Ansatz-Agnostic Exponential Resource Saving in Variational Quantum Algorithms Using Shallow Shadows

Afrad Basheer[1] , **Yuan Feng**[2] , **Christopher Ferrie**[1] and **Sanjiang Li**[1]
[1]Centre for Quantum Software and Information, University of Technology Sydney, NSW 2007, Australia
[2]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
{Sanjiang.Li, Christopher.Ferrie}@uts.edu.au, Afrad.M.Basheer@student.uts.edu.au,
yuan_feng@tsinghua.edu.cn

Abstract

Variational Quantum Algorithms (VQA) have been identified as a promising candidate for the demonstration of near-term quantum advantage in solving optimization tasks in chemical simulation, quantum information, and machine learning. The standard training model requires a significant amount of quantum resources, which led researchers to explore classical shadows as an alternative that consumes exponentially fewer quantum resources. However, the existing approach only works when dealing with local observables and shallow Alternating Layered Ansatz (ALA), thus severely limiting its potential in solving problems such as quantum state preparation, where the ideal state might not be approximable with an ALA. In this work, we present a protocol based on shallow shadows that achieves similar levels of savings for almost any shallow ansatz studied in the literature, when combined with observables of low Frobenius norm. We show that two important applications in quantum information for which VQAs can be a powerful option, namely variational quantum state preparation and variational quantum circuit synthesis, are compatible with our protocol. We also experimentally demonstrate orders of magnitude improvement in comparison to the standard VQA model.

1 Introduction

The field of quantum computing has witnessed significant progress over the past decade. In the current Noisy Intermediate Scale Quantum (NISQ) [Preskill, 2018] era, quantum devices are small and error-prone. Despite this, several research groups have successfully demonstrated quantum advantage over classical computers in synthetic but well-defined sampling problems [Arute *et al.*, 2019; Zhong *et al.*, 2020; Madsen *et al.*, 2022]. The next major breakthrough in this field will involve extending these advantages to solve practically valuable problems.

Many proposals have been put forward, and one that stands out is Variational Quantum Algorithms (VQAs) [McClean *et al.*, 2016]. These algorithms are specifically designed to solve optimization problems involving quantum information, stored as quantum states using quantum bits (i.e., qubits), manipulated through quantum circuits, and measured using quantum observables. Many important functions involving these objects, such as the expectation values of observables with respect to different states, are notoriously hard or intractable to evaluate on classical computers, as the required classical computational resources increases exponentially in the number of qubits involved. By using parameterized quantum circuits, such functions can be estimated with polynomially many quantum resources on quantum devices, thereby enabling optimization using iterative optimization algorithms. Potentially useful applications include Variational Quantum Eigensolver [Peruzzo *et al.*, 2014], Quantum Support Vector Machines [Havlíček *et al.*, 2019], Quantum Approximate Optimization Algorithm [Farhi *et al.*, 2014], etc.

Unlike classical computing, the absence of quantum memory devices, combined with the no-cloning theorem, implies that each use of a quantum state necessitates preparing it from scratch. In the context of VQAs, we refer to the term *sample complexity* to denote the total number of executions of the quantum preparation device required (equivalently, the total number of copies of quantum states consumed). In the standard VQA model, this scales linearly with the total number of function evaluations needed throughout the optimization. When additional factors such as hyperparameter tuning, model and ansatz selection are introduced, the scale of this number becomes notably significant. Moreover, in the near term, only very few capable quantum computers would be available, making the implementation of VQAs with reduced sample complexity crucial.

Classical shadow tomography [Huang *et al.*, 2020] provides an exponentially improved method to estimate linear functionals that involve quantum states. This advancement has been incorporated into the VQA training protocols, leading to an exponential reduction in quantum resources, as demonstrated in [Basheer *et al.*, 2023]. However, the method, titled *Alternating Layered Shadow Optimization* (ALSO), uses a version of shadow tomography that requires local target observables. This constraint restricts the ansatzes to require simple entanglement structures, such as the Alternating Layered Ansatz (ALA) illustrated in Figure 1(a). This limitation becomes significant when the optimal circuit or state cannot be approximated with ALAs.

The recently proposed shallow shadow technique [Bertoni

et al., 2023] describes a tomography procedure similar to classical shadow tomography but designed for easy implementation in NISQ devices. Even better, it does not rely directly on the locality of the observables. Building upon this, we introduce *Ansatz Independent Shadow Optimization* (AISO) in this work – a method that achieves an exponential reduction in quantum resources for VQA training. AISO is compatible with almost any shallow (depth logarithmic in the number of qubits) quantum circuit structure found in the literature when used in conjunction with observables of low Frobenius norm. We demonstrate these resource savings for two important problems in quantum information where VQAs are applicable: Variational Quantum State Preparation (VQSP) and Variational Quantum Circuit Synthesis (VQCS). Both problems involve determining the optimal circuit parameters for an ansatz that best approximates unknown quantum states or circuits.

The benefits of AISO can be summarized as follows:

1. *Exponential reduction in input state copies:* AISO achieves arbitrarily precise estimates of all function evaluations encountered during iterative optimization of the VQA cost function while consuming exponentially fewer copies of the input state compared to standard VQA. This enables more iterations, better approximations, and facilitates extensive hyperparameter tuning.
2. *Ansatz-agnostic implementation on quantum hardware:* Our method ensures a reduction in input state copies for almost any shallow ansatz studied in the literature. Additionally, the operations executed on the quantum device remain independent of the chosen ansatz.
3. *Optimization with different ansatzes:* The combination of the above two advantages implies that, for a given unknown input state or circuit, optimization can be performed over various types of ansatzes, using the same shallow shadows generated in an ansatz agnostic manner, unlike standard methods where multiple ansatzes must be implemented individually. This flexibility allows one to choose the most suitable ansatz with substantial savings in the utilization of quantum devices.
4. *Compatibility with VQCS:* Solving VQCS requires the utilization of maximally entangled states. Due to the requirement of ansatzes with limited entanglement for ALSO, it is not suitable for efficiently implementing VQCS. In contrast, AISO is ansatz independent, allowing its effective use in VQCS.

The advantage is experimentally demonstrated in both use cases of interest, where we show that AISO significantly outperforms standard VQA with the same number of copies across four different ansatzes: Alternating Layered Ansatz [Cerezo *et al.*, 2021], Multi-Entanglement Renormalization Ansatz (MERA) [Bridgeman *et al.*, 2015], Hardware Efficient Ansatz (HEA) [Leone *et al.*, 2022], and Tree Tensor Networks (TTN) [Shi *et al.*, 2006] (cf. Figure 1).

We also establish that the sample complexity of AISO, and consequently shallow shadows, can be enhanced when the input state being sampled is from a 2-design instead of a 1-design [Harrow and Low, 2009]. Finally, we discuss how AISO aligns with many heuristic methods commonly used to tackle trainability issues, such as barren plateaus, that may arise during optimization.

