Classical Simulability and Trainability of Quantum Machine Learning

by

Afrad Muhamed Basheer

A dissertation submitted in fulfillment
of the requirements for the degree of
Doctor of Philosophy

University of Technology Sydney
2024

Certificate of Original Authorship

I, Afrad Muhamed Basheer, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed prior to publication.

Afrad Muhamed Basheer
24-09-2024

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Prof. Yuan Feng and Prof. Sanjiang Li, for their invaluable guidance, support, and encouragement throughout my PhD journey. Their expertise and insights have been instrumental in shaping my research and bringing this thesis to fruition. I am also grateful to Distinguished Prof. Mingsheng Ying, who stepped in as my co-supervisor towards the end of my PhD, providing essential support and guidance during a crucial phase of my research.

Financially, my research was primarily supported by my supervisors' Australian Research Council research grant. I was also supported by the University of Technology Sydney (UTS) International Research Scholarship and the Sydney Quantum Academy (SQA) supplementary scholarship, which provided additional funding and resources.

I am profoundly thankful to A/Prof. Christopher Ferrie, who has been a co-author on all the papers included in this thesis, and has supported me in numerous ways beyond our collaborations. His mentorship and generous contributions have significantly enriched my academic experience.

I would also like to extend my heartfelt thanks to my childhood friend Afham, who also embarked on a PhD journey in the same field at the same university. His academic and non-academic support has been invaluable, and I am grateful for the many ways he has helped me throughout this journey.

A special thank you goes to Dr. Hakop Pashayan for his guidance and collaboration during my PhD, including co-authoring one of my papers. His advice and support at critical moments have been greatly appreciated.

I would like to thank Christian Bertoni, Guangxi Li, and Dr. Richard Kueng for their help with academic discussions which have been immensely beneficial to my work.

I also wish to acknowledge the administrative staff of UTS and SQA, who provided assistance at various stages of my PhD. Their support has been instrumental in navigating the many logistical aspects of my research journey.

During my PhD, I had the pleasure of collaborating with Dr. Eric Howard, Prof. Gavin Brennen, and Christopher Tam from BTQ, which greatly expanded my perspectives in quantum computing. I am especially grateful for the opportunity to work with Dr. Eric Howard and Iftekher Chowdhury on experimental physics projects, even though these works are not part of my thesis.

Finally, I would like to acknowledge all my friends, family, and colleagues who have supported me throughout this journey. While I choose not to name individuals to bind the space complexity of this section and avoid the risk of leaving anyone out, I deeply appreciate their encouragement and for being with me every step of the way.

Abstract

Variational Quantum Algorithms (VQA) form an important class of quantum machine learning and optimization algorithms, with potential applications in supervised learning, combinatorial optimization, chemical simulation, dimensionality reduction, etc. At its core, it is a class of optimization algorithms where parameterized quantum circuits are used to estimate functions that are typically extremely expensive for classical computers, and algorithms such as gradient descent are used to optimize the parameters classically.

Given the limited availability of quantum devices in the near future, it is essential to minimize their usage within VQAs. Recent research has highlighted two challenges that could increase quantum device demands: trainability issues akin to vanishing gradients such as barren plateaus and sample complexity problems, often exacerbated by practical demands such as circuit design, hyperparameter tuning, etc. This thesis presents new algorithms and theoretical insights to address these challenges, enhancing the efficiency of VQAs.

The contributions of this thesis can be broadly categorized into three parts. The first part introduces a novel training algorithm tailored for shallow alternating layered VQAs, called Alternating Layered Shadow Optimization (ALSO). By harnessing classical shadows of quantum input data, the algorithm achieves exponential reductions in quantum resources required for training. The optimization part can be completely carried out on classical computers efficiently with rigorous performance guarantees. Moreover, ALSO is easier to implement compared to standard VQA training methods, requiring only single qubit measurements and classical post-processing. We also experimentally demonstrate orders of magnitude improvement for ALSO in common quantum machine learning applications.

Building upon these advancements, the second part addresses similar computational demands of a different class of VQAs that can involve almost any shallow circuit and low Frobenius norm observables. This new training algorithm, called Ansatz Independent Shadow Optimization, extends the applicability of shadow tomography for VQAs to a diverse range of ansatzes, showcasing exponential savings in quantum resources. Beyond rigorous performance guarantees, we experimentally demonstrated successful applications in state preparation and variational quantum circuit synthesis, validating its superiority over traditional training methods.

The third part delves into the notorious phenomenon of barren plateaus observed in VQAs tasked with finding weakly entangled state approximations. Through theoretical analysis and rigorous experimentation, we elucidate how the choice of global versus local observables impacts gradient scaling, with exponentially low gradients and cost functions being present and absent in the former and latter scenarios respectively. On top of that, we also discuss how one could potentially be able to classically simulate the local observable version, with minimal usage of quantum resources. Moreover, all our results and claims are experimentally validated across different scenarios.

Collectively, these contributions underscore important advancements in VQAs, particularly in the context of near-term quantum devices, paving the way for their integration into diverse quantum machine learning applications.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Quantum computing is a new paradigm of computing that can potentially revolutionize high-performance computing in the upcoming future [Cer+21a; Cle+97; Bia+16; Ben+19]. The core idea is built upon the fact that quantum mechanics is a model that is inherently hard to simulate using normal, so-called, classical computers, intuitively implying that nature can carry out certain computations that we cannot hope to emulate using classical computers in a reasonable time. Unlike classical computers, quantum computers use quantum resources such as quantum bits, also known as, qubits, to store quantum information described using quantum states. Similarly, quantum gates are used to manipulate these quantum states, and information is read out (observed) using quantum measurements (quantum observables), which typically output a classical bit string according to some output distribution.

The early '90s witnessed the birth of a few quantum algorithms that can solve certain practically insignificant problems exponentially faster than any known classical algorithm [DJ92; BV97; Sim97]. That is, quantum algorithms can be used to solve these problems by consuming quantum resources that are exponentially less than the classical resources that *any* classical algorithm will consume.

The poster boy of quantum algorithms, Shor's algorithm for polynomial-time integer

factorization was discovered in 1994 [Sho94]. This turned many heads as integer factorization was a problem that, even now, does not have a "classical" solution with complexity polynomial in the number of bits involved [MOV96; Yas+16]. This inherent hardness of the problem made it convenient for applications in cryptography [MOV96], which is a field that features extensively in our daily lives.

Lov Grover demonstrated that quantum algorithms could search in an unstructured database with a quadratic speedup over classical methods [Gro96]. This foundational work has since been extended to various applications, including maxima finding [DH99; Dür+06] and quantum analogs of random walks and their associated uses [MB18; Chi+03; Kem03; Chi09].

While these algorithms offer advantages over their classical counterparts, practical large-scale implementation may still be years away, as current devices possess only a limited number of qubits, which are also highly prone to errors. Hence along with advancements in research in quantum hardware, the 2000s and 2010s witnessed significant effort in demonstrating quantum supremacy [LBR17]. That is, to physically implement a quantum algorithm on currently available small noisy quantum computers, also called Noisy Intermediate Scale Quantum (NISQ) devices [Pre18], and produce results in reasonable time that would take an extremely large amount of time and resources on a classical computer. This was recently accomplished in works such as [Aru+19; Zho+20a; Mad+22], where the supremacy of quantum computers over classical computers was demonstrated for certain specific sampling problems.

Now, there is considerable effort being put into research aiming to demonstrate similar advantages for practically useful problems. Although this is yet to be demonstrated, several areas are being investigated, such as physical and chemical simulation [Kas+08; Per+14; Bau+20; Til+22b], optimization [FGG14; Cha+20; BS17], machine learning [Hav+19; Hua+22; Ker+19; MP15], etc. The latter use cases are particularly significant due to the ubiquitous nature of their applications.

Machine learning (ML), a subset of artificial intelligence (AI), involves algorithms

that enable computers to learn from and make decisions based on data. Its impact on our daily lives is huge as its applications include personalized recommendations [Zha+19; KBV09; Bob+13], fraud detections [Nga+11; Kou+04; Phu+10], spam filters [Sak+03; GC09; Li+19], AI assistants [Hoy18; MCG16; Bér+24], disease diagnosis [Top19; Jia+17; Est+17], etc. However, in the upcoming era of *bigger* data, machine learning will face significant challenges despite the abundance of available information. Handling enormous datasets requires substantial computational power and memory, which can strain existing infrastructure and slow down processing times.

This is the reason why several research groups are looking at ML as an area where one could achieve practical and useful quantum supremacy. Several Quantum Machine Learning (QML) protocols and algorithm designs have already been put forward including quantum SVM [Bia+16], quantum Boltzmann machines [Bia+16], quantum clustering [ABG07], quantum neural networks [Bia+16; Ben+19], quantum persistent homology [LGZ14], quantum transformers [LF24; LZW23], quantum reinforcement learning [Jer+21], quantum natural language processing [Mei+20; Coe+20], quantum kNN [ABG20; WKS15; Rua+17; CGZ15], quantum algorithms for knowledge graphs [MWT20], etc.

Within ML, due to many reasons such as the complexity of the models, the vastness of potential data inputs, and the intricate nature of real-world problems, advantages over other algorithms are often demonstrated *empirically* on benchmark datasets [Vap00; Dom12; Mit97]. However, such empirical evaluation of QML models is infeasible as most of the aforementioned QML algorithms require quantum devices with thousands of qubits and noiseless functioning, placing them firmly out of the reach of current NISQ devices.

The most hopeful class of algorithms is Variational Quantum Algorithms (VQAs) [Cer+21c; Ben+19]. At their core, VQAs are general-purpose optimization algorithms where, we use quantum computers to estimate objective functions involving qubits and parameterized quantum circuits (also called *ansatzes*, typically designed in a layerwise fashion akin to how neural networks are designed), and update the parameters of these func-

tions classically towards their optimum using algorithms such as gradient descent [Rud17], ADAM [KB17], etc. Many such functions are notoriously hard to evaluate on classical devices [Hav+19], thus opening the door towards potential practical quantum supremacy. Popular examples include variational quantum eigensolver [Per+14], quantum approximate optimization algorithm (QAOA) [FGG14], quantum support vector machines [Hav+19], quantum autoencoder [ROA17], quantum neural networks [Ben+19], with latter three being examples of applications in machine learning.

One major issue hindering the success of VQAs is *barren plateaus* [Lar+24; Qi+23]. This is a property that the objective function exhibits where all partial derivatives for almost all inputs are exponentially small. Barren plateaus were first theoretically demonstrated for VQAs that use deep (number of layers scaling at least linearly in the number of qubits involved) parameterized circuits in [McC+18] and then for shallow ones (number of layers scaling at most logarithmically in the number of qubits), induced by the choice of the observables, by [Cer+21b]. However, several heuristic methods that address barren plateaus have been proposed, including [Pat+21; Mel+22; RSL22; Sko+21; Gri+23a; Gri+23b; FM22; Ver+19; Gra+19a; KS22; Zha+22a].

Another issue that has received comparatively lesser attention is the sample complexity of VQAs [Cer+23; Fon+22; BK22]. In quantum information, the no-cloning theorem states that a quantum device that acts as a universal copier of quantum states cannot exist [WWZ82]. Also, unlike classical computing, a reliable quantum memory device [GLM08] is yet to be developed. These factors imply 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 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

will be available, making the implementation of VQAs with reduced sample complexity crucial.

One method that has gained recent attention is to use the quantum device to gather just the right amount of information from the input quantum states, using a number of samples that is at most logarithmic in the total number of function evaluations required, so as to be able to then classically simulate the whole optimization procedure efficiently (using classical resources polynomial in the number of qubits involved). Although it is clear that this need not be possible for all kinds of VQA tasks, there have been some notable works that achieved this when some specific classes of ansatzes and/or measurements are involved [Cer+23; BK22; Oka+22; Fon+22; Bas+23; Bas+24a]. We use the phrase *classical simulation* of VQAs to mean these kinds of protocols, where few quantum resources are used to develop models that can be used to completely simulate the optimization procedure classically efficiently, thus exponentially improving the sample complexity of VQAs. Other approaches aimed at reducing sample complexity of VQAs include classically simulating the optimized (learned) VQA models [SEM22], developing optimization algorithms specific to VQAs, that require fewer iterations and function evaluations compared to standard classical methods [KB22; Sto+20], etc.

This thesis aims to study the trainability and classical simulability of different VQA ansatzes when used in combination with different types of observables. More specifically, the contributions of this thesis are as follows:

1. We introduce a training protocol whose sample complexity is exponentially lesser than the standard method for VQA objective functions that use the Alternating Layered Ansatz (ALA) [Cer+20] (cf. Figure 3.5 (a)) and local observables (observables which restricts measurements to only a small number of qubits).

2. We introduce a training protocol whose sample complexity is exponentially lesser than the standard method for VQA objective functions that use almost any shallow ansatz and observables of low Frobenius norm (observables which typically involve measuring nearly all qubits and consider only a few outcomes).

	Sto+20	BK22	Oka+22	Fon+22 KB22	ALSO/AISO
Agnostic to choice of optimizer	No	No	Yes	Yes	Yes
Independent of input state	Yes	Yes	No	Yes	Yes
Rigorous complexity guarantees	No	Yes	Yes	No	Yes

Table 1.1: Comparison with previous works on classical simulation and improving the sample complexity of VQAs.

3. We theoretically study the trainability of learning weakly entangled approximations of states variationally using the Matrix Product State (MPS) ansatz (cf. Figure 3.7). Specifically, we rigorously prove that the usage of global observables will induce barren plateaus, while the usage of local observables will avoid them. We also provide strong evidence that suggests the existence of a protocol that can greatly improve the sample complexity when using local observables.

The classical simulation methods introduced in this thesis have many advantages over other state-of-the-art methods providing sample complexity advantages for VQAs, as illustrated in Table 1.1. These advantages are provided in detail in the related works sections of the corresponding chapters (cf. Sections 4.7 and 6.7). Broadly, our methods demonstrate greater compatibility with classical optimizers than those in [BK22; Sto+20], support a wider class of input states compared to works such as [Oka+22], and provide more rigorous sample complexity guarantees than [Fon+22; KB22; Sto+20]. [Cer+23] is a comprehensive work on classical simulation that was released after ALSO and AISO which shares a lot of similarities with them.

1.2 Publications Related to This Thesis

- [Bas+23] 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, 2023, 37(6):6770–6778.

- [Bas+24a] Afrad Basheer, Yuan Feng, Christopher Ferrie, and Sanjiang Li. Ansatz-

Agnostic Exponential Resource Saving in Variational Quantum Algorithms Using Shallow Shadow. Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence Main Track, 2024, Pages 3706-3714.

- [Bas+24b] Afrad Basheer, Yuan Feng, Christopher Ferrie, Sanjiang Li, and Hakop Pashayan. On the Trainability and Classical Simulability of Learning Matrix Product States Variationally, arXiv: 2409.10055 (Accepted for publication in proceedings of the AAAI Conference on Artificial Intelligence, 2025).

1.3 Thesis Organization

This thesis is organized as follows:

Chapter 2 - Preliminaries: In this chapter, we introduce the notations used in this thesis and explain the required background concepts in detail. This includes quantum computing and quantum information, MPS, unitary t-designs, and classical shadow tomography.

Chapter 3 - Variational Quantum Algorithms: In this chapter, we introduce and explain VQAs in detail. This includes an overview, examples ansatzes, various applications, and trainability issues.

Chapter 4 - Alternating Layered Shadow Optimization: In this chapter, we introduce Alternating Layered Shadow Optimization (ALSO) — an efficient method to train alternating layered VQAs (ones that use the ALA) that is exponentially better than the standard way of training VQAs in terms of sample complexity, when used in combination with local observables. The saving of state copies is especially useful when multiple rounds of the same optimization algorithm are required for various choices of hyperparameters, or when one has to experiment with different VQAs altogether. Moreover, ALSO is implementable using fewer and simpler quantum operations; in fact, only single-qubit measurements according to Pauli bases are required in ALSO. Our algorithm uses classical shadows (cf. Section 4.2) of quantum input data, and can hence run on a classical computer with rigorous performance guarantees. Another interesting benefit is that the

produced classical shadows can be reused in different (independent) tasks. For example, the same set of classical shadows can be used in both finding the state preparation circuits and building quantum autoencoders. We demonstrate 2–3 orders of magnitude improvement in the training cost using our algorithm for the example problems of finding state preparation circuits (cf. Section 3.4.1) and the quantum autoencoder (cf. Section 3.4.3).

Chapter 5 - Ansatz Independent Shadow Optimization: In this chapter, we proposed Ansatz Independent Shadow Optimization (AISO) — a training algorithm that leverages shallow shadows (cf. Section 5.2) to achieve an exponential reduction in quantum resources required to train VQA objective 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: state preparation and Variational Quantum Circuit Synthesis (VQCS) (cf. Section 3.4).

Chapter 6 - Trainability and Classical Simulability of Learning MPS Approximations Using VQAs: In this chapter, we introduce new results regarding cost concentration, trainability, and classical simulability of learning state approximations of quantum states variationally using the MPS ansatz. This ansatz leverages the MPS data structure, which stores quantum states with space complexity that scales polynomially with the *bond dimensions* (parameters that measure entanglement between neighboring qubits). 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. We prove 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 demonstrate that using the ansatz with local observables reveals effective subspaces (cf.

Section 6.4.1) within the Pauli basis, paving the way for a potential classical simulation of this VQA. Also, all our results are experimentally validated across various scenarios.

Chapter 7 - Conclusion and Future Direction: In this chapter, we conclude this thesis by summarizing all the contributions made as part of this thesis and discuss the future directions of research that stem from the ideas developed as part of this work.

Chapter 8 - Appendix: In this chapter, we provide additional theorems, lemmas, definitions and more discussions on tensors and MPS.

Chapter 2

Preliminaries

First, we introduce some notations and terminologies used in this thesis.

$\lvert\cdot\rangle$ ($\langle\cdot\rvert$)	Column vectors (Row vectors)
$\lvert i\rangle$	i^{th} computational basis vector
$\mathcal{L}(V)$	Set of all operators acting on the vector space V
$\overline{X}(X^\dagger)$	Conjugate (conjugate transpose)
$\mathbb{1}$ ($\mathbb{1}_V$)	Identity matrix acting on $\mathbb{C}^2$ (acting on V)
$\prod_{t=1}^{p} A_t$	$A_1 A_2 \dots A_p$
$\lvert\mathbf{b}\rangle$	$\bigotimes_{i=1}^{n} \lvert b\rangle$, where $b \in \{0,1\}$ and n is implicitly understood from context
$\lVert A\rVert_p$	$\sqrt{\sum_{ij} \lvert A_{ij}\rvert^p}$, where $A = \sum_{ij} A_{ij} \lvert i\rangle\langle j\rvert$
$\lVert A\rVert_\infty$	Largest singular value of A.
$\lVert A\rVert_{\text{tr}}$	$\sum_i \omega_i$, where $\{\omega_i\}$ are the eigenvalues of $\sqrt{A^\dagger A}$
A_B	$BAB^\dagger$ if $A, B \in \mathcal{L}(\mathbb{C}^d)$
$\mathbb{N}, \mathbb{R}, \mathbb{C}$	The set of all natural, real and complex numbers respectively
$\mathbb{U}_t, \mathbb{H}_t$	The set of all unitaries, and Hermitians acting on $\mathbb{C}^t$ respectively
$\{A^{(j)}\}_j$	$\{A^{(1)}, A^{(2)}, A^{(3)}, \dots\}$, where $A^{(j)}$s are indexed set of entities

2.1 Matrix Basics

In this section, we recall some basic definitions regarding matrices.

- For any square matrix $A \in \mathbb{C}^{n\times n}$, its *trace* is defined as the sum of its diagonal elements. That is, if $A = \sum_{i,j=0}^{n-1} A_{ij} |i\rangle\langle j|$, then $\mathrm{tr}(A) = \sum_{i=0}^{n-1} A_{ii}$, where $\mathrm{tr}(A)$ is its trace.

- A *Hermitian* matrix is a square matrix $A \in \mathbb{C}^{n\times n}$ such that $A^\dagger = A$. These matrices admit only real eigenvalues.

- A *positive semi-definite* (*definite*) matrix is a Hermitian matrix $A \in \mathbb{C}^{n\times n}$ such that $\langle\psi| A |\psi\rangle \geqslant 0 \ \forall |\psi\rangle \in \mathbb{C}^n$ ($\langle\psi| A |\psi\rangle > 0 \ \forall |\psi\rangle \in \mathbb{C}^n$). These matrices admit only non-negative (positive) eigenvalues.

- The *matrix square root* of a square matrix $A \in \mathbb{C}^{n\times n}$ is any matrix $B \in \mathbb{C}^{n\times n}$ such that $A = B^2$.

- The *rank* of a matrix A is the dimension of the vector space spanned by its columns (equivalently, its rows).

2.2 The Kronecker Product

Let $A \in \mathbb{C}^{m_1\times n_1}$ and $B \in \mathbb{C}^{m_2\times n_2}$. Then the *Kronecker product* $A \otimes B \in \mathbb{C}^{m_1 m_2 \times n_1 n_2}$ is given as

$$A \otimes B = \begin{bmatrix} a_{11}B & \dots & a_{1n_1}B \\ \vdots & \vdots & \vdots \\ a_{m_1 1}B & \dots & a_{m_1 n_1}B \end{bmatrix}. \tag{2.1}$$

Throughout this work, for a set of matrices $\{A^{(i)}|$ for $i = 1, \dots, k\}$, $A^{(1)} \otimes A^{(2)} \otimes \cdots \otimes A^{(k)}$ is denoted as $\left(\bigotimes_{i=1}^{k} A^{(i)}\right)$. Similarly, for any matrix A, $A^{\otimes t} = A \otimes A \otimes \cdots \otimes A$ (t times). In the case of vectors $|\psi\rangle, |\phi\rangle$, for convenience, we denote $|\psi\rangle \otimes |\phi\rangle$ as $|\psi\rangle|\phi\rangle$. Also, we define $A_t^{(n)} := \mathbb{1}^{\otimes n-t} \otimes A \otimes \mathbb{1}^{\otimes t-1}$. In many cases, we omit the dependence on n and simply

write A_t, as this dependence is often implicitly understood. Some important properties of Kronecker products are

- For matrices $A \in \mathbb{C}^{t_1 \times t_2}$, $B \in \mathbb{C}^{t_2 \times t_3}$, $C \in \mathbb{C}^{s_1 \times s_2}$, $D \in \mathbb{C}^{s_2 \times s_3}$, we have $(A \otimes C)(B \otimes D) = AB \otimes CD$.
- For matrices $A \in \mathbb{C}^{t_1 \times t_1}$, $B \in \mathbb{C}^{t_2 \times t_2}$, we have $\mathrm{tr}(A \otimes B) = \mathrm{tr}(A)\mathrm{tr}(B)$.
- For matrices $A \in \mathbb{C}^{t_1 \times t_2}$, $B \in \mathbb{C}^{s_1 \times s_2}$ and $C \in \mathbb{C}^{s_1 \times s_2}$, we have $A \otimes (B + C) = (A \otimes B) + (A \otimes C)$.

2.3 Quantum Computing and Quantum Information

2.3.1 Quantum State, Gates, and Measurement

A quantum state $\sigma \in \mathcal{L}(\mathbb{C}^d)$ is a positive semi-definite operator acting on $\mathbb{C}^d$ with trace 1. When the rank of σ is 1, we say that σ is a *pure state*. Otherwise, we say that it is a *mixed state*. A *qubit* is the fundamental implementable entity in quantum computing and can admit any state $\sigma \in \mathcal{L}(\mathbb{C}^2)$ as its value, similar to how a *bit* in classical computing can admit any value in $\{0, 1\}$. When a qubit q takes a value σ, we say that q *is in the state* σ, or the operator σ *describes the state* that q is in. One can also define *qudits* as similar entities that can be in any state $\sigma \in \mathcal{L}(\mathbb{C}^d)$. Define a *system* or *register* to be a tuple of such entities. It is better to fix the ordering of the entities and hence a tuple is preferred. To describe the state of a system $\mathcal{S} = (q_n, q_m)$, where q_n is a qu-n-it and q_m is a qu-m-it, we use operators acting on the *tensor product of the vector spaces* $\mathbb{C}^n$ and $\mathbb{C}^m$, denoted as $\mathbb{C}^n \otimes \mathbb{C}^m$. This is the vector space defined as the span of all vectors of the form $|v_1\rangle \otimes |v_2\rangle$, where $|v_1\rangle \in \mathbb{C}^n$ and $|v_2\rangle \in \mathbb{C}^m$. This vector space is the nm-dimensional vector space $\mathbb{C}^{mn}$. So, a system of n-qubits can be in any state in $\mathcal{L}(\mathbb{C})$. A state $\sigma \in \mathcal{L}(\mathbb{C}^{2^n})$ describing a system of n-qubits, is sometimes referred to as an *n-qubit state.*

A *quantum gate* is defined as a unitary operator $U \in \mathcal{L}(\mathbb{C}^d)$. The application of such a quantum gate on a system in the state $\sigma \in \mathcal{L}(\mathbb{C}^d)$ transforms the state of the system as

follows:

$$\sigma \xrightarrow{U} \sigma_U, \tag{2.2}$$

where $\sigma_U := U\sigma U^\dagger$. One can see that this is a reversible operation and if we follow this operation up with an application of $U^\dagger$, which is also a quantum gate, on this system, we get the system back to the original state σ.

Given a system, generally, it is impossible to accurately compute the state that the system is in. The standard way with which we read an observation from a system in some state is through *quantum measurements*. A quantum measurement is defined by a set of measurement operators $\mathcal{M} = \{M^{(j)} \in \mathcal{L}(\mathbb{C}^d)\}_j$ such that $\sum_j {M^{(j)}}^\dagger M^{(j)} = \mathbb{1}_{\mathbb{C}^d}$. When a system in a state σ is "measured", the state of the system undergoes a transformation given as

$$\sigma \xrightarrow{\mathcal{M}} \frac{\sigma_{M^{(j)}}}{\mathrm{tr}\Big(\sigma_{M^{(j)}}\Big)} \text{ with probability } \mathrm{tr}\Big(\sigma_{M^{(j)}}\Big). \tag{2.3}$$

This is what one can observe from quantum systems. When a measurement using a measurement set $\mathcal{M}$ is carried out, we will observe the index j of the operator $M^{(j)}$ that was chosen as part of the probabilistic transformation. If one is only interested in the index, and not interested in what the state of the system is after measurement, we can denote the measurement protocol as

$$\sigma \xRightarrow{\mathcal{M}} j \text{ with probability } \mathrm{tr}\Big(\sigma_{M^{(j)}}\Big). \tag{2.4}$$

Due to the nature of the probabilistic transformation of the measurement protocol, one can say that measurement "destroys the state". In general, unlike the application of a quantum gate, the measurement is an irreversible operation.

An *observable* is defined as any Hermitian operator. Any Hermitian operator $O \in \mathcal{L}(\mathbb{C}^d)$ can be decomposed in terms of its eigenbasis using spectral decomposition as $O = \sum_j \lambda_j P^{(j)}$, where λ_j and $P^{(j)}$ are eigenvalues and associated projections on to the

corresponding eigenspace. Given a system in a state $\sigma \in \mathcal{L}(\mathbb{C}^d)$, the *expectation* of an observable O is defined as $\mathrm{tr}(O\sigma) = \sum_j \lambda_j \mathrm{tr}\left(\sigma P^{(j)}\right)$.

The probabilistic interpretation of this can be found in [Par05]. If we know the full spectral information of O, and have the capability of preparing a system in the state σ multiple times (or we are provided with multiple systems all prepared in the state σ), then $\mathrm{tr}(O\sigma)$ can be estimated by multiple measurements using the measurement set $\{P^{(j)}\}_j$.

The number of measurements (equivalently the number of copies of σ since each measurement destroys a copy) required to get an accurate estimate of $\mathrm{tr}(\sigma O)$ can be computed using Hoeffding's inequality [Hoe63], which says that for independent random variables $\eta^{(1)}, \eta^{(2)}, \ldots, \eta^{(T)}$ with $\eta^{(i)} \in [a_i, b_i]$, and for any $\epsilon \in (0,1)$, we have

$$\mathrm{Prob}\left(|\hat{\eta} - \mathbb{E}\left(\hat{\eta}\right)| \geqslant \epsilon\right) \leqslant 2e^{\Delta}, \tag{2.5}$$

where $\Delta = \left(\frac{-T^2\epsilon^2}{\sum_{i=1}^{T}(b_i - a_i)^2}\right)$ and $\hat{\eta} = \frac{1}{T}\sum_{i=1}^{T}\eta^{(i)}$. Let the outcome of the measurement be a random variable η. So the range of η is $[\lambda_{\min}, \lambda_{\max}]$, where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues of O respectively. If we measure T times, we are essentially getting the outcomes of T independent and identically distributed random variables $\eta^{(1)}, \ldots, \eta^{(T)}$ all distributed according to η. We also have $\mathbb{E}(\eta) = \mathrm{tr}(\sigma O)$. Putting all this is Eq (2.5) means that for any $\epsilon, \delta \in (0,1)$, if we measure the state $T \geqslant \frac{2(\lambda_{\max} - \lambda_{\min})^2}{\epsilon^2}\log\left(\frac{2}{\delta}\right)$ times, we will have

$$\mathrm{Prob}\left(|\hat{\eta} - \mathrm{tr}\left(\sigma O\right)| \leqslant \epsilon\right) \geqslant 1 - \delta, \tag{2.6}$$

where $\hat{\eta} = \frac{1}{T}\sum_{i=1}^{T}\eta^{(i)}$ is the sample means estimator. This can be derived by setting δ to be the right-hand side of Eq (2.5). Throughout this thesis, in similar contexts, the parameters ϵ and δ are called *precision* and *confidence* parameters.

These kinds of measurements are called *projective measurements*. Measurements in the computational basis, that is, using the set $\{|j\rangle\langle j|\}_j$ is called *standard projective mea-*

surement. This is the general type of measurement carried out in quantum circuits. To measure using any orthonormal basis given by the columns of a unitary matrix V, one simply applies the quantum gate $V^\dagger$ on the system and carries out standard projective measurement.

Quantum fidelity is a measure of similarity between two quantum states. For any two states σ, ρ, the (squared) fidelity is defined as

$$F(\sigma, \rho) = \left(\mathrm{tr}\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right)^2. \tag{2.7}$$

The higher the fidelity, the higher the similarity between the states, with a fidelity of 1 if and only if $\rho = \sigma$. A fidelity of 0 implies that the states are orthogonal, or equivalently, as "different" as possible. If any one of the input states is pure, then we have $F(\rho, \sigma) = \mathrm{tr}(\rho\sigma)$. Also, *quantum infidelity* is defined as $1 - F(\rho, \sigma)$.

2.3.2 Pure State Dynamics

Let $\mathcal{S}$ be a system in the pure state $\sigma \in \mathcal{L}(\mathbb{C}^d)$. Pure quantum states will admit only one non-zero eigenvalue. So, rather than using the matrix σ to describe the state of $\mathcal{S}$, one can also use any normalized (with respect to Euclidean 2-norm) eigenvector associated with its non-zero eigenvalue. That is, let $\sigma = |\psi\rangle\langle\psi|$ be an eigendecomposition of σ. Then, the application of any quantum gate $U \in \mathcal{L}(\mathbb{C}^d)$ on $\mathcal{S}$ can be modeled by the action of U as an operator on $|\psi\rangle$. From (2.2), we can see that the resulting state is the unique pure state whose eigenspace corresponding to the only non-zero eigenvalue is given by the span of $\{U|\psi\rangle\}$. This means that the action of U can be seen as

$$|\psi\rangle \xrightarrow{U} U|\psi\rangle. \tag{2.8}$$

Similarly, measuring $\mathcal{S}$ using the measurement set $\mathcal{M} = \left\{M^{(j)} \in \mathcal{L}(\mathbb{C}^d)\right\}$ can also be described in terms of $|\psi\rangle$ alone as

$$|\psi\rangle \xrightarrow{\mathcal{M}} \frac{M^{(j)}|\psi\rangle}{\left\|M^{(j)}|\psi\rangle\right\|_2} \text{ with probability } \left\|M^{(j)}|\psi\rangle\right\|_2^2 \tag{2.9}$$

or

$$|\psi\rangle \xRightarrow{\mathcal{M}} j \text{ with probability } \left\|M^{(j)}|\psi\rangle\right\|_2^2. \tag{2.10}$$

When a state is represented using an operator, we call that representation of the state the *density matrix* representation. So, given a pure state $|\psi\rangle$ in vector form, its density matrix representation can be computed as $|\psi\rangle\langle\psi|$. One can also see that $|\psi\rangle$ and $e^{i\gamma}|\psi\rangle$ are indistinguishable for any pure state $|\psi\rangle$ and $\gamma \in \mathbb{C}$, since by $|\psi\rangle$, what we are interested in is not the vector but the space spanned by the vector. Both of them are essentially the same quantum state $\sigma = |\psi\rangle\langle\psi| = e^{i\gamma}|\psi\rangle\langle\psi|e^{-i\gamma}$.

It is also convenient to define pure quantum states as tensors in a tensor product space [KB09; Nic13] (cf. Section 8.1 for a brief introduction). Let $\{|0\rangle, |1\rangle\}$ be the computational basis of the 2-dimensional Hilbert space (in the finite dimensional setting, a Hilbert space is a vector space endowed with an inner product) $\mathbb{C}^2$, where

$$|0\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad |1\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{2.11}$$

Consider n 2-dimensional Hilbert spaces $\{\mathcal{V}_j = \mathbb{C}^2 |\ j = 1, 2, \ldots, n\}$. So, the set

$$\left\{|j\rangle = |j_1\rangle|j_2\rangle\ldots|j_n\rangle \,\middle|\, j = 0, \ldots, 2^n - 1\right\}, \tag{2.12}$$

where $j = j_1 j_2 \ldots j_n$ is the binary representation of the integer j, forms an orthonormal basis (computational basis) for the 2^n-dimensional Hilbert space $\bigotimes_{j=1}^{n} \mathcal{V}_j = \mathbb{C}^{2^n}$. This

means that any n-qubit pure quantum state $|\psi\rangle \in \mathbb{C}^{2^n}$ can be written as

$$|\psi\rangle = \sum_{j=0}^{2^n-1} \psi_j |j\rangle = \sum_{j_1,j_2,\ldots,j_n=0}^{1} \psi_{j_1 j_2 \ldots j_n} |j_1\rangle |j_2\rangle \ldots |j_n\rangle, \tag{2.13}$$

where $\psi_j \in \mathbb{C}$. Also, since this is a normalized vector, we have that $\sum_{j=0}^{2^n-1} |\psi_j|^2 = 1$.

One can also see that any normalized vector $|\psi\rangle \in \mathbb{C}^d$ is a pure quantum state since $|\psi\rangle\langle\psi|$ is a rank 1 positive semi-definite matrix with trace 1. When a system $\mathcal{S} = (q_1, q_2, \ldots, q_n)$ of n qubits is in such a state $|\psi\rangle$ in a tensor product space, we consider the j^{th} mode of the tensor to correspond to the qubit q_j. That is, the Hilbert space $\mathcal{V}_j$ corresponds to q_j. A consequence of this arrangement is that if you prepare each individual qubit q_j in the state $|\psi_j\rangle$, then the state of $\mathcal{S}$ is given as $\bigotimes_{j=1}^{n} |\psi_j\rangle$. Note that not all systems can be decomposed as a product of small tensors in this nice manner. One can prepare a system of qubits in such a manner that this decomposition in terms of 2-dimensional complex vectors is not possible. In the case of a 2-qubit system $\mathcal{S} = (q_1, q_2)$, if we can decompose the state $|\psi\rangle$ of the system as $|\psi\rangle = |\psi_1\rangle|\psi_2\rangle$, we say that the qubits q_1 and q_2 are *separable*. In this case, we can say that the state of q_1 is $|\psi_1\rangle$ and the state of q_2 is $|\psi_2\rangle$. If we cannot decompose $|\psi\rangle$ in this manner, we say that q_1 and q_2 are *entangled*. This concept can be extended to multiple qubits as well as density matrices [Wat18]. Quantum entanglement is a fundamental resource in many quantum protocols [NC11; Hor+09; Eke91; Bri+98; Ben+93].

2.3.3 Partial Trace

Let A be a register of dimension d_A and B be a register of dimension d_B. Given any state $\sigma \in \mathcal{L}(\mathbb{C}^{d_A d_B}), \sigma = \sum_{i_1,j_1=0}^{d_A-1} \sum_{i_2,j_2=0}^{d_B-1} \sigma_{i_1 i_2, j_1 j_2} |i_1\rangle|i_2\rangle\langle j_1|\langle j_2|$ defined on these two registers, the state of the subsystem A is computed using a operation linear operation called the *partial trace*. This operation can be explained as applying the trace operation on one register

alone. Formally, we define $\mathrm{tr}_B(\sigma)$, applying partial trace on B or *tracing out* B, as

$$\mathrm{tr}_B(\sigma) = \sum_{q=0}^{d_B-1} \left(\mathbb{1}_{\mathbb{C}^{d_A}} \otimes \langle q|\right) \sigma \left(\mathbb{1}_{\mathbb{C}^{d_A}} \otimes |q\rangle\right) \tag{2.14}$$

$$= \sum_{q=0}^{d_B-1} \sum_{i_1,j_1=0}^{d_A-1} \sum_{i_2,j_2=0}^{d_B-1} \sigma_{i_1 i_2, j_1 j_2} \left(\mathbb{1}_{\mathbb{C}^{d_A}} \otimes \langle q|\right) |i_1\rangle |i_2\rangle\langle j_1| \langle j_2| \left(\mathbb{1}_{\mathbb{C}^{d_A}} \otimes |q\rangle\right) \tag{2.15}$$

$$= \sum_{q=0}^{d_B-1} \sum_{i_1,j_1=0}^{d_A-1} \sum_{i_2,j_2=0}^{d_B-1} \sigma_{i_1 i_2, j_1 j_2} \left(\mathbb{1}_{\mathbb{C}^{d_A}} \otimes \langle q|\right) |i_1\rangle\langle j_1| \otimes |i_2\rangle\langle j_2| \left(\mathbb{1}_{\mathbb{C}^{d_A}} \otimes |q\rangle\right) \tag{2.16}$$

$$= \sum_{i_1,j_1=0}^{d_A-1} \sum_{q=0}^{d_B-1} \sigma_{i_1 q, j_1 q} |i_1\rangle\langle j_1| . \tag{2.17}$$

It is easy to see that partial trace is a linear operation. The state of the subsystem A is $\mathrm{tr}_B(\sigma)$ while the trace of the subsystem B is $\mathrm{tr}_A(\sigma)$. To see why these are valid quantum states, refer to Appendix 8.3. In many instances, we call $\sigma_A := \mathrm{tr}_B(\sigma)$ the *reduced density matrix* of σ on A.

Also, for any matrix $C = C^{(a)} \otimes C^{(b)}$, defined on the same two register Hilbert space, we have $\mathrm{tr}_B(C) = C^{(a)}\mathrm{tr}\left(C^{(b)}\right)$. Hence, for any arbitrary matrix C defined on this Hilbert space, since there always exists a finite number of matrix $\{C^{(a_k)}\}_{a_k}, \{C^{(b_k)}\}_{b_k}$ such that $C = \sum_k C^{(a_k)} \otimes C^{(b_k)}$, we have $\mathrm{tr}_B(C) = \sum_k C^{(a_k)}\mathrm{tr}(C^{(b_k)})$.

We can also extend this definition to a state σ defined on many registers $A_1, A_2, \ldots, A_t$ of dimensions $d_1, d_2, \ldots, d_t$. Application of partial trace on registers $\mathcal{J} = \{A_{J_1}, A_{J_2}, \ldots, A_{J_{t'}}\}$ results in the state $\mathrm{tr}_{\mathcal{J}}(\sigma)$ where

$$\mathrm{tr}_{\mathcal{J}}(\sigma) = \sum_{j_1=0}^{d_1-1} \cdots \sum_{j_{t'}=0}^{d_{t'}-1} \left(\delta^{(1)\dagger} \otimes \cdots \otimes \delta^{(t)\dagger}\right) \sigma \left(\delta^{(1)} \otimes \cdots \otimes \delta^{(t)}\right), \tag{2.18}$$

where $\delta^{(i)} = |j_i\rangle$ if $i \in \{J_1, J_2, \ldots, J_{t'}\}$ and $\delta^{(i)} = \mathbb{1}_{\mathbb{C}^{d_i}}$.

2.3.4 Quantum Circuits

Similar to how a classical circuit is defined as a set of (classical) gates acting on n-bits which encodes the n-bit long input values, a *quantum circuit* is defined by a set of quantum

gates acting on a system of n-qubits, described by an *input state* (pure in most cases). The output of a classical circuit can be easily read from the output bit string. However, the output of a quantum circuit cannot be read directly from the final state that the system is in after all gates have been applied. We call the state of the system after all gates have been applied, the *output state*. What one can do, is measure the output state, usually a standard projective measurement, and the probabilistic output is the observed output of a quantum circuit. So that means, if the input state is $|\psi\rangle$ and the circuit is given by the collection of quantum gates $\left\{U^{(j)} \in \mathcal{L}(\mathbb{C}^{2^n}) |\ j = 1, \ldots, m\right\}$ (where $U^{(1)}$ is applied first, then $U^{(2)}$ and so on), then the output of the quantum circuit will be

$$|\psi\rangle \xrightarrow{U_1, U_2, \ldots, U_m} |\phi\rangle \Rightarrow j \text{ with probability } |\phi_j|^2, \tag{2.19}$$

where $|\phi\rangle = U_m \ldots U_2 U_1 |\psi\rangle = \sum_{j=0}^{2^n-1} \phi_j |j\rangle$. In this scenario, the post-measurement state of the system will always be a computational basis vector $|j\rangle$. So, even though an n-qubit pure quantum state is described by or "stores" 2^n complex numbers, we cannot access these numbers as we can do from a classically stored n-bit string.

For a system of qubits $\mathcal{S}$, define a *subsystem* or *subregister* to be a tuple of qubits selected without replacement from $\mathcal{S}$. Let $\mathcal{S} = (q_1, q_2, \ldots, q_n)$ be a system of n qubits in a state $|\psi\rangle \in \bigotimes_{j=1}^{n} \mathcal{V}_j$. Let $\mathcal{J} = (q_{j_1}, q_{j_2}, \ldots, q_{j_t})$ be a subsystem of $\mathcal{S}$. One can also apply a quantum gate $U \in \mathcal{L}(\mathbb{C}^{2^t})$ on the subsystem $\mathcal{J}$ alone. Such an application can be modeled using tensor operations. The resultant state can be computed as the matrix U acting on the tuple of t indices $(j_1, j_2, \ldots, j_t)$ of the tensor $|\psi\rangle$. That is, we contract these indices of $|\psi\rangle$ (viewed as a tensor with n indices of length 2) with the column index of U (after splitting its row and column indices each into t indices of length 2). The exact definitions of tensor contraction and index splitting are given in the Appendix (cf. Definitions 4 and 5. Then the state of $\mathcal{S}$ after applying U on $\mathcal{J}$, $|\phi\rangle = U_{\mathcal{J}} |\psi\rangle$, is given as

$$|\phi\rangle = \sum_{p_1, \ldots, p_n = 0}^{1} \left(\sum_{q_1, \ldots, q_t = 0}^{1} U_{p_1 \ldots p_t, q_1 \ldots q_t} \psi_{p'_1 \ldots p'_n} \right) |p_1 \ldots p_n\rangle, \tag{2.20}$$

where

$$p'_k = q_k \text{ if } k \in \mathcal{J} \tag{2.21}$$

$$p'_k = p_k \text{ otherwise.} \tag{2.22}$$

This can concisely be written as $|\phi\rangle = U_{\mathcal{J}} |\psi\rangle$ or $|\psi\rangle \xrightarrow{U_{\mathcal{J}}} |\phi\rangle$. Another way of describing this operation using permutation matrices is given in Section 8.2 in the Appendix. A quantum gate $U \in \mathbb{C}^{2^t}$ is sometimes referred to as a t-qubit quantum gate.

Similarly, one can also measure a subsystem $\mathcal{J}$ of an n-qubit system $\mathcal{S}$ that is in the state $|\psi\rangle$. This can be described using the measurement set $\mathcal{M} = \left\{M^{(j)}\right\}_j$ and the output of the measurement is given as

$$|\psi\rangle \xrightarrow{\mathcal{M}} \frac{{M^{(j)}}_{\mathcal{J}} |\psi\rangle}{\left\|{M^{(j)}}_{\mathcal{J}} |\psi\rangle\right\|_2} \text{ with probability } \left\|{M^{(j)}}_{\mathcal{J}} |\psi\rangle\right\|_2^2 \tag{2.23}$$

or

$$|\psi\rangle \xRightarrow{\mathcal{M}} j \text{ with probability } \left\|{M^{(j)}}_{\mathcal{J}} |\psi\rangle\right\|_2^2. \tag{2.24}$$

One can also easily model the same "partial" operations in the density matrix framework by extending these tensor network contractions to matrix formalisms in a similar way. That is, the application of any operator G (gate or measurement) can be computed by using the previous method to first compute the result of the right multiplication of σ with G, followed by a left multiplication of the result with $G^\dagger$. Denote such a gate application as $\sigma_{G_{\mathcal{J}}}$.

Another way of representing this operation is using reduced density matrices. Let σ be an n-qubit state defined on $(q_1, q_2, \ldots, q_n)$. Application of a t-qubit gate U and/or measurement t-qubit measurement operators $\left\{M^{(i)}\right\}_i$ on a subsystem $\mathcal{J} = (q_{j_1}, \ldots, q_{j_t})$, is equivalent to application of the same operations on $\sigma_{\mathcal{J}} = \text{tr}_{\overline{\mathcal{J}}}(\sigma)$, where $\overline{\mathcal{J}}$ is the register

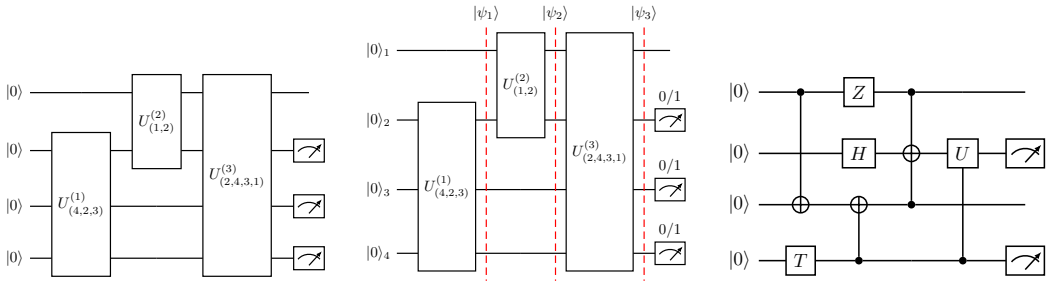


Figure 2.1: (a), (b) An example of a quantum circuit. The initial state of the system is $|0000\rangle = |\mathbf{0}\rangle$. Throughout this thesis, we label qubits as $1, 2, \ldots$ starting from the top. The subscript below each $|0\rangle$ in (b) depicts that. The quantum gate $U^{(1)}$ is applied to the tuple of qubits $(4, 2, 3)$. So, $|\psi_1\rangle = U^{(1)}_{(4,2,3)} |\mathbf{0}\rangle$. Similarly $|\psi_2\rangle = U^{(2)}_{(1,2)} U^{(1)}_{(4,2,3)} |\mathbf{0}\rangle$ and $|\psi_3\rangle = U^{(3)}_{(2,4,3,1)} U^{(2)}_{(1,2)} U^{(1)}_{(4,2,3)} |\mathbf{0}\rangle$. $|\psi_3\rangle$ is the output state and finally we measure qubits $(2, 3, 4)$. The meter symbol is used to denote computational basis measurement. Then, as the output of the circuit, we will see a 3 bit string. (c) Common representations of quantum gates. The single qubit gates are represented by small boxes labeled with their names. When it comes to CNOT and Toffoli gates, it is easier to denote the order of the tuple of qubits that the gate acts on. The $\text{CNOT}_{(q_1,q_2)}$ gate is depicted by a line segment with a black dot and a circle as ends. The ends determine which qubits the gate acts on. The black dot points to the qubit q_1 (control qubit), and the circle points to the qubit q_2 (target qubit). Generally, any $C\text{-}U_{(q_1,q_2)}$ gate can be depicted by a line segment with a black dot and the gate U as its ends, the black dot pointing to q_1 and U on q_2. Similarly the $\text{Toffoli}_{(q_1,q_2,q_3)}$ is depicted by two black dots pointing to both the control qubits q_1, q_2 and the circle on the target qubit q_3.

of all qubits not in $\mathcal{J}$. That is, the probabilities for all i, we have an equivalence of the measurement probabilities of the form $\text{tr}\left(\sigma_{G^{(i)}_{\mathcal{J}}}\right) = \text{tr}\left(\sigma_{\mathcal{J}G^{(i)}}\right)$, where $G^{(i)} = M^{(i)}U$.

Some of the most popular quantum gates include

$$H = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad T = \begin{bmatrix} 1 & 0 \\ 0 & e^{\frac{i\pi}{4}} \end{bmatrix}, \quad \text{CNOT} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad Y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \text{SWAP} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\text{Toffoli} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

H gate is often referred to as *Hadamard* gate. The gates $\{X, Y, Z\}$ are called *Pauli* operators. The action of some of these gates on computational basis vectors reveals nice analogies to their classical counterparts. The X gate maps $|0\rangle$ to $|1\rangle$ and vice-versa. This is the quantum version of the classical NOT gate. The CNOT gate is simply a controlled-NOT gate. It is a two qubit gate which maps $|j_1\rangle|j_2\rangle$ to $|j_1\rangle|j_2 \oplus j_1\rangle$. That is, CNOT being applied to (q_1, q_2) can be seen as follows: if qubit q_1 is in state $|1\rangle$, apply an X gate to qubit q_2. The qubit q_1 is called the *control qubit* and the qubit q_2 is called the *target qubit*. Similar controlled gates can be defined for other gates as well. An application of a general controlled gate C-U on (q_1, q_2) can be seen as an application of the gate U on the target qubit q_2 if the control qubit q_1 is in the state $|1\rangle$.

The SWAP gate can be used to swap the information contained in two qubits. Its effect can be neatly seen when applied to computational basis vectors. We can see that it leaves both $|00\rangle$ and $|11\rangle$ unchanged while mapping $|01\rangle$ to $|10\rangle$ and vice versa. Application of a Toffoli gate [Tof80] on (q_1, q_2, q_3) can be seen as a double-controlled X gate which will

apply an X gate to the target qubit q_3 if both the control qubits q_1 and q_2 are in the state $|1\rangle$. In classical computing, any circuit can be implemented equivalently using just Toffoli gates, with only polynomially growing additional resources required, in a reversible manner [Ben73; FT02]. So in quantum computing, the fact that a Toffoli gate can be efficiently constructed as a quantum gate using CNOT and 1-qubit gates [Bar+95] means that any classical circuit can be efficiently implemented as a reversible circuit in a quantum setting using qubits and quantum gates [NC11; SBM06].

Since the action on computational basis vectors of these classically analogous gates can be seen like this, their action on any state can be easily computed by writing the state in the computational basis and applying these operations linearly.

A typical n-qubit quantum circuit with input state $|\psi\rangle$ can be written as follows:

$$|\psi\rangle \xrightarrow{U^{(1)}_{\mathcal{J}_1}, U^{(2)}_{\mathcal{J}_2}, \ldots, U^{(m)}_{\mathcal{J}_m}} |\phi\rangle \xRightarrow{\mathcal{M}} j \text{ with probability } \left\| M^{(j)}{}_{\mathcal{F}} |\phi\rangle \right\|_2^2. \tag{2.25}$$

We start with a system $\mathcal{S}$ of n qubits in an input quantum state $|\psi\rangle \in \bigotimes_{j=1}^{n} \mathcal{V}_j$ as input. Then a series of quantum gates $U^{(1)}, U^{(2)}, \ldots, U^{(m)}$ will be applied on subsystems $\mathcal{J}_1, \mathcal{J}_2, \ldots, \mathcal{J}_m$ respectively. Finally, subsystem $\mathcal{F}$ is measured and we observe a $|\mathcal{F}|$ length bit string as the probabilistic output. A simple example is presented and discussed in detail in Figures 2.1 (a) and (b).

The common way of representing a quantum circuit is illustrated in Figure 2.1 (c). This circuit can be read as follows: first, we apply a CNOT gate with control as the first qubit and target as the third qubit and apply a T gate on the fourth qubit. Then we apply a CNOT gate with control as the fourth qubit and the target as the third qubit as well as Z and H gates on the first and second qubits. Then we apply a Toffoli gate with the first and third qubits as control qubits and the second qubit as the target qubit. Finally, we apply a controlled-U gate with the fourth qubit as the control and the second qubit as the target and then measure the second and fourth qubits.

Notice that in the circuit given in Figure 2.1 (c), when we start implementing the

circuit, we can either start with $\text{CNOT}_{(1,3)}$ or T_4. Since these two gates are acting on different qubits, the order in which they are implemented is irrelevant. The same goes with Z_1, H_2 and $\text{CNOT}_{(4,3)}$ after the application of $\text{CNOT}_{(1,3)}$ and T. We say that such gates can be implemented in the same *timestep*. But this is not the case for $\text{CNOT}_{(4,3)}$ and $\text{CNOT}_{(1,3)}$. This is because the state on which $\text{CNOT}_{(4,3)}$ is applied is directly dependent on the output of $\text{CNOT}_{(1,3)}$. The total number of such timesteps required for the whole circuit is called the *depth* of the circuit. For example, for the circuit in Figure 2.1 (c), the depth is 4. More detailed definitions and explanations of this concept can be found in [Ge+24].

A *universal* gate set is defined as a set of quantum gates that can be used to approximate the action of an arbitrary quantum gate to any degree. That is, a set $\mathcal{G}$ is a universal set if for any t-qubit quantum gate V and a precision parameter $\epsilon \in (0, 1)$, we have a finite number of gates $U^{(1)}, U^{(2)}, \ldots, U^{(m)}$, all selected from $\mathcal{G}$, and a finite number of tuples of qubits $\mathcal{J}_1, \mathcal{J}_2, \ldots, \mathcal{J}_m$ such that the operation ${U^{(m)}}_{\mathcal{I}_m} {U^{(m-1)}}_{\mathcal{I}_{m-1}} \ldots {U^{(1)}}_{\mathcal{I}_1}$ is ϵ close to V in terms of operator norm. Typically, the dependency of m on ϵ should be in $\mathcal{O}(\text{poly}(\log \frac{1}{\epsilon}))$. For example, the set $\{H, T, \text{CNOT}\}$ is a universal gate set [Boy+99].

