and cold start simulations, where a smaller number indicates that the set of initial parameters is better than those with more iteration numbers.

• Number of tries to achieve 100 successful runs. This quantifies the successful probability for different parameter initialization cases. A smaller number of tries also indicates a better parameter initialization. This can be affected by local minimum points besides BP.

• The sampled variance of \(\partial_0 C\) with different initial parameters, where 0 refers to the first effective parameter. (In HVA, it is \(\theta_{1,1}\). While in HEA, the first \(R_x\) gate acting on the state \(|0\rangle\) is useless and we will compute the information for the parameter corresponding to the first \(R_x\) gate.) As described in Lemma. 1, the variance determines the scaling of the sampled gradient. Then a larger sampled variance is preferred.

• The sampled average value of the normalized gradient norm \(G = \frac{||\nabla C||_2}{L} = \frac{1}{L} \sqrt{\sum_{j=1}^{L} (\partial_j C)^2}\) for different initial parameters, As only the variance of \(\partial_0 C\) is insufficient to acquire the whole scaling of the gradient, we computed the gradient norm as well, which is a direct comparison for the resources required to sample the gradient between different parameter initialization methods. And similar to the variance, a larger sampled norm indicates a better parameter initialization method. Both the variance and the gradient norm will be sampled 500 times.

Note that whether resources are saved is based on the comparison between the number of iteration numbers and tries for different parameter initializations. However, when simulating hot start parameters, the cost of training base networks to obtain the optimal parameters for transfer should be considered. Here, we will show that the cost can be neglected, which will not affect our result. We consider the cost of evaluating one step of gradient via the parameter-shift rule [47] or the finite difference method:

\[
t_{\text{per-step}} = 2 \times t_n \times L_n, \tag{13}
\]

where \(t_n\) and \(L_n\) are the time to evaluate the cost function and number of parameters when the qubit number is \(n\). In classical simulation, as the simulation cost increases exponentially with the number of qubits, then \(t_{\text{per-step}} \in O(2^n)\). In quantum simulations, we take the time to run a circuit simply as the number of layers. Then it is easy to see that \(t_{\text{per-step}} \in O(n^2)\) for HVA and \(t_{\text{per-step}} \in O(n^3)\) for HEA, where the number of layers is set to be \(O(n)\). It is also apparent that, under the same ansatz and question, the number of iterations also increases with the number of parameters. Combining these points with the simulation results, it is reasonable that the cost for training the base network can be neglected when comparing the parameter initialization methods.

A. Transverse field Ising model, HEA, and network transfer

First, we consider the transverse field Ising model, whose Hamiltonian is given as:

\[
H = -J_{\text{ising}} \sum_{\{i,j\}} Z_i Z_j - h \sum_i X_i, \tag{14}
\]

where \(P_j (P = X/Y/Z)\) is the Pauli operator on the \(j\)-th qubit. \(\{i, j\}\) represents the nearest iteration pairs. \(J\) and \(h\) models the spin-coupling strength in the \(z\)-axis and the transverse field along the \(x\)-axis, respectively. We perform numerical simulations on the one-dimensional model and the model parameters are set to be \(J_{\text{ising}} = 1\) and \(h = 2\). The base task is to solve a 4-qubit model and the base network is a 4-layer HEA. For target tasks, we consider solving 6- and 8-qubit models, respectively. The network transfer method will be applied here. In TABLE I, we list the number of iterations and number of tries for the two different target tasks. Hot start parameters provide better performances compared to cold start ones.
FIG. 3. Sampled variance of $\partial_0 C$ and normalized gradient norm $G = \frac{1}{L} \sqrt{\sum_{j=1}^{L} (\partial_j C)^2}$ in the simulations. The top and bottom three figures represent the sampled variance and gradient norm, respectively. The hot start parameters mitigated the speed of the decay of the gradient norm with the number of qubits, indicating that the transfer-learning-based initial parameters can mitigate BP. The variance of $\partial_0 C$ with the “TT” method in the Hydrogen chain and the XXZ model does not show a tendency of the decay of the variance, meaning that we can start training in this direction.

