On the Trainability and Classical Simulability of Learning Matrix Product States Variationally

Afrad Basheer[1], Yuan Feng[2], Christopher Ferrie[1], Sanjiang Li[1] and Hakop Pashayan[3]

[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
[3] Dahlem Center for Complex Quantum Systems, Freie Universität Berlin 14195, Germany
afrad.m.basheer@student.uts.edu.au, yuan_feng@tsinghua.edu.cn,
{christopher.ferrie, sanjiang.li}@uts.edu.au, hakopsadventures@gmail.com

Abstract

We prove that using global observables to train the matrix product state ansatz results in the vanishing of all partial derivatives, also known as barren plateaus, while using local observables avoids this. This ansatz is widely used in quantum machine learning for learning weakly entangled state approximations. Additionally, we empirically demonstrate that in many cases, the objective function is an inner product of almost sparse operators, highlighting the potential for classically simulating such a learning problem with few quantum resources. All our results are experimentally validated across various scenarios.

Introduction

Quantum computing has made major strides in the past few years in terms of qubit capacity (Chow, Dial, and Gambetta 2021; Gambetta 2022) and error correction (da Silva et al. 2024), prompting significant efforts in demonstrating concrete advantages over classical computers. Although this has been achieved in some cases (Arute et al. 2019; Zhong et al. 2020; Madsen et al. 2022), superiority in solving practically useful problems remains to be seen. With the emergence of Noisy Intermediate Scale Quantum (NISQ) devices (Preskill 2018), recent focus has shifted to quantum optimization algorithms, especially Variational Quantum Algorithms (VQAs) (Cerezo et al. 2021a). In VQAs, we use quantum computers to estimate objective functions involving quantum *states* and parameterized quantum circuits (also called *ansatzes*) and update the parameters of these functions classically. Popular examples include variational quantum eigensolver (Peruzzo et al. 2014), quantum support vector machines (Havlíček et al. 2019), quantum approximate optimization algorithm (Farhi, Goldstone, and Gutmann 2014), etc.

An important example that features extensively in quantum information is learning cost-effective approximations of quantum states (Cerezo et al. 2021b; Matos, Johri, and Papić 2021; Gard et al. 2020). These techniques can be used to learn ansatzes that can approximately prepare states using fewer gates than initially required. Among the many choices of ansatzes used, interest has surged in the Matrix Product State (MPS) ansatz (cf. Figure 1 (a)) (Rudolph et al. 2022; Ran 2020; Ben-Dov et al. 2022; Lin et al. 2021; Rieser, Köster, and Raulf 2023). The MPS ansatz leverages the MPS data structure, which stores quantum states with space complexity that scales polynomially with the *bond dimensions* (parameters that measure entanglement in linear qubit arrays). Consequently, the MPS ansatz is particularly effective for learning weakly entangled approximations, potentially resulting in fewer gate counts and simpler gate connectivity requirements compared to other approaches.

Learning state approximations variationally can be performed using either global or local observables (Cerezo et al. 2021b). Experimental results in (Ben-Dov et al. 2022) showed that using the MPS ansatz along with global observables for state approximations can result in barren plateaus, where all partial derivatives become exponentially small in the number of qubits. This makes estimating these derivatives using quantum devices require exponentially many executions. Moreover, the parameter updates also become exponentially small. In contrast, it was empirically shown that using local observables can help mitigate this issue.

Theoretically proving this phenomenon is a challenging task. Although this has been achieved for similar problems such as optimization over the Hardware Efficient Ansatz (HEA) (Cerezo et al. 2021b) (cf. Figure 1 (b)) and tensor network-based optimization in quantum information (Liu et al. 2022), these results cannot be used to explain exponentially vanishing objective functions and gradients for the MPS ansatz.

This work provides rigorous trainability proofs for the MPS ansatz. We prove that under uniformly random initialization of the circuit parameters, when using global observables, the variance of the objective function decreases exponentially while the usage of local observables ensures that the same variance is lower bounded by a quantity whose dependence on the number of qubits is linear and scales exponentially only in the width of the subcircuit involved. We also relate this with the variance of the partial derivatives and show that similar results hold for them as well.

Trainability is closely interrelated with classical simulability. In (Cerezo et al. 2024), it was conjectured, with evidence, that provably avoiding barren plateaus in this manner could imply classical simulability with few quantum resources. That is, for all provably trainable VQA objective functions, one can simulate the whole optimization classically using

the outputs of a few quantum measurements implemented beforehand on the input state. By proving the trainability of the MPS ansatz and local observable combination, our work prepares the groundwork for studying its classical simulability.

On this side, we demonstrate that these trainable VQA objective functions exhibit *effective subspaces*. These subspaces are loosely defined as the subspaces where the observables, when conjugated with the ansatzes, tend to be mostly concentrated, for almost all input parameters (Cerezo et al. 2024). If the objective function exhibits this property, then most function evaluations, which are nothing but inner products of the state with these conjugated observables (cf. Eq (2)), could potentially be classically estimated using the input state's coefficients in this subspace estimated beforehand using a quantum device. We first characterize the property of exhibiting effective subspaces by introducing an efficiently estimable norm for observables, the C-$\mathbb{K}$ norm, which we use to experimentally show that the MPS ansatz and local observable combination exhibits an effective subspace within the Pauli basis.

Our main contributions can be summarized as follows:

1. For the problem of learning weakly entangled state approximations variationally, we rigorously prove that the usage of global observables will induce barren plateaus, while the usage of local observables will avoid them.
2. We empirically show that the MPS ansatz, when used in combination with local observables, exhibits an effective subspace within the Pauli basis, which is conjectured to be a consequence of avoiding cost concentration and a sufficient condition for the ansatz to be classically simulable using few quantum resources as per (Cerezo et al. 2024).

Finally, we experimentally validate our results across various scenarios, including the impact of observable choices on MPS ansatz optimization and the detection of effective subspaces in MPS ansatz as well as other ansatzes such as HEA, and Quantum Convolutional Neural Network (QCNN) (cf. Figure 1 (c)).

Related Works

The impact of the locality of the observables on trainability was first observed in (Cerezo et al. 2021b), wherein the usage of global (local) observables were shown to induce (avoid) barren plateaus. But, the results of this work only hold for HEAs, and not for the MPS ansatz. Also, our inductive proof technique, which requires integration of only one subcircuit to start the induction, is very different from the techniques used in (Cerezo et al. 2021b), where the proof requires more complicated integrations over many subcircuits.

In (Liu et al. 2022; Garcia et al. 2023; Barthel and Miao 2024), the theoretical study of barren plateaus in tensor-network-based machine learning with MPS inputs reveals that using global observables in the objective function introduces barren plateaus, whereas local observables avoid them. However, as mentioned in (Liu et al. 2022), their model and assumptions differ from a variational circuit model. They model the input using the unitary decomposition of MPS, where each component tensor is reshaped into a $2D \times 2D$ unitary matrix, with D as the bond dimension. The randomness is modeled by assuming these unitaries form *unitary* 2-*designs* (Dankert et al. 2009), which are ensembles of unitaries such that integrating quadratic polynomials over them is equivalent to integrating the same over the Haar measure. In contrast, we assume that the subcircuits are sampled from unitary 2-designs, which is more natural for circuit-based problems as a circuit depth polynomial in the width of the subcircuits suffices for them to behave like a unitary 2-design under uniformly random parameter initialization (Harrow and Low 2009).

