THE TRAINABILITY AND EXPRESSIVITY OF QUANTUM MACHINE LEARNING MODELS

by

Eric R. Anschuetz
A.B., Harvard University (2017)
A.M., Harvard University (2017)

Submitted to the Department of Physics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Physics
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2023

Author .....
Department of Physics
May 15, 2023

Certified by .....
Aram W. Harrow
Professor of Physics
Thesis Supervisor

Certified by .....
Mikhail D. Lukin
Professor of Physics, Harvard University
Thesis Supervisor

Accepted by .....
Lindley Winslow
Associate Department Head of Physics

THE TRAINABILITY AND EXPRESSIVITY OF QUANTUM MACHINE LEARNING MODELS

by

Eric R. Anschuetz

Submitted to the Department of Physics
on May 15, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Physics

Abstract

Research over the last few decades has provided more and more evidence that precise control of many-body quantum systems yields a method of computation more powerful than what is achievable using conventional models of computation. This culminated in recent years with experimental demonstrations on quantum devices of computational tasks on the verge of classical intractability.

These current generation quantum devices are, however, too noisy and small to perform any means of error correction. These limitations motivate the study of hybrid quantum-classical algorithms as potential practical use-cases of these devices in the near-term. This thesis is concerned with studying potential use-cases of these hybrid algorithms, determining limitations of algorithms constructed via this framework, and giving provable guarantees on the performance of such algorithms. Specifically:

1. We consider quantum-classical hybrid algorithms which are framed as optimization problems with quantum-evaluated loss functions. We show that when the operations implemented on the quantum device are drawn from a certain problem-independent distribution, the loss landscapes (in expectation) exhibit a phase transition in trainability. We argue that the trainable phase is typically unachievable, and thus that such algorithms are not practical to implement.

2. We sharpen these arguments and show similar behavior for local, shallow, variational quantum algorithms. We also study the impact of noise on such algorithms when they are in the trainable phase, and show that such noise is capable of making otherwise trainable algorithms untrainable in a statistical query setting.

3. We give efficient classical algorithms to simulate certain variational quantum algorithms that circumvented the assumptions of our previous results and were known to be trainable, demonstrating that care must be taken to strike a balance between quantum implementability and classical intractability.

4. We prove unconditionally that certain quantum neural networks are more expressive than a wide class of classical neural networks and demonstrate that quantum contextuality is the resource for this separation. We also give arguments (along with numerical evidence) that such models are efficiently trainable, thus showing that there exists a regime where hybrid quantum-classical algorithms outperform their purely classical counterparts.

Thesis Supervisor: Aram W. Harrow
Title: Professor of Physics

Thesis Supervisor: Mikhail D. Lukin
Title: Professor of Physics, Harvard University

Acknowledgments

This thesis would not have been possible without the continued support of my advisors, friends, mentors, collaborators, and family.

I would first like to thank Aram for being a wonderful advisor these past few years—his encouragement to go and travel, meet other researchers, and make connections proved invaluable in helping me find collaborations and problems to think about over the course of my graduate studies. He also gave me the freedom to pursue whatever research directions I thought were interesting, which gave me the space to formulate my own independent research interests.

I am extremely grateful for the guidance, teaching, and advice of Misha, who at this point I have known for almost ten years. If not for the guidance of Misha during my undergraduate years, I probably would not be doing quantum information research today! Thanks for all of the advice, support, and encouragement.

I would also like to thank my many collaborators who made this thesis possible, including Hong-Ye Hu, Jin-Long Huang, Andreas Bauer, and Seth Lloyd. I would especially like to thank Xun Gao and Bobak T. Kiani, both of whom I have been lucky enough to work with on multiple occasions and who have given me invaluable guidance, ideas, and encouragement. It can be hard finding collaborators you gel with; I have been lucky enough to encounter it multiple times.

Throughout my time in the quantum information group at MIT I have been lucky enough to work alongside and be able to look up to Adam Bene-Watts, Tongyang Li, John Napp, Mehdi Soleimanifar, Ryuji Takagi, and John Wright. Thanks for being great mentors.

I was lucky enough to visit the IPM in Brno, Czechia for a three week visit in the fall of 2022; the trip would not have been nearly as wonderful as it was without the gracious hospitality of Martin Friák and Ivana Miháliková. Thanks for showing me around and making me feel welcome!

The latter half of my graduate studies were, of course, achieved during the COVID-19 pandemic. I was lucky enough to have been able to quarantine with my wonderful

partner Rin Zuo—whose love and support these past few years I would not have been able to do without—and our friend Elaine Spencer. Our frequent Domino's deliveries (special shout-out to our consistent pizzaiolo Luciano) and beer sessions made a dark time a whole lot brighter.

My many friendships throughout my last six years here managed to keep my weekends fun and full of life. I would particularly like to thank Pierre Barral, Jenni Crawford, Åsmund Folkestad, Simon Grosse-Holz, Thomas Hartke, Sam Leutheusser, Dimitra Pefkou, Greg Ridgway, Shreya Vardhan, Grace Weller, Ryan Weller, Sean Weller, and Cris Zanoci, (almost) all of whom were in my cohort here at MIT. Our frequent parties, board game nights, and outings sure have been a lot of fun.

This thesis also would not have been possible without the continued support of my wonderful family. My parents (well, and video games) inspired a love of computers from a young age, which is what got me into the sciences in the first place; my sister Elsa is my greatest cheerleader; and my grandparents, aunts, uncles, and cousins have also given much love and support throughout the years. Thank you all.

Finally, I would like to thank Ike Chuang and Soonwon Choi for giving me invaluable teaching experience and advice this past semester, along with Joe Formaggio for taking the time to be on my committee.

As for the rest: I never knew half of you half as well as I should have liked; and I liked less than half of you half as well as you deserved. So long, and thanks for all the fish!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Classically, physical states were believed to be completely determined by a number of canonical variables (namely, positions and momenta) linear in the system size n, and the measurable properties of such a system were thought to be deterministic (in principle) functions of these variables. The classical Hamiltonian dynamics governing the evolution of these canonical variables are simple; Turing machines can simulate these dynamics in time growing polynomially with n. This same principle has been true not only for mechanics as known by Newton but also for most physical theories introduced since. Einstein's equations introduce a curvature tensor but still fundamentally are deterministic and efficiently describable using Lagrangian dynamics. Classically, the formalism of statistical mechanics also follows directly from treating (in this case, many) classical canonical variables in a system.

With the formulation of quantum mechanics as a physical theory a century ago, humanity's understanding of the world around us changed remarkably. It has been realized that nature is inherently probabilistic. Rather than being described by an extensive number of scalar quantities, a physical state is described (at constant precision) by a *state vector* $|\psi\rangle$ with dimension exponentially large in n, and to date there are no known methods for efficiently classically simulating the (generic) Hamiltonian dynamics of such a state.

Inspired by this difficulty in modeling complex quantum systems, Richard Feynman in a talk given in 1982 [1] posed the question: why should we restrict ourselves

to classical Turing machines as a model of computation? If the world is quantum mechanical, we should model it using quantum mechanical systems—*quantum computers*—not classical ones. What sorts of computational tasks can be efficiently solved using this model of computation?

Even though it has been over forty years since Feynman initially proposed computing using quantum systems, we do not yet have a full understanding of exactly which tasks are efficiently solvable by quantum computers, which can be solved efficiently using (perhaps as of yet unknown) classical algorithms, and which are difficult for both classical and quantum computers. That said, there has been remarkable evidence that there exist practical tasks that lie in the first category and not the last two. Not only have efficient quantum algorithms been formulated for simulating quantum mechanical systems [2]—as originally proposed by Feynman—but also quantum algorithms more efficient than their known classical counterparts have been found for factoring large numbers [3], finding a marked element in an unstructured database [4], solving linear systems [5], and much more. Furthermore, quantum devices implementing these algorithms can be efficiently error corrected [6]. A general overview of quantum computation and quantum algorithms can be found in Reference [7].

Though these algorithms are still perhaps decades away from being implemented at a scale unachievable by current classical computers, the field of quantum information processing has reached a watershed in recent years: multiple experiments [8–13] have demonstrated quantum processors performing tasks on the verge of classical intractability. Spurred on by these recent experimental developments, there has been a push for finding algorithms that can be executed using these near-term quantum devices [14], potentially with some form of error mitigation. However, due to the constraints on these systems in terms of both size and coherence time, it is not at all obvious whether they are capable of outperforming state-of-the-art classical algorithms on practical tasks.

Motivated by these concerns, a subfield of quantum algorithms research has studied *hybrid quantum-classical algorithms*. In this framework, algorithms are con-

structed such that classically easy tasks are offloaded to a classical computer while any classically difficult (and quantumly easy) inner loop is run on a quantum device. The most commonly studied problems in this setting are optimization problems, where the solution to some problem is encoded as the minimum of some function $f(\boldsymbol{\theta})$ that is classically difficult to evaluate yet quantumly easy to evaluate. Then, a classical outer loop consisting of an optimization algorithm—utilizing a quantum computer as a "black box" to access f and its derivatives—can, in principle, optimize this function. In analogy with classical machine learning, this setting is often called *quantum machine learning* (QML) [15–18] or, depending on the form of $f(\boldsymbol{\theta})$, a *variational quantum algorithm* (VQA) [16, 19–26].

Some of the excitement around QML is due to the widespread success of classical machine learning algorithms in recent years. Thirty years ago, the gold standard for variationally learning complicated probability distributions was through the use of *Bayesian networks*; these ansatzes for probability distributions found widespread use in time series analysis [27], machine translation [28], natural language processing [29], and more [30–33]. In these probabilistic models, a distribution is modeled via products of conditional distributions associated with a directed acyclic graph:

$$p_{\text{model}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \prod_{i=1}^{n} p_i(\boldsymbol{x}_i \mid P(i)), \tag{1.1}$$

where here $P(i)$ is the set of direct predecessors of vertex i in the underlying graph. Given an underlying graph structure and a target distribution p_{target}, a (local) maximum likelihood estimate of p_{model} can be efficiently found using the expectation-maximization algorithm [34, 35]. Unfortunately, the graph structure itself is difficult to optimize over without *a priori* intuition as to what it should be. Furthermore, though training of these models is efficient in the dimension of the combined state space of the direct predecessors of vertices, these state spaces may need to be extremely large to fully capture more complex distributions with highly nonlocal correlations.

Due to these shortcomings, the last few decades have seen the field of machine

learning shift to *artificial neural network* (ANN)-based approaches. At finite precision, these models (when used for generative modeling) are a subclass of Bayesian networks that are efficiently trainable using gradient-based methods [36, 37]. This restriction to efficiently trainable models has allowed for the construction of extremely large neural networks—with parameter counts up to the order of a hundred billion—capable of extremely accurate sequence modeling [38, 39], video game playing at a high level [40], realistic text-to-speech [41], and much more [42, 43]. Given the ability of quantum devices to naturally sample from distributions believed to be beyond the reach of efficient classical algorithms, there is hope that quantum devices might one day be useful as a tool in the machine learner's toolkit [20, 21].

Unfortunately, there are many open questions regarding the usefulness of implementing QML algorithms on a hybrid quantum-classical device, even beyond noise concerns that typically arise when considering near-term quantum devices. For one, the optimization of QML loss landscapes is not guaranteed to work. One "miracle" of classical machine learning is the widespread trainability of neural network architectures: somewhat counterintuitively, neural networks become *easier* to train even as their widths (i.e., numbers of parameters per layer) grow [44, 45]. This behavior not only holds for simple toy models of neural networks but also—at least empirically—more generally. State-of-the-art neural networks today have hundreds of billions of parameters and are capable of generating strings of text almost indistinguishable from those generated by humans, all while being efficiently trainable [38, 39].

One might optimistically hope that the widespread success of machine learning algorithms in the classical setting would "port over" to their quantum counterparts. QML models generically explore exponentially large Hilbert spaces, in some sense probing larger model dimensions than do classical machine learning models. If the general heuristic that "wide models are easy to train" were to still hold true, this would imply the generic trainability of QML models. Realistically, however, any intuition gained from studying machine learning architectures in a classical setting has to be closely examined before hoping it applies to the quantum setting. Specifically, does this intuition for the efficient trainability of machine learning models still hold true

in the quantum setting? If not, is there a restriction of QML models to an efficiently trainable subspace, just as generative ANNs are for Bayesian networks? Do these classes of trainable QML models exhibit any post-classical behavior, or are the classes of trainable and classical machine learning models essentially identical?

This thesis answers these questions by studying QML models from a theoretical perspective. From this general theory we then attempt to build intuition for constructing QML models that not only are trainable but are also beyond the capabilities of classical machine learning algorithms. We hope that the presented intuition aids in the construction of new and exciting QML algorithms, just as the development of heuristics in classical machine learning led to an explosion in the number of practical and useful neural networks.

In a broad sense, we here show that QML models are generally *not* trainable. We show that a generic lack of trainability arises from a "phase transition" in the loss landscapes of QML models. More specifically, we show for the first time that when these models have a number of parameters fewer than the *degrees of freedom* m of the system—generically exponentially large in the system size n—a proliferation of poor local minima lead to loss landscapes unamenable to efficient training algorithms. We call this phase the *underparameterized phase*. In contrast, when the number of parameters of the model is at least m, the loss landscape of the model is nearly convex and efficiently trainable in terms of the number of optimization steps; we call this phase the *overparameterized phase*. Unfortunately, since this phase requires a number of parameters at least $m \sim \exp(n)$, QML models that are constructed to exist in this phase are generally inefficient to implement.

Though these results are generally pessimistic, one benefit of the existence of this phase transition is that it gives intuition for constructing *trainable* QML models. Namely, model classes with m growing only polynomially with n can easily be made to exist in the trainable phase. One natural way to restrict the effective dimension of such models is by requiring that the models respect *symmetries* [46]. Intuitively, if the dimension of the symmetry group G a model respects is large enough, the model should be trainable as the resultant degrees of freedom m should be sufficiently small

that the overparameterized phase is eminently reachable. This was recently formalized with $G = \mathrm{S}_n$, where it was shown that QML models equivariant under permutations of qubits that have a number of parameters $\gtrsim n^3$ are efficiently trainable [47]. Unfortunately, this solution is not a silver bullet. We here show that it is often the case that such trainable QML models are efficiently classically simulable, and give classical algorithms for simulating evolution symmetric under dynamics respecting such large symmetries.

The trick, then, is to balance restricting QML architectures in some way—such that they are trainable—while still maintaining an advantage over classical machine learning architectures. Luckily, we here show that this is indeed possible. We quantize a class of recurrent neural networks [48] and show the existence of a class of sequence learning problems that this model can efficiently represent, but *no* classical neural network with a subquadratic memory overhead can represent. We end with some preliminary thoughts on extending this quadratic separation to larger polynomial separations, and also with a general discussion of the trade-offs between the trainability, expressivity, and practicality of constructing quantum machine learning architectures.

1.1 Outline of Results

The results presented are drawn from a variety of papers [49–52] and would not have been possible without the valuable contributions of my coauthors and collaborators. The final result is edited together in an effort to present two aspects of QML research—the efficiency of training such algorithms, as well as their expressive power—as two sides of the same coin, with natural trade-offs between the two. We have also attempted to make each Chapter self-contained for readers primarily interested in a subset of the results presented here.

The remainder of this thesis is organized as follows:

- For the remainder of this Chapter, we give background on quantum generative modeling and variational quantum algorithms and give a summary of our re-

sults. Along the way we also briefly introduce "physics" notation for concepts in quantum computing to make them accessible to a classical computing audience.

- In Chapter 2, we consider the loss landscapes of a randomized class of variational quantum algorithms (with global interactions) and derive a phase transition in their trainability using techniques from Morse theory. To keep the exposition straightforward, we contain the technical results and proofs to Appendix A. These results are featured in Reference [49].

- In Chapter 3, we extend these results and consider the loss landscapes of *local* variational quantum algorithms, deriving a similar phase transition as in Chapter 2. We also consider the noisy optimization of such algorithms, and show more general untrainability results in this setting. Once again, we defer all technical details to Appendix B. These results are featured in Reference [50], joint work with Bobak T. Kiani.

- In Chapter 4, we consider variational quantum algorithms shown to circumvent these untrainability results; we show that, unfortunately, it is often the case that such learning algorithms can be dequantized to yield equally efficient classical algorithms. In Appendix C, we give technical background on properties of the symmetric group, which as a symmetry group is considered as a special case of our results. These results are featured in Reference [51], joint work with Andreas Bauer, Bobak T. Kiani, and Seth Lloyd.