2.3.5 Quantum Channels

Let $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3, \mathcal{H}_4$ be finite dimensional complex Hilbert spaces. A *super operator* is a map whose domain and co-domain are sets of operators themselves. Given two super operators $\Phi_1 : \mathcal{L}(\mathcal{H}_1) \to \mathcal{L}(\mathcal{H}_2)$ and $\Phi_2 : \mathcal{L}(\mathcal{H}_3) \to \mathcal{L}(\mathcal{H}_4)$, the action of the super operator $\Phi_1 \otimes \Phi_2 : \mathcal{L}(\mathcal{H}_1 \otimes \mathcal{H}_3) \to \mathcal{L}(\mathcal{H}_2 \otimes \mathcal{H}_4)$ can be seen as follows: for any input matrix $W \in \mathcal{L}(\mathcal{H}_1 \otimes \mathcal{H}_3)$, there always exist two finite sets of matrices, $\left\{A^{(i)} \in \mathcal{L}(\mathcal{H}_1)\right\}_i$ and $\left\{B^{(j)} \in \mathcal{L}(\mathcal{H}_3)\right\}_j$, such that $W = \sum_i \sum_j A^{(i)} \otimes B^{(j)}$ (we can construct on such set from the computational basis of $\mathcal{L}(\mathcal{H}_1 \otimes \mathcal{H}_3)$). Then $(\Phi_1 \otimes \Phi_2)(W) = \sum_{i,j} \Phi_1\left(A^{(i)}\right) \otimes \Phi_2\left(B^{(j)}\right)$. A map $\Phi : \mathcal{L}(\mathcal{H}_1) \to \mathcal{L}(\mathcal{H}_2)$ is a positive map if for all positive semi-definite matrices $W \in \mathcal{L}(\mathcal{H}_1)$, $\Phi(W)$ is positive semi-definite in $\mathcal{L}(\mathcal{H}_2)$.

Quantum channels constitute a broad class of implementable quantum operations. A

map $\Phi : \mathcal{L}(\mathcal{H}_1) \rightarrow \mathcal{L}(\mathcal{H}_2)$ is a quantum channel if Φ satisfying two conditions:

- Trace preserving: $\text{tr}(W) = \text{tr}(\Phi(W)) \ \forall \ W \in \mathcal{L}(\mathcal{H}_1)$
- Completely positive: $\Phi \otimes \mathbb{1}_{\mathcal{L}(\mathcal{H})}$ is a positive map for every finite-dimensional complex Hilbert space $\mathcal{H}$. Here, $\mathbb{1}_{\mathcal{L}(\mathcal{H})} : \mathcal{L}(\mathcal{H}) \rightarrow \mathcal{L}(\mathcal{H})$, such that $\mathbb{1}_{\mathcal{L}(\mathcal{H})}(W) = W$.

If an operation preserves trace as well as positivity, it will map density matrices to density matrices. But from the point of view of quantum circuits, we should be able to apply this operation on a subsystem of a system of qubits as well. This is why complete positivity is required. The channels associated with any quantum gate U is simply $\Phi(W) = UWU^\dagger$, while the channel associated with computational basis measurements is $\Phi(W) = \sum_i |i\rangle\langle i| \, W \, |i\rangle\langle i| = \sum_i W_{ii} \, |i\rangle\langle i|$.

Super operators can be conveniently described using different types of representations. One of them is called the *Kraus representation*. For any super operator $\Phi : \mathcal{L}(\mathcal{H}_1) \rightarrow \mathcal{L}(\mathcal{H}_2)$, there exist two collections of operators $\left\{A^{(i)} : \mathcal{H}_1 \rightarrow \mathcal{H}_2\right\}$ and $\left\{B^{(i)} : \mathcal{H}_1 \rightarrow \mathcal{H}_2\right\}$, not necessarily unique, such that $\Phi(W) = \sum_i A^{(i)} W B^{(i)\dagger}$. This representation of Φ is called the Kraus representation and the operators $\left\{A^{(i)}\right\}$ and $\left\{B^{(i)}\right\}$ are called *Kraus operators*.

One can define quantum channels in terms of Kraus representation also. A super operator $\Phi : \mathcal{L}(\mathcal{H}_1) \rightarrow \mathcal{L}(\mathcal{H}_2)$ is a quantum channel if and only if there exists a Kraus representation which satisfies the following two conditions:

- $A^{(i)} = B^{(i)} \ \forall \ i$. This makes the super operator completely positive.
- $\sum_i A^{(i)\dagger} A^{(i)} = \mathbb{1}_{\mathcal{H}_1}$. This makes the super operator trace-preserving.

So, a quantum channel Φ can be defined by a single collection of operators $\left\{A^{(i)}\right\}_i$, such that $\sum_i A^{(i)\dagger} A^{(i)} = \mathbb{1}$ and $\Phi(W) = A^{(i)} W A^{(i)\dagger}$. The quantum channel representation of gates and computational basis measurements that we provided earlier are examples of Kraus representations.

2.4 Matrix Product State

The MPS decomposition [Sch11; Vid03; Oru13; Hae+16] is well studied in quantum information, mostly in fields that concern efficient classical simulations of quantum processes, specifically of those involving weak entanglement structures. Although there are many variants of MPS decompositions, we mostly stick to linear MPS decompositions in this section, with other extensions discussed at the end. Under such an MPS decomposition, any n-qubit state $|\psi\rangle$ can be decomposed using n third order tensors as

$$|\psi\rangle = \sum_{i_1,i_2,\ldots,i_n=0}^{1} G_{i_1}^{(1)} G_{i_2}^{(2)} \ldots G_{i_n}^{(n)} |i_1 i_2 \ldots i_n\rangle \tag{2.26}$$

$$= \sum_{i_1,i_2,\ldots,i_n=0}^{1} \left(\sum_{j_2,j_3,\ldots,j_n=0}^{r_2-1,r_3-1,\ldots,r_n-1} G_{i_1,j_2}^{(1)} G_{i_2,j_2,j_3}^{(2)} \ldots G_{i_n,j_{n-1},j_n}^{(n)} G_{i_n,j_n}^{(n)} \right) |i_1 i_2 \ldots i_n\rangle, \tag{2.27}$$

where $G^{(j)} \in \mathbb{C}^{2\times r_j \times r_{j+1}}$ are called *core tensors*, $G_{i_j}^{(j)} \in \mathbb{C}^{r_j \times r_{j+1}}$ is the matrix we get when we set the first index of $G^{(j)}$ to i_j, and $r_1 = r_{n+1} = 1$.

The first equality can be understood as follows. Each third order tensor $G_{i_j}^{(j)}$ is simply the pair of matrices $G_0^{(j)}$ and $G_1^{(j)}$. Once we are provided with these n pairs of matrices, to find the $\psi_{i_1 i_2 \ldots i_n}$, we simply choose the matrices $G_{i_1}^{(1)}, G_{i_2}^{(2)}, \ldots G_{i_n}^{(n)}$ and compute their product. We get the second equality by expanding all the matrix multiplications involved.

The numbers $r_2, r_3, \ldots, r_n$ are called *bond dimensions*. The bond dimension r_j has a direct relationship with the entanglement between qubits j and $j+1$. Although one can have multiple MPS decompositions for the same state, with varying bond dimensions, typically, the higher the r_j, the higher the entanglement between these qubits. More technical details regarding this can be found in [Cir+21; Sau+19]. Hence, typically, for states where the entanglement between nearest neighbor qubits is weak, r_j will be very small, scaling as $\mathcal{O}(\text{poly}(n))$. This implies that storing all tensors $G^{(j)}$, which will have space complexity $\mathcal{O}(\text{poly}(n))$, will be way more efficient than storing the full state vector, which has space complexity $\mathcal{O}(2^n)$. Moreover, as mentioned earlier, each entry of $|\psi\rangle$ can be computed using the core tensors with computational cost scaling as $\mathcal{O}(\text{poly}(n))$ as

well. This makes MPS decompositions an attractive data structure to store information regarding such weakly entangled states.

As an example, we give the MPS decomposition of the GHZ state, defined as $|\text{GHZ}_n\rangle = \frac{1}{\sqrt{2^n}}(|\mathbf{0}\rangle + |\mathbf{1}\rangle)$. This is a well-studied and useful state in several quantum information [LWZ14; NLW13; GYW05; EP14; Żuk+98]. The MPS decomposition of $|\text{GHZ}_4\rangle$ is

$$
\begin{aligned}
G_0^{(1)} &= \begin{bmatrix} \frac{1}{\sqrt{2^4}} & 0 \end{bmatrix}, \quad G_0^{(2)} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad G_0^{(3)} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad G_0^{(4)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \\
G_1^{(1)} &= \begin{bmatrix} 0 & \frac{1}{\sqrt{2^4}} \end{bmatrix}, \quad G_1^{(2)} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad G_1^{(3)} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad G_1^{(4)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.
\end{aligned} \tag{2.28}
$$

In general, the core tensors of $|\text{GHZ}_n\rangle$ are

$$
G_i^{(j)} = \begin{cases} |i\rangle\langle i|, & \text{if } j \notin \{1, n\} \\ |i\rangle, & \text{if } j = n \\ \frac{1}{\sqrt{2^4}}\langle i|, & \text{if } j = 1 \end{cases} \tag{2.29}
$$

The formulation provided in Eq (2.26) can be visualized using tensor network diagrams (cf. Section 8.1.2) as given in Figure 2.2. Each box here represents a third order tensor $G^{(j)}$, or equivalently, pairs of matrices $G_0^{(j)}$ and $G_1^{(j)}$. The index on the right denotes the choice of the matrix being selected. The r_j's represent the dimension of the matrices, or equivalently, the bond dimension. So, for any $G^{(j)}$, the line on its right represents its first index (the index of length 2), the line going up represents its second index (the index of length r_j) and the line going down represents its third index (the index of length r_{j+1}). The fact that the third index of one tensor is connected to the second index of another one situated below it represents a tensor contraction of these indices. The nature of the contractions that give rise to each element of the vector as per Eq (2.26) can be visualized here as a linear chain of contractions.

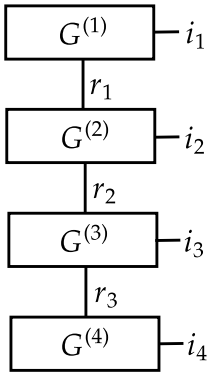


Figure 2.2: (a) Linear tensor network structure of MPS representation of state vectors. Each box represented a 3^{rd} order tensor, or equivalently, pairs of matrices. The dimension of the matrices also called the bond dimensions, is given as r_j. The indices on the right, which can take values in $\{0, 1\}$, represent the qubit indices, or equivalently, the index of the matrices contained in the white box.

The concept of MPS can also be seen as a generalization of product states. The details of this can be found in Section 8.5 in the Appendix.

2.4.1 Basic Operations Involving MPS

Here, we discuss some basic operations that one can do using MPS decompositions of vectors, specifically using their core tensors alone without requiring to expand them as full vectors.

- *Addition* of two vectors $|\psi_0\rangle$ and $|\psi_1\rangle$, given as MPS decompositions with core tensors $G^{(1)}, \ldots, G^{(n)}$ and $H^{(1)}, \ldots, H^{(n)}$ results in the vector $|\phi\rangle$ which has an MPS

decomposition with core tensors $F^{(1)}, \ldots, F^{(n)}$, where

$$F_{i_j}^{(j)} = \begin{cases} \begin{bmatrix} G_{i_j} & H_{i_j} \end{bmatrix}, & \text{if } j = 1 \\ \begin{bmatrix} G_{i_j} \\ H_{i_j} \end{bmatrix}, & \text{if } j = n \\ \begin{bmatrix} G_{i_j} & \mathbf{0} \\ \mathbf{0} & H_{i_j} \end{bmatrix}, & \text{otherwise} \end{cases} \tag{2.30}$$

because

$$|\phi\rangle = \sum_{i_1,\ldots,i_n=0}^{1} F_{i_1}^{(1)} \ldots F_{i_n}^{(n)} |i_1 \ldots i_n\rangle \tag{2.31}$$

$$= \sum_{i_1,\ldots,i_n=0}^{1} \left(G_{i_1}^{(1)} \ldots G_{i_n}^{(n)} + H_{i_1}^{(1)} \ldots H_{i_n}^{(n)} \right) |i_1 \ldots i_n\rangle \tag{2.32}$$

$$= |\psi_0\rangle + |\psi_1\rangle, \tag{2.33}$$

- *Hadamard product* (entrywise multiplication), denoted as $*$, of two vectors $|\psi_0\rangle$ and $|\psi_1\rangle$, given as MPS decompositions with core tensors $G^{(1)}, \ldots, G^{(n)}$ and $H^{(1)}, \ldots, H^{(n)}$ respectively, results in a vector $|\phi\rangle$ with MPS decompostion given by core tensors

$F^{(1)}, \ldots, F^{(n)}$ where $F^{(j)}_{i_j} = G^{(j)}_{i_j} \otimes H^{(j)}_{i_j}$, because

$$|\phi\rangle = \sum_{i_1,\ldots,i_n=0}^{1} F^{(1)}_{i_1} \ldots F^{(n)}_{i_n} |i_1 \ldots i_n\rangle \tag{2.34}$$

$$= \sum_{i_1,\ldots,i_n=0}^{1} \left(G^{(1)}_{i_1} \otimes H^{(1)}_{i_1}\right) \ldots \left(G^{(n)}_{i_n} \otimes H^{(n)}_{i_n}\right) |i_1 \ldots i_n\rangle \tag{2.35}$$

$$= \sum_{i_1,\ldots,i_n=0}^{1} \left(G^{(1)}_{i_1} \ldots G^{(n)}_{i_n}\right) \left(H^{(1)}_{i_1} \ldots H^{(n)}_{i_n}\right) |i_1 \ldots i_n\rangle \tag{2.36}$$

$$= |\psi_0\rangle * |\psi_1\rangle . \tag{2.37}$$

- *Standard inner Product* of two vectors $|\psi\rangle$ and $|\phi\rangle$, given as MPS decompositions with core tensors $G^{(1)}, \ldots, G^{(n)}$ and $H^{(1)}, \ldots, H^{(n)}$ respectively, can be computed as follows:

$$\langle\psi|\phi\rangle = \sum_{i_i,\ldots,i_n=0}^{1} \psi_{i_1 \ldots i_n} \phi_{i_1 \ldots i_n} \tag{2.38}$$

$$= \sum_{i_1 \ldots i_n=0}^{1} G^{(1)}_{i_1} \ldots G^{(n)}_{i_n} H^{(1)}_{i_1} \ldots H^{(n)}_{i_n} \tag{2.39}$$

$$= \prod_{j=1}^{n} \left(\sum_{i_j=0}^{1} G^{(j)}_{i_j} \otimes H^{(j)}_{i_j} \right). \tag{2.40}$$

- *Multiplication* of a vector $|\psi\rangle$, with MPS decomposition given by core tensors $G^{(1)}$, $\ldots, G^{(n)}$, with a scalar γ results in a vector $|\phi\rangle$ with MPS decomposition given by core tensors $F^{(1)}, \ldots, F^{(n)}$, where

$$F^{(j)} = \begin{cases} \gamma G^{(j)}, & \text{if } j = 1 \\ F^{(j)}, & \text{otherwise} \end{cases} . \tag{2.41}$$

because

$$|\phi\rangle = \sum_{i_1,\ldots,i_n=0}^{1} F^{(1)}_{i_1} \ldots F^{(n)}_{i_n} |i_1 \ldots i_n\rangle = \gamma \sum_{i_1,\ldots,i_n=0}^{1} G^{(1)}_{i_1} \ldots G^{(n)}_{i_n} |i_1 \ldots i_n\rangle = \gamma |\psi\rangle .$$

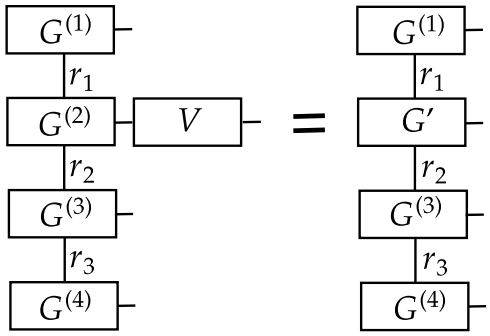


Figure 2.3: Tensor network diagram illustrating an example of a single-qubit gate V being applied on an MPS. The resulting MPS differs from the original MPS only in the core tensor associated with the qubit on which the gate was applied. Also, the new core tensor has the same bond dimension as the one it replaced.

2.4.2 Classical Simulation of Quantum Circuits Using MPS

One important application of MPS is in the classical simulation of quantum circuits. To see how this can be done, we require the ability to do two things: classically simulate the application of a single-qubit gate or a two-qubit gate being applied on nearest neighbor qubits, and classically simulate computational basis measurement, all without needing to know the full state description in some basis. Moreover, we also require the cost of all the operations scaling polynomially in the bond dimension and the total number of qubits involved.

We start with the application of single-qubit gates. Let $\left\{G^{(j)} \in \mathbb{C}^{2\times r_j \times r_{j+1}}\right\}_j$ constitute an MPS description of some state $|\psi\rangle$, and let $V = \sum_{p,q=0}^{1} V_{pq} |p\rangle\langle q|$ be a single qubit gate. We will discuss how the application of V on a qubit i other than the first or last can be simulated. It is trivial to extend this to those qubits. The application of V on the i^{th}

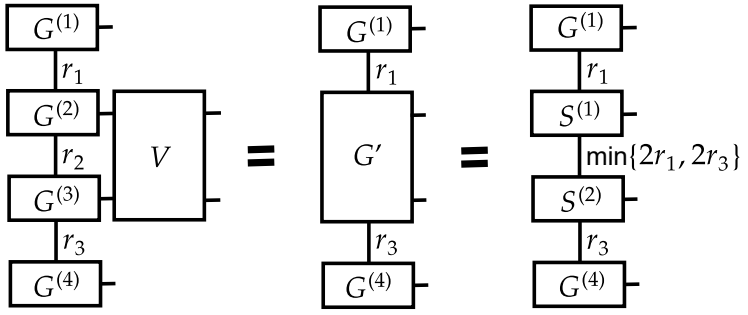


Figure 2.4: Tensor network diagram illustrating an example of a two-qubit gate V being applied on an MPS. After the tensor contractions, we can replace the core tensors associated with both qubits on which the gate is being applied with a single order 4 tensor G'. Then, we use SVD to factorize or "break" G' into the contraction of two order 3 tensors which are the reshaped versions of the resulting matrices $S^{(1)}$ and $S^{(2)}$. The final MPS differs from the original MPS only in the two core tensors associated with the qubits on which the gate was applied. Also, the new bond dimension is the minimum of twice the neighboring bond dimensions, which is a result of the SVD factorization involved.

qubit of $|\psi\rangle$ gives us

$$V_i |\psi\rangle = \sum_{j_1,\ldots,j_n=0}^{1} G^{(1)}_{j_1} \ldots G^{(n)}_{j_n} |j_1 \ldots j_{i-1}\rangle \left(\sum_{k=0}^{1} V_{kj_i} |k\rangle \right) |j_{i+1} \ldots j_n\rangle \tag{2.42}$$

$$= \sum_{j_1,\ldots,j_n,k=0}^{1} G^{(1)}_{j_1} \ldots G^{(i-1)}_{j_{i-1}} V_{kj_i} G^{(i)}_{j_i} G^{(i-1)}_{j_{i+1}} \ldots G^{(n)}_{j_n} |j_1 \ldots j_{i-1} k j_{i+1} \ldots j_n\rangle \tag{2.43}$$

$$= \sum_{j_1,\ldots,j_n=0}^{1} G^{(1)}_{j_1} \ldots G^{(i-1)}_{j_{i-1}} G'_{j_i} G^{(i+1)}_{j_{i+1}} \ldots G^{(n)}_{j_n} |j_1 j_2, \ldots j_n\rangle, \tag{2.44}$$

where $G' \in \mathbb{C}^{2 \times r_i \times r_{i+1}}$ with $G'_{j_i,p,q} = \sum_{j=0}^{1} V_{j_i j} G^{(i)}_{j,p,q}$. All we have done here is that we have contracted the second index of V and the first index of $G^{(i)}$. So the tensors $G^{(1)}, \ldots, G^{(i-1)}, G', G^{(i+1)}, \ldots, G^{(n)}$ make up an MPS decompostion of $V_i |\psi\rangle$. All we had to do was replace $G^{(j)}$ with G'. Computation of G' is dependent only on $G^{(i)}$ and V and the cost of this operation scales as $\mathcal{O}(r_i r_{i+1})$. This process is illustrated using tensor networks in Figure 2.3.

Now, let's move on to the application of two-qubit gates on nearest neighbor qubits. We shall only explain the case when it is applied to any nearest neighbor pair that does

not involve the first or last qubits. As mentioned earlier, the same approach can be easily extended when the gate is applied on the first or last qubit. Let

$$V = \sum_{p_1,p_2,q_1,q_2=0}^{1} V_{p_1p_2,q_1q_2} \, |p_1p_2\rangle\langle q_1q_2| \tag{2.45}$$

be a two-qubit gate. Then, the application of V on qubits $i, i+1$ gives us

$$\begin{aligned}
& V_{(i,i+1)} \, |\psi\rangle \\
&= \sum_{j_1,\ldots,j_n=0}^{1} G^{(1)}_{j_1} \ldots G^{(n)}_{j_n} \, |j_1 \ldots j_{i-1}\rangle \left(\sum_{k_1,k_2=0}^{1} V_{k_1k_2,j_ij_{i+1}} \, |j_i\rangle |j_{i+1}\rangle \right) |j_{i+2} \ldots j_n\rangle \\
&= \sum_{j_1,\ldots,j_n,k_1,k_2=0}^{1} G^{(1)}_{j_1} \ldots G^{(i-1)}_{j_{i-1}} V_{k_1k_2,j_ij_{i+1}} G^{(i)}_{j_i} G^{(i+1)}_{j_{i+1}} \ldots G^{(n)}_{j_n} \, |j_1 \ldots j_{i-1}k_1k_2j_{i+2}, \ldots j_n\rangle \\
&= \sum_{j_1,\ldots,j_n=0}^{1} G^{(1)}_{j_1} \ldots G^{(i-1)}_{j_{i-1}} G'_{j_i,j_{i+1}} G^{(i+1)}_{j_{i+1}} \ldots G^{(n)}_{j_n} \, |j_1 \ldots j_n\rangle,
\end{aligned}$$

where $G' \in \mathbb{C}^{2\times 2\times r_i \times r_{i+2}}$. In this case, to get G' we have contracted the first indices of $G_{(i)}$ and $G_{(i+1)}$ with the column index of V, split into two indices of length 2. But the issue now is that G' is not an order 3 tensor. To get an MPS structure, we need to split G' into two order 3 tensors lying in $\mathbb{C}^{2\times r_i \times r}$ and $\mathbb{C}^{2\times r\times r_{i+2}}$ such that contracting the third index of the first and the second index of the second would give G', for some $r \in \mathbb{N}$.

One way to do this is by computing the Singular Value Decomposition (SVD) of the matrix $G'' \in \mathbb{C}^{2r_i\times 2r_{i+2}}$ that we get by reshaping G' into a matrix, where we combine the odd and even indices of G'. The SVD of this matrix allows us to decompose G'' as $G'' = S^{(1)}S^{(2)}$, where $S^{(1)} \in \mathbb{C}^{2r_i\times r}$ and $S^{(2)} \in \mathbb{C}^{r\times 2r_{i+2}}$, and $r \leqslant \min\{2r_i, 2r_{i+2}\}$. By reshaping $S^{(1)}$ into an order 3 tensor of the form $S^{(1)'} \in \mathbb{C}^{2\times r_i\times r}$ (by splitting the row index into two) and $S^{(2)}$ into an order 3 tensor of the form $S^{(2)'} \in \mathbb{C}^{2\times r\times r_{i+2}}$ (by splitting the column index and permuting the indices accordingly), we get the required two order 3 tensors. So, to get the MPS decomposition after the application of the two-qubit gate, we simply need to replace $G^{(i)}$ and $G^{(i+1)}$ with $S^{(1)'}$ and $S^{(2)'}$ respectively. The cost of doing this operation scales as $\mathcal{O}(\text{poly}(r_i, r_{i+1}, r_{i+2}))$. This process is illustrated using

tensor networks in Figure 2.4.

So far we have discussed a method that can be used to apply a two-qubit gate V on neighboring qubits. If we want to apply it on qubits i and j with $i < j$ that are not neighbors, we can first apply a sequence of SWAP gates on pairs of neighboring qubits $(i, i+1), (i+1, i+2), \ldots, (j-1, j)$. Now, the qubits i and j have become neighbors and we can simulate the application of V. After that, we apply SWAP gates on $(j-1, j), (j-2, j-1), \ldots, (i, i+1)$ to get i back to its original position.

Finally, we move on to measurement. Notice that the single qubit gate operation we described previously does not necessitate the matrix being applied to be unitary. That is, we can compute the MPS decomposition of $V_i |\psi\rangle$ for any $V \in \mathcal{L}(\mathbb{C}^2)$ not necessarily a unitary efficiently. This is crucial in simulating computational basis measurements classically. Assume we want to measure qubit i. This means that after the measurement, we should get

$$\frac{1}{\| |0\rangle\langle 0|_i |\psi\rangle \|_2} |0\rangle\langle 0|_i |\psi\rangle \text{ with probability } \| |0\rangle\langle 0|_i |\psi\rangle \|_2^2, \tag{2.46}$$

$$\frac{1}{\| |1\rangle\langle 1|_i |\psi\rangle \|_2} |1\rangle\langle 1|_i |\psi\rangle \text{ with probability } \| |1\rangle\langle 1|_i |\psi\rangle \|_2^2. \tag{2.47}$$

So, to simulate measurement, we must be able to compute the probabilities $\left\{\| |0\rangle\langle 0|_i |\psi\rangle \|_2^2, \| |1\rangle\langle 1|_i |\psi\rangle \|_2^2\right\}$ and compute the associated post-measurement states as well. To compute the probability associated with $|0\rangle\langle 0|_i |\psi\rangle$, we first compute the MPS decomposition of the state using the method we described for single-qubit gates. After that, we have to compute the 2-norm of the resultant vector, which is simply an inner product between two MPS decompositions. Moreover, to compute the post-measurement states, all we have to do is divide these vectors by the square root of their associated probabilities. Both these operations can be done by the methods we described in Section 2.4.1.

The initial state in many quantum circuits is assumed to $|\mathbf{0}\rangle$, which is a separable state. Now, given 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 $j \leqslant i \leqslant k$. Let $R_V = \max_i R_{V,i}$. Then we have the following theorem, first introduced in [Joz06].

Theorem 1. *Every quantum circuit V that has $|\mathbf{0}\rangle$ as the initial state and involves computational basis measurements can be classically simulated using MPS with cost $\mathcal{O}(n \cdot poly(2^{R_V}))$.*

Proof. The input state has a trivial MPS decomposition with all bond dimensions 1. First, we shall absorb all single-qubit gates into their closest two-qubit gates so that the resultant circuit contains only two-qubit gates. That is, when a single qubit gate A has to be applied right after (or right before) a two-qubit gate B has to be applied, on a qubit that B is being applied on, we can combine A and B as $(A \otimes \mathbb{1})B$ or $(\mathbb{1} \otimes A)B$ ($B(A \otimes \mathbb{1})$ or $B(\mathbb{1} \otimes A)$). Now we replace all two-qubit gates that are not being applied on neighboring qubits with only gates acting on neighboring qubits using SWAP gates. For every such replacement, R_V can only increase by a maximum of 4.

Then we shall start applying each two-qubit gate one by one on the input state $|\mathbf{0}\rangle$. Let $G^{(1)}, \ldots, G^{(n)}$ make up the MPS decomposition of the state at some point in this process, with bond dimensions $r_1, \ldots r_n$. Assume that we have to apply a two-qubit gate on some pair of qubits $i, i+1$ on this state. The resultant state will have an MPS decomposition with bond dimensions $r_1, \ldots, r_i, r, r_{i+2}, \ldots, r_n$, with $r = \min\{2r_i, 2r_{i+2}\}$. So each application of a two-qubit gate on a qubit can potentially double the resources required to store the core tensors associated with the qubits on which it was applied. R_{V_i} tracks how many two-qubit gates are applied on the qubit i, after introducing SWAP gates into the circuit. Since that introduction can only increase all the R_{V_i} by a constant number, the overall cost of implementing all the gates will be $\mathcal{O}(2^{R_V})$. The fact that measurement can be carried out efficiently completes the proof. □

2.4.3 MPS Decomposition Algorithm

In this section, we discuss a method that can be used to find an MPS decomposition of a quantum state $|\psi\rangle$ whose description in the computational basis is given. For that, we first

define the concept of matricization. Given a tensor $X \in \mathbb{C}^{I_1 \times \cdots \times I_k}$, its j^{th} matricisation $X_{[j]} \in \mathbb{C}^{I_1 I_2 \ldots I_j \times I_{j+1} I_{j+2} \ldots I_k}$ is the matrix that we get when we reshape X into a matrix with indices $1, 2, \ldots, j$ featuring as rows and the rest as columns.

The decomposition algorithm can be seen as an extended version of the SVD subroutine featured in the two-qubit gate application procedure. The SVD was used to break a tensor of order 4 into a contraction of two order 3 tensors. Similarly, we shall sequentially break the state, which is an order n tensor, into multiple contractions of order 3 tensors.

Let $\psi_{[j]}$ be the j^{th} matricisation of ψ. The procedure starts with an SVD of $\psi_{[1]}$, giving a decomposition of the form $\psi_{[1]} = U^{(1)} U^{(2)}$, where $U^{(1)} \in \mathbb{C}^{2 \times r_2}$ and $U^{(2)} \in \mathbb{C}^{r_2 \times 2^{n-1}}$. The $U^{(1)}$ here is the first core tensor. To get the subsequent core tensors, we simply carry out SVD of $U^{(2)}$, and the left singular matrix that we get can be reshaped into the next core tensor and so on. The full algorithm is provided in Algorithm 1 and is illustrated using tensor networks in Figure 2.5.

So far, we have discussed how the MPS data structure is used to store and work with pure state described in vector form efficiently. MPS can be similarly used for density matrices as well. The details of this can be found in Section 8.4 in the Appendix.

Algorithm 1 MPS decompostion algorithm

Require: A tensor $|\psi\rangle \in \mathbb{C}^{2 \times \cdots \times 2}$
Ensure: Core tensors $G^{(1)}, G^{(2)}, G^{(3)}, \ldots, G^{(N-1)}, G^{(N)}$

1: Compute the reduced SVD of $\psi_{[1]}$ to get a decomposition of the form $\psi_{[1]} = U^{(1)} U^{(2)}$, where $U^{(1)} \in \mathbb{C}^{2 \times r_2}$ and $U^{(2)} \in \mathbb{C}^{r_2 \times 2^{n-1}}$ and $r_2 \leqslant 2$.
2: Set $G^{(1)}$ as $U^{(1)}$ reshaped into a tensor of shape $2 \times 1 \times r_2$.
3: Split the columns indices of $U^{(2)}$ into $n-1$ indices of length 2 to get $A \in \mathbb{C}^{r_2 \times 2 \times \cdots \times 2}$ such that $A_{[1]} = U^{(2)}$.
4: **for** $k = 2 \ldots n-1$ **do**
5: Compute the reduced SVD of $A_{[2]}$ to get a decomposition of the form $A_{[2]} = U^{(1)} U^{(2)}$, where $U^{(1)} \in \mathbb{C}^{2 r_k \times r_{k+1}}$, $U^{(2)} \in \mathbb{C}^{r_{k+1} \times 2^{n-k}}$ and $r_{k+1} \leqslant \min\{2 r_k, 2^{n-k}\}$.
6: Set $G^{(k)}$ as $U^{(1)}$ reshaped as a order 3 tensor with shape $2 \times r_k \times r_{k+1}$, by splitting its row index.
7: Split the columns indices of $U^{(2)}$ into $n-k$ indices of length 2 to get $A \in \mathbb{C}^{r_{k+1} \times 2 \times \cdots \times 2}$ such that $A_{[1]} = U^{(2)}$.
8: **end for**
9: Set $G^{(n)}$ as A reshaped into a order 3 tensor of shape $2 \times r_{n-1} \times 1$.
10: **return** $G^{(1)}, G^{(2)}, \ldots, G^{(n-1)}, G^{(n)}$.

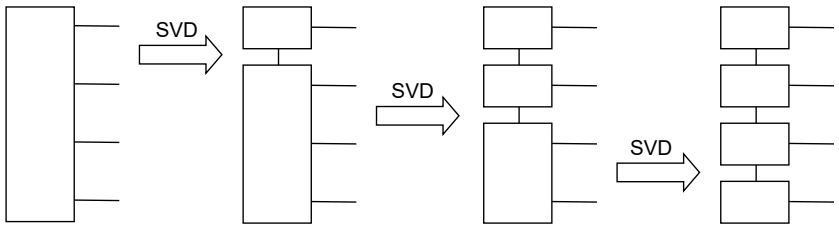


Figure 2.5: Tensor network diagram illustrating an example of the MPS decomposition algorithm. SVD is used to sequentially break the order n tensor into a linear chain of order 3 tensors, similar to how SVD was used to break an order 4 tensor into a chain of two order 3 tensors within the two-qubit gate application procedure.

2.4.4 MPS Tomography

Quantum tomography is a subfield of quantum information that deals with techniques designed to compute the classical description of an unknown state σ, when multiple copies are provided [Gro+10; OW16; Haa+16; FBK21; Guţ+20]. It has already been shown that in general, to get a classical description of an arbitrary state σ, to any fidelity, we require a number of copies of σ that scales exponentially in the number of qubits involved [Yue23].

But if the target state is pure and admits an MPS description with all bond dimensions small, say bounded by a small number B, then [Cra+10] introduces two tomography procedures that would consume a number of copies scaling linearly in n and polynomial in B [Cra+10]. In this section, we shall explain one of them in detail.

Let $|\psi\rangle$ be the target pure state, which has such an efficient MPS decomposition. Define $Q = \lceil \log B \rceil + 1$. The method starts with estimating the state of the first Q qubits $\mathcal{J} = (1, 2, \ldots, Q)$ using any conventional quantum tomography technique. This would require $\mathcal{O}(4^Q)(\mathcal{O}(\text{poly}(B)))$ copies of σ.

For simplicity, assume we have a perfect estimation of $\sigma_{\mathcal{J}}$. Let $\sigma_{\mathcal{J}} = \sum_{j=1}^{\text{rank}(\sigma_{\mathcal{J}})} \omega_j |w_j\rangle\langle w_j|$ be its eigendecomposition. From Algorithm 1, we can see that the i^{th} bond dimension is the rank of $\psi_{[i]}$. Also, notice that $\sigma_{\mathcal{J}} = \psi_{[i]}\psi_{[i]}^\dagger$, meaning that the i^{th} bond dimension is the same as $\text{rank}(\sigma_{\mathcal{J}})$. Since $\text{rank}(\sigma_{\mathcal{J}}) \leqslant 2^{Q-1}$, there exist a Q-qubit unitary $U^{(1)}$ such that $U^{(1)}\sigma_{\mathcal{J}}{U^{(1)}}^\dagger = |0\rangle\langle 0| \otimes \sigma'$ for some state $\sigma' \in \mathcal{L}(\mathbb{C}^{2^{Q-1}})$. That is, application of

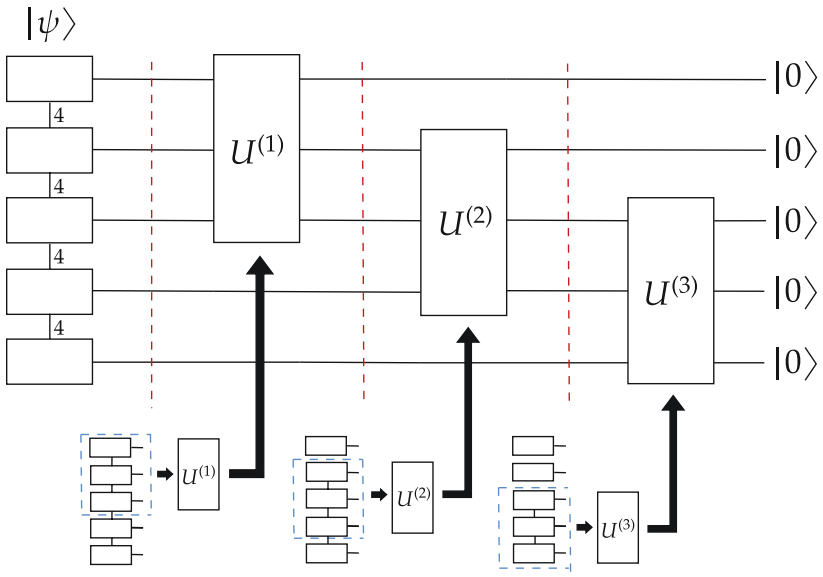


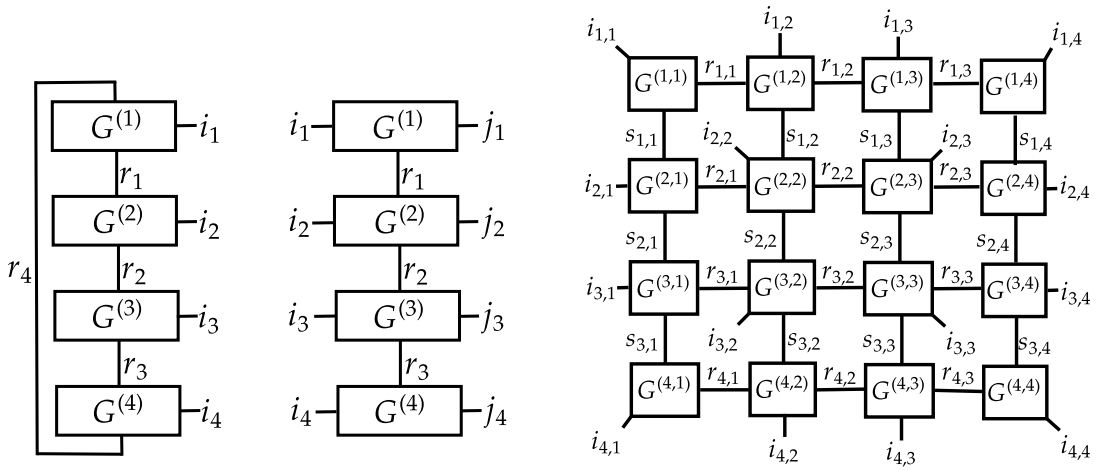
Figure 2.6: Tensor network diagram illustrating a 5-qubit example of the MPS tomography protocol. All the bond dimensions of the input state $|\psi\rangle$ is 4. Hence, we only require tomography of 3-qubit subsystems of $|\psi\rangle$. After the first tomography subroutine, we can compute $U^{(1)}$. Application of $U^{(1)}$ results in a state with the first qubit being $|0\rangle$. Then, we carry out tomography of the next 3 qubits, from which we can compute $U^{(2)}$. Application of $U^{(1)}$ followed by $U^{(2)}$ results in a state with the first two qubits being in $|0\rangle$. Since the resulting part is simply a 3-qubit state, we carry out tomography of that part to find a circuit $U^{(3)}$ that prepares it. Hence, application of the inverse of all these gates in reverse, that is, ${U^{(3)}}^\dagger$ on the bottom three qubits, ${U^{(2)}}^\dagger$ on the next three, and ${U^{(1)}}^\dagger$ on the first three qubits, of $|\mathbf{0}\rangle$, will prepare $|\psi\rangle$.

$U^{(1)}$ on the first Q qubits "disentangles" the first qubit and results in a state of the form $|0\rangle \otimes |\psi'\rangle$, for some $|\psi'\rangle \in \mathbb{C}^{2^{n-1}}$.

Repeat this entire process again for $|\psi'\rangle$. Then, repeat the whole thing for the corresponding resultant states $n-Q$ times, to get unitaries $U^{(1)}, U^{(2)}, \ldots, U^{(n-Q)}$. After $n-Q$ rounds, we will get

$$U^{(n-Q)}_{(n-Q,n-Q+1\ldots,n-1)} \cdots U^{(2)}_{(2,3\ldots,Q+1)} U^{(1)}_{(1,2\ldots,Q)} |\psi\rangle = |0\rangle^{\otimes n-Q} \otimes |\psi''\rangle, \tag{2.48}$$

where $|\psi''\rangle$ is a Q qubit state. Now, do full tomography on $|\psi''\rangle$ to find a unitary $U^{(n-Q+1)}$ such that $U^{(n-Q+1)} |\psi''\rangle = |0\rangle^{\otimes Q}$. Hence, we see that if we apply the inverses of all these



(a) Cyclic MPS (b) MPO (c) PEPS

Figure 2.7: Extensions of MPS. (a) Cyclic MPS, which is a typical MPS with the first and last core tensors connected with an additional edge. i_k, r_k represent the free qubit indices and bond dimensions respectively. (b) MPO, where each core tensor has an extra index j_k making it a $2^n \times 2^n$ dimensional matrix. (c) PEPS representing a grid of qubits, where r_{k_1,k_2}, s_{k_1,k_2} are bond dimensions.

gates in reverse, we will get a circuit that prepares $|\psi\rangle$.

Let U be the combined circuit. Then, since R_U is bounded by B (cf. Theorem 1), classically simulating this circuit using MPS simulation will result in an MPS decomposition of $|\psi\rangle$ with all bond dimensions bounded by B. A 5-qubit example of the protocol is illustrated in Figure 2.6.

The method described previously assumed that each of the tomography procedures was perfect. In reality, this is not possible. But it turns out that all errors accrued as part of such tomography procedures add up only linearly in n. More details regarding this can be found in [Cra+10].

2.4.5 Cyclic MPS and Beyond

There are many extensions and variations of MPS used within quantum information and tensor networks. One example that features in many areas within this thesis is the *cyclic*

MPS. In this framework, an n-qubit state $|\psi\rangle$ is decomposed as

$$|\psi\rangle = \sum_{i_1,\ldots,i_n=0}^{1} \operatorname{tr}\left(G_{i_1}^{(1)} \ldots G_{i_n}^{(n)}\right) |i_1 \ldots i_n\rangle \tag{2.49}$$

$$= \sum_{i_1,\ldots,i_n=0}^{1} \left(\sum_{j_1,j_2,\ldots,j_{n-1},j_n=0}^{r_1-1,r_2-1,\ldots,r_{n-1}-1,r_n-1} G_{i_1,j_1,j_2}^{(1)} \ldots G_{i_n,j_{n-1},j_n}^{(n-1)} G_{i_n,j_n,j_1}^{(n)} \right) |i_1 \ldots i_n\rangle, \tag{2.50}$$

where $G^{(j)} \in \mathbb{C}^{2\times r_j \times r_{j+1}}$ are the core tensors and $r_1 = r_{n+1}$ but need not be 1. In Figure 2.7 (a), we illustrate this structure using tensor networks.

Two other related examples of extensions in quantum information are the Matrix Product Operator (MPO) [Pir+10; HMS17; Kel+15; Cha+16; KGE14] and Projected Entangled Pair States (PEPS) [SPC11; Sch+13; Per+07; Sch+12; LCB14]. The former, illustrated in Figure 2.7 (b), is a decomposition of a $2^n \times 2^n$ matrix with each core tensor having an extra free index. The latter, illustrated in Figure 2.7 (c), can be used to decompose a system of qubits arranged (interacting) in a grid-like structure, with multiple bond dimensions representing horizontal and vertical interactions. All the properties and operations involving MPS can be easily extended to these scenarios using trivial modifications.

2.5 Unitary t-Designs

Consider a finite collection of m unitary operators $\mathcal{F} = \left\{U^{(i)} \in \mathbb{C}^d\right\}_i$. Associated with this collection, we can define a super operator $\Phi_{\mathcal{F}}^{(t)} : \mathcal{L}\left(\mathbb{C}^{d^t}\right) \to \mathcal{L}\left(\mathbb{C}^{d^t}\right)$ on quantum states describing t-qudit systems, where

$$\Phi_{\mathcal{F}}^{(t)}(W) = \frac{1}{m} \sum_{i=1}^{m} \left(\bigotimes_{j=1}^{t} U^{(i)} \right) W \left(\bigotimes_{j=1}^{t} U^{(i)\dagger} \right). \tag{2.51}$$

One can see that Φ is a valid quantum channel with Kraus operates $\left\{\frac{1}{\sqrt{m}} U^{(i)}\right\}_i$. This is a channel that can be easily implemented as well. One just has to sample a unitary from the collection uniformly and then apply the unitary on each of the t-qudits.

One can also consider the collection to be all unitary matrices from the unitary group $\mathbb{U}_d$. In this case, we sample unitaries based on the uniform Haar measure on the unitary group [CŚ06; Wat18]. So, the super operator associated with this collection $\Phi_{\mathbb{U}_d}$ is defined as

$$\Phi_{\mathbb{U}_d}^{(t)}(W) = \int_U \left(\bigotimes_{j=1}^{t} U\right) W \left(\bigotimes_{j=1}^{t} U^\dagger\right) \mathrm{d}U, \tag{2.52}$$

where the integral is taken over the uniform Haar measure. One way of implementing $\Phi_{\mathbb{U}_d}^{(t)}$ is by sampling a unitary $U \in \mathbb{U}_d$ according to the Haar measure and applying it on all t-registers. One can classically carry out this sampling, or generate a uniformly random unitary matrix using QR factorization [Mez06]. Carrying out such sampling in a quantum computer is inefficient, that is, it will require resources exponential in the number of qubits involved [Dan+09].

But $\Phi_{\mathbb{U}_d}^{(t)}$ can be implemented without sampling from the unitary group according to the Haar measure. To be precise, we can have other finite sets $\mathcal{F}$ of unitary operators such that

$$\Phi_{\mathcal{F}}^{(t)} = \Phi_{\mathbb{U}_d}^{(t)}. \tag{2.53}$$

These sets are called *unitary t-designs*. It is clear that a unitary t-design is always a unitary $(t-1)$-design as well, since the fact that it should work for all $W \in \mathcal{L}(\mathbb{C}^{d^t})$ will imply that it should work for all operators of the form $\mathbb{1}_{\mathbb{C}^d} \otimes W'$ as well, where $W' \in \mathcal{L}(\mathbb{C}^{d^{t-1}})$. From (2.53), one derives another useful equivalent characterization of unitary t-designs. A finite set $\mathcal{F}$ of m unitary operators is a t-design if and only if for every homogeneous polynomial $P_{(t,t)}$ which has degree at most t in the matrix elements of U and at most t is the complex conjugate of these elements, we have

$$\frac{1}{m}\sum_{i=1}^{m} P_{(t,t)}\left(U^{(i)}\right) = \int_U P_{(t,t)}(U)\ \mathrm{d}U. \tag{2.54}$$

To see why, first assume that $\Phi_{\mathbb{U}_d}^{(t)} = \Phi_{\mathcal{F}}^{(t)}$. Let $|i\rangle \in \mathbb{C}^{d^t}$ be a computational basis vector with $|i\rangle = |i_1\rangle|i_2\rangle \ldots |i_t\rangle$ and $|i_j\rangle$ is a computational basis vector in $\mathbb{C}^d$, for all $j \in \{1, \ldots, t\}$.

Then, we know that

$$\langle i| \int_U \left(\bigotimes_{p=1}^{t} U \right) |j\rangle\langle k| \left(\bigotimes_{p=1}^{t} U^\dagger \right) \mathrm{d}U \, |l\rangle = \langle i| \frac{1}{m} \sum_{q=1}^{m} \left(\bigotimes_{p=1}^{t} U^{(q)} \right) |j\rangle\langle k| \left(\bigotimes_{p=1}^{t} U^{(q)^\dagger} \right) |l\rangle . \tag{2.55}$$

This implies

$$\int_U \left(\bigotimes_{p=1}^{t} \langle i_p| U |j_p\rangle \right) \left(\bigotimes_{p=1}^{t} \langle k_p| U^\dagger |l_p\rangle \right) \mathrm{d}U = \frac{1}{m} \sum_{q=1}^{m} \left(\bigotimes_{p=1}^{t} \langle i_p| U^{(q)} |j_p\rangle \right) \left(\bigotimes_{p=1}^{t} \langle k_p| U^{(q)^\dagger} |l_p\rangle \right), \tag{2.56}$$

and furthermore,

$$\int_U \left(\prod_{p=1}^{t} U(i_p, j_p) \right) \left(\prod_{p=1}^{t} \overline{U(l_p, k_p)} \right) \mathrm{d}U = \frac{1}{m} \sum_{q=1}^{m} \left(\prod_{p=1}^{t} U^{(q)}(i_p, j_p) \right) \left(\prod_{p=1}^{t} \overline{U^{(q)}(l_p, k_p)} \right). \tag{2.57}$$

This means that the required condition is satisfied for every monomial $M_{(t,t)}$ and hence by linearity of sum and integral, it will be satisfied for every homogeneous polynomial $P_{(t,t)}$. To prove the converse, since the condition is satisfied for every homogenous polynomial $P_{(t,t)}$, it should be satisfied for every monomial $M_{(t,t)}$, which means that it should be satisfied for every computational basis element in $\mathcal{L}(\mathbb{C}^{d^t})$, which combined with linearity, completes the proof.

From this definition, we can see that sampling uniformly from a unitary t-design is equivalent to sampling from the unitary group according to Haar measure up to t statistical moments, since the t^{th} statistical moment is a polynomial of degree t. This property makes unitary t-designs particularly useful in many areas of quantum information where generation of a uniformly sampled random unitary is required, such as [AS04; RRS06; Bou+19; Sze+11; Gro11].

There is a similar concept for quantum states called the *state t-design* [HL09a], which uses the spherical measure. We define a state t-design as a set of states $\{|\phi_1\rangle, |\phi_2\rangle, \ldots, |\phi_m\rangle\}$

such that

$$\int_{\psi} \left(\bigotimes_{j=1}^{t} |\psi\rangle\langle\psi| \right) \mathrm{d}\psi = \frac{1}{m} \sum_{i=1}^{m} \left(\bigotimes_{j=1}^{m} |\psi_i\rangle\langle\psi_i| \right), \tag{2.58}$$

where $\mathrm{d}\psi$ is the uniform spherical measure [Wat18] defined over normalized vectors. Also, a set $\{\psi_i\}_i$ is a state t-design if and only if for every homogeneous polynomial $P_{(t,t)}$ which has degree at most t in elements of $|\psi\rangle$ and at most t in the complex conjugate of these elements, we have

$$\frac{1}{m} \sum_{i=1}^{m} P_{(t,t)} \left(|\psi_i\rangle\langle\psi_i| \right) = \int_{\psi} P_{(t,t)} \left(|\psi_i\rangle\langle\psi_i| \right) \ \mathrm{d}\psi. \tag{2.59}$$

An important result connecting unitary t-designs and spherical t-designs is that for a unitary t-design $\left\{U^{(i)}\right\}_i$, the set $\left\{U^{(i)} |\mathbf{0}\rangle\right\}_i$ is a spherical t-design [Wat18]. Hence, unitary t-designs are more "powerful" than state t-designs in the sense that they can be used to generate state t-designs and do other things as well. [SZ84] have proved the existence of unitary t-designs for any value of d and t, But for arbitrary values of t, sampling from such t-designs is still an exponentially expensive procedure in terms of the number of qubits [Nak+21], except for some specific cases [RS09; Ban+18; Ban+19].

A slightly different definition of unitary t-designs is considered in [HL08]. Consider a set of unitaries $\mathcal{F} = \{U_1, U_2 \ldots, U_m\}$ and a probability vector $p = [p_1 \ p_2 \ \ldots \ p_m]$. Let

$$\Phi^{(t)}_{(\mathcal{F},p)}(W) = \sum_{i=1}^{m} p_i \left(\bigotimes_{j=1}^{t} U^{(i)} \right) W \left(\bigotimes_{j=1}^{t} {U^{(i)}}^{\dagger} \right). \tag{2.60}$$

The ensemble $(\mathcal{F}, p)$ is a unitary t-design if and only if $\Phi^{(t)}_{(\mathcal{F},p)} = \Phi^{(t)}_{\mathbb{U}_d}$. The ensemble can be generalized to probability distributions on the group of unitaries. Consider a probability distribution $g(U)$ defined over the unitary group and let

$$\Phi^{(t)}_{g}(W) = \int_{U} \left(\bigotimes_{j=1}^{t} U \right) W \left(\bigotimes_{j=1}^{t} U^{\dagger} \right) \mathrm{d}g(U). \tag{2.61}$$

Then, the ensemble generated by g is a unitary t-design if and only if $\Phi_g^{(t)} = \Phi_h^{(t)}$. Here, h is the Haar measure over the unitary group. When we talk about randomized circuit constructions which generate a unitary sample from a distribution that is, up to t moments, similar to the Haar distribution, we generally use this definition of a unitary t-design to analyze the protocol.

In many applications, one might not necessarily require exact unitary t-designs, but a good controllable approximation will do. An ϵ-approximate unitary t-design is an ensemble generated by a distribution g, such that $\left\|\Phi_g^{(t)} - \Phi_h^{(t)}\right\|_\diamond \leqslant \epsilon$. Here, $\|\cdot\|_\diamond$ is the diamond norm [Wat18], which is generally used to assess the "distance" between different quantum channels, and as a fundamental measure in problems involving distinguishing quantum channels. Approximate unitary t-designs can be constructed with much fewer resources compared to exact unitary t-designs [HL09b; Haf+20]. Another interesting line of research is that certain types of randomly generated quantum circuits eventually converge to approximate unitary t-designs [Ćwi+12; HM18], at depth $\mathcal{O}(\text{poly}(n))$ with degree bounded by t.

Now, we shall recall a useful and important property of the Haar measure from [Sep06]: for any (integrable) function η, we have

$$\int_U \eta(U)\mathrm{d}U = \int_U \eta(U^\dagger)\mathrm{d}U. \tag{2.62}$$

This means that for any matrices $A, B \in \mathcal{L}(\mathbb{C}^{d^t})$, we have

$$\operatorname{tr}\left(A^\dagger \Phi_h^{(t)}(B)\right) = \int_U \operatorname{tr}\left(A^\dagger \left(\bigotimes_{j=1}^t U\right) B \left(\bigotimes_{j=1}^t U^\dagger\right)\right) \mathrm{d}g(U) \tag{2.63}$$

$$= \int_U \operatorname{tr}\left(A^\dagger \left(\bigotimes_{j=1}^t U\right) B \left(\bigotimes_{j=1}^t U^\dagger\right)\right) \mathrm{d}U \tag{2.64}$$

$$= \int_U \operatorname{tr}\left(\left(\bigotimes_{j=1}^t U^\dagger\right) A^\dagger \left(\bigotimes_{j=1}^t U\right) B\right) \mathrm{d}U \tag{2.65}$$

$$= \int_U \operatorname{tr}\left(\left(\bigotimes_{j=1}^t U\right) A^\dagger \left(\bigotimes_{j=1}^t U^\dagger\right) B\right) \mathrm{d}U \tag{2.66}$$

$$= \operatorname{tr}\left(\Phi_h^{(t)}(A)^\dagger B\right), \tag{2.67}$$

implying that $\Phi_h^{(t)}$ is a Hermitian operator. Moreover, with a bit more effort, we can also show that $\Phi_h^{(t)}$ is a projection [HL09a].