2 Related Works

Classical shadows have been used to improve the sample complexity of VQAs in [Basheer *et al.*, 2023]. This method, termed ALSO, utilizes a form of shadow tomography relying on local target observables and assumes that the ansatz has a weak entanglement structure, such as an ALA. Consequently, for applications such as VQSP, results can be poor if the optimal state is not approximable by ALAs. Additionally, this method is unsuitable for VQCS since it necessitates working with the maximally entangled state. In contrast, our method AISO uses shallow shadows, allowing it to be used with almost any shallow ansatz studied in the literature, addressing both VQSP and VQCS.

In [Schreiber *et al.*, 2022], classical shadows have been employed to reduce the number of quantum computer calls (i.e., sample complexity) in quantum machine learning applications. Given an already learned VQA model, the approach uses a quantum computer to generate classical shadows so that predictions can be made of the learned model using a classical computer. It is important to note that, in this approach, the learning procedure is still carried out on a quantum computer. In contrast, in AISO, the entire learning procedure takes place on a classical computer.

New classical optimization algorithms, as introduced in [Boyd and Koczor, 2022; Wierichs *et al.*, 2020; Stokes *et al.*, 2020], demonstrate faster convergence rates with significantly fewer iterations than traditional gradient descent. For applications involving shallow circuits and low Frobenius norm observables, AISO, being agnostic towards the choice of classical optimizer, can enhance the performance of these methods by significantly reducing the required number of state copies.

A new related work appears after the completion of ours. In [Cerezo *et al.*, 2023], the authors conjecture that VQA models that can avoid barren plateaus are also classically simulable (with quantum experiments polynomial in the number of qubits). Strong evidence is also provided to support their conjecture. In their terms, our approach actually shows that VQA problems with shallow ansatz and low Frobenius norm observables are also classically simulable, but it is still unclear if these models are barren plateau-free.

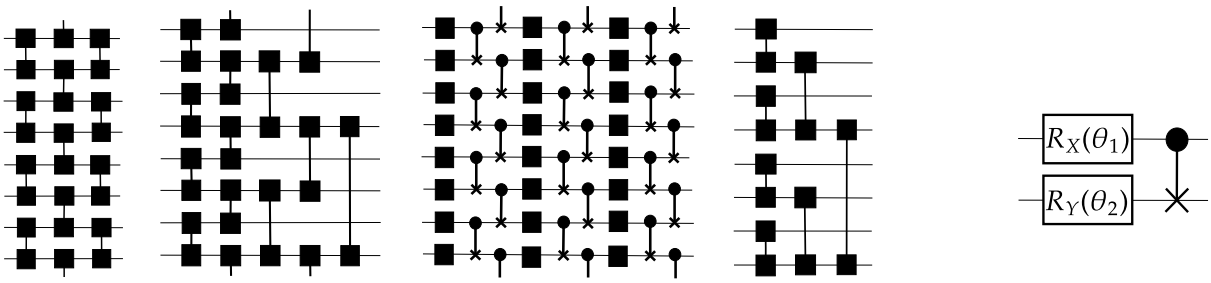
3 Background

In this section, we review quantum computing, shallow shadows, and VQAs.

3.1 Quantum Computing

Throughout this work, we use the 'ket' and 'bra' notations to denote column vectors $|\psi\rangle$ and their conjugate transposes $\langle\psi|$ respectively. $|i\rangle \in \mathbb{C}^d$ is the i^{th} standard basis vector. We use $\mathcal{L}(\mathbb{C}^d)$ to denote the set of all linear operators that act on $\mathbb{C}^d$, and for any operator A, $A^\dagger$ is its conjugate transpose.

A quantum *state* is defined as a positive semidefinite operator $\rho \in \mathcal{L}(\mathbb{C}^d)$ with $\text{tr}(\rho) = 1$. In quantum computing, a



(a) ALA (b) MERA (c) HEA (d) TTN (e) Subcircuit used in our experiments.

Figure 1: Ansatzes used in our simulations. In (a), (b), (d), each connected pair of black boxes represent a two-qubit subcircuit. In (c), each black box is a single qubit subcircuit while the two-qubit gate is the CNOT gate.

qubit is the analog of a bit in classical computing and can admit any quantum state in $\mathcal{L}(\mathbb{C}^2)$ as its value. The state of an n-qubit system can be described using states that act on the tensor product of n two-dimensional vector spaces, denoted as $\mathbb{C}^2 \otimes \cdots \otimes \mathbb{C}^2 \cong \mathbb{C}^{2^n}$.

A *quantum gate* acting on n qubits is a unitary operator $U \in \mathcal{L}(\mathbb{C}^{2^n})$. Such a gate transforms the state of an n-qubit system from ρ to $U\rho U^\dagger$. The *Pauli gates* are defined as

$$X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \; Y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \; Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \quad (1)$$

Let $P = \{\mathbb{1}, X, Y, Z\}$. For $\gamma \in \{1, -1, i, -i\}$, we write $\mathcal{P}_n^{(\gamma)}$ for the set of all n-fold tensor products of the elements in P, with the scalar γ multiplied to them. Clearly, $\mathcal{P}_n = \mathcal{P}_n^{(1)} \cup \mathcal{P}_n^{(-1)} \cup \mathcal{P}_n^{(i)} \cup \mathcal{P}_n^{(-i)}$ forms a group under matrix multiplication. The n-qubit *Clifford gates* are those contained in the normalizer of $\mathcal{P}_n$ in the group of unitary operators acting on $\mathbb{C}^{2^n}$. A *quantum circuit* is defined as a composition of multiple quantum gates.

To extract information from a quantum system in a state ρ, we perform a measurement using an *observable*, defined as any Hermitian operator O. Let the spectral decomposition of O be $O = \sum_i \lambda_i |u_i\rangle \langle u_i|$. The measurement yields an outcome λ_i with probability $\langle u_i| \rho |u_i\rangle$. Subsequently, the post-measurement state becomes $|u_i\rangle$. In addition, the expected value of this random procedure is $\mathrm{tr}(\rho O)$, concisely written as $\langle O \rangle_\rho$. Measurements using diagonal observables are called *standard basis measurements*.

Pure states are rank one states. In this case, gate operations and measurements can be fully described by any normalized eigenvector in its support.

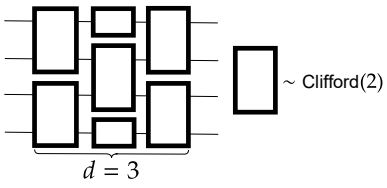


Figure 2: The structure of the unitary ensemble used to generate shallow shadows.

3.2 Shallow Shadows

For an arbitrary state ρ and known observables $O_1, O_2, \ldots, O_M$, it requires $\mathcal{O}(2^n \cdot M)$ copies of ρ to estimate $\langle O_i \rangle_\rho$ for each i using conventional quantum tomography techniques. Using classical shadow tomography [Huang *et al.*, 2020], the dependence on M can be reduced exponentially. Furthermore, for certain classes of observables, the dependence on n is $\mathcal{O}(\mathrm{poly}(n))$, and thus the total number of copies required in these cases can be reduced to $\mathcal{O}(\mathrm{poly}(n) \cdot \log M)$.