In (Ben-Dov et al. 2022), it was experimentally observed that while using the MPS ansatz, the usage of global observables leads to exponentially decreasing gradient magnitudes, whereas local observables avoid this issue. In our work, we study this phenomenon theoretically as well as similar trends in cost concentration. The existence of exponentially decreasing partial derivatives in MPS ansatz-based VQAs is proved using ZX-calculus in (Zhao and Gao 2021). However, the method can only be used to prove this for individual examples of the MPS ansatz, with pre-defined structures for the subcircuits. In contrast, our proofs consider the most generalized form of the ansatz, with the only assumption being that the subcircuits form unitary 2 designs. Also in (Zhao and Gao 2021), there are no discussions regarding the impact that observables and subcircuit widths can have on trainability, which we theoretically demonstrate in the case of cost concentration as well as barren plateaus.

Background

We denote column vectors as $|\psi\rangle$ ('ket' notation) and their conjugate transposes as $\langle\psi|$ ('bra' notation). The vector $|i\rangle \in \mathbb{C}^d$ represents the i^{th} computational basis vector. We also define $|\mathbf{0}\rangle \coloneqq |0\rangle^{\otimes n}$, where n is the number of qubits involved. We define $\mathbb{1}_t$ to be the identity matrix acting on $\mathbb{C}^{2^t}$, with $\mathbb{1} \coloneqq \mathbb{1}_1$, and for any $A \in \mathbb{C}^{2\times 2}$, we define $A_t^{(n)} \coloneqq \mathbb{1}_{n-t} \otimes A \otimes \mathbb{1}_{t-1}$. For a set of matrices $\{A^{(1)}, \ldots, A^{(p)}\}$, we define $\prod_{i=1}^{p} A^{(i)} \coloneqq A^{(1)} \ldots A^{(p)}$. For any complex matrix $A = \sum_{ij} A_{ij} |i\rangle \langle j|$, $\|A\|_1 \coloneqq \sum_{ij} |A_{ij}|$ and $\|A\|_{\text{tr}} \coloneqq \text{tr}(\sqrt{A^\dagger A})$. Any matrix $A \in \mathbb{C}^{2^t \times 2^t}$ such that $A = A^{(1)} \otimes \cdots \otimes A^{(t)}$ for some $A^{(1)}, \ldots, A^{(n)} \in \mathbb{C}^{2\times 2}$ is called a *product* matrix.

Quantum Computing

A (pure) quantum *state* is a (rank one) positive semidefinite operator $\sigma \in \mathbb{C}^{d\times d}$, such that $\text{tr}(\sigma) = 1$. In quantum computing, a *qubit* is the quantum counterpart of a classical bit. An n-qubit system's state can be represented by a quantum state in $\mathbb{C}^{2^n \times 2^n}$.

A quantum *gate* operating on n qubits is represented by a unitary operator $U \in \mathbb{C}^{2^n \times 2^n}$, which transforms the state of an n-qubit system from σ to $\sigma_U \coloneqq U\sigma U^\dagger$. In particular, the one-qubit *Pauli gates* are defined as

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

We define $\mathbb{P}_n$ as the set of all n-fold tensor products of the elements in $\{\mathbb{1}, X, Y, Z\}$.

A quantum *circuit* is defined as a composition of multiple quantum gates. The *width* of a quantum circuit is defined as the number of qubits on which it is defined. A Hermitian

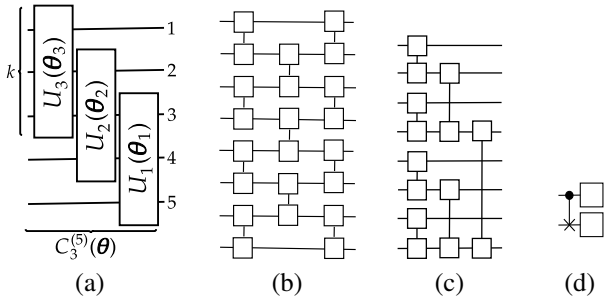

Figure 1: (a) Example of an MPS ansatz with $n = 5$ and $t = 3$. The numbers on the end of each qubit wire are the indices of the qubits. (b) HEA, where each connected pair of boxes are 2-qubit subcircuits. (c) QCNN ansatz (d) The 2-qubit subcircuit used in our simulations within HEA and QCNN, where each box is an arbitrary single qubit gate

operator, also known as an *observable*, can be used to define a *measurement* and an *expectation value*. A measurement of a system in a state σ, using an observable W with eigendecomposition $W = \sum_i \omega_i |w_i\rangle \langle w_i|$ results in an outcome ω_i with probability $\langle w_i| \sigma |w_i\rangle$. Thus, the expected value of the measurement outcome is $\text{tr}(W\sigma)$.

We define $\mathbb{D}_n, \mathbb{U}_n$ and $\mathbb{H}_n$ as the set of all n-qubit pure states, gates, and observables. If a matrix A acting on an n-qubit system acts as $\mathbb{1}_{n-k}$ on $n-k$ qubits, then we call the matrix a k-local matrix. For any two pure n-qubit states σ and ρ, the quantum *fidelity*, defined as $F(\rho, \sigma) \coloneqq \text{tr}(\rho\sigma)$, serves as a measure of similarity between the states with a fidelity of 1 indicating that $\sigma = \rho$.

Variational Quantum Algorithms

Many optimization problems in quantum information can be formulated as the optimization of outputs from parameterized quantum circuits, also called *ansatzes*, denoted as $C(\boldsymbol{\theta}) \in \mathbb{U}_n$, where $\boldsymbol{\theta}$ is a vector of parameters. Typically, these problems involve optimizing (over $\boldsymbol{\theta}$) the function defined as

$$f_{\sigma,W,C}(\boldsymbol{\theta}) = \text{tr}\left(WC(\boldsymbol{\theta})\sigma C(\boldsymbol{\theta})^\dagger\right) = \text{tr}\left(W\sigma_{C(\boldsymbol{\theta})}\right), \quad (2)$$

where σ is the input quantum state, $\sigma_{C(\boldsymbol{\theta})}$ is $C(\boldsymbol{\theta})\sigma C(\boldsymbol{\theta})^\dagger$ by our notation, and W is an output observable. One can estimate this function for any input $\boldsymbol{\theta}$ by sample means estimation of the expected value of measuring the state $\sigma_{C(\boldsymbol{\theta})}$ using W. With this capability, all partial derivatives can also be estimated using standard methods such as finite differencing or quantum-specific ones such as the parameter shift rule (Mitarai et al. 2018). The $\boldsymbol{\theta}$ search space has $\mathcal{O}(\text{poly}(n))$ dimensions. Hence, $\boldsymbol{\theta}$ can be updated towards the optimum classically. Throughout this work, we omit the subscript C and use the notation $f_{\sigma,W}$ since the choice of ansatz C is usually implicitly understood from the context.