The sampled variance of $\partial_0 C$ with randomly initial parameters is shown in Fig. 3(a), where the gradient does not always decay. The sampled gradient norm is shown in Fig. 3(d), which shows that there establishes a decay of the gradient norm and hot start parameters have successfully mitigated the speed of that decay. We can see that only the variance is insufficient to diagnose BP. The result indicates that hot start parameters have successfully mitigated BP.

B. Hydrogen chain, HEA and structure transfer

Here we consider the molecular system in quantum chemistry. The second-quantized Hamiltonian is:

$$H = \sum_{pq} h_{pq} a_p^\dagger a_q + \sum_{pqrs} V_{pqrs} a_p^\dagger a_q^\dagger a_r a_s,$$

where $a_p^\dagger (a_p)$ represents the creation (annihilation) operators, $h_{pq}$ and $V_{pqrs}$ are one- and two-electron integrals. We consider the Hydrogen atoms in a linear structure with atom-atom distance at 0.74Å. The STO-3G basis and the Jordan-Wigner transformation [48] is applied, which will result in a $2m$-qubit Hamiltonian for $m$ atoms. Finally, the spin Hamiltonian is expressed as a linear combination of Pauli operators: $H = \sum_I c_I P_I$.

| Initial methods | Average iterations | Number of tries |
|-----------------|--------------------|----------------|
| TTTT            | 105                | 101            |
| RRT             | 104                | 105            |
| TTR             | 123                | 113            |
| RRR             | 125                | 115            |

(a) Transfer learning test from 4-qubit to 6-qubit

(b) Transfer learning test from 4-qubit to 8-qubit

We use ChemiQ [49] to generate the spin Hamiltonian. The base task is to solve $H_2$. We use a 4-qubit 4-layer HEA as the base network. The target task is to solve
$H_3$. We will use a 6-qubit 8-layer circuit as the target network. The top 4 qubits are transferred.

In TABLE II we give the number of iterations and tries. We can see that hot start parameters have reduced the average number of iterations. And the number of tries for hot start parameters is nearly half of the cold start one. Hot start parameters perform better.

We plot the sampled variance of $\partial_0 C$ and the gradient norm in Fig. 3(b) and Fig. 3(e). Results show there establishes a decay for both the variance and the gradient norm for cold start parameters and hot start parameters have successfully mitigated that speed. We observe that the variance of $\partial_0 C$ with the "TT" method does not show a tendency of decay. This indicates that we can start training with this direction and make the cost function decrease such that we can away from the BP region as analyzed in Sec. III A.

C. Heisenberg XXZ model

Then we consider the Heisenberg XXZ model, whose Hamiltonian is:

$$H = -J_{XXZ} \sum_{\{ij\}} X_i X_j + Y_i Y_j + \Delta Z_i Z_j,$$

where $J_{XXZ}$ represents the in-plane spin coupling strength and $\Delta$ models the dimensionless anisotropy factor between the $xy$ plane and $z$ axis coupling strength.

In the numerical simulation, we will set the model parameters as $J_{XXZ} = 1$ and $\Delta = 2$. The base task is a 4-qubit model in one dimension. For target tasks, we consider the one-dimensional model and two-dimensional model, respectively.

For the one-dimensional target model, We will consider the HVA and use the structure transfer method. The Hamiltonian can be divided into three parts as $H = H_X + H_Y + H_Z$, which will result in three parameters per layer. The number of layers for the base network and the target network will be 4 and 8, respectively. In the target task of the two-dimensional model with size $2 \times 4$, we will use HEA and structure transfer. The number of layers is the same as above.

Similar to the former simulations, we give the number of iterations and tries in TABLE III. The hot start parameters outperform the cold start ones. For the two-dimensional case, even though the cold start parameters perform well, the hot start ones perform better.

The sampled variance of $\partial_0 C$ and gradient norm is shown in Fig. 3(c) and Fig. 3(f). We can see that hot start parameters have successfully mitigated BP. And similar to the above model, the variance with the "TT" method does not show a tendency of decay.

V. CONCLUSION AND DISCUSSION

In this paper, based on transfer learning, we proposed a parameter initialization method for VQAs to mitigate BP. The task with a small size is trained as the base network, then the information is transferred to the target tasks with larger sizes. Numerical simulations showed that this method could save resources compared to randomly initial parameters and mitigate BP. Therefore, VQAs might be applied to more practical problems with this method.