To generate a shadow, we first apply a circuit U sampled from an ensemble $\mathcal{U}$ of n-qubit circuits, and then measure the resultant state according to the standard basis to obtain an n-bit string u. A classical shadow is computed classically as

$$\hat{\rho}_{U,u} = \Delta_{\mathcal{U}}^{-1}(U^\dagger |u\rangle \langle u| U), \quad (2)$$

where

$$\Delta_{\mathcal{U}}(\rho) = \mathbb{E}_{U \sim \mathcal{U}} \sum_{u \in \{0,1\}^n} \langle u| U\rho U^\dagger |u\rangle U^\dagger |u\rangle \langle u| U. \quad (3)$$

Furthermore, $\hat{\rho}_{U,u}$ gives an unbiased estimator of ρ and hence $\langle O_i \rangle_{\hat{\rho}_{U,u}}$ is an unbiased estimator of $\langle O_i \rangle_\rho$ for all i.

The number of such shadows required for precise estimation is dominated by the *state-dependent shadow norm* of the traceless part of the observables, defined as

$$\|\widetilde{O}\|_{\rho,\mathcal{U}}^2 = \mathbb{E}_{U \sim \mathcal{U}} \sum_{u \in \{0,1\}^n} \langle u| U\rho U^\dagger |u\rangle \langle \widetilde{O} \rangle_{\hat{\rho}_{U,u}}^2, \quad (4)$$

where $\widetilde{O} = O - \frac{\mathrm{tr}(O)}{2^n}\mathbb{1}$. Using this, the sample complexity of the protocol is given by the following theorem.

Theorem 1. *[Huang* et al.*, 2020] Let $\mathcal{U}$ be an ensemble of circuits such that $\Delta_{\mathcal{U}}^{-1}$ exists, and $O_1, O_2, \ldots, O_M$ be n-qubit observables. For any $\delta, \epsilon \in (0,1)$, let $T_1 = 2\log(2M/\delta)$ and $T_2 = (34/\epsilon^2)\max_i \|\widetilde{O_i}\|_{\rho,\mathcal{U}}^2$. Let ρ be a state with classical shadows (generated using $\mathcal{U}$) $\hat{\rho}_1, \hat{\rho}_2, \ldots, \hat{\rho}_{T_1 T_2}$, where $\hat{\rho}_i = \hat{\rho}_{U_i,u_i}$. Define $\widehat{\langle O_i \rangle}_\rho = \mu_{T_1,T_2}(\{\langle O_i \rangle_{\hat{\rho}_j},\ 1 \le j \le T_1 T_2\})$, where μ_{T_1,T_2} is the median-of-means estimator (median of T_1 means of T_2 values each). Then, with probability at least $1 - \delta$, we have $|\widehat{\langle O_i \rangle}_\rho - \langle O_i \rangle_\rho| \le \epsilon$ for all i.*

Define *shadow norm* as $\|\cdot\|_{\mathcal{U}} \coloneqq \max_{\sigma:\text{state}} \|\cdot\|_{\sigma,\mathcal{U}}$. In Theorem 1, we can also replace $\|\widetilde{O_i}\|_{\rho,\mathcal{U}}$ with $\|\widetilde{O_i}\|_{\mathcal{U}}$ [Huang *et al.*, 2020], which removes the dependency on ρ and provides the worst-case sample complexity of the protocol.

In [Huang *et al.*, 2020], it was shown that when the ensemble is the Clifford group over n qubits, the shadow norm of the observables, and hence the sample complexity, are proportional to the Frobenius norm. But the implementation requires very deep circuits, ruling itself out for NISQ devices.

[Bertoni *et al.*, 2023] proposed an ensemble of shallow-depth circuits $\mathcal{U}_d$ (with depth d), given in Figure 2, that achieves similar performance guarantees. Each two-qubit subcircuit here is a uniformly randomly sampled two-qubit Clifford gate. The shadow can be classically computed and stored in the matrix product state form, with cost $\mathcal{O}(2^d)$. Formally, we have

Theorem 2. *[Bertoni* et al.*, 2023] If $d = \Theta(\log n)$, then $\|O\|^2_{\mathbb{1}/2^n,\mathcal{U}_d} \leq 4\|O\|^2_F$ for any observable O with tr$(O) = 0$, where $\|\cdot\|_F$ is the Frobenius norm.*

The term $\|O\|^2_{\mathbb{1}/2^n,\mathcal{U}_d}$ in Theorem 2 is called the *locally scrambled shadow norm*. For any ensemble $\mathcal{D}_1$ of states, if $\mathbb{E}_{\rho\sim\mathcal{D}_1}(\rho) = \mathbb{1}/2^n$ (also called *state* 1*-designs* when all states are pure), we have $\|O\|^2_{\mathbb{1}/2^n,\mathcal{U}} = \mathbb{E}_{\rho\sim\mathcal{D}_1}\|O\|^2_{\rho,\mathcal{U}}$. Hence, we can view $\|O\|_{\mathbb{1}/2^n,\mathcal{U}}$ as a quantity that intuitively characterizes the sample complexity of a shadow protocol for a "typical" state or the performance of the protocol on average, similar to how the shadow norm describes the worst-case performance. This is more apparent when all states in $\mathcal{D}_1$ are pure. Then, sampling from $\mathcal{D}_1$ is equivalent in mean to sampling uniformly (according to the spherical measure) from the set of all pure states. Throughout this work, we set $d = \Theta(\log n)$.

3.3 Variational Quantum Algorithms

Parameterized quantum circuits can be used to encode various optimization problems that one encounters in quantum information. The circuit structure used is called an *ansatz*. We use $U(\boldsymbol{\theta})$ to denote a parameterized circuit, where $\boldsymbol{\theta}$ is a vector of parameters. In standard VQA, we use $U(\boldsymbol{\theta})$ to estimate the value of a target function and then optimize the parameters by feeding the output to a classical iterative optimizer.

For any ansatz U, we define $\rho(\boldsymbol{\theta}) \coloneqq U(\boldsymbol{\theta})\rho U(\boldsymbol{\theta})^\dagger$. Our focus in this paper is on the function defined (over $\boldsymbol{\theta}$) as

$$\langle O\rangle_{\rho(\boldsymbol{\theta})} = \text{tr}(U(\boldsymbol{\theta})\rho U(\boldsymbol{\theta})^\dagger O), \tag{5}$$

where ρ is the input quantum state and O is an output observable, and we aim to find the parameters that maximize it. One can estimate $\langle O\rangle_{\rho(\boldsymbol{\theta})}$ for any $\boldsymbol{\theta}$ by repeated measurements, after the application of $U(\boldsymbol{\theta})$ on ρ. Given this ability, the gradient of $\langle O\rangle_{\rho(\boldsymbol{\theta})}$ can also be estimated using standard methods such as finite differencing or quantum-specific approaches such as the parameter shift rule [Mitarai *et al.*, 2018]. Problems in quantum information that can be reduced to an instance of optimization of Eq. (5) include variational quantum eigensolver [Peruzzo *et al.*, 2014], quantum autoencoder [Romero *et al.*, 2017], as well as VQSP and VQCS.