MPS Ansatz

The MPS ansatz is given in Figure 1 (a). It is built using t cascading layers of smaller parameterized subcircuits $U_i(\boldsymbol{\theta}_i)$ for $i = 1, \ldots, t$, each with width k. Mathematically, the ansatz is defined as

$$C_t^{(n)}(\boldsymbol{\theta}) = \prod_{p=1}^{t} \mathbb{1}^{\otimes n-k-p+1} \otimes U_p(\boldsymbol{\theta}_p) \otimes \mathbb{1}^{\otimes p-1}, \quad (3)$$

where $t \leq n-k+1$, $\boldsymbol{\theta} = \boldsymbol{\theta}_1 \oplus \cdots \oplus \boldsymbol{\theta}_t$ with $\boldsymbol{\theta}_p = [\theta_{p1}, \theta_{p2}, \ldots, \theta_{pm}]$ and $U_p(\boldsymbol{\theta}_p) = \prod_{q=1}^{m} e^{-i\theta_{pq}H_{pq}}$ are k-qubit parameterized subcircuits, with $H_{pq} \in \mathbb{H}_k$.

This ansatz has a close relationship with the MPS data structure, which can store quantum states with space complexity that is positively correlated with the entanglement between neighboring qubits measured using their bond dimensions (Cirac et al. 2021). From (Cramer et al. 2010), we can see that every state that can be represented efficiently as an MPS with bond dimensions at most 2^{k-1} can be implemented using this ansatz (assuming that U_p can implement any k-qubit unitary). This is what led many works to use the MPS ansatz to solve state approximation problems variationally (Lin et al. 2021; Rudolph et al. 2022; Ben-Dov et al. 2022; Ran 2020; Rieser, Köster, and Raulf 2023).

Throughout this work, we set $T = n-k+1$, and our focus is on $C_T^{(n)}$. Also, in appropriate contexts, we denote $C_T^{(n)}$ as C_T, as the dependency on n is implied by the system's size.

State Approximation Using MPS Ansatz

The problem that we consider here is as follows: given many copies of an n-qubit pure state σ (or stronger access to a circuit that can prepare σ), output the parameters of an MPS ansatz such that we can approximately prepare σ using these parameters. We present two different approaches to solve this problem, differing only in their measurement strategies.

In the first method, the idea is to maximize the fidelity between $|\mathbf{0}\rangle \langle \mathbf{0}|_{C_T(\boldsymbol{\theta})^\dagger}$ and σ, over $\boldsymbol{\theta}$. That is, find $\arg\max_{\boldsymbol{\theta}} f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta})$. This fidelity can be estimated by applying $C_T(\boldsymbol{\theta})$ on σ, measuring all qubits simultaneously using the observable Z, and estimating the probability of all measurements resulting in $+1$. So the observable whose expectation features within the objective function is the *global observable* $|\mathbf{0}\rangle \langle \mathbf{0}|$.

Alternatively, in the second method, we employ the expectation of $O \coloneqq 1/n \sum_{i=1}^{n} |0\rangle \langle 0|_i$, which is a sum of 1-local observables. The intuition here is that if $\boldsymbol{\theta}^*$ maximizes $f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}$, then $\sigma_{C_T(\boldsymbol{\theta}^*)} = |\mathbf{0}\rangle \langle \mathbf{0}|$ and hence $f_{\sigma,O}$ also attains its maximum on $\boldsymbol{\theta}^*$. Moreover, for any $\boldsymbol{\theta}$, within $f_{\sigma,O}(\boldsymbol{\theta}) = 1/n \sum_{i=1}^{n} f_{\sigma,|0\rangle\langle 0|_i}(\boldsymbol{\theta})$, the i^{th} summand can be estimated by measuring the i^{th} qubit of $\sigma_{C_T(\boldsymbol{\theta})}$ using Z and estimating the probability that it yields $+1$.

Trainability

In this section, we present our theoretical results regarding cost concentration and barren plateaus of state approximation carried out using global and local observables.

Cost Concentration

We start with a formal definition of cost concentration.

Definition 1. *Let $\sigma \in \mathbb{D}_n$, $W \in \mathbb{H}_n$, and $\boldsymbol{\theta}$ be distributed over a compact parameter space. For any ansatz C, $f_{\sigma,W}$ exhibits cost concentration if*

$$Var_{\boldsymbol{\theta}}\left(f_{\sigma,W}(\boldsymbol{\theta})\right) \in \mathcal{O}\left(\frac{1}{b^n}\right) \tag{4}$$

for some $b > 1$.

From the above definition, we can see that for any VQA objective function that exhibits cost concentration, the output would be exponentially concentrated around $\mathbb{E}_{\boldsymbol{\theta}}\left(f_{\sigma,W}(\boldsymbol{\theta})\right)$. Thus, if there exists a space $\mathbb{V}$ such that $\forall\, \boldsymbol{\theta} \in \mathbb{V}$, $f_{\sigma,W}(\boldsymbol{\theta}) \geq \mathbb{E}_{\boldsymbol{\theta}}\left(f_{\sigma,W}(\boldsymbol{\theta})\right) + \Omega(1)$, then, $\mathbb{V}$ must have $\mathcal{O}(1/b^n)$ measure. Further, cost concentration ensures that the landscape of $f_{\sigma,W}(\boldsymbol{\theta})$ is extremely flat for almost all $\boldsymbol{\theta}$. Thus gradient variations are too small to distinguish from zero using at most $\mathcal{O}(\text{poly}(n))$ samples ensuring there is no strong gradient to follow to reach $\mathbb{V}$. Therefore, to estimate these outputs with meaningful precision, one would require an exponentially large number of samples, or equivalently, copies of σ.

Now, we present our theoretical results regarding cost concentration in learning MPS approximations variationally using C_T. Many trainability results in the literature assume one of two assumptions on the input state (Cerezo et al. 2024); either they are "close" to product states (Pesah et al. 2021; Cerezo et al. 2021b) or they are sparse (Monbroussou et al. 2023; Larocca et al. 2022; Cherrat et al. 2023). Our results also make such assumptions and hence use $h_1(\sigma) \coloneqq \min_{V_1,\ldots,V_n \in \mathbb{U}_2} \|\sigma_{V_1 \otimes \cdots \otimes V_n}\|_1^2$ and $h_2(\sigma) = \min_{\rho_1,\ldots,\rho_n \in \mathbb{D}_1} \|\rho_1 \otimes \cdots \otimes \rho_n - \sigma\|_{\text{tr}}$, to quantify sparsity and proximity to product states respectively.

We start by proving that using global observables for state approximation can give rise to an objective function that exhibits cost concentration.

Theorem 1. *Let $\sigma \in \mathbb{D}_n$ and $C_T^{(n)}$ be an MPS ansatz where each parameterized subcircuit U_i forms a unitary 2-design. Then, we have*

$$Var_{\boldsymbol{\theta}}\left(f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta})\right) \leq \frac{h_1(\sigma)}{4^{n-k-1}}. \tag{5}$$

Hence, for states with $h_1(\sigma) \in \mathcal{O}(4^{n/p})$ with $p > 1$, we see that the upper bound will decrease exponentially.