Identifying the eigenspace with eigenvalue1 1 of $\Phi_h^{(t)}$, which is simply the space spanned by all the permutation operators of the form W given in (8.2) defined for qudits [HL08], is key to solving integrals of polynomials of Haar random unitaries of degree (t, t). The recipe is as follows, let the eigendecompostion of the matrix form of $\Phi_h^{(t)}$ be $\sum_i |W_i\rangle\langle W_i|$, where W_i are the eigenvectors. Then, we know that every monomial will have the form $\Phi_h^{(t)}(|i\rangle\langle j|)$ and replicating this operation in the bigger vector space with $\sum_i |W_i\rangle\langle W_i|$ will give us the answer. The solutions for common polynomials for $t = 1$ and $t = 2$ are given in Lemma 6 and Lemma 5. More details about such integration techniques can be found in [HL08; CŚ06].

An important class of unitary t-designs used and studied extensively in the literature is the unitary 2-design. These sets of unitaries are capable of simulating sampling from Haar distribution over the unitaries up to 2 moments. So, they are similar in terms of mean and variance. They are the most commonly used approximations for sampling from the unitary group.

Let $\Psi : \mathcal{L}(\mathbb{C}^d) \to \mathcal{L}(\mathbb{C}^d)$ be a quantum channel and g a distribution over $\mathbb{U}_d$. Consider

the function $\mathrm{Tw}_{(g,\Psi)} : \mathcal{L}(\mathbb{C}^d) \to \mathcal{L}(\mathbb{C}^d)$ defined as

$$\mathrm{Tw}_{(g,\Psi)}(W) = \int_U U^\dagger \Psi(UWU^\dagger)U \, \mathrm{d}f(U) = \int_U \sum_i U^\dagger A^{(i)} UWU^\dagger A^{(i)\dagger} U \, \mathrm{d}g(U),$$

where $\left\{A^{(i)}\right\}$ is the Kraus operators of Ψ. Condition (2.54), which works for continuous distributions also, clearly tells us that if g is a 2-design, then $\mathrm{Tw}_{(g,\Psi)} = \mathrm{Tw}_{(h,\Psi)}$, even if Ψ was an arbitrary super operator and not necessarily a quantum channel. So, to show the converse, similar to the proof of the previous condition, we just apply $\mathrm{Tw}_{(g,\Psi_{ij})}$ to computational basis vectors, where $\Psi_{ij}(W) = |i_1\rangle\langle i_2| W |j_1\rangle\langle j_2|$ and $|i\rangle = |i_1\rangle|i_2\rangle, |j\rangle = |j_1\rangle|j_2\rangle$.

The channel $\mathrm{Tw}_{(g,\Psi)}$ can be implemented as follows: apply a unitary U sampled according to the distribution g, to the input W. Then we apply the channel Ψ and then finally apply the inverse of U to the resultant state. This process is called *g-twirling*, or twirling the channel Ψ using the ensemble generated by g. A unitary 2-design is generated by a distribution g if and only if $\mathrm{Tw}_{(g,\Psi)} = \mathrm{Tw}_{(h,\Psi)}$. Another interesting characterization of unitary 2-designs is presented in [GAE07]. A distribution g generates a unitary 2-design if and only if

$$\int_U \int_V \left|\mathrm{tr}(U^\dagger V)\right|^4 \, \mathrm{d}g(V) \, \mathrm{d}g(U) = 2. \tag{2.68}$$

Let $t\text{-SWAP} \in \mathcal{L}(\mathbb{C}^{d^2})$ such that $t\text{-SWAP}\,|x\rangle|y\rangle = |y\rangle|x\rangle$. This operator is the SWAP gate defined for qudit systems. Condition (2.68) can also be derived as follows: for any distribution g, $\Phi_g^{(2)}$ will have $\mathbb{1}_{\mathcal{C}^{d^2}}$ and t-SWAP as singular vectors, with singular value 1. Moreover, if g is a 2-design, then $\Phi_g^{(2)}$ is a rank 2 projection [HL09a; Guţ+20]. That means that the square of the Frobenius norm ($\|\cdot\|_2$) of $\Phi_g^{(2)}$, seen as a matrix defined on $\mathbb{C}^{d^4}$, is 2 if and only if g is a 2 design. This implies that

$$\mathrm{tr}\left(\int_U \int_V U^\dagger V \otimes U^\dagger V \otimes \overline{U}^{\mathrm{T}} \overline{V} \otimes \overline{U}^{\mathrm{T}} \overline{V}\right) \mathrm{d}g(U) = 2 \iff g = h. \tag{2.69}$$

Hence,

$$\int_U \int_V \left|\mathrm{tr}(U^\dagger V)\right|^4 \, \mathrm{d}g(V) \, \mathrm{d}g(U) = 2 \iff g = h. \tag{2.70}$$

Let $P = \{\mathbb{1}, X, Y, Z\}$ and let $\mathbb{P}_n$ contain all possible distinct 4^n n-fold tensor product of the elements in P. That is, $\mathbb{P}_n = \{P_1 \otimes \cdots \otimes P_n \mid \forall\ P_1, \ldots, P_n \in P\}$. $\mathbb{P}_n$ is an orthogonal (with respect to trace inner product) basis for $\mathcal{L}(\mathbb{C}^{2^n})$, called the *Pauli basis*, and hence any operator $A \in \mathcal{L}(\mathbb{C}^{2^n})$ can be written as

$$A = \sum_{P \in \mathbb{P}_n} \frac{\mathrm{tr}(AP)}{\sqrt{2^n}} \cdot \frac{P}{\sqrt{2^n}}. \tag{2.71}$$

For every quantum gate U, define Φ_U as its quantum channel representation, that is, a channel with the sole Kraus operator $\{U\}$. This is the unique way of representing a quantum gate, that is if for any two gates U, V their action on states will be equivalent if and only if $\Phi_U = \Phi_V$. This representation neatly avoids the impact that global phases can have when representing quantum gates as unitaries, and hence, is the more rigorous representation of a quantum gate.

Notice that the set $\{\Phi_P \mid \forall P \in \mathbb{P}_n\}$, along with the composition of channels, forms a group. Let $\mathcal{C}_n$ be the normalizer of this group in $\{\Phi_U \mid U \in \mathbb{U}_{2^n}\}$. $\mathcal{C}_n$ is called the *Clifford group* over n qubits. This set is a unitary 2-design of dimension 2^n [Dan+09]. In fact, they form a unitary 3-design [Web16; Zhu17], but fails to be a unitary 4-design [Zhu+16]. This means that the protocols such as [Ber20; BM21] which are efficient methods that can be used to sample uniformly from the Clifford group, can be used to sample efficiently from a unitary 3-design. Other constructions of exact and approximate unitary 2-designs include [Cle+15; Dan+09; BWV08; GAE07; Nak+17].

Chapter 3

Variational Quantum Algorithms

VQAs [Cer+21c; Ben+19] are quantum-classical hybrid algorithms whose goal is to use quantum devices to encode and solve optimization algorithms that are typically believed to be extremely difficult or intractable to do so classically. A VQA encodes the task under consideration using parameterized quantum circuits, also called *ansatzes*. Write $C(\boldsymbol{\theta})$ for the ansatz, where $\boldsymbol{\theta}$ is a real-valued vector of parameters. The VQA uses $C(\boldsymbol{\theta})$ to estimate a target function's outputs and gradients for different inputs, and then optimizes the parameters of the ansatz by feeding the circuit's output to a classical optimizer. We explain this in more detail in the coming sections.

3.1 Quantum Ansatzes

We explain the concept of quantum ansatzes using a simple example. Consider the rotation gates defined as

$$R_X(\theta) = e^{\frac{iX\theta}{2}} = \begin{bmatrix} \cos\frac{\theta}{2} & -i\sin\frac{\theta}{2} \\ -i\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{bmatrix}$$
$$R_Y(\theta) = e^{\frac{iY\theta}{2}} = \begin{bmatrix} \cos\frac{\theta}{2} & -\sin\frac{\theta}{2} \\ \sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{bmatrix}$$
$$R_Z(\theta) = e^{\frac{iZ\theta}{2}} = \begin{bmatrix} e^{-i\frac{\theta}{2}} & 0 \\ 0 & e^{i\frac{\theta}{2}} \end{bmatrix},$$

where $\theta \in \mathbb{R}$ is a tunable parameter. These gates combined with a CNOT gate make up a universal gate set. One can build parameterized circuits using such universal gate sets [Bar+95]. The universality of this gate set implies that any unitary can be approximated by such a parameterized circuit of sufficient depth. Hence, we can use these gates to design and implement optimization algorithms over unitaries, by optimizing their parameters.

A general ansatz defined on n qubits has the form

$$C(\boldsymbol{\theta}) = G_0 e^{\frac{iH_1\theta_1}{2}} G_1 e^{\frac{iH_2\theta_2}{2}} G_2 \dots e^{\frac{iH_p\theta_p}{2}} G_p, \tag{3.1}$$

where $H_1, \dots, H_p \in \mathbb{H}_{2^n}$, $\boldsymbol{\theta} \in \mathbb{R}^p$ is a vector of parameters, $G_0, G_1, \dots, G_p \in \mathbb{U}_{2^n}$ are gates independent of $\boldsymbol{\theta}$. Some examples of ansatzes are given in Figures 3.5, 3.6, and 3.7.

3.2 Overview of VQAs

Now, consider a parametrized circuit $C(\boldsymbol{\theta})$. If we apply this circuit to a system prepared in a quantum state σ (that can be prepared multiple times), and then estimate the expectation of an observable O through measurements, we obtain an estimate of the output of

the function

$$f_{\sigma,O,C}(\boldsymbol{\theta}) = \mathrm{tr}\left(OC(\boldsymbol{\theta})\sigma C(\boldsymbol{\theta})^\dagger\right). \tag{3.2}$$

One can also estimate its gradient at any point, with respect to individual parameters through measurements and minor changes to the parameters of the circuit using standard methods such as finite differencing [LeV07], or a quantum-specific one called the *parameter shift rule* [Mit+18] (cf. Lemma 7 in Appendix).

VQAs aim to optimize $f_{\sigma,O,C}$ over $\boldsymbol{\theta}$. Note that although most VQAs involve objective functions of this form, this is not true for all cases. But, in this thesis, we stick to VQAs that optimize objective functions of this form. The idea is to use quantum devices as black boxes that can estimate the function or its partial derivatives and update the parameters classically as per the update rules of any classical optimization algorithm such as gradient descent or ADAM. It turns out that many optimization problems in combinatorial optimization, quantum machine learning, and quantum chemistry can be framed as optimization of $f_{\sigma,O,C}$, for suitable choices of σ, O and ansatz C [Per+14; FGG14; ROA17]. For the remainder of this thesis, we shall omit C from this notation and use only $f_{\sigma,O}$, since the choice of ansatz C can be implicitly understood from the context.

We explain this with a simple example: computing the smallest eigenvalue $\lambda_{\min}$ of an observable $O \in \mathbb{H}_{2^n}$, and an associated eigenvector. Within quantum information, this problem is called finding the ground state or ground state energy of a Hamiltonian [SN20; SO82; Per+14]. Here, Hamiltonian is any observable with interesting physical properties such as the Ising Model Hamiltonian [GS18], Bose-Hubbard Hamiltonian [Sac11], Hydrogen Atom Hamiltonian [Sha11], etc, the ground state energy is simply $\lambda_{\min}$ while the ground state is an associated eigenvector. One can see that minimizing $f_{\sigma,O}$ for some initial state σ, can give us a good heuristic approximation of $\lambda_{\min}$. The choice of the ansatz C and the initial state σ generally depends on O. Consider a typical classical gradient descent optimization algorithm with its update rule given as $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \nabla f_{\sigma,O}(\boldsymbol{\theta}^{(t)})$. Hence, we can start with a randomly chosen $\boldsymbol{\theta}^{(0)}$, use the quantum device to estimate

$\nabla f_{\sigma,O}(\boldsymbol{\theta}^{(0)})$, compute $\boldsymbol{\theta}^{(1)}$ as per the update rule, and repeat this.

In practice, an n-qubit VQA is designed to carry out optimization of a function that would require resources exponential in n to evaluate or estimate classically. In the previous example, if we do not use a quantum circuit, we will have to carry out computation involving matrix multiplication with cost exponential in n, for a single update step of the optimization. Although there are more advanced classical simulation techniques such as MPS, stabilizer formalism [AG04], match gates[JM08], Lie algebraic simulation [Goh+23], Hamming bound circuits [Che+23], permutation equivariant circuits [Sch+24], etc, there exist many VQA objective functions that we strongly believe cannot be classically simulated efficiently using any such techniques [Hav+19; Cer+21c].

One thing to keep in mind here is that when optimizing the variational circuit, one might not be able to search through all of the search space, since the family of unitaries that the circuit can simulate might not contain all the elements of the search space. Therefore, these algorithms are generally suited for NISQ devices, where, as an example, with around 40-50 qubits, one can optimize functions defined over some subset of unitaries acting on spaces of dimension $\approx 10^9$, which classically might be extremely expensive or intractable. One also has to keep in mind that these circuits estimate the function values and gradients using measurements by estimating the sample proportion. This technique will require $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ executions of the circuit to estimate each of them up to a precision of ϵ.

3.3 Trainability Issues

Although by leveraging the universality of certain parameterized gate sets such as the example given in Section 3.1 one could argue that given enough gates or depth in the circuit, we must be able to estimate most of the unitaries, this approach brings severe trainability issues into the circuit such as the ones that we will discuss in this section.

3.3.1 Barren Plateaus

Recent works on general variational circuits have revealed a major obstacle regarding the trainability of quantum circuits called *barren plateaus* in the training landscape [Lar+24; Wan+21; OKW21; Cer+20]. This can, in some ways, be seen as an analog of the vanishing gradient phenomenon in classical neural networks [AG17; Hoc98]. It has been proved that in general, most of the objective function landscape could end up being barren plateaus, even with modest circuit depth linear in the number of qubits [McC+18]. The existence of these regions in the objective function landscape is usually characterized by the variance of all the partial derivatives when the parameters of the variational circuit are (uniformly) randomly initialized, being exponentially small.

We can formalize this and define barren plateaus as follows:

Definition 1. *Let σ be an n-qubit state and let $O \in \mathbb{H}_{2^n}$. For any ansatz $C(\boldsymbol{\theta}) = \prod_{p=1}^{t} U_p(\boldsymbol{\theta}_p)$, where $U_p(\boldsymbol{\theta}_p) = \prod_{q=1}^{m} e^{-i\theta_{pq}H_{pq}}$, $\boldsymbol{\theta}_p = [\theta_1 \dots \theta_m]^T$, $H_{pq} \in \mathbb{H}_{2^n}$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_1 \oplus \cdots \oplus \boldsymbol{\theta}_t$, and for any p, q, define*

$$U_p^{(L,q)}(\boldsymbol{\theta}_p) = \prod_{j=1}^{q-1} e^{-i\theta_{pj}H_{pj}}, \quad U_p^{(R,q)}(\boldsymbol{\theta}_p) = \prod_{j=q+1}^{m} e^{-i\theta_{pj}H_{pj}}. \tag{3.3}$$

Then, $f_{\sigma,O}(\boldsymbol{\theta})$ exhibits a barren plateau if $\forall\ p, q$ such that $1 \leqslant p \leqslant t, 1 \leqslant q \leqslant m$

$$Var_{\boldsymbol{\theta}}\left(\partial_{\theta_{pq}} f_{\sigma,O}(\boldsymbol{\theta})\right) \in \mathcal{O}\left(\frac{1}{b^n}\right), \tag{3.4}$$

for some constant $b > 1$, where $\partial_{\theta_{pq}} f_{\sigma,O}$ is its partial derivative with respect to θ_{pq} and $U_1, \dots U_{p-1}, U_{p+1}, \dots U_t$, 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.

Now, we shall explain the reason and consequence of this definition using Chebyschev's inequality. The expectation of any partial derivative of $f_{\sigma,O}$ under uniform initial param-

eterization is 0 [McC+18] (cf. Lemma 8 in Appendix). Hence, we have

$$\text{Prob}_{\boldsymbol{\theta}}(|\partial_k f_{\sigma,O}(\boldsymbol{\theta})| \leqslant \epsilon) \geqslant 1 - \text{Var}_{\boldsymbol{\theta}}(\partial_k f_{\sigma,O}(\boldsymbol{\theta}))/\epsilon^2, \tag{3.5}$$

for any k with $\partial_k f_{\sigma,O}(\boldsymbol{\theta})$ being the partial derivative of $f_{\sigma,O}(\boldsymbol{\theta})$ along its k^{th} direction. So, if $\forall\ k, \text{Var}_{\boldsymbol{\theta}}(\partial_k f_{\sigma,O}(\boldsymbol{\theta}))$ for uniformly randomly sampled $\boldsymbol{\theta}$ is scaling as $\mathcal{O}(1/b^n)$ for some $b > 1$, then the probability that $\partial_k f_{\sigma,O}(\boldsymbol{\theta})$ will be exponentially small is very high.

Many algorithms typically initialize the parameters uniformly at random, and can thus end up facing barren plateaus. If the partial derivatives are exponentially small, the number of samples required to get a meaningful estimate of them will be exponentially high since sample means estimation requires $\mathcal{O}(1/\epsilon^2)$ samples to estimate the expectation to precision ϵ (cf. Eq. (2.6)). Also, even if we somehow estimate them very well, the fact that they are exponentially small means that from the perspective of a gradient-based algorithm, the updates that one makes to the parameters will also be exponentially small, effectively rendering the training procedure extremely slow. Recent works have shown that barren plateaus can affect the training of non-gradient-based optimization algorithms as well [Arr+21].

Moreover, the choice of uniform distribution in this context intuitively means that most areas of the objective function landscape will have exponentially small partial derivatives, with the minima encountered in narrow gorges. Also, it turns out that there are many reasons other than circuit depth that can induce barren plateaus into the objective function such as simple hardware error models [Wan+21], choice of observables [Cer+20], level of entanglement [OKW21], etc.

But later works have identified the existence of certain circuit structures and objective functions that can have a polynomially vanishing lower bound for the variance of the gradient [Cer+20; LSW21a; Gra+19b; Pes+21; Mon+23; Sch+24; Dia+23; PKH24; Wes+24; Rud+23], which makes them trainable and provably avoid barren plateaus. These works have put forward directives that future circuit designs should follow, to be trainable. One example of a strategy that generally avoids barren plateaus is to use objective functions

that feature local ansatzes and local observables. These are circuits or components that act only on subsystems rather than the whole system of qubits. An example of such a circuit is shown in Figure 3.5. The locality of the circuits, or more importantly, the fact that all parameterized subcircuits are acting on a number of qubits independent of n, is crucial to avoiding barren plateaus. Using this, researchers have shown that the variance of the gradient can be shown to be exponentially small only in the number of qubits that the local circuits and observables act on [Cer+20; Wan+21], which means that the number of executions of the circuit that one would require to estimate the gradient is exponentially small only in the same number of qubits, rather than the total number of qubits in the system. Other strategies include smarter initialization techniques [Zha+22b], advanced quantum tomography techniques [Sac+22], etc.

3.3.2 Cost Concentration

Another related trainability issue that has received comparatively lesser attention is called *cost concentration* [Arr+22]. This happens when the objective function values themselves are exponentially small in most areas of the objective function landscape. Similar to the case of barren plateaus, this phenomenon is identified using the variance of the objective function itself. That is, we will have

$$\text{Prob}(|f_{\sigma,O}(\boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{\theta}}(f_{\sigma,O}(\boldsymbol{\theta}))| \geqslant \epsilon) \leqslant \frac{\text{Var}_{\boldsymbol{\theta}}(f_{\sigma,O}(\boldsymbol{\theta}))}{\epsilon^2}, \tag{3.6}$$

implying that if $\text{Var}_{\boldsymbol{\theta}}\left(f_{\sigma,O}(\boldsymbol{\theta})\right) \in \mathcal{O}(1/b^n)$ for some $b > 1$, the objective function will be exponentially close to its mean with very high probability for a uniformly sampled input parameter. This further implies that an exponentially large number of executions will be required to estimate the objective function to meaningful precision. Formally, we shall define cost concentration as follows:

Definition 2. *Let σ be an n-qubit state and $O \in \mathbb{H}_{2^n}$. For any ansatz $C(\boldsymbol{\theta})$, the function*

$f_{\sigma,O}$ exhibits cost concentration if

$$Var_{\boldsymbol{\theta}}\left(f_{\sigma,O}(\boldsymbol{\theta})\right) \in \mathcal{O}\left(\frac{1}{b^n}\right) \tag{3.7}$$

for some $b > 1$.

We refer the reader to [Arr+22] for a comprehensive work relating barren plateaus and cost concentration.

3.4 VQA Applications Discussed in This Thesis

VQAs have been proposed for various kinds of applications including combinatorial optimization [FGG14; FH19; Zho+20b], variational quantum eigensolver [Per+14; Til+22a; Cer+22], quantum autoencoder [ROA17; Wan+17; VPB18], quantum classifiers [Hav+19; Sch21; CCL19], etc. In this section, we delve a little bit deeper into those specific applications that we have used in this thesis mostly to demonstrate the capabilities of the protocols developed.

3.4.1 State Preparation

State preparation is a problem that is extensively studied and used in quantum information [Cer+20; MJP21; Gar+20]. The problem that we consider here is as follows: given many copies of an n-qubit pure state σ (or access to a circuit that can prepare σ), output the parameters of an ansatz $C(\boldsymbol{\theta})$ such that we can approximately prepare σ using these parameters. Broadly speaking, there are two different approaches to solving this problem, differing only in their measurement strategies.

- *Global Observables*: In this method, the idea is to maximize the fidelity between $|\mathbf{0}\rangle\langle\mathbf{0}|_{C(\boldsymbol{\theta})^\dagger}$ and σ, over $\boldsymbol{\theta}$. That is, find

$$\arg\max_{\boldsymbol{\theta}} f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta}). \tag{3.8}$$

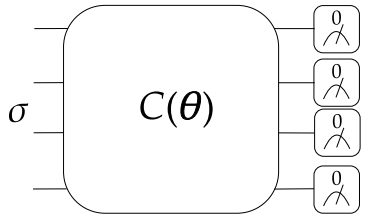


Figure 3.1: An example of a state preparation circuit using global observables. Here, we apply $C(\boldsymbol{\theta})$ to the input state σ, estimate the probability of all qubits being in 0, which is the fidelity between σ and $|\mathbf{0}\rangle\langle\mathbf{0}|_{C(\boldsymbol{\theta})^\dagger}$ and maximize it.

This fidelity can be estimated by applying $C(\boldsymbol{\theta})$ 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 as the objective function is the *global observable* $|\mathbf{0}\rangle\langle\mathbf{0}|$. The circuit is given in Figure 3.1.

In many instances, this approach has been shown to induce barren plateaus in the training landscape [Cer+20; Liu+22]. However, methods such as [Pat+21; Mel+22; RSL22; Sko+21; Gri+23a; Gri+23b; FM22; Ver+19; Gra+19a; KS22; Zha+22a] could be used to heuristically mitigate this issue.

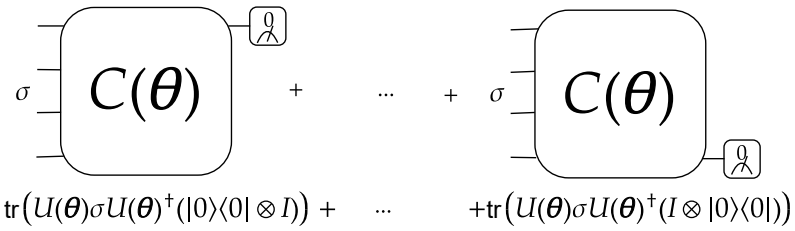


Figure 3.2: An example of quantum state preparation circuit using local observables. Here, we apply $C(\boldsymbol{\theta})$ to the input state σ, estimate the sum of probabilities of each qubit being in 0, and maximize it.

- *Local Observables*: Alternatively, one can employ the expectation of $F := 1/n \sum_{i=1}^{n} |0\rangle\langle 0|_i$, which is a sum of 1-local observables, as the objective function,

since as per [Cer+21b], we have

$$\arg\max_{U\in\mathbb{U}_{2^n}} \langle \mathbf{0}| \sigma_U |\mathbf{0}\rangle = \arg\max_{U\in\mathbb{U}_{2^n}} \operatorname{tr}(F\sigma_U) = \arg\max_{U\in\mathbb{U}_{2^n}} \frac{1}{n}\sum_{i=1}^{n} \operatorname{tr}(|0\rangle\langle 0|_i\, \sigma_U). \tag{3.9}$$

The intuition here is that if $\boldsymbol{\theta}^*$ maximizes $f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}$, then $\sigma_{C(\boldsymbol{\theta}^*)} = |\mathbf{0}\rangle\langle\mathbf{0}|$ and hence $f_{\sigma,F}$ also attains its maximum on $\boldsymbol{\theta}^*$.

Hence, from Eqs. (3.8) and (3.9), we can see that maximizing the sum of probabilities of each qubit of $\sigma_{C(\boldsymbol{\theta})}$ yielding $+1$ when measured using Z is (heuristically) approximately equivalent to maximizing the probability of all qubits simultaneously yielding $+1$. Hence, in this regime, we try to find $\arg\max_{\boldsymbol{\theta}} f_{\sigma,F}(\boldsymbol{\theta})$. The circuit can be seen in Figure 3.2.

The advantage of this approach is that in works such as [Cer+20; Wan+21; Liu+22], it has been shown that this measurement strategy coupled with certain shallow ansatzes (ansatzes with depth logarithmic in the number of qubits) can provably avoid barren plateaus.

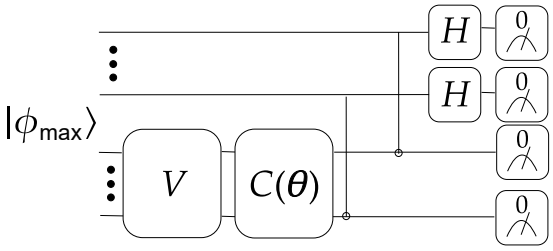


Figure 3.3: An example of VQCS. Initially, we prepare the maximally entangled state $|\phi_{\max}\rangle = \frac{1}{\sqrt{2^n}}\sum_{i=0}^{2^n-1} |i\rangle|i\rangle$ on two n qubit registers by applying H gates on all n qubits in the first register and following it up with a series of n CNOT gates on corresponding qubits in both registers. Then we apply the unknown gate V on it to prepare its vectorized version $|V\rangle$. Then we implement a state preparation circuit on it (global measurement method). That is, we would like to maximize the fidelity between $|V\rangle$ and $|C(\boldsymbol{\theta})\rangle$. To do this, we should apply the inverse of the circuit that prepares $|C(\boldsymbol{\theta})\rangle$ on $|V\rangle$ ($C(\boldsymbol{\theta})$ on the second register followed by application of the inverse of the gates that prepared $|\phi_{\max}\rangle$ which are CNOTs followed by the H gates), estimate the probability of all qubits simultaneously being in 0 and maximize it.

3.4.2 Variational Quantum Circuit Synthesis (VQCS)

VQCS is a natural extension of state preparation to quantum circuits. Here, our goal is to learn the parameters of an n-qubit ansatz $C(\boldsymbol{\theta})$ that best approximates a given unknown quantum gate V. Similar to how we use fidelity or infidelity for quantum states, we can use the Hilbert-Schmidt cost function defined for unitaries in [Kha+19]. For any $\boldsymbol{\theta}$, this is computed as $H(\boldsymbol{\theta}) = 1-|\mathrm{tr}(C(\boldsymbol{\theta})^\dagger V)|^2/4^n$ 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 $\left\| C(\boldsymbol{\theta}) \otimes \overline{C(\boldsymbol{\theta})} - V \otimes \overline{V} \right\|_2^2$.

To evaluate $H(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$, we start with the maximally entangled state on two n-qubit systems, defined as $|\phi_{\max}\rangle = \frac{1}{\sqrt{2^n}} \sum_{i=0}^{2^n-1} |i\rangle |i\rangle$. To see how we can prepare this state, define the two n qubit systems as $(1, 2, \ldots, 2n-1, 2n)$. $|\phi_{\max}\rangle$ can be prepared by initializing the whole system in $|\mathbf{0}\rangle$, applying H on the first n qubits, and applying $\mathrm{CNOT}_{(q,q+n)}$ for all $q = 1, 2, \ldots, n$.

Once $|\phi_{\max}\rangle$ is prepared, apply V on qubits $(n+1, \ldots, 2n)$ to obtain $\frac{1}{\sqrt{2^n}} |V\rangle$, where

$$|V\rangle = \sum_{i=0}^{2^n-1} |i\rangle |v_{\bullet i}\rangle, \tag{3.10}$$

where $|v_{\bullet i}\rangle$ is the i^{th} column of V. Also, applying $C(\boldsymbol{\theta})$ on $|\phi_{\max}\rangle$ in a similar manner will produce $\frac{1}{\sqrt{2^n}} |C(\boldsymbol{\theta})\rangle$. Then we have $H(\boldsymbol{\theta}) = 1 - \frac{1}{4^n}\mathrm{tr}(|C(\boldsymbol{\theta})\rangle\langle C(\boldsymbol{\theta})| \, |V\rangle\langle V|)$, which is the infidelity between the states $\frac{1}{\sqrt{2^n}} |C(\boldsymbol{\theta})\rangle$ and $\frac{1}{\sqrt{2^n}} |V\rangle$. To estimate the infidelity between any two states $|\psi\rangle$ and $W|0\rangle$ for some unitary W, we simply apply $W^\dagger$ on $|\psi\rangle$ and estimate the probability of all qubits being in 0. Hence, to estimate the Hilbert Schmidt cost, we simply apply the inverse of the circuit that prepares $\frac{1}{\sqrt{2^n}} |C(\boldsymbol{\theta})\rangle$ on $\frac{1}{\sqrt{2^n}} |V\rangle$, that is, $C(\boldsymbol{\theta})^\dagger$ on qubits $(n+1, \ldots, 2n)$ followed by $\mathrm{CNOT}_{(q,q+n)}$ for all $q = 1, 2, \ldots n$ and then H on qubits $(1, 2, \ldots, n)$, and then estimate the probability of all $2n$ qubits returning 0. The

circuit is given in Figure 3.3.

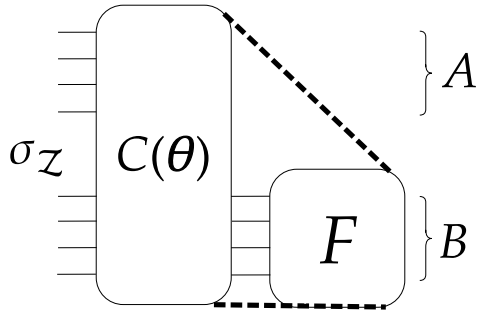
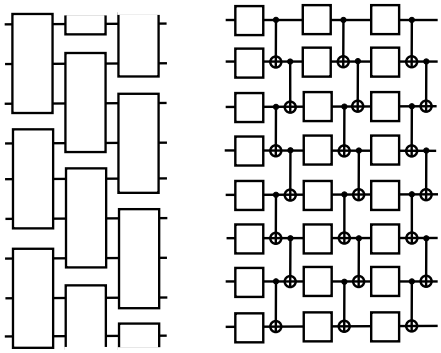


Figure 3.4: An example of a quantum autoencoder circuit. We first apply $C(\boldsymbol{\theta})$ to $\sigma_{\mathcal{Z}} = \sum_i p_i \left|\psi_i\right\rangle\!\left\langle\psi_i\right|$. Then, we measure and estimate the sum of the probability of each qubit in B being in 0. This probability is then used to maximize the objective function $f_{\sigma_{\mathcal{Z}},F}(\boldsymbol{\theta})$, where $F = \frac{1}{n_B}\mathbb{1}_{C^{2^{n_A}}} \otimes \sum_{i=1}^{n} \left|0\right\rangle\!\left\langle 0\right|_i$. This forces all the population (probabilities) of the σ (equivalently, of all input states $\left|\psi_i\right\rangle$) to be in register B.

3.4.3 Quantum Autoencoder

Autoencoder is a popular dimensionality reduction technique in classical machine learning [HZ93]. Using deep neural networks, autoencoders learn low-dimensional representations of high-dimensional input data, which should ideally keep hold of the original characteristics of the data. This can also be seen as a form of data compression. Recently, there have been numerous works on extending this concept to quantum data [ROA17; Wan+17; VPB18; PTP19; Lam+18]. Although there are many variations of quantum autoencoders, we shall focus on the one presented in [ROA17], as that is the version that we use in Chapter 4.

The idea behind this version of the quantum autoencoder is to compress n-qubit quantum states into $n_B < n$ qubit states. Consider an ensemble of n-qubit states $\mathcal{Z} = \{(p_i, \left|\psi_i\right\rangle)\}$ with each state being prepared in registers A and B having n_A and n_B qubits respectively, where $n = n_A + n_B$. Let $\sigma_{\mathcal{Z}} = \sum_i p_i \left|\psi_i\right\rangle\!\left\langle\psi_i\right|$. As a measure of the compression effect, we consider $f_{\sigma_{\mathcal{Z}},\mathbb{1}_{\mathbb{C}^{2^{n_A}}}\otimes F}(\boldsymbol{\theta})$, where $F = \frac{1}{n_B}\sum_{i=1}^{n_B} \left|0\right\rangle\!\left\langle 0\right|_i$ is defined, as in the state preparation problem, to be a sensible objective function that not only forces



(a) ALA (b) HEA

Figure 3.5: (a) ALA: Here each block represents a parameterized subcircuit acting on a subset of qubits, in a brick-like manner. The number of vertical columns of these subcircuits is called the depth of the circuit. (b) HEA: The circuit structure is very similar to the ALA. In this architecture, we apply layers of single qubit parameterized gates (represented by boxes), followed by CNOT gates being applied for two layers in a brick-like manner to introduce entanglement into the circuit.

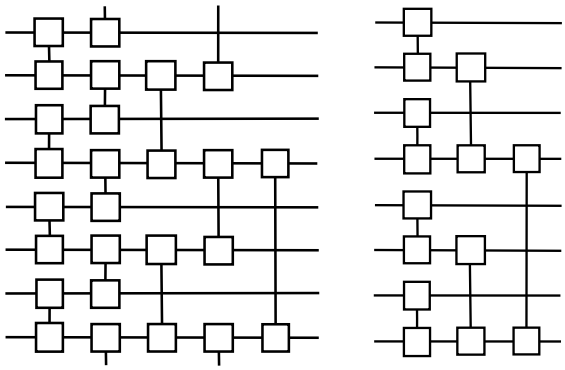
the population (probabilities) of all the states to be in the qubits in the register B, but also involves the same 1-local observables that we have used for the more trainable version of state preparation given in Section 3.4.1. Again, this objective function has been used in [Cer+20]. The circuit can be seen in Figure 3.4.

3.5 Ansatzes Used in This Thesis

In this section, we introduce the ansatzes we use in this thesis.

3.5.1 Alternating Layered Ansatz

The Alternating Layered Ansatz (ALA) is the brick-like circuit structure presented in Figure 3.5 (a), where each subcircuit is a parameterized circuit acting on a small number of qubits. A simple example of a subcircuit, the one we have used for the experiments in Chapter 4 regarding ALA, is given in Figure 4.1 (b). ALA is well studied in the literature for its expressivity [NY21] (the volume of unitaries that it can represent) and trainability [Cer+20] when used in combination with local observables. The Hardware Efficient Ansatz (HEA) given in Figure 3.5 (b) is a slightly modified variant of ALA,



(a) QCNN (MERA) (b) TTN

Figure 3.6: (a) The QCNN ansatz. Each pair of connected boxes represents a parameterized two-qubit subcircuit. Multiple layers of ALA are applied, with half the qubits omitted as we progress, thus resulting in an ansatz of depth $\mathcal{O}(\log n)$. This ansatz can also be seen as a special case of the MERA. (b) Tree Tensor Network: a special case of QCNN with the ALA being used having depth 1.

where we apply single qubit gates on all qubits, followed by 2 layers of CNOT gates in a brick-like manner to introduce entanglement into the ansatz. Other examples of works that have used or studied these ansatzes include [Leo+24; Tan21; PKH24; Kar+21; Kan+17]. ALA will be discussed in more detail in Chapter 4.

3.5.2 Quantum Convolutional Neural Network

The Quantum Convolutional Neural Network Ansatz (QCNN), introduced in [CCL19] is a variant of the popular Multiscale Entanglement Renormalization Ansatz (MERA) [MV18; EV13; FV12] and is given in Figure 3.6 (a). It is the quantum analog of the classical Convolutional Neural Network (CNN) [Den+88; Li+22]. Each layer is an ALA, with subcircuit width 2. After the application of ALA on each layer, in the subsequent layer, we only apply the next ALA on half the number of qubits on which the previous ALA was applied. Thus, the total circuit consists of $\mathcal{O}(\log n)$ layers. In the first layer, the ALA is applied to all qubits $[1, 2, \ldots, n]$. Then in the l^{th} layer, a similar ALA is applied to qubits $[2^{(l-1)}, 2 \cdot 2^{(l-1)}, 3 \cdot 2^{(l-1)}, \ldots, n]$.

To see the relationship with classical CNN, we look at the original version of QCNN presented in [CCL19]. In that version, each subcircuit within the ALAs is the quantum

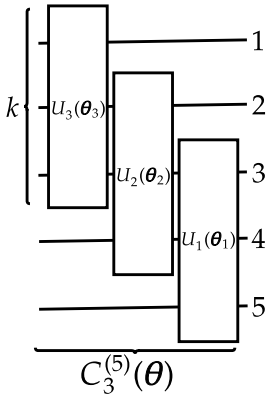


Figure 3.7: Example of an MPS ansatz with $n = 5$ and $t = 3$. The numbers on the end of each qubit wire are the indices of the qubits. Each box is a parameterized subcircuit, which is applied in a staircase manner.

version of a convolutional layer. Similar to the classical case, where the convolutional layer consists of filters applied on different parts of the input, here the subcircuits act on different parts of the input state. Each ALA layer is then followed by a pooling layer. In the l^{th} pooling layer, between each pair of adjacent qubits on which the ALA is being applied, say $k \cdot 2^{l-1}$ and $(k+1) \cdot 2^{l-1}$, we apply a classically controlled (an operation where the control qubit is measured and a controlled gate based on the classical output is applied on the target qubit) single qubit parameterized gate. Since half the qubits featured in that layer get measured, the number of qubits gets halved after each layer, and the next ALA is applied to all qubits that featured as target qubits in the previous pooling layer. Similar to how pooling introduces a non-linear operation in the classical CNN, here the pooling layer introduces a non-unitary operation. The version of QCNN that we described in the previous paragraph is simply an extension of this. The Tree Tensor Network (TTN) ansatz [SDV06; TEV09; Mur+10; NC13] given in Figure 3.6 (b) is a simple case of MERA with the ALAs having depth 1.

3.5.3 MPS Ansatz

The MPS ansatz is given in Figure 3.7. It is built using cascading layers of smaller k-qubit parameterized subcircuits, in a staircase manner. From the MPS tomography procedure

explained in Section 2.4.4, we can see that every state that can be represented using an MPS representation with bond dimensions at most 2^{k-1} can be implemented using this ansatz (assuming that each of the subcircuits can implement any k-qubit unitary). This is what led many works to use the MPS ansatz to solve state preparation problems variationally [Lin+21; Rud+22; Dov+22; Ran20]. More details on this ansatz can be found in Chapter 6.

Chapter 4

Alternating Layered Shadow Optimization

4.1 Overview

In this chapter, we explain the first major contribution of this thesis, an efficient training algorithm for VQAs that involves ALAs and local observables which is exponentially better than the standard method. We start with a small review of the properties of this ansatz-observable combination.

In [Cer+21b], it is proved that barren plateaus can be avoided for VQAs involving ALAs provided that the depth of the ansatz is $\mathcal{O}(\log n)$, where n is the number of qubits and the objective function is defined with local observables. Surprisingly, in [NY21], it was recently proved that shallow ALAs are almost as expressive as the more widely used HEA. Thus, the ALA is both expressive and trainable. In addition, this ansatz has been investigated or implemented in works such as [Hin+21; Wu+21; Arr+21; SVC22].

Our work introduces a training algorithm with an exponential improvement in the number of copies of input states consumed during training an alternating layered VQA with shallow depth and local observables. Moreover, the training can be done entirely on a classical computer efficiently (with computational cost depending only polynomially on

n) without the need to implement the ansatz on a quantum device. This result is achieved by using the classical shadow technique and working in the Heisenberg picture rather than the Schröndinger model.

Specifically, for an ALA $C(\boldsymbol{\theta})$, an input state σ and an observable O, the VQAs of our interest estimate each evaluation of $f_{\sigma,O}$ using quantum computers. In contrast, our method can efficiently compute this classically using classical shadows of σ. But note that all VQAs that use ALAs need not have this specific form.

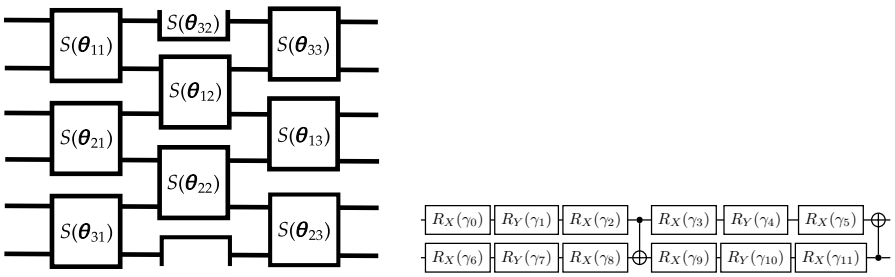
Our method, called *Alternating Layered Shadow Optimization*, or simply ALSO, outperforms standard alternating layered VQA in two aspects:

1. *Exponential savings on input state copies.* Note that the number of copies of the input state needed in the standard VQA scales linearly in the total number of function evaluations required. In contrast, to achieve a similar precision, ALSO only uses logarithmically many copies. This allows for more iterations and better approximations in the classical optimization algorithm for a given ansatz. In addition, it allows for more hyperparameter tuning with very few copies of the input state, and the same set of shadows can be used for multiple similar optimization problems and ALAs.

2. *Easy implementation on quantum hardware.* ALSO only requires the quantum device to be able to carry out single-qubit Pauli basis measurements on the input states. But standard VQA requires the ability to apply CNOT gates and rotation gates on them, and measurement also.

We demonstrate the practical efficacy of our result with two important examples: finding state preparation circuits and quantum data compression using a quantum autoencoder. In both cases, we demonstrate that ALSO can match the results of the impossible ideal VQA that uses infinite copies, using a comparatively small number of copies of the input quantum state. We also show that, with the same number of copies of the input state, ALSO outperforms the standard VQA significantly.

The ALA is the brick-like circuit structure presented in Figure 4.1 (a), where each $S(\boldsymbol{\theta}_{ij})$ is a parameterized circuit acting on a small number of qubits. A brief introduction



(a) (b)

Figure 4.1: (a) A detialed illustration of ALAs where the parameterized sub-circuit $S(\boldsymbol{\theta}_{32})$ is applied on the first and the last qubits. Here, $\boldsymbol{\theta}$ is an order 3 tensor with each $\boldsymbol{\theta}_{ij}$ being vectors of real parameters. (b) The structure of the parameterized subcircuit $S(\boldsymbol{\gamma})$ used in the simulation.

is given in Section 3.5.1. A simple example of S is given in Figure 4.1 (b). This work assumes that each S acts on two qubits and has p real parameters, but our idea can be easily extended to the general case. In Figure 4.1 (a), the total number of vertical blocks of S gates, written d, is the depth of the ansatz. The circuit depicted in the figure has $d = 3$. In a specific vertical block j, each circuit $S(\boldsymbol{\theta}_{ij})$ acts on qubits $2(i-1) \oplus j$ and its neighbor $2(i-1) \oplus j \oplus 1$. So, the final form of the circuit is given as

$$C(\boldsymbol{\theta}) = \prod_{j=1}^{d} \prod_{i=1}^{n/2} S(\boldsymbol{\theta}_{ij})_{(2(i-1)\oplus j, 2(i-1)\oplus j\oplus 1)}, \tag{4.1}$$

where $\oplus$ denotes addition modulo n, $\boldsymbol{\theta} \in \mathbb{R}^{\frac{n}{2} \times d \times p}$ is a tensor of real parameters where $\boldsymbol{\theta}_{ij}$ is a p-dimensional real vector of parameters.

ALA is one of the most researched ansatzes in the literature [Leo+24; Tan21; PKH24; Kar+21; Kan+17; Cer+20; NY21]. The fact that the subcircuits act on a small number of neighbouring qubits means that the physical qubits of the hardware need less connectivity to implement this. But, the drawback of the nearest neighbour structure is the lack of ability to represent quantum circuits that induce a lot of entanglement. Also, it has already been shown that this ansatz can induce barren plateaus if used with depth $\Omega(\text{poly}(n))$, while $\mathcal{O}(\log n)$ depth avoids it. But, at the latter small depth, if the initial state is also one that can be prepared by a (known) circuit with similar depth, then the whole VQA

circuit can be simulated classically efficiently using MPS (cf. Section 2.4.2). Hence, it is necessary to have an input state that requires a $\mathcal{O}(\text{poly}(n))$ depth circuit to implement to take the whole VQA protocol beyond the realm of (full) classical simulation.

4.2 Classical Shadow Tomography

We start by introducing classical shadows. Let $O^{(1)}, O^{(2)}, \ldots, O^{(M)}$ be arbitrary n-qubit observables the classical descriptions of which are given. As mentioned in Section 2.4.4, using conventional quantum tomography techniques, $\mathcal{O}(2^n)$ copies of σ are required to estimate $\text{tr}\left(O^{(i)}\sigma\right)$ for each $O^{(i)}$.

Many techniques have already been developed in the literature that try to circumvent this issue, that is, estimate these expectations without carrying out full quantum state tomography [Pai+21; Yu20; HKP20; Aar18; Aar07; FL11; GA22]. The classical shadow technique [HKP20], developed from shadow tomography [Aar18], provides succinct classical descriptions of quantum states. Using this technique, $\text{tr}(O^{(i)}\sigma)$ can be collectively predicted by consuming only $\mathcal{O}(\log M)$ copies of σ. Moreover, when these observables belong to certain classes, the dependency on n is polynomial or constant.

Now let us explain the procedure in more detail. Let Φ be the measurement (in the computational basis) channel and let g be a distribution on unitaries such that $\text{Tw}_{(g,\Phi)}$ (channel twirling cf. Section 2.5) is invertible. Consider a randomized protocol implementing $\text{Tw}_{(g,\Phi)}$ defined as follows: (i) sample an n-qubit unitary U according to g, (ii) apply U on σ, (iii) measure all qubits of the resulting state, giving the classical n-bit string i as output, (iv) compute a *classical shadow*, defined as $\hat{\sigma} := \text{Tw}^{-1}_{(g,\Phi)}\left(U^\dagger |i\rangle\langle i| U\right)$, classically. The whole protocol is illustrated in Figure 4.2.

The classical shadow is a random variable that reproduces the original state in expectation. To see this, we first have to prove that $\text{Tw}_{g,\Phi}$ is a Hermitian map. To see that, we further require proving that Φ is a Hermitian map. This can be seen as follows; for any

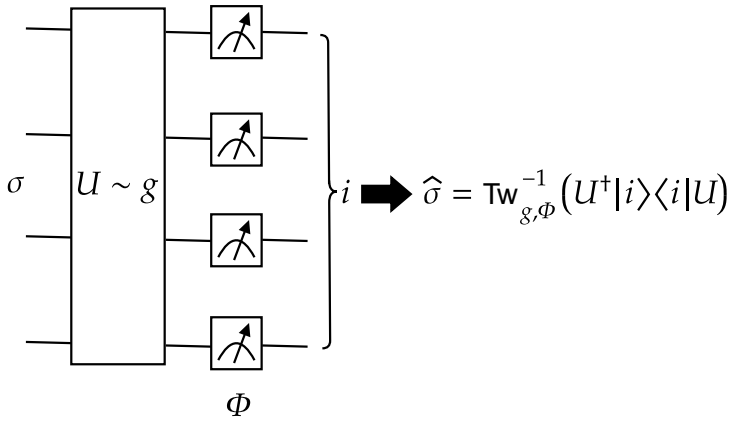


Figure 4.2: Protocol to generate a classical shadow of a state σ, given a distribution g defined over all unitaries. The only prerequisite is that $\mathrm{Tw}_{g,\Phi}$ should be an invertible map, where Φ is the computational basis measurement channel. First, we apply a unitary U sampled according to g on σ. Then, we measure the resultant state in the computational basis, resulting in a classical bit string i. Finally, we compute the shadow $\hat{\sigma} = \mathrm{Tw}^{-1}_{g,\Phi}(U^\dagger |i\rangle\langle i| U)$ classically.

two operators A, B, we have

$$\mathrm{tr}\left(\Phi(A)^\dagger B\right) = \sum_i \overline{A_{ii}} B_{ii} = \mathrm{tr}\left(A^\dagger, \Phi(B)\right). \tag{4.2}$$

Now, we move on to $\mathrm{Tw}_{g,\Phi}$. The reason why $\mathrm{Tw}_{g,\Phi}$ is Hermitian is

$$\mathrm{tr}(\mathrm{Tw}_{g,\Phi}(X)^\dagger Y) = \mathrm{tr}\left(\int_U U^\dagger \Phi(UXU^\dagger)^\dagger UY \mathrm{d}g(U)\right) \tag{4.3}$$

$$= \mathrm{tr}\left(\int_U \Phi(UXU^\dagger)^\dagger UYU^\dagger \mathrm{d}g(U)\right) \tag{4.4}$$

$$= \mathrm{tr}\left(\int_U UX^\dagger U^\dagger \Phi(UYU^\dagger) \mathrm{d}g(U)\right) \tag{4.5}$$

$$= \mathrm{tr}\left(\int_U X^\dagger U^\dagger \Phi(UYU^\dagger) U \mathrm{d}g(U)\right) \tag{4.6}$$

$$= \mathrm{tr}\left(\int_U X^\dagger \mathrm{Tw}_{g,\Phi}(Y)\right). \tag{4.7}$$

This also implies that $\mathrm{Tw}_{g,\Phi}^{-1}$ is Hermitian. Hence, for any observable O, we have

$$\mathbb{E}_{g,\Phi}\left(\mathrm{tr}(O\hat{\sigma})\right) = \mathbb{E}_{U,i}\left(\mathrm{tr}\left(O\mathrm{Tw}_{g,\Phi}^{-1}\left(U^\dagger|i\rangle\langle i|U\right)\right)\right) \tag{4.8}$$

$$= \int_U \sum_i \langle i|U\sigma U^\dagger|i\rangle \,\mathrm{tr}\left(O\mathrm{Tw}_{g,\Phi}^{-1}\left(U^\dagger|i\rangle\langle i|U\right)\right)\, \mathrm{d}g(U) \tag{4.9}$$

$$= \int_U \mathrm{tr}\left(\mathrm{Tw}_{g,\Phi}^{-1}(O)U^\dagger\left[\sum_i \left(\langle i|U\sigma U^\dagger|i\rangle\right)|i\rangle\langle i|\right]U\right)\, \mathrm{d}g(U) \tag{4.10}$$

$$= \mathrm{tr}\left(\mathrm{Tw}_{g,\Phi}^{-1}(O)\mathrm{Tw}_{g,\Phi}(\sigma)\right) \tag{4.11}$$

$$= \mathrm{tr}(O\sigma). \tag{4.12}$$

A classical shadow is a unit trace matrix but not necessarily positive semidefinite. To see why it has unit trace, notice that $\mathrm{Tw}_{g,\Phi}(\mathbb{1}_{2^n}) = \mathbb{1}_{2^n}$. Since $\mathrm{Tw}_{g,\Phi}$ is linear and invertible, it should be an injective map, meaning that $\mathrm{Tw}_{g,\Phi}^{-1}(\mathbb{1}_{2^n}) = \mathbb{1}_{2^n}$. Hence, we have

$$\mathrm{tr}\left(\mathrm{Tw}_{g,\Phi}^{-1}\left(U^\dagger|i\rangle\langle i|U\right)\right) = \mathrm{tr}\left(\mathrm{Tw}_{g,\Phi}^{-1}\left(U^\dagger|i\rangle\langle i|U\right)\mathbb{1}_{2^n}\right) \tag{4.13}$$

$$= \mathrm{tr}\left(U^\dagger|i\rangle\langle i|U\mathrm{Tw}_{g,\Phi}^{-1}\left(\mathbb{1}_{2^n}\right)\right) \tag{4.14}$$

$$= \mathrm{tr}\left(U^\dagger|i\rangle\langle i|U\right) \tag{4.15}$$

$$= 1. \tag{4.16}$$

Eq 4.8 means that classical shadows can be used to estimate the expectation of any observable unbiasedly. To see how good the estimate is, we shall try to derive a bound on the variance, or the second moment.

$$\mathrm{Var}_{g,\Phi}\left(\mathrm{tr}(O\hat{\sigma})\right) = \mathbb{E}_{g,\Phi}(\mathrm{tr}(O\hat{\sigma}) - \mathrm{tr}(O\sigma))^2 \tag{4.17}$$

$$= \mathbb{E}_{g,\Phi}(\mathrm{tr}(O_{\mathrm{TL}}\hat{\sigma}) - \mathrm{tr}(O_{\mathrm{TL}}\sigma))^2 \tag{4.18}$$

$$\leqslant \mathbb{E}_{g,\Phi}(\mathrm{tr}\left(O_{\mathrm{TL}}\hat{\sigma}\right)^2), \tag{4.19}$$

where $O_{\mathrm{TL}} = O - \frac{\mathrm{tr}(O)}{2^n}\mathbb{1}_{\mathbb{C}^{2^n}}$, the traceless part of O. This can be seen when we consider the trace inner product as a standard inner product in the orthonormal Pauli basis, that

is,

$$\operatorname{tr}(A^\dagger B) = \sum_{P\in\mathbb{P}_n} \frac{\operatorname{tr}\left(A^\dagger P\right)\operatorname{tr}\left(B^\dagger P\right)}{2^n} = \frac{\operatorname{tr}\left(A^\dagger\right)\operatorname{tr}\left(B^\dagger\right)}{2^n} + \sum_{P\in\mathbb{P}_n/\{\mathbb{1}/\sqrt{2^n}\}} \frac{\operatorname{tr}\left(A^\dagger P\right)\operatorname{tr}\left(B^\dagger P\right)}{2^n} \tag{4.20}$$

$\forall\ A, B \in \mathcal{L}\left(\mathbb{C}^{2^n}\right)$, and the fact that shadows have unit trace. Then we have,

$$\begin{aligned}
\operatorname{Var}_{g,\Phi}\left(\operatorname{tr}\left(O\hat{\sigma}\right)\right) &\leqslant \mathbb{E}_{g,\Phi}\left(\operatorname{tr}\left(O\hat{\sigma}\right)^2\right) && (4.21)\\
&= \int_U \sum_i \langle i|\, U\sigma U^\dagger\, |i\rangle \operatorname{tr}\left(\operatorname{Tw}^{-1}_{g,\Phi}(W_{\mathrm{TL}})U^\dagger\, |i\rangle\langle i|\, U\right)^2 \mathrm{d}g(U) && (4.22)\\
&= \|O_{\mathrm{TL}}\|^2_{\sigma,\mathrm{sh}}. && (4.23)
\end{aligned}$$

This state-dependent function $\|\cdot\|_{\sigma,\mathrm{sh}}$, which is also a norm, is called the *state-dependent shadow norm*. Hence, the norm is defined as $\|\cdot\|_{\mathrm{sh}} := \max_\sigma \|\cdot\|_{\sigma,\mathrm{sh}}$, called the *shadow norm*, gives us an upper bound on the variance of the estimator.