4 Ansatz Independent Shadow Optimization

In this section, we explain the main idea and theoretical results behind AISO.

For any quantum circuit V and any qubit i, we define the number of times a gate touches or crosses the qubit wire as $R_{V,i}$. Formally, this is the number of 2-qubit gates being applied on any qubits j, k such that $j \leq i \leq k$. Let $R_V = \max_i R_{V,i}$. We require our ansatz U to have $R_U \in \mathcal{O}(\log n)$. Most shallow ansatzes used in the literature satisfy this. Let $\langle O\rangle_{\rho(\boldsymbol{\theta}^{(1)})}, \langle O\rangle_{\rho(\boldsymbol{\theta}^{(2)})}, \ldots, \langle O\rangle_{\rho(\boldsymbol{\theta}^{(C)})}$ be function evaluations that one encountered while optimizing Eq. (5) using an iterative optimization algorithm.

Define $W_O(\boldsymbol{\theta}) = U(\boldsymbol{\theta})^\dagger O U(\boldsymbol{\theta})$. Each function evaluation can be seen as estimating the expectation of ρ with these parameterized observables because

$$\langle O\rangle_{\rho(\boldsymbol{\theta})} = \text{tr}(U(\boldsymbol{\theta})\rho U(\boldsymbol{\theta})^\dagger O) = \langle W_O(\boldsymbol{\theta})\rangle_\rho. \tag{6}$$

Moreover, the Frobenius norm remains invariant since $\|O\|^2_F = \|VOV^\dagger\|^2_F$ for any unitary V.

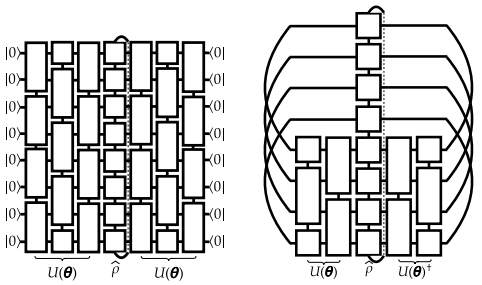
Now, using Theorems 1 and 2, we can estimate all C function evaluations using shallow shadows, and the AISO protocol goes as follows.

1. Choose precision and confidence parameters $\epsilon, \delta \in (0, 1)$. Let $m \geq 1/\delta$. Generate T_1T_2 shallow shadows of ρ, where
$$T_1 \geq 2\log\left(\frac{2(m-1)C}{m\delta - 1}\right), T_2 \geq \frac{136}{\epsilon^2} m\|O\|^2_F. \tag{7}$$
Let them be $\hat{\rho}_{U_1,u_1}, \hat{\rho}_{U_2,u_2}, \ldots, \hat{\rho}_{U_{T_1T_2},u_{T_1T_2}}$.
2. Use the iterative optimization algorithm to optimize the target function
$$\widehat{\langle W_O(\boldsymbol{\theta})\rangle}_\rho \coloneqq \mu_{T_1,T_2}(\{\langle W_O(\boldsymbol{\theta})\rangle_{\hat{\rho}_{U_j,u_j}} \; 1 \leq j \leq T_1T_2\}). \tag{8}$$

Now, we shall prove that when T_1 and T_2 satisfy Eq. (7). the AISO protocol achieves the desired precision and confidence.

Theorem 3. *Let ρ be an n-qubit pure state sampled from a state 1-design $\mathcal{D}_1$. For any $\delta, \epsilon \in (0, 1)$, $m > 1/\delta$, and any $C > 0$, let T_1 and T_2 satisfy Eq.* (7)*. Then, for any parameter vectors $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(C)}$, with probability at least $1-\delta$, we have $|\langle W_O(\boldsymbol{\theta}^{(c)})\rangle_\rho - \widehat{\langle W_O(\boldsymbol{\theta}^{(c)})\rangle}_\rho| \leq \epsilon$ for all $1 \leq c \leq C$, where $\langle W_O(\boldsymbol{\theta}^{(c)})\rangle_\rho$ and $\widehat{\langle W_O(\boldsymbol{\theta}^{(c)})\rangle}_\rho$ are defined in Eq.s* (6) *and* (8)*, respectively.*

The rationale behind AISO's ability to yield exponential savings in estimating the cost function can be intuitively grasped as follows. In standard VQA, estimating C evaluations requires preparing $U(\boldsymbol{\theta}^{(c)})$ for all c and conducting multiple measurements for each. Therefore, the total number of required copies would be $\mathcal{O}(C)$. One key limitation arises from the inability to reuse measurement results, as each measurement is conducted specifically to estimate $\langle W_O(\boldsymbol{\theta}^{(c)})\rangle_\rho$ for a particular c. In contrast, in AISO, all quantum measurements made are *independent* of $\boldsymbol{\theta}^{(c)}$, and these measurements are used when estimating all the expectations.



(a) State preparation (b) Circuit Synthesis

Figure 3: Tensor networks to compute $\langle W_O(\boldsymbol{\theta})\rangle_{\hat{\rho}}$.

Although the constants in Eq. (7) appear large, due to the use of union bounds as well as a few loose constants, in practice significantly lesser copies than what is suggested there suffice. We explore this in detail in our experimental results.

The cost of classical computation is dominated by the cost of computing $\langle\widehat{W_O(\boldsymbol{\theta})}\rangle_\rho$ classically. Thus we have the following theorem.

Theorem 4. *In AISO, for any quantum ansatz U with $R_U \in \mathcal{O}(\log n)$, $\langle\widehat{W_O(\boldsymbol{\theta})}\rangle_\rho$ can be classically evaluated with cost $\mathcal{O}(poly(n)\cdot\log C\cdot\|O\|_F^2)$ for VQSP and VQCS. The overall classical computational cost for C function evaluations is thus $\mathcal{O}(poly(n)\cdot C\log C\cdot\|O\|_F^2)$.*

Note that since the circuit is being implemented classically, R_U has no physical relevance here. The space complexity of the protocol is dominated by the storage of shallow shadows. Each shadow is an MPS with maximum bond dimension at most 2^{d-1}. This means that each shadow can be stored using at most $n2^d$ complex numbers and hence the total space complexity is at most $nT_1T_2 2^d$. So, when $d=\mathcal{O}(\log n)$, the space complexity is $\mathcal{O}(\text{poly}(n)\cdot T_1T_2)$.

5 Applications

In this section, we discuss how AISO can be used to tackle two important problems in quantum information.