In contrast, the next theorem shows that the alternative method leveraging local observables provably avoids cost concentration.

Theorem 2. *Let $\sigma \in \mathbb{D}_n$, $O \coloneqq 1/n \sum_{i=1}^n |0\rangle\langle 0|_i$, and $C_T^{(n)}$ be an MPS ansatz, where each parameterized subcircuit U_i forms a unitary 2-design. Then, we have*

$$Var_{\boldsymbol{\theta}}\left(f_{\sigma,O}(\boldsymbol{\theta})\right) \geq \frac{1}{n(2^{2k+1}+4)} - \frac{h_2(\sigma)}{2n}. \tag{6}$$

We note that for any product state σ, $h_2(\sigma) = 0$. More generally, when $h_2(\sigma) \ll 1/(2^{2k}+2)$, the lower bound scales linearly in n and exponentially only in k.

The core idea behind both proofs is to integrate each U_t starting from U_T using standard Haar random integration methods. Typically, this would yield a linear combination of multiple terms, each being an expectation of MPS ansatz circuit outputs with the same observables but defined over $n-T+t-1$ qubit systems and different input states that were *dependent* on the previous state. Thus, naively integrating each U_t one at a time requires integrating a number of terms exponential in T. However, we demonstrate that for the MPS ansatz and the state classes in Theorems 1 and 2, integrating any U_t results in a linear combination of such terms that are *independent* of the previous state, with such state dependency only in the coefficients. This allows us to compute all T integrations using products of T matrices, whose dimension is the number of terms in the linear combination, which in our case, is 2.

Our experimental results discussed later in this work used input states with h_1 and h_2 not necessarily small, suggesting this or similar bounds for a wider variety of states may hold.

Some works in the literature that use the MPS ansatz consider efficient MPS descriptions of states as input, rather than actual quantum states. In such cases, the entire VQA optimization can be efficiently implemented on classical computers using tensor network simulation (Jozsa 2006). Within such methods, the objective function is evaluated exactly, not estimated, so cost concentration is not an issue. However, as we will see in the next section, cost concentration also leads to barren plateaus, which can cause parameter updates to be exponentially small, thus hindering even fully classical optimization protocols.

From Cost Concentration to Barren Plateaus

In this section, we discuss the relationship of Theorems 1 and 2 to barren plateaus. First, we formally define barren plateaus, as per (Arrasmith et al. 2022).

Definition 2. *Let $\sigma \in \mathbb{D}_n$ and let $W \in \mathbb{H}_n$. For any ansatz $C(\boldsymbol{\theta}) = \prod_{p=1}^{t} U_p(\boldsymbol{\theta}_p)$, where $U_p(\boldsymbol{\theta}_p) = \prod_{q=1}^{m} e^{-i\theta_{pq}H_{pq}}$, $\boldsymbol{\theta}_p = [\theta_{p1}\ldots\theta_{pm}]$, $H_{pq} \in \mathbb{H}_n$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_1 \oplus \cdots \oplus \boldsymbol{\theta}_t$, and for any p, q, define*

$$U_p^{(L,q)}\left(\boldsymbol{\theta}_p\right) = \prod_{j=1}^{q-1} e^{-i\theta_{pj}H_{pj}}, \tag{7}$$

$$U_p^{(R,q)}\left(\boldsymbol{\theta}_p\right) = \prod_{j=q+1}^{m} e^{-i\theta_{pj}H_{pj}}. \tag{8}$$

Then, $f_{\sigma,W}$ exhibits a barren plateau if $\forall\, p, q$ satisfying $1 \leq p \leq t, 1 \leq q \leq m$, we have

$$Var_{\boldsymbol{\theta}}\left(\partial_{\theta_{pq}} f_{\sigma,W}(\boldsymbol{\theta})\right) \in \mathcal{O}\left(\frac{1}{b^n}\right), \tag{9}$$

for some constant $b > 1$, where $\partial_{\theta_{pq}} f_{\sigma,W}(\boldsymbol{\theta})$ is it's partial derivative with respect to θ_{pq} and $U_1, \ldots U_{p-1}, U_{p+1}, \ldots U_t$, along with one of $U_p^{(L,q)}$ or $U_p^{(R,q)}$ are distributed according to the Haar measure and θ_{pq} is distributed uniformly.

Just as cost concentration makes the outputs of most inputs exponentially concentrated, barren plateaus cause most partial derivatives to be exponentially small, since $\mathbb{E}_{\boldsymbol{\theta}}\left(\partial_{\theta_{pq}}\text{tr}(W\sigma_{C(\boldsymbol{\theta})})\right) = 0\ \forall\ p, q$ (Cerezo et al. 2021b). This means that estimating these derivatives will require exponential resources, and in most cases, the parameter updates that gradient-based classical optimizers make will be exponentially small.

Next, we will use Theorem 1 to demonstrate that employing the MPS ansatz for learning state approximations leads to barren plateaus when global observables are used.

Corollary 1. *Let $\sigma \in \mathbb{D}_n$ and $C_T^{(n)}$ be an MPS ansatz. Then, $f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}$ exhibitis barren plateaus if $h_1(\sigma) \in \mathcal{O}(4^{n/p})$.*

Similarly, we extend Theorem 2 to demonstrate that using the MPS ansatz with local observables prevents barren plateaus.

Corollary 2. *Let $\sigma \in \mathbb{D}_n$, $C_T^{(n)}$ be an MPS ansatz, and $O \coloneqq 1/n \sum_{i=1}^{n} |0\rangle\langle 0|_i$. Then, $f_{\sigma,O}$ does not exhibit barren plateaus if $h_2(\sigma) \ll 1/(2^{2k}+2)$.*

Towards Classical Simulation Through Effective Subspaces

In this section, we discuss the possibility of designing an efficient classical algorithm capable of simulating state approximation VQAs involving MPS ansatzes and local observables, using very few copies of the input quantum state.

The idea builds on the conjecture from (Cerezo et al. 2024) which says that any objective function avoiding cost concentration exhibits effective subspaces, a property useful for designing classical simulation algorithms with minimal quantum resources. Our simulations demonstrate that objective functions involving MPS ansatz and local observables, which we previously proved to avoid cost concentration, indeed exhibit effective subspaces within the Pauli basis, further supporting this conjecture.

Note that in this work, we do not present an explicit algorithm for the aforementioned classical simulation, but rather present evidence that such a protocol could exist. First, we introduce effective subspaces as outlined in (Cerezo et al. 2024).