- In Chapter 5, we consider an explicit class of quantum machine learning models that sidesteps the pessimism of the previous chapters. Namely, we prove that this class of quantum machine learing models has a provable expressivity separation over a wide class of quantum neural networks, and give heuristic (and numerical) evidence that this class is trainable. Technical details are once again deferred to Appendix D. These results are featured in Reference [52], work done in collaboration with Hong-Ye Hu, Jin-Long Huang, and Xun Gao.

- In Chapter 6, we give an outlook for this line of research and conclude.

1.2 Preliminaries

Before discussing the details of our results, we first provide some basic background on quantum machine learning (QML) in the context of hybrid quantum-classical algorithms, and also more specifically discuss variational quantum algorithms. More technical background that more specifically applies to individual Chapters are given there.

1.2.1 Classical Machine Learning Models

The field of classical machine learning has grown to cover so many concepts in the past few years that it would be impossible to give a complete background here. We thus here focus on the models and tasks we more specifically consider in later Chapters; a general introduction to the field is given in Reference [53].

Typically, the task one considers in machine learning contexts is the search for a function f within a given function class $\mathcal{F}$ that minimizes some risk:

$$\mathcal{R}[f] = \mathbb{E}_{\boldsymbol{x}}[\ell(\boldsymbol{x} \mid f)] \tag{1.2}$$

given a distribution of inputs $\boldsymbol{x}$ and a loss function ℓ. To perform learning, one searches over a parameterized subset functions $\hat{f}_{\boldsymbol{w}} \in \mathcal{F}$ termed *models*. When this expectation cannot be efficiently taken, one instead minimizes the *empirical risk* (or *training error*) $\hat{\mathcal{R}}(\boldsymbol{w})$ over a given training data set D:

$$\hat{\mathcal{R}}(\boldsymbol{w}) = \frac{1}{|D|} \sum_{\boldsymbol{x}_i \in D} \ell(\boldsymbol{x}_i \mid f_{\boldsymbol{w}}). \tag{1.3}$$

There are many model classes often considered in machine learning, and which is used typically depends on the specifics of the task being performed by the model. Broadly, state-of-the-art models typically are *feedforward neural networks* (FNNs) [36]; here, functions are modeled as alternating trained linear functions $\lambda_{i;\boldsymbol{w}}$ and fixed (up

to, perhaps, tuned hyperparameters) nonlinearities ν_i:

$$\hat{f}_{\boldsymbol{w}} = \bigcirc_i \left(\nu_i \circ \lambda_{i;\boldsymbol{w}}\right). \tag{1.4}$$

Typically, these $\lambda_{i;\boldsymbol{w}}$ are just simple matrix multiplication by a matrix completely parameterized by $\boldsymbol{w}$. Somewhat miraculously, the empirical risk of typical machine learning loss functions when the models take this simple layered form can be shown to be (typically) efficient to optimize with simple gradient descent [44, 45]. The widespread success of FNNs can, at least partially, be attributed to the tractability of these loss landscapes along with the simplicity of calculating gradients via backpropagation [37]. When implemented on specialized hardware such as graphical processing units, this efficient training has allowed models with on the order of a hundred billion parameters [38] to be trained for tasks with real-world applications.

1.2.2 Quantum Machine Learning Models

A quantum system of size n is naturally represented by a *quantum state*, which is a normalized vector—for simplicity, assumed to be over n qubits—$|\psi\rangle \in \mathbb{C}^{2^n}$. Here, we use the typical physics notation $|\psi\rangle$ to denote a column vector instead of (say) $\boldsymbol{\psi}$ when we are describing a quantum state, with the notation $\langle\psi|$ used to denote its conjugate transpose. A quantum state in $\mathbb{C}^{2^n}$ can be considered a generalization of probability distributions over 2^n states (i.e., over states described by n bits), where the norm squared of entries of $|\psi\rangle$ give the *measurement distribution* over these states. A general overview of quantum mechanics and, more specifically, its applications to quantum computation can be found in Reference [7].

Just as operations that map probability distributions to probability distributions are naturally described by stochastic matrices, operations that map (pure, n qubit) quantum states to (pure, n qubit) quantum states are naturally described by unitary matrices; equivalently, they are described by the matrix exponentials of $2^n \times 2^n$ skew-Hermitian matrices. Multiple quantum operations can then be described by the sequential matrix multiplication of various matrix exponentials. It is then apparent

that one can construct a QML model (often called an *ansatz* in the physics literature) by parameterizing these matrix exponentials, giving rise to the layered structure:

$$|\boldsymbol{\theta}\rangle \equiv \prod_{i=1}^{q} U_i(\boldsymbol{\theta}) |\psi_0\rangle \equiv \prod_{i=1}^{q} \mathrm{e}^{-\mathrm{i}\theta_i Q_i} |\psi_0\rangle , \tag{1.5}$$

where here q is the depth of the model and Q_i are fixed Hermitian matrices. Typically q is polynomial in n, i.e., logarithmic in the dimension of the initial state vector $|\psi_0\rangle$. Though the linear nature of the model may be surprising, this model structure (for large enough depth q) is known to be a universal approximator of all (pure) quantum states, even for a fixed number of allowed Q_i [54, 55]. Note that unlike typical classical machine learning models the unitaries in Equation (1.5) are parameterized by only a *single* parameter each even though they are each of dimension $2^n \times 2^n$.

For completely general (i.e., dense) Q_i in Equation (1.5), this model is not efficient to implement on a quantum computer. Thus, due to their efficiency in physical implementation, these Q_i are typically taken to be members of the n *qubit Pauli group* $\mathbb{P}_n$ which forms a basis for all $2^n \times 2^n$ Hermitian matrices. The Pauli group is also convenient to study analytically, as is the normalizer of the group (called the *Clifford group*). The assumption that each Q_i is a Pauli operator also allows us to use a single parameter θ_i for each layer of the model without loss of generality; in principle, more parameters can describe each layer by parameterizing sparse Q_i. However, as the Pauli group is a basis for Hermitian matrices, this is a special case of the class of models we consider here (at the expense of larger q, new dependencies among the θ_i, and a controllable approximation error).

Just as in classical machine learning, QML algorithms are most often tasked with minimizing some risk over (for our purposes, assumed pure) quantum states $\rho = |\psi\rangle\langle\psi|$:

$$\mathcal{R}(\rho) = \mathbb{E}_{\boldsymbol{x}}[\ell(\boldsymbol{x} \mid \rho)] \tag{1.6}$$

given a distribution of inputs $\boldsymbol{x}$ and a loss function ℓ. To perform learning, one optimizes this risk over parameterized models $|\boldsymbol{\theta}\rangle$. In a completely analogous fashion to classical machine learning, when this expectation cannot be efficiently taken one

instead minimizes the empirical risk $\hat{\mathcal{R}}(\boldsymbol{\theta})$ over a given training data set D:

$$\hat{\mathcal{R}}(\boldsymbol{\theta}) = \frac{1}{|D|} \sum_{\boldsymbol{x}_i \in D} \ell\left(\boldsymbol{x}_i \mid |\boldsymbol{\theta}\rangle \langle\boldsymbol{\theta}|\right). \tag{1.7}$$

Of course, just because machine learning can be performed using inherently quantum models does not mean that necessarily it is a good idea to do so. Luckily, it is known that under certain complexity theoretic assumptions there are certain tasks—even on classical data sets—where there are known expressivity separations between quantum and classical model classes [56–60]. These results essentially rely on reducing certain machine learning tasks to problems believed to be in BQP but not in BPP. However, the feasibility of performing an optimization over models which essentially comprise all of efficient quantum computation was not considered—we will discuss this in more detail in Section 1.3, along with Chapters 2 and 3.

1.2.3 Variational Quantum Algorithms

Perhaps the most well-studied class of quantum machine learning algorithm consists of those classified as variational quantum algorithms (VQAs) [19]. This is a class of generative modeling problems where the task is to prepare a state close to some target state $|\psi_{\text{target}}\rangle$. In this setting, we are not given $|\psi_{\text{target}}\rangle$ directly; instead, we are given a $2^n \times 2^n$ Hermitian matrix H (called the *problem Hamiltonian*) where $|\psi_{\text{target}}\rangle$ is the eigenvector associated with the smallest eigenvalue of H (called the *ground state*). In this formulation, optimization proceeds via the minimization of:

$$F_{\text{VQA}}(\boldsymbol{\theta}) = \langle\boldsymbol{\theta}| H |\boldsymbol{\theta}\rangle. \tag{1.8}$$

In the language of risks typical in classical machine learning this can be expressed as the minimization of:

$$\hat{\mathcal{R}}_{\text{VQA}}(\boldsymbol{\theta}) = \sum_{i=1}^{A} \alpha_i \langle\boldsymbol{\theta}| P_i |\boldsymbol{\theta}\rangle, \tag{1.9}$$

where:

$$H = \sum_{i=1}^{A} \alpha_i P_i \tag{1.10}$$

is the Pauli decomposition of H. Assuming no degeneracies in the eigenspectrum of H and a sufficiently expressive model, the minimizer $|\boldsymbol{\theta}^*\rangle$ of Equation (1.8) is the ground state of H up to an overall phase due to the quadratic nature of Equation (1.8). Assuming H is efficiently expressible as the weighted sum of $\mathrm{O}\left(\mathrm{poly}\left(n\right)\right)$ Pauli matrices, Equation (1.8) and its gradients can be efficiently measured on a quantum computer [19, 61]. This and similar formulations of quantum generative modeling with Equation (1.8) as the loss function are called *variational quantum algorithms* (VQAs) [19]. Generally, the goal of these algorithms is to find the state $|\boldsymbol{\theta}\rangle$ that optimizes Equation (1.8) up to some constant additive error in loss. Though there are other formulations of quantum generative modeling, we here focus on VQAs as they do not require coherent access to data $|\psi_{\text{target}}\rangle$ which is generally believed to be difficult [62].

Typically, models in VQAs come in one of two flavors: Hamiltonian agnostic models and Hamiltonian informed models. Hamiltonian agnostic models are constructed such that the Q_i present in the model definition are independent of H, and are generally more efficient to implement. This is most analogous to the case in classical generative modeling, where the model structure is usually independent from the specific choice of data distribution. This will be the setting we mostly consider throughout this thesis, though we give some discussion on Hamiltonian informed ansatzes in Chapters 2 and 3.

1.3 Summary of Results

1.3.1 Quantum Machine Learning Models Are Generically Untrainable

The close analogy between quantum and classical machine learning models gives one hope that the same trainability results from the classical machine learning litera-

ture [44, 45] might immediately apply to the quantum setting. Particularly, one might hope that optimization of the loss landscape of Equation (1.9) is tractable. Unfortunately, Chapters 2 and 3 are dedicated to showing that unlike the classical setting, the performance of QML models is often dominated by poor performance in the training procedure (see Appendix B.1 for a breakdown of potential sources of error in the training of QML models). Typically the trainability of QML models is studied more specifically in the context of VQAs, so in this summary we focus on that setting. In Chapter 3 we also consider the trainability of more general QML models when there is noise present by working in the *statistical query* setting, but we exclude discussion in this summary for brevity.

Prior to the publication of the results featured in Chapters 2 and 3, research on the trainability of VQAs was mainly focused on the *deep model* regime. It was previously known that in this regime gradients of deep variational quantum circuits vanish exponentially with the problem size in many settings [63–65]. Unlike the vanishing gradients that sometimes appear in classical machine learning contexts, these "barren plateaus" in the quantum setting cannot be solved via clever initialization strategies (like those found classically [66]) that limit their decay with depth; rather, they are endemic to the models themselves and (for instance) can worsen even if one only increases the width of a quantum model. Problematic training in the deep model regime has also been studied beyond gradient descent [67, 68]. In short, these previous results left the door open for generic, trainable, shallow QML models. See Table 1.1 for a summary of these previous results compared with the results presented in this thesis.

For the first time, the results presented in Chapters 2 and 3 demonstrate analytically the untrainability of QML models in the shallow regime. More specifically, we analytically show the presence of a "phase transition" in the training of VQAs for certain classes of randomized ansatzes inspired by the hardware-efficient class of ansatzes, qualitatively similar to a trainability phase transition previously observed numerically [69–72]. In particular, we demonstrate the convergence of these ansatzes to a certain class of random fields on the hypertorus; we are then able to analytically

Result	Dimension	Locality	Depth	Barren plateaus?	Poor minima?
Reference [63]	d	2	$\Omega\left(n^{\frac{1}{d}}\right)$	✓	?
Reference [64]	1	2	$\omega\left(\log\left(n\right)\right)$	✓	?
Reference [65]	d	2	$\omega\left(\log\left(n\right)^{\frac{1}{d}}\right)$	✓	?
Chapter 2	N/A	n	$\Omega\left(1\right)$	✓/✗	✓/✗
Chapter 3	d	2	$\Omega\left(1\right)$	✗	✓

Table 1.1: A summary of previous results on the untrainability of variational quantum algorithms. A label of "✓/✗" denotes that certain regimes were studied where the phenomenon was present, and certain regimes where it was not. A label of "?" denotes that the phenomenon was not studied. "Dimension" indicates the locality structure of the ansatzes study. For instance, Dimension = 1 denotes ansatzes with nearest-neighbor interactions for qubits on a line.

calculate the distribution of critical points of a specified index as a function of loss function value. Asymptotically, we are able to show that there is a "phase transition" from an *underparameterized regime*—where all local minima are exponentially concentrated near half the mean eigenvalue of H, far from the constant additive error to the ground state energy that is often the goal in VQAs [19]—and an *overparameterized regime*—where all local minima are exponentially concentrated near the global minimum of H. Numerically, we also observe this behavior at a modest depth p. Additionally, in the process we prove novel results in the distribution of local minima for this class of random fields which may be of independent interest. In Chapter 2, this randomized class of ansatzes includes those with nonlocal interactions and essentially is a result of a scrambling behavior in the ansatzes; we strengthen our arguments in Chapter 3 to show that a similar phase transition exists even when the interactions in the ansatzes are constrained to be local due to a "local scrambling" behavior. We give an abbreviated overview of our results and techniques in the remainder of this Section.

Mapping Loss Landscapes to Random Fields

Previous results in the classical machine learning literature [44, 45] on the distribution of critical points in the loss landscape begin by mapping their class of machine learn-

ing models of interest to a class of random fields known as *Gaussian hyperspherical random fields*. These are random fields of the form:

$$F_{\text{GHRF}}(\boldsymbol{\theta}) \propto \sum_{i_1,\ldots,i_r,i'_1,\ldots,i'_r=1}^{\Lambda} \sigma_{i_1} \ldots \sigma_{i_r} J_{i_1,\ldots,i_r,i'_1,\ldots,i'_r} \sigma_{i'_1} \ldots \sigma_{i'_r}, \tag{1.11}$$

where $\boldsymbol{\sigma}$ is a point on the hypersphere S^{Λ} parameterized by $\boldsymbol{\theta}$ (for details, see Equation (2.4)). Here, each $J_{i_1,\ldots,i_r,i'_1,\ldots,i'_r}$ is an i.i.d. random Gaussian variable. These mappings rely on nonlinearities in the model providing the effective Gaussian interaction in Equation (1.11), and the fact that the linear transformations in the model are completely parameterized gives the product of σ_i in Equation (1.11). Known results on the loss landscape of this class of random fields [73–75] can then be used to infer the loss landscape of the studied machine learning models; namely, that the local minima of the landscape concentrate near the global minimum.

Our mapping of Equation (1.8) with a class of Hamiltonian agnostic ansatzes differs from the classical machine learning construction in two major ways. First, the variational ansatz is linear; the nonlinearity of the classical construction was crucial in giving the effective Gaussian couplings $\boldsymbol{J}$. Second, the unitaries in variational ansatzes are (usually) parameterized rotations by simple Pauli strings; this means the product of unitaries in the ansatz do not give a simple product of the ansatz parameters as in Equation (1.11), and each layer is vastly underparameterized. We will later find that this latter fact gives rise to the qualitative differences in the loss landscapes of deep neural networks and variational quantum algorithms.