5.1 Variational Quantum State Preparation

In VQSP, our goal is to find a circuit that is capable of (approximately) preparing a pure state $\rho=|\psi\rangle\langle\psi|$, given access to multiple copies of it. That is, we would like to find a parameter vector $\boldsymbol{\theta}$ that minimizes the *infidelity* between $U(\boldsymbol{\theta})^\dagger|0\rangle$ and $|\psi\rangle$, defined as $1-|\langle\psi|U(\boldsymbol{\theta})^\dagger|0\rangle|^2$, where U is a heuristically chosen ansatz. Infidelity assumes values in $[0,1]$ and is widely used in quantum information to measure how far apart two states are, with 1 implying orthogonality and 0 implying equality. Note that the minimization of infidelity is the same as the maximization of $\langle|0\rangle\langle 0|\rangle_{\rho(\boldsymbol{\theta})}$. Since $|0\rangle\langle 0|$ has unit Frobenius norm, this objective function is compatible with AISO. Moreover, with AISO, it becomes possible to explore optimal parameters for a diverse range of circuit ansatzes through multiple optimization procedures while minimizing the number of required copies.

Last but not least, for any shallow shadow $\hat{\rho}$, $\langle W_O(\boldsymbol{\theta})\rangle_{\hat{\rho}}$ can be computed classically efficiently by contracting the tensor network given in Figure 3(a). Even though the example given here is the ALA, using Theorem 4, one can easily replace it with any ansatz with $R_U\in\mathcal{O}(\log n)$. The reasoning is explained in detail in the proof of Theorem 4 in the Appendix.

5.2 Variational Quantum Circuit Synthesis

VQCS is a natural extension of VQSP to quantum circuits. Here, our goal is to learn the parameters of an n-qubit ansatz $U(\boldsymbol{\theta})$ that best approximates a given unknown quantum gate V. Similar to how we use infidelity for quantum states, we can use the Hilbert-Schmidt cost function defined for unitaries in [Khatri *et al.*, 2019]. For any $\boldsymbol{\theta}$, this is computed as $H(\boldsymbol{\theta})=1-|\text{tr}(U(\boldsymbol{\theta})^\dagger V)|^2/4^n$ and minimizing H gives us the set of parameters that (approximately) prepares V.

To see why, first note that any quantum gate W can be uniquely identified using a representation given as $W\otimes\overline{W}$, where $\overline{W}$ is the complex conjugate of W. This can be derived from its action on the vectorized version of elements in $\mathcal{L}(\mathbb{C}^{2^n})$. Then we see that $H(\boldsymbol{\theta})$ is proportional to $\|U(\boldsymbol{\theta})\otimes\overline{U(\boldsymbol{\theta})}-V\otimes\overline{V}\|_F^2$. To evaluate $H(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$, we start with the maximally entangled state on two n-qubit systems, defined as $|\Phi\rangle=1/\sqrt{2^n}\sum_{i=0}^{2^n-1}|i\rangle|i\rangle$. Then, we apply V on the second register to obtain $|V\rangle=1/\sqrt{2^n}\sum_{i=0}^{2^n-1}|i\rangle|v_{\bullet i}\rangle$, where $|v_{\bullet i}\rangle$ is the i^{th} column of V. Then, one can see that $H(\boldsymbol{\theta})=1-\langle|U(\boldsymbol{\theta})\rangle\langle U(\boldsymbol{\theta})|\rangle_{|V\rangle\langle V|}$. Therefore, we can use shallow shadows of $|V\rangle$ to estimate $H(\boldsymbol{\theta})$. Since $\|\,|U(\boldsymbol{\theta})\rangle\langle U(\boldsymbol{\theta})|\,\|_F=1$ for all $\boldsymbol{\theta}$, the number of shadows, or equivalently, the number of applications of V, is independent of n.

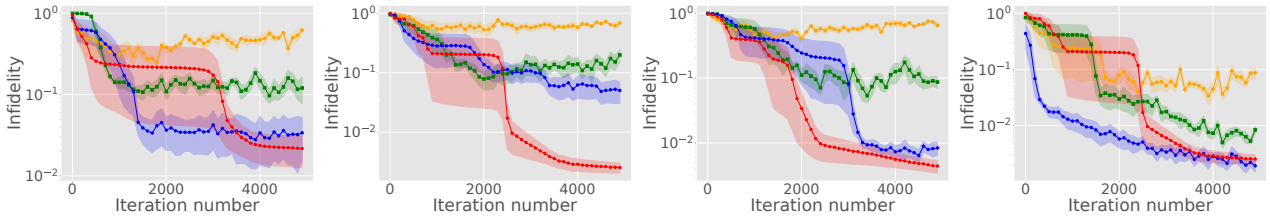
In terms of classical computational complexity, $\langle|U(\boldsymbol{\theta})\rangle\langle U(\boldsymbol{\theta})|\rangle_{\hat{\rho}}$ for any shallow shadow $\hat{\rho}$ can be computed by contracting the tensor network given in Figure 3(b), the cost of which is polynomial in n. The explanation regarding the usage of ALA in this figure is the same as the one for VQSP. From now on, when discussing the sample complexity of VQCS, the "number of copies" will mean the number of copies of $|V\rangle$ consumed (equivalently, the number of applications of V).

6 Simulation Results

Here we elaborate on the experimental results by comparing the sample complexity of AISO and the standard VQA in the two use cases discussed above.

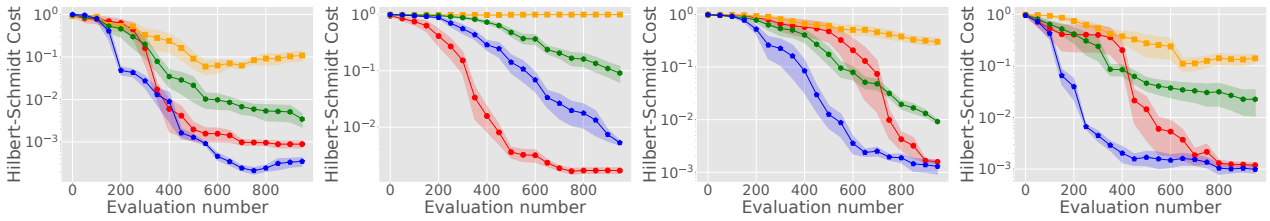
The depth d of the shallow shadow ensemble (cf. Figure 2) is set to 3 throughout the experiments. The viability of AISO in solving both problems is tested across four different ansatzes that are widely used in the literature, whose structures are given in Figure 1(a,b,c,d). Except in HEA, all two-qubit gates can be arbitrary two-qubit subcircuits. The specific ones used in our simulation are given in Figure 1(e). Also, for VQCS, each two-qubit subcircuit is a combination of two of these. In HEA, the two-qubit gate used is the CNOT gate.

For VQSP, we have used the Simultaneous Perturbation Stochastic Approximation [Spall, 1992] (SPSA), where the converging sequences used are, respectively, $c_r=a_r=r^{-0.4}$ and the total number of iterations is 5000. On the other



(a) ALA (b) MERA (c) HEA (d) TTN

Figure 4: Simulation results for state preparation. Here, the red curve represents AISO (10^4) while the orange, green, and blue curves represent VQA (5×10^5), VQA (10^6) and VQA (2.5×10^6) respectively.