Since we have an upper bound on the variance, we can use Chebyshev inequality [Knu97], which states that for any (integrable) random variable η,

$$\operatorname{Prob}(|\eta - \mathbb{E}(\eta)| \geqslant \epsilon) \leqslant \frac{\operatorname{Var}(\eta)}{\epsilon^2}. \tag{4.24}$$

for any $\epsilon > 0$, to assess the performance of a sample means estimator. Let $\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_T$ be independently generated classical shadows. Then, using Chebychev inequality, we have

$$\operatorname{Prob}\left(\left|\frac{1}{T}\sum_{i=1}^{T}\operatorname{tr}(O\hat{\sigma}_i) - \operatorname{tr}(O\sigma)\right| \geqslant \epsilon\right) \leqslant \frac{\operatorname{Var}_{g,\Phi}(\operatorname{tr}\left(O\hat{\sigma}\right))}{T\epsilon^2} \leqslant \frac{\|O_{\mathrm{TL}}\|^2_{\mathrm{sh}}}{T\epsilon^2}.$$

Set $\delta = \frac{\|O\|^2_{\mathrm{sh}}}{T^2\epsilon^2}$. Then we can say that for any precision and confidence parameters $\epsilon, \delta \in (0,1)$, if we use $T = \frac{\|O_{\mathrm{TL}}\|^2_{\mathrm{sh}}}{\delta\epsilon^2}$ shadows, we will have

$$\operatorname{Prob}\left(\left|\frac{1}{T}\sum_{i=1}^{T}\operatorname{tr}(O\hat{\sigma}_i) - \operatorname{tr}(O\sigma)\right| \leqslant \epsilon\right) \geqslant 1 - \frac{\operatorname{Var}_{g,\Phi}(\operatorname{tr}(O\hat{\sigma}))}{T\epsilon^2} \geqslant 1 - \delta.$$

Now, if we had multiple observables $O^{(1)}, O^{(2)}, \ldots, O^{(M)}$, then, for any $\delta', \epsilon \in (0,1)$, we get

$$\text{Prob}\left(\bigcap_{j=1}^{M}\left|\frac{1}{T}\sum_{i=1}^{T}\text{tr}\left(O^{(j)}\hat{\sigma}_i\right)-\text{tr}\left(O^{(j)}\sigma\right)\right|\geqslant\epsilon\right) \tag{4.25}$$

$$\leqslant \sum_{j=1}^{M}\text{Prob}\left(\left|\frac{1}{T}\sum_{i=1}^{T}\text{tr}\left(O^{(j)}\hat{\sigma}_i\right)-\text{tr}\left(O^{(j)}\sigma\right)\right|\geqslant\epsilon\right) \tag{4.26}$$

$$\leqslant M\delta' \tag{4.27}$$

if we choose $T \geqslant \frac{\max_j \|O_{\text{TL}}^{(j)}\|_{\text{sh}}}{\delta'\epsilon^2}$. So if we set, $\delta = \delta'/M$, we get the desired precision and confidence.

But the issue here is that this sample complexity is linearly dependent on M. Now we shall explain how this dependency can be improved to $\log M$. The key is to use the median of means estimation instead of sample means estimation. Given T_1T_2 samples $\eta^{(1)}, \eta^{(2)}, \ldots, \eta^{(T_1T_2)}$ of an arbitrary random variable η, the median of means estimator $\mu_{T_1T_2}$ is defined as

$$\mu_{T_1T_2}(\{\eta^{(i)}\}_i) \coloneqq \text{median}\left\{\frac{1}{T_1}\sum_{k=1}^{T_1}\eta^{(k)}, \frac{1}{T_1}\sum_{k=T_1+1}^{2T_1}\eta^{(k)}, \ldots, \frac{1}{T_1}\sum_{k=(T_2-1)T_1+1}^{T_2T_1}\eta^{(k)}\right\}. \tag{4.28}$$

If we use median of means estimation, then a $\mu_{T_1T_2}$ estimate can achieve any precision ϵ and confidence δ if $T_1 \geqslant \frac{34}{\epsilon^2}\text{Var}(\eta)$ and $T_2 \geqslant 2\log(\frac{2}{\delta})$ [JVV86; Bla85].

So, if we use the median of means estimation for shadow estimation, we can achieve precision ϵ and confidence δ' as per Eq 4.2, using a number of samples T_1T_2, where $T_1 \geqslant \frac{34}{\epsilon^2}\max_j \left\|O_{\text{TL}}^{(j)}\right\|_{\text{sh}}$ and $T_2 \geqslant 2\log\left(\frac{2}{\delta'}\right)$. In this setting, if we introduce $\delta = \delta'/M$, we see that the dependency of the total sample of complexity on M is only logarithmic. Hence, we get the following result.

Theorem 2. *[HKP20] Let $\sigma \in \mathcal{L}(\mathbb{C}^{2^n})$ be a quantum state, $O^{(1)}, O^{(2)}, \ldots, O^{(M)} \in \mathbb{H}_{2^n}$, and $\epsilon, \delta \in [0,1]$ be precision and confidence parameters respectively. By using T_1T_2 classical shadows $\left\{\hat{\sigma}^{(1)}, \hat{\sigma}^{(2)}, \ldots, \hat{\sigma}^{(T_1T_2)}\right\}$ where $T_1 \geqslant \frac{34}{\epsilon^2}\max_i \left\|O^{(j)} - \frac{tr(O^{(j)})}{2^n}\mathbb{1}_{2^n}\right\|_{sh}$ and $T_2 \geqslant 2\log\left(\frac{2M}{\delta}\right)$, we have $\left|\mu_{T_1T_2}\left(\{tr\left(O^{(j)}\hat{\sigma}^{(i)}\right)\}_i\right) - tr\left(\sigma O^{(j)}\right)\right| \leqslant \epsilon \;\forall\; j$ with probability*

at least $1-\delta$.

This is remarkable because the number of shadows required to estimate the expectation of the observables depends only on (i) the shadow norm of the observable, (ii) the precision to which we want to estimate the expectation, and (iii) the error probability, and not on the state itself. Also, to estimate the expectation of M observables, we only require $\mathcal{O}(\log M)$ classical shadows loaded. To make this procedure efficient, we should have g defined over unitaries that enables computing $\mathrm{Tw}^{-1}_{(g,\Phi)}(U^\dagger |b\rangle\langle b| U)$ classically feasible, with classical cost polynomial in n. Two such distributions g have been proposed in the original paper [HKP20], each resulting in different ranges of values for the shadow norm.

1. *Pauli ensemble*: In this case, g is the discrete uniform distribution over the set $\left\{Q^{(1)} \otimes \cdots \otimes Q^{(n)} \mid \forall\, Q^{(1)}, \ldots, Q^{(n)} \in \left\{\mathbb{1}, H, HS^\dagger\right\}\right\}$. Here, it turns out that the quantum operations required to generate a shadow are simply measuring each qubit in the eigenbasis of a uniformly randomly sampled non-identity Pauli. In this case, for any observable for the form $O = \bigotimes_{i=1}^{n} O^{(i)}$ with $O^{(i)} \in \mathcal{L}\left(\mathbb{C}^2\right)$, we have $\mathrm{Tw}^{-1}_{(g,\Phi)}(O) = \bigotimes_{i=1}^{n} \left(3O^{(i)} - \mathbb{1}_{\mathbb{C}^2}\right)$ and $\|O\|_{\mathrm{sh}} = \mathcal{O}\left(4^k\right)$, where k is the total number of qubits which is non-trivially acted upon by O. Interestingly, for this ensemble, it has also been proved in [Sac+22] that the usage of sample means estimation will also guarantee a similar exponentially low dependence on M.

   Note that one can achieve similar results by simply carrying out full tomography of the input state on the subsystem where the target local observables act non-trivially. But classical shadow tomography is a simpler protocol and more importantly, can be extended to other classes of observables by using different choices of unitary ensembles.

2. *Clifford ensemble*: In this case, g is the discrete uniform distribution over all n-qubit Clifford gates. Here, $\mathrm{Tw}^{-1}_{(g,\Phi)}(O) = (2^n+1)\, O - \mathbb{1}_{\mathbb{C}^{2^n}}$. The stabilizer formalism [AG04] can be used to work with classical shadows in this scenario. In this case, $\|O\|_{\mathrm{sh}} = \mathcal{O}\left(\|O\|_2^2\right)$.

If the observable is a quantum state (occurs when we want to compute the fidelity with a pure state), one can choose to use the Clifford ensemble, and if the observable is a local observable (or can be written as a linear combination of a few local observables), then we can use the Pauli ensemble. Many relevant observables in nature can be written as linear combinations of a few observables that act non-trivially on only a few qubits [BL08]. This makes classical shadows a very powerful "storage facility" for states as well. In many contexts, this is why classical shadows are considered an alternative to conventional quantum state tomography, which requires resources exponential in the number of qubits. The key fact here is that to estimate many properties of quantum states, up to acceptable precision and quality, one does not need to know the full classical description of the states. Similar complexity guarantees are extended to other important problems involving quantum states such as finding entanglement witnesses [GT09], direct fidelity estimation [FL11], estimating the output of non-linear functions involving quantum states, etc. It was also proved that this protocol saturates the lower bound (in terms of the number of copies of the state required) over all protocols that can be used to predict the outputs of M linear functions defined on quantum states.

Many variants of classical shadows can be found in the literature, such as optimal classical shadows for pure states [GPS24], error mitigated classical shadows [Jna+24], classical shadows for continuous-variable quantum systems [Bec+24], classical shadows for process tomography [Kun+23], classical shadows generated using shallow circuit ensembles [Ber+23], etc. We shall explain the latter in detail in the next section since it is an important tool used in Chapter 5.

4.3 Method

We first explain our approach in a simpler model, with 1-local observables and ALAs built with 2-local circuits, and then extend the results to circuits and observables with arbitrary localities. The detailed proof of our results can be found in Section 4.6 in the Appendix.

For ALSO, we will be using classical shadows generated using the Pauli ensemble. We

first discuss this particular shadow-generating process in more detail and then introduce an improved version of Theorem 2 for the Pauli ensemble that does not use the median of means estimation.

To generate a shadow, the first step is to measure the individual qubits of σ on a random Pauli basis. To this end, for each qubit i, we apply a gate U_i uniformly randomly chosen from $\left\{\mathbb{1}, H, HS^\dagger\right\}$, and then measure it in the computational basis. Let the measurement outcome be $u_i \in \{0, 1\}$. Then a classical shadow of σ is calculated (classically) as

$$\hat{\sigma} = \Phi\left(U_1^\dagger |u_1\rangle\langle u_1| U_1\right) \otimes \cdots \otimes \Phi\left(U_n^\dagger |u_n\rangle\langle u_n| U_n\right), \tag{4.29}$$

where $\Phi(A) = 3A - \mathbb{1}$. As a fully separable matrix, $\hat{\sigma}$ can be stored efficiently as n 2×2 matrices. Furthermore, $\hat{\sigma}$ gives an unbiased estimation of the unknown state σ and hence $\mathrm{tr}\left(O^{(i)}\hat{\sigma}\right)$ is an unbiased estimator of $\mathrm{tr}\left(O^{(i)}\sigma\right)$ for all i.

Specifically, we have:

Theorem 3. *[Sac+22] Let $\sigma \in \mathcal{L}(\mathbb{C}^{2^n})$ be a quantum state. Suppose $O^{(1)}, O^{(2)}, \ldots, O^{(M)} \in \mathcal{L}(\mathbb{C}^{2^n})$ are M k-local observables. For any $\delta, \epsilon \in (0, 1)$, let T be any integer not smaller than $\frac{4^{k+1}}{\epsilon^2} \cdot \log(\frac{2M}{\delta}) \max_i \|O^{(i)}\|_\infty^2$ and define shadow state $\hat{\sigma}_T$ as $\hat{\sigma}_T = \frac{1}{T}\sum_{j=1}^{T} \hat{\sigma}^{(j)}$, where $\hat{\sigma}^{(j)}$ are single-qubit classical shadows as in Eq. (4.29). Then, with probability at least $1 - \delta$ and for all i, we have $\left|tr(O^{(i)}\hat{\sigma}_T) - tr(O^{(i)}\sigma)\right| \leqslant \epsilon$.*

Note that the original version of this theorem required $\|O^{(i)}\|_\infty \leqslant 1$ for all i. However, this can be relaxed by dividing every matrix by $\max_i \|O^{(i)}\|_\infty$ and then estimating with precision $\epsilon / \max_i \|O^{(i)}\|_\infty$.

Moreover, each estimation $\mathrm{tr}\left(O^{(l)}\hat{\sigma}_T\right)$ can be classically computed very efficiently. Let $A_l = (q_{l_1}, \ldots, q_{l_k})$ be the sub-register that $O^{(l)}$ acts non-trivially on and $O^{(l)} = \widetilde{O}^{(l)} \otimes \mathbb{1}_{\mathbb{C}^{2^{n-k}}}$ with $\widetilde{O}^{(l)} \in \mathcal{L}(\mathbb{C}^{2^k})$. To compute expectation with this observable using the shadow $\hat{\sigma}$ in Eq. (4.29), we only need to use the k 2×2 matrices corresponding to the sub-register A_l. Denote by $\hat{\sigma}_T\big|_{A_l}$ the classical shadow obtained by taking an average of T such reduced shadows. Then we have $\mathrm{tr}\left(O^{(l)}\hat{\sigma}_T\right) = \mathrm{tr}\left(\widetilde{O}^{(l)}\hat{\sigma}_T\big|_{A_l}\right)$ and hence it can be computed with

cost exponential only in k and independent of n.

We start with a lemma that forms the backbone of ALSO. Note that the proofs of all theorems and lemmas introduced in this section can be found in Section 4.6.

Lemma 1. *Let d, S and C be defined as in Eq. (4.1), and $\boldsymbol{\theta} \in \mathbb{R}^{\frac{n}{2} \times d \times p}$. For any n-qubit 1-local observable O, we have $\left\| O_{C(\boldsymbol{\theta})^\dagger} \right\|_\infty = \|O\|_\infty$ and $O_{C(\boldsymbol{\theta})^\dagger}$ is $2d$-local, that is, $O_{C(\boldsymbol{\theta})^\dagger} = \widetilde{O}_{C(\boldsymbol{\theta})^\dagger_A}$ for some sub-register A of $2d$ qubits.*

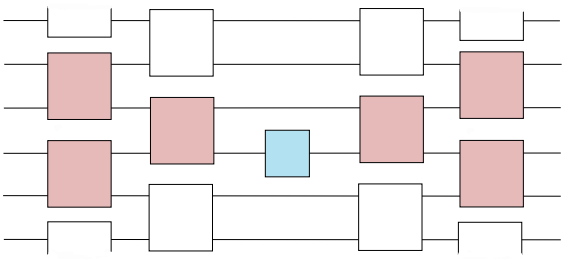


Figure 4.3: The structure of $O_{C(\boldsymbol{\theta})^\dagger} = U(\boldsymbol{\theta})^\dagger O U(\boldsymbol{\theta})$ where the blue box is a 1-local observable, and all other boxes are S sub-circuits. Except for the red ones, all other sub-circuits cancel each other out, resulting in $O_{C(\boldsymbol{\theta})^\dagger}$ being $2d$-local.

Moreover, from Figure 4.3, we can see that for any $\boldsymbol{\theta}$, $\widetilde{O}_{C(\boldsymbol{\theta})^\dagger}$ can be computed with cost exponential only in d using tensor contractions of the observable with all the S gates marked in red.

Let $O = \sum_{i=1}^{D} O^{(i)}$ be an observable, where each $O^{(i)}$ is 1-local. Assume that we are using an iterative optimization algorithm, one that takes as input a target function and outputs its optimizer, to find the maximizer of Eq. (3.2), and the whole optimization procedure requires M function evaluations of the form $f_{\sigma,O}\left(\boldsymbol{\theta}^{(1)}\right), f_{\sigma,O}\left(\boldsymbol{\theta}^{(2)}\right), \ldots, f_{\sigma,O}\left(\boldsymbol{\theta}^{(M)}\right)$. Lemma 1 says that each function evaluation can be seen as estimating the expectation of $2d$-local observables, because $f_{\sigma,O}(\boldsymbol{\theta}) = \sum_{i=1}^{D} \operatorname{tr}\left(O^{(i)} \sigma_{C(\boldsymbol{\theta})}\right) = \sum_{i=1}^{D} \operatorname{tr}\left(O^{(i)}_{C(\boldsymbol{\theta})^\dagger} \sigma\right)$.

Now using Theorem 3, we can estimate all the M function evaluations and the whole ALSO algorithm goes as follows:

1. Load $T = \mathcal{O}(\log(M) \cdot \operatorname{poly}(n))$ classical shadows of σ. Let $\hat{\sigma}_T$ be the shadow state (cf. Theorem 3).

2. For all i, compute $\hat{\sigma_T}\big|_{A_i}$, where A_i is the sub-register that $O^{(i)}_{C(\boldsymbol{\theta}^\dagger)}$ acts non-trivially on.

3. Use the iterative optimization algorithm to optimize the target function $\hat{f}_{\sigma,O}(\boldsymbol{\theta}) = \sum_{i=1}^{D} \mathrm{tr}\left(\widetilde{O}^{(i)}_{C(\boldsymbol{\theta}^\dagger)}\hat{\sigma_T}\big|_{A_i}\right)$.

Note that the cost of classical computation is dominated by the computation of $\sum_{i=1}^{D} \mathrm{tr}\left(\widetilde{O}^{(i)}_{C(\boldsymbol{\theta}^\dagger)}\hat{\sigma_T}\big|_{A_i}\right)$ and so it scales exponentially only in d. Hence, when $d \in \mathcal{O}(\log n)$, the classical computational cost scales polynomially on n.

4.4 Sample Complexity

In this section, we discuss the sample complexity of the protocol, that is, the range of values of T that guarantee good estimations of all the function evaluations. We show that when $d \in \mathcal{O}(\log n)$, the sample complexity is $\mathcal{O}(\log(M) \cdot \mathrm{poly}(n))$.

Theorem 4. *Let d, S and C be defined as in Eq. (4.1). Suppose σ is an arbitrary n-qubit state and $O = \sum_{i=1}^{D} O^{(i)}$, where each $O^{(i)}$ is an n-qubit 1-local observable. Then, for any $\delta, \epsilon \in (0,1)$ and any M parameter tensors $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(M)}$, all values $f_{\sigma,O}\left(\boldsymbol{\theta}^{(m)}\right)$ can be estimated using $\hat{f}_{\sigma,O}(\boldsymbol{\theta}^{(m)}) := tr\left(O_{C\left(\boldsymbol{\theta}^{(m)}\right)^\dagger}\hat{\sigma}_T\right)$ with the guarantee*

$$Prob\left(\bigcap_{m=1}^{M}\left[\left|f_{\sigma,O}\left(\boldsymbol{\theta}^{(m)}\right) - \hat{f}_{\sigma,O}\left(\boldsymbol{\theta}^{(m)}\right)\right| \leqslant \epsilon\right]\right) \geqslant 1-\delta, \tag{4.30}$$

where $T \geqslant D^2 \log\left(\frac{2MD}{\delta}\right) \cdot \frac{4^{2d+1}}{\epsilon^2} \max_i \|O_i\|_\infty^2$.

This is remarkable as, without using classical shadows, we may need to estimate $f_{\sigma,O}(\boldsymbol{\theta})$ for any parameter tensor $\boldsymbol{\theta}$ through measurements. Suppose K copies of σ are consumed to estimate each of these values. Then, we end up consuming MK copies of σ, which can be exponentially larger than the number consumed by ALSO.

In our method, the measurements that we have to make are solely for computing the classical shadows and hence are independent of all $\boldsymbol{\theta}^{(m)}$. Moreover, each measurement

outcome can be reused multiple times. In the standard method of training VQAs, we are not able to reuse the measurement outcomes that are made as part of the optimization, because each measurement outcome is dependent on the input parameter $\boldsymbol{\theta}^{(m)}$. This is a crucial reason why ALSO is a much more appealing option to optimize these functions, especially from a practical perspective where one has to do hyperparameter tuning, find the right classical optimizer, etc.

One important point to note is that even though the constants look large, in practice, we need not necessarily require this many copies (classical shadows) of σ. This is illustrated in our experimental results, where we are able to match the results of ideal VQA simulations (simulations that use infinite copies of the input state σ) by using a number of copies of σ orders of magnitude fewer than the number suggested by Theorem 4.

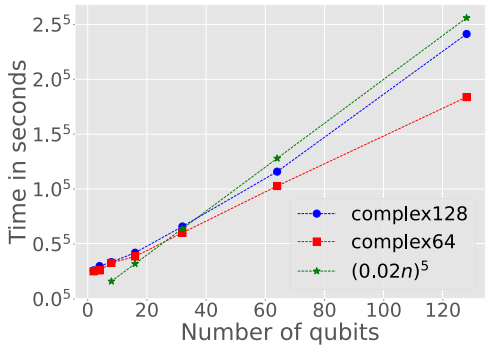


Figure 4.4: Plot showing the time (in seconds) taken for a single function evaluation using ALSO. Here, $d \in \lfloor \log n \rfloor$, S is a 2-qubit parameterized sub-circuit and O is a 1-qubit observable. Along with the execution times, we plot the function $(0.02n)^5$ to highlight the polynomial dependence of time on the number of qubits.

The space complexity of the protocol is dominated by the storage of the matrices $\hat{\sigma}_T\big|_{A_i}$. Since the dimension of each of these matrices is 4^d, we need $16^d M$ complex numbers to store all of them. For example, let $n, M = 50$, $d = 5$. Then we see that if we are using 128 bits to store each complex number, then we only require 838MB to store all matrices $\hat{\sigma}_T\big|_{A_i}$. Time taken for single function evaluations is plotted in Figure 4.4. On the x-axis, we have the number of qubits, and on the y-axis, we have the time (in seconds) taken to compute a single function evaluation, averaged over 5 cases. In each case, $d \in \lfloor \log n \rfloor$

and the observable O is a 1-local observable, with S being the circuit in Figure 4.1 (b). We plot the results showing a polynomial dependence of time on the number of qubits for both 'complex128' and 'complex64' being used as datatypes in Python. The simulation was carried out on a laptop with 16GB RAM and 2.6GHz Intel i7 processor.

One can easily generalize Theorem 4 for arbitrarily local observables and circuits. In a similar setting, if we use k_0-local parameterized circuits and an observable that is a sum of k_1-local observables, then we can carry out an iterative optimization algorithm with all function evaluations satisfying Eq. (4.32) using

$$T \geqslant \frac{D^2}{\epsilon^2} \cdot \log\left(\frac{2MD}{\delta}\right) \cdot 4^{k_1+(2k_0-2)d-1} \cdot \max_i \|O_i\|_\infty^2$$

copies of the input state. This is because for each increment in-depth, the locality of $O_{C(\boldsymbol{\theta})^\dagger}^{(i)}$ increases by at most $2k_0 - 2$, starting from k_1.

4.5 Simulation Results

In this section, we discuss the experimental results comparing the performance of ALSO and the standard VQA in two applications: state preparation and quantum autoencoder.

4.5.1 Experiments Set-Up

For all experiments, each brick-like sub-circuit $S(\boldsymbol{\theta}_{ij})$ (cf. Figure 4.1 (a)) has the form given in Figure 4.1 (b). The simulation results presented in this section (except for Table 4.1) have used Simultaneous Perturbation Stochastic Approximation [Spa92] (SPSA), where the converging sequences used for state preparation and quantum autoencoder are, respectively, $c_r = a_r = r^{-0.5}$ and $c_r = a_r = r^{-0.3}$.

In the following, we denote by ALSO-T the ALSO algorithm that uses T shadows and by VQA-K the VQA algorithm that consumes per function evaluation K state copies (for state preparation) or K samples from $\mathcal{Z} = \{(p_i, |\psi_i\rangle)\}$ (for quantum autoencoder). In addition, we write VQA-infinite for the VQA algorithm which has access to an infinite

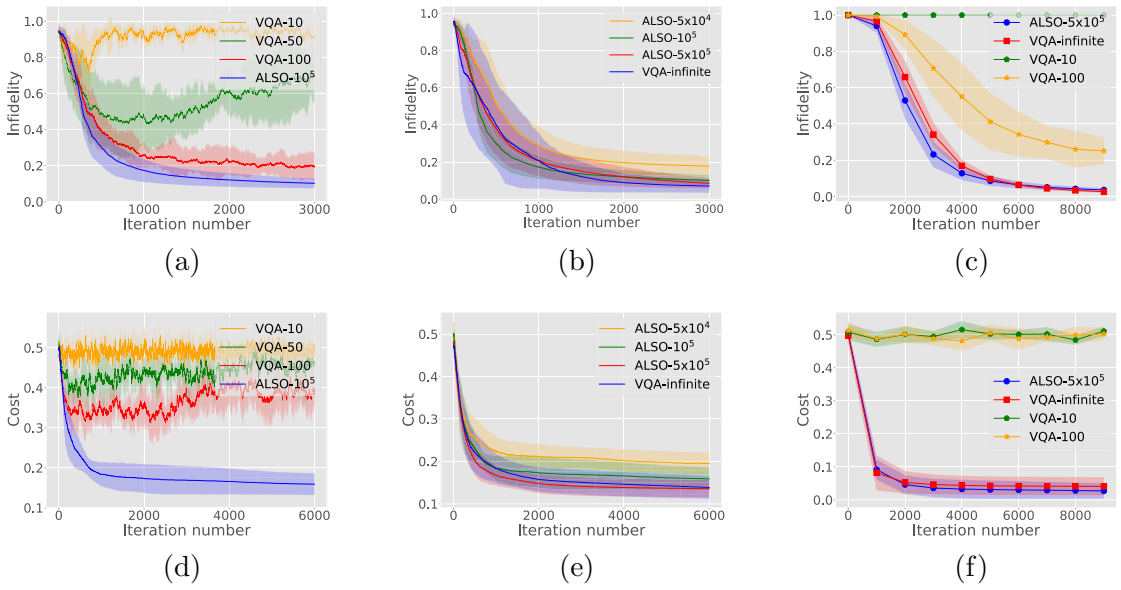


Figure 4.5: Simulation results for state preparation (a-c) and quantum autoencoder (d-f) using SPSA. Each graph corresponds to 5 instances of a problem. VQA-K consumes K copies (samples) per function evaluation while ALSO-T consumes T copies (samples) in total. In (a) and (c), we compare the performance of ALSO with standard VQA in the case of 8-qubit problems. Here, VQA-10, VQA-50, and VQA-100 will consume 4.8×10^5 (4.8×10^5), 2.4×10^6 (2.4×10^6) and 4.8×10^6 (4.8×10^6) copies (samples) respectively while ALSO-10^5 consumes only 10^5 (10^5) copies (samples), and still outperforms VQA considerably. Continuing in the 8-qubit scenario, In (b) and (d), we compare the performance of ALSO with the ideal VQA that consumes infinite copies, and we see that ALSO is able to almost match the results of this ideal VQA using a modest 5×10^5 (5×10^5) copies (samples). In (c) and (f), we plot the results of similar experiments carried out on 30-qubit states. In this case, VQA consumes 5.4×10^6 (1.8×10^6) and 5.4×10^7 (1.8×10^7) copies (samples) respectively. Note that here, only iteration numbers that are multiples of 1000 are plotted.

number of state copies.

Let R be the total number of iterations of SPSA. Since SPSA requires 2 function evaluations per iteration, for state preparation and quantum autoencoder, VQA-K will consume $2KRn$ and $2KRn_B$ state copies respectively.

4.5.2 State Preparation Experiments

For the state preparation problem, we first consider the case when $n = 8$, i.e., the target state is an 8-qubit state. In each experiment, the target state is compatible with an ALA.

We repeat the experiments for five different target states and our results are shown in Figure 4.5 (a,b), where each plot corresponds to five different instances of a problem. At any value on the x-axis, we plot the mean of infidelity/cost values across the five different experiments that were carried out. The colored area of the plot is marked on top and bottom by the mean plus and minus the standard deviation of the 5 values at each point respectively.

In Figure 4.5 (a), VQA-10 consumes $2KRn = 2 \times 10 \times 3000 \times 8 = 4.8 \times 10^5$ state copies and in a similar manner, the other VQA algorithms consume 2.4×10^6 and 4.8×10^6 state copies, which are 4.8x, 24x, and 48x of that ALSO consumes. Furthermore, from Fig. 4.5 (b), we can see that ALSO closely matches the outcome of VQA-infinite with only 5×10^5 state copies.

Moving on from the 8-qubit scenario, we then carry out similar experiments for 30-qubit systems. Fig. 4.5 (c) shows the results, where, as in [Cer+21b] (2021), all states involved are computational basis states.

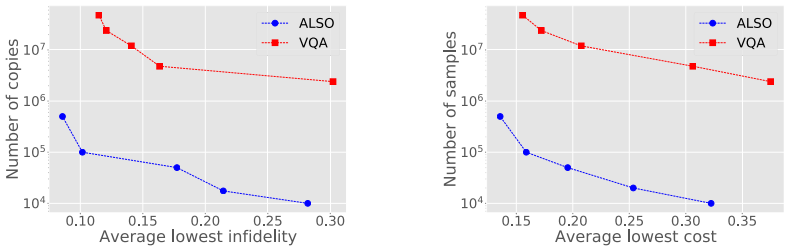
We note in this case, VQA-10 consumes $2 \times 10 \times 9000 \times 30 = 5.4 \times 10^6$ state copies while ALSO-T remains unchanged with the change of n from 8 to 30. From the figure, it is clear that a similar conclusion for 8-qubit state preparation also applies to 30-qubit state preparation. In particular, ALSO (with 5×10^5 samples) significantly outperforms VQA with 100x more samples.

4.5.3 Quantum Autoencoder Experiments

For quantum autoencoder, similar experiments are carried out for both 8- and 30-qubit systems. Ensembles containing two pure states $|\psi_1\rangle$ and $|\psi_2\rangle$ are chosen with $p_1 = 1/3$ and $p_2 = 2/3$, and n_B is set as 4 and 10, respectively, for 8- and 30-qubit systems. We repeat the experiments for five different ensembles. The results are summarised in Figures 4.5 (d), (e), and (f), which have the same explanation as those in Figures 4.5 (a), (b), and (c). Note that the cost values plotted here are $1 - f_{\sigma,F}(\boldsymbol{\theta})$ (cf. Section 3.4.3). Moreover, it is the actual true cost and not their estimations. From the figures, we can see similar

conclusion we have obtained for state preparation also holds for quantum autoencoder. It seems that in this case ALSO with 10^5 samples significantly outperforms VQA with 48×10^5 samples and ALSO with 5×10^5 samples can often match VQA-infinite.

4.5.4 Resource Consumption for the Same Objective



(a) 8-qubit state preparation (b) 8-qubit quantum autoencoder

Figure 4.6: Resource requirement for different objectives. On the x-axis, we plot the least average infidelity (cost) of 5 instances of the corresponding problem. On the y-axis, we plot the number of copies (samples) that were required to achieve them using SPSA. We see that ALSO achieves an order of magnitude savings in the number of copies (samples).

In the above, we compared the performance of ALSO and VQA algorithms with predetermined resources. The efficiency of ALSO over VQA can also be illustrated by comparing the resource consumption required for the same objective. Given an objective, which can be either the average lowest infidelity or the average lowest cost, we carry out experiments to check how many state copies or samples are required for ALSO or VQA to achieve the objective. The results are presented in Figure 4.6, where each point represents the average of five instances. It is clear that ALSO achieves a huge advantage in the number of state copies or samples that were required to achieve the specific levels of quality.

4.5.5 More Iterations by Using Powell's Method

All the simulation results discussed above have used SPSA to find the optimal parameters. In each case, we set the maximum iterations to be the same for ALSO and VQA. From a practical point of view, this is unfair as ALSO can carry out more iterations with the given number of state copies or samples.

	state preparation		quantum autoencoder	
	#copies	infidelity	#samples	cost
VQA-10^2	5×10^5	0.921	5.6×10^5	0.494
VQA-10^3	1.3×10^7	0.348	4.6×10^6	0.408
VQA-10^4	3.3×10^8	0.094	10^8	0.250
VQA-10^5	4×10^9	0.069	2.1×10^9	0.188
ALSO	5×10^5	0.004	5×10^5	0.117

Table 4.1: Simulation results comparing the performance of ALSO and the standard VQA when using Powell's method, where infidelity (cost) is the average lowest infidelity (cost) of five instances, and #copies (#samples) is the number of state copies (samples) used by the algorithm.

To further demonstrate this advantage of ALSO, we turn to Powell's method [Pow64] to optimize the parameters. We carry out 8-qubit state preparation as well as quantum autoencoder optimizations and the results are presented in Table 4.1. In the infidelity (cost) columns, each entry is an average optimal infidelity (cost) of 5 instances of the problem, and we give an approximation of the average number of copies consumed (except for ALSO where exactly 5×10^5 copies are consumed) to achieve these values in the #copies (#samples) columns.

We set 5×10^4 as an upper limit on the total number of function evaluations for VQA. But, since ALSO does not consume any copies for more iterations, we don't set any limit in the case of ALSO. As we can see, our approach greatly outperforms VQA in this case. Interestingly, in the case of VQA, the optimizers terminated in $5 \times 10^3 - 3 \times 10^4$ function evaluations in most cases. Only for the state preparation problem and with 10^5 copies consumed per function evaluation, we saw the optimizer exceeding the 5×10^4 limit. We also observe that VQA with Powell's method performs very poorly compared to VQA with SPSA when K is small, which is possibly due to the inherent ability of SPSA to deal with noisy functions.

4.6 Proofs of All Theorems

Here, we present the proofs of Lemma 1 and Theorem 4. First, we recall

Theorem 4. *[Sac+22] Let $\sigma \in \mathcal{L}(\mathbb{C}^{2^n})$ be a quantum state. Suppose $O^{(1)}, O^{(2)}, \ldots, O^{(M)} \in \mathcal{L}(\mathbb{C}^{2^n})$ are M k-local observables. For any $\delta, \epsilon \in (0,1)$, let T be any integer not smaller than $\frac{4^{k+1}}{\epsilon^2} \cdot \log(\frac{2M}{\delta}) \max_i \|O^{(i)}\|_\infty^2$ and define shadow state $\hat{\sigma}_T$ as $\hat{\sigma}_T = \frac{1}{T}\sum_{j=1}^{T} \hat{\sigma}^{(j)}$, where $\hat{\sigma}^{(j)}$ are single-qubit classical shadows as in Eq. (4.29). Then, with probability at least $1-\delta$ and for all i, we have $|tr(O_i\hat{\sigma}_T) - tr(O_i\sigma)| \leqslant \epsilon$.*

Note that the ALA used here has the form

$$U(\boldsymbol{\theta}) = \prod_{j=1}^{d}\prod_{i=1}^{n/2} S(\boldsymbol{\theta}_{ij})[2(i-1)\oplus j, 2(i-1)\oplus j\oplus 1]. \tag{4.31}$$

Lemma 1. *Let d, S and C be defined as in Eq. (4.1), and $\boldsymbol{\theta} \in \mathbb{R}^{\frac{n}{2}\times d\times p}$. For any n-qubit 1-local observable O, we have $\left\|O_{C(\boldsymbol{\theta})^\dagger}\right\|_\infty = \|O\|_\infty$ and $O_{C(\boldsymbol{\theta})^\dagger}$ is $2d$-local, that is, $O_{C(\boldsymbol{\theta})^\dagger} = \widetilde{O}_{C(\boldsymbol{\theta})^\dagger[A]}$ for some sub-register A of $2d$ qubits.*

Proof. For any $\boldsymbol{\theta}$, $O_{C(\boldsymbol{\theta})^\dagger}$ is obtained by conjugating O with a unitary matrix. This means that the eigenvalues of O and $O_{C(\boldsymbol{\theta})^\dagger}$ are the same. So, $\left\|O_{C(\boldsymbol{\theta})^\dagger}\right\|_\infty = \|O\|_\infty$. Figure 4.3 shows the structure of $O_{C(\boldsymbol{\theta})^\dagger}$. If $d = 1$, then $O_{C(\boldsymbol{\theta})^\dagger}$ will be an observable with locality 2 as all blocks of the parameterized circuit except the ones acting on the qubit where the observable acts on will cancel out. Similarly, if $d = 2$, then $O_{C(\boldsymbol{\theta})^\dagger}$ will have locality 4. For each increment in d, the locality of $O_{C(\boldsymbol{\theta})^\dagger}$ increases by 2. So, the locality of $O_{C(\boldsymbol{\theta})^\dagger}$ is $2d$. □

Theorem 5. *Let d, S and C be defined as in Eq. (4.1). Suppose σ is an arbitrary n-qubit state and $O = \sum_{i=1}^{D} O^{(i)}$, where each $O^{(i)}$ is an n-qubit 1-local observable. Then, for any $\delta, \epsilon \in (0,1)$ and any M parameter tensors $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(M)}$, all values $f_{\sigma,O}\left(\boldsymbol{\theta}^{(m)}\right)$ can be estimated using $\hat{f}_{\sigma,O}(\boldsymbol{\theta}^{(m)}) := tr\left(O_{C\left(\boldsymbol{\theta}^{(m)}\right)^\dagger}\hat{\sigma}_T\right)$ with the guarantee*

$$Prob\left(\bigcap_{m=1}^{M}\left[\left|f_{\sigma,O}\left(\boldsymbol{\theta}^{(m)}\right) - \hat{f}_{\sigma,O}\left(\boldsymbol{\theta}^{(m)}\right)\right| \leqslant \epsilon\right]\right) \geqslant 1-\delta, \tag{4.32}$$

where $T \geqslant D^2 \log\left(\frac{2MD}{\delta}\right) \cdot \frac{4^{2d+1}}{\epsilon^2} \max_i \|O_i\|_\infty^2$.

Proof. Note that for any $\boldsymbol{\theta} \in \mathbb{R}^{\frac{n}{2} \times d \times p}$, we have $f_{\sigma,O}(\boldsymbol{\theta}) = \operatorname{tr}(O\sigma_{C(\boldsymbol{\theta})} = \operatorname{tr}(O_{C(\boldsymbol{\theta})^\dagger}\sigma)$.

First, consider the case when $D = 1$. From Lemma 1, we know that $O_{C(\boldsymbol{\theta})^\dagger}$ is a $2d$-local operator, with $\left\|O_{C(\boldsymbol{\theta})^\dagger}\right\|_\infty = \|O\|_\infty$. This means that all the function evaluations can be seen as computing expectations of observables with locality $2d$ and $\|\cdot\|_\infty$ the same as O.

The function evaluations that we have to approximate are $f_{\sigma,O}\left(\boldsymbol{\theta}^{(1)}\right), f_{\sigma,O}\left(\boldsymbol{\theta}^{(2)}\right), \ldots,$ $f_{\sigma,O}\left(\boldsymbol{\theta}^{(M)}\right)$ and hence the observables whose expectation that we have to find are $O_{C\left(\boldsymbol{\theta}^{(1)}\right)^\dagger}$, $O_{C\left(\boldsymbol{\theta}^{(2)}\right)^\dagger}, \ldots, O_{C\left(\boldsymbol{\theta}^{(M)}\right)^\dagger}$. Then from Theorem 4, by using

$$T \geqslant \log\left(\frac{2M}{\delta}\right) \cdot \frac{4^{2d+1}}{\epsilon^2}\|O\|_\infty^2 \tag{4.33}$$

classical shadows of σ, we can estimate the expectation of all of these observables to precision ϵ with a probability of at least $1 - \delta$.

When $D > 1$, we no longer have O being necessarily a 1-local observable. So, for every m we have to estimate the expectation of σ with the observables $O^{(i)}_{C\left(\boldsymbol{\theta}^{(m)}\right)^\dagger}$ for all i, and compute their sum. Hence, the total number of observables that we have to compute expectations with is now MD.

By using

$$T \geqslant \log\left(\frac{2MD}{\delta}\right) \cdot \frac{4^{2d+1}}{(\epsilon/D)^2} \max_i \|O^{(i)}\|_\infty^2 \tag{4.34}$$

classical shadows, we will have approximations of $f_{\sigma,O^{(i)}}(\boldsymbol{\theta}^{(m)})$ given as $\hat{f}_{\sigma,O^{(i)}}(\boldsymbol{\theta}^{(m)}) = \operatorname{tr}(O^{(i)}_{C(\boldsymbol{\theta})^\dagger}\hat{\sigma}_T)$ such that $\operatorname{Prob}(E) \geqslant 1 - \delta$ for

$$E = \bigcap_{m=1}^{M}\bigcap_{i=1}^{D}\left[\left|f_{\sigma,O^{(i)}}(\boldsymbol{\theta}^{(m)}) - \hat{f}_{\sigma,O^{(i)}}(\boldsymbol{\theta}^{(m)})\right| \leqslant \epsilon/D\right]. \tag{4.35}$$

This is because, if we consider the difference $\left|f_{\sigma,O^{(i)}}\left(\boldsymbol{\theta}^{(m)}\right) - \hat{f}_{\sigma,O^{(i)}}\left(\boldsymbol{\theta}^{(m)}\right)\right|$ being less than or equal to ϵ/D to be an event, then Theorem 4 says that the intersection of these events for all i, m occurs with probability at least $1 - \delta$.

Then, with probability at least $1 - \delta$, for all m we have

$$\sum_{i=1}^{D} \left| f_{\sigma,O^{(i)}}(\boldsymbol{\theta}^{(m)}) - \hat{f}_{\sigma,O^{(i)}}(\boldsymbol{\theta}^{(m)}) \right| \leqslant \epsilon$$

and thus

$$\left| f_{\sigma,O}(\boldsymbol{\theta}^{(m)}) - \hat{f}_{\sigma,O}(\boldsymbol{\theta}^{(m)}) \right| = \left| \sum_{i=1}^{D} \left[f_{\sigma,O^{(i)}}(\boldsymbol{\theta}^{(m)}) - \hat{f}_{\sigma,O^{(i)}}(\boldsymbol{\theta}^{(m)}) \right] \right| \leqslant \epsilon,$$

where $\hat{f}_{\sigma,O}\left(\boldsymbol{\theta}^{(m)}\right) = \sum_{i=1}^{D} \hat{f}_{\sigma,O^{(i)}}\left(\boldsymbol{\theta}^{(m)}\right)$. So all approximations $\hat{f}_{\sigma,O}\left(\boldsymbol{\theta}^{(1)}\right), \ldots, \hat{f}_{\sigma,O}\left(\boldsymbol{\theta}^{(M)}\right)$ satisfy Eq (4.32). Furthermore, since each single-copy classical shadow requires only 1 copy of the input state σ, we can estimate all the function evaluations to the required quality with T copies of σ. □

4.7 Related Works

Studying alternating layered VQAs as optimizations of local parameterized observables is already considered in [Oka+22]. Here, the locality of $O_{C(\boldsymbol{\theta})^\dagger}$ is leveraged to implement variational quantum eigensolvers. ALSO can be seen as a generalization of this method because the input state can be arbitrary in our setting (due to the use of classical shadows).

[Fon+22] have experimentally shown that in certain cases, the quantum alternating operator ansatz [FGG14] and the Hamiltonian variataionl ansatz [Wie+20] can be trained by estimations in the Fourier basis [SSM21], with $\mathcal{O}(\text{poly}(n))$ copies of the input state and $\mathcal{O}(\text{poly}(n))$ classical computational cost. Similarly, [YBL20] showed that reinforcement learning [DN08] can be used to optimize the learning process of variational parameters in the former ansatz. In our work, we focus on the ALA and prove theoretically the existence of sample efficient training methods for this ansatz.

In [Sto+20] and [BK22], new classical optimization algorithms are introduced and analyzed (the latter also uses classical shadows) and are shown to converge using much fewer iterations compared to the standard gradient descent. However, since ALSO is agnostic towards the choice of the classical optimizer, our method can be used to boost the performance of these methods by significantly reducing the number of state copies

used. Variational shadow quantum circuits developed by [LSW21b], extract local classical features by focusing on a series of local subcircuits. Although inspired by the classical shadow work, the approach itself does not use classical shadows.

Chapter 5

Ansatz Independent Shadow Optimization

5.1 Overview

Recall that 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 ALA. This limitation becomes significant when the optimal circuit or state cannot be approximated with ALAs.

Shallow shadow technique [Ber+23] (cf. Section 5.2) 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), another method that achieves an exponential reduction in quantum resources for VQA training. AISO is compatible with almost any shallow 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: state preparation and 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 objective function while consuming exponentially fewer copies of the input state compared to standard VQA. This enables more iterations, and 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 using various types of ansatzes. 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 advantages are experimentally demonstrated in state preparation and VQCS, where we show that AISO significantly outperforms standard VQA with the same number of copies across four different ansatzes: ALA, MERA, HEA, and TTN. These shallow ansatzes have been researched and their potential have been studied in many works in the literature [Leo+24; Tan21; PKH24; Kar+21; Kan+17; Cer+20; NY21; CCL19; MV18; EV13; FV12]. Such shallow ansatzes induce less hardware related noise since the depth and gate counts are small. Also, all of these ansatzes are proven to avoid barren plateaus

when used in combination with local observables [Cer+20; Pes+21]. But, the key feature that enables these benefits, the shallowness of the circuit, also restricts the class of unitaries that these circuits can represent. More specifically, the types of entanglement that these circuits can generate are greatly limited.

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. 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.

5.2 Shallow Shadows

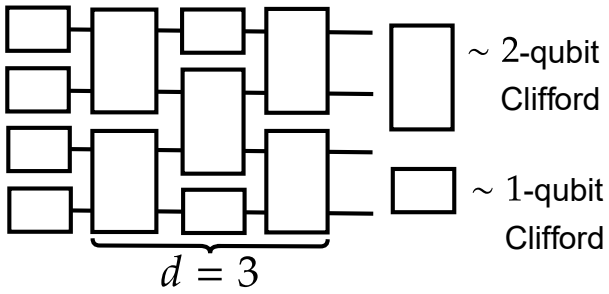


Figure 5.1: The structure of the unitary ensemble used to generate shallow shadows. Each 1-qubit gate is a uniformly randomly sampled 1-qubit Clifford gate, while each 2-qubit block here is a uniformly randomly sampled 2-qubit Clifford gate. d is the number of vertical layers of 2-qubit gates in the circuit.

Shallow shadows were introduced in [Ber+23]. These shadows are generated using an ensemble of shallow-depth circuits $\mathcal{U}_d$ (with depth d), given in Figure 5.1. Each 1-qubit gate is a uniformly randomly sampled 1-qubit Clifford gate, while each two-qubit subcircuit here is a uniformly randomly sampled 2-qubit Clifford gate. The shadow can be classically computed and stored in an MPS form, with cost $\mathcal{O}\left(2^d\right)$. This is enabled by first showing that $\mathrm{Tw}_{\mathcal{U}_d,\Phi}$ is a diagonal operator in the Pauli basis, and the diagonal elements seen as a vector has a known MPS structure with bond dimension $\mathcal{O}\left(2^d\right)$. Since it is a diagonal operator, its inverse must simply be its elementwise inverse. To find this

inverse, we minimize the objective function of the form

$$\zeta(A) = \|\mathrm{diag}(\mathrm{Tw}_{\mathcal{U}_d,\Phi}) * A - \mathbf{1}\|_2^2, \tag{5.1}$$

where $\mathbf{1}$ is the vector with all entries 1. All objects featured in the above expression can be computed using MPS methods with cost scaling as $\mathcal{O}\left(2^d\right)$. Moreover, one can see that as $\zeta(A)$ gets smaller, A will get closer to the diagonal part of $\mathrm{Tw}^{-1}_{\mathcal{U}_d,\Phi}$. Also, for any observable whose Pauli basis coordinates admit an MPS decomposition with bond dimension $\mathcal{O}\left(2^d\right)$, computation of expectations with shallow shadows can also be carried out using cost $\mathcal{O}\left(\mathrm{poly}\left(2^d\right)\right)$ since all circuits in the ensemble have depth d.

Its effect can be seen as a generalization of both the ensembles that we have discussed in the previous section. When $d = 0$, we get the Pauli ensemble, and when $d \to \infty$, we get the Clifford ensemble. Moreover, it is shown that when $d \in \Theta(\log n)$, the shallow shadow ensemble exhibits the properties of both the Pauli as well as the Clifford ensembles to some degree. That is, shallow shadows can be used to estimate expectations of local observables as well as observables of low Frobenius norm.

Formally, we have

Theorem 5. *[Ber+23] If $d \in \Theta(\log n)$, then $\|O\|^2_{\mathbb{1}/2^n,\mathcal{U}_d} \leqslant 4\|O\|_2^2$ for any observable O with $tr(O) = 0$. Furthermore, let l be the maximum distance between two qubits on which O is not acting as $\mathbb{1}_{\mathbb{C}^2}$. Then*

$$\|O\|^2_{\mathbb{1}/2^n,\mathcal{U}_d} \leqslant n^{\mathcal{O}(1)} \frac{\|O\|_2^2}{2^n}. \tag{5.2}$$

The term $\|O\|^2_{\mathbb{1}/2^n,\mathcal{U}_d}$ in Theorem 5 is called the *locally scrambled shadow norm*. To see the properties of this norm, notice that for any state 1-design $\mathcal{D}_1$,

$$\mathbb{E}_{|\psi\rangle\sim\mathcal{D}_1}(|\psi\rangle\langle\psi|) = \int_U U|0\rangle\langle 0|U^\dagger \mathrm{d}U = \frac{\mathbb{1}}{2^n}. \tag{5.3}$$

The first equality can be seen from the relationship between state 1-design and unitary

1-design discussed in Section 2.5 and the last equality can be seen from Lemma 6 in the Appendix.

Then, we have $\|O\|^2_{\mathbb{1}/2^n,\mathcal{U}} = \mathbb{E}_{\sigma\sim\mathcal{D}_1}\|O\|^2_{\sigma,\mathcal{U}}$. 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 because of the presence of the maximization over all states.

To generate a shallow shadow, we first apply a circuit U sampled from the ensemble $\mathcal{U}_d$ (cf. Figure 5.1), and then measure the resultant state according to the computational basis to obtain an n-bit string u. A shallow shadow is computed classically as

$$\hat{\sigma}_{U,u} = \mathrm{Tw}^{-1}_{\mathcal{U}_d,\Phi}\left(U^\dagger |u\rangle\langle u| U\right), \tag{5.4}$$

5.3 Ansatz Independent Shadow Optimization

Now, we shall explain the main idea and theoretical results behind AISO.

For any quantum circuit V, we recall the definition of R_V given in Section 2.4.2. For any qubit i, this is the number of 2-qubit gates being applied on any qubits j, k such that $j \leqslant i \leqslant k$. Let $R_V = \max_i R_{V,i}$. We require our ansatz C to have $R_C \in \mathcal{O}(\log n)$. Most shallow ansatzes used in the literature satisfy this. Let $f_{\sigma,O}\left(\boldsymbol{\theta}^{(1)}\right), f_{\sigma,O}\left(\boldsymbol{\theta}^{(2)}\right), \ldots, f_{\sigma,O}\left(\boldsymbol{\theta}^{(M)}\right)$, be function evaluations that one encountered while optimizing Eq. (3.2) using an iterative optimization algorithm.

Each function evaluation can be seen as estimating the expectation of σ with parameterized observables of the form $O_{C(\boldsymbol{\theta})^\dagger}$ because

$$f_{\sigma,O}(\boldsymbol{\theta}) = \mathrm{tr}\left(O\sigma_{C(\boldsymbol{\theta})}\right) = \mathrm{tr}\left(O_{C(\boldsymbol{\theta})^\dagger}\sigma\right) \tag{5.5}$$

Moreover, the Frobenius norm remains invariant since $\|O\|_2^2 = \|O_V\|_2^2$ for any unitary V.

Now, using Theorems 2 and 5, we can estimate all M function evaluations using shallow

shadows, and the AISO protocol goes as follows.

1. Choose precision and confidence parameters $\epsilon, \delta \in (0,1)$. Let $\gamma \geqslant 1/\delta$. Generate $T_1 T_2$ shallow shadows of σ (with $d \in \Theta(\log n)$), where

$$T_1 \geqslant 2 \log\left(\frac{2(\gamma-1)M}{\gamma\delta - 1}\right), \; T_2 \geqslant \frac{136}{\epsilon^2}\gamma \|O\|_2^2. \tag{5.6}$$

   Let them be $\hat{\sigma}_{U_1,u_1}, \hat{\sigma}_{U_2,u_2}, \ldots, \hat{\sigma}_{U_{T_1T_2},u_{T_1T_2}}$.

2. Use the iterative optimization algorithm to optimize the target function

$$\hat{f}_{\sigma,O}(\boldsymbol{\theta}) \coloneqq \mu_{T_1,T_2}\left(\left\{ f_{\hat{\sigma}_{U_j,u_j},O}(\boldsymbol{\theta}) \;\middle|\; 1 \leqslant j \leqslant T_1T_2 \right\}\right). \tag{5.7}$$

Now, we shall prove that when T_1 and T_2 satisfy Eq. (5.6), the AISO protocol achieves the desired precision and confidence. Proofs of all theorems discussed in this chapter can be found in Section 5.7.

Theorem 6. *Let σ be an n-qubit pure state sampled from a state 1-design $\mathcal{D}_1$. For any $\delta, \epsilon \in (0,1)$, $\gamma > 1/\delta$, and any $M \in \mathbb{N}$, let T_1 and T_2 satisfy Eq. (5.6). Then, for any parameter vectors $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(C)}$, with probability at least $1-\delta$, we have $|f_{\sigma,O}\left(\boldsymbol{\theta}^{(m)}\right) - \hat{f}_{\sigma,O}\left(\boldsymbol{\theta}^{(m)}\right)| \leqslant \epsilon$ for all $1 \leqslant m \leqslant M$, where $f_{\sigma,O}\left(\boldsymbol{\theta}^{(m)}\right)$ and $\hat{f}_{\sigma,O}\left(\boldsymbol{\theta}^{(m)}\right)$ are defined in Eq.s (5.5) and (5.7), respectively.*

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

Although the constants in Eq. (5.6) appear large, due to the use of union bounds as well as a few loose constants, in practice significantly fewer 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 $f_{\sigma,O}(\boldsymbol{\theta})$ classically. Thus, we have the following theorem.