Thus, our mapping relies on a different strategy. In Chapter 2, to make the mapping tractable, we consider ansatzes built from rotations by uniformly random Pauli strings. This has the added benefit of making the parameters of our ansatz naturally described as points on the hypertorus. We then consider the path integral expansion of Equation (1.8), where each path is weighted by the parameters of the anstaz and various matrix elements of H. Then, we show at fixed H that these matrix elements are approximately a sum of (shifted) *Wishart random variables*. These are just simple multivariate generalizations of the gamma distribution; for more details on

Wishart random matrices as well as random fields constructed from Wishart matrices, see Section 2.3.3. Finally, we express this sum of Wishart random variables as a single Wishart random variable which is close in distribution under reasonable assumptions on H. This means that, unlike the classical case, the natural random field to study for VQAs is the *Wishart hypertoroidal random field* (WHRF) which we introduce in Chapter 2. These results are summarized in the following (informal) theorem:

Theorem 1.1 (Nonlocal VQAs as WHRFs, informal statement of Theorems A.1 and A.2)**.** *Consider the class of ansatzes:*

$$|\boldsymbol{\theta}\rangle \equiv \prod_{i=1}^{q} U_i(\boldsymbol{\theta}) |\psi_0\rangle \equiv \prod_{i=1}^{q} \mathrm{e}^{-\mathrm{i}\theta_i Q_i} |\psi_0\rangle , \tag{1.12}$$

where each Q_i is drawn uniformly from the Pauli group $\mathbb{P}_n$ and $|\psi_0\rangle$ is a uniformly random stabilizer state. Let p be the number of distinct θ_i, and let $r = q/p$. Under reasonable assumptions on the eigenvalues of H (with ground state energy λ_1), the random variational objective function

$$F_H(\boldsymbol{\theta}) = \frac{\langle\boldsymbol{\theta}| H |\boldsymbol{\theta}\rangle - \lambda_1}{2^{-n} \|H - \lambda_1\|_*} \tag{1.13}$$

converges in distribution to the random field

$$F_{WHRF}(\boldsymbol{\theta}) = \sum_{i_1,\ldots,i_r,i'_1,\ldots,i'_r=1}^{2^p} w_{i_1} \ldots w_{i_r} J_{i_1,\ldots,i_r,i'_1,\ldots,i'_r} w_{i'_1} \ldots w_{i'_r}, \tag{1.14}$$

where $\boldsymbol{w}$ are points on the hypertorus $(S^1)^{\times p}$ parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{J}$ is a complex Wishart random matrix normalized by its number of degrees of freedom. $\|\cdot\|_*$ denotes the nuclear norm of $\cdot$.

For the class of ansatzes we consider and for physically relevant H, $\boldsymbol{J}$ has $m = \Theta(2^n)$ degrees of freedom.

In Chapter 3, we extend these results to also include local VQAs (assumed to be shallow, as untrainability results for deep local VQAs are already known [63–65]). These results can be summarized by the following (informal) theorem:

Theorem 1.2 (Local VQAs as WHRFs, informal statement of Theorem B.18)**.** *Consider a class of ansatzes as in Equation 1.12 but with the Q_i now constrained to be local and $r = 1$ for simplicity. Let l be the number of parameters in the reverse lightcone of any observable in the Pauli decomposition of H. Then, F_H converges in distribution to the random field*

$$F_{WHRF}(\boldsymbol{\theta}) = \sum_{i,i'=1}^{2^l} w_i J_{i,i'} w_{i'}, \tag{1.15}$$

where $\boldsymbol{w}$ are points on the hypertorus $(S^1)^{\times p}$ parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{J}$ is a complex Wishart random matrix normalized by its number of degrees of freedom.

We find that in this setting, $\boldsymbol{J}$ has $m = \Theta\left(n2^l\right) = \omega(l)$ degrees of freedom.

The Loss Landscapes of Wishart Hypertoroidal Random Fields

Once mapped to a standard random field on a compact manifold, we can utilize results from Morse theory to find the expected distribution of critical points of the WHRF loss function (and therefore those of Equation (1.8)) by studying the joint distribution of the loss function, its gradient, and its Hessian. In the classical machine learning case these random fields are all Gaussian. This follows from the fact that sums of the independent Gaussian coefficients in Equation (1.11) are also Gaussian. However, the entries of Wishart $\boldsymbol{J}$ are *not* independent—this can easily be seen as $\boldsymbol{J}$ is positive semidefinite. Luckily, through explicit calculation we are able to show that the Hessian (conditioned on being at a critical point) takes the simple form of the sum of a Wishart matrix and an independent Gaussian matrix. The joint distribution with the loss and gradient takes on a similarly simple form.

With the joint distribution of the loss function and its derivatives in hand, we are able to use results from Morse theory to find the expected distribution of critical points of the WHRF loss function at various energies E (in units of the mean eigenvalue $2^{-n}\left\|H - \lambda_1\right\|_*$). Though the results are unwieldy—involving an expectation over the eigenvalues of the sum $\boldsymbol{C}$ of independent Wishart and Gaussian matrices—they are

exact. Furthermore, the exact formula allows us to probe the expected distribution of critical points of various indices k, where k labels the number of negative eigenvalues of the Hessian at the critical point (e.g., $k = 0$ probes local minima). The informal result (for WHRFs on the p-torus) is as follows:

Theorem 1.3 (WHRF critical point distribution, informal statement of Theorem A.5)**.** *Let $\mu_{\boldsymbol{C}(E)}$ be the eigenvalue distribution of a random matrix $\boldsymbol{C}(E)$ drawn from a certain distribution of random matrices dependent on E. Then, the expected number of critical points of index k at an energy E is*

$$\begin{aligned}&\mathbb{E}\left[\mathrm{Crt}_k(E)\right]\\&= \left(\frac{\pi}{r}\right)^{\frac{p}{2}} \Gamma(m)^{-1} m^{(1+\gamma)m} \mathbb{E}_{\boldsymbol{C}(E)}\left[\mathrm{e}^{p\int \ln(|\lambda-2rE|)\mathrm{d}\mu_{\boldsymbol{C}(E)}} \mathbf{1}\left\{\lambda_{k+1}^{\boldsymbol{C}(E)} \geq 2rE\right\}\right] E^{(1-\gamma)m-1}\mathrm{e}^{-mE},\end{aligned} \tag{1.16}$$

where

$$\gamma = \frac{p}{2m} \tag{1.17}$$

and $\lambda_i^{\boldsymbol{C}}$ is the ith smallest eigenvalue of $\boldsymbol{C}$.

The precise statement of this theorem is given in Appendix A, along with the distribution from which $\boldsymbol{C}(E)$ is drawn. We call γ in Equation (1.17) the *overparameterization factor*; it is the (scaled) ratio of the number of independent parameters of the loss function to the number of degrees of freedom m of the WHRF. As discussed in Section 1.3.1, m is generically exponential in n for nonlocal ansatzes, so unless the Hamiltonian agnostic ansatz has exponentially many parameters γ is very small. Similarly, for shallow local ansatzes m is generically exponential in $l + \log(n)$; here, γ will *always* be very small.

Asymptotic Limits of the Critical Point Distribution

Though Equation (1.16) gives the exact distribution of critical points, it is difficult to use in practice. As mentioned in Section 1.3.1, this difficulty comes from the expectation over eigenvalues of the sum of independent Wishart and Gaussian matrices. Surprisingly, however, the eigenvalues of both Wishart and Gaussian orthogonal ma-

trices converge in distribution to *fixed* distributions. Roughly, asymptotically in the size of the matrix, the eigenvalue distributions of all normalized Wishart matrices are the same (given by the *Marchenko–Pastur distribution*) and the eigenvalue distributions of all Gaussian orthogonal matrices are the same (given by the *Wigner semicircle distribution*). Putting aside deviations from this convergence for the moment, an asymptotic treatment of Equation (1.16) can be given when considering the asymptotic behavior of the eigenvalue distribution of the sum of these matrices.

Luckily, we can characterize the asymptotic distribution of eigenvalues of the sum well using the tools of *free probability theory*. Roughly, free probability theory is the probability theory of noncommutative random variables (e.g., random matrices). As the distribution of the sum of two random variables in commutative probability theory can be described by the convolution of the distributions of the two independent random variables, so can the *free convolution* of the distributions of two *freely independent* noncommutative random variables. Using the asymptotic free independence of Wishart and Gaussian orthogonal random variables, we are able to show that asymptotically the eigenvalue distribution of their sum weakly converges to the free convolution of a Marchenko–Pastur distribution with a semicircle distribution.

However, weak convergence is not enough; due to the exponential factor in the expectation, any large deviations from the asymptotic convergence—even if they occur with exponentially vanishing probability—can potentially cause large deviations from the naive application of free probability theory. In Appendix A we are able to bound the probability of these large deviations, and show that unlike the Gaussian case [73] to (logarithmic) leading order they do not contribute to the final result. This is due to the contribution to the expectation from the deviations being dominated by what is predicted by free probability theory.

Armed with an asymptotic expression for the distribution of critical points, we specialize to two limits for WHRFs on the p-torus: $p \geq 2m$ (i.e., the overparameterized regime) and $p \ll m$ (i.e., the underparameterized regime). In the former, we show to leading multiplicative order in $p \gg 1$ that all local minima are located at the global minimum. Though the classes of ansatzes differ, we believe that a similar phenomenon

gives rise to the phase transition in trainability observed in References [69–72]. In the underparameterized regime, we show that the density of local minima approximately follows a compound confluent hypergeometric (CCH) distribution [76]:

$$f_{\mathrm{CCH}}\left(E \mid p, m\right) \propto \mathrm{e}^{-E} E^{m-\frac{p}{2}} \left(1-2E\right)^{p}, \tag{1.18}$$

which has a width $\sim m^{-1}$ and is centered near $E = \frac{1}{2} - \gamma$ (i.e., near half of the mean eigenvalue of the objective observable when units are restored) when $\gamma \ll 1$. In other words, these local minima are exponentially concentrated far from the global minimum, rendering such models untrainable when in the underparameterized machine. As reaching the overparameterized regime requires large-depth quantum models—a regime shown by References [63–65] to be inconducive to training due to the presence of barren plateaus—this result implies that *all* generic VQAs are untrainable.

In the time since the results featured in Chapters 2 and 3 were made public, others have studied the trainability of VQAs in the shallow model regime. In a similar line of research, Reference [77] showed that for certain quantum variational ansatzes or quantum neural networks there exist data sets and loss functions which induce exponentially many local minima in the loss landscape. References [78, 79] both showed that in an overparameterized regime these models experience good local minima, though this transition to trainability typically occurs at an intractable number of parameters. Finally, assuming the presence of a constant rate of noise per ansatz gate, Reference [80] showed convergence of the loss landscape to the uniform distribution at a rate exponential in the circuit depth.

1.3.2 Classical Simulability of Symmetric Quantum Machine Learning Models

The introduction of these untrainability results motivates the construction of *non-generic* QML models. The most straightforward way to achieve this is through the introduction of *symmetry equivariant models* [46]. Given a representation $R\left(G\right)$ of a

(here assumed finite) symmetry group G, these models undergo parameterized time evolution under Hermitian operators H_i that commute with the representation:

$$[R(G), H_i] = 0. \tag{1.19}$$

As the untrainability results of Chapters 2 and 3—as well as the results mentioned in Section 1.3.1—are due to the exponential scaling (with the system size) of the degrees of freedom of generic quantum machine learning models, considering symmetry-restricted models that have fewer effective degrees of freedom intuitively yields trainable QML models. Indeed, for the symmetric group $G = \mathrm{S}_n$ with representation $R(G)$ given by products of SWAP operators acting on an n qubit Hilbert space this has been shown analytically [47].

However, in Chapter 4, we show that this reduction in the effective degrees of freedom of the QML model can yield efficient classical simulation algorithms. We first state very general results for general (finite) symmetry groups G and then specialize to the symmetric group as an explicit example.

General Symmetry Groups

Our general results are essentially a consequence of the (potential) existence of more efficient representations of symmetry equivariant models than ones that act on a 2^n-dimensional Hilbert space. As a warm-up, let us first consider finding the ground state energy of a Hamiltonian H that commutes with a group representation R of G acting on 2^n-dimensional Hilbert space. We can consider the isotypic decomposition of this representation into *irreducible representations* (irreps):

$$R = \bigoplus_{\lambda} V_{\lambda}^{\oplus q_\lambda}. \tag{1.20}$$

Schur's lemma [81] then implies that any element a of (the natural representation of) the commutant X of R must be of the form:

$$a = \bigoplus_{\lambda} U_\lambda \otimes I_{\dim(V_\lambda)}, \tag{1.21}$$

where U_λ acts on a q_λ-dimensional vector space. However, X has another representation B where the element a represents in X has representation:

$$b = \bigoplus_{\lambda} U_\lambda. \tag{1.22}$$

Note that this representation is only $\sum_\lambda q_\lambda$-dimensional rather than $\sum_\lambda \dim(V_\lambda)\, q_\lambda$-dimensional, and that the smallest eigenvalue of b is identical to that of a. In particular, considering $H \in X$ in a more efficient representation immediately yields an algorithm for determining the ground state energy of H more efficiently than the naive algorithm. Luckily, there exists an efficiently (in the dimension of X) calculable representation of X acting on a vector space only of dimension $\dim(X)$ called the *regular representation*. This observation immediately yields the following (informally stated) theorem:

Theorem 1.4 (Algorithm for finding the ground state energy of symmetric Hamiltonians, informal statement of Theorem 4.2)**.** *Let H be a Hamiltonian that represents an element of the commutant X of R. The ground state energy of H can be found in time* $\mathrm{O}\left(\dim(X)^3\right)$*. We give a classical algorithm to do so in Chapter 4.*

We also give a variant of this algorithm that allows one to find the ground state itself in a given preferred basis.

These same ideas can also be used to simulate equivariant dynamics followed by measurement of an operator O commuting with the group representation R. Notably, these results do *not* require the initial state of the dynamics to commute with the group representation. Instead, we show that given a certain classical description (more specifically, a *classical shadow* [82] description) of a general initial state, evolution under the given equivariant dynamics followed by measuring O can be sim-

ulated to small additive error with high probability. These results informally can be summarized in the following way:

Theorem 1.5 (Algorithm for simulating equivariant dynamics, informal statement of Theorem 4.4)**.** *Equivariant dynamics acting on an arbitrary quantum state followed by measurement of a symmetry invariant operator O can be simulated with high probability to low additive error in time polynomial in* $\dim(X)$*. We give a classical algorithm to do so in Chapter 4.*

Specialization to the Symmetric Group

These theorems immediately specialize to G equaling the symmetric group. In particular—optimizing the time complexities slightly from the general statements—we have that:

Corollary 1.6 (Algorithm for finding the ground state energy of S_n-invariant Hamiltonians, informal statement of Corollary 4.5)**.** *Let H be a Hamiltonian on n qubits invariant under permutations of these qubits. The ground state energy of H can be found in time* $O(n^8)$*. We give a classical algorithm to do so in Chapter 4.*

and:

Corollary 1.7 (Algorithm for simulating S_n-equivariant dynamics, informal statement of Corollary 4.8)**.** *S_n-equivariant dynamics acting on an arbitrary quantum state followed by measurement of a symmetry invariant operator O can be simulated with high probability to low additive error in time* $O(n^4)$*. We give a classical algorithm to do so in Chapter 4.*

Interestingly, we show that this latter algorithm can be parallelized in a way such that its time complexity is effectively less than the quantum algorithm given in Reference [47] for performing inference using S_n-equivariant QML models; our result in this instance can be thought of as a "fast-forwarding" of the quantum time evolution as noticed in Reference [83]. Our results here demonstrate the nontrivial balancing act one has to perform when constructing QML models: not only must they be constructed in a way to avoid the untrainability results discussed in Section 1.3.1, but

also (ideally) one should prove a separation over classical machine learning algorithms to avoid dequantization results similar to those presented here.

1.3.3 Provable Expressivity Advantage in Trainable Quantum Machine Learning Models

As mentioned in Section 1.2.2, all previous proofs of advantage in the expressivity of QML models over classical models rely on results from computational complexity theory, themselves conditional on complexity theoretic assumptions [56–60]. As the proofs of separation are abstract, it is unclear what realistic classical data sets one should expect a separation to hold in practice. Also, due to the universality of many of these models, they are very likely to be untrainable due to phenomena outlined in Section 1.3.1; and, as Section 1.3.2 demonstrated, it is nontrivial to find classes of quantum models that exhibit both trainability and still yield a quantum advantage over classical models. Because of these concerns, it has become increasingly clear that quantum models should be carefully constructed to fit the task at hand. Above all else, the *interpretability* of any expressivity separation achieved by a QML model has become increasingly important. Interpretability reveals which features of quantum mechanics yield more expressive models compared to classical models and, armed with this knowledge, allows one to find classes of problems where a practical quantum advantage on real data is potentially achievable.