Effective Subspace

Let $C(\boldsymbol{\theta})$ be an n-qubit ansatz and let $W \in \mathbb{H}_n, \sigma \in \mathbb{D}_n$. Effective subspaces are loosely defined as follows:

Definition 3. *(Cerezo et al. 2024) For any orthonormal basis $\mathbb{K} = \{K_1, K_2, \dots K_{4^n}\}$ of $\mathbb{C}^{2^n \times 2^n}$, and for any $\boldsymbol{\theta}$, define a distribution $\mathcal{P}_{\boldsymbol{\theta},W,\mathbb{K}}$ over $\mathbb{K}$ as*

$$\mathcal{P}_{\boldsymbol{\theta},W,\mathbb{K}}(K_j) = \frac{f_{K_j,W}(\boldsymbol{\theta})^2}{\|W\|_2^2}. \tag{10}$$

An ansatz and observable combination exhibits an effective subspace if there exists a basis $\mathbb{K}$ such that for almost all $\boldsymbol{\theta}$, $\mathcal{P}_{\boldsymbol{\theta},W,\mathbb{K}}(K_j)$ is large only for K_j in a subset $\mathbb{K}_s \subset \mathbb{K}$, that is independent of $\boldsymbol{\theta}$ and has $|\mathbb{K}_s| \in \mathcal{O}(poly(n))$.

Note that the elements of $\mathbb{K}$ are not restricted to quantum states. Also, in appropriate contexts, we denote $\mathcal{P}_{\boldsymbol{\theta},W,\mathbb{K}}(K_j)$ as $\mathcal{P}_{\boldsymbol{\theta},W}(K_j)$ as the dependency on $\mathbb{K}$ is implicitly understood from the context.

In (Cerezo et al. 2024), it is conjectured, with evidence, that all ansatz-observable combinations that have been shown to provably avoid barren plateaus exhibit an effective subspace, at least for some subset of input states. Popular examples involving shallow ($\mathcal{O}(\log n)$-depth) ansatzes include HEA-local observable and the QCNN-local observable combinations. In both these cases, the basis $\mathbb{K}$ can be $\mathbb{P}_n$. The presence of effective subspaces means that if we estimate $\text{tr}(K\sigma)\ \forall\ K \in \mathbb{K}_s$ as a preprocessing step, and if we can classically compute $\text{tr}(EW_{C(\boldsymbol{\theta})^\dagger})\ \forall\ K \in \mathbb{K}_s$ and $\forall\ \boldsymbol{\theta}$ efficiently, then in many cases, $f_{\sigma,W}(\boldsymbol{\theta})$ can be classically estimated with good precision, because

$$\begin{aligned} f_{\sigma,W}(\boldsymbol{\theta}) = \text{tr}\left(W\sigma_{C(\boldsymbol{\theta})}\right) &= \text{tr}\left(W_{C(\boldsymbol{\theta})^\dagger}\sigma\right) \\ &= \sum_{K\in\mathbb{K}} \text{tr}\left(KW_{C(\boldsymbol{\theta})^\dagger}\right)\text{tr}(K\sigma), \end{aligned}$$

and if most $\text{tr}\left(KW_{C(\boldsymbol{\theta})^\dagger}\right)$ is large only for those $K \in \mathbb{K}_s$, then

$$f_{\sigma,W}(\boldsymbol{\theta}) \approx \sum_{K\in\mathbb{K}_s} \text{tr}\left(KW_{C(\boldsymbol{\theta})^\dagger}\right)\text{tr}(K\sigma). \tag{11}$$

This is the underlying principle behind designing classical simulations using effective subspaces.

When it comes to the classical simulation of $f_{\sigma,O,C_T}(\boldsymbol{\theta}) = 1/n\sum_i f_{\sigma,|0\rangle\langle 0|_i,C_T}(\boldsymbol{\theta})$, it is sufficient to be able to classically estimate each $f_{\sigma,|0\rangle\langle 0|_i}(\boldsymbol{\theta})$ efficiently. From Figure 1 (a), it is easy to see that for any product state σ, among these n terms, the hardest to estimate are $f_{\sigma,|0\rangle\langle 0|_i}(\boldsymbol{\theta})$ for $i \in \{n-k+1, \dots, n\}$, because for all other i, at least one subcircuit within $C_T(\boldsymbol{\theta})$ will be canceled leaving an expression of the same form on fewer qubits. When q subcircuits are canceled, at least $4^{n-q}(4^q-1)$ outcomes of $\mathcal{P}_{\boldsymbol{\theta},|0\rangle\langle 0|_i,\mathbb{P}_n}$ will be zero for any $\boldsymbol{\theta}$, making the distribution very concentrated. Also, for any $i, j \in \{n-k+1, \dots, n\}$, $\mathbb{E}_{\boldsymbol{\theta}}(f_{\sigma,|0\rangle\langle 0|_i}(\boldsymbol{\theta})) = \mathbb{E}_{\boldsymbol{\theta}}(f_{\sigma,|0\rangle\langle 0|_j}(\boldsymbol{\theta}))$ (proved in the Appendix). Therefore, we focus on $f_{\sigma,|0\rangle\langle 0|_n}(\boldsymbol{\theta})$ and aim to show that the C_T-$|0\rangle\langle 0|_n$ combination also exhibits an effective subspace with $\mathbb{K} = \mathbb{P}_n$.

C-$\mathbb{K}$ Norm

Now, we introduce a norm that can be used to measure how concentrated the distributions $\mathcal{P}_{\boldsymbol{\theta},W}$ would be, for typical values of $\boldsymbol{\theta}$. Given any discrete distribution (probability vector) $\mathcal{P}$, $\|\mathcal{P}\|_2$ can be used to measure how concentrated the distribution is. A higher $\|\mathcal{P}\|_2$ indicates that the distribution is concentrated among a few outcomes with high probability. Hence, we can use the 2-norm of the distributions $\mathcal{P}_{\boldsymbol{\theta},W}$ to assess how concentrated these distributions are. So, we define the $\mathbb{K}$-norm (in $\mathbb{H}_n$) as this 2-norm, that is

$$\|W\|_{\mathbb{K}} \coloneqq \frac{1}{\|W\|_2}\left[\sum_{K\in\mathbb{K}} \text{tr}(KW)^4\right]^{1/4} \tag{12}$$