(a) ALA (b) MERA (c) HEA (d) TTN

Figure 5: Simulation results for circuit synthesis. Here, the red curve represents AISO (10^4) while the orange, green, and blue curves represent VQA (10^5), VQA (10^6) and VQA (10^7) respectively.

hand, the results of VQCS have used Powell's method [Powell, 1964] with a maximum of 10^3 function evaluations allowed. We denote by AISO/VQA (T) the AISO/VQA algorithm that uses T copies in total. This means that VQA (T) will consume $T/10^4$ copies per function evaluation in SPSA and $T/10^3$ copies in Powell's method. This is because SPSA requires two function evaluations to produce estimates of the gradient.

The unknown target states considered in the VQSP are 8-qubit states, which are also compatible with the corresponding ansatzes being used. In each setting, the experiment is carried out across five different states and the results are shown in Figure 4. Here, we have plotted the mean of infidelity values achieved at different iterations across the five different experiments that were carried out. The shaded region comprises the mean plus and minus 0.3 times the standard deviation of the five different infidelities.

In Figure 4, VQA (5×10^5), which utlizes 5×10^5 copies in total, consumes 50 state copies per function evaluation. Similarly, the other VQA algorithms consume 100 and 250 state copies per evaluation. One can see that AISO closely matches or outperforms the results of VQA by consuming only 10^4 copies in total.

Moving on to VQCS, similar experiments are carried out for 4-qubit quantum gates (meaning 8-qubits used in total). The results are summarized in Figure 5. Here, the minimum $H(\boldsymbol{\theta})$ in each interval of 10^2 function evaluations out of the total allowed 10^3 is plotted. The three VQA algorithms used here consume $10^2, 10^3$ and 10^4 copies per function evaluation respectively. It is clear from the plots that AISO can match the performance of standard VQA similarly using considerably fewer copies to what we saw in the case of VQSP.

In Figures 6 and 7, we present the superiority of AISO over VQA in a different light. On the x-axis, we plot different infidelity or Hilbert-Schmidt cost values, and on the y-axis, we plot the number of copies required to achieve them, which are exponentially better for AISO.

7 Improved Bounds Using 2-Design Assumption

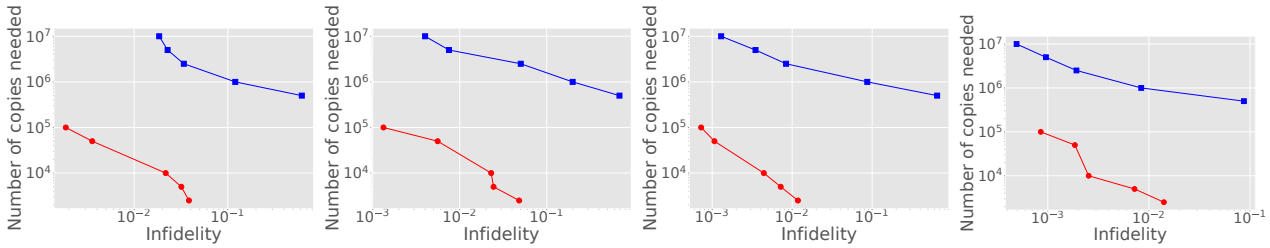
In this section, we analyze the assumption of the input state in more detail. The assumption that the state is sampled from a 1-design merely says that the input state is the maximally mixed state. So, to further understand the notion of a "typical input state" and to get closer to the notion of the input state being an average state or a randomly generated state, we make a stronger assumption on the distribution. More precisely, we assume that the input state is sampled from a *state 2-design* $\mathcal{D}_2$. These are ensembles such that sampling from them is equivalent to sampling a pure state uniformly (according to Haar measure) up to two statistical moments. 2-designs are extensively used in quantum information to generate pseudorandomness and to analyze average case complexities [Dankert *et al.*, 2009; Scott, 2008; Ambainis *et al.*, 2009].

In this regime, we derive two results, starting with an upper bound on the variance of the state-dependent shadow norm when the state is sampled from a state 2-design.

Theorem 5. *Let $\mathcal{D}_2$ be a state 2-design and $d = \Theta(\log n)$. Then, for any observable O, we have*

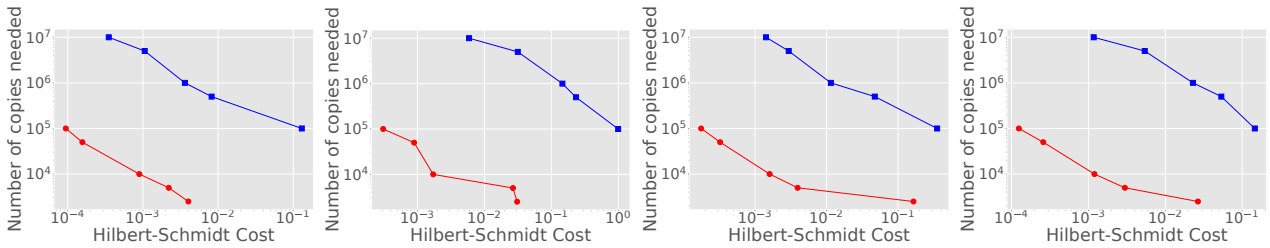
$$Var_{\sigma\sim\mathcal{D}_2}\left(\|O\|^2_{\sigma,\mathcal{U}_d}\right) \leq 64\|O\|^2_F. \tag{9}$$

Using this result, we can derive a result similar to Theorem 3, with better constants.



(a) ALA (b) MERA (c) HEA (d) TTN

Figure 6: Resource needs for different infidelity objectives. Here, the red curve represents AISO, while the blue curve represents VQA.



(a) ALA (b) MERA (c) HEA (d) TTN

Figure 7: Resource needs for different Hilbert Schmidt Cost objectives. Here, the red curve represents AISO, while the blue curve represents VQA.

Theorem 6. *Let $d = \Theta(\log n)$ and ρ be an n-qubit pure state sampled from a state 2-design $\mathcal{D}_2$. For any $\delta, \epsilon \in (0,1)$, $m > 1/\sqrt{\delta}$, and any $C > 0$, let*

$$T_1 \geq 2\log\left(\frac{2(m^2-1)C}{m^2\delta - 1}\right), \; T_2 \geq \frac{136}{\epsilon^2}(2m+1)\|O\|_F^2. \tag{10}$$

For any parameter vectors $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(C)}$, with probability at least $1-\delta$, we have $|\langle W_O(\boldsymbol{\theta}^{(c)})\rangle_\rho - \langle \widehat{W_O}(\boldsymbol{\theta}^{(c)})\rangle_\rho| \leq \epsilon$ for all $1 \leq c \leq C$, where $\langle W_O(\boldsymbol{\theta}^{(c)})\rangle_\rho$ and $\langle \widehat{W_O}(\boldsymbol{\theta}^{(c)})\rangle_\rho$ are defined in Eq.s (6) *and* (8)*, respectively.*

Hence, we see that the lower bound on T_1 in Eq. (10) is a constant time better than the lower bound on T_1 in Eq. (7). By replacing the function evaluations in Theorem 6 with expectations with arbitrary observables, one can see that similar advantages can be gained for regular shallow shadow estimation also when the input is sampled from a 2-design.