In Chapter 5, we give the first unconditional separation in the expressive power of quantum generative models and a wide class of classical neural networks on *sequence-to-sequence* learning tasks [84]. We consider a quantization of linear recurrent neural networks, where time evolution is performed under a Hamiltonian quadratic in the canonical operators $\hat{q}_i$ and $\hat{p}_i$. To measure properties of the state of the system, the most natural choice is to perform *homodyne measurement*; that is, measure linear combinations of the canonical operators. This yields a quantum generative model where all operations are Gaussian. However, as all operations are Gaussian, there are efficient Wigner function based simulations of sampling from such a system [85]. In

other words, such models on n modes are equivalent to deep belief networks [86]—a class of commonly used classical models—with $2n$ latent variables.

Instead, we extend this model slightly further by allowing for measurements of the canonical operators *modulo* 2π. We call this introduced class of models *contextual recurrent neural networks* (CRNNs). Our main result is that CRNNs are more memory efficient at expressing certain distributions than essentially all trainable classical sequence models, even though CRNNs are not universal for continuous variable (CV) quantum computation. Concretely, we show unconditionally that there exists a class of CRNNs with $\mathrm{O}(n)$ quantum neurons that can express certain distributions that no "reasonable" classical model is able to represent without an $\Omega(n^2)$-dimensional latent space. Though this is only a quadratic separation in memory, the time complexity of inference for classical models is typically superlinear in the model size [48, 87–89], often yielding a superquadratic time separation. Surprisingly, this separation is true even when the classical model is *nonlinear* even though the quantum model is a quantization of a linear model.

Moreover, we are able to show directly that this quantum advantage is due to quantum contextuality [90–94] present in our quantum model. Intuitively, the source of our separation is the ability for the quantum model to efficiently store certain correlations that can be probed via different *measurement contexts*. More formally, we demonstrate that CV stabilizer states with contextual stabilizers have one-shot "partially" distinguishing measurement sequences in the sense that one of many (certain) hypotheses for a state can be ruled out with certainty using only one copy of the state. This phenomenon is summarized via the following (informally stated) lemma, proven in Appendix D:

Lemma 1.8 (Quantum contextuality yields partially distinguishing measurement sequences, informal statement of Lemma D.1)**.** *Consider three CV stabilizer states with contextual stabilizers exhibiting certain properties. There exists a one-shot "partially" distinguishing measurement sequence given by certain stabilizers of these states.*

This property essentially requires any classical simulation of intermediate mea-

surements on these stabilizer states to explicitly memorize the measurement context at any given point, as otherwise a tester can with certainty determine that the classical simulation algorithm is not accurately simulating the quantum system.

This inspires a method to prove a concrete memory separation between classical and quantum sequence models. We consider the modeling of a conditional distribution $p_{\text{data}}(\boldsymbol{m} \mid \boldsymbol{s})$ of measurement results $\boldsymbol{m}$ given a classical description of a sequence of displacement operators $\boldsymbol{s}$. The goal of the translation task is to output a sequence of real numbers that are consistent with quantum mechanics when sequentially performing phase estimation on these displacement operators when beginning in a fixed initial GKP state [95]. We call this task (when the sequence length is k and the displacement operators are on n qumodes) the (k, n) *stabilizer measurement translation task*. For arbitrary k, it is straightforward to see that a CRNN can sample from this distribution with n qumodes of memory. We first prove in Chapter 5 a separation on this task over *online* models—namely, classical machine learning model which act on the input sequence one word at a time.

Theorem 1.9 (Online stabilizer measurement translation memory lower bound, informal statement of Theorem D.2)**.** *Consider a "reasonable" online model with latent space L. If $\dim(L) < \frac{n(n-3)}{2}$, this model incorrectly translates an input sequence in the (k, n) stabilizer measurement translation task for all $k \geq n + 2$.*

We also show a more general separation over essentially *all* classical machine learning models at the cost of a longer input sequence length. Interestingly, the class of classical models this separation holds over includes Transformers [89], an example of which is the state-of-the-art sequence model GPT-4 [39].

Theorem 1.10 (General stabilizer measurement translation memory lower bound, informal statement of Theorem D.4)**.** *Consider a "reasonable" classical machine learning model with latent space L. If $\dim(L) < \frac{n(n-3)}{2}$, this model incorrectly translates an input sequence in the (k, n) stabilizer measurement translation task for all $k \geq n^2$.*

The strategy of these proofs involves demonstrating that classical machine learning models must not be injective over certain inputs if $\dim(L)$ is subquadratic in n, and

then showing that the associated states must satisfy the assumptions of Lemma 1.8. In Appendix D we also give an interpretation of our results as a memory lower bound on the classical simulability of certain CV circuits.

Finally, we also show empirically that these models are more efficient at performing a real-world translation task than classical sequence models of the same dimension, even when compared with state-of-the-art neural sequence models such as GRU RNNs [88] and Transformers [89]. This separation holds even when both the classical and quantum models have very similar numbers of parameters. These results taken together provide a promising avenue for showcasing a real (polynomial) quantum advantage on a practical task in a near-term experiment.

To our knowledge, these results are the first unconditional proofs of an expressivity separation between a quantum neural network and classical neural networks on classical data. By explicitly demonstrating that quantum contextuality is the source of this advantage, we are also able to provide intuition as to which classes of problems CRNNs are able to outperform traditional machine learning models in solving. Our numerics confirm the intuition that CRNNs perform extremely well on problems exhibiting linguistic contextuality.

CRNNs demonstrate that even the quantization of a very simple class of classical architectures—here, the class of linear recurrent neural networks (LRNNs)—is able to outperform a wide range of classical models on certain tasks, even if the classical models are much more powerful than LRNNs. In Chapter 6, we consider future directions for constructing quantum models that exhibit a larger memory separation over classical models while maintaining trainability.

Chapter 2

Critical Points in Quantum Generative Models

The results of this chapter were featured in Reference [49].

2.1 Introduction

One of the great successes of neural networks is the efficiency at which they are trained via gradient-based methods. Though training algorithms often involve the optimization of complicated, non-convex, high-dimensional functions, training via gradient descent in many contexts manages to converge to local minima that are good approximations of the global minimum in loss function value. This phenomenon has begun to be understood in the context of random matrix theory, particularly when applied to dense classifiers [44, 45].

Quantum generative models hold great promise in their ability to sample from probability distributions out of the reach of classical models [8, 96]. Though deep quantum generative models are believed to be difficult to train due to vanishing gradients [63, 64, 97], the situation for shallow models is murkier. Some shallow quantum models have empirically shown great promise in being trainable [70–72], while others have empirically been shown to suffer from poor distributions of local minima [69, 98]. Numerically, all of these models have been seen to experience a

phase transition in trainability: below some critical depth, local minima are poor approximators of the global minimum. Above this critical depth, they are good approximators. This transition has been poorly understood analytically, as typically the distribution of local minima monotonically improves as the size of the model increases [44, 45].

In this Chapter, we (for the first time) analytically show the presence of a computational phase transition in the training of a certain class of quantum generative models. To achieve this, we first show that this class of randomized quantum models is approximated in distribution by a Wishart random field on the hypertorus. We are then able to use techniques from Morse theory to exactly calculate the distribution of local minima (and general critical points) of this random field. Finally, we analyze this distribution in the limit of large model size, and analytically show the presence of this trainability phase transition. Roughly, we show that in this limit the expected density of local minima for a model with p parameters and Hilbert space dimension $\sim m$ (exponential in the problem size) at loss value $0 \leq E \leq \frac{1}{2}$ follows a generalization of the beta distribution [76]:

$$\mathbb{E}\left[\mathrm{Crt}_0\left(E\right)\right] \sim \mathrm{e}^{-mE} E^{m-\frac{p}{2}} \left(1-2E\right)^p . \tag{2.1}$$

This distribution experiences a transition in behavior at $p = 2m$: when $p < 2m$, local minima are exponentially concentrated (i.e. with width m^{-1}) far away from the global minimum $E = 0$, implying poor optimization performance in this regime. When $p \geq 2m$, this distribution is exponentially concentrated at $E = 0$, implying good optimization performance. We also verify our results numerically, demonstrating this concentration of minima even at small problem sizes. More specifically, when

$$\gamma \equiv \frac{p}{2m} = \mathrm{o}\left(\frac{1}{\log\left(n\right)}\right), \tag{2.2}$$

a superpolynomially small (in n) fraction of the local minima are within any constant additive energy error to the global minimum. This is typically extensive in the problem size when units are restored to Equation 2.1. We thus call γ the *overparame-*

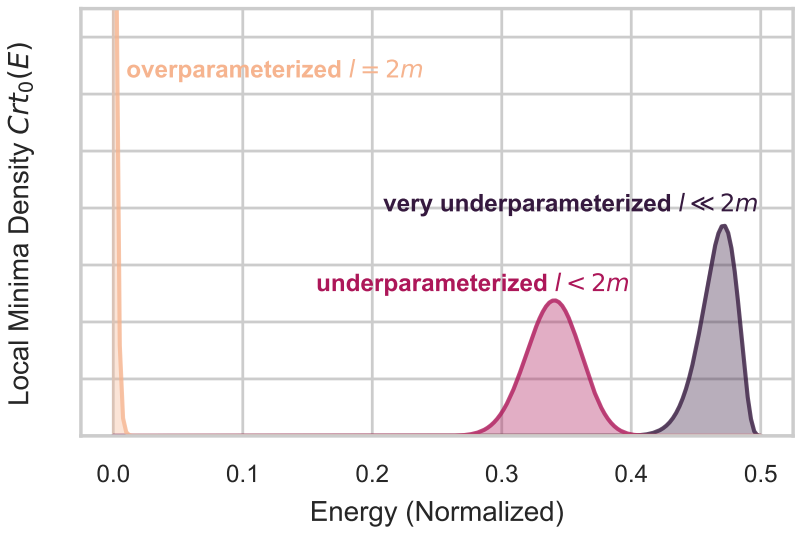


Figure 2-1: Plot of the asymptotic distribution of local minima of WHRFs with m degrees of freedom on the l-torus in: the extremely underparameterized regime, where $l \ll 2m$; the moderately underparameterized regime, where l is a finite fraction of $2m$; and at the critical overparameterization regime, where $l = 2m$. Here, the energy is scaled and shifted as per Equation (3.6) so that global minima have zero energy. In the underparameterized regime, only a fraction $\sim \exp(-m)$ of the critical points are within any constant additive error of the global minimum. In the overparameterized regime, local minima are exponentially concentrated at the global minimum.

terization factor. Representative plots of this distribution in various parameterization regimes are shown in Figure 2-1.

For the class of quantum generative models we consider, our results mirror the empirical results of [69, 98] in that only unreasonably overparameterized quantum models have good local minima. Though these results are pessimistic, we emphasize here that our results only apply to a certain class of quantum generative models. We are also able to give a heuristic explanation based on our proof techniques as to how one may be able to construct models of a reasonable size that are still trainable at the expense of computational overhead in implementing the model, as seen empirically in References [70–72].

2.2 Machine Learning Loss Landscapes as Random Fields

2.2.1 Random Fields on Manifolds

As in previous results on the loss landscapes of machine learning models [44, 45], we will map the distribution of a randomized class of quantum generative models to a random field on a manifold. This will then enable us to use standard mathematical techniques to study the distribution of critical points of the model.

Though they can be expressed in many ways, here we will be interested in random fields of the form:

$$F_{\mathrm{RF}}\left(\boldsymbol{\sigma}\right) \propto \sum_{i_1,\ldots,i_r,i'_1,\ldots,i'_r=1}^{\Lambda} \sigma_{i_1}\ldots\sigma_{i_r} J_{i_1,\ldots,i_r,i'_1,\ldots,i'_r}\sigma_{i'_1}\ldots\sigma_{i'_r}. \tag{2.3}$$

Here, $\boldsymbol{\sigma} \in M$ is some point on a manifold, and $\boldsymbol{J}$ is a random variable. In the context of most studies of machine learning loss landscapes, M is typically the hypersphere and $\boldsymbol{J}$ a symmetric matrix of i.i.d. Gaussian random variables, i.e. a Gaussian orthogonal ensemble (GOE) matrix.

We will instead find that variational quantum algorithms (VQAs) are naturally

described as *Wishart hypertoroidal random fields* (WHRFs). For these models, the manifold M is a tensor product embedding of the hypertorus into exponentially large Euclidean space; that is, points on this embedding are described by the Kronecker product:

$$\boldsymbol{\sigma} = \bigotimes_i \begin{pmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{pmatrix} \tag{2.4}$$

for angles $-\pi \leq \theta_i \leq \pi$. Furthermore, in these models, $\boldsymbol{J}$ is drawn from a normalized complex Wishart distribution. The complex Wishart distribution is a natural multivariate generalization of the gamma distribution, and is given by the distribution of the square of a complex Gaussian random matrix. Specifically, for $\boldsymbol{X} \in \mathbb{C}^{n \times m}$ a matrix with i.i.d. complex Gaussian columns with covariance matrix $\boldsymbol{\Sigma}$, the matrix

$$\boldsymbol{W} = \frac{1}{m} \boldsymbol{X} \cdot \boldsymbol{X}^\dagger \tag{2.5}$$

is normalized complex Wishart distributed with scale matrix $\boldsymbol{\Sigma}$ and m degrees of freedom. Throughout this thesis we use the notation $\boldsymbol{W} \sim \mathcal{CW}_n(m, \boldsymbol{\Sigma})$ to denote a $\boldsymbol{W}$ drawn from such a complex Wishart distribution, and similarly $\boldsymbol{W} \sim \mathcal{W}_n(m, \boldsymbol{\Sigma})$ when drawn from a real Wishart distribution. We will find that the degrees of freedom m will greatly affect the distribution of local minima of the WHRF, and thus also of the class of quantum generative models that we consider.

2.2.2 Quantum Generative Models as Random Fields

We first show that a certain randomized class of Hamiltonian agnostic VQAs—as described in Section 1.2.2—can be expressed as WHRFs. In most quantum generative models, various θ_i are completely dependent, e.g. $\theta_i = \theta_{i+5}$ for all i [61]. For simplicity, we assume throughout this Chapter that each independent parameter appears a constant number r times in the model, and that the total number of independent parameters is given by $p = q/r$.

This demonstration of the convergence of certain VQAs to WHRFs will allow us to more easily study the critical points of the model using techniques from random

matrix theory. Though we leave the full statement and proof for Appendix A.1, we give an informal statement and discussion here.

Theorem 2.1 (VQAs as WHRFs, informal statement of Theorems A.1 and A.2)**.** *Consider the class of models*

$$|\boldsymbol{\theta}\rangle \equiv \prod_{i=1}^{q} U_i(\boldsymbol{\theta}) |\psi_0\rangle \equiv \prod_{i=1}^{q} \mathrm{e}^{-\mathrm{i}\theta_i Q_i} |\psi_0\rangle, \tag{2.6}$$

where each Q_i is drawn uniformly from the Pauli group $\mathbb{P}_n$ and $|\psi_0\rangle$ is the first column of a uniformly random member of the Clifford group. Let p be the number of distinct θ_i, and let $r = q/p$. Under reasonable assumptions on the eigenvalues of H (with minimum eigenvalue λ_1 and mean eigenvalue $\overline{\lambda}$), the random loss function

$$F_H(\boldsymbol{\theta}) = \frac{\langle\boldsymbol{\theta}| H |\boldsymbol{\theta}\rangle - \lambda_1}{\overline{\lambda} - \lambda_1} \tag{2.7}$$

converges in distribution to the random field

$$F_{WHRF}(\boldsymbol{\theta}) = \sum_{i_1,\ldots,i_r,i'_1,\ldots,i'_r=1}^{2^p} w_{i_1} \ldots w_{i_r} J_{i_1,\ldots,i_r,i'_1,\ldots,i'_r} w_{i'_1} \ldots w_{i'_r}, \tag{2.8}$$

where $\boldsymbol{w}$ are points on the hypertorus $(S^1)^{\times p}$ parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{J}$ is a complex Wishart random matrix normalized by its number of degrees of freedom m.