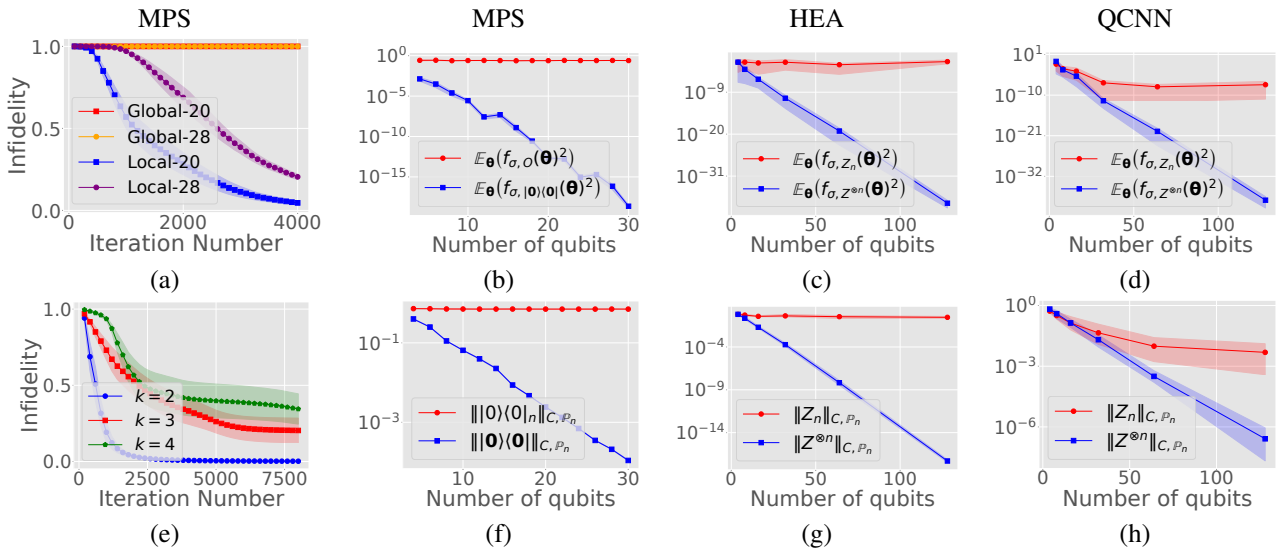

Figure 2: Simulation results of state approximation, C-$\mathbb{K}$ norms and second moments.

Note that for any $\boldsymbol{\theta}$, $\|W_{C(\boldsymbol{\theta})^\dagger}\|_{\mathbb{K}}$ is simply the 2-norm of the distribution $\mathcal{P}_{\boldsymbol{\theta},W}$. We first prove the following result regarding the cost of computing $\left\|W_{C(\boldsymbol{\theta})^\dagger}\right\|_{\mathbb{K}}$ for any $\boldsymbol{\theta}$.

Theorem 3. *For any n-qubit quantum circuit V, where the qubits are arranged in a line, let $R_V = \max_i R_{V,i}$, where $R_{V,i}$ is the number of 2-qubit gates being applied on any qubits j,k such that $j \leq i \leq k$. Then, for any product observable W, $\|W_{V^\dagger}\|_{\mathbb{K}}$ can be classically evaluated with cost $\mathcal{O}\left(2^{R_V}\right)$.*

Although the MPS ansatz is defined using k-qubit parameterized gates, Theorem 3 concerns only 2-qubit gates because, in practice, we always decompose each k-qubit unitary into smaller 2-qubit parameterized gates. As an example, in all our simulations, all k-qubit unitaries are HEAs (cf. Figure 1 (b)).

From Figure 1 (a), we can see that R_V is independent of n. Typically, it scales as $\mathcal{O}(\text{poly}(k))$ meaning that the cost of evaluating $\left\|W_{C_T(\boldsymbol{\theta})^\dagger}\right\|_{\mathbb{K}}$ will be $\mathcal{O}\left(2^{\text{poly}(k)}\right)$.

Now, as mentioned earlier, we would like to analyze $\left\||0\rangle\langle 0|_{nC_T(\boldsymbol{\theta})^\dagger}\right\|_{\mathbb{K}}$ for typical values of $\boldsymbol{\theta}$. Hence, we introduce the C-$\mathbb{K}$ norm in the following theorem.

Theorem 4. *For any parameterized circuit C, and an orthonormal basis $\mathbb{K}$ of $\mathbb{C}^{2^n\times 2^n}$, define*

$$\|W\|_{C,\mathbb{K}} := \int_{\boldsymbol{\theta}} \left\|W_{C(\boldsymbol{\theta})^\dagger}\right\|_{\mathbb{K}} d\boldsymbol{\theta}, \tag{13}$$

for any $W \in \mathbb{H}_n$. Then, $\|\cdot\|_{C,\mathbb{K}}$ is a norm on $\mathbb{H}_n$.

This norm can be numerically estimated by sampling various parameter vectors $\boldsymbol{\theta}$ and taking the average of their $\|W_{C(\boldsymbol{\theta})^\dagger}\|_{\mathbb{K}}$. Intuitively, if $\|W\|_{C,\mathbb{K}}$ remains constant or reduces only polynomially with respect to n, then we can expect the C-W combination to exhibit an effect subspace since the distribution $\mathcal{P}_{\boldsymbol{\theta},W}$ is defined over 4^n outcomes. Conversely, if $\|W\|_{C,\mathbb{K}}$ reduces exponentially with respect to n, then the C-W combination need not exhibit one.

We first numerically test this hypothesis on some instances where the presence and absence of effective subspaces are known. To do this, we choose two ansatzes; shallow HEA and QCNN, in combination with local and global observables.

It is known that effective subspaces exist when both these ansatzes are used in combination with local observables. The results (presented in Figure 2) strongly support the hypothesis and hence we carry out the same experiments for C_T. We discuss these simulation results in detail in the following section.

Finally, the effective subspace for C_T-$|0\rangle\langle 0|_n$ can be roughly identified by considering the cancellation of subcircuits. Typically, the probability $\mathcal{P}_{C_T,|0\rangle\langle 0|_n,\mathbb{P}_n}(P_j)$ increases when more subcircuits are canceled within its expression, as this forces some qubits to have no circuits being acted on them and hence contribute the maximum that any qubit can to the expectation. This is also true for shallow HEA and QCNN ansatzes when used with local observables, where higher probabilities are associated with 1-local Paulis, regardless of the position of its non-local component. For Paulis with a higher locality, one can always find an upper bound on the total number of non-canceled subcircuits that is independent of n and dependent only on the locality. However, for the C_T-$|0\rangle\langle 0|_n$ combination, the position of the non-local part of the Pauli is crucial. The closer it is to the last qubit, the more subcircuits are canceled, resulting in higher probabilities. Similarly, if the non-local component is on the first qubit, unlike the other ansatzes, even a 1-local observable can have no subcircuits getting canceled in the expression of the probability. Thus, the concentration of probabilities should be towards Paulis where non-local components occur near the last qubit. This hypothesis is also validated using experiments discussed in the next section.

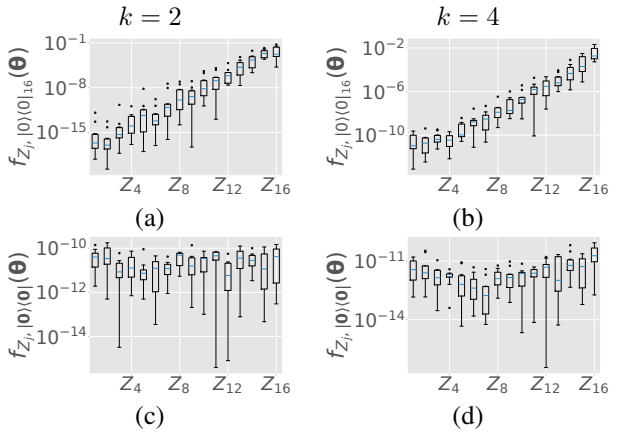

Figure 3: Boxplots of distributions $\mathcal{P}_{\boldsymbol{\theta},|0\rangle\langle 0|_{16},\mathbb{P}_{16}}$ in (a) and (b), and distributions $\mathcal{P}_{\boldsymbol{\theta},|\mathbf{0}\rangle\langle\mathbf{0}|,\mathbb{P}_{16}}$ in (c) and (d), with subcircuit widths $k = 2, 4$.