8 Dealing with Barren Plateaus

Global observables may lead to barren plateaus (regions with gradients exponentially small in the number of qubits) occurring in the training landscape [Cerezo *et al.*, 2021; Liu *et al.*, 2022], which makes evaluating them using quantum devices extremely difficult. Although the gradients are evaluated classically in AISO, since we may encounter global observables and require $\mathcal{O}(1/\epsilon^2)$ shadows to additively approximate the gradients to precision ϵ, in some cases, we might end up requiring exponentially many shadows for meaningful approximations. However, several heuristic approaches have been proposed, which have been experimentally shown to reduce barren plateaus in many cases. We note that our method is compatible with almost all barren plateau mitigating methods that have been proposed in the literature. For example, [Patti *et al.*, 2021; Mele *et al.*, 2022; Rad *et al.*, 2022; Skolik *et al.*, 2021; Grimsley *et al.*, 2023a; Grimsley *et al.*, 2023b; Friedrich and Maziero, 2022; Verdon *et al.*, 2019; Grant *et al.*, 2019; Kulshrestha and Safro, 2022; Zhang *et al.*, 2022] are methods that ultimately use the quantum device only to estimate $\langle W_O(\boldsymbol{\theta})\rangle_\rho$ at certain carefully chosen inputs $\boldsymbol{\theta}$. So, it is clear that if we use shadows to estimate them, then exponential advantages similar to the ones discussed in this paper can be achieved.

9 Conclusion and Future Direction

In this work, we proposed AISO — a training algorithm that leverages shallow shadows to achieve an exponential reduction in quantum resources required to train VQA cost functions. AISO is a very general approach that works with almost all of the popular shallow quantum circuit structures in the literature, when used in combination with observables of low Frobenius norm. It allows one to do more iterations of the classical optimizer, more hyperparameter tuning, and experiment with various ansatzes and optimizers with very few executions of the quantum device. We demonstrated this advantage in two important use cases of interest in quantum information: Variational Quantum State Preparation and Variational Quantum Circuit Synthesis.

For future work, we aim to design similar resource-efficient and ansatz-agnostic protocols for local observables, by leveraging classical machine learning with classical shadows similar to [Huang *et al.*, 2021].

Acknowledgements

We thank Afham for pointing out important references and related works. We also thank Hakop Pashayan and Christian Bertoni for helping us clear up any questions we had about shallow shadows. This work is partially supported by the Australian Research Council (Grant No: DP220102059). AB was partially supported by the Sydney Quantum Academy PhD scholarship.

References

[Ambainis *et al.*, 2009] Andris Ambainis, Jan Bouda, and Andreas Winter. Nonmalleable encryption of quantum information. *Journal of Mathematical Physics*, 50(4), apr 2009.

[Arute *et al.*, 2019] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando Brandao, David Buell, Brian Burkett, Yu Chen, Jimmy Chen, Ben Chiaro, Roberto Collins, William Courtney, Andrew Dunsworth, Edward Farhi, Brooks Foxen, Austin Fowler, Craig Michael Gidney, Marissa Giustina, Rob Graff, Keith Guerin, Steve Habegger, Matthew Harrigan, Michael Hartmann, Alan Ho, Markus Rudolf Hoffmann, Trent Huang, Travis Humble, Sergei Isakov, Evan Jeffrey, Zhang Jiang, Dvir Kafri, Kostyantyn Kechedzhi, Julian Kelly, Paul Klimov, Sergey Knysh, Alexander Korotkov, Fedor Kostritsa, Dave Landhuis, Mike Lindmark, Erik Lucero, Dmitry Lyakh, Salvatore Mandrà, Jarrod Ryan McClean, Matthew McEwen, Anthony Megrant, Xiao Mi, Kristel Michielsen, Masoud Mohseni, Josh Mutus, Ofer Naaman, Matthew Neeley, Charles Neill, Murphy Yuezhen Niu, Eric Ostby, Andre Petukhov, John Platt, Chris Quintana, Eleanor G. Rieffel, Pedram Roushan, Nicholas Rubin, Daniel Sank, Kevin J. Satzinger, Vadim Smelyanskiy, Kevin Jeffery Sung, Matt Trevithick, Amit Vainsencher, Benjamin Villalonga, Ted White, Z. Jamie Yao, Ping Yeh, Adam Zalcman, Hartmut Neven, and John Martinis. Quantum supremacy using a programmable superconducting processor. *Nature*, 574:505–510, 2019.

[Basheer *et al.*, 2023] Afrad Basheer, Yuan Feng, Christopher Ferrie, and Sanjiang Li. Alternating layered variational quantum circuits can be classically optimized efficiently using classical shadows. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6770–6778, Jun. 2023.

[Bertoni *et al.*, 2023] Christian Bertoni, Jonas Haferkamp, Marcel Hinsche, Marios Ioannou, Jens Eisert, and Hakop Pashayan. Shallow shadows: Expectation estimation using low-depth random clifford circuits. *arXiv:2209.12924*, 2023.

[Boyd and Koczor, 2022] Gregory Boyd and Bálint Koczor. Training variational quantum circuits with covar: covariance root finding with classical shadows. *arXiv:2204.08494*, 2022.

[Bridgeman *et al.*, 2015] Jacob C. Bridgeman, Aroon O'Brien, Stephen D. Bartlett, and Andrew C. Doherty. Multiscale entanglement renormalization ansatz for spin chains with continuously varying criticality. *Phys. Rev. B*, 91:165129, Apr 2015.

[Cerezo *et al.*, 2021] M. 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), mar 2021.

[Cerezo *et al.*, 2023] M. Cerezo, Martin Larocca, Diego García-Martín, N. L. Diaz, Paolo Braccia, Enrico Fontana, Manuel S. Rudolph, Pablo Bermejo, Aroosa Ijaz, Supanut Thanasilp, Eric R. Anschuetz, and Zoë Holmes. Does provable absence of barren plateaus imply classical simulability? or, why we need to rethink variational quantum computing. *arXiv:2312.09121*, 2023.

[Dankert *et al.*, 2009] Christoph Dankert, Richard Cleve, Joseph Emerson, and Etera Livine. Exact and approximate unitary 2-designs and their application to fidelity estimation. *Phys. Rev. A*, 80:012304, Jul 2009.

[Farhi *et al.*, 2014] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm. *arXiv:1411.4028*, 2014.

[Friedrich and Maziero, 2022] Lucas Friedrich and Jonas Maziero. Avoiding barren plateaus with classical deep neural networks. *Physical Review A*, 106(4), oct 2022.