Note that for convenience, we have shifted and scaled the typical VQA loss such that it is always greater than zero and independent from overall scalings of the problem Hamiltonian.

In the course of this mapping, we find that the *degrees of freedom* of the Wishart matrix $\boldsymbol{J}$ (formally a real number) is given by the ratio:

$$m \equiv \frac{\|H - \lambda_1\|_*^2}{\left\|H - \overline{\lambda}\right\|_{\mathrm{F}}^2}. \tag{2.9}$$

Here, $\|\cdot\|_*$ denotes the nuclear norm, and $\|\cdot\|_{\mathrm{F}}$ the Frobenius norm. Generally, this ratio is exponential in n, particularly when modeling the class of ground states typi-

cally represented by VQAs [19, 61, 99]. Though we are unable to prove Theorem 2.1 for Hamiltonian informed models, there are heuristic reasons to believe that they are described by a similar random field with $m = \mathrm{O}(\mathrm{poly}(n))$, as opposed to Equation (2.9) (see Appendix A.1.3 and empirical evidence in Section 2.4). We will later find that a number of independent model parameters p that is twice the degrees of freedom m of the matrix $\boldsymbol{J}$ marks the transition from the underparameterized to the overparameterized regime of F_{WHRF}, where the quality of local minima improves.

The general idea for showing this equivalence relies on the *path integral expansion* of the VQA loss function. Effectively, this is just a Taylor expansion of the unitary matrices composing the model, which is exact even at a finite number of terms. One can then show that terms in this expansion can be assumed independent with negligible error in distribution, and then show that the resulting random process is asymptotically a WHRF. The reasonable assumptions on the eigenvalues of H are essentially just a requirement that the eigenvalues of H are not "unnaturally" spread out; for the quantum states VQAs typically model, this is never the case. We give a full description of these requirements with the full proof in Appendix A.1.

2.3 The Loss Landscape of Wishart Hypertoroidal Random Fields

2.3.1 Exact Results

Having shown that VQAs can be described as WHRFs, we now focus discussion entirely on WHRFs. Our strategy for showing the distribution of critical points of this random field will be similar to that in Reference [73], where similar results were shown for Gaussian spherical random fields. Namely, we will lean heavily on the *Kac–Rice formula*, which gives the expected number of critical points of a certain index at a given range of function values for random fields on manifolds. We give an informal description of the Kac–Rice formula here, with the formal version given in Appendix A.2.

Lemma 2.2 (Kac–Rice formula [100], informal statement of Lemma A.3)**.** *Let M be a compact, oriented manifold. Assume a random field $F(\boldsymbol{\sigma})$ on M is sufficiently nice. Then, the number of critical points of index at most k with $F(\boldsymbol{\sigma}) \in B$ for an open set $B \subset \mathbb{R}$ is*

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{Crt}_k(B)\right] \\
&\quad = \int_M \mathbb{E}\left[\left|\det\left(\boldsymbol{\nabla}^2 F(\boldsymbol{\sigma})\right)\right| \mathbf{1}\left\{F(\boldsymbol{\sigma}) \in B\right\} \mathbf{1}\left\{\iota\left(\boldsymbol{\nabla}^2 F(\boldsymbol{\sigma})\right) \leq k\right\} \mid \boldsymbol{\nabla} F(\boldsymbol{\sigma}) = 0\right] \\
&\quad \times p_{\boldsymbol{\sigma}}\left(\boldsymbol{\nabla} F(\boldsymbol{\sigma}) = 0\right) \mathrm{d}\boldsymbol{\sigma}\,,
\end{aligned} \tag{2.10}$$

where $\boldsymbol{\nabla}\cdot$ is the covariant gradient, $\iota(\cdot)$ is the index of $\cdot$, $p_{\boldsymbol{\sigma}}$ is the probability density of $\boldsymbol{\nabla} F(\boldsymbol{\sigma})$ at $\boldsymbol{\sigma}$, and $\mathrm{d}\boldsymbol{\sigma}$ is the volume element on M.

From Lemma 2.2, we see that when the joint distribution of $\boldsymbol{\nabla}^2 F$, $\boldsymbol{\nabla} F$, and F is known, then the expected number of critical points with function values in an open set B can be calculated. Perhaps surprisingly, as in the Gaussian case, the joint distribution of these derivatives for WHRFs is fairly simple. Once again leaving the full proof for Appendix A.3, we show the distribution of the Hessian conditioned to be at a critical point of function value x can be described by the shifted sum of a Wishart matrix with an independent GOE matrix. Similarly, the distribution of the gradient conditioned on the function value being x is given by a normal distribution.

Lemma 2.3 (Hessian and gradient distributions, informal statement of Lemma A.4)**.** *The scaled Hessian $m\partial_i\partial_j F_{WHRF}(\boldsymbol{w})$ conditioned on $F_{WHRF}(\boldsymbol{w}) = x$ and $\partial_k F_{WHRF}(\boldsymbol{w}) = 0$ is distributed as*

$$m\tilde{C}_{ij}(x) = -2rmx\delta_{ij} + rW_{ij} + r\sqrt{2mx}N_{ij}, \tag{2.11}$$

where $\boldsymbol{W}$ is Wishart distributed with $2m$ degrees of freedom, $\boldsymbol{N}$ GOE distributed, and they are independent. Furthermore, the scaled gradient $m\partial_k F_{WHRF}(\boldsymbol{w})$ conditioned on $F_{WHRF}(\boldsymbol{w}) = x$ is distributed as

$$m\tilde{G}_k(x) = \sqrt{2mrx}N_k, \tag{2.12}$$

where N_k are i.i.d. standard normal distributions independent from all W_{ij} and N_{ij}.

With all of the pieces in place, we are able to explicitly calculate the expected distribution of local minima in WHRFs via the Kac–Rice formula (with full calculations left for Appendix A.3). In Section 2.4 we find empirical evidence that these results hold not only in expectation, but in distribution; we leave further analytic investigation of this to future work.

Theorem 2.4 (Distribution of critical points in WHRFs, informal statement of Theorem A.5)**.** *Let*

$$\mu_{\boldsymbol{C}(x)} = \frac{1}{p}\sum_{i=1}^{p} \delta\left(\lambda_i^{\boldsymbol{C}(x)}\right) \tag{2.13}$$

be the empirical spectral measure of the random matrix

$$\boldsymbol{C}(x) = \frac{r}{m}\left(\boldsymbol{W} + \sqrt{2mx}\boldsymbol{N}\right), \tag{2.14}$$

where $\boldsymbol{W}$ is Wishart distributed with $2m$ degrees of freedom, $\boldsymbol{N}$ GOE distributed, and they are independent. $\lambda_i^{\boldsymbol{C}}(x)$ is the ith smallest eigenvalue of $\boldsymbol{C}(x)$. Then, the distribution of the expected number of critical points of index k of a WHRF at a function value $E > 0$ is given by

$$\begin{aligned}&\mathbb{E}\left[\mathrm{Crt}_k(E)\right]\\ &= \left(\frac{\pi}{r}\right)^{\frac{p}{2}} \Gamma(m)^{-1} m^{(1+\gamma)m} \mathbb{E}_{\boldsymbol{C}(E)}\left[\mathrm{e}^{p\int \ln(|\lambda-2rE|)\mathrm{d}\mu_{\boldsymbol{C}(E)}} \mathbf{1}\left\{\lambda_{k+1}^{\boldsymbol{C}(E)} \geq 2rE\right\}\right] E^{(1-\gamma)m-1}\mathrm{e}^{-mE},\end{aligned} \tag{2.15}$$

where

$$\gamma = \frac{p}{2m}. \tag{2.16}$$

We call the parameter γ the *overparameterization factor*. It describes the ratio between the number of parameters of the model p and twice the degrees of freedom m of the model. We will later find that γ governs the phase transition between an *underparameterized phase* of the model—where local minima are far from the

global minimum—and an *overparameterized phase*—where local minima are good approximators of the global minimum.

2.3.2 Asymptotic Results as $p \to \infty$

Though Theorem 2.4 gives the exact distribution of critical points, it is difficult to use in practice. This difficulty comes from the expectation over eigenvalues of the sum of independent Wishart and Gaussian matrices. Surprisingly, however, the eigenvalues of both Wishart and Gaussian orthogonal matrices converge to *fixed* distributions. Essentially, asymptotically in the size of the matrix, the eigenvalue distribution of all normalized Wishart matrices are the same (given by the *Marchenko–Pastur distribution*) and the eigenvalue distribution of all Gaussian orthogonal matrices are the same (given by the *Wigner semicircle distribution*).

Luckily, we can characterize the asymptotic distribution of eigenvalues of the sums of these matrices using the tools of *free probability theory*. Roughly, free probability theory is the probability theory of noncommutative random variables (e.g. random matrices). As the distribution of the sum of two random variables in commutative probability theory can be described by the convolution of the distributions of the two independent random variables, so can the *free convolution* of the distributions of two *freely independent* noncommutative random variables. Using the asymptotic free independence of Wishart and Gaussian orthogonal random variables, we are able to show that asymptotically the eigenvalue distribution of their sum weakly converges to the free convolution of a Marchenko–Pastur distribution with a semicircle distribution.

However, weak convergence is not enough; due to the exponential factor in the expectation in Theorem 2.4, any large deviations from the asymptotic convergence—even if they occur with exponentially vanishing probability—can potentially cause large deviations from the naive application of free probability theory. Thus, our results rely on using large deviations theory to show that to (logarithmic) leading order these deviations do not contribute to the final result. This is due to the contribution to the expectation from the deviations being dominated by what is predicted by free probability theory. These results can be summarized via the following theorem

(proved in Appendix A.4):

Theorem 2.5 (Logarithmic asymptotics of the local minima distribution, informal statement of Theorem A.10)**.** *Let* $\mathrm{d}\mu_E^*$ *be the free convolution of a scaled Marchenko–Pastur and scaled Wigner semicircle distribution, with* $\lambda_{E,1}^*$ *the infimum of its support. Let* $p, m \gg 1$ *with* $\frac{p}{2m} = \gamma = \mathrm{O}(1)$*. Then, the expected distribution of local minima of a WHRF at a fixed function value* $E > 0$ *is given by*

$$\begin{aligned}\frac{1}{p}\ln\left(\mathbb{E}\left[\mathrm{Crt}_0\left(E\right)\right]\right) = \frac{1}{2}\ln\left(\frac{\pi q}{2\gamma}\right) + \frac{1}{2\gamma}\left(1-E\right) + \frac{1}{2}\left(\gamma^{-1}-1\right)\ln\left(E\right) \\ + \int \ln\left(\left|\frac{\lambda}{r} - 2E\right| \mathbf{1}\left\{\frac{\lambda_{E,1}^*}{r} \geq 2E\right\}\right)\mathrm{d}\mu_E^* + \mathrm{o}\left(1\right).\end{aligned} \tag{2.17}$$

Note that, though we only prove the asymptotic distribution of local minima in Theorem 2.5, we expect similar theorems to also hold for critical points of constant index k (taking $\lambda_{E,1}^* \mapsto \lambda_{E,k}^*$ in the integrand). The only difference in the derivation is the exact form of the large deviations of the kth smallest eigenvalue of $\boldsymbol{C}\left(x\right)$. This is similar to the case in Gaussian hyperspherical random fields, which are often used to model neural network loss functions [44, 45, 73].

2.3.3 Discussion of the Critical Point Distribution

Let us now discuss the implications of Theorem 2.5. Note first that the rescaled logarithmic number of critical points diverges when $p \to \infty$. Following the derivation closely, one finds that this is due to an exponentially suppressed (when m is exponential in n and r is fixed) gradient. We believe that this is a manifestation of the "barren plateau" phenomenon, where for many deep VQA models it can be shown that there is an exponentially vanishing variance of the gradient over the loss landscape [63, 64, 97]. This interpretation suggests that these barren plateau regions are filled with many small "bumps" that are exponentially shallow. Our methods extend the typical barren plateau analysis, though, as we are also able to study a regime without barren plateaus by considering models with sufficiently large r (see

Lemma 2.3). Furthermore, note that this class of random fields exhibits banded behavior in the eigenvalues. That is, local minima only exist in the band $0 \leq E \leq E_0$, where E_0 is the solution to

$$\lambda^*_{E_0,1} = 2rE_0. \tag{2.18}$$

This banded behavior is similar to that in the Gaussian spherical case. We will see, however, that this does not give necessarily good guarantees on the distribution of local minima. This is due to E_0 being generally far from 0 when $\gamma < 1$ as $p, m \to \infty$. To illustrate this, we focus now on two cases: $p \geq 2m$ (the *overparameterized regime*) and $p \ll m$ (the *underparameterized regime*).

$p \geq 2m$

First, let us consider when $p \geq 2m$, i.e. $\gamma \geq 1$. In this limit, the Wishart term of $\boldsymbol{C}$ is low-rank, and μ^*_E has support on eigenvalues ≤ 0 for all $E \geq 0$. Therefore, the condition $\lambda^*_{E,1} \geq 2Er$ is never satisfied, and to leading order in p there is a vanishing fraction of local minima at any function value $E > 0$. That is, all local minima are global minima in the $p \to \infty$ limit when $\gamma \geq 1$. Though the choice of model is slightly different, we suspect that a related phenomenon may be what gives rise to the phase transition in training numerically observed in References [69–72, 98].

$p \ll m$

When the number of distinct parameters p is poly (n) and considering a physically relevant problem Hamiltonian such that the number of degrees of freedom m is exp (n), we have that $p \ll m$ (i.e. $\gamma \ll 1$) for large n. In this limit, the spectral distribution μ^*_E is dominated by the Wishart term of $\boldsymbol{C}$, as its eigenvalues are O (1) while the eigenvalues of the GOE term are O $\left(\sqrt{\gamma}\right)$. Furthermore, the Marchenko–Pastur distribution in this limit only has support at $\lambda = 1 + \mathrm{O}\left(\sqrt{\gamma}\right)$. Therefore, the expected

number of local minima at a function value E will be proportional to

$$\mathbb{E}\left[\mathrm{Crt}_0\left(E\right)\right] \propto \mathrm{e}^{-mE+\mathrm{o}(p)} E^{m-\frac{p}{2}} \left(1-2E+\mathrm{O}\left(\sqrt{\gamma}\right)\right)^p \mathbf{1}\left\{0 \leq E+\mathrm{O}\left(\sqrt{\gamma}\right) \leq \frac{1}{2}\right\}. \tag{2.19}$$

In particular, up to shifts on the order of $\sqrt{\gamma}$, the distribution of local minima is roughly that of a compound confluent hypergeometric (CCH) distribution [76]. The CCH distribution can be considered a generalization of the beta distribution, and for our parameters has mean on the order of $\frac{1}{2}-\gamma$ and standard deviation on the order of m^{-1}. Restoring the overall scaling in Equation (2.7), this implies that in this limit the local minima of the variational loss function exponentially concentrate (in expectation) near half the mean eigenvalue of $H-\lambda_1$ instead of the smallest eigenvalue. Even worse, the CCH distributed form of the local minima implies that, even when beginning at an initial function value well below half of the mean eigenvalue of $H-\lambda_1$, the found loss will only improve by a fraction of the initial function value before the optimization algorithm finds a local minimum. This is insufficient to find the optimal loss to constant additive error when beginning training at a random point, as is often the goal in VQAs [19]. Empirically, we find that this occurs not just in expectation but also for individual model instances in Section 2.4.

2.4 Numerical Experiments

We now test our analytic predictions using numerical simulations. First, we investigate the empirical performance of the class of randomized models we study theoretically, and give numerical evidence of things we were unable to prove. Then, we give numerical evidence that, for models dependent on the objective Hamiltonian, the effective degrees of freedom parameter m can be much smaller than predicted. In all cases, we numerically test the predictions of our results by modeling the ground state of the 1D n site *spinless Fermi–Hubbard Hamiltonian* [101] at half filling. Here, we take units such that the mean eigenvalue of the considered Hamiltonian (minus its smallest eigenvalue) is $E = 1$. We give further details of our numerical simulations

in Appendix A.5.