Due to the limitation of our computational resource, tasks requiring more qubits are not simulated here. The variance of $\partial_0 C$ and the gradient norm in the simulations showed that this method can mitigate BP. Therefore, the trainability of VQAs is improved, which is our main focus in this work. However, whether training the QNN can reach the target is not still straightforward, where trainability is just one factor. There would be more efficient transfer methods depending on the problem, like different arrangements of the base network and parameter initializations, which will be studied further.
ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (Grant No. 2016YFA0301700) and the Anhui Initiative in Quantum Information Technologies (Grant No. AHY080000). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of University of Science and Technology of China.

[1] D. Gottesman, Theory of fault-tolerant quantum computation, Phys. Rev. A 57, 127 (1998).
[2] J. Preskill, Quantum computing in the nisq era and beyond, Quantum 2, 79 (2018).
[3] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heinonen, J. S. Kottmann, T. Menke, W.-K. Mok, S. Sim, L.-C. Kwek, and A. Aspuru-Guzik, Noisy intermediate-scale quantum (nisq) algorithms, (2021), arXiv:2101.08448 [quant-ph].
[4] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, et al., Variational quantum algorithms, Nature Reviews Physics, 1 (2021).
[5] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, The theory of variational hybrid quantum-classical algorithms, New Journal of Physics 18, 023023 (2016).
[6] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O’Brien, A variational eigenvalue solver on a photonic quantum processor, Nature Communications 5, 4213 (2014).
[7] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets, Nature 549, 242 (2017).
[8] K. M. Nakanishi, K. Mitarai, and K. Fujii, Subspace-search variational quantum eigensolver for excited states, Phys. Rev. Research 1, 033062 (2019).
[9] K. Seki, T. Shirakawa, and S. Yunoki, Symmetry-adapted variational quantum eigensolver, Phys. Rev. A 101, 052440 (2020).
[10] X. Xu, J. Sun, S. Endo, Y. Li, S. C. Benjamin, and X. Yuan, Variational algorithms for linear algebra, Science Bulletin 66, 2181 (2021).
[11] M. Lubasch, J. Joo, P. Moinier, M. Kiffner, and D. Jaksch, Variational quantum algorithms for nonlinear problems, Phys. Rev. A 101, 010301 (2020).
[12] H.-L. Liu, Y.-S. Wu, L.-C. Wan, S.-J. Pan, S.-J. Qin, F. Gao, and Q.-Y. Wen, Variational quantum algorithm for the poisson equation, Phys. Rev. A 104, 022418 (2021).
[13] M. Schuld and N. Killoran, Quantum machine learning in feature hilbert spaces, Phys. Rev. Lett. 122, 040504 (2019).
[14] A. Blance and M. Spannowsky, Quantum machine learning for particle physics using a variational quantum classifier, Journal of High Energy Physics 2021, 1 (2021).
[15] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, Supervised learning with quantum-enhanced feature spaces, Nature 567, 209 (2019).
[16] H.-Y. Liu, T.-P. Sun, Y.-C. Wu, and G.-P. Guo, Variational quantum algorithms for the steady states of open quantum systems, Chinese Physics Letters 38, 080301 (2021).
[17] M. Mahdian and H. D. Yeganeh, Hybrid quantum variational algorithm for simulating open quantum systems with near-term devices, Journal of Physics A: Mathematical and Theoretical 53, 415301 (2020).
[18] U. Dorner, R. Demkowicz-Dobrzański, B. J. Smith, J. S. Lundeen, W. Wasilewski, K. Banaszek, and I. A. Walmsley, Optimal quantum phase estimation, Phys. Rev. Lett. 102, 040403 (2009).
[19] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nature Communications 9, 4812 (2018).
[20] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Cost function dependent barren plateaus in shallow parametrized quantum circuits, Nature Communications 12, 1791 (2021).
[21] C. Zhao and X.-S. Gao, Analyzing the barren plateau phenomenon in training quantum neural networks with the ZX-calculus, Quantum 5, 466 (2021).
[22] S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, and P. J. Coles, Noise-induced barren plateaus in variational quantum algorithms, Nature Communications 12, 6961 (2021).
[23] A. Wu, G. Li, Y. Ding, and Y. Xie, Mitigating noise-induced gradient vanishing in variational quantum algorithm training (2021), arXiv:2111.13209 [quant-ph].
[24] S. J. Pan and Q. Yang, A survey on transfer learning, IEEE Transactions on knowledge and data engineering 22, 1345 (2009).
[25] A. Mari, T. R. Bromley, J. Izaac, M. Schuld, and N. Killoran, Transfer learning in hybrid classical-quantum neural networks, Quantum 4, 340 (2020).
[26] R. Zen, L. My, R. Tan, F. Hébert, M. Gattobigio, C. Miniatura, D. Poletti, and S. Bressan, Transfer learning for scalability of neural-network quantum states, Phys. Rev. E 101, 053301 (2020).
[27] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, Self-taught learning: Transfer learning from unlabeled data, in Proceedings of the 24th International Conference on Machine Learning, ICML '07 (Association for Computing Machinery, New York, NY, USA, 2007) p. 759–766.
[28] H. L. Tang, V. Shkolnikov, G. S. Barron, H. R. Grimsmo, N. J. Mayhall, E. Barnes, and S. E. Economou, Qubit-adapt-vqe: An adaptive algorithm for constructing hardware-efficient ansätze on a quantum processor, PRX Quantum 2, 020310 (2021).
[29] R. Wiersma, C. Zhou, Y. de Sereville, J. F. Carrasquilla, Y. B. Kim, and H. Yuen, Exploring entanglement and optimization within the hamiltonian variational ansatz,
E. Dagotto and A. Moreo, Improved hamiltonian variational technique for lattice models, Phys. Rev. D 31, 865 (1985).