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

[Grimsley *et al.*, 2023a] Harper Grimsley, Nicholas Mayhall, George Barron, Edwin Barnes, and Sophia Economou. Adaptive, problem-tailored variational quantum eigensolver mitigates rough parameter landscapes and barren plateaus. *npj Quantum Information*, 9, 03 2023.

[Grimsley *et al.*, 2023b] Harper R. Grimsley, George S. Barron, Edwin Barnes, Sophia E. Economou, and Nicholas J. Mayhall. Adaptive, problem-tailored variational quantum eigensolver mitigates rough parameter landscapes and barren plateaus. *npj Quantum Information*, 9(1), mar 2023.

[Harrow and Low, 2009] Aram W. Harrow and Richard A. Low. Random quantum circuits are approximate 2-designs. *Communications in Mathematical Physics*, 291(1):257–302, July 2009.

[Havlíček *et al.*, 2019] Vojtěch Havlíček, Antonio D. Córcoles, Kristan Temme, Aram W. Harrow, Abhinav Kandala, Jerry M. Chow, and Jay M. Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, mar 2019.

[Huang *et al.*, 2020] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, jun 2020.

[Huang *et al.*, 2021] Hsin-Yuan Huang, Richard Kueng, Giacomo Torlai, Victor V. Albert, and John Preskill. Provably

efficient machine learning for quantum many-body problems. *arXiv:2106.12627*, 2021.

[Khatri *et al.*, 2019] Sumeet Khatri, Ryan LaRose, Alexander Poremba, Lukasz Cincio, Andrew T. Sornborger, and Patrick J. Coles. Quantum-assisted quantum compiling. *Quantum*, 3:140, May 2019.

[Kulshrestha and Safro, 2022] Ankit Kulshrestha and Ilya Safro. Beinit: Avoiding barren plateaus in variational quantum algorithms. *arXiv:2204.13751*, 2022.

[Leone *et al.*, 2022] Lorenzo Leone, Salvatore F. E. Oliviero, Lukasz Cincio, and M. Cerezo. On the practical usefulness of the hardware efficient ansatz. *arXiv:2211.01477*, 2022.

[Liu *et al.*, 2022] Zidu Liu, Li-Wei Yu, L.-M. Duan, and Dong-Ling Deng. Presence and absence of barren plateaus in tensor-network based machine learning. *Phys. Rev. Lett.*, 129:270501, Dec 2022.

[Madsen *et al.*, 2022] Lars Skovgaard Madsen, Fabian Laudenbach, Mohsen Falamarzi Askarani, Fabien Rortais, Trevor Vincent, Jacob F F Bulmer, Filippo M. Miatto, Leonhard Neuhaus, Lukas G Helt, Matthew J Collins, Adriana E. Lita, Thomas Gerrits, Sae Woo Nam, Varun Vaidya, Matteo Menotti, Ish Dhand, Zachary Vernon, Nicolás Quesada, and Jonathan Lavoie. Quantum computational advantage with a programmable photonic processor. *Nature*, 606:75 – 81, 2022.

[McClean *et al.*, 2016] Jarrod R McClean, Jonathan Romero, Ryan Babbush, and Alá n Aspuru-Guzik. The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics*, 18(2):023023, feb 2016.

[Mele *et al.*, 2022] Antonio A. Mele, Glen B. Mbeng, Giuseppe E. Santoro, Mario Collura, and Pietro Torta. Avoiding barren plateaus via transferability of smooth solutions in a hamiltonian variational ansatz. *Physical Review A*, 106(6), dec 2022.

[Mitarai *et al.*, 2018] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii. Quantum circuit learning. *Phys. Rev. A*, 98:032309, Sep 2018.

[Patti *et al.*, 2021] Taylor L. Patti, Khadijeh Najafi, Xun Gao, and Susanne F. Yelin. Entanglement devised barren plateau mitigation. *Physical Review Research*, 3(3), jul 2021.

[Peruzzo *et al.*, 2014] 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(1), jul 2014.

[Powell, 1964] M. J. D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2):155–162, 01 1964.

[Preskill, 2018] John Preskill. Quantum Computing in the NISQ era and beyond. *Quantum*, 2:79, August 2018.

[Rad *et al.*, 2022] Ali Rad, Alireza Seif, and Norbert M. Linke. Surviving the barren plateau in variational quantum circuits with bayesian learning initialization. *arXiv:2203.02464*, 2022.

[Romero *et al.*, 2017] Jonathan Romero, Jonathan P Olson, and Alan Aspuru-Guzik. Quantum autoencoders for efficient compression of quantum data. *Quantum Science and Technology*, 2(4):045001, aug 2017.

[Schreiber *et al.*, 2022] Franz J. Schreiber, Jens Eisert, and Johannes Jakob Meyer. Classical surrogates for quantum learning models. *arXiv:2206.11740*, 2022.

[Scott, 2008] A J Scott. Optimizing quantum process tomography with unitary 2-designs. *Journal of Physics A: Mathematical and Theoretical*, 41(5):055308, jan 2008.

[Shi *et al.*, 2006] Y.-Y. Shi, L.-M. Duan, and G. Vidal. Classical simulation of quantum many-body systems with a tree tensor network. *Phys. Rev. A*, 74:022320, Aug 2006.

[Skolik *et al.*, 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), jan 2021.

[Spall, 1992] J.C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

[Stokes *et al.*, 2020] James Stokes, Josh Izaac, Nathan Killoran, and Giuseppe Carleo. Quantum Natural Gradient. *Quantum*, 4:269, May 2020.

[Verdon *et al.*, 2019] Guillaume Verdon, Michael Broughton, Jarrod R. McClean, Kevin J. Sung, Ryan Babbush, Zhang Jiang, Hartmut Neven, and Masoud Mohseni. Learning to learn with quantum neural networks via classical neural networks. *arXiv:1907.05415*, 2019.

[Wierichs *et al.*, 2020] David Wierichs, Christian Gogolin, and Michael Kastoryano. Avoiding local minima in variational quantum eigensolvers with the natural gradient optimizer. *Phys. Rev. Research*, 2:043246, Nov 2020.

[Zhang *et al.*, 2022] Kaining Zhang, Liu Liu, Min-Hsiu Hsieh, and Dacheng Tao. Escaping from the barren plateau via gaussian initializations in deep variational quantum circuits. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 18612–18627. Curran Associates, Inc., 2022.

[Zhong *et al.*, 2020] Han-Sen Zhong, Hui Wang, Yu-Hao Deng, Ming-Cheng Chen, Li-Chao Peng, Yi-Han Luo, Jian Qin, Dian Wu, Xing Ding, Yi Hu, Peng Hu, Xiao-Yan Yang, Wei-Jun Zhang, Hao Li, Yuxuan Li, Xiao Jiang, Lin Gan, Guangwen Yang, Lixing You, Zhen Wang, Li Li, Nai-Le Liu, Chao-Yang Lu, and Jian-Wei Pan. Quantum computational advantage using photons. *Science*, 370(6523):1460–1463, 2020.