2.4.1 Empirical Performance of Random Ansatzes

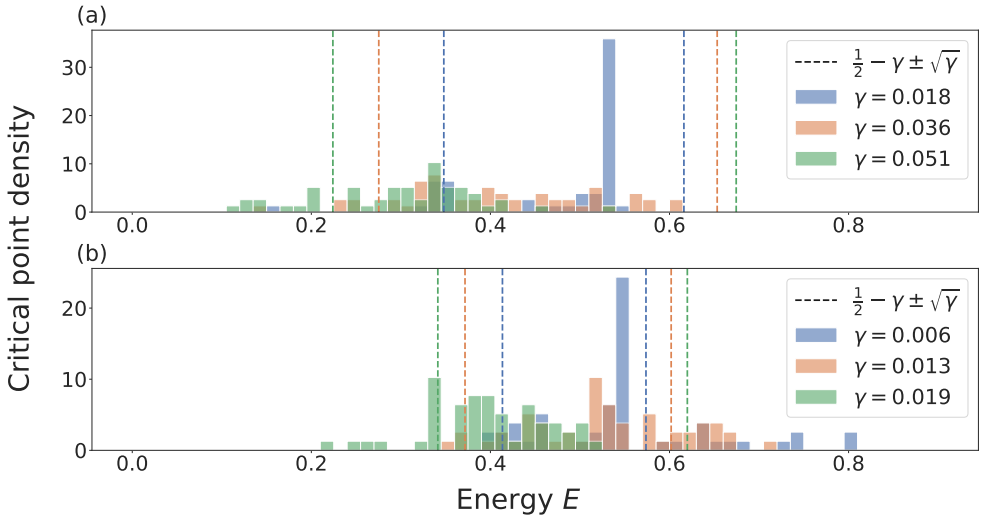


Figure 2-2: Here we plot the distribution of found local minima found after 52 separate training instances using the randomized model on (a) $2^n = 64$- and (b) $2^n = 256$-dimensional models. Dashed lines denote the predicted region local minima will lie. Note the clustering of local minima at a finite function value when $\gamma \ll 1$.

First, we analyzed the performance of a VQA on this loss function via the random model construction procedure defined in Theorem 2.1. Previous numerical results on related Hamiltonian agnostic ansatzes have already shown the concentration of local minima far away in loss value from the global minimum below some degrees-of-freedom transition, and concentration at the global minimum above this transition [69, 98]. Here, we tracked where our analysis predicts the local minima to lie as a function of γ for $\gamma \ll 1$, up to deviations on the order of $\sqrt{\gamma}$ that arise from numerically considering the problem at finite size (as discussed in Section 2.3.3).

Concretely, for a given training instance and depth $q = p$, we generated an ansatz $|\boldsymbol{\theta}\rangle$ composed of p layers of Pauli rotations, where each Pauli rotation was chosen uniformly from all nonidentity Pauli matrices on n qubits. The numbers of model layers we consider are typical of current physical implementations of Hamiltonian agnostic VQAs [102]. A summary of the normalized distribution of found local minima

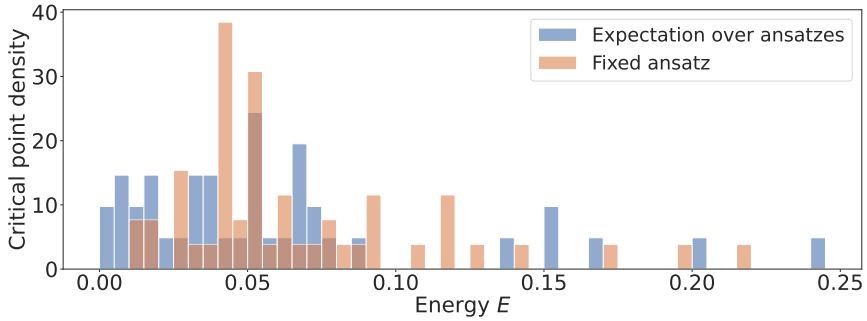


Figure 2-3: Here we plot the distribution of found local minima found after 52 separate training instances using the randomized model, with $p = 48$ and $2^n = 64$ model dimension. For even a small model size, qualitatively the expected distribution of critical points and the distribution of critical points for a fixed random ansatz are in agreement.

for the randomized model with model dimension $2^n = 64,256$ is given in Figure 2-2, along with the predicted region in which all local minima should lie in the $p \to \infty$ limit as discussed in Theorem 2.5. See Appendix A.5 for details on how this distribution was generated.

We see that almost all found local minima lie within the predicted region, even at small p, n. In particular, for small γ, the distribution of local minima is almost entirely localized within $\sqrt{\gamma}$ of the predicted $\frac{1}{2} - \gamma$ (in units of the mean eigenvalue of $H - \lambda_1$). Finally, we numerically observe that the distribution of local minima are qualitatively similar in expectation and for a single choice of random model in Figure 2-3.

2.4.2 Empirical Performance of a Hamiltonian Informed Model

Previous numerical results [70] on VQAs have shown that only a moderate number of model parameters suffices for efficient training when using a Hamiltonian informed model. As discussed in Section 2.2.2, we believe this is due to this class of models effectively limiting the degrees of freedom m of the associated WHRF model; to test this, we performed more numerical experiments using a Hamiltonian informed ansatz. We once again tracked where our analysis predicts the local minima to lie as a function of γ for $\gamma \ll 1$, up to deviations on the order of $\sqrt{\gamma}$.

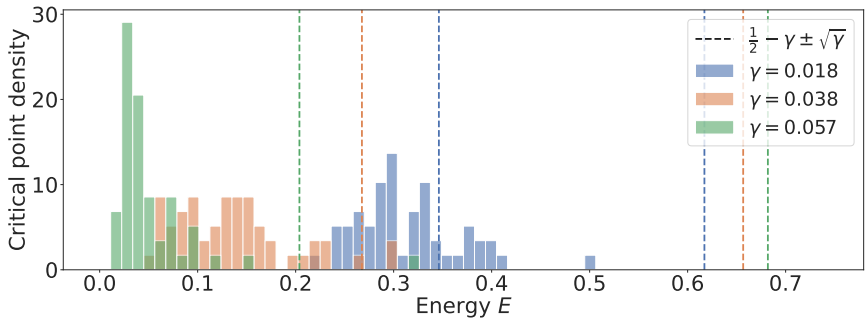


Figure 2-4: Here we plot the distribution of found local minima after 52 separate training instances using a Hamiltonian informed model. Dashed lines denote the predicted region local minima will lie. We see that the predicted region is overly pessimistic. We believe that this is due to the Hamiltonian informed model lowering the effective degrees of freedom m of the WHRF instance the ansatz maps to; see Section 2.2.2.

We show the empirical distribution of local minima in Figure 2-4 for $2^n = 256$, along with the predicted region local minima should lie as discussed in Section 2.3.3. The predicted local minima distribution is overly pessimistic (particularly at larger p). We suspect this is due to the fact that the ansatz is constructed in a way that minimizes the effective degrees of freedom of the WHRF m such that γ is close to 1 for smaller p than is predicted analytically.

2.5 Conclusion

Though variational quantum algorithms are perhaps the most promising way to use the error-prone quantum devices of today for practical computational tasks, there are many caveats with regard to their trainability. In particular, previous work has shown that utilizing deep quantum models that are independent of the problem Hamiltonian can introduce a vanishing gradient phenomenon where, though the model is expressive enough to capture the ground state of interest, in practice optimizing the loss function is infeasible [63, 64, 97]. We extended these results by showing a particular class of random models independent of the problem instance not only can exhibit these vanishing gradients at large depth, but also has a concentration of local minima near

the mean eigenvalue of the objective Hamiltonian. This is in contrast to the case in traditional neural networks, where even generic model structure tends to lead to a concentration of local minima in a band near the global minimum of the loss function.

Though our results may not seem encouraging for quantum generative models, we emphasize that we expect our analytic results to hold only when the model is independent of the problem Hamiltonian. Indeed, we found empirically (and heuristically) good performance for a particular Hamiltonian informed ansatz, where our analytic results seem much too pessimistic. In principle, this new way of thinking about variational quantum algorithms may inform future quantum generative model design; we leave for future work the study of how various model choices may impact the distribution of critical points of the loss function positively, and how practical considerations such as noisy model implementations may play a role.

Chapter 3

Quantum Variational Algorithms Are Swamped With Traps

The results of this chapter were featured in Reference [50], work done in collaboration with Bobak T. Kiani.

3.1 Introduction

The trainability of classical neural networks via simple gradient-based methods is one of the most important factors leading to their general success on a wide variety of problems. This is particularly exciting given the variety of no-go results via statistical learning theory, which demonstrate that in the worst case these models are not trainable via stochastic gradient-based methods [103–106]. There has been recent hope that variational quantum algorithms—the quantum analog of traditional neural networks—may inherit these nice trainability properties from classical neural networks. Indeed, in certain regimes [107], training algorithms exist such that the resulting quantum model provably outperforms certain classical algorithms. This would potentially enable the use of quantum models to efficiently represent complex distributions which are provably inefficient to express using classical networks [96].

Unfortunately, such good training behavior is not always the case in quantum models. There have been previous untrainability results for deep variational quan-

tum algorithms due to vanishing gradients [63–65, 97], and in Chapter 2 we showed the untrainability of shallow nonlocal models due to poor local minima; however, no such results were known for shallow, local quantum models with local cost functions. Indeed, there have been promising preliminary numerical experiments on the performance of variational quantum algorithms in these regimes, but typically have relied on good initialization [108] or highly symmetric problem settings [70–72] to show convergence to a good approximation of the global optimum.

Here, we show that generally such models are not trainable, particularly when a good choice of initial point is not known and when the model does not exhibit a high amount of symmetry. We first prove general untrainability results in the presence of noise using techniques from statistical query learning theory. Surprisingly, these results hold for all learning problems in a wide range of variational learning settings, and in many scenarios even when the magnitude of the noise is exponentially small in the problem size. We then consider the trainability of models that may not have noise by studying their typical loss landscapes. We prove that, for typical model instances, local minima concentrate far from the global optimum even for certain local shallow circuits that do not suffer from barren plateaus. This phenomenon can be visualized in Figure 3-1, where the training landscape for a shallow QCNN learning a random instance of itself is shown to concentrate far from the global optimum. As in Chapter 2, this phenomenon is the result of a trainability phase transition in the loss landscape of the quantum model. In Chapter 2, this transition was governed by the ratio of the number of parameters to the Hilbert space dimension; we show in the shallow case that instead, this transition is governed by the ratio of the local number of parameters to the local Hilbert space dimension, in the reverse light cone of a given measured observable. As this is typically much less than 1 for local variational ansatzes, these models are typically untrainable. We then give numerical evidence of this fact, and conclude by studying where there may be reason for optimism in the training of certain variational quantum models.

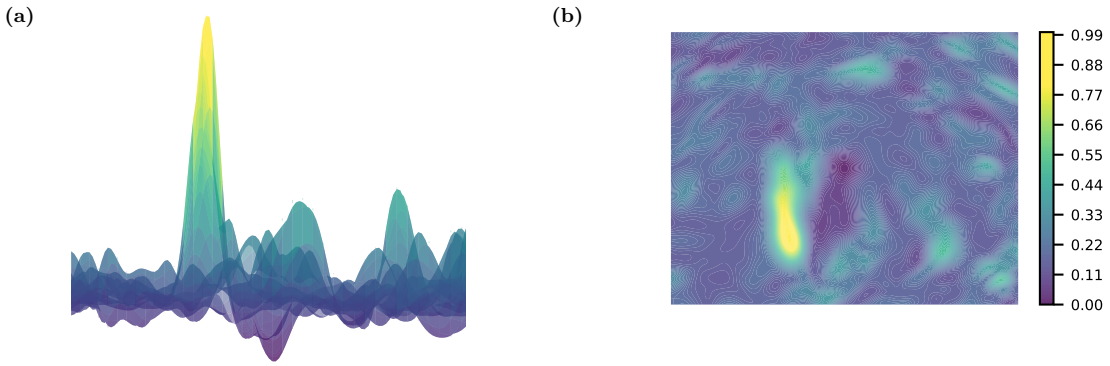


Figure 3-1: Loss landscapes of underparameterized quantum variational algorithms generally appear "bumpy," filled with various local minima and traps. Here, we plot the loss landscape as a (a) surface and (b) contour plot along two random normalized directions for the teacher-student learning task of the QCNN for 14 qubits. Though a global minimum is located at the center of the plot, finding this global minima is generally challenging due to the shape of the loss landscape. Details of this visualization are given in Appendix B.6.

3.2 Statistical Query Learning

3.2.1 The Statistical Query Learning Framework

We first give a brief overview of the classical statistical query (SQ) setting here. A more detailed review is provided in Appendix B.2.

Let $\mathcal{D}$ be a distribution on an input space $\mathcal{X}$. Consider an output space $\mathcal{Y}$, and let $c : \mathcal{X} \to \mathcal{Y}$ be a target function. In the classical SQ setting, one queries the SQ model by inputting a function f and receiving an estimate of $\mathbb{E}_{x\sim\mathcal{D}}\left[f\left(x, c\left(x\right)\right)\right]$ within a given tolerance τ. As an example, one can query a loss function ℓ for a model $m_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$ by querying the function $\ell\left(m_{\boldsymbol{\theta}}\left(x\right), c\left(x\right)\right)$.

A special class of statistical queries are *inner product queries*, where query functions g are defined only on $\mathcal{X}$ and the correlational statistical query returns an estimate:

$$\langle g, c\rangle_{\mathcal{D}} \equiv \mathbb{E}_{x\sim\mathcal{D}}\left[g\left(x\right) c\left(x\right)\right] \tag{3.1}$$

within a specified tolerance τ. Correspondingly, we define the norm:

$$\|g\|_{\mathcal{D}} \equiv \sqrt{\langle g, g \rangle_{\mathcal{D}}}. \tag{3.2}$$

This inner product and norm lead to a natural definition of the *statistical query dimension* [103, 109], which intuitively is the number of distinguishable elements in a concept class under this inner product.

Definition 3.1 (Statistical query dimension [103, 109])**.** For a distribution $\mathcal{D}$ and concept class $\mathcal{H}$ where $\|M\|_{\mathcal{D}}^2 \leq C_{\max}$ for all $M \in \mathcal{H}$, the statistical query dimension ($\text{SQ-DIM}_{\mathcal{D}}(\mathcal{H})$) is the largest positive integer d such that there exists d elements $M_1, M_2, \ldots, M_d \in \mathcal{H}$ such that for all $i \neq j : \left|\langle M_i, M_j \rangle_{\mathcal{D}}\right| \leq C_{\max}/d$.

Standard results in SQ theory directly relate this quantity to the difficulty in learning a hypothesis class. We use the term "learn" here loosely, but give a formal definition in Appendix B.2.1.

Theorem 3.2 (Query complexity of learning [103, 104])**.** *Given a distribution $\mathcal{D}$ on inputs and a hypothesis class $\mathcal{H}$ where $\|M\|_{\mathcal{D}}^2 \leq C_{max}$ for all $M \in \mathcal{H}$, let $d = \text{SQ-DIM}_{\mathcal{D}}(\mathcal{H})$ be the statistical query dimension of $\mathcal{H}$. Any learner making queries with tolerance $C_{max}\tau$ must make at least $(d\tau^2 - 1)/2$ queries to learn $\mathcal{H}$ up to error $C_{max}\tau$.*

3.2.2 Quantum Machine Learning in the Statistical Query Framework

Quantum machine learning algorithms are inherently noisy due to both unavoidable sources of error—such as shot noise from sampling outputs—or potentially correctable sources of error such as gate errors and state preparation noise. In such noisy settings, the SQ model provides a useful framework for quantifying the complexity of learning a class of functions by considering how many query calls to a noisy oracle are needed to learn any function in that class [104, 110, 111]. In this setting, we consider the optimization of a risk of the form of Equation (1.7). We assume there

Setting (n qubits, L layers)	Query complexity ($\beta < 1/2^*$)
$L = 1$, global measurement, 1-local gates	$2^{\Omega(n)}$ if $\tau \geq 3^{-\beta n}$
$L = \lceil \log_2(n) \rceil$, single qubit measurement, global 1- and 2-local gates	$2^{\Omega(n)}$ if $\tau \geq 4^{-\beta n}$
$L \ll n$, single qubit measurement, neighboring 1- and 2-local gates on a d-dim. lattice	$2^{\Omega\left(L^d\right)}$ if $\tau = \Omega(1)^{**}$
$L = 1$, single qubit gates, unitary learning	$2^{\Omega(n)}$ if $\tau \geq 4^{-\beta n}$

* Technically, we require $\beta = 1/2 - \Omega(1)$; ** $\tau = 2^{-\,\omega\left(\min\left(2L, n^{1/d}\right)^d\right)}$ is sufficient.