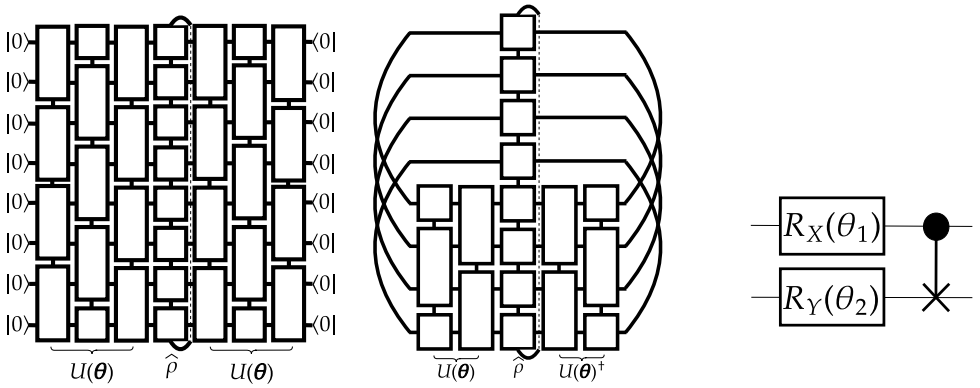
Theorem 7. *In AISO, for any quantum ansatz C with $R_C \in \mathcal{O}(\log n)$, $f_{\sigma,O}(\boldsymbol{\theta})$ can be classically evaluated with cost $\mathcal{O}(poly(n) \cdot \log M \cdot \|O\|_2^2)$ for state preparation and VQCS. The overall classical computational cost for M function evaluations is thus $\mathcal{O}(poly(n) \cdot M \log M \cdot \|O\|_2^2)$.*

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 $d \in \mathcal{O}(\log n)$, the space complexity is $\mathcal{O}(\text{poly}(n) \cdot T_1T_2)$.

Last but not least, in the state preparation problem, for any shallow shadow $\hat{\sigma}$, $f_{\hat{\sigma},O}(\boldsymbol{\theta})$ can be computed classically efficiently by contracting the tensor network given in Figure 5.2 (a). Even though the example given here is the ALA, using Theorem 7, one can easily replace it with any ansatz with $R_C \in \mathcal{O}(\log n)$. The reasoning is explained in detail in the proof of Theorem 7 in the Appendix.

Regarding VQCS, in terms of classical computational complexity, $\frac{1}{2^n}\text{tr}(|C(\boldsymbol{\theta})\rangle\langle C(\boldsymbol{\theta})|\,\hat{\sigma})$ for any shallow shadow $\hat{\sigma}$ can be computed by contracting the tensor network given in Figure 5.2 (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 state preparation. From now on, when discussing the sample complexity of VQCS, the "number of copies" will mean the number of copies of $\frac{1}{\sqrt{2^n}}|V\rangle$ consumed (equivalently, the number of applications of V).

For VQCS, from Section 3.4.2, recall that $H(\boldsymbol{\theta}) = 1 - \frac{1}{4^n}\text{tr}(|C(\boldsymbol{\theta})\rangle\langle C(\boldsymbol{\theta})|\,|V\rangle\langle V|)$, where $\frac{1}{\sqrt{2^n}}|C(\boldsymbol{\theta})\rangle$ version of $C(\boldsymbol{\theta})$. Therefore, we can use shallow shadows of $\frac{1}{\sqrt{2^n}}|V\rangle$ to estimate $H(\boldsymbol{\theta})$. Since $\frac{1}{2^n}\|\,|C(\boldsymbol{\theta})\rangle\langle C(\boldsymbol{\theta})|\,\|_2 = 1$ for all $\boldsymbol{\theta}$, the number of shadows, or



(a) State preparation (b) Circuit synthesis (c) Subcircuit used in our experiments.

Figure 5.2: (a,b) Tensor networks to compute $f_{\hat{\sigma},O}(\boldsymbol{\theta})$. The examples used here use the ALA. (a) corresponds to state preparation while (b) corresponds to VQCS. To contract (a) efficiently, we can start from the top qubit wire and contract wire by wire. One can see that, at every step, the total number of free indices the tensor will have is $\mathcal{O}(\log n)$, thus the cost of contraction is $\mathcal{O}(\text{poly}(n))$. Note that this is true for any ansatz with $R_C \in \mathcal{O}(\log n)$. A similar argument can be made for (b) when we start contracting ring by ring from the top. (c) Structure of the two-qubit subcircuits used in our simulations concerning AISO. Here, each black box is a single-qubit subcircuit while the two-qubit gate is the CNOT gate.

equivalently, the number of applications of V, is independent of n.

Also, note that the sample complexity will remain the same for any ansatz of any depth. But for deeper circuits, especially those with $R_C \in \Omega(\text{poly}(n))$ and beyond, the classical computational complexity will suffer as it scales exponentially in R_C.

Finally, we comment on the impact of the dimensionality of the lattice on AISO. The dimensionality describes how the qubits are physically arranged and connected within the device. For example, in a one dimensional lattice, the qubits are arranged linearly and we can only apply two qubit gates between neighbouring qubits. Similarly, in a two-dimensional lattice, the qubits are arranged in a grid-like manner. As mentioned previously, the sample complexity is agnsotic to the nature of the ansatz. The only factor to consider is the classical computational complexity. We have already seen that this is exponential in R_C for linear qubit arrays. However, for two dimensional lattices also, we can derive similar bounds by using PEPS simulation (cf. Section!2.4.5) within the classical

computation. Note that for ALSO, since it is only defined for ALA, which is in turn only defined for linear qubit arrays, this discussion is not relevant.

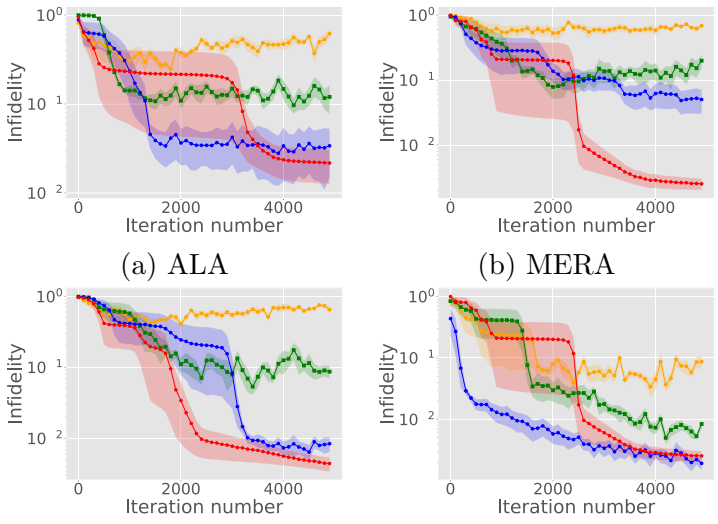
5.4 Simulation Results

Here, we elaborate on the experimental results by comparing the sample complexity of AISO and the standard VQA in 8-qubit state preparation and VQCS experiments.

The depth d of the shallow shadow ensemble (cf. Figure 5.1) 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 Figures 3.5 (a,b) and 3.6 (a,b) . Except in HEA, all two-qubit gates can be arbitrary two-qubit subcircuits. The specific ones used in our simulation are given in Figure 5.2 (c). 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 state preparation, we have used the Simultaneous Perturbation Stochastic Approximation [Spa92] (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 hand, the results of VQCS have used Powell's method [Pow64] 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 state preparation 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 5.3. 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.



(a) ALA (b) MERA

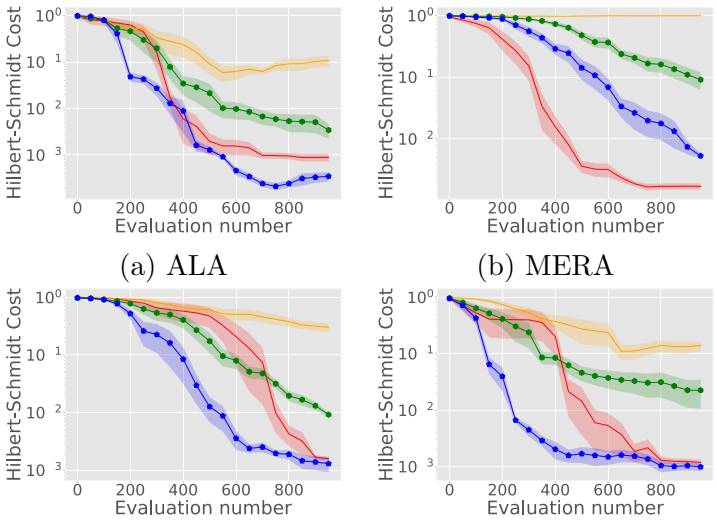
(c) HEA (d) TTN

Figure 5.3: Simulation results comparing the learning curves of AISO with the standard VQA, when solving state preparation problem. Each shaded region corresponds to 5 instances of a problem. VQA/AISO (T) consumes T copies in total throughout the optimization. Plots (a), (b), (c), (d) correspond to ALA, MERA, HEA, and TTN being used as the ansatz, respectively. The classical optimizer used is the SPSA algorithm, with 5×10^3 iterations. 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) consuming $50, 100$,and 250 copies per function evaluation respectively. We can see that AISO can closely match or outperform standard VQA by consuming orders of magnitude fewer copies in total.

In Figure 5.3, VQA (5×10^5), which utilizes 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 5.4. 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 state preparation.

In Figures 5.5 and 5.6, we present the superiority of AISO over VQA in a different



(a) ALA (b) MERA

(e) HEA (f) TTN

Figure 5.4: Simulation results comparing the learning curves of AISO with the standard VQA, when solving VQCS. Each shaded region corresponds to 5 instances of a problem. VQA/AISO (T) consumes T copies in total throughout the optimization. Plots (a), (b), (c), (d) correspond to ALA, MERA, HEA, and TTN being used as the ansatz, respectively. The classical optimizer used is Powell's method, with a total of 10^3 function evaluations allowed. In these plots, the minimum Hilbert-Schmidt cost in each interval of 100 function evaluations is plotted. 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) consuming $10, 10^2$, and 10^3 copies per function evaluation respectively. We can see that AISO can closely match or outperform standard VQA by consuming orders of magnitude fewer copies in total.

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.

5.5 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* $\mathcal{D}_2$.

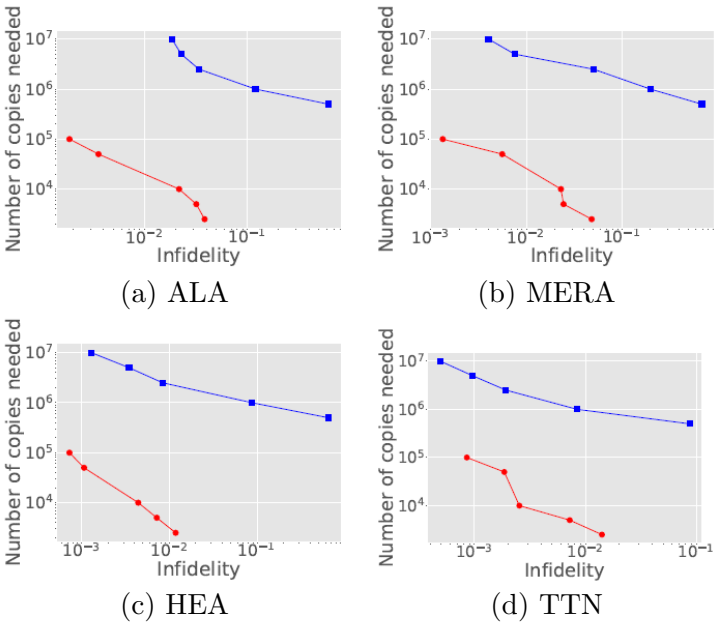


Figure 5.5: Resource needs for different infidelity objectives. All points plotted correspond to the mean of 5 instances of a state preparation problem, with the x-axis representing the average lowest infidelity achieved and the y-axis representing the total number of copies consumed to achieve it. The classical optimizers used are the same as Figure 5.3. Plots (a), (b), (c), (d) correspond to ALA, MERA, HEA, and TTN being used as the ansatz respectively. The order of magnitude savings in the number of copies when using AISO is evident.

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

$$Var_{\sigma\sim\mathcal{D}_2}\left(\|O\|^2_{\sigma,\mathcal{U}_d}\right) \leqslant 64\|O\|_2^2. \tag{5.8}$$

Using this result, we can derive a result similar to Theorem 6, with better constants.

Theorem 9. *Let $d \in \Theta(\log n)$ and σ be an n-qubit pure state sampled from a state 2-design $\mathcal{D}_2$. For any $\delta, \epsilon \in (0,1)$, $\gamma > 1/\sqrt{\delta}$, and any $M \in \mathbb{N}$, let*

$$T_1 \geqslant 2\log\left(\frac{2(\gamma^2-1)M}{\gamma^2\delta-1}\right),\ T_2 \geqslant \frac{136}{\epsilon^2}(2\gamma+1)\|O\|_2^2. \tag{5.9}$$

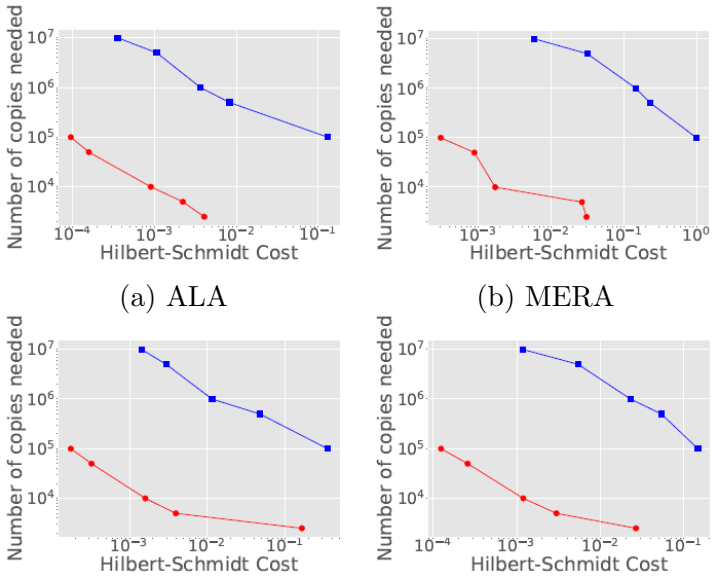


Figure 5.6: Resource needs for different Hilbert Schmidt Cost objectives. All points plotted correspond to the mean of 5 instances of VQCS, with the x-axis representing the average lowest Hilbert-Schmidt Cost achieved and the y-axis representing the total number of copies consumed to achieve it. The classical optimizers used are the same as Figure 5.4. Plots (a), (b), (c), and (d) correspond to ALA, MERA, HEA, and TTN being used as the ansatz respectively. The order of magnitude savings in the number of copies when using AISO is evident.

For any parameter vectors $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}$*, with probability at least* $1-\delta$*, we have* $|f_{\sigma,O}(\boldsymbol{\theta}^{(m)}) - \hat{f}_{\sigma,O}(\boldsymbol{\theta}^{(m)})| \leqslant \epsilon$ *for all* $1 \leqslant m \leqslant M$*, where* $f_{\sigma,O}(\boldsymbol{\theta}^{(m)})$ *and* $\hat{f}_{\sigma,O}(\boldsymbol{\theta}^{(m)})$ *are defined in Eq.s* (5.5) *and* (5.7)*, respectively.*

Hence, we see that the lower bound on T_1 in Eq. (5.32) is a constant time better than the lower bound on T_1 in Eq. (5.6). By replacing the function evaluations in Theorem 9 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.

5.6 Dealing With Barren Plateaus

Global observables may lead to barren plateaus occurring in the training landscape [Cer+21b; Liu+22], 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, [Pat+21; Mel+22; RSL22; Sko+21; Gri+23a; Gri+23b; FM22; Ver+19; Gra+19a; KS22; Zha+22a] are methods that ultimately use the quantum device only to estimate $f_{\sigma,O}(\boldsymbol{\theta})$ 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.

5.7 Proofs of All Theorems

Here, we present the proofs of Theorems 6, 7, 8 and 9. For brevity, we recall two important definitions here;

$$\hat{f}_{\sigma,O}(\boldsymbol{\theta}) \coloneqq \mu_{T_1,T_2}\left(\left\{f_{\hat{\sigma}_{U_j,u_j},O}(\boldsymbol{\theta}) \;\middle|\; 1 \leqslant j \leqslant T_1T_2\right\}\right), \tag{5.10}$$

where μ_{T_1,T_2} is the median-of-means estimator (median of T_1 means of T_2 values each), and

$$f_{\sigma,O}(\boldsymbol{\theta}) \coloneqq \mathrm{tr}(\sigma_{C(\boldsymbol{\theta})}O) = \mathrm{tr}\left(O_{C(\boldsymbol{\theta})^\dagger}\sigma\right). \tag{5.11}$$

Theorem 6. *Let σ be an n-qubit pure state sampled from a state 1-design $\mathcal{D}_1$. For any $\delta, \epsilon \in (0,1)$, $\gamma > 1/\delta$, and any $M \in \mathbb{N}$, let T_1 and T_2 satisfy Eq. (5.6). Then, for any*

parameter vectors $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(C)}$*, with probability at least* $1-\delta$*, we have* $|f_{\sigma,O}\left(\boldsymbol{\theta}^{(m)}\right) - \hat{f}_{\sigma,O}\left(\boldsymbol{\theta}^{(m)}\right)| \leqslant \epsilon$ *for all* $1 \leqslant m \leqslant M$*, where* $f_{\sigma,O}\left(\boldsymbol{\theta}^{(m)}\right)$ *and* $\hat{f}_{\sigma,O}\left(\boldsymbol{\theta}^{(m)}\right)$ *are defined in Eq.s* (5.5) *and* (5.7)*, respectively.*

Proof. As we have already established, all M function evaluations are expectations of σ with M parameterized observables $O_{C\left(\boldsymbol{\theta}^{(m)}\right)^\dagger}$, each with $\left\|O_{C\left(\boldsymbol{\theta}^{(m)}\right)^\dagger}\right\|_2 = \|O\|_2$.

From Theorem 5, we know that

$$\|O_{\mathrm{TL}}\|^2_{\mathbb{1}/2^n,\mathcal{U}_d} \leqslant 4\|O_{\mathrm{TL}}\|_2^2, \tag{5.12}$$

implying that

$$\mathbb{E}_{\sigma\sim\mathcal{D}_1}\|O_{\mathrm{TL}}\|^2_{\sigma,\mathcal{U}_d} \leqslant 4\|O_{\mathrm{TL}}\|_2^2. \tag{5.13}$$

Now, We recall the Markov inequality [Sha05] here, which says that for any non-negative random variable η, we have

$$\mathrm{Prob}(\eta \leqslant \gamma\mathbb{E}(\eta)) \geqslant 1 - \frac{1}{\gamma}. \tag{5.14}$$

So by using Markov Inequality, we have

$$\mathrm{Prob}\left[\|O_{\mathrm{TL}}\|^2_{\sigma,\mathcal{U}_d} \leqslant \gamma\mathbb{E}_{\sigma\sim\mathcal{D}_1}\|O_{\mathrm{TL}}\|^2_{\sigma,\mathcal{U}_d}\right] \geqslant 1 - 1/\gamma, \tag{5.15}$$

implying that

$$\mathrm{Prob}\left[\|O_{\mathrm{TL}}\|^2_{\sigma,\mathcal{U}_d} \leqslant 4\gamma\|O_{\mathrm{TL}}\|_2^2\right] \geqslant 1 - 1/\gamma. \tag{5.16}$$

So with probability at least $1-1/\gamma$, the state-dependent shadow norm is bounded by $4\gamma\|O_{\mathrm{TL}}\|_2^2$. So, for any $\delta', \epsilon \in (0,1)$, if we use T_1T_2 shallow shadows, where $T_1 = 2\log(2M/\delta')$, $T_2 = (136\gamma/\epsilon^2)\|O_{i\mathrm{TL}}\|_2^2$, with probability at least $(1-\delta')(1-1/\gamma)$, for all m, we will have $|f_{\sigma,O}(\boldsymbol{\theta}) - \hat{f}_{\sigma,O}(\boldsymbol{\theta})| \leqslant \epsilon$.

Set $1-\delta = (1-\delta')(1-1/\gamma)$. So, we have

$$1-\delta = (1-\delta')(1-1/\gamma), \tag{5.17}$$

which implies that

$$\delta' = 1 - \left(\frac{1-\delta}{1-1/\gamma}\right) = 1 + \frac{\gamma(\delta-1)}{\gamma-1} = \frac{\gamma\delta-1}{\gamma-1}. \tag{5.18}$$

This completes the proof. □

Theorem 7. *In AISO, for any quantum ansatz C with $R_C \in \mathcal{O}(\log n)$, $f_{\sigma,O}(\boldsymbol{\theta})$ can be classically evaluated with cost $\mathcal{O}(poly(n) \cdot \log M \cdot \|O\|_2^2)$ for state preparation and VQCS. The overall classical computational cost for M function evaluations is thus $\mathcal{O}(poly(n) \cdot M \log M \cdot \|O\|_2^2)$.*

Proof. The proof is based on the proof of classical simulation of quantum circuits in [Joz06], which we have already discussed in Section 2.4.2 (Theorem 1).

Earlier we explained a proof for this theorem using MPS theory. Another reasoning, from a tensor network perspective, is that when we start contracting the tensor network from the top qubit-wire, the maximum number of free indices the tensor can have at any point will be $\mathcal{O}(R_C)$. This means that the size of the tensor at every point will be $\mathcal{O}(2^{R_C})$. So, such contractions for all n qubit-wires can be done using cost $\mathcal{O}(n \cdot \text{poly}(2^{R_C}))$.

For state preparation of an unkwown state σ, the required classical computation is mainly computing $f_{\hat{\sigma},|0\rangle\langle 0|}(\boldsymbol{\theta})$ for some shallow shadow $\hat{\sigma}$. An example of this can be seen in Figure 5.2(a). But one can see that this is almost the same as a quantum circuit tensor network. The only difference is the fact that the core tensors of $\hat{\sigma}$ may not be valid quantum operations. But that does not affect the complexity of tensor contraction. Since the shadows are generated using an ensemble with $d = \mathcal{O}(\log n)$, the core tensors will have dimension $\mathcal{O}(\text{poly}(n))$. Combining this with the fact that there are $\mathcal{O}(\log M \cdot \|O\|_2^2)$ shadows, the complexity of classical computation is $\mathcal{O}(\text{poly}(n) \cdot \log M \cdot \|O\|_2^2)$.

For VQCS of an unknown circuit V, the expectation that we are estimating is given as

$$\frac{\left|\mathrm{tr}(C(\boldsymbol{\theta})^\dagger V)\right|^2}{4^n} = \frac{1}{4^n}\left|\langle C(\boldsymbol{\theta})|V\rangle\right|^2 = \frac{1}{4^n}\langle C(\boldsymbol{\theta})|\,|V\rangle\langle V|\,|C(\boldsymbol{\theta})\rangle\,. \tag{5.19}$$

So for any shadow $\hat{\sigma}$ of the state $\frac{1}{\sqrt{2^n}}|V\rangle$, we have

$$\langle |C(\boldsymbol{\theta})\rangle\langle C(\boldsymbol{\theta})|\rangle_{\hat{\sigma}} = \langle C(\boldsymbol{\theta})|\,\hat{\sigma}\,|C(\boldsymbol{\theta})\rangle\,. \tag{5.20}$$

Note that for any unitary matrix W, $\frac{1}{\sqrt{2^n}}|W\rangle$ is simply the vectorized and normalized version of W. If any unitary is depicted as a tensor block, we can always get a vectorized version of it by bending the output wire as shown in Figure 8.1 (d). So, given the circuit description of $C(\boldsymbol{\theta})$, we can get a tensor network depicting $|C(\boldsymbol{\theta})\rangle$ by bending the output wires in the manner shown in Figure 5.2 (b). Similar to how Figure 5.2 (a) can be contracted efficiently if we start from the top and go in a line-by-line manner, this network can also be contracted efficiently if we start from the top and contract ring by ring. That is, contract tensors along the top ring, then contract tensors along the second ring, and so on. Finally, we divide the answer by 2^n. Similar to the previous case, we see that at every instance, the number of free indices will be $\mathcal{O}(\log n)$, and hence the complexity of contracting this tensor is $\mathcal{O}(\mathrm{poly}(n))$. Hence, since we have $\mathcal{O}(\log M \cdot \|O\|_2^2)$ shadows, the complexity of a single function evaluation is $\mathcal{O}(\log M \cdot \|O\|_2^2 \cdot \mathrm{poly}(n))$.

It is easy to see that one can generalize this to arbitrary observables that can be represented as a quantum circuit-like tensor network with $R_O \in \mathcal{O}(\log n)$.

□

Theorem 8. *Let $\mathcal{D}_2$ be a state 2-design and $d \in \Theta(\log n)$. Then, for any observable O, we have*

$$Var_{\sigma\sim\mathcal{D}_2}\left(\|O\|^2_{\sigma,\mathcal{U}_d}\right) \leqslant 64\|O\|_2^2. \tag{5.21}$$

Proof. Let $\mathcal{U}_d = \{U_1, U_2, \ldots, U_{|\mathcal{U}_d|}\}$. Given any unitary 2-design $\mathcal{W}$, $\{W|\mathbf{0}\rangle \mid \forall\ W \in \mathcal{W}\}$ is a state 2-design [Wat18] (Section 2.5). Using this, we have

$$
\begin{aligned}
&\mathrm{Var}_{\sigma\sim\mathcal{D}_2}\left(\|O\|^2_{\sigma,\mathcal{U}_d}\right) \\
&\leqslant \mathbb{E}_{\sigma\sim\mathcal{D}_2}\left(\|O\|^4_{\sigma,\mathcal{U}_d}\right) && (5.22)\\
&= \mathbb{E}_{W\sim\mathcal{W}}\left(\|O\|^4_{W|\mathbf{0}\rangle\langle\mathbf{0}|W^\dagger,\mathcal{U}_d}\right) && (5.23)\\
&= \mathbb{E}_{W\sim\mathcal{W}}\left[\mathbb{E}_{U\sim\mathcal{U}_d}\sum_{u=0}^{2^n-1}\langle u|\,|\mathbf{0}\rangle\langle\mathbf{0}|_{UW}\,|u\rangle\,\mathrm{tr}\left(\hat{\sigma}_{U,u}O\right)^2\right]^2 && (5.24)\\
&= \mathbb{E}_{W\sim\mathcal{W}}\mathbb{E}_{U_1\sim\mathcal{U}_d}\mathbb{E}_{U_2\sim\mathcal{U}_d}\sum_{u_1,u_2=0}^{2^n-1}\langle u_1|\,|\mathbf{0}\rangle\langle\mathbf{0}|_{U_1W}\,|u_1\rangle\,\mathrm{tr}\left(\hat{\sigma}_{U_1,u_1}O\right)^2 \\
&\qquad\qquad \langle u_2|\,|\mathbf{0}\rangle\langle\mathbf{0}|_{U_2W}\,|u_2\rangle\,\mathrm{tr}\left(\hat{\sigma}_{U_2,u_2}O\right)^2 && (5.25)\\
&= \mathbb{E}_{U_1\sim\mathcal{U}_d}\mathbb{E}_{U_2\sim\mathcal{U}_d}\sum_{u_1,u_2=0}^{2^n-1}\mathrm{tr}\left(\hat{\sigma}_{U_1,u_1}O\right)^2\mathrm{tr}\left(\hat{\sigma}_{U_2,u_2}O\right)^2 \\
&\qquad\qquad \mathbb{E}_{W\sim\mathcal{W}}\mathrm{tr}\left[|\mathbf{0}\rangle\langle\mathbf{0}|_W\,|u_1\rangle\langle u_1|_{U_1^\dagger}\right]\mathrm{tr}\left[|\mathbf{0}\rangle\langle\mathbf{0}|_W\,|u_2\rangle\langle u_2|_{U_2^\dagger}\right]. && (5.26)
\end{aligned}
$$

Now, we shall derive a simple corollary of Lemma 5 which will be useful in the proof.

Corollary 1. *Let $\mathcal{W}$ be a unitary 2-design of operators acting on $\mathbb{C}^{2^n}$ and let $A, B, C, D \in \mathcal{L}(\mathbb{C}^{2^n})$ be pure states. Then, we have*

$$\mathbb{E}_{W\sim\mathcal{W}}tr[A_WB]\,tr[C_WD] \leqslant \frac{4}{4^n}. \tag{5.27}$$

Proof. The third and fourth terms are non-negative and can be dropped since they involve only traces of states and fidelity between states. The first two terms are upper bounded by 1 due to the same reason as well. Finally, consider the fact that $1/(4^n-1) \leqslant 2/4^n$ for any $n \geqslant 1$. □

Plugging this in Eq (5.22) gives us

$$\mathrm{Var}_{\sigma\sim\mathcal{D}_2}\left(\|O\|^2_{\sigma,\mathcal{U}_d}\right) \leqslant \frac{4}{4^n}\mathbb{E}_{U_1\sim\mathcal{U}_d}\mathbb{E}_{U_2\sim\mathcal{U}_d}\sum_{u_1,u_2=0}^{2^n-1}\mathrm{tr}\left(\hat{\sigma}_{U_1,u_1}O\right)^2\mathrm{tr}\left(\hat{\sigma}_{U_2,u_2}O\right)^2. \tag{5.28}$$

Since a state 2-design is also a state 1-design, we have

$$\mathbb{E}_{\sigma\sim\mathcal{D}_2}\mathbb{E}_{U\sim\mathcal{U}_d}\sum_{u=0}^{2^n-1}\langle u|\,U\sigma U^\dagger\,|u\rangle\,\mathrm{tr}\left(\hat{\sigma}_{U,u}O\right)^2 = \frac{1}{2^n}\mathbb{E}_{U\sim\mathcal{U}_d}\sum_{u=0}^{2^n-1}\mathrm{tr}\left(\hat{\sigma}_{U,u}O\right)^2 \tag{5.29}$$

$$= \|O\|^2_{\mathbb{1}/2^n,\mathcal{U}_d} \tag{5.30}$$

$$\leqslant 4\|O\|_2^2. \tag{5.31}$$

This completes the proof. □

Theorem 9. *Let $d \in \Theta(\log n)$ and σ be an n-qubit pure state sampled from a state 2-design $\mathcal{D}_2$. For any $\delta, \epsilon \in (0,1)$, $\gamma > 1/\sqrt{\delta}$, and any $M \in \mathbb{N}$, let*

$$T_1 \geqslant 2\log\left(\frac{2(\gamma^2-1)M}{\gamma^2\delta-1}\right),\ T_2 \geqslant \frac{136}{\epsilon^2}(2\gamma+1)\|O\|_2^2. \tag{5.32}$$

For any parameter vectors $\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(M)}$, with probability at least $1-\delta$, we have $|f_{\sigma,O}(\boldsymbol{\theta}^{(m)}) - \hat{f}_{\sigma,O}(\boldsymbol{\theta}^{(m)})| \leqslant \epsilon$ for all $1 \leqslant m \leqslant M$, where $f_{\sigma,O}(\boldsymbol{\theta}^{(m)})$ and $\hat{f}_{\sigma,O}(\boldsymbol{\theta}^{(m)})$ are defined in Eq.s (5.5) and (5.7), respectively.

Proof. Using Chebychev's Inequality, when we sample a state σ from a 2 design, we have

$$\mathrm{Prob}\Big[\left|\|O_{\mathrm{TL}}\|^2_{\sigma,\mathcal{U}_d} - \|O_{\mathrm{TL}}\|^2_{\mathbb{1}/2^n,\mathcal{U}_d}\right| \leqslant \gamma\sqrt{\mathrm{Var}_{\sigma\sim\mathcal{D}_2}\|O_{\mathrm{TL}}\|^2_{\sigma,\mathcal{U}_d}}\Big] \geqslant 1-\frac{1}{\gamma^2}, \tag{5.33}$$

implying that

$$\mathrm{Prob}\left[\|O_{\mathrm{TL}}\|^2_{\sigma,\mathcal{U}_d} - 4\|O\|_2^2 \leqslant 8\|O_{\mathrm{TL}}\|_2^2\gamma\right] \geqslant 1-\frac{1}{m^2}, \tag{5.34}$$

and hence, we have

$$\mathrm{Prob}\left[\|O_{\mathrm{TL}}\|^2_{\sigma,\mathcal{U}_d} \leqslant 4(2\gamma+1)\|O_{\mathrm{TL}}\|_2^2\right] \geqslant 1-\frac{1}{\gamma^2}. \tag{5.35}$$

So with probability at least $1-1/\gamma^2$, the state dependent shadow norm is bounded by $4(2\gamma+1)\|O_{\mathrm{TL}}\|_2^2$. So, for any $\delta', \epsilon \in (0,1)$, if we use T_1T_2 shallow shadows, where $T_1 =$

$2\log(2M/\delta')$, $T_2 = (136(2\gamma+1)/\epsilon^2)\|O_{i\mathrm{TL}}\|_2^2$, with probability $(1-\delta)(1-1/\gamma^2)$, for all m, we will have $|f_{\sigma,O}(\boldsymbol{\theta}^{(c)}) - \hat{f}_{\sigma,O}(\boldsymbol{\theta}^{(c)})\rangle_\sigma| \leqslant \epsilon$.

Set $1-\delta = (1-\delta')(1-1/\gamma^2)$. So, we have

$$1-\delta = (1-\delta')(1-1/\gamma^2), \tag{5.36}$$

implying that

$$\delta' = 1 - \left(\frac{1-\delta}{1-1/\gamma^2}\right) = 1 + \frac{\gamma^2(\delta-1)}{\gamma^2-1} = \frac{\gamma^2\delta - 1}{\gamma^2-1}. \tag{5.37}$$

This completes the proof. □

5.8 Related Works

In [SEM22], classical shadows have been employed to reduce the 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.

In [Cer+23], 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.

Chapter 6

Trainability and Classical Simulability of Learning MPS Variationally

6.1 Overview

One VQA application that features in quantum information is learning weakly entangled MPS approximations of target states [Rud+22; Ran20; Dov+22; Lin+21; RKR23] variationally using the MPS ansatz (cf. Figure 3.7). This method can be used to learn simpler circuits capable of preparing (approximating) states for which the previously known generation methods are inefficient in terms of gate count or require high connectivity within the qubit topology.

As detailed in Section 3.4.1, learning state approximations variationally can be performed using either global or local observables [Cer+21b]. Experimental results in [Dov+22] 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, using local observables can help mitigate this issue.

The usage of global observables inducing barren plateaus is expected for most ansatzes (even for an ansatz with only a single layer of rotation gates as shown in [Cer+20]). However, theoretically proving this phenomenon is a challenging task. Although this has been achieved for similar problems such as optimization over the HEA [Cer+21b] and tensor network-based optimization in quantum information [Liu+22], these results cannot be used to explain exponentially vanishing objective functions and gradients for the MPS ansatz.

This work aims to provide rigorous trainability proofs for MPS ansatzes. 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 [Cer+23], 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-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 [Cer+23]. 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 (3.2)), 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-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 [Cer+23].

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 QCNN. Note that proofs of all theorems introduced in this section can be found in Section 6.6.

6.2 Mathematical Formulation of the Ansatz

The MPS ansatz is given in Figure 3.7. Let k be the width of the subcircuit that one is using to build the MPS ansatz. Then, the MPS ansatz is defined as

$$C_t^{(n)}(\boldsymbol{\theta}) = \prod_{p=1}^{t} U_p(\boldsymbol{\theta}_p)_{(n-k-p+2,\ldots,n-p+1)},$$

where $t \leqslant n-k+1$, $\boldsymbol{\theta} = \boldsymbol{\theta}_1 \oplus \cdots \oplus \boldsymbol{\theta}_t$ with $\boldsymbol{\theta}_p = [\theta_{p1}, \theta_{p2}, \ldots, \theta_{pm}]^T$ and $U_p(\boldsymbol{\theta}_p) = \prod_{q=1}^{m} e^{-i\theta_{pq}H_{pq}}$ are k-qubit parameterized subcircuits, with $H_{pq} \in \mathbb{H}_{2^k}$.

This ansatz has a close relationship with the MPS data structure. From Section 2.4.4, 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+21; Rud+22; Dov+22; Ran20; RKR23].

Throughout this section, 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. Also, in this section, we restrict all the discussion to pure states, and so we define $\mathbb{D}_n$ to be the set of all n-qubit pure density matrices.

6.3 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.

6.3.1 Cost Concentration

Here, we present our theoretical results regarding cost concentration in learning MPS approximations variationally using $C_T^{(n)}$. Many trainability results in the literature assume one of two assumptions on the input state [Cer+23]; either they are "close" to product states [Pes+21; Cer+21b] or they are sparse [Mon+23; Lar+22; Che+23]. Our results also make such assumptions and hence use $h_1(\sigma) := \min_{V_1,\ldots,V_n \in \mathbb{U}_2} \|\sigma_{V_1 \otimes \cdots \otimes V_n}\|_1^2$ 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 characterize 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 10. *Let σ be an n-qubit pure state and $C_T^{(n)}$ be an MPS ansatz where each*

parameterized subcircuit U_i forms a unitary 2-design. Then, we have

$$Var_{\boldsymbol{\theta}}\left(f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta})\right) \leqslant \frac{h_1(\sigma)}{4^{n-k-1}}. \tag{6.1}$$

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 11. *Let σ be an n-qubit pure state, $O := 1/n \sum_{i=1}^{n} |0\rangle\langle 0|_i$, and $C_T^{(n)}$ be an MPS ansatz, where each parameterized subcircuit U_i forms a unitary 2-design. Then we have*

$$Var_{\boldsymbol{\theta}}\left(f_{\sigma,O}(\boldsymbol{\theta})\right) \geqslant \frac{1}{n(2^{2k+1}+4)} - \frac{h_2(\sigma)}{2n}. \tag{6.2}$$

So, when $h_2(\sigma) \ll 1/(2^{2k}+2)$, the lower bound scales linearly in n and exponentially only in k.

The proofs regarding trainability in works such as [Cer+20; Pes+21] also uses integration of subcircuits over the Haar measure. But in these works, the function being integrated is the partial derivative of the cost function. In our work, we first prove similar results for cost concentration and leverage the relationship that it has with barren plateaus introduced in [Arr+21] to extend the result to barren plateaus. We adopt this strategy since integrating the VQA cost function is much easier than integrating the partial derivatives, as evidenced in [Cer+20; Pes+21] where the integrations are very complex and sometimes involve heuristic approximations. In addition, we employ a method that makes use of the specific structure of the MPS ansatz to derive analytical expressions for cost concentration when different types of observables are used.

The core idea behind both cost concentration proofs is to integrate each U_t starting from U_T using standard Haar random integration methods (cf. Lemma 5 in Appendix). 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 10 and 11, 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 allowed 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 the existence of similar bounds for a wide variety of states.

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. 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.

6.3.2 From Cost Concentration to Barren Plateaus

In this section, we discuss the relationship of Theorems 10 and 11 to barren plateaus. We will use Theorem 10 to demonstrate that employing the MPS ansatz for learning state approximations leads to barren plateaus when global observables are used.

Corollary 2. *Let σ be an n-qubit pure state and $C_T^{(n)}$ be an MPS ansatz. Then, we have*

$$Var_{\boldsymbol{\theta}}\left(\partial_{\theta_{pq}} f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta})\right) \leqslant \frac{h(\sigma)}{4^{n-k-1}} \tag{6.3}$$

$\forall\ p,q$ such that $1 \leqslant p \leqslant T, 1 \leqslant q \leqslant m$, where $h(\sigma)$ is defined in Theorem 10 and $U_1, \ldots U_{p-1}, U_{p+1}, \ldots U_T$, along with one of $U_p^{(L,q)}$ or $U_p^{(R,q)}$ form unitary 2-designs and θ_{pq} is distributed uniformly.

Similarly, we extend Theorem 11 to demonstrate that using the MPS ansatz with local observables prevents barren plateaus.

Corollary 3. *Let σ be an n-qubit pure state and $C_T^{(n)}$ be an MPS ansatz. Let $O := 1/n \sum_{i=1}^{n} |0\rangle\langle 0|_i$. Then, there exist p, q with $1 \leqslant p \leqslant T, 1 \leqslant q \leqslant m$ such that*

$$Var_{\boldsymbol{\theta}} \left(\partial_{\theta_{pq}} f_{\sigma,O}(\boldsymbol{\theta}) \right) \notin \mathcal{O}\left(\frac{1}{b^n} \right), \tag{6.4}$$

where $U_1, \ldots U_{p-1}, U_{p+1}, \ldots U_T$, along with one of $U_p^{(L,q)}$ or $U_p^{(R,q)}$ form unitary 2-designs and θ_{pq} is distributed uniformly.

But keep in mind that this does not necessarily mean that the variance of all partial derivatives will escape exponential upper bounds. In fact, [ZG21] gives us an example of an MPS ansatz-local observable combination having a partial derivative whose variance is exponentially small.

6.4 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 [Cer+23] 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 [Cer+23].

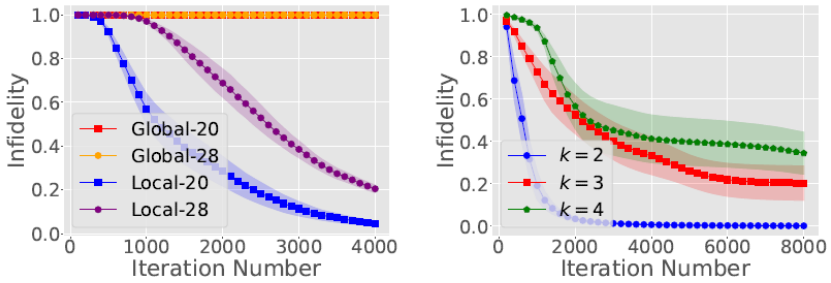
6.4.1 Effective Subspace

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. *[Cer+23] 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*

$$\mathcal{P}_{\boldsymbol{\theta},W,\mathbb{K}}(K_j) = \frac{f_{K_j,W}(\boldsymbol{\theta})^2}{\|W\|_2^2}. \tag{6.5}$$

An ansatz-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 those K_j contained in a subset $\mathbb{K}_s \subset \mathbb{K}$, that is independent of $\boldsymbol{\theta}$ and has $|\mathbb{K}_s| \in \mathcal{O}(poly(n))$.



(a) (b)

Figure 6.1: Simulation results of state approximation using MPS ansatz. In (a), we plot the learning curves of state approximation using the MPS ansatz with subcircuit width 2 for the target state $|\mathbf{0}\rangle\langle\mathbf{0}|$, optimized by SPSA with $n = 20$ and $n = 28$. The results demonstrate that global observables significantly hinder the learning process. In (b) state approximation results for the same target state using the MPS ansatz with local observables are plotted for $n = 12$ with varying subcircuit widths k. The plots indicate that increasing the subcircuit width progressively impairs learning efficiency.

In [Cer+23], 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 non-trivial 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 $\text{tr}\left(KW_{C_T(\boldsymbol{\theta})^\dagger}\right)\ \forall\ K \in \mathbb{K}_s$ and $\forall\ \boldsymbol{\theta}$ efficiently, then in many cases, $f_{\sigma,W}(\boldsymbol{\theta})$ can be classically estimated with good precision, because

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

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

$$f_{\sigma,W}(\boldsymbol{\theta}) \approx \sum_{K\in\mathbb{K}_s} \text{tr}\left(KW_{C_T(\boldsymbol{\theta})^\dagger}\right)\text{tr}(K\sigma). \tag{6.6}$$

This is the underlying principle behind designing classical simulations using effective subspaces. One can also use powerful tomography protocols such as classical shadow tomography to gain exponentially better sample efficiency.

When it comes to the classical simulation of $f_{\sigma,O}(\boldsymbol{\theta}) = 1/n\sum_i f_{\sigma,|0\rangle\langle 0|_i}(\boldsymbol{\theta})$, it is sufficient to be able to classically estimate each $f_{\sigma,|0\rangle\langle 0|_i}(\boldsymbol{\theta})$ efficiently. Among these n terms, the hardest to estimate are $f_{\sigma,|0\rangle\langle 0|_i}(\boldsymbol{\theta})$ for $i \in \{n-k+1,\ldots,n\}$, because for all other i, at least one subcircuit within $C_T^{(n)}(\boldsymbol{\theta})$ will be canceled. When t subcircuits are canceled, at least $4^{n-t}(4^t-1)$ 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. Using Lemma 12 in the Appendix, we find that for any $i,j \in \{n-k+1,\ldots,n\}$, $\mathbb{E}_{\boldsymbol{\theta}}\left(f_{\sigma,|0\rangle\langle 0|_i}(\boldsymbol{\theta})\right) = \mathbb{E}_{\boldsymbol{\theta}}\left(f_{\sigma,|0\rangle\langle 0|_j}(\boldsymbol{\theta})\right)$. 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 $\mathbb{K} = \mathbb{P}_n$.

6.4.2 C-$\mathbb{K}$ Norm

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

$$\|W\|_{\mathbb{K}} := \frac{1}{\|W\|_2}\left[\sum_{K\in\mathbb{K}} \operatorname{tr}(KW)^4\right]^{1/4} \tag{6.7}$$

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 12. *For any n-qubit quantum circuit V, 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 $j \leqslant i \leqslant k$. Then, for any product observable W, $\|W_{V^\dagger}\|_{\mathbb{K}}$ can be classically evaluated with cost $\mathcal{O}(2^{R_V})$.*

From Figure 3.7, 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 13. *For any parameterized circuit C, and an orthonormal basis $\mathbb{K}$ of $\mathbb{C}^{2^n\times 2^n}$, define*

$$\|W\|_{C,\mathbb{K}} := \int_{\boldsymbol{\theta}} \left\|W_{C(\boldsymbol{\theta})^\dagger}\right\|_{\mathbb{K}} d\boldsymbol{\theta}. \tag{6.8}$$

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

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 $\mathcal{P}_{\boldsymbol{\theta},W,\mathbb{K}}$ 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 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 6.2) strongly support

the hypothesis and hence we carry out the same experiments for $C_T^{(n)}$. 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.

6.5 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 10 and 11 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 6.2 is given in Figure 5.2 (c), but instead of R_X and R_Y, the single qubit gates are Haar random gates.

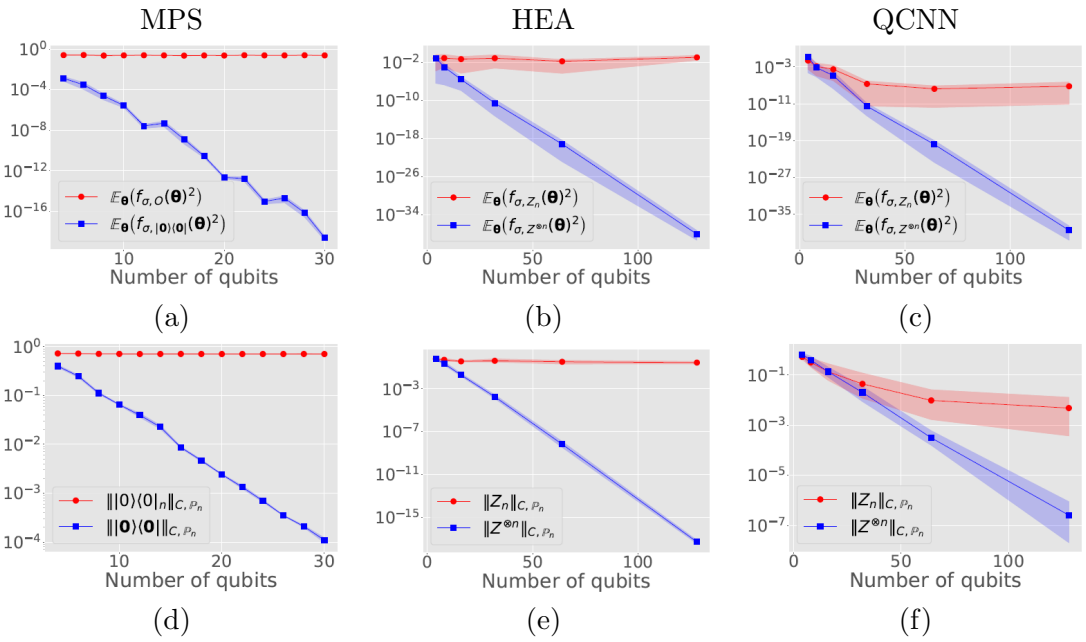
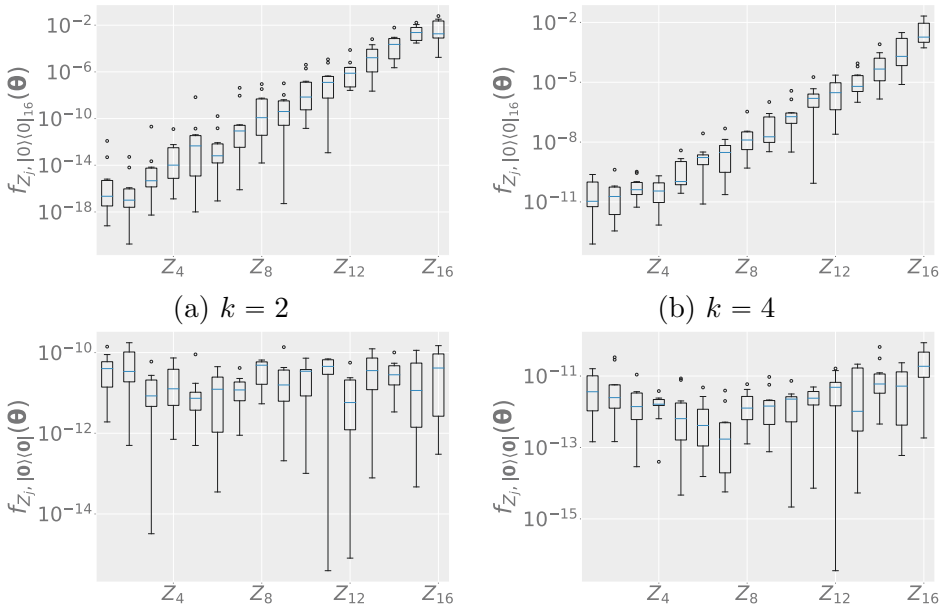


Figure 6.2: Simulation results of C-$\mathbb{K}$ norms and second moments. In all the plots, the x-axis represents the number of qubits. Plots (a-c) show the second moment of objective functions while plots (d-f) show the C-$\mathbb{P}_n$ norms. In (a), the ansatz used is the MPS ansatz with subcircuits being HEAs with depth $\lfloor \log n \rfloor$. Here, we plot the second moment of $\langle \mathbf{0} | \sigma_{C(\boldsymbol{\theta})} | \mathbf{0} \rangle$ and $\mathrm{tr}(O\sigma_{C(\boldsymbol{\theta})})$, estimated using 10 different $\boldsymbol{\theta}$ and averaged over five different input states randomly generated from HEAs of depth $\lfloor \log n \rfloor$. We can see that global observables induce cost concentration while local observables avoid it. In (b) and (c), we plot similar second moments for the shallow HEA and QCNN ansatzes respectively, which are following similar trends as well. In (e) and (f), we plot estimated C-$\mathbb{P}_n$ norms for these 2 ansatzes. From plots (b), (c), (e), and (f), we see that larger C-$\mathbb{P}_n$ norms are associated with trainable ansatz-observable combinations known to exhibit effective subspaces. Hence, in (d), we plot the C-$\mathbb{P}_n$ norms of the MPS ansatz, with subcircuit width $\lfloor \log n \rfloor$, showing a trend similar to the other ansatzes.

We start with the learning curves presented in Figure 6.1. 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 classical optimizer used here is SPSA [Spa92], 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, respectively. In (a), we have plotted results for $n = 20, 28$, with $k = 2$. We can see global observables hindering the optimization. In (b), we plot the results of similar experiments carried out for $n = 12$, 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 6.2 (a). 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 $\lfloor \log n \rfloor$. 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 10 and 11.



(a) $k = 2$ (b) $k = 4$

(c) $k = 2$ (d) $k = 4$

Figure 6.3: Boxplots of distributions $\mathcal{P}_{\boldsymbol{\theta},|0\rangle\langle 0|_{16},\mathbb{P}_{16}}$ in (a), (b) $\mathcal{P}_{\boldsymbol{\theta},|\mathbf{0}\rangle\langle\mathbf{0}|,\mathbb{P}_{16}}$ in (c), (d). with subcircuit widths $k = 2, 4$. For every Z_i on the x-axis, we plot a boxplot of the probabilities computed across 10 different values of $\boldsymbol{\theta}$. The subcircuit used in each plot is an HEA with width and depth $k = 2, 4$. In (a) and (b), we can see that higher probabilities are associated Z_i, with i close to 16, suggesting the presence of an effective subspace consisting of these terms. Also, the distribution gets flatter as we increase k. In (c) and (d), we see that the distribution remains flat for all values of k, suggesting the absence of any effective subspace.

Next, we move on to Figures 6.2 (b-f). The idea here is to show that the C-$\mathbb{K}$ norm can be used to detect the presence of effective subspace. From [Cer+23], 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 (b), (c), (e), and (f). In (b) and (c), 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 (e) and (f), 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 6.2 (d), we plot the C_T-$\mathbb{P}_n$ norms, with subcircuits having the same structure as in (a). 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 very 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 6.3. 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 $\mathcal{P}_{\boldsymbol{\theta},|0\rangle\langle 0|_{16},\mathbb{P}_{16}}(Z_i)$, 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.

6.6 Proofs of All Theorems

We start with some notation used throughout the proofs.

- For any function η defined on Haar random matrices $U_1, U_2, \ldots, U_T$, we define

$$\int_{\mathbf{U}_T} \eta(U_1, \ldots, U_T) \mathrm{d}\mathbf{U}_T \coloneqq \int_{U_1} \cdots \int_{U_T} \eta(U_1, \ldots, U_T) \mathrm{d}U_1 \ldots \mathrm{d}U_T, \tag{6.9}$$

  where $\mathbf{U}_T = \{U_1, \ldots, U_T\}$.

- For any string $i = i_1 i_2, \dots i_t$, $i_{t_1:t_2} = i_{t_1} i_{t_1+1} \dots i_{t_2}$.
- For any $t_1 \in \mathbb{N}$ such that $1 \leqslant t_1 \leqslant T-1$ and $W \in \mathbb{C}^{2^{t_2} \times 2^{t_2}}$ with $1 \leqslant t_2 \leqslant n$, define $\mu_{t_1}^{(W)} : \mathbb{C}^{2^{t_2} \times 2^{t_2}} \times \mathbb{C}^{2^{t_2} \times 2^{t_2}} \to \mathbb{C}$, where

$$\mu_{t_1}^{(W)}(X, Y) \coloneqq \int_{\mathbf{U}_{t_1}} \operatorname{tr}\left(W X_{C_{t_1}}\right) \operatorname{tr}\left(W Y_{C_{t_1}}\right) \mathrm{d}\mathbf{U}_{t_1}, \tag{6.10}$$

- $\mathbb{P}_n \coloneqq \{P_{i_1 i_2 \dots i_n} \mid i_1, i_2, \dots, i_n \in \{0,1,2,3\}\}$, where we define $P_0 \coloneqq \mathbb{1}, P_1 \coloneqq X, P_2 \coloneqq Z, P_3 \coloneqq Y$ and

Next, we introduce some definitions and lemmas that will be useful throughout our proofs.