Simulation Results

In this section, we discuss and present the numerical simulations that we have conducted as part of this work. The main aims of the simulations are threefold: visualize the impact of Theorems 1 and 2 using learning curves, argue that similar results could also hold for most states not necessarily satisfying the criteria mentioned in these theorems and demonstrate the presence (absence) of effective subspaces when MPS ansatzes are used with local (global) observables. The structure of all two-qubit subcircuits used in Figure 2 is given in Figure 1 (d).

We start with the learning curves presented in Figures 2 (a) and (e). Here, we have carried out state approximation using the MPS ansatz with subcircuit width 2, and target state $|\mathbf{0}\rangle\langle\mathbf{0}|$. The optimization algorithm used here is SPSA (Spall 1992), where the converging sequences are $a_j = c_j = 0.4$ and all parameters are initialized uniformly from $[0, \pi/2]$. The x-axis and y-axis represent the iteration number and corresponding *infidelity*, defined as $1 - F$, respectively. In (a), we have plotted results for $n = 20, 28$, with $k = 2$. We can see global observables hindering the optimization. In (e), we plot the results of similar experiments carried out for $n = 12$ and local observables, but with $k = 2, 3$ and 4. The subcircuits are HEA with depth k. From this, we can see that increasing k negatively impacts the optimization.

Now, we move on to Figure 2 (b). Here, the x and y axes represent the number of qubits and the estimated second moments of the objective functions $f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta})$ and $f_{\sigma,O}(\boldsymbol{\theta})$ averaged over 5 input states randomly generated using HEA ansatz of depth $\lfloor \log n \rfloor$, with the single qubit gates being Haar random. The subcircuit used here is HEA with width and depth $\lfloor \log n \rfloor$. We can see global observables inducing cost concentration, and local observables avoiding it, even though the input states do not necessarily satisfy the conditions required as per Theorems 1 and 2.

Next, we move on to Figures 2 (c,d,f,g,h). The idea here is to show that the C-$\mathbb{K}$ norm can be used to detect the presence of effective subspace. From (Cerezo et al. 2024), we know that shallow HEA and QCNN ansatzes exhibit effective subspaces when used in combination with local observables. This can be seen from the plots (c), (d), (g), and (h). In (c) and (d), we have plotted the estimated second moments of the objective functions $f_{\sigma,Z^{\otimes n}}(\boldsymbol{\theta})$ and $f_{\sigma,Z_n}(\boldsymbol{\theta})$, averaged over 5 states generated in the same manner as in the previous experiment, for the shallow HEA and QCNN ansatzes respectively. In (g) and (h), we have plotted estimated C-$\mathbb{P}_n$ norms of these combinations. From these four plots, we can see that the C-$\mathbb{P}_n$ norms are behaving as we expected. So, in Figure 2 (f), we plot the C_T-$\mathbb{P}_n$ norms, with subcircuits having the same structure as in (b). The observable is chosen to be $|0\rangle\langle 0|_n$ since as mentioned earlier when it comes to classical simulation, it suffices to estimate the C_T-$\mathbb{P}_n$ norms of $|0\rangle\langle 0|_n$. We can see that when we use local observables, we get exponentially high C_T-$\mathbb{P}_n$ norms, thus suggesting the presence of effective subspaces.

As noted at the end of Section C-$\mathbb{K}$ Norm, this effective subspace is the one that is spanned by Paulis whose non-identity components are near the last qubit. This is experimentally verified using 16-qubit simulations whose results are shown in Figure 3. In (a) and (b) we plot a portion of the distribution $\mathcal{P}_{\boldsymbol{\theta},|0\rangle\langle 0|_{16},\mathbb{P}_{16}}$ with subcircuits being HEA built using 2 qubit Haar random gates, with depths and widths of $k = 2, 4$. Although there are 4^{16} possible outcomes, we focus on 16, specifically the 1-local Paulis $\{Z_i \mid i = 1, \ldots, n\}$, shown on the x-axis. In these figures, boxplots of probabilities $\mathcal{P}_{\boldsymbol{\theta},|0\rangle\langle 0|_{16}}(Z_i)$, computed across 10 different $\boldsymbol{\theta}$ values are plotted. We can see that as the Z component in the observables on the x-axis is closer to the last qubit, the probability is exponentially higher. In (d) and (e), similar experiments are carried out for the distribution $\mathcal{P}_{\boldsymbol{\theta},|\mathbf{0}\rangle\langle\mathbf{0}|,\mathbb{P}_{16}}$, but we notice no such concentration of probabilities, indicating the absence of effective subspaces.

Conclusion and Future Direction

In this work, we have introduced new results regarding trainability and classical simulability of learning MPS approximations of quantum states variationally. We have proven that the usage of global observables forces the variance of the objective function and all its partial derivatives to be exponentially small in the number of qubits, while the usage of local observables avoids this. Moreover, we have demonstrated that using this ansatz with local observables reveals effective subspaces within the Pauli basis, paving the way for a potential classical simulation of the MPS ansatz.

For future directions, we aim to generalize and enhance our results by theoretically proving similar trainability results when multiple layers of C_T are used, extending the current proofs to all quantum states, and theoretically analyzing the effective subspaces. Additionally, we plan to develop an efficient classical simulation with rigorous performance guarantees.

Acknowledgments

This work was partially supported by the Australian Research Council (Grant No: DP220102059) and the National Science Foundation of China (12071271). AB was partially supported

by the Sydney Quantum Academy PhD scholarship. HP acknowledges the Centre for Quantum Software and Information at the University of Technology Sydney for hosting him as a visiting scholar and the support of the University of Sydney Nano Institute (Sydney Nano). YF was partly supported by the National Natural Science Foundation of China under Grant No. 92465202.

References

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.

Arute, F.; Arya, K.; Babbush, R.; Bacon, D.; Bardin, J.; Barends, R.; Biswas, R.; Boixo, S.; Brandao, F.; Buell, D.; Burkett, B.; Chen, Y.; Chen, J.; Chiaro, B.; Collins, R.; Courtney, W.; Dunsworth, A.; Farhi, E.; Foxen, B.; Fowler, A.; Gidney, C. M.; Giustina, M.; Graff, R.; Guerin, K.; Habegger, S.; Harrigan, M.; Hartmann, M.; Ho, A.; Hoffmann, M. R.; Huang, T.; Humble, T.; Isakov, S.; Jeffrey, E.; Jiang, Z.; Kafri, D.; Kechedzhi, K.; Kelly, J.; Klimov, P.; Knysh, S.; Korotkov, A.; Kostritsa, F.; Landhuis, D.; Lindmark, M.; Lucero, E.; Lyakh, D.; Mandrà, S.; McClean, J. R.; McEwen, M.; Megrant, A.; Mi, X.; Michielsen, K.; Mohseni, M.; Mutus, J.; Naaman, O.; Neeley, M.; Neill, C.; Niu, M. Y.; Ostby, E.; Petukhov, A.; Platt, J.; Quintana, C.; Rieffel, E. G.; Roushan, P.; Rubin, N.; Sank, D.; Satzinger, K. J.; Smelyanskiy, V.; Sung, K. J.; Trevithick, M.; Vainsencher, A.; Villalonga, B.; White, T.; Yao, Z. J.; Yeh, P.; Zalcman, A.; Neven, H.; and Martinis, J. 2019. Quantum Supremacy using a Programmable Superconducting Processor. *Nature*, 574: 505–510.