Table 3.1: Relatively simple classes of functions require exponentially many statistical queries to learn using any naive algorithm that reduces to statistical queries. The table above quantifies the number of queries needed to identify a target function in the function class, over a distribution of states that forms a 2-design and with queries that have tolerance $C_{\max}\tau$ (query tolerance lower bounded by a constant times $C_{\max}$ suffices in all cases).

is a target observable M that we would like to learn on some distribution over states $\mathcal{D}$. We define a correlational statistical query $\mathrm{qCSQ}(O, \tau)$, which takes in a bounded observable O with $\|O\|_\infty \leq 1$ (where $\|\cdot\|_\infty$ is the operator norm) and a tolerance τ and returns a value in the range:

$$\mathbb{E}_{\rho\sim\mathcal{D}}\left[\mathrm{tr}\left(O\rho\right)\mathrm{tr}\left(M\rho\right) - \tau\right] \leq \mathrm{qCSQ}\left(O, \tau\right) \leq \mathbb{E}_{\rho\sim\mathcal{D}}\left[\mathrm{tr}\left(O\rho\right)\mathrm{tr}\left(M\rho\right) + \tau\right]. \tag{3.3}$$

Note that there are no guarantees on the form of the additive error other than it is within the tolerance τ, and may for instance depend on the observable being queried O. Though SQ oracle calls may at first appear unrelated to variational algorithms, we show in Appendix B.2.1 that many common variational optimizers in the presence of noise of the magnitude τ reduce to calls to an SQ oracle; for instance, commonly used first and second order optimization algorithms fall within the framework of the SQ model we consider, like the parameter shift rule for analytically evaluating gradients [112, 113]. In Appendix B.2.1, we also describe an analogous SQ model for learning unitaries.

To quantify the hardness of learning variational circuits, we consider the task of learning certain function classes generated by shallow variational circuits over a distribution of inputs $\mathcal{D}$ which forms a 2-design. Our results also generally hold

for distributions that are uniform over states in the computational basis, recovering the statistical query setting for classical Boolean functions. Table 3.1 summarizes the number of queries needed to learn various function classes which are generated by variational circuits, with proofs deferred to Appendix B.3. In all settings we consider, an exponential number of queries (in either n or the light cone size) are needed to learn simple classes, such as the class of functions generated by single qubit gates followed by a fixed global measurement. This hardness intuitively arises because each individual query can only obtain information about a few of the exponentially many orthogonal elements in the function class. More formally, we lower bound the SQ dimension of the function classes considered in Table 3.1 to show our query lower bounds.

Our hardness results hold for any target observable M, as long as the learning setting is one we consider in Table 3.1. Furthermore, they hold for any variational ansatz—not just on average—provided it is in one of the settings of Table 3.1. Finally, our results hold for any constant error τ in the statistical queries; indeed, the majority of our results hold even if this noise were only exponentially small in the problem size. For instance, training via gradient descent where the gradient is estimated using polynomially many samples fits into this framework immediately just from the induced shot noise. From the third row of Table 3.1, then, we find that certain barren plateau untrainability results—those which suggest that gradient-based methods for optimizing typical d-dimensional variational quantum ansatzes of depth $L \ll n$ take $2^{\Omega\left(L^d\right)}$ problem queries [65]—can be generalized to *all* training algorithms that fall under the statistical query framework.

In a more positive light, learning local Hamiltonians generated by shallow depth circuits can potentially be efficiently performed as the complexity grows exponentially only with locality or depth in this setting. In fact, prior results have provably shown that certain classes of Hamiltonians are efficiently learnable using properly chosen algorithms [114, 115]. Nevertheless, this does not correspond to efficient learnability of the ground state of a given Hamiltonian, as learnability of the properties of a Hamiltonian is not the same as the learnability of its ground state. Indeed, we will

see in Section 3.3 that typically, even in this setting, learning the ground state of such a local Hamiltonian is difficult.

Though noise during optimization may appear unnatural in classical settings, noise in quantum settings is rather endemic and the SQ model allows one to rigorously analyze the complexity of learning in the presence of noise. One important caveat of these results, however, is that in the SQ setting learning must succeed for *all* values of the query within the given tolerance τ. Noise in quantum settings, which can arise from sampling a finite data set, gate error, state preparation error, measurement sampling noise, or other means does not exactly coincide with the assumed tolerance of an SQ model as the SQ model assumes adversarial noise. We circumvent these strong assumptions on the noise in Section 3.3 by instead considering untrainability in terms of loss landscapes, though this is at the expense of the very strong no-go results we prove here.

Our hardness results do not indicate that simple classes of functions like those generated by single qubit rotations are hard to learn for all algorithms, but only those whose steps reduce to statistical queries. For example, the class of Pauli channels is not learnable in the SQ setting, but there exist simple, carefully constructed, algorithms which can learn Pauli channels [116–118]. This is analogous to the classical setting where parity functions are hard to learn in the noisy SQ setting, but efficient to learn using simple linear regression [111]. Similarly, the related work of Reference [119] showed that output distributions of Clifford circuits can be hard to learn using statistical queries, but efficient using a technique that resorts to linear regression on a matrix formed from samples of the overall distribution. More loosely, our results provide support to the basic maxim that algorithms which apply too broadly will work very rarely [120]; more careful construction of learning algorithms tailored to the problem at hand is generally necessary. One straightforward way to avoid the hardness of the SQ setting is to construct algorithms whose basic steps do not reduce to statistical queries, e.g. via the construction of non-global metrics [121–123]. However, such a fix is by no means guaranteed to avoid the more general issues of poor landscapes and noise that also make learning in the SQ setting so difficult, as we now

examine.

3.3 Loss Landscapes of Local Variational Quantum Algorithms

We now consider the trainability of VQAs in the noise-free regime, beyond optimization algorithms that reduce to statistical queries. Though we are unable to prove the very strong no-go results proved in the SQ framework, we are able to show that the loss landscapes of typical local variational algorithms with Hamiltonian agnostic ansatzes are unamenable to optimization. We achieve this by showing that typically, the loss landscapes of shallow, local VQAs are swamped with poor local minima.

As discussed in Table 1.1, it is already known that deep Hamiltonian agnostic ansatzes are typically untrainable due to the presence of barren plateaus [63–65]; hence, here we focus on shallow ansatzes. In Chapter 2, we have also shown that shallow, nonlocal models are untrainable, by showing that the scrambling of variational ansatzes over the entire system in these instances induce poor local minima. These techniques are not extendable to shallow, local ansatzes, however, which do not scramble globally.

Instead, here, we show that ansatzes that approximately scramble locally are difficult to train. As we will later show, this includes common classes of variational ansatzes, such as Hamiltonian agnostic checkerboard ansatzes on a d-dimensional lattice. We show that this approximate, local scrambling suffices to imply that the loss landscapes of these VQAs are close to those of Wishart hypertoroidal random fields (WHRFs), introduced in Chapter 2. To review, these are random fields parameterized by l, m of the form:

$$F_{\mathrm{WHRF}}(\boldsymbol{w}) = m^{-1} \sum_{i,j=1}^{2^l} w_i J_{i,j} w_j, \tag{3.4}$$

where $\boldsymbol{J}$ is drawn from a Wishart distribution with m degrees of freedom, and $\boldsymbol{w}$ are points on a certain embedding of the hypertorus $(S^1)^{\times l}$ in $\mathbb{R}^{2^l}$. We demonstrate this convergence via new techniques, directly bounding the error in the joint characteristic

function of the function value, gradient, and Hessian components of the variational loss from those of WHRFs. As the typical loss landscapes of WHRFs are known given these random variables, by demonstrating sufficient convergence of these random variables to those of WHRFs, we will be able to infer the distribution of critical points for local VQAs.

To begin, we take our (assumed traceless) problem Hamiltonian to have Pauli decomposition:

$$H = \sum_{i=1}^{A} \alpha_i P_i, \tag{3.5}$$

and for simplicity scale and shift the loss landscape of Equation (1.9) to be of the form:

$$\hat{\mathcal{R}}_{\mathrm{VQE}}(\boldsymbol{\theta}) = 1 + \|\boldsymbol{\alpha}\|_1^{-1} \sum_{i=1}^{A} \alpha_i \langle \boldsymbol{\theta}| P_i |\boldsymbol{\theta}\rangle , \tag{3.6}$$

where $\boldsymbol{\alpha}$ is the vector of all α_i and the ansatz $|\boldsymbol{\theta}\rangle$ is as given in Equation (1.5). As this ansatz is assumed to be shallow and local, we assume that the reverse light cone of each P_i under the ansatz is of size $l \ll n$.

As in most analytic treatments of Hamiltonian agnostic VQAs, we consider certain randomized classes of ansatzes [49, 63–65]. Roughly, we assume that in a local region around each measured Pauli observable P_i, the ansatz is an ϵ-approximate t-design; that is, its first t moments are ϵ-close to those of the Haar distribution. This is a much weaker assumption than global scrambling of the ansatz. For instance, for P_i of constant weight, such approximately locally scrambling circuits include constant depth local circuits with random local gates [124]. We discuss in more detail when this assumption holds in practice when specializing to common variational quantum learning scenarios, and defer technical details to Appendix B.4.

Our main result, informally, is that the random field given by Equation (3.6) under this approximate, local scrambling assumption converges in distribution to that of a WHRF. The formal statement and derivation of this result are given in Appendix B.4, where we also lay out our assumptions more explicitly. Informally, the result follows by deriving a bound on the error in the joint characteristic function of the loss function and its first two derivatives from that of a WHRF. We then use this

to bound the error in distribution that is incurred by the induced scrambling being only approximate. Finally, we show using properties of local Haar random gates and the locality of the problem Hamiltonian that this suffices to prove convergence of these random variables to those of a WHRF.

Theorem 3.3 (Approximately locally scrambled variational loss functions converge to WHRFs, informal statement of Theorem B.18)**.** *Let*

$$m \equiv \frac{\|\boldsymbol{\alpha}\|_1^2}{\|\boldsymbol{\alpha}\|_2^2} 2^{l-1} \tag{3.7}$$

be the degrees of freedom parameter. Assume $q \log(q) = \mathrm{o}(m)$*, where* q *is the number of ansatz parameters in the reverse light cone of each* P_i*. Then, the distribution of Equation* (3.6) *and its first two derivatives are equal to those of a WHRF*

$$F_{WHRF}(\boldsymbol{\theta}) = m^{-1} \sum_{i,j=1}^{2^l} w_i J_{i,j} w_j \tag{3.8}$$

with m *degrees of freedom, up to an error in distribution on the order of* $\tilde{\mathrm{O}}\left(\mathrm{poly}\left(\frac{1}{t} + \epsilon + \exp(-l)\right)\right)$*. Here,* $\boldsymbol{w}$ *are points on the hypertorus* $(S^1)^{\times l}$ *parameterized by* $\tilde{\boldsymbol{\theta}}$*, where* $\tilde{\theta}_i$ *is the sum of all* θ_j *on qubit* i*.*

We interpret this result as the degrees of freedom m of the model being given by roughly the sum of the local Hilbert space dimensions of the reverse light cones of terms in the Pauli decomposition of H. We interpret this as the local underparameterization of the model, to be contrasted with the global underparameterization interpretation when m is exponentially large in n. Using known properties of the loss landscapes of WHRFs (see Chapter 2), we are then able to prove the following result on the loss landscapes of local VQAs:

Corollary 3.4 (Shallow, local VQAs have poor loss landscapes, informal statement of Corollary B.21)**.** *Let* $\hat{\mathcal{R}}_{VQE}$ *be a local VQA loss function of the form of Equation* (3.6)*.*

Assume all coefficients α_i of the Pauli decomposition of H are $\Theta(1)$, and

$$l \log(n) + q \log(q) = \mathrm{o}\left(2^l A\right). \tag{3.9}$$

Then $\hat{\mathcal{R}}_{VQE}$ has a fraction superpolynomially small in n of local minima within any constant additive error of the ground state energy.

Optimizing loss landscapes where only a superpolynomially small (in n) fraction of the local minima are near the global minimum in energy is expected to be difficult. Indeed, algorithms such as gradient descent would then expect to have to be restarted a superpolynomial number of times before finding a good approximation to the global minimum; we also give heuristic reasons why this should continue to be true for other local optimizers in Appendix B.7. Our results stand in stark contrast with the loss landscapes typically found in classical machine learning, where almost all local minima closely approximate the global minimum in function value [44, 45].

In the shallow ansatz regime—where $q, l = \mathrm{O}(\mathrm{polylog}(n))$—and assuming an extensive problem Hamiltonian such that $A = \Omega(n)$, the condition given by Equation (3.9) is always satisfied. Interestingly, this is a regime where barren plateaus are known not to occur [65], demonstrating that poor local minima can give rise to poor optimization performance even when the loss function features large gradients. We now specialize to common variational quantum learning scenarios, and consider the implications of Corollary 3.4.

First, let us consider d-dimensional checkerboard ansatzes of constant depth. Fix p, t to be sufficiently large constants. We assume that the initial state forms an $\mathrm{O}\left(\frac{1}{\mathrm{poly}(t)}\right)$-approximate t design on l qubits around each Pauli observable of weight k; this can be done via a depth p, d-dimensional circuit of 2-local Haar random unitaries when $l = \mathrm{O}\left(\frac{(p+k)^d}{\mathrm{poly}(t)}\right) \geq k$ for some fixed polynomial in t [124, 125]. After this state preparation circuit, a traditional depth $\Theta\left(l^{\frac{1}{d}}\right)$ (i.e. independent of n), d-dimensional, n qubit checkerboard circuit is applied, with observable reverse light cones of size at greatest l. By Corollary 3.4, these variational ansatzes are untrainable due to poor local minima, yet by the results of Reference [65] do not suffer from barren

plateaus.

One interesting consideration is extending this result to traditional checkerboard ansatzes, without the special state preparation procedure we have considered. There, the $l = \mathrm{O}\left(\frac{(p+k)^d}{\mathrm{poly}(t)}\right)$ qubit local state is mixed, and our results therefore do not directly apply. However, we expect no reason for the mixedness of the initial state to improve training performance in any way. We validate this intuition numerically in Section 3.4.

We also consider a class of models similar to quantum convolutional neural networks (QCNNs) [108] previously shown not to suffer from barren plateaus [126]. Though these models are in full generality trained on arbitrary loss functions, for learning various physical models the loss may take the form of Equation (1.9). QCNNs are defined by their measurement of a subset of qubits at periodic intervals, via so-called pooling layers; for sufficiently deep (i.e., large constant depth) convolutional layers, then, at some point in the model, the number of remaining qubits will be sufficiently small such that the remaining convolutional layers are approximately scrambling. If one then assumes that the initial states are adversarially chosen such that they remain pure by this layer, this scenario reduces to the shallow checkerboard ansatz scenario, and once again we expect poor local minima by Corollary 3.4. Even if the initial states are not adversarially chosen and the input to the scrambling convolutional layers is mixed, we expect by similar intuition the model to remain untrainable; we will see this numerically, where we also observe that this poor training occurs when training on loss functions beyond Equation (1.9).

3.4 Numerical Results

To numerically validate our theoretical findings, we perform numerical simulations showing that learning in various settings cannot be guaranteed unless exponentially many parameters are included in an ansatz. We only consider problems and ansatzes where the existence of a zero loss global minima is guaranteed to study whether or not optimizers can actually find the global minimum or a similarly good critical point. We parameterize all trainable 2-qubit gates in the Lie algebra of the 4-dimensional unitary

group, and implement the resulting unitary matrix via the exponential map which is surjective and capable of expressing any local 4×4 unitary gate. In all cases, we perform simulations using calculations with computer precision and analytic forms of the gradient (see Appendix B.6 for more details). In practice, actual quantum implementations will be hampered by various sources of inefficiency such as the lack of an analogous method of backpropogation for calculating gradients, sampling noise, or even gate errors. Thus, our numerical analysis can be interpreted as a "best case" setting for quantum computation where we disregard such inefficiencies and focus solely on learnability. In Appendix B.5, we further study variations of the teacher-student learning and random variational quantum eigensolver (VQE) [19] settings discussed here. We also consider the training performance of VQE in finding the ground state of a Heisenberg XYZ Hamiltonian [127]. Our supplemental results reinforce our findings here.