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

K. Nakaji and N. Yamamoto, Expressibility of the alternating layered ansatz for quantum computation, Quantum 5, 434 (2021).

S. Sim, P. D. Johnson, and A. Aspuru-Guzik, Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms, Advanced Quantum Technologies 2, 1900070 (2019).

Y. Du, Z. Tu, X. Yuan, and D. Tao, An efficient measure for the expressivity of variational quantum algorithms, (2021), arXiv:2104.09961 [quant-ph].

Y. Shen, X. Zhang, S. Zhang, J.-N. Zhang, M.-H. Yung, and K. Kim, Quantum implementation of the unitary coupled cluster for simulating molecular electronic structure, Physical Review A 95, 020501 (2017).

J. Lee, W. J. Huggins, M. Head-Gordon, and K. B. Whaley, Generalized unitary coupled cluster wave functions for quantum computation, Journal of Chemical Theory and Computation 15, 311 (2019).

V. Akshay, H. Philathong, M. E. S. Morales, and J. D. Biamonte, Reachability deficits in quantum approximate optimization, Phys. Rev. Lett. 124, 090504 (2020).

G. Mates, S. Johri, and Z. Papić, Quantifying the efficiency of state preparation via quantum variational eigensolvers, PRX Quantum 2, 010309 (2021).

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

H. Zhu, Multiqubit clifford groups are unitary 3-designs, Phys. Rev. A 96, 062336 (2017).

A. W. Harrow and R. A. Low, Random quantum circuits are approximate 2-designs, Communications in Mathematical Physics 291, 257 (2009).

Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, (2021), arXiv:2101.02138 [quant-ph].

A. Arrasmith, Z. Holmes, M. Cerezo, and P. J. Coles, Equivalence of quantum barren plateaus to cost concentration and narrow gorges, (2021), arXiv:2104.05868 [quant-ph].

E. Grant, L. Wossnig, M. Ostaszewski, and M. Benedetti, An initialization strategy for addressing barren plateaus in parametrized quantum circuits, Quantum 5, 558 (2021).

R. Battiti and F. Masulli, Bfgs optimization for faster and automated supervised learning, in International neural network conference (Springer, 1990) pp. 757–760.

Q.-S. Li, H.-Y. Liu, Q. Wang, Y.-C. Wu, and G.-P. Guo, A unified framework of transformations based on jordan-wigner transformation, (2021), arXiv:2108.01725 [quant-ph].