Lemma 2. *Let $k \leqslant t \leqslant n$. For any $W, X, Y \in \mathbb{C}^{2^t \times 2^t}$, we have*

$$\mu_{t-k+1}^{(W)}(X, Y) = \mu_{t-k+1}^{(W)}\left(X_{V_1 \otimes \dots \otimes V_{t-k+1}}, Y_{V_1 \otimes \dots \otimes V_{t-k+1}}\right), \tag{6.11}$$

for any k-qubit unitary V_1 and 2 qubit unitaries $V_2 \dots V_{t-k+1}$.

Proof. This can be seen from the structure of $C_T^{(n)}$ in Figure 3.7 and Lemma 4. □

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*

$$Var_{\boldsymbol{\theta}}\left(f_{\sigma, |\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta})\right) \leqslant \frac{h_1(\sigma)}{4^{n-k-1}}. \tag{6.12}$$

Proof. Note that since

$$\operatorname{Var}_{\boldsymbol{\theta}}\left(f_{\sigma, |\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta})\right) \leqslant \mathbb{E}_{\boldsymbol{\theta}}\left(f_{\sigma, |\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta})\right)^2 \tag{6.13}$$

it is sufficient to prove $\mu_T^{(|\mathbf{0}\rangle\langle\mathbf{0}|)}(\sigma, \sigma) \leqslant \left(h_1(\sigma)/4^{n-k-1}\right)$.

Let $\sigma = \sum_{ij} \sigma_{ij} |i\rangle\langle j|$. Then, we have

$$\begin{aligned}\mu_T^{(|\mathbf{0}\rangle\langle\mathbf{0}|)}(\sigma,\sigma) &= \sum_{pqrs} \sigma_{pq}\sigma_{rs}\mu_T^{(|\mathbf{0}\rangle\langle\mathbf{0}|)}(|p\rangle\langle q|, |r\rangle\langle s|) \\ &\leqslant \sum_{pqrs} |\sigma_{pq}||\sigma_{rs}||\mu_T^{(|\mathbf{0}\rangle\langle\mathbf{0}|)}(|p\rangle\langle q|, |r\rangle\langle s|)|. \end{aligned} \tag{6.14}$$

Our next step is to prove that $|\mu_T^{(|\mathbf{0}\rangle\langle\mathbf{0}|)}(|p\rangle\langle q|, |r\rangle\langle s|)|$ is upper bounded by $1/4^{n-k-1}$. Let $p = p_1 p_2 \dots p_n$ be the binary expansion of p (similar definitions for q, r and s). Then, we have

$$\begin{aligned}
&\mu_T^{(|\mathbf{0}\rangle\langle\mathbf{0}|)}(|p\rangle\langle q|, |r\rangle\langle s|) \\
=&\int_{\mathbf{U}_T} \langle\mathbf{0}| \left(|p\rangle\langle q|\right)_{C_T} |\mathbf{0}\rangle\langle\mathbf{0}| \left(|r\rangle\langle s|\right)_{C_T} |\mathbf{0}\rangle \, \mathrm{d}\mathbf{U}_T && (6.15)\\
=&\int_{\mathbf{U}_T} \langle\mathbf{0}| \left(|p_{1:k}\rangle\langle q_{1:k}|_{U_T} \otimes |p_{k+1:n}\rangle\langle q_{k+1:n}|\right)_{C_{T-1}} |\mathbf{0}\rangle \\
&\qquad \langle\mathbf{0}| \left(|r_{1:k}\rangle\langle s_{1:k}|_{U_T} \otimes |r_{k+1:n}\rangle\langle s_{k+1:n}|\right)_{C_{T-1}} |\mathbf{0}\rangle \, \mathrm{d}\mathbf{U}_T && (6.16)\\
=&\frac{1}{2^{2k}} \int_{\mathbf{U}_T} \sum_{\substack{i_{1:k}\\ j_{1:k}}} \langle\mathbf{0}| \left(P_{i_{1:k}} \otimes |p_{k+1:n}\rangle\langle q_{k+1:n}|\right)_{C_{T-1}} |\mathbf{0}\rangle \operatorname{tr}\left(P_{i_{1:k}} |p_{1:k}\rangle\langle q_{1:k}|_{U_T}\right) \\
&\qquad \langle\mathbf{0}| \left(P_{j_{1:k}} \otimes |r_{k+1:n}\rangle\langle s_{k+1:n}|\right)_{C_{T-1}} |\mathbf{0}\rangle \operatorname{tr}\left(P_{j_{1:k}} |r_{1:k}\rangle\langle s_{1:k}|_{U_T}\right) \mathrm{d}\mathbf{U}_T && (6.17)\\
=&\frac{1}{2^k} \int_{\mathbf{U}_{T-1}} \sum_{\substack{i_{1:k}\\ j_{1:k}}} \langle\mathbf{0}| \left(P_{i_{1:k}} \otimes |p_{k+1:n}\rangle\langle q_{k+1:n}|\right)_{C_{T-1}} |\mathbf{0}\rangle\langle\mathbf{0}| \left(P_{j_{1:k}} \otimes |r_{k+1:n}\rangle\langle s_{k+1:n}|\right)_{C_{T-1}} |\mathbf{0}\rangle \\
&\qquad \times \frac{1}{2^k} \int_{U_T} \operatorname{tr}\left(P_{i_{1:k}} |p_{1:k}\rangle\langle q_{1:k}|_{U_T}\right) \operatorname{tr}\left(P_{j_{1:k}} |r_{1:k}\rangle\langle s_{1:k}|_{U_T}\right) \mathrm{d}\mathbf{U}_T && (6.18)\\
=&\frac{1}{2^k} \sum_{\substack{i_{1:k}\\ j_{1:k}}} \mu_{T-1}^{(|\mathbf{0}\rangle\langle\mathbf{0}|)}(P_{i_{1:k}} \otimes |p_{k+1:n}\rangle\langle q_{k+1:n}|, P_{j_{1:k}} \otimes |r_{k+1:n}\rangle\langle s_{k+1:n}|) \\
&\qquad \times \frac{1}{2^k} \int_{U_T} \operatorname{tr}\left(P_{i_{1:k}} |p_{1:k}\rangle\langle q_{1:k}|_{U_T}\right) \operatorname{tr}\left(P_{j_{1:k}} |r_{1:k}\rangle\langle s_{1:k}|_{U_T}\right) \mathrm{d}U_T. && (6.19)
\end{aligned}$$

Our goal is to integrate over U_T and get expressions that look like $\mu_T^{(|\mathbf{0}\rangle\langle\mathbf{0}|)}(|p\rangle\langle q|, |r\rangle\langle s|)$, but with different bit strings, defined over an $n-1$ qubit system, and as an integral of

$T-1$ Haar random gates.

We shall use Lemma 5 to integrate over U_T in the previous equation. Notice that when $b_1 \notin \{0,2\}$, $\langle \mathbf{0}| \left(P_{b_{1:k}} \otimes |p_{k+1:n}\rangle\langle q_{k+1:n}|\right)_{C_{T-1}} |\mathbf{0}\rangle = 0$. Similarly, using Lemma 5, we can see that when $P_{i_{1:k}}$ and $P_{j_{1:k}}$ are not equal, $\int\limits_{U_T} \operatorname{tr}\left(P_{i_{1:k}} |p_{1:k}\rangle\langle q_{1:k}|_{U_T}\right) \operatorname{tr}\left(P_{j_{1:k}} |r_{1:k}\rangle\langle s_{1:k}|_{U_T}\right) = 0$.

When $i_{1:k} = j_{1:k} = 0 \ldots 0$, we have

$$\frac{1}{2^k} \int\limits_{U_T} \operatorname{tr}\left(P_{i_{1:k}} |p_{1:k}\rangle\langle q_{1:k}|_{U_T}\right) \operatorname{tr}\left(P_{j_{1:k}} |r_{1:k}\rangle\langle s_{1:k}|_{U_T}\right) \mathrm{d}U_T = \frac{\delta_{\mathrm{tr}}^{(1)}}{2^k} \tag{6.20}$$

and when $i_{1:k} = j_{1:k} \neq 0 \ldots 0$ with $i_1 \in \{0,2\}$, we have

$$\frac{1}{2^k} \int\limits_{U_T} \operatorname{tr}\left(P_{i_{1:k}} |p_{1:k}\rangle\langle q_{1:k}|_{U_T}\right) \operatorname{tr}\left(P_{j_{1:k}} |r_{1:k}\rangle\langle s_{1:k}|_{U_T}\right) \mathrm{d}U_T = \frac{2^k \delta_{\mathrm{ip}}^{(1)} - \delta_{\mathrm{tr}}^{(1)}}{2^k(2^{2k}-1)} = \tau. \tag{6.21}$$

where $\delta_{\mathrm{ip}}^{(1)} = \delta_{p_{1:k},r_{1:k}} \delta_{q_{1:k},s_{1:k}}, \delta_{\mathrm{tr}}^{(1)} = \delta_{p_{1:k},q_{1:k}} \delta_{r_{1:k},s_{1:k}}$. Hence, we have

$$\begin{aligned}
&\mu_T^{(|\mathbf{0}\rangle\langle\mathbf{0}|)}(|p\rangle\langle q|, |r\rangle\langle s|) \\
=&\frac{\delta_{\mathrm{tr}}^{(1)}}{2^{2k}} \mu_{T-1}^{(|0\rangle\langle 0|^{\otimes n-1})}(\mathbb{1}_{2^{k-1}} \otimes |p_{k+1:n}\rangle\langle q_{k+1:n}|, \mathbb{1}_{2^{k-1}} \otimes |r_{k+1:n}\rangle\langle s_{k+1:n}|) \\
&+ \frac{\tau}{2^k} \sum_{\substack{P \in \mathbb{P}_{k-1} \\ P \neq \mathbb{1}_{2^{k-1}}}} \mu_{T-1}^{(|0\rangle\langle 0|^{\otimes n-1})}(P \otimes |p_{k+1:n}\rangle\langle q_{k+1:n}|, P \otimes |r_{k+1:n}\rangle\langle s_{k+1:n}|).
\end{aligned} \tag{6.22}$$

Now, using Lemma 2, we have

$$\begin{aligned}
&\mu_T^{(|\mathbf{0}\rangle\langle\mathbf{0}|)}(|p\rangle\langle q|, |r\rangle\langle s|) \\
=&\frac{\delta_{\mathrm{tr}}^{(1)}}{2^{2k}} \mu_{T-1}^{(|0\rangle\langle 0|^{\otimes n-1})}\left(\mathbb{1}_{2^{k-1}} \otimes |p_{k+1:n}\rangle\langle q_{k+1:n}|, \mathbb{1}_{2^{k-1}} \otimes |r_{k+1:n}\rangle\langle s_{k+1:n}|\right) \\
&+ \frac{(2^{2k-2}-1)\tau}{2^k} \mu_{T-1}^{(|0\rangle\langle 0|^{\otimes n-1})}\left(Z^{\otimes k-1} \otimes |p_{k+1:n}\rangle\langle q_{k+1:n}|, Z^{\otimes k-1} \otimes |r_{k+1:n}\rangle\langle s_{k+1:n}|\right).
\end{aligned} \tag{6.23}$$

This is because there always exists a unitary that will map any non-identity Pauli to

any other non-identity Pauli. Let $\mathbb{1}_{2^{k-1}} = \sum_{i=0}^{2^{k-1}-1} |i\rangle\langle i|$ and $Z^{\otimes(k-1)} = \sum_{i}^{2^{k-1}-1} \lambda_i |i\rangle\langle i|$ be spectral decompositions. Then, we have

$$\begin{aligned}
&\mu_T^{(|\mathbf{0}\rangle\langle\mathbf{0}|)}(|p\rangle\langle q|, |r\rangle\langle s|) \\
=&\frac{\delta_{\text{tr}}^{(1)}}{2^{2k}} \sum_{i,j=0}^{2^{k-1}-1} \mu_{T-1}^{(|0\rangle\langle 0|^{\otimes n-1})}(|i\rangle\langle i| \otimes |p_{k+1:n}\rangle\langle q_{k+1:n}|, |j\rangle\langle j| \otimes |r_{k+1:n}\rangle\langle s_{k+1:n}|) \\
&+ \frac{(2^{2k-2}-1)\tau}{2^k} \sum_{i,j=0}^{2^{k-1}-1} \lambda_i\lambda_j \mu_{T-1}^{(|0\rangle\langle 0|^{\otimes n-1})}(|i\rangle\langle i| \otimes |p_{k+1:n}\rangle\langle q_{k+1:n}|, \\
&\qquad\qquad |j\rangle\langle j| \otimes |r_{k+1:n}\rangle\langle s_{k+1:n}|). \qquad (6.24)
\end{aligned}$$

Now, using Lemma 2, we also have that

$$\begin{aligned}
&\mu_{T-1}^{(|0\rangle\langle 0|^{\otimes n-1})}(|i\rangle\langle i| \otimes |p_{k+1:n}\rangle\langle q_{k+1:n}|, |i\rangle\langle i| \otimes |r_{k+1:n}\rangle\langle s_{k+1:n}|) \\
=&\mu_{T-1}^{(|0\rangle\langle 0|^{\otimes n-1})}(|0\rangle\langle 0|^{\otimes k-1} \otimes |p_{k+1:n}\rangle\langle q_{k+1:n}|, |0\rangle\langle 0|^{\otimes k-1} \otimes |r_{k+1:n}\rangle\langle s_{k+1:n}|) \qquad (6.25)
\end{aligned}$$

and

$$\begin{aligned}
&\mu_{T-1}^{(|0\rangle\langle 0|^{\otimes n-1})}(|i\rangle\langle i| \otimes |p_{k+1:n}\rangle\langle q_{k+1:n}|, |j\rangle\langle j| \otimes |r_{k+1:n}\rangle\langle s_{k+1:n}|) \\
=&\mu_{T-1}^{(|0\rangle\langle 0|^{\otimes n-1})}(|0\rangle\langle 0|^{\otimes k-1} \otimes |p_{k+1:n}\rangle\langle q_{k+1:n}|, |1\rangle\langle 1|^{\otimes k-1} \otimes |r_{k+1:n}\rangle\langle s_{k+1:n}|). \qquad (6.26)
\end{aligned}$$

The reason for the second equation is that we can always find a $k-1$-qubit unitary V such that $V|i\rangle\langle i|V^\dagger = |0\rangle\langle 0|^{\otimes k-1}$ and $V|j\rangle\langle j|V^\dagger = |1\rangle\langle 1|^{\otimes k-1}$. A similar explanation for the first equation as well.

Now, define ${\Delta_t}^{(=)}$ and ${\Delta_t}^{(\neq)}$ as

$$\begin{aligned}
{\Delta_t}^{(=)} &= \mu_{T-t}^{(|0\rangle\langle 0|^{\otimes n-t})}(|0\rangle\langle 0|^{\otimes k-1} \otimes |p_{k+t:n}\rangle\langle q_{k+t:n}|, |0\rangle\langle 0|^{\otimes k-1} \otimes |r_{k+t:n}\rangle\langle s_{k+t:n}|) \qquad (6.27) \\
{\Delta_t}^{(\neq)} &= \mu_{T-t}^{(|0\rangle\langle 0|^{\otimes n-t})}(|0\rangle\langle 0|^{\otimes k-1} \otimes |p_{k+t:n}\rangle\langle q_{k+t:n}|, |1\rangle\langle 1|^{\otimes k-1} \otimes |r_{k+t:n}\rangle\langle s_{k+t:n}|). \qquad (6.28)
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\mu_{T-1}^{(|0\rangle\langle 0|^{\otimes n-1})}(\mathbb{1}_{2^{k-1}} \otimes |p_{k+1:n}\rangle\langle q_{k+1:n}|, \mathbb{1}_{2^{k-1}} \otimes |r_{k+1:n}\rangle\langle s_{k+1:n}|) \\
&= 2^{k-1}\Delta_1^{(=)} + 2^{k-1}(2^{k-1}-1)\Delta_1^{(\neq)}.
\end{aligned} \tag{6.29}$$

Similarly, we have

$$\begin{aligned}
&\mu_{T-1}^{(|0\rangle\langle 0|^{\otimes n-1})}\left(Z^{\otimes k-1} \otimes |p_{k+1:n}\rangle\langle q_{k+1:n}|, Z^{\otimes k-1} \otimes |r_{k+1:n}\rangle\langle s_{k+1:n}|\right) \\
&= 2^{k-1}\Delta_1^{(=)} + \sum_{i,j=0, i\neq j}^{2^{k-1}-1} \lambda_i\lambda_j\Delta_1^{(\neq)} \\
&= 2^{k-1}\Delta_1^{(=)} - 2^{k-1}\Delta_1^{(\neq)}.
\end{aligned} \tag{6.30}$$

The reason for the last equality is as follows. Notice that the set $\{\lambda_i \mid i = 0, \ldots, 2^{k-1}\}$ has 2^{k-2} 1s and 2^{k-2} -1s. So, using Lemma 11 we see that $\sum_{i\neq j} \lambda_i\lambda_j = -2^{k-1}$.

Hence, we have

$$\mu_T^{(|\mathbf{0}\rangle\langle\mathbf{0}|)}(|p\rangle\langle q|, |r\rangle\langle s|) = \alpha_1\Delta_1^{(=)} + \beta_1\Delta_1^{(\neq)}, \tag{6.31}$$

where

$$\alpha_1 = \frac{\delta_{\text{tr}}^{(1)}}{2^{k+1}} + \frac{\left(2^{2k-2}-1\right)\left(2^k\delta_{\text{ip}}^{(1)} - \delta_{\text{tr}}^{(1)}\right)}{2^{k+1}\left(2^{2k}-1\right)}, \tag{6.32}$$

$$\beta_1 = \frac{\delta_{\text{tr}}^{(1)}\left(2^{k-1}-1\right)}{2^{k+1}} - \frac{\left(2^{2k-2}-1\right)\left(2^k\delta_{\text{ip}}^{(1)} - \delta_{\text{tr}}^{(1)}\right)}{2^{k+1}\left(2^{2k}-1\right)}. \tag{6.33}$$

So we have achieved the goal of reducing the Eq (6.15) to a similar integral of $n-1$ qubit systems and $T-1$ Haar random gates.

Now we can integrate $\Delta_1^{(=)}$ and $\Delta_1^{(\neq)}$ in the same way. For that, we first define

$$\alpha_t^{(=)} = \frac{\delta_{\text{tr}}^{(1)}}{2^{k+1}} + \frac{\left(2^{2k-2}-1\right)\left(2^k\delta_{\text{ip}}^{(t)} - \delta_{\text{tr}}^{(t)}\right)}{2^{k+1}\left(2^{2k}-1\right)}, \tag{6.34}$$

$$\beta_t^{(=)} = \frac{\delta_{\text{tr}}^{(1)}\left(2^{k-1}-1\right)}{2^{k+1}} - \frac{\left(2^{2k-2}-1\right)\left(2^k\delta_{\text{ip}}^{(t)} - \delta_{\text{tr}}^{(t)}\right)}{2^{k+1}(2^{2k}-1)} \tag{6.35}$$

$$\alpha_t^{(\neq)} = \frac{\delta_{\text{tr}}^{(1)}}{2^{k+1}} + \frac{\left(2^{2k-2}-1\right)\left(-\delta_{\text{tr}}^{(t)}\right)}{2^{k+1}\left(2^{2k}-1\right)} \tag{6.36}$$

$$\beta_t^{(\neq)} = \frac{\delta_{\text{tr}}^{(1)}\left(2^{k-1}-1\right)}{2^{k+1}} - \frac{\left(2^{2k-2}-1\right)\left(-\delta_{\text{tr}}^{(t)}\right)}{2^{k+1}\left(2^{2k}-1\right)} \tag{6.37}$$

for $2 \leqslant t \leqslant T$ and

$$\delta_{\text{ip}}^{(t)} = \delta_{p_{k+t-1},r_{k+t-1}}\delta_{q_{k+t-1},s_{k+t-1}} \tag{6.38}$$

$$\delta_{\text{tr}}^{(t)} = \delta_{p_{k+t-1},q_{k+t-1}}\delta_{r_{k+t-1},s_{k+t-1}}. \tag{6.39}$$

Integrating $\Delta_1^{(=)}$ will result in $\alpha_2^{(=)}\Delta_2^{(=)} + \beta_2^{(=)}\Delta_2^{(\neq)}$ and integrating $\Delta_1^{(\neq)}$ will result in $\alpha_2^{(\neq)}\Delta_2^{(=)} + \beta_2^{(\neq)}\Delta_2^{(\neq)}$.

Assume that after integration over unitaries $U_T, \ldots U_{T-t+1}$, we get $\gamma^{(=)}\Delta_t^{(=)} + \gamma^{(\neq)}\Delta_t^{(\neq)}$. Now, if we integrate over U_{T-t}, the coefficients of $\Delta_{t+1}^{(=)}$ and $\Delta_{t+1}^{(\neq)}$ will be $\alpha_{t+1}^{(=)}\gamma^{(=)} + \alpha_{t+1}^{(\neq)}\gamma^{(\neq)}$ and $\beta_{t+1}^{(=)}\gamma^{(=)} + \beta_{t+1}^{(\neq)}\gamma^{(\neq)}$ respectively. Therefore, we have

$$\mu_T^{(|\mathbf{0}\rangle\langle\mathbf{0}|)}(|p\rangle\langle q|, |r\rangle\langle s|) = M_T M_{T-1} \ldots M_1 \tag{6.40}$$

where

$$M_1 = \begin{bmatrix} \alpha_1 \\ \beta_1 \end{bmatrix}, \; M_T = \begin{bmatrix} \alpha_t & \beta_t \end{bmatrix}, \; M_t = \begin{bmatrix} \alpha_t & \alpha_t' \\ \beta_t & \beta_t' \end{bmatrix}. \tag{6.41}$$

$$\alpha_t = \Delta_{T-1}^{(=)} = \frac{\delta_{\text{tr}}^{(T)} + \delta_{\text{ip}}^{(T)}}{2^k+1}, \quad \beta_t = \Delta_{T-1}^{(\neq)} = \frac{\delta_{\text{tr}}^{(T)}}{2^k+1} \tag{6.42}$$

and $2 \leqslant t \leqslant T-1$. α_t and β_t can be evaluated directly using Lemma 5. Since each M_t is a 2×2 matrix, its eigenvalues can be computed analytically. We use SymPy for this computation and the eigenvalues are

$$\frac{\delta_{\mathrm{tr}}^{(t)}}{4}, \quad \frac{\delta_{\mathrm{tr}}^{(t)}\left(4^k-4\right)}{8\left(4^k-1\right)}. \tag{6.43}$$

We can see that the absolute values of all these eigenvalues are upper bounded by $1/4$.

Also, we have

$$\|M_T\|_2 \|M_1\|_2 \leqslant \sqrt{\frac{10 \cdot 2^{6k}+140 \cdot 2^{4k}-240 \cdot 2^{3k}-220 \cdot 2^{2k}+240 \cdot 2^k+160}{2^{2k+6}\left(2^k+1\right)^2\left(2^{2k}-1\right)^2}} \leqslant 1 \tag{6.44}$$

(computed using SymPy). Combining Eq (6.44) with Eq (6.14) gives us

$$\mu_T^{(|\mathbf{0}\rangle\langle\mathbf{0}|)}(\sigma,\sigma) \leqslant \sum_{pqrs}|\sigma_{pq}||\sigma_{rs}|\frac{1}{4^{n-k-1}} = \frac{\|\sigma\|_1^2}{4^{n-k-1}}. \tag{6.45}$$

Combining Eq (6.45) with Lemma 2 completes the proof. □

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*

$$Var_{\boldsymbol{\theta}}\left(f_{\sigma,O}(\boldsymbol{\theta})\right) \geqslant \frac{1}{n(2^{2k+1}+4)} - \frac{h_2(\sigma)}{2n}. \tag{6.46}$$

Proof. First, notice that for any $i \in \{1,\ldots,n\}$,

$$f_{\sigma,|0\rangle\langle 0|_i}(\boldsymbol{\theta}) = \frac{1}{2} + \frac{f_{\sigma,Z_i}(\boldsymbol{\theta})}{2}, \tag{6.47}$$

implying that

$$\mathrm{Var}_{\boldsymbol{\theta}}\left(f_{\sigma,|0\rangle\langle 0|_i}(\boldsymbol{\theta})\right) = \frac{1}{4}\mathrm{Var}_{\boldsymbol{\theta}}\left(f_{\sigma,Z_i}(\boldsymbol{\theta})\right). \tag{6.48}$$

So, we have

$$\begin{aligned}
\mathrm{Var}_{\boldsymbol{\theta}}\left(f_{\sigma,O}(\boldsymbol{\theta})\right) =& \frac{1}{n^2}\mathrm{Var}_{\boldsymbol{\theta}}\left(\sum_{i=1}^{n} f_{\sigma,|0\rangle\langle 0|_i}(\boldsymbol{\theta})\right) && (6.49)\\
=& \frac{1}{4n^2}\mathrm{Var}_{\boldsymbol{\theta}}\left(\sum_{i=1}^{n} f_{\sigma,Z_i}(\boldsymbol{\theta})\right) && (6.50)\\
=& \frac{1}{4n^2}\mathbb{E}_{\boldsymbol{\theta}}\left(\sum_{i=1}^{n} f_{\sigma,Z_i}(\boldsymbol{\theta})\right)^2 - \frac{1}{4n^2}\left(\mathbb{E}_{\boldsymbol{\theta}}\sum_{i=1}^{n} f_{\sigma,Z_i}(\boldsymbol{\theta})\right)^2 && (6.51)\\
=& \frac{1}{4n^2}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta}}\left(f_{\sigma,Z_i}(\boldsymbol{\theta})\right)^2 + \frac{2}{4n^2}\sum_{\substack{i,j=1\\ i>j}}^{n}\mathbb{E}_{\boldsymbol{\theta}}\left(f_{\sigma,Z_i}(\boldsymbol{\theta})f_{\sigma,Z_j}(\boldsymbol{\theta})\right)\\
& - \frac{1}{4n^2}\left(\mathbb{E}_{\boldsymbol{\theta}}\sum_{i=1}^{n} f_{\sigma,Z_i}(\boldsymbol{\theta})\right)^2 && (6.52)
\end{aligned}$$

Hence, from Lemmas 12 and 13, we have

$$\mathrm{Var}_{\boldsymbol{\theta}}\left(f_{\sigma,O}(\boldsymbol{\theta})\right) = \frac{1}{4n^2}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta}}\left(f_{\sigma,Z_i}(\boldsymbol{\theta})\right)^2. \tag{6.53}$$

Similar to the beginning of the proof of Theorem 10, for any pure product state ρ, using Lemmas 2, we can see that for any i

$$\mathbb{E}_{\boldsymbol{\theta}}\left(f_{\rho,Z_i}(\boldsymbol{\theta})\right)^2 = \mu_T^{(Z_i)}(\rho,\rho) = \mu_T^{(Z_i)}(|\mathbf{0}\rangle\langle\mathbf{0}|, |\mathbf{0}\rangle\langle\mathbf{0}|). \tag{6.54}$$

Now, we shall derive a lower bound for Eq (6.54) $\forall\ i$. We only derive this for $i \leqslant k$, since we will see that the same lower bound works for any $i > k$ as well. Hence, assume $i \leqslant k$. First, we compute

$$\mu_T^{(Z_i)}(|p\rangle\langle p|, |q\rangle\langle q|) \tag{6.55}$$

This will be used later on to compute $\mu_T^{(Z_i)}(|\mathbf{0}\rangle\langle\mathbf{0}|, |\mathbf{0}\rangle\langle\mathbf{0}|)$.

Our goal is to integrate over U_T and get expressions that look like Eq (6.55), but with different bit strings, defined over an $n-1$ qubit system, and as an integral of $T-1$ Haar random gates.

Following the proof of Theorem 10, from Eq (6.15), we have

$$\begin{aligned}
&\mu_T^{(Z_i)}\left(|p\rangle\langle p|, |q\rangle\langle q|\right) \\
=&\frac{1}{2^k}\sum_{i_{1:k}j_{1:k}} \mu_{T-1}^{(Z_i)}\left(P_{i_{1:k}}\otimes|p_{k+1:n}\rangle\langle p_{k+1:n}|, P_{j_{1:k}}\otimes|q_{k+1:n}\rangle\langle q_{k+1:n}|\right) \\
&\times\frac{1}{2^k}\int_{U_T}\mathrm{tr}\left(P_{i_{1:k}}|p_{1:k}\rangle\langle p_{1:k}|_{U_T}\right)\mathrm{tr}\left(P_{j_{1:k}}|q_{1:k}\rangle\langle q_{1:k}|_{U_T}\right)\mathrm{d}U_T. \qquad (6.56)
\end{aligned}$$

We see that whenever $i_1 \neq 0$, the integral drops to 0. Similar to the proof of Theorem 10, we see that when $i_{2:k} \neq j_{2:k}$, the integral drops to 0. Hence, when $P_{01_2\ldots i_k} = P_{0j_2\ldots j_k} \neq \mathbb{1}_{2^k}$, we can directly use Eq (6.15) to get

$$\frac{1}{2^k}\int_{U_T}\mathrm{tr}\left(P_{0i_{2:k}}|p_{1:k}\rangle\langle p_{1:k}|_{U_T}\right)\mathrm{tr}\left(P_{0i_{2:k}}|q_{1:k}\rangle\langle q_{1:k}|_{U_T}\right)\mathrm{d}U_T = \frac{2^k\delta_{q_{1:k}q_{1:k}}-1}{2^k(2^{2k}-1)}. \qquad (6.57)$$

Similarly, when $P_{01_2\ldots i_k} = P_{0j_2\ldots j_k} = \mathbb{1}_{2^k}$, we have

$$\frac{1}{2^k}\int_{U_T}\mathrm{tr}\left(P_{0i_{2:k}}|p_{1:k}\rangle\langle p_{1:k}|_{U_T}\right)\mathrm{tr}\left(P_{0j_{2:k}}|q_{1:k}\rangle\langle q_{1:k}|_{U_T}\right)\mathrm{d}U_T = \frac{1}{2^k}. \qquad (6.58)$$

Given Eqs (6.58), (6.57), and (6.56), we have

$$\begin{aligned}
&\mu_T^{(Z_i)}\left(|p\rangle\langle p|, |q\rangle\langle q|\right) \\
=&\frac{1}{2^{2k-2}}\mu_{T-1}^{(Z_i)}\left(\mathbb{1}_{2^{k-1}}\otimes|p_{k+1:n}\rangle\langle p_{k+1:n}|, \mathbb{1}_{2^{k-1}}\otimes|q_{k+1:n}\rangle\langle q_{k+1:n}|\right) \\
&+\frac{(2^{k+1}\delta_{p_{1:k}q_{1:k}}-2)(2^{2k-2}-1)}{2^{2k-1}(2^{2k}-1)}\mu_{T-1}^{(Z_i)}\Big(Z^{\otimes(k-1)}\otimes|p_{k+1:n}\rangle\langle p_{k+1:n}|, \\
&\qquad\qquad Z^{\otimes k-1}\otimes|q_{k+1:n}\rangle\langle q_{k+1:n}|\Big). \qquad (6.59)
\end{aligned}$$

Let $\mathbb{1}_{2^{k-1}} = \sum_{i=0}^{2^{k-1}-1}|i\rangle\langle i|$ and $Z^{\otimes(k-1)} = \sum_{i}^{2^{k-1}-1}\lambda_i|i\rangle\langle i|$ be spectral decompositions. Then,

we have

$$\begin{aligned}
&\mu_T^{(Z_i)}\left(|p\rangle\langle p|, |q\rangle\langle q|\right) \\
=&\frac{1}{2^{2k-2}}\sum_{i,j=0}^{2^{k-1}-1}\mu_{T-1}^{(Z_i)}(|i\rangle\langle i|\otimes|p_{k+1:n}\rangle\langle p_{k+1:n}|, |j\rangle\langle j|\otimes|q_{k+1:n}\rangle\langle q_{k+1:n}|) \\
&+\frac{(2^{k+1}\delta_{p_{1:k}q_{1:k}}-2)(2^{2k-2}-1)}{2^{2k-1}(2^{2k}-1)}\sum_{i,j=0}^{2^{k-1}-1}\lambda_i\lambda_j\mu_{T-1}^{(Z_i)}(|i\rangle\langle i|\otimes|p_{k+1:n}\rangle\langle p_{k+1:n}|, \\
&\qquad\qquad |j\rangle\langle j|\otimes|q_{k+1:n}\rangle\langle q_{k+1:n}|). \qquad (6.60)
\end{aligned}$$

Next, we define

$$\Delta_t^{(=)} = \int_{\mathbf{U}_{T-t}} \mu_{T-t}^{(Z_i)}(|0\rangle\langle 0|^{\otimes k-1}\otimes|p_{k+t:n}\rangle\langle p_{k+t:n}|, |0\rangle\langle 0|^{\otimes k-1}\otimes|q_{k+t:n}\rangle\langle q_{k+t:n}|), \quad (6.61)$$

$$\Delta_t^{(\neq)} = \int_{\mathbf{U}_{T-t}} \mu_{T-t}^{(Z_i)}(|0\rangle\langle 0|^{\otimes k-1}\otimes|p_{k+t:n}\rangle\langle p_{k+t:n}|, |1\rangle\langle 1|^{\otimes k-1}\otimes|q_{k+t:n}\rangle\langle q_{k+t:n}|). \quad (6.62)$$

In a similar manner to how we proceeded in Theorem 10, using Lemmas 2 and 11, we have

$$\mu_T^{(Z_i)}(|p\rangle\langle p|, |q\rangle\langle q|) = \alpha\Delta_1^{(=)} + \beta\Delta_1^{(\neq)}, \quad (6.63)$$

where

$$\alpha = \frac{1}{2^{k-1}} + \frac{(2^{2k-2}-1)(2^k\delta_{p_{1:k}q_{1:k}}-1)}{2^{k-1}(2^{2k}-1)}, \quad (6.64)$$

$$\beta = \frac{2^{k-1}-1}{2^{k-1}} - \frac{(2^{2k-2}-1)(2^k\delta_{p_{1:k}q_{1:k}}-1)}{2^{k-1}(2^{2k}-1)}. \quad (6.65)$$

Now, let us consider these values when $p = q$ and $p \neq q$. Define

$$\alpha^{(=)} = \frac{1}{2^{k-1}} + \frac{2^{2k-2}-1}{2^{k-1}(2^k+1)}, \tag{6.66}$$

$$\beta^{(=)} = \frac{2^{k-1}-1}{2^{k-1}} - \frac{2^{2k-2}-1}{2^{k-1}(2^k+1)}, \tag{6.67}$$

$$\alpha^{(\neq)} = \frac{1}{2^{k-1}} - \frac{2^{2k-2}-1}{2^{k-1}(2^{2k}-1)}, \tag{6.68}$$

$$\beta^{(\neq)} = \frac{2^{k-1}-1}{2^{k-1}} + \frac{2^{2k-2}-1}{2^{k-1}(2^{2k}-1)}. \tag{6.69}$$

So we have

$$\mu_T^{(Z_i)}(|\mathbf{0}\rangle\langle\mathbf{0}|, |\mathbf{0}\rangle\langle\mathbf{0}|) = \alpha^{(=)}\Delta_1^{(=)} + \beta^{(=)}\Delta_1^{(\neq)}. \tag{6.70}$$

Similar to the proof of Theorem 10, we can see that when we integrate $\Delta_1^{(=)}$ and $\Delta_1^{(\neq)}$ with respect to U_{T-1}, we get linear combinations of $\Delta_2^{(=)}$ and $\Delta_2^{(\neq)}$.

Assume that after integration over unitaries $U_T, \ldots U_{T-t+1}$, we get $\gamma^{(=)}\Delta_t^{(=)} + \gamma^{(\neq)}\Delta_t^{(\neq)}$. Now, if we integrate over U_{T-t}, the coefficients of $\Delta_{t+1}^{(=)}$ and $\Delta_{t+1}^{(\neq)}$ will be $\alpha^{(=)}\gamma^{(=)} + \alpha^{(\neq)}\gamma^{(\neq)}$ and $\beta^{(=)}\gamma^{(=)} + \beta^{(\neq)}\gamma^{(\neq)}$ respectively. Unlike in the proof of Theorem 10, these coefficients are independent of t.

So we have

$$\mu_T^{(Z_i)}(|\mathbf{0}\rangle\langle\mathbf{0}|, |\mathbf{0}\rangle\langle\mathbf{0}|) = M_T M_{n-k-1} M_1 \tag{6.71}$$

where

$$M_1 = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \; M_T = \begin{bmatrix} \alpha_T & \beta_T \end{bmatrix}, \; M = \begin{bmatrix} \alpha^{(=)} & \alpha^{(\neq)} \\ \beta^{(=)} & \beta^{(\neq)} \end{bmatrix}, \tag{6.72}$$

$$\alpha_T = \Delta_{T-1}^{(=)} = \frac{1}{2^k+1}, \qquad \beta_T = \Delta_{T-1}^{(\neq)} = \frac{-1}{2^{2k}-1}. \tag{6.73}$$

α_t and β_t can be evaluated directly using Lemma 5. The eigenvalues of the matrix M are 1 and

$$B_2 = \frac{2^{2k-2}-1}{2^{2k}-1}. \tag{6.74}$$

Therefore,

$$\mu_T^{(Z_i)}(|\mathbf{0}\rangle\langle\mathbf{0}|, |\mathbf{0}\rangle\langle\mathbf{0}|) = B_0 + B_1 B_2^{n-k-1} \geqslant B_0, \tag{6.75}$$

where

$$B_0 = \frac{1}{2^{2k-1}+1}, \tag{6.76}$$

$$B_1 = \frac{2^{5k} - 2^{4k} - 6\cdot 2^{3k} + 4\cdot 2^{2k} + 2\cdot 2^k}{2(2^k+1)^2\cdot(2^{2k}-1)(2^{2k}+2)}. \tag{6.77}$$

Now, notice that whenever $i > k$, we have

$$\int_{\mathbf{U}_T} \left(\mathrm{tr}\left(Z_i|\mathbf{0}\rangle\langle\mathbf{0}|_{C_T}\right)\right)^2 \mathrm{d}\mathbf{U}_T = \int_{\mathbf{U}_{T-i+k}} \mathrm{tr}\left(Z_k\,|0\rangle\langle 0|^{\otimes n-i+k}_{C_{T-i+k}}\right)^2 \mathrm{d}\mathbf{U}_{T-i+k}, \tag{6.78}$$

which is also an instance of the previous case defined over $n-i+k$ qubits. Hence, from Eq (6.75), we can see that the same lower bound shall apply in this case as well. So, we have

$$\mathrm{Var}_{\boldsymbol{\theta}}\left(f_{\rho,O}(\boldsymbol{\theta})\right) \geqslant \frac{1}{4n} B_0. \tag{6.79}$$

Now, consider $\sigma \in \mathbb{D}_n$ such that $\|\rho - \sigma\|_{\mathrm{tr}} \leqslant \epsilon$ for some pure product state ρ. Then, for

any i, we have

$$\begin{aligned}
&\left| \int_{\mathbf{U}_T} \operatorname{tr}\left(Z_i \rho_{C_T}\right)^2 \mathrm{d}\mathbf{U}_T - \int_{\mathbf{U}_T} \operatorname{tr}\left(Z_i \sigma_{C_T}\right)^2 \mathrm{d}\mathbf{U}_T \right| \\
&\leqslant \int_{\mathbf{U}_T} \left| \operatorname{tr}\left(Z_i \rho_{C_T}\right)^2 - \operatorname{tr}\left(Z_i \sigma_{C_T}\right)^2 \right| \mathrm{d}\mathbf{U}_T \qquad (6.80) \\
&= \int_{\mathbf{U}_T} \left| \operatorname{tr}\left(Z_i \rho_{C_T}\right) - \operatorname{tr}\left(Z_i \sigma_{C_T}\right) \right| \times \left| \operatorname{tr}\left(Z_i \rho_{C_T}\right) + \operatorname{tr}\left(Z_i \sigma_{C_T}\right) \right| \, \mathrm{d}\mathbf{U}_T \qquad (6.81) \\
&= \int_{\mathbf{U}_T} \left| \operatorname{tr}\left(Z_i (\rho - \sigma)_{C_T}\right) \right| \times \left| \operatorname{tr}\left(Z_i (\rho + \sigma)_{C_T}\right) \right| \, \mathrm{d}\mathbf{U}_T \qquad (6.82) \\
&\leqslant \|\rho - \sigma\|_{\mathrm{tr}} \times \|\rho + \sigma\|_{\mathrm{tr}} \qquad (6.83) \\
&\leqslant 2\|\rho - \sigma\|_{\mathrm{tr}}, \qquad (6.84)
\end{aligned}$$

where $\|\cdot\|_{\mathrm{tr}}$ is the trace norm and the second-last inequality follows from Tracial Matrix Hölder's Inequality (Lemma 9). Then, we have

$$\sum_{i=1}^{n} \int_{\mathbf{U}_T} \operatorname{tr}\left(Z_i \rho_{C_T}\right)^2 \mathrm{d}\mathbf{U}_T - 2n\|\sigma - \rho\|_{\mathrm{tr}} \leqslant \sum_{i=1}^{n} \int_{\mathbf{U}_T} \operatorname{tr}\left(Z_i \sigma_{C_T}\right)^2 \mathrm{d}\mathbf{U}_T. \qquad (6.85)$$

This implies that

$$\frac{1}{4n^2} \sum_{i=1}^{n} \int_{\mathbf{U}_T} \operatorname{tr}\left(Z_i \rho_{C_T}\right)^2 \mathrm{d}\mathbf{U}_T - \frac{1}{2n}\|\sigma - \rho\|_{\mathrm{tr}} \leqslant \frac{1}{4n^2} \sum_{i=1}^{n} \int_{\mathbf{U}_T} \operatorname{tr}\left(Z_i \sigma_{C_T}\right)^2 \mathrm{d}\mathbf{U}_T \qquad (6.86)$$

further implying that

$$\operatorname{Var}_{\boldsymbol{\theta}}\left(f_{\rho,O}(\boldsymbol{\theta})\right) - \frac{1}{2n}\|\sigma - \rho\|_{\mathrm{tr}} \leqslant \operatorname{Var}_{\boldsymbol{\theta}}\left(f_{\sigma,O}(\boldsymbol{\theta})\right) \qquad (6.87)$$

and hence we have

$$\frac{B_0}{4n} - \frac{\epsilon}{2n} \leqslant \operatorname{Var}_{\boldsymbol{\theta}}\left(f_{\sigma,O}(\boldsymbol{\theta})\right). \qquad (6.88)$$

Now, for this σ, minimization over all ρ as required by the theorem completes the proof. □

Corollary 1. *Let $\sigma \in \mathbb{D}_n$ and $C_T^{(n)}$ be an MPS ansatz. Then, we have*

$$Var_{\boldsymbol{\theta}}\left(\partial_{\theta_{pq}} f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta})\right) \leqslant \frac{h(\sigma)}{4^{n-k-1}} \tag{6.89}$$

$\forall\ p,q$ such that $1 \leqslant p \leqslant T, 1 \leqslant q \leqslant m$, where $h(\sigma)$ is defined in Theorem 10 and $U_1, \ldots U_{p-1}, U_{p+1}, \ldots U_T$, along with one of $U_p^{(L,q)}$ or $U_p^{(R,q)}$ form unitary 2-designs and θ_{pq} is distributed uniformly.

Proof. Throughout this proof, we assume that within the computation of the variance, $U_p^{(L,q)}$ is distributed according to the Haar measure since trivial changes to the proof are sufficient to prove the same when $U_p^{(R,q)}$ is distributed according to Haar measure. Let $\mathbf{U}_T^{(p,q)} = \{U_1, \ldots, U_{p-1}, \theta_{pq}, U_p^{(L,q)}, U_{p+1}, \ldots, U_T\}$. Using Lemmas 8 and 7, we have

$$\begin{aligned}
\mathrm{Var}_{\boldsymbol{\theta}}\left(\partial_{\theta_{pq}} f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta})\right) =& \mathbb{E}_{\boldsymbol{\theta}}\left(\partial_{\theta_{pq}} f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta})\right)^2 && (6.90)\\
=& \frac{1}{4}\mathbb{E}_{\boldsymbol{\theta}}\left(f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta}_{pq+}) - f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta}_{pq-})\right)^2 && (6.91)\\
=& \frac{1}{4}\mathbb{E}_{\boldsymbol{\theta}}\Big(\left(f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta}_{pq+})\right)^2 + \left(f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta}_{pq-})\right)^2 \\
& - 2 f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta}_{pq+}) f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta}_{pq-})\Big) && (6.92)\\
\leqslant& \frac{1}{4}\mathbb{E}_{\boldsymbol{\theta}}\left(\left(f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta}_{pq+})\right)^2 + \left(f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta}_{pq-})\right)^2\right) \\
& + \frac{1}{2}\left|\mathbb{E}_{\boldsymbol{\theta}}\left(f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta}_{pq+}) f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta}_{pq-})\right)\right|. && (6.93)
\end{aligned}$$

Now, notice that

$$\mathbb{E}_{\boldsymbol{\theta}}\left(f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta}_{pq+})\right)^2 = \mathbb{E}_{\boldsymbol{\theta}}\left(f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta}_{pq-})\right)^2 = \mathbb{E}_{\boldsymbol{\theta}}\left(f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta})\right)^2. \tag{6.94}$$

This follows from combining Lemma 4 with the fact that

$$U_p(\boldsymbol{\theta}_{pq\pm}) = U_p^{(L,q)}(\boldsymbol{\theta}_p) e^{-i(\theta_{pq} \pm \pi/2) H_{pq}} U_p^{(R,q)}(\boldsymbol{\theta}_p) = U_p^{(L,q)}(\boldsymbol{\theta}_p) e^{-i\theta_{pq} H_{pq}} e^{\pm \frac{i\pi H_{pq}}{2}} U_p^{(R,q)}(\boldsymbol{\theta}_p) \tag{6.95}$$

Also,

$$\begin{aligned}
&\left| \int_{\mathbf{U}_T^{(\theta)}} f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta}_{pq+}) f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta}_{pq-}) \left(\prod_{j=1,j\neq p}^{T} \mathrm{d}U_j \right) \mathrm{d}U_p^{(L,q)} \mathrm{d}\theta_{pq} \right| \\
&\leqslant \sqrt{\int_{\mathbf{U}_T^{(\theta)}} \left(f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta}_{pq+}) \right)^2 \left(\prod_{j=1,j\neq p}^{T} \mathrm{d}U_j \right) \mathrm{d}U_p^{(L,q)} \mathrm{d}\theta_{pq}} \\
&\quad \times \sqrt{\int_{\mathbf{U}_T^{(\theta)}} \left(f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta}_{pq-}) \right)^2 \left(\prod_{j=1,j\neq p}^{T} \mathrm{d}U_j \right) \mathrm{d}U_p^{(L,q)} \mathrm{d}\theta_{pq}} \qquad (6.96) \\
&= \mathbb{E}_{\boldsymbol{\theta}} \left(f_{\sigma,|\mathbf{0}\rangle\langle\mathbf{0}|}(\boldsymbol{\theta}) \right)^2 . \qquad (6.97)
\end{aligned}$$

This follows directly from Cauchy Schwarz inequality (cf. Lemma 10). Plugging Eqs (6.96) and (6.94) into Eq (6.90) completes the proof. □

Corollary 2. *Let $\sigma \in \mathbb{D}_n$ and $C_T^{(n)}$ be an MPS ansatz. Let $O = 1/n \sum_{i=1}^{n} |0\rangle\langle 0|_i$. Then, there exist p, q with $1 \leqslant p \leqslant T, 1 \leqslant q \leqslant m$ such that*

$$Var_{\boldsymbol{\theta}} \left(\partial_{\theta_{pq}} f_{\sigma,O}(\boldsymbol{\theta}) \right) \notin \mathcal{O}\left(\frac{1}{b^n} \right), \qquad (6.98)$$

where $U_1, \ldots U_{q-1}, U_{q+1}, \ldots U_T$, along with one of $U_p^{(L,q)}$ or $U_p^{(R,q)}$ form unitary 2-designs and θ_{pq} is distributed uniformly.

Proof. First, we recall an important lemma relating cost concentration with barren plateaus from [Arr+22].

Lemma 3. *[Arr+22] For any ansatz $C(\boldsymbol{\theta})$, $\sigma \in \mathbb{D}_n$ and $W \in \mathbb{H}_n$, if $\forall\ p, q$, where $1 \leqslant p \leqslant T, 1 \leqslant q \leqslant m$,*

$$Var_{\boldsymbol{\theta}} \left(\partial_{pq} f_{\sigma,W}(\boldsymbol{\theta}) \right) \in \mathcal{O}(1/b^n) \qquad (6.99)$$

for some $b > 1$, then

$$Var_{\boldsymbol{\theta}}\left(f_{\sigma,W}(\boldsymbol{\theta})\right) \in \mathcal{O}(1/b^n). \tag{6.100}$$

The result then follows from the contrapositive of Theorem 3. □

Theorem 3. *For any n-qubit quantum circuit V, 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 $j \leqslant i \leqslant k$. Then, for any product observable W, $\|W_{V^\dagger}\|_{\mathbb{K}}$ can be classically evaluated with cost $\mathcal{O}(2^{R_V})$.*

Proof. From [Joz06], we can see that given a classical description of a circuit V, an MPS description of $W_{V^\dagger}$ with bond dimension $\mathcal{O}(2^{R_V})$ can be computed using with computational cost scaling as $\mathcal{O}(2^{R_V})$. The change from the standard basis to the orthonormal Pauli basis is efficient, involving only local rotations. Let $\widehat{W}_{V^\dagger}$ be a vector of Pauli basis coefficients of $W_{V^\dagger}$. Then, $\|W\|_{\mathbb{K}} = \sqrt{\|\widehat{W}_{V^\dagger} * \widehat{W}_{V^\dagger}\|_2 / 2^n}$, where $*$ is the Hadamard product. From [Ose11], the Hadamard and inner products of tensors represented as MPS can be computed efficiently, with cost scaling polynomially in their bond dimension, thus completing the proof. □

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

$$\|W\|_{C,\mathbb{K}} := \int_{\boldsymbol{\theta}} \left\| W_{C(\boldsymbol{\theta})^\dagger} \right\|_{\mathbb{K}} d\boldsymbol{\theta}. \tag{6.101}$$

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

Proof. First notice that $\|W\|_{C,\mathbb{K}}$ is non-negative since it is an average of norms. When $W = 0$, then $\|W\|_{C,\mathbb{K}} = 0$. Similarly, when $\|W\|_{C,\mathbb{K}} = 0$, we will have $\left\| W_{C(\boldsymbol{\theta})^\dagger} \right\|_{\mathbb{K}} = 0 \ \forall \ \boldsymbol{\theta}$ and hence $W = 0$.

Next, we prove the triangle inequality.

$$\left\|W^{(1)}\right\|_{C,\mathbb{K}} + \left\|W^{(2)}\right\|_{C,\mathbb{K}} = \int_{\boldsymbol{\theta}} \left\|W^{(1)}_{C(\boldsymbol{\theta})^\dagger}\right\|_{\mathbb{K}} + \left\|W^{(2)}_{C(\boldsymbol{\theta})^\dagger}\right\|_{\mathbb{K}} \mathrm{d}\boldsymbol{\theta} \tag{6.102}$$

$$\geqslant \int_{\boldsymbol{\theta}} \left\|W^{(1)}_{C(\boldsymbol{\theta})^\dagger} + W^{(2)}_{C(\boldsymbol{\theta})^\dagger}\right\|_{\mathbb{K}} \mathrm{d}\boldsymbol{\theta} \tag{6.103}$$

$$= \int_{\boldsymbol{\theta}} \left\|\left(W^{(1)} + W^{(2)}\right)_{C(\boldsymbol{\theta})^\dagger}\right\|_{\mathbb{K}} \mathrm{d}\boldsymbol{\theta} \tag{6.104}$$

$$= \left\|W^{(1)} + W^{(2)}\right\|_{C,\mathbb{K}}. \tag{6.105}$$

□

6.7 Related Works

In [Liu+22; Gar+23; BM24], 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+22], 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. 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 [HL09a].

In [Dov+22], it was experimentally observed that the usage of global observables leads to exponentially decreasing gradient magnitudes, whereas local observables avoid this issue. In our work, we study this phenomenon as well as similar trends in cost concentration theoretically. The existence of exponentially decreasing partial derivatives in MPS ansatz-based VQAs is proved using ZX-calculus in [ZG21]. 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 [ZG21], 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.

Chapter 7

Conclusion and Future Direction

In this thesis, we have presented advancements in the training and optimization of VQAs. Our contributions span from introducing novel algorithms to providing theoretical insights and experimental validations, all aimed at making VQAs more efficient and scalable.

Firstly, we introduced ALSO, an efficient method for training alternating layered VQAs. By leveraging classical shadows, ALSO significantly reduces the number of input state copies required, providing exponential improvements over traditional methods, with rigorous performance guarantees. This is particularly beneficial in practical scenarios that necessitate multiple optimization rounds such as when hyperparameter tuning or finding the right optimizer is required. Additionally, the simplicity of ALSO's implementation, requiring only single-qubit measurements, further enhances its practicality. Another notable advantage is that the generated classical shadows can be reused for various independent tasks. For instance, the same set can be employed to find state preparation circuits and to build quantum autoencoders. We also experimentally demonstrate 2-3 orders of magnitude reduction in training costs for these tasks.

Secondly, we proposed AISO, a training algorithm that achieves similar exponential reductions in quantum resource requirements through the use of shallow shadows. AISO's general applicability to various shallow quantum circuit structures and observables with low Frobenius norms allows for extensive optimization with minimal quantum device ex-

ecutions. Our demonstrations in state preparation and VQCS underscore the practical advantages of AISO in important quantum information use cases.

Lastly, we explored the trainability and classical simulability of learning MPS approximations of quantum states using VQAs. Our theoretical results highlight the importance of local observables in avoiding the exponential decay in the variance of the cost function and its derivatives, which is induced by global observables. We further demonstrated that local observables induce effective subspaces within the Pauli basis, enabling the potential classical simulation of the MPS ansatz. These findings were experimentally validated, confirming their applicability across various scenarios.

In conclusion, the algorithms and theoretical insights presented in this thesis offer improvements in the efficiency and feasibility of VQAs. By addressing key challenges in the training and optimization of VQAs, our work paves the way for more scalable quantum technologies. Future research can build upon these foundations to explore further applications and refinements, ultimately advancing the field of VQAs.