Barthel, T.; and Miao, Q. 2024. Absence of barren plateaus and scaling of gradients in the energy optimization of isometric tensor network states. arXiv:2304.00161.

Ben-Dov, M.; Shnaiderov, D.; Makmal, A.; and Torre, E. G. D. 2022. Approximate encoding of quantum states using shallow circuits. *npj Quantum Information*, 10: 1–8.

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.; Larocca, M.; García-Martín, D.; Diaz, N. L.; Braccia, P.; Fontana, E.; Rudolph, M. S.; Bermejo, P.; Ijaz, A.; Thanasilp, S.; Anschuetz, E. R.; and Holmes, Z. 2024. Does provable absence of barren plateaus imply classical simulability? Or, why we need to rethink variational quantum computing. arXiv:2312.09121.

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).

Cherrat, E. A.; Raj, S.; Kerenidis, I.; Shekhar, A.; Wood, B.; Dee, J.; Chakrabarti, S.; Chen, R.; Herman, D.; Hu, S.; Minssen, P.; Shaydulin, R.; Sun, Y.; Yalovetzky, R.; and Pistoia, M. 2023. Quantum Deep Hedging. *Quantum*, 7: 1191.

Chow, J.; Dial, O.; and Gambetta, J. 2021. IBM Quantum breaks the 100-qubit processor barrier. *IBM Research Blog*.

Cirac, J. I.; Pérez-García, D.; Schuch, N.; and Verstraete, F. 2021. Matrix product states and projected entangled pair states: Concepts, symmetries, theorems. *Rev. Mod. Phys.*, 93: 045003.

Cramer, M.; Plenio, M. B.; Flammia, S. T.; Somma, R.; Gross, D.; Bartlett, S. D.; Landon-Cardinal, O.; Poulin, D.; and Liu, Y.-K. 2010. Efficient quantum state tomography. *Nature Communications*, 1(1).

da Silva, M. P.; Ryan-Anderson, C.; Bello-Rivas, J. M.; Chernoguzov, A.; Dreiling, J. M.; Foltz, C.; Frachon, F.; Gaebler, J. P.; Gatterman, T. M.; Grans-Samuelsson, L.; Hayes, D.; Hewitt, N.; Johansen, J.; Lucchetti, D.; Mills, M.; Moses, S. A.; Neyenhuis, B.; Paz, A.; Pino, J.; Siegfried, P.; Strabley, J.; Sundaram, A.; Tom, D.; Wernli, S. J.; Zanner, M.; Stutz, R. P.; and Svore, K. M. 2024. Demonstration of logical qubits and repeated error correction with better-than-physical error rates. arXiv:2404.02280.

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).

Farhi, E.; Goldstone, J.; and Gutmann, S. 2014. A Quantum Approximate Optimization Algorithm. arXiv:1411.4028.

Gambetta, J. 2022. Quantum-centric supercomputing: The next wave of computing. *IBM Research Blog*.

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).

Gard, B. T.; Zhu, L.; Barron, G. S.; Mayhall, N. J.; Economou, S. E.; and Barnes, E. 2020. Efficient symmetry-preserving state preparation circuits for the variational quantum eigensolver algorithm. *npj Quantum Information*, 6(1).

Harrow, A. W.; and Low, R. A. 2009. Random Quantum Circuits are Approximate 2-designs. *Communications in Mathematical Physics*, 291(1): 257–302.

Havlíček, V.; Córcoles, A. D.; Temme, K.; Harrow, A. W.; Kandala, A.; Chow, J. M.; and Gambetta, J. M. 2019. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747): 209–212.

Jozsa, R. 2006. On the simulation of quantum circuits. arXiv:quant-ph/0603163.

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.

Lin, S.-H.; Dilip, R.; Green, A. G.; Smith, A.; and Pollmann, F. 2021. Real- and Imaginary-Time Evolution with Compressed Quantum Circuits. *PRX Quantum*, 2(1).

Liu, Z.; Yu, L.-W.; Duan, L.-M.; and Deng, D.-L. 2022. Presence and Absence of Barren Plateaus in Tensor-Network Based Machine Learning. *Physical Review Letters*, 129(27).

Madsen, L. S.; Laudenbach, F.; Askarani, M. F.; Rortais, F.; Vincent, T.; Bulmer, J. F. F.; Miatto, F. M.; Neuhaus, L.; Helt, L. G.; Collins, M. J.; Lita, A. E.; Gerrits, T.; Nam, S. W.; Vaidya, V.; Menotti, M.; Dhand, I.; Vernon, Z.; Quesada, N.;

and Lavoie, J. 2022. Quantum computational advantage with a programmable photonic processor. *Nature*, 606: 75 – 81.

Matos, G.; Johri, S.; and Papić, Z. 2021. Quantifying the Efficiency of State Preparation via Quantum Variational Eigensolvers. *PRX Quantum*, 2: 010309.

Mitarai, K.; Negoro, M.; Kitagawa, M.; and Fujii, K. 2018. Quantum circuit learning. *Phys. Rev. A*, 98: 032309.

Monbroussou, L.; Landman, J.; Grilo, A. B.; Kukla, R.; and Kashefi, E. 2023. Trainability and Expressivity of Hamming-Weight Preserving Quantum Circuits for Machine Learning. arXiv:2309.15547.

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).

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. *Phys. Rev. X*, 11: 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. *Phys. Rev. A*, 101: 032310.

Rieser, H.-M.; Köster, F.; and Raulf, A. P. 2023. Tensor networks for quantum machine learning. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 479(2275).

Rudolph, M. S.; Chen, J.; Miller, J.; Acharya, A.; and Perdomo-Ortiz, A. 2022. Decomposition of matrix product states into shallow quantum circuits. *Quantum Science and Technology*, 9.

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

Zhao, C.; and Gao, X.-S. 2021. Analyzing the barren plateau phenomenon in training quantum neural networks with the ZX-calculus. *Quantum*, 5: 466.

Zhong, H.-S.; Wang, H.; Deng, Y.-H.; Chen, M.-C.; Peng, L.-C.; Luo, Y.-H.; Qin, J.; Wu, D.; Ding, X.; Hu, Y.; Hu, P.; Yang, X.-Y.; Zhang, W.-J.; Li, H.; Li, Y.; Jiang, X.; Gan, L.; Yang, G.; You, L.; Wang, Z.; Li, L.; Liu, N.-L.; Lu, C.-Y.; and Pan, J.-W. 2020. Quantum computational advantage using photons. *Science*, 370(6523): 1460–1463.