For future work, we aim to extend our resource-efficient VQA protocols to other trainable ansatz-observable combinations, such as the QCNN ansatz-local observable combination. Leveraging classical machine learning techniques with classical shadows, similar to the approach in [Hua+21], will be a key focus area. Another significant avenue for future research is to apply similar shadow tomography methods to other important domains within quantum information, such as error correction [KL97; Fow+12; Kit03] and device calibration [Li+13; Jai+11; Wit+21], which typically require large amounts of copies or executions.

Also, we plan to conduct a more extensive study of the performance of these methods in the presence of noise in real quantum devices, and compare these results with other classical simulation strategies available in the literature. Hypothetically, interesting challenges could arise in the presence of noise. For example, when using a standard VQA to solve state preparation on a real quantum device, the learned parameters adapt to the device's specific noise profile. In contrast, classical simulation methods optimize parameters in an

ideal, noiseless setting, potentially leading to different outcomes. Moreover, if the data acquisition phase was significantly affected by noise, then the results could vary even more. Hence, this area is an intriguing and promising area of study which we would like to embark on in the future.

Regarding the trainability of learning MPS approximations using VQAs, we plan to generalize and enhance our results by theoretically proving the trainability of protocols when multiple layers are used. Also, we aim to extend the current proofs to encompass all quantum states and perform a more rigorous theoretical analysis of effective subspaces. Furthermore, since we have already presented strong evidence for the existence of a classical simulation algorithm that consumes only a few quantum resources, we aim to develop such an efficient algorithm with rigorous performance guarantees.

Chapter 8

Appendix

8.1 Tensors

A tensor can be seen as a multidimensional array. These are data structures with multiple number of indices allowed. For example, a matrix M with entries $M_{i,j}$ can be seen as a tensor with 2 indices. Every tensor can be mathematically defined using the Kronecker product operation. Let $|x\rangle \in \mathcal{V} \subset \mathbb{C}^m$, where $|x\rangle = \sum_{i=0}^{m-1} x_i |i\rangle$ and $|y\rangle \in \mathcal{W} \subset \mathbb{C}^n$, where $|y\rangle = \sum_{j=0}^{n-1} y_j |j\rangle$. Then $|x\rangle \otimes |y\rangle \in \mathbb{C}^{mn}$ and $|x\rangle \otimes |y\rangle = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} x_i y_j |i\rangle \otimes |j\rangle = \sum_{k=0}^{mn-1} x_{\lfloor k/n \rfloor} y_{k \bmod n} |k\rangle$. A tensor product of two finite dimensional vector spaces $\mathcal{V}$ and $\mathcal{W}$, defined by $\mathcal{V} \otimes \mathcal{W}$ is the vector space of all possible linear combinations of vectors of the form $|v\rangle \otimes |w\rangle$ for all $|v\rangle \in V$ and $|w\rangle \in W$. If $\{|v_1\rangle, |v_2\rangle, \ldots, |v_n\rangle\}$ is a basis of $\mathcal{V}$ and $\{|w_1\rangle, |w_2\rangle, \ldots, |w_m\rangle\}$ is a basis of $\mathcal{W}$, then $\{|v_i\rangle \otimes |w_j\rangle | i = 1, \ldots, n, j = 1, \ldots, m\}$ is a basis of $\mathcal{V} \otimes \mathcal{W}$. Hence, the dimension of $\mathcal{V} \otimes \mathcal{W}$ is the product of the dimension of $\mathcal{V}$ and dimension of $\mathcal{W}$.

Let $\mathcal{V} = \bigotimes_{n=1}^{N} \mathbb{C}^{I_n} = \mathbb{C}^{I_1 \times \cdots \times I_N}$ be a tensor product of vector spaces. Let $|v\rangle \in \mathcal{V}$. $|v\rangle$ can always be written as

$$|v\rangle = \sum_{i_1=0}^{I_1-1} \cdots \sum_{i_N=0}^{I_N-1} v_{i_1,\ldots,i_N} |i_1^{(1)}\rangle \ldots |i_N^{(N)}\rangle \tag{8.1}$$

Here, $|i_j^{(j)}\rangle$ is the i_j^{th} standard basis vector of $\mathbb{C}^{I_j}$ and each of the vector spaces are called *modes*. These tensors are also sometimes called order N tensors or N^{th} order tensors. For notational convenience, this is also written as

$$|v\rangle = \sum_{\substack{i_k=0 \\ k=1,\ldots,N}}^{I_k-1} v_{i_1,\ldots,i_N} |i_1\rangle \ldots |i_N\rangle \tag{8.2}$$

From a multidimensional array perspective, the coordinate of $|v\rangle$ in the standard basis of V, associated with the vector $|i_1\rangle\otimes\cdots\otimes|i_N\rangle$, that is, $v_{i_1,\ldots,i_N}$, is the $(i_1,\ldots,i_N)$ element of the multidimensional array v associated with $|v\rangle$. So, throughout this section, we might define tensors simply as $A \in \mathbb{C}^{I_1\times\ldots I_k}$, where $A_{i_1,\ldots i_k}$ is its $(i_1,\ldots,i_k)$ element.

Now, we shall define *tensor splitting*, which we will be using in multiple parts of the thesis.

Definition 4. *Given a tensor $A \in \mathbb{C}^{I_1\times I_2\times\cdots\times I_t}$, splitting the k^{th} index into two indices of length I'_k and I''_k (such that $I'_k I''_k = I_k$) results in a new tensor*

$$A' \in \mathbb{C}^{I_1\times I_{k-1}\times I'_k\times I''_k\times\cdots\times I_t} \tag{8.3}$$

where

$$A'_{i_1,\ldots,i'_k,i''_k,\ldots i_t} = A_{i_1,\ldots,i_{k-1},I''_k i'_k+i''_k,i_{k+1},\ldots,i_t}. \tag{8.4}$$

Similarly, for any tensor $A \in \mathbb{C}^{I_1\times I_2\times\ldots I_{k-1}\times p^m\times I_{k+1}\times\cdots\times I_t}$, splitting its k^{th} index into m indices of length p results in the tensor

$$A' \in \mathbb{C}^{I_1\times I_2\times\ldots I_{k-1}\times p\times p\times\cdots\times p\times I_{k+1}\times\cdots\times I_t} \tag{8.5}$$

where

$$A'_{i_1,\ldots,i_{k-1}p_0,p_1,\ldots,p_{m-1},i_{k+1},\ldots,i_t} = A_{i_1,\ldots,i_{k-1},\left(\sum_{j=0}^{m-1} p_{m-1-j}p^j\right),i_{k+1},\ldots,i_t}. \tag{8.6}$$

Here, $p_0p_1\ldots p_{m-1}$ *is simply a p-nary expansion of* $\sum_{j=0}^{m-1} p_{m-1-j}p^j$.

Also, $\forall A \in \mathbb{C}^{I_1\times\cdots\times I_k}$ we define $A_{i_1\ldots i_{k'}} \in \mathbb{C}^{I_{k'+1}\times\cdots\times I_k}$ such that its $(i_{k'+1},\ldots,i_k)^{\text{th}}$ entry is $A_{i_1\ldots i_k}$.

8.1.1 Tensor Contraction

Similar to how we can view tensors as an extension of the concept of matrices to multiple dimensions, one can also extend matrix multiplication to a more general operation with tensors. It is called *tensor contraction*. Within matrix multiplication of two matrices $A \in \mathbb{C}^{I_1\times I_2}$ and $B \in \mathbb{C}^{I_2\times I_3}$, what essentially happens is that we get a matrix $C \in \mathbb{C}^{I_1\times I_3}$, whose each entry is a summation of the form $C_{i_1i_3} = \sum_{i_2=0}^{I_2-1} A_{i_1,i_2}B_{i_2,i_3}$. Here, it is crucial that the second index of A should have the same length as the first index of B. The natural way to extend this to tensors is to allow one to sum over multiple indices of two tensors, having the same lengths. For example, if we are given two tensors $A \in \mathbb{C}^{I_1\times I_2\times I_3\times I_4}$ and $B \in \mathbb{C}^{I_6\times I_4\times I_3\times I_7\times I_8}$, we can carry out this extended matrix multiplication with the last two indices of A and the third and second indices of B. The resultant tensor $C \in \mathbb{C}^{I_6\times I_7\times I_8\times I_1\times I_2}$ will contain all the remaining indices (ordered as the remaining indices of B first, then the remaining indices of A) and would then have the form

$$C_{i_6,i_7,i_8,i_1,i_2} = \sum_{i_3=0}^{I_3-1}\sum_{i_4=0}^{I_4-1} A_{i_1,i_2,i_3,i_4}B_{i_6,i_4,i_3,i_7,i_8}. \tag{8.7}$$

Notice that unlike matrix multiplication, this is an index-dependent operation, in the sense that we can choose to contract any pairs of indices (one from each tensor), as long as they have the same length. The formal definition of this operation is as follows:

Definition 5. *Let $A \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_{t_1}}$ and $B \in \mathbb{C}^{J_1 \times J_2 \times \cdots \times J_{t_2}}$ be two tensors. Tensor contraction of the indices $\mathcal{I} = [p_1, \ldots, p_t]$ of A with indices $\mathcal{J} = [q_1, \ldots, q_t]$ of B requires $I_{p_k} = J_{q_k}\ \forall\ k$, is depicted as $(A, \mathcal{I}) \star (B, \mathcal{J})$ and results in a tensor*

$$C \in \mathbb{C}^{\left(\underset{\substack{k=1 \\ J_k \notin \mathcal{J}}}{\overset{t_2}{\times}} J_k\right) \times \left(\underset{\substack{k=1 \\ I_k \notin \mathcal{I}}}{\overset{t_1}{\times}} I_k\right)} \tag{8.8}$$

with

$$C^{j_1,\ldots,j_{q_1-1}j_{q_1+1},\ldots,j_{q_2-1}j_{q_2+1},\ldots,j_{q_t-1}j_{q_t+1},\ldots,j_{t_2}}_{i_1,\ldots,i_{p_1-1}i_{p_1+1},\ldots,i_{p_2-1}i_{p_2+1},\ldots,i_{p_t-1}i_{p_t+1},\ldots,i_{t_1}}$$
$$= \sum_{i_{p_1}, i_{p_2}, \ldots, i_{p_t}} A_{i_1,\ldots,i_{t_1}} B_{j_1,\ldots,j_{q_1-1},i_{p_1},j_{q_1+1}\ldots,j_{q_2-1},i_{p_2},j_{q_2+1},\ldots,j_{q_t-1}i_{p_t}j_{q_t+1},\ldots j_{q_{t_2}}}.$$

The cost of multiplying two matrices (carried out naively) $A \in \mathbb{C}^{I_1 \times I_2}, B \in \mathbb{C}^{I_2 \times I_3}$ is $\mathcal{O}(I_1 I_2 I_3)$ since we have to compute $I_1 I_3$ new values, each requiring computation of inner product of vectors of length I_2. A similar logic can be used to argue that the cost of $(A, [p_1, \ldots, p_t]) \star (B, [q_1, \ldots, q_t])$ $A \in \mathbb{C}^{I_1 \times \cdots \times I_N}$ and $B \in \mathbb{C}^{J_1 \times \cdots \times J_M}$ will be $\mathcal{O}\left(\left(\prod_{\substack{j=1 \\ j \notin \mathcal{J}}}^{M} J_j\right)\left(\prod_{\substack{i=1 \\ i \notin \mathcal{I}}}^{N} I_i\right)\left(\prod_{k=1}^{t} I_{p_k}\right)\right)$.

8.1.2 Tensor Networks

Tensor networks are diagrammatical representations of tensors and contractions. Any order N tensor is represented using a box (or any other shape) with n lines sticking out, where each line represents an index (mode). Examples are given in Figure 8.1. Given any two tensors $A \in \mathbb{C}^{I_1 \times \cdots \times I_N}, B \in \mathbb{C}^{J_1 \times \cdots \times J_N}$, to depict the contraction of a pair of modes, we simply connect the two lines representing them. Examples are given in Figure 8.2.

In some sense, one can also view quantum circuits as tensor networks, by viewing all the $|0\rangle$s in the beginning as vectors and k-qubit gates as $2k$ dimensional tensors. To compute the probability of any specific outcome $b_1 b_2 \ldots b_n$, one simply connects $|b_1\rangle |b_2\rangle, |b_n\rangle$ to the end of the circuit. Also, if we wish to compute the expectation of an observable of the

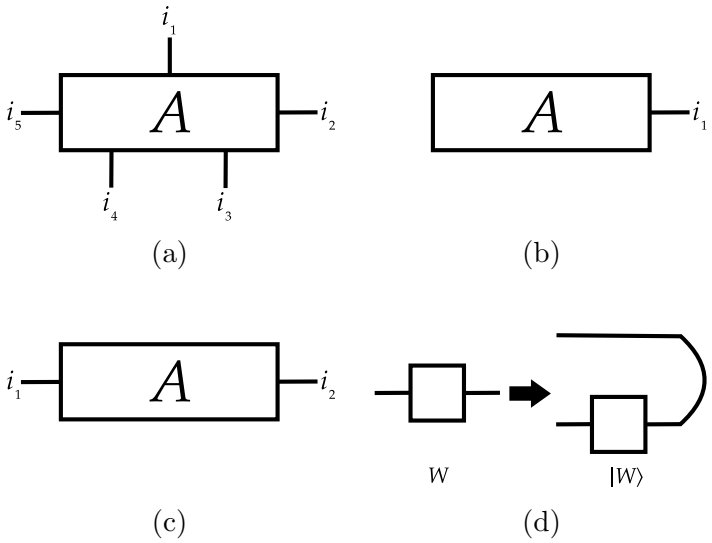


Figure 8.1: Tensor network examples: (a) An order 5 tensor A with each of its five indices represented using the black lines. (b) Tensor network representation of a vector. (c) Tensor network representation of a matrix. (d) Vectorization of a matrix W.

form $O = O^{(1)} \otimes \cdots \otimes O^{(n)}$, where $O^{(i)} \in \mathcal{L}(\mathbb{C}^2)$, we have to contract a tensor network that depicts $\langle \mathbf{0} | U^\dagger O U | \mathbf{0} \rangle$, where U is the quantum circuit. First, depict this using n single qubit matrices, then attach all free indices of the circuit to all the second indices of all the matrices $O^{(i)}$. Then add another tensor network circuit depicting $U^\dagger$, which is the same tensor network written in reverse, with all gates replaced with their inverses. Then connect all the free indices of this circuit to all the first indices of all $O^{(i)}$s. An example is provided in Figure 8.3.

8.2 Qubit Permutation

Another way of describing the application of a t-qubit gate U on an n-qubit state $|\psi\rangle$ is by changing the indices of the tensor $|\psi\rangle$. Let $\pi \in S_n$, where S_n is the symmetric group on n elements. Consider the operator $W_\pi \in \mathcal{L}(\mathbb{C}^{2^n})$ defined as

$$W_\pi \left(\bigotimes_{k=1}^{n} |j_k\rangle \right) = \bigotimes_{k=1}^{n} |j_{\pi^{-1}(k)}\rangle . \tag{8.9}$$

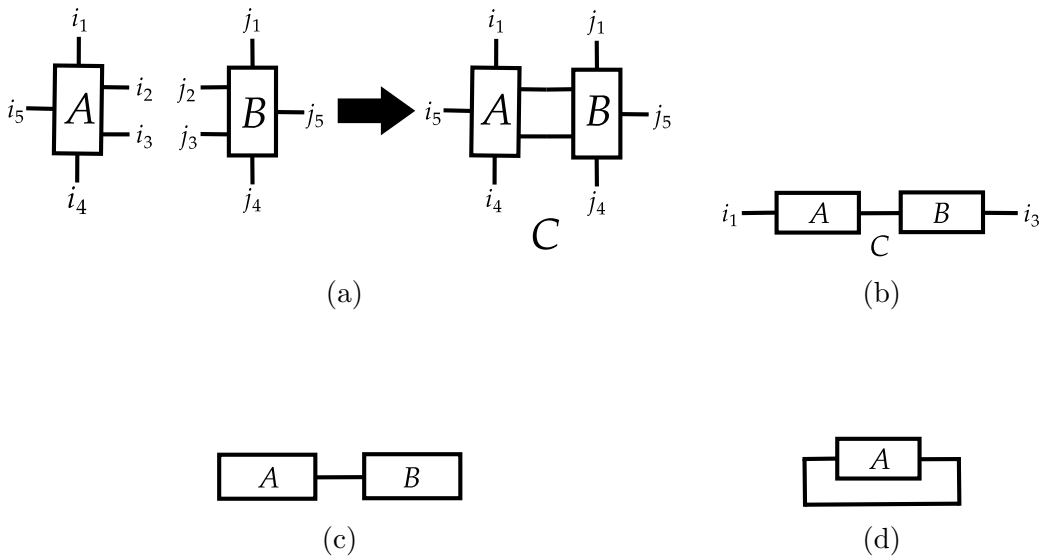


Figure 8.2: Tensor contraction examples. (a) Contraction of two pairs of modes, $[2,3]$ and $[2,3]$, of two order 5 tensors $A \in \mathbb{C}^{I_1 \times I_2 \times I_3 \times I_4 \times I_5}$ and $B \in \mathbb{C}^{J_1 \times J_2 \times J_3 \times J_4 \times J_5}$ resulting in an order 6 tensor $C \in \mathbb{C}^{J_1 \times J_4 \times J_5 \times I_1 \times I_4 \times I_5}$. The k^{th} mode of A, B is depicted using indices i_k, j_k respectively. (b) Matrix multiplication of $A \in \mathbb{C}^{I_1 \times I_2}, B \in \mathbb{C}^{I_2 \times I_3}$ resulting in a matrix $C \in \mathbb{C}^{I_1 \times I_3}$. (c) The standard inner product of two vectors A, B. (d) Trace of a matrix A.

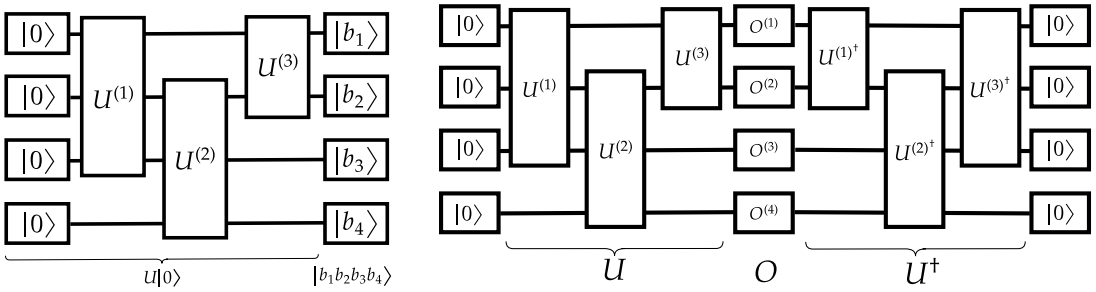
Since this is a norm-preserving transformation, it is a unitary transformation and $|\phi\rangle$ can be computed as

$$|\phi\rangle = W_\pi \left(U \otimes \mathbb{1}_{\mathbb{C}^{2^{n-k}}}\right) W_{\pi^{-1}} |\psi\rangle = Q_{U,\pi} |\psi\rangle, \tag{8.10}$$

where

$$\pi(j) = \begin{cases} j_k & \text{if } \ k \in \{1, 2, \ldots, t\} \\ \text{any permissible value otherwise,} \end{cases} \tag{8.11}$$

and $Q_{U,\pi} = W_\pi \left(U \otimes \mathbb{1}_{\mathbb{C}^{2^{n-k}}}\right) W_{\pi^{-1}}$. This can be interpreted as follows: the gate $W_{\pi^{-1}}$ will rearrange the qubits (or the modes of the tensor) in such a manner that the qubits in $\mathcal{J}$ are together and in order. Then, we apply U on them and leave other qubits untouched. Finally, we bring the qubits back to its original position.



(a) (b)

Figure 8.3: Tensor network circuit examples. (a) Tensor network to compute the probability of the outcome $b_1b_2b_3b_4$ of a circuit $U = U^{(3)}_{[1,2]}U^{(2)}_{[2,3,4]}U^{(1)}_{[1,2,3]}$ and input state $|\mathbf{0}\rangle$. The application of all three gates gives us the state $U|\mathbf{0}\rangle$. Then, compute the inner product with the vector $|b_1b_2b_3b_4\rangle$ (cf. Figure 8.2 (d)), whose squared absolute value is the probability in question. (b) Tensor network to compute the expectation of a product observable $O = O^{(1)} \otimes O^{(2)} \otimes O^{(3)} \otimes O^{(4)}$. We have used the same circuit as previously. We can interpret this as computing the inner product between $U^\dagger OU|\mathbf{0}\rangle$ and $|\mathbf{0}\rangle$, which is the required expectation $\langle\mathbf{0}|U^\dagger OU|\mathbf{0}\rangle$.

8.3 Partial Trace of Quantum States

Let A be a register of dimension d_A and B be a register of dimension d_B. Given any state $\sigma \in \mathcal{L}(\mathbb{C}^{d_Ad_B})$, the aim of this section is to show that $\mathrm{tr}_B(\sigma)$ is a valid quantum state, that is, it is a trace 1 positive semidefinite matrix. The same proof can be easily used for $\mathrm{tr}_A(\sigma)$.

First, we shall show that it is a unit trace matrix. From Eq (2.14), we have

$$\mathrm{tr}(\mathrm{tr}_B(\sigma)) = \sum_{p=0}^{d_A-1} \langle p| \left(\sum_{i_1,j_1=0}^{d_A-1} \sum_{q=0}^{d_B-1} \sigma_{i_1q,j_1q} |i_1\rangle\langle j_1| \right) |p\rangle \tag{8.12}$$

$$= \sum_{p=0}^{d_A-1} \sum_{q=0}^{d_B-1} \sigma_{pq,pq} \tag{8.13}$$

$$= \mathrm{tr}(\sigma) \tag{8.14}$$

$$= 1,. \tag{8.15}$$

Next, we shall show that $\mathrm{tr}_B(\sigma)$ is positive semidefinite. To see this, let $|\psi\rangle$ be an arbitrary

vector in $\mathbb{C}^{d_A}$. Then, from Eq (2.14), we have

$$
\begin{aligned}
&\langle\psi|\,\mathrm{tr}_B(\sigma)\,|\psi\rangle \\
&= \sum_{q=0}^{d_B-1}\sum_{i_1,j_1=0}^{d_A-1}\sum_{i_2,j_2=0}^{d_B-1}\sigma_{i_1i_2,j_1j_2}\langle\psi|\left(\mathbb{1}_{\mathbb{C}^{d_A}}\otimes\langle q|\right)|i_1\rangle\langle j_1|\otimes|i_2\rangle\langle j_2|\left(\mathbb{1}_{\mathbb{C}^{d_B}}\otimes|q\rangle\right)|\psi\rangle \\
&= \sum_{q=0}^{d_B-1}\sum_{i_1,j_1=0}^{d_A-1}\sum_{i_2,j_2=0}^{d_B-1}\sigma_{i_1i_2,j_1j_2}\left(\langle\psi|\otimes\langle q|\right)|i_1\rangle\langle j_1|\otimes|i_2\rangle\langle j_2|\left(|\psi\rangle\otimes|q\rangle\right) \\
&= \sum_{q=0}^{d_B-1}\langle\psi|\otimes\langle q|\left(\sum_{i_1,j_1=0}^{d_A-1}\sum_{i_2,j_2=0}^{d_B-1}\sigma_{i_1i_2,j_1j_2}|i_1\rangle\langle j_1|\otimes|i_2\rangle\langle j_2|\right)|\psi\rangle\otimes|q\rangle \\
&= \sum_{q=0}^{d_B-1}\left(\langle\psi|\otimes\langle q|\right)\sigma\left(|\psi\rangle\otimes|q\rangle\right) \\
&\geqslant 0.
\end{aligned}
$$

The last inequality follows since σ is positive semi-definite.

8.4 MPS Decomposition for Density Matrices

We can extend MPS decompositions for pure state vectors to density matrices as well. The only changes required is to view these objects as tensors lying in $\mathbb{C}^{4\times\cdots\times4}$. The core tensors will then have their first indices having length 4. All the operations that we have discussed can be easily extended to the case of density matrices in such a decomposition.

For example, let

$$\sigma = \sum_{i_1,\ldots,i_n=0}^{1}\sum_{j_1,\ldots,j_n=0}^{1} = \sigma_{i_1\ldots i_n,j_1\ldots j_n}|i_1\ldots i_n\rangle\langle j_1\ldots j_n| \tag{8.16}$$

be a density matrix. Define *vectorization* as an operation that converts such a matrix σ to

$$|\sigma\rangle = \sum_{i_1,\ldots,i_n=0}^{1}\sum_{j_1,\ldots,j_n=0}^{1}\sigma_{i_1\ldots i_n,j_1\ldots j_n}|i_1\ldots i_n\rangle|j_1\ldots j_n\rangle. \tag{8.17}$$

Although we can work with the MPS decomposition of $|\sigma\rangle$, there is an issue there. The problem is that, in that case, 2-dimensional vector spaces associated with the q^{th} qubit, represented by $|i_q\rangle$ and $|j_q\rangle$, are situated at the q^{th} and $(n+q)^{\text{th}}$ positions. This means that any operations applied on the q^{th} qubit is no longer a local or nearest neighbor operation, as is required in a standard MPS regime.

One easy way to resolve this is to vectorize σ and reorder the indices as

$$|\sigma_{\text{RO}}\rangle = \sum_{i_1,\ldots,i_n=0}^{1} \sum_{j_1,\ldots,j_n=0}^{1} = \sigma_{i_1\ldots i_n,j_1\ldots j_n} |i_1 j_1\rangle |i_2 j_2\rangle \ldots |i_n j_n\rangle . \tag{8.18}$$

Now, each pair of indices $|i_k j_k\rangle$ can be combined into an index of length 4. In this basis, the application of any single qubit gate V on the q^{th} qubit of σ can be implemented as $\left(V \otimes \overline{V}\right)_q |\sigma_{\text{RO}}\rangle$. The reason is that for any matrices A, B, we have $|A_B\rangle = \left(B \otimes \overline{B}\right) |A\rangle$.

Application of a two-qubit gate $V = \sum_{i_1,i_2=0}^{1} \sum_{j_1,j_2=0}^{1} V_{i_1 i_2, j_1 j_2} |i_1 i_2\rangle\langle j_1 j_2|$ will require a small permutation of the indices of $V \otimes \overline{V}$. This is because

$$V \otimes \overline{V} = \sum_{i_1,i_2=0}^{1} \sum_{j_1,j_2=0}^{1} \sum_{k_1,k_2=0}^{1} \sum_{l_1,l_2=0}^{1} V_{i_1 i_2, j_1 j_2} \overline{V_{k_1 k_2, l_1 l_2}} |i_1 i_2 k_1 k_2\rangle\langle j_1 j_2 l_1 l_2| . \tag{8.19}$$

Since we have applied a permutation of 1-qubit vector spaces on the input state, we have to apply the same permutation here. Mimicking the same permutation on $V \otimes \overline{V}$ will give us

$$(V \otimes \overline{V})_{\text{RO}} = \sum_{i_1,i_2=0}^{1} \sum_{j_1,j_2=0}^{1} \sum_{k_1,k_2=0}^{1} \sum_{l_1,l_2=0}^{1} V_{i_1 i_2, j_1 j_2} \overline{V_{k_1 k_2, l_1 l_2}} |i_1 k_1 i_2 k_2\rangle\langle j_1 l_1 j_2 l_2| . \tag{8.20}$$

Note that in the single-qubit gate case, we don't require such a permutation of indices. Now, we can say that the application of V on the q^{th} can be simulated by computing $(V \otimes \overline{V})_{\text{RO}_q} |\sigma_{\text{RO}}\rangle$. So both single-qubit and two-qubit operations have been localized and the same techniques we have used for vector-based MPS simulations can be used for the density matrix-based MPS as well.

8.5 MPS as an Extension of Product States

The concept of MPS can be seen as a generalization of product states. Note that for any product state $|\psi\rangle = |\psi^{(1)}\rangle \otimes \cdots \otimes |\psi^{(n)}\rangle$, with $|\psi^{(j)}\rangle = \psi_0^{(j)} |0\rangle + \psi_1^{(j)} |1\rangle$ being single qubit pure states, we have $\psi_{i_1 i_2, \ldots i_n} = \prod_{j=1}^{n} \psi_{i_j}^{(j)}$. Hence, a valid MPS decompostion of $|\psi\rangle$ is given by core tensors $G^{(j)}$ with $G_0^{(j)} = \psi_0^{(j)}$ and $G_1^{(j)} = \psi_1^{(j)}$. All the bond dimensions here are 1, which is the least possible. As mentioned earlier, the bond dimensions are positively correlated with the entanglement. So, for separable states, we get the lowest possible bond dimension. When the scalars in the product state decomposition are replaced with matrices, we get an MPS decomposition.

8.6 Additionally Required Lemmas

Lemma 4. *For any unitary V, we have*

$$\int_U \eta(UV)\, dU = \int_U \eta(VU)\, dU = \int_U \eta(U)\, dU \tag{8.21}$$

for any integrable functional η where dU is the Haar measure.

Lemma 5. *Let $A, B, C, D \in \mathbb{C}^{N \times N}$ be arbitrary matrices. Then, we have*

$$\begin{aligned} \int_U tr(BA_U)tr(DC_U)\, dU = {} & \frac{1}{N^2-1} \left(tr(A)tr(B)tr(C)tr(D) + tr(AC)tr(BD) \right) \\ & - \frac{1}{N(N^2-1)} \left(tr(AC)tr(B)tr(D) + tr(A)tr(C)tr(BD) \right), \end{aligned} \tag{8.22}$$

where dU is the Haar measure.

Lemma 6. *Let $A, B \in \mathbb{C}^{N \times N}$ be arbitrary matrices. Then, we have*

$$\int_U tr(BA_U)\, dU = \frac{tr(A)\, tr(B)}{N},$$

where dU is the Haar measure.

Lemma 7. *[Mit+18] Parameter Shift Rule: Let $\sigma \in \mathbb{D}_n$ and let $O \in \mathbb{H}_n$. Then, for any ansatz $C(\boldsymbol{\theta}) = \prod_{p=1}^{m} e^{-i\theta_p H_p}$, where $\boldsymbol{\theta} = [\theta_1 \dots \theta_m]^T$ and $H_p \in \mathbb{H}_n \ \forall \ p$, we have*

$$\partial_{\theta_p} f_{\sigma,O}(\boldsymbol{\theta}) = \frac{f_{\sigma,O}(\boldsymbol{\theta}_{p+}) - f_{\sigma,O}(\boldsymbol{\theta}_{p-})}{2}, \tag{8.23}$$

where ∂_{θ_p} is the partial derivative with respect to θ_p, $\boldsymbol{\theta}_{p\pm} = [\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{p-1}, \boldsymbol{\theta}_p \pm \pi/2, \boldsymbol{\theta}_{p+1}, \dots, \boldsymbol{\theta}_m]^T$.

Lemma 8. *[Cer+21b] Let $\sigma \in \mathbb{D}_n$ and let $O \in \mathbb{H}_n$. For any ansatz $C(\boldsymbol{\theta}) = \prod_{p=1}^{t} U_p(\boldsymbol{\theta}_p)$ where $U_p(\boldsymbol{\theta}_p) = \prod_{q=1}^{m} e^{-i\theta_{pq} H_{pq}}$, where $\boldsymbol{\theta}_p = [\theta_1 \dots \theta_m]^T$, $H_{pq} \in \mathbb{H}_n$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_1 \oplus \dots \oplus \boldsymbol{\theta}_t$, and for any p, q, define*

$$U_p^{(L,q)}(\boldsymbol{\theta}_p) = \prod_{j=1}^{q-1} e^{-i\theta_{pj} H_{pj}}, \tag{8.24}$$

$$U_p^{(R,q)}(\boldsymbol{\theta}_p) = \prod_{j=q+1}^{m} e^{-i\theta_{pj} H_{pj}}. \tag{8.25}$$

Then, we have

$$\mathbb{E}_{\boldsymbol{\theta}} \left(\partial_{\theta_{pq}} f_{\sigma,O}(\boldsymbol{\theta}) \right) = 0, \tag{8.26}$$

where $U_1(\boldsymbol{\theta}_1), U_2(\boldsymbol{\theta}_2), \dots, U_{p-1}(\boldsymbol{\theta}_{p-1}), U_{p+1}(\boldsymbol{\theta}_{p+1}), \dots, U_t(\boldsymbol{\theta}_t)$ along with either $U_p^{(L,q)}$ or $U_p^{(R,q)}$ form unitary 2-designs.

Lemma 9. *Tracial Matrix Hölder's Inequality [Bau11]: For any $A, B \in \mathbb{C}^{t \times t}$, and any $1 \leqslant p, q \leqslant \infty$ such that $\frac{1}{p} + \frac{1}{q} = 1$, we have,*

$$\left| tr(A^\dagger B) \right| \leqslant \|A\|_p \|B\|_q. \tag{8.27}$$

Lemma 10. *Cauchy-Schwarz Inequality: Let x, y be vectors in some inner product space with inner product $\langle \ , \ \rangle$. Then, we have*

$$|\langle x, y \rangle|^2 \leqslant \langle x, x \rangle \langle y, y \rangle. \tag{8.28}$$

Lemma 11. *For a set of numbers $\lambda_0, \lambda_1, \ldots, \lambda_{D-1}$, where D is even, with $\lambda_i \in \{1, -1\}$ and $|\{\lambda_i = 1 \mid i = 0, \ldots, D-1\}| = D/2$, we have $\sum_{i,j=0, i\neq j}^{D-1} \lambda_i \lambda_j = -D$.*

Proof. Let $\mathbb{I} = \{(i,j) \mid i,j = 0, 1, 2, \ldots, D-1\}$. Define $\mathbb{I}_{\pm} = \{(i,j) \in \mathbb{I} \; | \lambda_i \lambda_j = \pm 1\}$. So $|\mathbb{I}_{\pm}| = D^2/2$. All $\frac{D^2}{2}$ elements in $\mathbb{I}_-$ should have $i \neq j$. But only $\frac{D^2}{2} - D$ elements in $\mathbb{I}_+$ can have $i \neq j$ since whenever $i = j$, $(i,j) \in \mathbb{I}_+$. Hence, $\sum_{i \neq j} \lambda_i \lambda_j = -D$. □

Lemma 12. *Let $A \in \mathbb{C}^{2^n \times 2^n}$. Then, we have,*

$$\int_U tr\left(Z_i A_{\mathbb{1}_{n-k} \otimes U}\right) dU = 0, \tag{8.29}$$

for any $i \in \{n-k+1, \ldots, n\}$, where dU is the Haar measure.

Proof. Let $A = \sum_{p,q=0}^{2^n-1} A_{pq} \, |p_{1:n-k}\rangle\langle q_{1:n-k}| \otimes |p_{n-k+1:n}\rangle\langle q_{n-k+1:n}|$. Then, we have

$$\int_U \mathrm{tr}\left(Z_i A_{\mathbb{1}_{n-k} \otimes U}\right) \mathrm{d}U = \sum_{p,q=0}^{2^n-1} A_{pq} \delta_{p_{1:n-k}, q_{1:n-k}} \int_U \mathrm{tr}\left(Z_i (|p_{n-k+1:n}\rangle\langle q_{n-k+1:n}|)_U\right) \mathrm{d}U = 0, \tag{8.30}$$

where the last equality follows from Lemma 6. □

Lemma 13. *Let $A \in \mathbb{C}^{2^n \times 2^n}$. Then, we have,*

$$\int_U tr(Z_i A_{\mathbb{1}_{n-k} \otimes U}) tr(Z_j A_{\mathbb{1}_{n-k} \otimes U}) dU = 0, \tag{8.31}$$

for any $i, j \in \{n-k+1, \ldots, n\}$ with $i \neq j$, where dU is the Haar measure.

Proof. Let $A = \sum_{p,q=0}^{2^n-1} A_{pq} \left|p_{1:n-k}\right\rangle\!\left\langle q_{1:n-k}\right| \otimes \left|p_{n-k+1:n}\right\rangle\!\left\langle q_{n-k+1:n}\right|$. Then, we have

$$\begin{aligned}
&\int_U \operatorname{tr}\left(Z_i A_{\mathbb{1}_{2^{n-k}} \otimes U}\right) \operatorname{tr}\left(Z_j A_{\mathbb{1}_{2^{n-k}} \otimes U}\right) \mathrm{d}U \\
&= \int_U \sum_{p,q,r,s=0}^{2^n-1} A_{pq} A_{rs} \delta_{p_{1:n-k},q_{1:n-k}} \delta_{r_{1:n-k},s_{1:n-k}} \times \operatorname{tr}(Z_i(\left|p_{n-k+1:n}\right\rangle\!\left\langle q_{n-k+1:n}\right|)_U) \\
&\qquad \times \operatorname{tr}(Z_j(\left|r_{n-k+1:n}\right\rangle\!\left\langle s_{n-k+1:n}\right|)_U) \mathrm{d}U \qquad (8.32) \\
&= 0, \qquad (8.33)
\end{aligned}$$

where the last equality follows from Lemma 5. □

Bibliography

[Aar07] Scott Aaronson. "The learnability of quantum states". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 463.2088 (Sept. 2007), pp. 3089–3114. DOI: 10.1098/rspa.2007.0113. URL: https://doi.org/10.1098%2Frspa.2007.0113.

[Aar18] Scott Aaronson. "Shadow Tomography of Quantum States". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. STOC 2018. Los Angeles, CA, USA: Association for Computing Machinery, 2018, pp. 325–338. ISBN: 9781450355599. DOI: 10.1145/3188745.3188802. URL: https://doi.org/10.1145/3188745.3188802.

[AG04] Scott Aaronson and Daniel Gottesman. "Improved simulation of stabilizer circuits". In: *Phys. Rev. A* 70 (5 Nov. 2004), p. 052328. DOI: 10.1103/PhysRevA.70.052328. URL: https://link.aps.org/doi/10.1103/PhysRevA.70.052328.

[ABG20] Afham, Afrad Basheer, and Sandeep Goyal. *Quantum k-nearest neighbor machine learning algorithm*. Mar. 2020.

[ABG07] Esma Aımeur, Gilles Brassard, and Sébastien Gambs. "Quantum Clustering Algorithms". In: *Proceedings of the 24th International Conference on Machine Learning*. ICML '07. Corvalis, Oregon, USA: Association for Computing Machinery, 2007, pp. 1–8. ISBN: 9781595937933. DOI: 10.1145/1273496.1273497. URL: https://doi.org/10.1145/1273496.1273497.

[AG17] Neena Aloysius and M. Geetha. "A review on deep convolutional neural networks". In: *2017 International Conference on Communication and Signal Processing (ICCSP)*. 2017, pp. 0588–0592. DOI: 10.1109/ICCSP.2017.8286426.

[AS04] Andris Ambainis and Adam Smith. "Small Pseudo-Random Families of Matrices: Derandomizing Approximate Quantum Encryption". In: *Lecture Notes in Computer Science* 3122 (May 2004). DOI: 10.1007/978-3-540-27821-4_23.

[Arr+21] Andrew Arrasmith et al. "Effect of barren plateaus on gradient-free optimization". In: *Quantum* 5 (Oct. 2021), p. 558. ISSN: 2521-327X. DOI: 10.22331/q-2021-10-05-558. URL: https://doi.org/10.22331/q-2021-10-05-558.

[Arr+22] Andrew Arrasmith et al. “Equivalence of quantum barren plateaus to cost concentration and narrow gorges”. In: *Quantum Science and Technology* 7.4 (Aug. 2022), p. 045015. ISSN: 2058-9565. DOI: 10.1088/2058-9565/ac7d06. URL: http://dx.doi.org/10.1088/2058-9565/ac7d06.

[Aru+19] Frank Arute et al. “Quantum Supremacy using a Programmable Superconducting Processor”. In: *Nature* 574 (2019), pp. 505–510. URL: https://www.nature.com/articles/s41586-019-1666-5.

[Ban+18] Eiichi Bannai et al. *Unitary t-groups*. Oct. 2018.

[Ban+19] Eiichi Bannai et al. “On the explicit constructions of certain unitary t-designs”. In: 52.49 (Nov. 2019), p. 495301. DOI: 10.1088/1751-8121/ab5009. URL: https://doi.org/10.1088/1751-8121/ab5009.

[Bar+95] Adriano Barenco et al. “Elementary gates for quantum computation”. In: *Physical Review A* 52.5 (Nov. 1995), pp. 3457–3467. ISSN: 1094-1622. DOI: 10.1103/physreva.52.3457. URL: http://dx.doi.org/10.1103/PhysRevA.52.3457.

[BM24] Thomas Barthel and Qiang Miao. *Absence of barren plateaus and scaling of gradients in the energy optimization of isometric tensor network states*. 2024. arXiv: 2304.00161 [quant-ph]. URL: https://arxiv.org/abs/2304.00161.

[Bas+23] Afrad Basheer et al. “Alternating Layered Variational Quantum Circuits Can Be Classically Optimized Efficiently Using Classical Shadows”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.6 (June 2023), pp. 6770–6778. DOI: 10.1609/aaai.v37i6.25830. URL: https://ojs.aaai.org/index.php/AAAI/article/view/25830.

[Bas+24a] Afrad Basheer et al. “Ansatz-Agnostic Exponential Resource Saving in Variational Quantum Algorithms Using Shallow Shadows”. In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*. Ed. by Kate Larson. Main Track. International Joint Conferences on Artificial Intelligence Organization, Aug. 2024, pp. 3706–3714. DOI: 10.24963/ijcai.2024/410. URL: https://doi.org/10.24963/ijcai.2024/410.

[Bas+24b] Afrad Basheer et al. *On the Trainability and Classical Simulability of Learning Matrix Product States Variationally*. 2024. arXiv: 2409.10055 [quant-ph]. URL: https://arxiv.org/abs/2409.10055.

[Bau+20] Bela Bauer et al. “Quantum Algorithms for Quantum Chemistry and Quantum Materials Science”. In: *Chemical Reviews* 120.22 (Oct. 2020), pp. 12685–12717. ISSN: 1520-6890. DOI: 10.1021/acs.chemrev.9b00829. URL: http://dx.doi.org/10.1021/acs.chemrev.9b00829.

[Bau11] Bernhard Baumgartner. *An inequality for the trace of matrix products, using absolute values*. 2011. arXiv: 1106.6189 [math-ph].

[Bec+24] Simon Becker et al. “Classical Shadow Tomography for Continuous Variables Quantum Systems”. In: *IEEE Transactions on Information Theory* 70.5 (2024), pp. 3427–3452. DOI: 10.1109/TIT.2024.3357972.

[Ben+19] Marcello Benedetti et al. "Parameterized quantum circuits as machine learning models". In: *Quantum Science and Technology* 4.4 (Nov. 2019), p. 043001. DOI: 10.1088/2058-9565/ab4eb5. URL: https://doi.org/10.1088/2058-9565/ab4eb5.

[Ben73] C. H. Bennett. "Logical Reversibility of Computation". In: *IBM Journal of Research and Development* 17.6 (1973), pp. 525–532. DOI: 10.1147/rd.176.0525.

[Ben+93] Charles H. Bennett et al. "Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels". In: *Phys. Rev. Lett.* 70 (13 Mar. 1993), pp. 1895–1899. DOI: 10.1103/PhysRevLett.70.1895. URL: https://link.aps.org/doi/10.1103/PhysRevLett.70.1895.

[Ber20] Ewout van den Berg. "A simple method for sampling random Clifford operators". In: *arXiv: Quantum Physics* (2020).

[BV97] Ethan Bernstein and Umesh Vazirani. "Quantum Complexity Theory". In: *SIAM Journal on Computing* 26.5 (1997), pp. 1411–1473. DOI: 10.1137/S0097539796300921. eprint: https://doi.org/10.1137/S0097539796300921. URL: https://doi.org/10.1137/S0097539796300921.

[Ber+23] Christian Bertoni et al. "Shallow shadows: Expectation estimation using low-depth random Clifford circuits". In: *arXiv:2209.12924* (2023). URL: https://arxiv.org/abs/2209.12924.

[Bér+24] Caterina Bérubé et al. "Proactive behavior in voice assistants: A systematic review and conceptual model". In: *Computers in Human Behavior Reports* 14 (2024), p. 100411. ISSN: 2451-9588. DOI: https://doi.org/10.1016/j.chbr.2024.100411. URL: https://www.sciencedirect.com/science/article/pii/S2451958824000447.

[Bia+16] Jacob Biamonte et al. "Quantum Machine Learning". In: *Nature* 549 (Nov. 2016). DOI: 10.1038/nature23474.

[BL08] Jacob D. Biamonte and Peter J. Love. "Realizable Hamiltonians for universal adiabatic quantum computers". In: *Physical Review A* 78.1 (July 2008). ISSN: 1094-1622. DOI: 10.1103/physreva.78.012352. URL: http://dx.doi.org/10.1103/PhysRevA.78.012352.

[Bla85] Charles E. Blair. "Problem Complexity and Method Efficiency in Optimization (A. S. Nemirovsky and D. B. Yudin)". In: *Siam Review* 27 (1985), pp. 264–265. URL: https://api.semanticscholar.org/CorpusID:122924498.

[Bob+13] J. Bobadilla et al. "Recommender systems survey". In: *Knowledge-Based Systems* 46 (2013), pp. 109–132. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2013.03.012. URL: https://www.sciencedirect.com/science/article/pii/S0950705113001044.

[Bou+19] Adam Bouland et al. "On the complexity and verification of quantum random circuit sampling". In: *Nature Physics* 15 (Feb. 2019). DOI: 10.1038/s41567-018-0318-2.

[BK22] Gregory Boyd and Bálint Koczor. "Training variational quantum circuits with CoVaR: covariance root finding with classical shadows". In: *arXiv:2204.08494* (2022). DOI: 10.48550/ARXIV.2204.08494. URL: https://arxiv.org/abs/2204.08494.

[Boy+99] P.O. Boykin et al. "On universal and fault-tolerant quantum computing: a novel basis and a new constructive proof of universality for Shor's basis". In: *40th Annual Symposium on Foundations of Computer Science (Cat. No.99CB37039)*. 1999, pp. 486–494. DOI: 10.1109/SFFCS.1999.814621.

[BS17] Fernando GSL Brandao and Krysta M Svore. "Quantum speed-ups for solving semidefinite programs". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 415–426.

[BM21] Sergey Bravyi and Dmitrii L. Maslov. "Hadamard-Free Circuits Expose the Structure of the Clifford Group". In: *IEEE Transactions on Information Theory* 67 (2021), pp. 4546–4563.

[Bri+98] H.-J. Briegel et al. "Quantum Repeaters: The Role of Imperfect Local Operations in Quantum Communication". In: *Phys. Rev. Lett.* 81 (26 Dec. 1998), pp. 5932–5935. DOI: 10.1103/PhysRevLett.81.5932. URL: https://link.aps.org/doi/10.1103/PhysRevLett.81.5932.

[BWV08] Winton G. Brown, Yaakov S. Weinstein, and Lorenza Viola. "Quantum pseudorandomness from cluster-state quantum computation". In: *Phys. Rev. A* 77 (4 Apr. 2008), p. 040303. DOI: 10.1103/PhysRevA.77.040303. URL: https://link.aps.org/doi/10.1103/PhysRevA.77.040303.

[Cer+20] M Cerezo et al. "Cost-Function-Dependent Barren Plateaus in Shallow Quantum Neural Networks". In: *ArXiv* abs/2001.00550 (2020).

[Cer+21a] M Cerezo et al. "Variational Quantum Algorithms". In: *ArXiv* abs/2012.09265 (2021).

[Cer+21b] M. Cerezo et al. "Cost function dependent barren plateaus in shallow parametrized quantum circuits". In: *Nature Communications* 12.1 (Mar. 2021). DOI: 10.1038/s41467-021-21728-w. URL: https://doi.org/10.1038%2Fs41467-021-21728-w.

[Cer+21c] M. Cerezo et al. "Variational quantum algorithms". In: *Nature Reviews Physics* 3.9 (Aug. 2021), pp. 625–644. DOI: 10.1038/s42254-021-00348-9. URL: https://doi.org/10.1038%2Fs42254-021-00348-9.

[Cer+22] M. Cerezo et al. "Variational quantum state eigensolver". In: *npj Quantum Information* 8.1 (Sept. 2022). ISSN: 2056-6387. DOI: 10.1038/s41534-022-00611-6. URL: http://dx.doi.org/10.1038/s41534-022-00611-6.

[Cer+23] M. Cerezo et al. "Does provable absence of barren plateaus imply classical simulability? Or, why we need to rethink variational quantum computing". In: *arXiv:2312.09121* (2023). URL: https://arxiv.org/abs/2312.09121.

[Cha+20] Shouvanik Chakrabarti et al. "Quantum algorithms and lower bounds for convex optimization". In: *Quantum* 4 (Jan. 2020), p. 221. ISSN: 2521-327X. DOI: 10.22331/q-2020-01-13-221. URL: https://doi.org/10.22331/q-2020-01-13-221.

[Cha+16] Garnet Kin-Lic Chan et al. "Matrix product operators, matrix product states, and ab initio density matrix renormalization group algorithms". In: *The Journal of Chemical Physics* 145.1 (July 2016), p. 014102. ISSN: 0021-9606. DOI: 10.1063/1.4955108. eprint: https://pubs.aip.org/aip/jcp/article-pdf/doi/10.1063/1.4955108/14000039/014102_1_online.pdf. URL: https://doi.org/10.1063/1.4955108.

[CGZ15] H. Chen, Y. Gao, and J. Zhang. "Quantum K-nearest neighbor algorithm". In: *Dongnan Daxue Xuebao (Ziran Kexue Ban)/Journal of Southeast University (Natural Science Edition)* 45 (July 2015), pp. 647–651. DOI: 10.3969/j.issn.1001-0505.2015.04.006.

[Che+23] El Amine Cherrat et al. "Quantum Deep Hedging". In: *Quantum* 7 (Nov. 2023), p. 1191. ISSN: 2521-327X. DOI: 10.22331/q-2023-11-29-1191. URL: https://doi.org/10.22331/q-2023-11-29-1191.

[Chi09] Andrew M. Childs. "Universal Computation by Quantum Walk". In: *Physical Review Letters* 102.18 (May 2009). ISSN: 1079-7114. DOI: 10.1103/physrevlett.102.180501. URL: http://dx.doi.org/10.1103/PhysRevLett.102.180501.

[Chi+03] Andrew M. Childs et al. "Exponential algorithmic speedup by a quantum walk". In: *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*. STOC03. ACM, June 2003. DOI: 10.1145/780542.780552. URL: http://dx.doi.org/10.1145/780542.780552.

[Cir+21] J. Ignacio Cirac et al. "Matrix product states and projected entangled pair states: Concepts, symmetries, theorems". In: *Rev. Mod. Phys.* 93 (4 Dec. 2021), p. 045003. DOI: 10.1103/RevModPhys.93.045003. URL: https://link.aps.org/doi/10.1103/RevModPhys.93.045003.

[Cle+97] Richard Cleve et al. "Quantum algorithms revisited". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 454 (Aug. 1997). DOI: 10.1098/rspa.1998.0164.

[Cle+15] Richard Cleve et al. "Near-linear constructions of exact unitary 2-designs". In: *Quantum Information and Computation* 16 (Jan. 2015). DOI: 10.26421/QIC16.9-10-1.

[Coe+20] Bob Coecke et al. *Foundations for Near-Term Quantum Natural Language Processing*. Dec. 2020.

[CŚ06] Benoıt Collins and Piotr Śniady. "Integration with respect to the Haar measure on unitary, orthogonal and symplectic group". In: *Communications in Mathematical Physics* 264.3 (2006), pp. 773–795.

[CCL19] Iris Cong, Soonwon Choi, and Mikhail D. Lukin. "Quantum convolutional neural networks". In: *Nature Physics* 15.12 (Aug. 2019), pp. 1273–1278. DOI: 10.1038/s41567-019-0648-8. URL: https://doi.org/10.1038%2Fs41567-019-0648-8.

[Cra+10] Marcus Cramer et al. "Efficient quantum state tomography". In: *Nature Communications* 1.1 (Dec. 2010). ISSN: 2041-1723. DOI: 10.1038/ncomms1147. URL: http://dx.doi.org/10.1038/ncomms1147.

[Ćwi+12] Piotr Ćwikliński et al. "Local random quantum circuits are approximate polynomial-designs: Numerical results". In: *Journal of Physics A: Mathematical and Theoretical* 46 (Dec. 2012). DOI: 10.1088/1751-8113/46/30/305301.

[Dan+09] Christoph Dankert et al. "Exact and approximate unitary 2-designs and their application to fidelity estimation". In: *Phys. Rev. A* 80 (1 July 2009), p. 012304. DOI: 10.1103/PhysRevA.80.012304. URL: https://link.aps.org/doi/10.1103/PhysRevA.80.012304.

[DN08] Peter Dayan and Yael Niv. "Reinforcement learning: The Good, The Bad and The Ugly". In: *Current Opinion in Neurobiology* 18.2 (2008). Cognitive neuroscience, pp. 185–196. ISSN: 0959-4388. DOI: https://doi.org/10.1016/j.conb.2008.08.003. URL: https://www.sciencedirect.com/science/article/pii/S0959438808000767.

[Den+88] J. S. Denker et al. "Neural network recognizer for hand-written zip code digits". In: *Proceedings of the 1st International Conference on Neural Information Processing Systems*. NIPS'88. Cambridge, MA, USA: MIT Press, 1988, pp. 323–331.

[DJ92] David Deutsch and Richard Jozsa. "Rapid solution of problems by quantum computation". In: *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences* 439 (1992), pp. 553–558.

[Dia+23] N. L. Diaz et al. *Showcasing a Barren Plateau Theory Beyond the Dynamical Lie Algebra*. 2023. arXiv: 2310.11505 [quant-ph]. URL: https://arxiv.org/abs/2310.11505.

[Dom12] Pedro Domingos. "A few useful things to know about machine learning". In: *Commun. ACM* 55.10 (Oct. 2012), pp. 78–87. ISSN: 0001-0782. DOI: 10.1145/2347736.2347755. URL: https://doi.org/10.1145/2347736.2347755.

[Dov+22] Matan Ben Dov et al. *Approximate encoding of quantum states using shallow circuits*. 2022. arXiv: 2207.00028 [quant-ph].

[DH99] Christoph Durr and Peter Hoyer. *A Quantum Algorithm for Finding the Minimum*. 1999. arXiv: quant-ph/9607014 [quant-ph]. URL: https://arxiv.org/abs/quant-ph/9607014.

[Dür+06] Christoph Dürr et al. "Quantum Query Complexity of Some Graph Problems". In: *SIAM Journal on Computing* 35.6 (2006), pp. 1310–1328. DOI: 10.1137/050644719. eprint: https://doi.org/10.1137/050644719. URL: https://doi.org/10.1137/050644719.

[Eke91] Artur Ekert. "Quantum cryptography based on Bell's theorem." In: *Physical review letters* 67 6 (1991), pp. 661–663. URL: https://api.semanticscholar.org/CorpusID:27683254.

[EP14] Pakhshan Espoukeh and Pouria Pedram. "Quantum teleportation through noisy channels with multi-qubit GHZ states". In: *Quantum Information Processing* 13.8 (June 2014), pp. 1789–1811. ISSN: 1573-1332. DOI: 10.1007/s11128-014-0766-2. URL: http://dx.doi.org/10.1007/s11128-014-0766-2.

[Est+17] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542 (2017), pp. 115–118. URL: https://api.semanticscholar.org/CorpusID:3767412.

[EV13] Glen Evenbly and Guifre Vidal. *Quantum Criticality with the Multi-scale Entanglement Renormalization Ansatz*. 2013. arXiv: 1109.5334 [quant-ph]. URL: https://arxiv.org/abs/1109.5334.

[FGG14] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. "A Quantum Approximate Optimization Algorithm". In: *arXiv:1411.4028* (2014). DOI: 10.48550/ARXIV.1411.4028. URL: https://arxiv.org/abs/1411.4028.

[FH19] Edward Farhi and Aram W Harrow. *Quantum Supremacy through the Quantum Approximate Optimization Algorithm*. 2019. arXiv: 1602.07674 [quant-ph]. URL: https://arxiv.org/abs/1602.07674.

[FV12] Andrew J. Ferris and Guifre Vidal. "Variational Monte Carlo with the multiscale entanglement renormalization ansatz". In: *Phys. Rev. B* 85 (16 Apr. 2012), p. 165147. DOI: 10.1103/PhysRevB.85.165147. URL: https://link.aps.org/doi/10.1103/PhysRevB.85.165147.

[FL11] Steven T. Flammia and Yi-Kai Liu. "Direct Fidelity Estimation from Few Pauli Measurements". In: *Phys. Rev. Lett.* 106 (23 June 2011), p. 230501. DOI: 10.1103/PhysRevLett.106.230501. URL: https://link.aps.org/doi/10.1103/PhysRevLett.106.230501.

[Fon+22] Enrico Fontana et al. "Efficient recovery of variational quantum algorithms landscapes using classical signal processing". In: *arXiv:2208.05958* (2022). DOI: 10.48550/ARXIV.2208.05958. URL: https://arxiv.org/abs/2208.05958.

[Fow+12] Austin G. Fowler et al. "Surface codes: Towards practical large-scale quantum computation". In: *Phys. Rev. A* 86 (3 Sept. 2012), p. 032324. DOI: 10.1103/PhysRevA.86.032324. URL: https://link.aps.org/doi/10.1103/PhysRevA.86.032324.

[FBK21] Daniel Stilck França, Fernando G.S L. Brandão, and Richard Kueng. "Fast and Robust Quantum State Tomography from Few Basis Measurements". en. In: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021. DOI: 10.4230/LIPICS.TQC.2021.7. URL: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.TQC.2021.7.

[FT02] Edward Fredkin and Tommaso Toffoli. "Conservative logic". In: *International Journal of Theoretical Physics* 21 (2002), pp. 219–253. URL: https://api.semanticscholar.org/CorpusID:37305161.

[FM22] Lucas Friedrich and Jonas Maziero. "Avoiding barren plateaus with classical deep neural networks". In: *Physical Review A* 106.4 (Oct. 2022). DOI: 10.1103/physreva.106.042433. URL: https://doi.org/10.1103%2Fphysreva.106.042433.

[GYW05] T Gao, F L Yan, and Z X Wang. "Deterministic secure direct communication using GHZ states and swapping quantum entanglement". In: *Journal of Physics A: Mathematical and General* 38.25 (June 2005), p. 5761. DOI: 10.1088/0305-4470/38/25/011. URL: https://dx.doi.org/10.1088/0305-4470/38/25/011.

[Gar+23] Roy J. Garcia et al. "Barren plateaus from learning scramblers with local cost functions". In: *Journal of High Energy Physics* 2023.1 (Jan. 2023). ISSN: 1029-8479. DOI: 10.1007/jhep01(2023)090. URL: http://dx.doi.org/10.1007/JHEP01(2023)090.

[Gar+20] Bryan T. Gard et al. "Efficient symmetry-preserving state preparation circuits for the variational quantum eigensolver algorithm". In: *npj Quantum Information* 6.1 (Jan. 2020). ISSN: 2056-6387. DOI: 10.1038/s41534-019-0240-1. URL: http://dx.doi.org/10.1038/s41534-019-0240-1.

[Ge+24] Yan Ge et al. *Quantum Circuit Synthesis and Compilation Optimization: Overview and Prospects*. 2024. arXiv: 2407.00736 [quant-ph]. URL: https://arxiv.org/abs/2407.00736.

[GLM08] Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. "Architectures for a quantum random access memory". In: *Phys. Rev. A* 78 (5 Nov. 2008), p. 052310. DOI: 10.1103/PhysRevA.78.052310. URL: https://link.aps.org/doi/10.1103/PhysRevA.78.052310.

[Goh+23] Matthew L. Goh et al. *Lie-algebraic classical simulations for variational quantum computing*. 2023. arXiv: 2308.01432 [quant-ph]. URL: https://arxiv.org/abs/2308.01432.

[GA22] Weiyuan Gong and Scott Aaronson. "Learning Distributions over Quantum Measurement Outcomes". In: *arXiv:2209.03007* (2022). DOI: 10.48550/ARXIV.2209.03007. URL: https://arxiv.org/abs/2209.03007.

[Gra+19a] Edward Grant et al. "An initialization strategy for addressing barren plateaus in parametrized quantum circuits". In: *Quantum* 3 (Dec. 2019), p. 214. DOI: 10.22331/q-2019-12-09-214. URL: https://doi.org/10.22331%2Fq-2019-12-09-214.

[Gra+19b] Edward Grant et al. "An initialization strategy for addressing barren plateaus in parametrized quantum circuits". In: *Quantum* 3 (Dec. 2019), p. 214. ISSN: 2521-327X. DOI: 10.22331/q-2019-12-09-214. URL: https://doi.org/10.22331/q-2019-12-09-214.

[GPS24] Daniel Grier, Hakop Pashayan, and Luke Schaeffer. "Sample-optimal classical shadows for pure states". In: *Quantum* 8 (June 2024), p. 1373. ISSN: 2521-327X. DOI: 10.22331/q-2024-06-17-1373. URL: https://doi.org/10.22331/q-2024-06-17-1373.

[GS18] David J. Griffiths and Darrell F. Schroeter. *Introduction to Quantum Mechanics*. 3rd ed. Cambridge University Press, 2018.

[Gri+23a] Harper Grimsley et al. "Adaptive, problem-tailored variational quantum eigensolver mitigates rough parameter landscapes and barren plateaus". In: *npj Quantum Information* 9 (Mar. 2023). DOI: 10.1038/s41534-023-00681-0.

[Gri+23b] Harper R. Grimsley et al. "Adaptive, problem-tailored variational quantum eigensolver mitigates rough parameter landscapes and barren plateaus". In: *npj Quantum Information* 9.1 (Mar. 2023). DOI: 10.1038/s41534-023-00681-0. URL: https://doi.org/10.1038%2Fs41534-023-00681-0.

[GAE07] D. Gross, K. Audenaert, and Jens Eisert. "Evenly distributed unitaries: On the structure of unitary designs". In: *Journal of Mathematical Physics* 48 (Oct. 2007), p. 052104. DOI: 10.1063/1.2716992.

[Gro11] David Gross. "Recovering Low-Rank Matrices From Few Coefficients in Any Basis". In: *IEEE Transactions on Information Theory* 57.3 (Mar. 2011), pp. 1548–1566. ISSN: 1557-9654. DOI: 10.1109/tit.2011.2104999. URL: http://dx.doi.org/10.1109/TIT.2011.2104999.

[Gro+10] David Gross et al. "Quantum State Tomography via Compressed Sensing". In: *Phys. Rev. Lett.* 105 (15 Oct. 2010), p. 150401. DOI: 10.1103/PhysRevLett.105.150401. URL: https://link.aps.org/doi/10.1103/PhysRevLett.105.150401.

[Gro96] Lov K. Grover. "A Fast Quantum Mechanical Algorithm for Database Search". In: *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*. STOC '96. Philadelphia, Pennsylvania, USA: Association for Computing Machinery, 1996, pp. 212–219. ISBN: 0897917855. DOI: 10.1145/237814.237866. URL: https://doi.org/10.1145/237814.237866.

[GT09] Otfried Gühne and Géza Tóth. "Entanglement detection". In: *Physics Reports* 474.1–6 (Apr. 2009), pp. 1–75. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2009.02.004. URL: http://dx.doi.org/10.1016/j.physrep.2009.02.004.

[Guţ+20] M Guţă et al. "Fast state tomography with optimal error bounds". In: *Journal of Physics A: Mathematical and Theoretical* 53.20 (Apr. 2020), p. 204001. DOI: 10.1088/1751-8121/ab8111. URL: https://doi.org/10.1088/1751-8121/ab8111.

[GC09] Thiago S. Guzella and Walmir M. Caminhas. "A review of machine learning approaches to Spam filtering". In: *Expert Systems with Applications* 36.7 (2009), pp. 10206–10222. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2009.02.037. URL: https://www.sciencedirect.com/science/article/pii/S095741740900181X.

[Haa+16] Jeongwan Haah et al. "Sample-optimal tomography of quantum states". In: *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*. STOC '16. Cambridge, MA, USA: Association for Computing Machinery, 2016, pp. 913–925. ISBN: 9781450341325. DOI: 10.1145/2897518.2897585. URL: https://doi.org/10.1145/2897518.2897585.

[Hae+16] Jutho Haegeman et al. "Unifying time evolution and optimization with matrix product states". In: *Phys. Rev. B* 94 (16 Oct. 2016), p. 165116. DOI: 10.1103/PhysRevB.94.165116. URL: https://link.aps.org/doi/10.1103/PhysRevB.94.165116.

[Haf+20] Jonas Haferkamp et al. *Quantum homeopathy works: Efficient unitary designs with a system-size independent number of non-Clifford gates*. 2020. arXiv: 2002.09524 [quant-ph].

[HL08] Aram Harrow and Richard Low. "Random Quantum Circuits are Approximate 2-designs". In: *Communications in Mathematical Physics* 291 (Feb. 2008). DOI: 10.1007/s00220-009-0873-6.

[HM18] Aram Harrow and Saeed Mehraban. *Approximate unitary t-designs by short random quantum circuits using nearest-neighbor and long-range gates*. Sept. 2018.

[HL09a] Aram W. Harrow and Richard A. Low. "Random Quantum Circuits are Approximate 2-designs". In: *Communications in Mathematical Physics* 291.1 (July 2009), pp. 257–302. ISSN: 1432-0916. DOI: 10.1007/s00220-009-0873-6. URL: http://dx.doi.org/10.1007/s00220-009-0873-6.

[HL09b] Aram Wettroth Harrow and Richard Andrew Low. "Efficient Quantum Tensor Product Expanders and k-Designs". In: *APPROX-RANDOM*. 2009.

[Hav+19] Vojtěch Havlíček et al. "Supervised learning with quantum-enhanced feature spaces". In: *Nature* 567.7747 (Mar. 2019), pp. 209–212. DOI: 10.1038/s41586-019-0980-2. URL: https://doi.org/10.1038%2Fs41586-019-0980-2.

[Hin+21] Marcel Hinsche et al. "Learnability of the output distributions of local quantum circuits". In: *arXiv:2110.05517* (2021). DOI: 10.48550/ARXIV.2110.05517. URL: https://arxiv.org/abs/2110.05517.

[HZ93] Geoffrey E. Hinton and Richard S. Zemel. "Autoencoders, Minimum Description Length and Helmholtz Free Energy". In: *Proceedings of the 6th International Conference on Neural Information Processing Systems*. NIPS'93. Denver, Colorado: Morgan Kaufmann Publishers Inc., 1993, pp. 3–10.

[Hoc98] Sepp Hochreiter. "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 06.02 (1998), pp. 107–116. DOI: 10.1142/S0218488598000094. eprint: https://doi.org/10.1142/S0218488598000094. URL: https://doi.org/10.1142/S0218488598000094.

[Hoe63] Wassily Hoeffding. "Probability Inequalities for Sums of Bounded Random Variables". In: *Journal of the American Statistical Association* 58.301 (1963), pp. 13–30. DOI: 10.1080/01621459.1963.10500830. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1963.10500830. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500830.

[Hor+09] Ryszard Horodecki et al. "Quantum entanglement". In: *Reviews of Modern Physics* 81.2 (June 2009), pp. 865–942. ISSN: 1539-0756. DOI: 10.1103/revmodphys.81.865. URL: http://dx.doi.org/10.1103/RevModPhys.81.865.

[Hoy18] Matthew B Hoy. "Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants". In: *Medical Reference Services Quarterly* 37 (2018), pp. 81–88. URL: https://api.semanticscholar.org/CorpusID:30809087.

[HKP20] Hsin-Yuan Huang, Richard Kueng, and John Preskill. "Predicting many properties of a quantum system from very few measurements". In: *Nature Physics* 16.10 (June 2020), pp. 1050–1057. DOI: 10.1038/s41567-020-0932-7. URL: https://doi.org/10.1038%2Fs41567-020-0932-7.

[Hua+21] Hsin-Yuan Huang et al. "Provably efficient machine learning for quantum many-body problems". In: *arXiv:2106.12627* (2021). DOI: 10.48550/ARXIV.2106.12627. URL: https://arxiv.org/abs/2106.12627.

[Hua+22] Hsin-Yuan Huang et al. "Provably efficient machine learning for quantum many-body problems". In: *Science* 377.6613 (2022), eabk3333. DOI: 10.1126/science.abk3333. eprint: https://www.science.org/doi/pdf/10.1126/science.abk3333. URL: https://www.science.org/doi/abs/10.1126/science.abk3333.

[HMS17] C. Hubig, I. P. McCulloch, and U. Schollwöck. "Generic construction of efficient matrix product operators". In: *Phys. Rev. B* 95 (3 Jan. 2017), p. 035129. DOI: 10.1103/PhysRevB.95.035129. URL: https://link.aps.org/doi/10.1103/PhysRevB.95.035129.

[Jai+11] Nitin Jain et al. "Device Calibration Impacts Security of Quantum Key Distribution". In: *Phys. Rev. Lett.* 107 (11 Sept. 2011), p. 110501. DOI: 10.1103/PhysRevLett.107.110501. URL: https://link.aps.org/doi/10.1103/PhysRevLett.107.110501.

[Jer+21] Sofiene Jerbi et al. "Quantum Enhancements for Deep Reinforcement Learning in Large Spaces". In: *PRX Quantum* 2 (Feb. 2021). DOI: 10.1103/PRXQuantum.2.010328.

[JVV86] Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. "Random generation of combinatorial structures from a uniform distribution". In: *Theoretical Computer Science* 43 (1986), pp. 169–188. ISSN: 0304-3975. DOI: https://doi.org/10.1016/0304-3975(86)90174-X. URL: https://www.sciencedirect.com/science/article/pii/030439758690174X.

[Jia+17] Fei Jiang et al. "Artificial intelligence in healthcare: past, present and future". In: *Stroke and Vascular Neurology* 2 (2017), pp. 230–243. URL: https://api.semanticscholar.org/CorpusID:3779750.

[Jna+24] Hamza Jnane et al. "Quantum Error Mitigated Classical Shadows". In: *PRX Quantum* 5 (1 Feb. 2024), p. 010324. DOI: 10.1103/PRXQuantum.5.010324. URL: https://link.aps.org/doi/10.1103/PRXQuantum.5.010324.

[Joz06] Richard Jozsa. *On the simulation of quantum circuits*. 2006. arXiv: quant-ph/0603163 [quant-ph].

[JM08] Richard Jozsa and Akimasa Miyake. "Matchgates and classical simulation of quantum circuits". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 464.2100 (July 2008), pp. 3089–3106. ISSN: 1471-2946. DOI: 10.1098/rspa.2008.0189. URL: http://dx.doi.org/10.1098/rspa.2008.0189.

[Kan+17] Abhinav Kandala et al. "Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets". In: *Nature* 549 (2017), pp. 242–246.

[Kar+21] Andrey Kardashin et al. "Numerical hardware-efficient variational quantum simulation of a soliton solution". In: *Phys. Rev. A* 104 (2 Aug. 2021), p. L020402. DOI: 10.1103/PhysRevA.104.L020402. URL: https://link.aps.org/doi/10.1103/PhysRevA.104.L020402.

[Kas+08] Ivan Kassal et al. "Polynomial-time quantum algorithm for the simulation of chemical dynamics". In: *Proceedings of the National Academy of Sciences* 105.48 (2008), pp. 18681–18686. DOI: 10.1073/pnas.0808245105. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.0808245105. URL: https://www.pnas.org/doi/abs/10.1073/pnas.0808245105.

[Kel+15] Sebastian Keller et al. "An efficient matrix product operator representation of the quantum chemical Hamiltonian". In: *The Journal of Chemical Physics* 143.24 (Dec. 2015), p. 244118. ISSN: 0021-9606. DOI: 10.1063/1.4939000. eprint: https://pubs.aip.org/aip/jcp/article-pdf/doi/10.1063/1.4939000/15513208/244118_1_online.pdf. URL: https://doi.org/10.1063/1.4939000.

[Kem03] J Kempe. "Quantum random walks: An introductory overview". In: *Contemporary Physics* 44.4 (July 2003), pp. 307–327. ISSN: 1366-5812. DOI: 10.1080/0010715103100011077 6. URL: http://dx.doi.org/10.1080/00107151031000110776.

[Ker+19] Iordanis Kerenidis et al. "q-means: A quantum algorithm for unsupervised machine learning". In: *Advances in neural information processing systems* 32 (2019).

[Kha+19] Sumeet Khatri et al. "Quantum-assisted quantum compiling". In: *Quantum* 3 (May 2019), p. 140. DOI: 10.22331/q-2019-05-13-140. URL: https://doi.org/10.22331%2Fq-2019-05-13-140.

[KB17] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG]. URL: https://arxiv.org/abs/1412.6980.

[Kit03] A.Yu. Kitaev. "Fault-tolerant quantum computation by anyons". In: *Annals of Physics* 303.1 (Jan. 2003), pp. 2–30. ISSN: 0003-4916. DOI: 10.1016/s0003-4916(02)00018-0. URL: http://dx.doi.org/10.1016/S0003-4916(02)00018-0.

[KGE14] M. Kliesch, D. Gross, and J. Eisert. "Matrix-Product Operators and States: NP-Hardness and Undecidability". In: *Phys. Rev. Lett.* 113 (16 Oct. 2014), p. 160503. DOI: 10.1103/PhysRevLett.113.160503. URL: https://link.aps.org/doi/10.1103/PhysRevLett.113.160503.

[KL97] Emanuel Knill and Raymond Laflamme. "Theory of quantum error-correcting codes". In: *Phys. Rev. A* 55 (2 Feb. 1997), pp. 900–911. DOI: 10.1103/PhysRevA.55.900. URL: https://link.aps.org/doi/10.1103/PhysRevA.55.900.

[Knu97] Donald Ervin Knuth. *The Art of Computer Programming*. Vol. 1. Addison-Wesley Professional, 1997.

[KB22] Bálint Koczor and Simon C. Benjamin. "Quantum analytic descent". In: *Physical Review Research* 4.2 (Apr. 2022). ISSN: 2643-1564. DOI: 10.1103/physrevresearch.4.023017. URL: http://dx.doi.org/10.1103/PhysRevResearch.4.023017.

[KB09] Tamara G. Kolda and Brett W. Bader. "Tensor Decompositions and Applications". In: *SIAM Rev.* 51 (2009), pp. 455–500.

[KBV09] Yehuda Koren, Robert Bell, and Chris Volinsky. "Matrix Factorization Techniques for Recommender Systems". In: *Computer* 42.8 (2009), pp. 30–37. DOI: 10.1109/MC.2009.263.

[Kou+04] Yufeng Kou et al. "Survey of fraud detection techniques". In: *IEEE International Conference on Networking, Sensing and Control, 2004*. Vol. 2. 2004, 749–754 Vol.2. DOI: 10.1109/ICNSC.2004.1297040.

[KS22] Ankit Kulshrestha and Ilya Safro. "BEINIT: Avoiding Barren Plateaus in Variational Quantum Algorithms". In: *arXiv:2204.13751* (2022). URL: https://arxiv.org/abs/2204.13751.

[Kun+23] Jonathan Kunjummen et al. "Shadow process tomography of quantum channels". In: *Phys. Rev. A* 107 (4 Apr. 2023), p. 042403. DOI: 10.1103/PhysRevA.107.042403. URL: https://link.aps.org/doi/10.1103/PhysRevA.107.042403.

[Lam+18] L Lamata et al. "Quantum autoencoders via quantum adders with genetic algorithms". In: *Quantum Science and Technology* 4.1 (Oct. 2018), p. 014007. DOI: 10.1088/2058-9565/aae22b. URL: https://doi.org/10.1088/2058-9565/aae22b.

[Lar+22] Martin Larocca et al. "Diagnosing Barren Plateaus with Tools from Quantum Optimal Control". In: *Quantum* 6 (Sept. 2022), p. 824. ISSN: 2521-327X. DOI: 10.22331/q-2022-09-29-824. URL: http://dx.doi.org/10.22331/q-2022-09-29-824.

[Lar+24] Martin Larocca et al. *A Review of Barren Plateaus in Variational Quantum Computing*. 2024. arXiv: 2405.00781 [quant-ph].

[Leo+24] Lorenzo Leone et al. "On the practical usefulness of the Hardware Efficient Ansatz". In: *Quantum* 8 (July 2024), p. 1395. ISSN: 2521-327X. DOI: 10.22331/q-2024-07-03-1395. URL: https://doi.org/10.22331/q-2024-07-03-1395.

[LeV07] Randall J. LeVeque. *Finite Difference Methods for Ordinary and Partial Differential Equations*. Society for Industrial and Applied Mathematics, 2007. DOI: 10.1137/1.9780898717839. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9780898717839. URL: https://epubs.siam.org/doi/abs/10.1137/1.9780898717839.

[Li+19] Ao Li et al. "Spam Review Detection with Graph Convolutional Networks". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. CIKM '19. Beijing, China: Association for Computing Machinery, 2019, pp. 2703–2711. ISBN: 9781450369763. DOI: 10.1145/3357384.3357820. URL: https://doi.org/10.1145/3357384.3357820.

[LWZ14] Dong-fen Li, Rui-jin Wang, and Feng Zhang. "Quantum Information Splitting of Arbitrary Three-Qubit State by Using Four-Qubit Cluster State and GHZ-State". In: *International Journal of Theoretical Physics* 54 (2014), pp. 1142–1153. URL: https://api.semanticscholar.org/CorpusID:120488883.

[LSW21a] Guangxi Li, Zhixin Song, and Xin Wang. "VSQL: Variational Shadow Quantum Learning for Classification". In: *AAAI*. 2021.

[LSW21b] Guangxi Li, Zhixin Song, and Xin Wang. "VSQL: Variational Shadow Quantum Learning for Classification". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.9 (May 2021), pp. 8357–8365. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17016.

[LZW23] Guangxi Li, Xuanqiang Zhao, and Xin Wang. *Quantum Self-Attention Neural Networks for Text Classification*. 2023. arXiv: 2205.05625 [quant-ph]. URL: https://arxiv.org/abs/2205.05625.

[Li+13] H W Li et al. "Calibration and high fidelity measurement of a quantum photonic chip". In: *New Journal of Physics* 15.6 (June 2013), p. 063017. ISSN: 1367-2630. DOI: 10.1088/1367-2630/15/6/063017. URL: http://dx.doi.org/10.1088/1367-2630/15/6/063017.

[Li+22] Zewen Li et al. "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects". In: *IEEE Transactions on Neural Networks and Learning Systems* 33.12 (2022), pp. 6999–7019. DOI: 10.1109/TNNLS.2021.3084827.

[LF24] Yidong Liao and Chris Ferrie. *GPT on a Quantum Computer*. 2024. arXiv: 2403.09418 [quant-ph]. URL: https://arxiv.org/abs/2403.09418.

[Lin+21] Sheng-Hsuan Lin et al. "Real- and Imaginary-Time Evolution with Compressed Quantum Circuits". In: *PRX Quantum* 2.1 (Mar. 2021). ISSN: 2691-3399. DOI: 10.1103/prxquantum.2.010342. URL: http://dx.doi.org/10.1103/PRXQuantum.2.010342.

[Liu+22] Zidu Liu et al. "Presence and Absence of Barren Plateaus in Tensor-Network Based Machine Learning". In: *Physical Review Letters* 129.27 (Dec. 2022). ISSN: 1079-7114. DOI: 10.1103/physrevlett.129.270501. URL: http://dx.doi.org/10.1103/PhysRevLett.129.270501.

[LGZ14] Seth Lloyd, Silvano Garnerone, and Paolo Zanardi. "Quantum algorithms for topological and geometric analysis of big data". In: *Nature Communications* (Aug. 2014). DOI: 10.1038/ncomms10138.

[LCB14] Michael Lubasch, J. Ignacio Cirac, and Mari-Carmen Bañuls. "Algorithms for finite projected entangled pair states". In: *Phys. Rev. B* 90 (6 Aug. 2014), p. 064425. DOI: 10.1103/PhysRevB.90.064425. URL: https://link.aps.org/doi/10.1103/PhysRevB.90.064425.

[LBR17] A. P. Lund, Michael J. Bremner, and T. C. Ralph. "Quantum sampling problems, BosonSampling and quantum supremacy". In: *npj Quantum Information* 3.1 (Apr. 2017). ISSN: 2056-6387. DOI: 10.1038/s41534-017-0018-2. URL: http://dx.doi.org/10.1038/s41534-017-0018-2.

[MWT20] Yunpu Ma, Yuyi Wang, and Volker Tresp. *Quantum Machine Learning Algorithm for Knowledge Graphs*. Jan. 2020.

[Mad+22] Lars Skovgaard Madsen et al. "Quantum computational advantage with a programmable photonic processor". In: *Nature* 606 (2022), pp. 75–81. URL: https://api.semanticscholar.org/CorpusID:249276257.

[MP15] Ilya Sinayskiy Maria Schuld and Francesco Petruccione. "An introduction to quantum machine learning". In: *Contemporary Physics* 56.2 (2015), pp. 172–185. DOI: 10.1080/00107514.2014.964942. eprint: https://doi.org/10.1080/00107514.2014.964942. URL: https://doi.org/10.1080/00107514.2014.964942.

[MJP21] Gabriel Matos, Sonika Johri, and Zlatko Papi ć. "Quantifying the Efficiency of State Preparation via Quantum Variational Eigensolvers". In: *PRX Quantum* 2 (1 Jan. 2021), p. 010309. DOI: 10.1103/PRXQuantum.2.010309. URL: https://link.aps.org/doi/10.1103/PRXQuantum.2.010309.

[McC+18] Jarrod R. McClean et al. "Barren plateaus in quantum neural network training landscapes". In: *Nature Communications* 9.1 (Nov. 2018). DOI: 10.1038/s41467-018-07090-4. URL: https://doi.org/10.1038%2Fs41467-018-07090-4.

[MCG16] Michael F. McTear, Zoraida Callejas, and David Griol. "The Conversational Interface: Talking to Smart Devices". In: 2016. URL: https://api.semanticscholar.org/CorpusID:57914840.

[Mei+20] Konstantinos Meichanetzidis et al. *Quantum Natural Language Processing on Near-Term Quantum Computers*. May 2020.

[Mel+22] Antonio A. Mele et al. "Avoiding barren plateaus via transferability of smooth solutions in a Hamiltonian variational ansatz". In: *Physical Review A* 106.6 (Dec. 2022). DOI: 10.1103/physreva.106.l060401. URL: https://doi.org/10.1103%2Fphysreva.106.l060401.

[MOV96] Alfred Menezes, Paul C. van Oorschot, and Scott A. Vanstone. "Handbook of Applied Cryptography". In: 1996.

[Mez06] Francesco Mezzadri. "How to generate random matrices from the classical compact groups". In: *Notices of the American Mathematical Society* 54 (Oct. 2006).

[MV18] Ashley Milsted and Guifre Vidal. "Geometric interpretation of the multi-scale entanglement renormalization ansatz". In: *arXiv preprint arXiv:1812.00529* (2018).

[Mit+18] K. Mitarai et al. "Quantum circuit learning". In: *Phys. Rev. A* 98 (3 Sept. 2018), p. 032309. DOI: 10.1103/PhysRevA.98.032309. URL: https://link.aps.org/doi/10.1103/PhysRevA.98.032309.

[Mit97] Tom Mitchell. *Machine learning*. Vol. 1. 9. 1997. URL: https://www.cs.cmu.edu/~tom/mlbook.html.

[MB18] Leonard Mlodinow and Todd A. Brun. "Discrete spacetime, quantum walks, and relativistic wave equations". In: *Phys. Rev. A* 97 (4 Apr. 2018), p. 042131. DOI: 10.1103/PhysRevA.97.042131. URL: https://link.aps.org/doi/10.1103/PhysRevA.97.042131.

[Mon+23] Léo Monbroussou et al. *Trainability and Expressivity of Hamming-Weight Preserving Quantum Circuits for Machine Learning*. 2023. arXiv: 2309.15547 [quant-ph]. URL: https://arxiv.org/abs/2309.15547.

[Mur+10] V. Murg et al. "Simulating strongly correlated quantum systems with tree tensor networks". In: *Phys. Rev. B* 82 (20 Nov. 2010), p. 205105. DOI: 10.1103/PhysRevB.82.205105. URL: https://link.aps.org/doi/10.1103/PhysRevB.82.205105.

[NY21] Kouhei Nakaji and Naoki Yamamoto. "Expressibility of the alternating layered ansatz for quantum computation". In: *Quantum* 5 (Apr. 2021), p. 434. DOI: 10.22331/q-2021-04-19-434. URL: https://doi.org/10.22331%2Fq-2021-04-19-434.

[Nak+17] Yoshifumi Nakata et al. "Unitary 2-designs from random X - and Z -diagonal unitaries". In: *Journal of Mathematical Physics* 58 (May 2017), p. 052203. DOI: 10.1063/1.4983266.

[Nak+21] Yoshifumi Nakata et al. "Quantum Circuits for Exact Unitary t-Designs and Applications to Higher-Order Randomized Benchmarking". In: *PRX Quantum* 2 (3 Sept. 2021), p. 030339. DOI: 10.1103/PRXQuantum.2.030339. URL: https://link.aps.org/doi/10.1103/PRXQuantum.2.030339.

[NC13] Naoki Nakatani and Garnet Kin Chan. "Efficient tree tensor network states (TTNS) for quantum chemistry: Generalizations of the density matrix renormalization group algorithm". In: *The Journal of chemical physics* 138.13 (2013).

[Nga+11] E.W.T. Ngai et al. "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature". In: *Decision Support Systems* 50.3 (2011). On quantitative methods for detection of financial fraud, pp. 559–569. ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2010.08.006. URL: https://www.sciencedirect.com/science/article/pii/S0167923610001302.

[Nic13] Maximilian Nickel. "Tensor factorization for relational learning". Aug. 2013. URL: http://nbn-resolving.de/urn:nbn:de:bvb:19-160568.

[NLW13] Yi-You Nie, Yuan-Hua Li, and Zi-Sheng Wang. "Semi-quantum information splitting using GHZ-type states". In: *Quantum Information Processing* 12.1 (Jan. 2013), pp. 437–448. ISSN: 1570-0755. DOI: 10.1007/s11128-012-0388-5. URL: https://doi.org/10.1007/s11128-012-0388-5.

[NC11] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. 10th. USA: Cambridge University Press, 2011. ISBN: 1107002176.

[OW16] Ryan O'Donnell and John Wright. "Efficient quantum tomography". In: *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*. STOC '16. Cambridge, MA, USA: Association for Computing Machinery, 2016, pp. 899–912. ISBN: 9781450341325. DOI: 10.1145/2897518.2897544. URL: https://doi.org/10.1145/2897518.2897544.

[Oka+22] Ken N. Okada et al. "Identification of topological phases using classically-optimized variational quantum eigensolver". In: *arXiv:2202.02909* (2022). DOI: 10.48550/ARXIV.2202.02909. URL: https://arxiv.org/abs/2202.02909.

[OKW21] Carlos Ortiz Marrero, Mária Kieferová, and Nathan Wiebe. "Entanglement-Induced Barren Plateaus". In: *PRX Quantum* 2 (4 Oct. 2021), p. 040316. DOI: 10.1103/PRXQuantum.2.040316. URL: https://link.aps.org/doi/10.1103/PRXQuantum.2.040316.

[Oru13] Roman Orus. "A Practical Introduction to Tensor Networks: Matrix Product States and Projected Entangled Pair States". In: *Annals of Physics* 349 (June 2013). DOI: 10.1016/j.aop.2014.06.013.

[Ose11] I. V. Oseledets. "Tensor-Train Decomposition". In: *SIAM Journal on Scientific Computing* 33.5 (2011), pp. 2295–2317. DOI: 10.1137/090752286. eprint: https://doi.org/10.1137/090752286. URL: https://doi.org/10.1137/090752286.

[Pai+21] Marco Paini et al. "Estimating expectation values using approximate quantum states". In: *Quantum* 5 (Mar. 2021), p. 413. ISSN: 2521-327X. DOI: 10.22331/q-2021-03-16-413. URL: https://doi.org/10.22331/q-2021-03-16-413.

[PKH24] Chae-Yeun Park, Minhyeok Kang, and Joonsuk Huh. *Hardware-efficient ansatz without barren plateaus in any depth*. 2024. arXiv: 2403.04844 [quant-ph]. URL: https://arxiv.org/abs/2403.04844.

[Par05] K. R. Parthasarathy. "Lectures on Quantum Computation, Quantum Error Correcting Codes And Information Theory". In: 2005.

[Pat+21] Taylor L. Patti et al. "Entanglement devised barren plateau mitigation". In: *Physical Review Research* 3.3 (July 2021). DOI: 10.1103/physrevresearch.3.033090. URL: https://doi.org/10.1103%2Fphysrevresearch.3.033090.

[PTP19] Alex Pepper, Nora Tischler, and Geoff J. Pryde. "Experimental Realization of a Quantum Autoencoder: The Compression of Qutrits via Machine Learning". In: *Phys. Rev. Lett.* 122 (6 Feb. 2019), p. 060501. DOI: 10.1103/PhysRevLett.122.060501. URL: https://link.aps.org/doi/10.1103/PhysRevLett.122.060501.

[Per+07] D. Perez-Garcia et al. *Matrix Product State Representations*. 2007. arXiv: quant-ph/0608197 [quant-ph].

[Per+14] Alberto Peruzzo et al. "A variational eigenvalue solver on a photonic quantum processor". In: *Nature Communications* 5.1 (July 2014). DOI: 10.1038/ncomms5213. URL: https://doi.org/10.1038%2Fncomms5213.

[Pes+21] Arthur Pesah et al. "Absence of Barren Plateaus in Quantum Convolutional Neural Networks". In: *Physical Review X* 11.4 (Oct. 2021). DOI: 10.1103/physrevx.11.041011. URL: https://doi.org/10.1103%2Fphysrevx.11.041011.

[Phu+10] Clifton Phua et al. "A Comprehensive Survey of Data Mining-based Fraud Detection Research". In: *CoRR* abs/1009.6119 (Sept. 2010).

[Pir+10] B Pirvu et al. "Matrix product operator representations". In: *New Journal of Physics* 12.2 (Feb. 2010), p. 025012. DOI: 10.1088/1367-2630/12/2/025012. URL: https://dx.doi.org/10.1088/1367-2630/12/2/025012.

[Pow64] M. J. D. Powell. "An efficient method for finding the minimum of a function of several variables without calculating derivatives". In: *The Computer Journal* 7.2 (Jan. 1964), pp. 155–162. ISSN: 0010-4620. DOI: 10.1093/comjnl/7.2.155. eprint: https://academic.oup.com/comjnl/article-pdf/7/2/155/959784/070155.pdf. URL: https://doi.org/10.1093/comjnl/7.2.155.

[Pre18] John Preskill. "Quantum Computing in the NISQ era and beyond". In: *Quantum* 2 (Aug. 2018), p. 79. ISSN: 2521-327X. DOI: 10.22331/q-2018-08-06-79. URL: http://dx.doi.org/10.22331/q-2018-08-06-79.

[Qi+23] Han Qi et al. "The barren plateaus of quantum neural networks: review, taxonomy and trends". In: *Quantum Information Processing* 22 (2023), pp. 1–26. URL: https://api.semanticscholar.org/CorpusID:266164696.

[RSL22] Ali Rad, Alireza Seif, and Norbert M. Linke. "Surviving The Barren Plateau in Variational Quantum Circuits with Bayesian Learning Initialization". In: *arXiv:2203.02464* (2022). URL: https://arxiv.org/abs/2203.02464.

[RRS06] Jaikumar Radhakrishnan, Martin Rötteler, and Pranab Kumar Sen. "Random Measurement Bases, Quantum State Distinction and Applications to the Hidden Subgroup Problem". In: *Algorithmica* 55 (2006), pp. 490–516.

[Ran20] Shi-Ju Ran. "Encoding of matrix product states into quantum circuits of one- and two-qubit gates". In: *Phys. Rev. A* 101 (3 Mar. 2020), p. 032310. DOI: 10.1103/PhysRevA.101.032310. URL: https://link.aps.org/doi/10.1103/PhysRevA.101.032310.

[RKR23] Hans-Martin Rieser, Frank Köster, and Arne Peter Raulf. "Tensor networks for quantum machine learning". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 479.2275 (July 2023). ISSN: 1471-2946. DOI: 10.1098/rspa.2023.0218. URL: http://dx.doi.org/10.1098/rspa.2023.0218.

[ROA17] Jonathan Romero, Jonathan P Olson, and Alan Aspuru-Guzik. "Quantum autoencoders for efficient compression of quantum data". In: *Quantum Science and Technology* 2.4 (Aug. 2017), p. 045001. DOI: 10.1088/2058-9565/aa8072. URL: https://doi.org/10.1088/2058-9565/aa8072.

[RS09] Aidan Roy and A. J. Scott. "Unitary designs and codes". In: *Designs, Codes and Cryptography* 53 (2009), pp. 13–31.

[Rua+17] Yue Ruan et al. "Quantum Algorithm for K-Nearest Neighbors Classification Based on the Metric of Hamming Distance". In: *International Journal of Theoretical Physics* 56 (Nov. 2017). DOI: 10.1007/s10773-017-3514-4.

[Rud17] Sebastian Ruder. *An overview of gradient descent optimization algorithms*. 2017. arXiv: 1609.04747 [cs.LG]. URL: https://arxiv.org/abs/1609.04747.

[Rud+22] Manuel S. Rudolph et al. *Decomposition of Matrix Product States into Shallow Quantum Circuits*. 2022. arXiv: 2209.00595 [quant-ph].

[Rud+23] Manuel S. Rudolph et al. *Trainability barriers and opportunities in quantum generative modeling*. 2023. arXiv: 2305.02881 [quant-ph]. URL: https://arxiv.org/abs/2305.02881.

[Sac11] Subir Sachdev. *Quantum Phase Transitions*. 2nd ed. Cambridge University Press, 2011.

[Sac+22] Stefan H. Sack et al. "Avoiding Barren Plateaus Using Classical Shadows". In: *PRX Quantum* 3.2 (June 2022). DOI: 10.1103/prxquantum.3.020365. URL: https://doi.org/10.1103%2Fprxquantum.3.020365.

[Sak+03] Georgios Sakkis et al. "A Memory-Based Approach to Anti-Spam Filtering for Mailing Lists". In: *Inf. Retr.* 6.1 (Jan. 2003), pp. 49–73. ISSN: 1386-4564. DOI: 10.1023/A:1022948414856. URL: https://doi.org/10.1023/A:1022948414856.

[SN20] J. J. Sakurai and Jim Napolitano. *Modern Quantum Mechanics*. 3rd ed. Cambridge University Press, 2020.

[Sau+19] David Sauerwein et al. "Matrix Product States: Entanglement, Symmetries, and State Transformations". In: *Phys. Rev. Lett.* 123 (17 Oct. 2019), p. 170504. DOI: 10.1103/PhysRevLett.123.170504. URL: https://link.aps.org/doi/10.1103/PhysRevLett.123.170504.

[Sch+24] Louis Schatzki et al. "Theoretical guarantees for permutation-equivariant quantum neural networks". In: *npj Quantum Information* 10.1 (Jan. 2024). ISSN: 2056-6387. DOI: 10.1038/s41534-024-00804-1. URL: http://dx.doi.org/10.1038/s41534-024-00804-1.

[Sch11] Ulrich Schollwöck. "The density-matrix renormalization group in the age of matrix product states". In: *Annals of Physics* 326.1 (Jan. 2011), pp. 96–192. ISSN: 0003-4916. DOI: 10.1016/j.aop.2010.09.012. URL: http://dx.doi.org/10.1016/j.aop.2010.09.012.

[SEM22] Franz J. Schreiber, Jens Eisert, and Johannes Jakob Meyer. "Classical surrogates for quantum learning models". In: *arXiv:2206.11740* (2022). URL: https://arxiv.org/abs/2206.11740.

[SPC11] Norbert Schuch, David Pérez-García, and Ignacio Cirac. "Classifying quantum phases using matrix product states and projected entangled pair states". In: *Physical Review B* 84.16 (Oct. 2011). ISSN: 1550-235X. DOI: 10.1103/physrevb.84.165139. URL: http://dx.doi.org/10.1103/PhysRevB.84.165139.

[Sch+12] Norbert Schuch et al. "Resonating valence bond states in the PEPS formalism". In: *Phys. Rev. B* 86 (11 Sept. 2012), p. 115108. DOI: 10.1103/PhysRevB.86.115108. URL: https://link.aps.org/doi/10.1103/PhysRevB.86.115108.

[Sch21] Maria Schuld. *Quantum machine learning models are kernel methods*. Jan. 2021.

[SSM21] Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. "Effect of data encoding on the expressive power of variational quantum-machine-learning models". In: *Physical Review A* 103.3 (Mar. 2021). DOI: 10.1103/physreva.103.032430. URL: https://doi.org/10.1103%2Fphysreva.103.032430.

[Sch+13] Martin Schwarz et al. "Preparing topological projected entangled pair states on a quantum computer". In: *Physical Review A* 88.3 (Sept. 2013). ISSN: 1094-1622. DOI: 10.1103/physreva.88.032321. URL: http://dx.doi.org/10.1103/PhysRevA.88.032321.

[Sep06] Mark Roger Sepanski. "Compact Lie Groups". In: 2006. URL: https://api.semanticscholar.org/CorpusID:117022255.

[SZ84] P.D Seymour and Thomas Zaslavsky. "Averaging sets: A generalization of mean values and spherical designs". In: *Advances in Mathematics* 52.3 (1984), pp. 213–240. ISSN: 0001-8708. DOI: https://doi.org/10.1016/0001-8708(84)90022-7. URL: https://www.sciencedirect.com/science/article/pii/0001870884900227.

[Sha05] A. Shadrin. "Twelve Proofs of the Markov Inequality". In: 2005. URL: https://api.semanticscholar.org/CorpusID:17028844.

[Sha11] R. Shankar. *Principles of Quantum Mechanics*. 2nd ed. Springer, 2011. ISBN: 978-1-4419-1309-3.

[SBM06] V.V. Shende, S.S. Bullock, and I.L. Markov. "Synthesis of quantum-logic circuits". In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 25.6 (June 2006), pp. 1000–1010. ISSN: 1937-4151. DOI: 10.1109/tcad.2005.855930. URL: http://dx.doi.org/10.1109/TCAD.2005.855930.

[SDV06] Y.-Y. Shi, L.-M. Duan, and G. Vidal. "Classical simulation of quantum many-body systems with a tree tensor network". In: *Phys. Rev. A* 74 (2 Aug. 2006), p. 022320. DOI: 10.1103/PhysRevA.74.022320. URL: https://link.aps.org/doi/10.1103/PhysRevA.74.022320.

[Sho94] P.W. Shor. "Algorithms for quantum computation: discrete logarithms and factoring". In: *Proceedings 35th Annual Symposium on Foundations of Computer Science*. 1994, pp. 124–134. DOI: 10.1109/SFCS.1994.365700.

[Sim97] Daniel R. Simon. "On the Power of Quantum Computation". In: *SIAM Journal on Computing* 26.5 (1997), pp. 1474–1483. DOI: 10.1137/S0097539796298637. eprint: https://doi.org/10.1137/S0097539796298637. URL: https://doi.org/10.1137/S0097539796298637.

[Sko+21] Andrea Skolik et al. "Layerwise learning for quantum neural networks". In: *Quantum Machine Intelligence* 3.1 (Jan. 2021). DOI: 10.1007/s42484-020-00036-4. URL: https://doi.org/10.1007%2Fs42484-020-00036-4.

[SVC22] Lucas Slattery, Benjamin Villalonga, and Bryan K. Clark. "Unitary block optimization for variational quantum algorithms". In: *Phys. Rev. Research* 4 (2 Apr. 2022), p. 023072. DOI: 10.1103/PhysRevResearch.4.023072. URL: https://link.aps.org/doi/10.1103/PhysRevResearch.4.023072.

[Spa92] J.C. Spall. "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation". In: *IEEE Transactions on Automatic Control* 37.3 (1992), pp. 332–341. DOI: 10.1109/9.119632.

[Sto+20] James Stokes et al. "Quantum Natural Gradient". In: *Quantum* 4 (May 2020), p. 269. ISSN: 2521-327X. DOI: 10.22331/q-2020-05-25-269. URL: http://dx.doi.org/10.22331/q-2020-05-25-269.

[SO82] Attila Szabó and Neil S. Ostlund. "Modern quantum chemistry : introduction to advanced electronic structure theory". In: 1982. URL: https://api.semanticscholar.org/CorpusID:94743139.

[Sze+11] Oleg Szehr et al. "Decoupling with unitary approximate two-designs". In: *New Journal of Physics* 15 (Sept. 2011). DOI: 10.1088/1367-2630/15/5/053022.

[TEV09] L. Tagliacozzo, G. Evenbly, and G. Vidal. "Simulation of two-dimensional quantum systems using a tree tensor network that exploits the entropic area law". In: *Phys. Rev. B* 80 (23 Dec. 2009), p. 235127. DOI: 10.1103/PhysRevB.80.235127. URL: https://link.aps.org/doi/10.1103/PhysRevB.80.235127.

[Tan21] Ewin Tang. "Quantum Principal Component Analysis Only Achieves an Exponential Speedup Because of Its State Preparation Assumptions". In: *Physical Review Letters* 127.6 (Aug. 2021). DOI: 10.1103/physrevlett.127.060503. URL: https://doi.org/10.1103%2Fphysrevlett.127.060503.

[Til+22a] Jules Tilly et al. "The Variational Quantum Eigensolver: A review of methods and best practices". In: *Physics Reports* 986 (2022). The Variational Quantum Eigensolver: a review of methods and best practices, pp. 1–128. ISSN: 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2022.08.003. URL: https://www.sciencedirect.com/science/article/pii/S0370157322003118.

[Til+22b] Jules Tilly et al. "The variational quantum eigensolver: a review of methods and best practices". In: *Physics Reports* 986 (2022), pp. 1–128.

[Tof80] Tommaso Toffoli. "Reversible Computing". In: *International Colloquium on Automata, Languages and Programming*. 1980. URL: https://api.semanticscholar.org/CorpusID:11680687.

[Top19] Eric J. Topol. "High-performance medicine: the convergence of human and artificial intelligence". In: *Nature Medicine* 25 (2019), pp. 44–56. URL: https://api.semanticscholar.org/CorpusID:57574615.

[Vap00] Vladimir Vapnik. "The Nature of Statistical Learning Theory". In: vol. 8. Jan. 2000, pp. 1–15. ISBN: 978-1-4419-3160-3. DOI: 10.1007/978-1-4757-3264-1_1.

[VPB18] Guillaume Verdon, Jason Pye, and Michael Broughton. "A Universal Training Algorithm for Quantum Deep Learning". In: *arXiv:1806.09729* (2018). DOI: 10.48550/ARXIV.1806.09729. URL: https://arxiv.org/abs/1806.09729.

[Ver+19] Guillaume Verdon et al. "Learning to learn with quantum neural networks via classical neural networks". In: *arXiv:1907.05415* (2019). URL: https://arxiv.org/abs/1907.05415.

[Vid03] Guifré Vidal. "Efficient Classical Simulation of Slightly Entangled Quantum Computations". In: *Physical Review Letters* 91.14 (Oct. 2003). ISSN: 1079-7114. DOI: 10.1103/physrevlett.91.147902. URL: http://dx.doi.org/10.1103/PhysRevLett.91.147902.

[Wan+17] Kwok Ho Wan et al. "Quantum generalisation of feedforward neural networks". In: *npj Quantum Information* 3.1 (Sept. 2017). DOI: 10.1038/s41534-017-0032-4. URL: https://doi.org/10.1038%2Fs41534-017-0032-4.

[Wan+21] Samson Wang et al. "Noise-induced barren plateaus in variational quantum algorithms". In: *Nature Communications* 12.1 (Nov. 2021). ISSN: 2041-1723. DOI: 10.1038/s41467-021-27045-6. URL: http://dx.doi.org/10.1038/s41467-021-27045-6.

[Wat18] John Watrous. *The Theory of Quantum Information*. 1st. USA: Cambridge University Press, 2018. ISBN: 1107180562.

[Web16] Zak Webb. "The Clifford Group Forms a Unitary 3-Design". In: *Quantum Info. Comput.* 16.15–16 (Nov. 2016), pp. 1379–1400. ISSN: 1533-7146.

[Wes+24] Maxwell T. West et al. *Provably Trainable Rotationally Equivariant Quantum Machine Learning*. 2024. arXiv: 2311.05873 [quant-ph]. URL: https://arxiv.org/abs/2311.05873.

[WKS15] Nathan Wiebe, Ashish Kapoor, and Krysta Svore. "Quantum algorithms for nearest-neighbor methods for supervised and unsupervised learning". In: *Quantum Information and Computation* 15 (Mar. 2015), pp. 318–358.

[Wie+20] Roeland Wiersema et al. "Exploring Entanglement and Optimization within the Hamiltonian Variational Ansatz". In: *PRX Quantum* 1.2 (Dec. 2020). DOI: 10.1103/prxquantum.1.020319. URL: https://doi.org/10.1103%2Fprxquantum.1.020319.

[Wit+21] Nicolas Wittler et al. "Integrated Tool Set for Control, Calibration, and Characterization of Quantum Devices Applied to Superconducting Qubits". In: *Phys. Rev. Appl.* 15 (3 Mar. 2021), p. 034080. DOI: 10.1103/PhysRevApplied.15.034080. URL: https://link.aps.org/doi/10.1103/PhysRevApplied.15.034080.

[WWZ82] William K. Wootters, William K. Wootters, and Wojciech H. Zurek. "A single quantum cannot be cloned". In: *Nature* 299 (1982), pp. 802–803. URL: https://api.semanticscholar.org/CorpusID:4339227.

[Wu+21] Anbang Wu et al. "Towards Efficient Ansatz Architecture for Variational Quantum Algorithms". In: *arXiv:2111.13730* (2021). DOI: 10.48550/ARXIV.2111.13730. URL: https://arxiv.org/abs/2111.13730.

[YBL20] Jiahao Yao, Marin Bukov, and Lin Lin. "Policy Gradient based Quantum Approximate Optimization Algorithm". In: *Proceedings of The First Mathematical and Scientific Machine Learning Conference*. Ed. by Jianfeng Lu and Rachel Ward. Vol. 107. Proceedings of Machine Learning Research. PMLR, 20–24 Jul 2020, pp. 605–634. URL: https://proceedings.mlr.press/v107/yao20a.html.

[Yas+16] Masaya Yasuda et al. "Computational hardness of IFP and ECDLP". In: *Applicable Algebra in Engineering, Communication and Computing* 27 (Dec. 2016). DOI: 10.1007/s00200-016-0291-x.

[Yu20] Nengkun Yu. "Sample efficient tomography via Pauli Measurements". In: *arXiv:2009.04610* (2020). DOI: 10.48550/ARXIV.2009.04610. URL: https://arxiv.org/abs/2009.04610.

[Yue23] Henry Yuen. "An Improved Sample Complexity Lower Bound for (Fidelity) Quantum State Tomography". In: *Quantum* 7 (Jan. 2023), p. 890. ISSN: 2521-327X. DOI: 10.22331/q-2023-01-03-890. URL: http://dx.doi.org/10.22331/q-2023-01-03-890.

[Zha+22a] Kaining Zhang et al. "Escaping from the Barren Plateau via Gaussian Initializations in Deep Variational Quantum Circuits". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 18612–18627. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/7611a3cb5d733e628081431445cb01fd-Paper-Conference.pdf.

[Zha+22b] Kaining Zhang et al. “Escaping from the Barren Plateau via Gaussian Initializations in Deep Variational Quantum Circuits”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 18612–18627. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/7611a3cb5d733e628081431445cb01fd-Paper-Conference.pdf.

[Zha+19] Shuai Zhang et al. “Deep Learning Based Recommender System: A Survey and New Perspectives”. In: *ACM Comput. Surv.* 52.1 (Feb. 2019). ISSN: 0360-0300. DOI: 10.1145/3285029. URL: https://doi.org/10.1145/3285029.

[ZG21] Chen Zhao and Xiao-Shan Gao. “Analyzing the barren plateau phenomenon in training quantum neural networks with the ZX-calculus”. In: *Quantum* 5 (June 2021), p. 466. ISSN: 2521-327X. DOI: 10.22331/q-2021-06-04-466. URL: https://doi.org/10.22331/q-2021-06-04-466.

[Zho+20a] Han-Sen Zhong et al. “Quantum computational advantage using photons”. In: *Science* 370.6523 (2020), pp. 1460–1463. DOI: 10.1126/science.abe8770. eprint: https://www.science.org/doi/pdf/10.1126/science.abe8770. URL: https://www.science.org/doi/abs/10.1126/science.abe8770.

[Zho+20b] Leo Zhou et al. “Quantum Approximate Optimization Algorithm: Performance, Mechanism, and Implementation on Near-Term Devices”. In: *Phys. Rev. X* 10 (2 June 2020), p. 021067. DOI: 10.1103/PhysRevX.10.021067. URL: https://link.aps.org/doi/10.1103/PhysRevX.10.021067.

[Zhu17] Huangjun Zhu. “Multiqubit Clifford groups are unitary 3-designs”. In: *Phys. Rev. A* 96 (6 Dec. 2017), p. 062336. DOI: 10.1103/PhysRevA.96.062336. URL: https://link.aps.org/doi/10.1103/PhysRevA.96.062336.

[Zhu+16] Huangjun Zhu et al. “The Clifford group fails gracefully to be a unitary 4-design”. In: *arXiv: Quantum Physics* (2016).

[Żuk+98] Marek Żukowski et al. “Quest for GHZ states”. In: *Acta Physica Polonica A* 93.1 (1998), pp. 187–195.