One may conjecture that it is plausible to learn the class of functions generated by relatively shallow depth variational teacher circuits by parameterizing a shallow-depth student circuit of the same form and training its parameters. In this so-called teacher-student setup, we are guaranteed the existence of a perfect global minimum since recovering the parameters of the teacher circuit achieves zero loss. In other words, the global minimum is guaranteed to be achievable in the setting we consider here. Still, we showed earlier that such circuits are typically have many poor local minima, and are always hard to learn in the statistical query setting. Here, we provide numerical evidence of these findings for the QCNN ansatz. Additional confirmation of these findings with a checkerboard ansatz is included in Appendix B.5.

The quantum convolutional neural network (QCNN) presents an interesting test bed for our analysis since it has been shown in prior work to avoid barren plateaus [126]. Nevertheless, the QCNN, like other models, is riddled with poor local minima in generic learning tasks. For our analysis, we attempt to learn randomly generated quantum convolutional neural networks (QCNNs) with a parameterized QCNN of the same form. In the QCNN, both student and teacher circuits have parameterized 2-qubit gates at each layer followed by 2-qubit pooling layers (see Appendix B.6 for

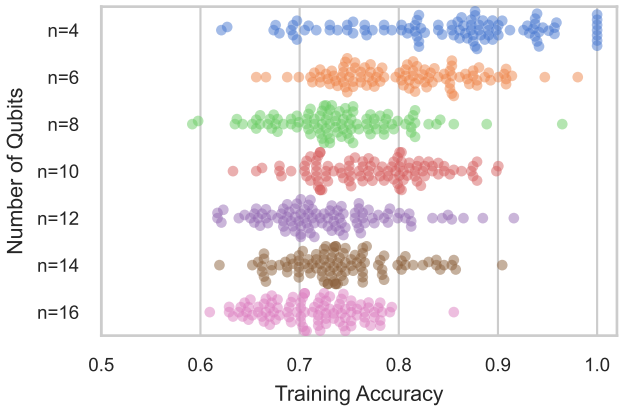


Figure 3-2: The student circuit is unable to learn the teacher circuit as the number of qubits grows, converging to a local minimum of the loss landscape. The existence of a global optimum is guaranteed as the teacher circuit is drawn from a random initialization of the same QCNN structure of the student circuit. Here, for a ranging number of qubits, 100 student circuits are trained to learn randomized teacher circuits of the same form and the resulting swarm plots of the final training accuracy are shown.

more details). Each 2-qubit gate is fully parameterized in the Lie algebra of the unitary group. Networks are trained to predict the probability of the measurement of the last qubit in the teacher circuit. In other words, the student network is trained on a classification problem defined by teacher network where, by construction, perfect classification accuracy is known to be achievable. We benchmark performance with the classification accuracy, where a prediction is considered correct when it predicts the most likely measurement of the last qubit correctly. Networks are trained via the Adam optimizer [128] to learn outputs of 512 randomly chosen computational basis states. QCNNs with 4, 8, 12, and 16 qubits have 32, 48, 64, and 64 trainable parameters, respectively.

Figure 3-2 plots the final training accuracy achieved over 100 random simulations for varying ranges of circuit sizes. For circuits with 4 qubits, the training is sometimes successful, often achieving an accuracy above 85 percent on the training dataset. However, as the number of qubits grows, even past 8 qubits, the optimizer is unable to recover parameters which match the outputs of the teacher circuit. The results here show that the QCNN circuit—which has O $(\log(n))$ depth—still locally scrambles

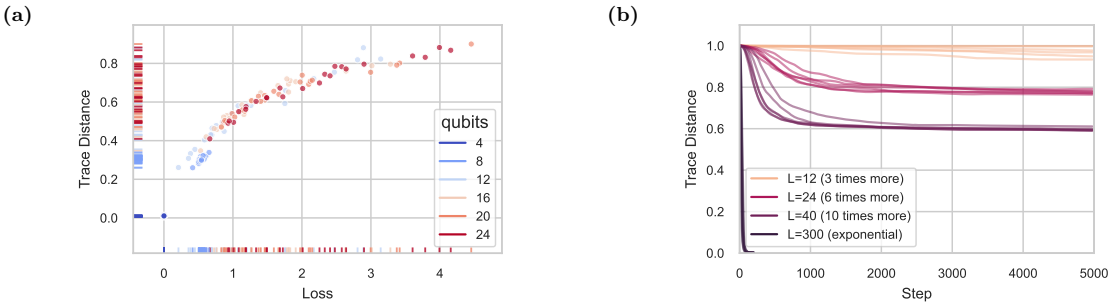


Figure 3-3: (a) Scatter plot of the final loss and trace distance of the VQE state after 30000 steps of gradient descent optimization shows that the algorithm converges to poorer local minima as the number of qubits grows. 24 simulations are performed for each value of n. The algorithm always succeeds at obtaining the ground state with 4 qubits, but progressively struggles more with added qubits. (b) The number of layers needed to guarantee convergence to the ground state empirically grows exponentially with the number of qubits. Here, we consider 4-layer Hamiltonians of the form of Equation (3.10) on 14 qubits where the number of layers L in the ansatz is varied. When the ansatz has 300 layers—enough that the number of ansatz parameters is larger than the explored Hilbert space dimension—the model successfully converges to the ground state, rather than remaining stuck in a poor local minimum.

outputs to hinder learnability.

We now consider VQE. To analyze the performance of variational optimizers, we consider problems and ansatzes which are capable of recovering the global minimum. We aim to find the ground states of local Hamiltonians H_t over n qubits that take the form of single qubit Pauli Z Hamiltonians conjugated by L^* layers of two alternating unitary operators U_1 and U_2 which are product unitaries on neighboring 2-local qubits:

$$H_t = \left(U_2^\dagger U_1^\dagger\right)^{L^*} \left[\sum_{i=1}^{n} Z_i\right] (U_1 U_2)^{L^*} + nI. \tag{3.10}$$

The added identity matrix normalizes the Hamiltonian to have ground state with energy 0. Since the ground state of $\sum_{i=1}^{n} Z_i$ is the state $|1\rangle^{\otimes n}$, we are guaranteed the existence of a global minima when using a checkerboard ansatz of at least depth L^*, since this ansatz can "undo" the conjugation by unitary operators. In the remainder of this Section, we consider Equation (3.10) with $L^* = 4$.

We measure the performance of optimization with two metrics. The first is the

loss function itself, which is the average energy $\langle\psi| H_t |\psi\rangle$ of the VQE ansatz state $|\psi\rangle$ for the given Hamiltonian H_t. The second is the trace distance to the ground state $|\phi_g\rangle$ of H_t, equal to $\| |\phi_g\rangle\langle\phi_g| - |\psi\rangle\langle\psi| \|_* /2$ (where $\|\cdot\|_*$ is the nuclear norm). Both of these metrics converge to zero at the global minimum.

We first aim to learn the ground state using a checkerboard ansatz by performing vanilla gradient descent on $L = L^* = 4$ parameterized layers, equal in depth to the Hamiltonian conjugation circuit and thus capable of recovering the ground state. In Figure 3-3(a), we plot the final values of the loss and trace distance for 24 randomly initialized VQE problems for a number of qubits ranging from 4 to 24. Similar results are observed when using more advanced optimizers such as Adam (see Appendix B.6) [128]. Consistent with our theoretical findings, convergence clusters around local minima far from the ground state, particularly as the number of qubits grows.

Our theoretical results also imply the difficulty of training beyond a finite fraction of the ground state energy in a VQE setting. Figure 3-3(b) illustrates this phenomenon when performing optimization on a 14 qubit ansatz. As more parameters are added to the ansatz via increasing its depth L, the VQE algorithm performs better, but it is not until the number of parameters is exponential in the problem size that convergence to a global minimum (or even within a small additive error of the global minimum) is guaranteed. This is true even though the ansatz is capable of expressing the ground state at $L = 4$. Simulations here are performed as before on random $L^* = 4$ Hamiltonians of the form of Equation (3.10).

3.5 Conclusion

Though variational quantum algorithms—and quantum machine learning models in general—have been cited as perhaps the most promising use case for quantum devices in the near future [14], theoretical guarantees of their training performance have been sparse. Here, we have excluded a wide class of variational algorithms by showing that in many settings, they are in fact not trainable. We showed this in two different

frameworks: first, in Section 3.2.2, we studied various classes of quantum models in the statistical query framework. We showed that in the presence of noise, exponentially many queries in the problem size are needed for these models to learn. As a complementary approach, we also examined the typical loss landscapes of variational quantum algorithms in the noiseless setting in Section 3.3, and showed that even at constant depth these models can have a number of poor local minima superpolynomially large in the problem size. We also numerically confirmed these results for a variety of problems in Section 3.4. These results go beyond the typical studies on the presence of barren plateaus, as many of the models we study here have gradients vanishing only polynomially quickly in the problem size. Our work demonstrates that showing that barren plateaus are not present in a model does not necessarily vindicate it as trainable.

These results, though they exclude a wide variety of variational quantum algorithms, still leave room for hope in the usefulness of these algorithms. Particularly, our analysis in the noiseless setting of landscapes of variational quantum algorithms focuses on very general, Hamiltonian agnostic ansatzes; in various instances, more focused ansatzes may be trainable. For instance, as previously shown in Reference [107], for certain classes of problems the quantum approximate optimization algorithm (QAOA) [99] is provably able to outperform the best unconditionally proven classical algorithms, even when taking into account the training of the model. This is due to parameter concentration, where the global optimum for small problem instances is close to the global optimum for large problem instances [129]. These results demonstrate the power of good model initialization in variational quantum algorithms: even if the total variational landscape is swamped with poor local minima, good initialization may ensure that the optimizer begins in the region of attraction of the global minimum. Though this is perhaps most relevant for the variational quantum eigensolver (VQE) [19] and QAOA [99], where there exists physical intuition for potentially performant parameter initializations, in more traditional machine learning settings this may manifest as good performance on certain inputs to the model.

Variationally studying models with many symmetries may also avoid our poor

performance guarantees. Intuitively, our results here are the consequence of underparameterization. Namely, unless the ansatz is parameterized such that the number of parameters grows with the (local) Hilbert space dimension, the model is not trainable. Typically, this Hilbert space dimension is exponentially larger than the number of parameters the ansatz uses to explore it. However, if the model is heavily constrained by symmetries, this dimension might be much smaller. Such models were studied numerically in References [70, 130], where it was shown that certain variational quantum algorithms optimize efficiently. Though often these models can be solved classically when the symmetries are known (see in particular Chapter 4), these symmetries may not be known *a priori*. Indeed, one may be able to test for the presence of symmetries in a given model by studying whether associated variational quantum algorithms are trainable. Similar to these general symmetry considerations, known structure in the problem may also allow one to build up hierarchical ansatzes that are able to be trained sequentially. We leave further investigation in these directions to future work.

Finally, though many variational models fit the framework of Equation (1.9), there exist other settings of variational quantum algorithms. One class of such models includes quantum Boltzmann machines, which attempt to model given quantum states via the training of quantum Gibbs states [16]. When the full quantum Gibbs state is observed, it is known that these models are efficiently trainable [114], and numerically it is known that these models are trainable even when the full state is not observed [16, 131]. Furthermore, though in full generality preparing quantum Gibbs states is difficult, state preparation has been shown to be efficient in certain regimes relevant to machine learning [131–133], potentially giving an end-to-end trainable quantum machine learning model. We leave further analytical investigation on the training landscapes of quantum Boltzmann machines to future work.

Our results contribute to the already vast library of literature on the trainability of variational quantum models in further culling the landscape of potentially trainable quantum models. We hope these results have the effect of focusing research efforts toward classes of models that have the potential for trainability, and whittle down the search for practical use cases of variational quantum algorithms.

Chapter 4

Efficient Classical Algorithms for Simulating Symmetric Quantum Systems

The results of this chapter were featured in Reference [51], work done in collaboration with Andreas Bauer, Bobak T. Kiani, and Seth Lloyd.

4.1 Introduction

In the physical sciences, symmetries are useful for simplifying difficult computational tasks by reducing the effective degrees of freedom of the problem. This general principle has been used to find exact solutions to many problems, such as integrable systems [134], topological fixed-point models [135], or conformal field theories [136]. There has been a hope that similar symmetries may enable the efficiency of quantum algorithms for simulating or finding the ground state of a symmetric Hamiltonian. Indeed, it is known that there exist theoretical guarantees for quantum algorithms for finding the ground state [47] and fast-forwarding quantum dynamics [83] of Hamiltonians which commute under the action of the symmetric group S_n on qubits. It has also numerically been shown that quantum algorithms are capable of finding the ground state of certain integrable systems [49, 70] even when the symmetry is not

explicitly given to the quantum algorithm *a priori*. Furthermore, prior work used Lie algebraic methods to efficiently classically simulate operators restricted to a Lie algebra whose dimension is polynomially large (independent of the potentially exponentially large Hilbert space dimension) [137, 138]. Quantum machine learning models that are symmetry equivariant are also believed to be more efficiently trainable than their general counterparts [46, 50, 79, 139–141]. These quantum models are partly inspired by classical neural network models that have enjoyed much recent success [142–144]. However, restricting quantum algorithms to problems obeying many symmetries potentially allows for efficient classical algorithms which also take advantage of these same symmetries. This raises the natural question: are there efficient classical algorithms capable of performing these tasks?

This is what we investigate here. Intuitively, we show that problems constrained by large symmetry groups yield efficient classical algorithms for computing many properties of interest, as illustrated in Figure 4-1(a). We first give a very general classical algorithm for finding the ground state and energy of Hamiltonians constrained by many symmetries. We also consider the problem of simulating dynamics under symmetric Hamiltonians. We then specialize to the case of systems invariant under permutations of its qubits. Finally, we dequantize an algorithm for performing binary classification problems using permutation-invariant systems on qubits.

4.2 Motivation and Setting

Our algorithms are motivated by the fact that symmetries significantly reduce the number of degrees of freedom for a given problem. For example, consider the classical setting of Boolean functions which are invariant under arbitrary permutations of the bits. Such functions are defined up to the orbits of the Boolean cube with respect to permutations of the bits. For a Boolean function on n bits, there are $n + 1$ orbits indexed by the Hamming weight of the bitstrings. Therefore, any problem over symmetric Boolean functions need only consider a given element of each of the $n + 1$ orbits to cover all possible degrees of freedom. As we will later show, the symmetric

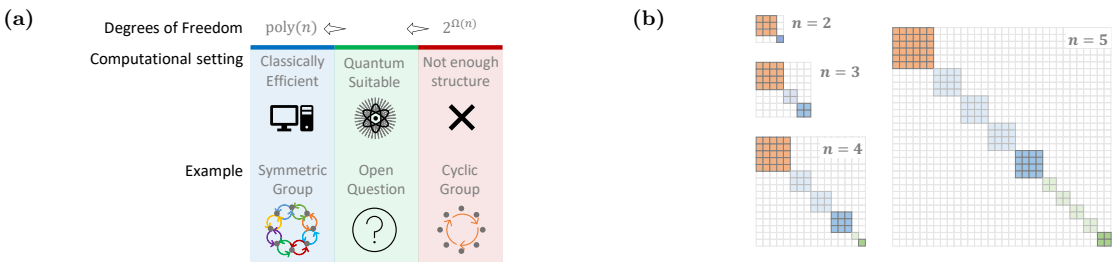


Figure 4-1: (a) Small groups of symmetry leave too large of an effective dimension for the problem to be tractable via quantum computation. On the contrary, very restrictive symmetries render a problem classically tractable. Between these two regions lies an area of promise where quantum computers may offer an advantage. (b) The Schur–Weyl decomposition shows that only a smaller representative subspace (indicated by darker colors) of the larger Hilbert space needs to be considered for permutation invariant operations. The size of this subspace grows as $\mathrm{O}(n^3)$ for n qubits.

group acting over n qubits similarly reduces systems to $\mathrm{O}(n^3)$ degrees of freedom. By considering the algebra of the symmetric group on the symmetric subspace of linear operators, we will show that all these degrees of freedom can be manipulated solely through classical computation.

Before proceeding, we need to introduce important functions and definitions that will be used in this setting. We first formalize the notion of symmetry by speaking of *invariant operators*, defined in the following way:

Definition 4.1 (Invariant operator)**.** Given a compact group G with unitary representation $R : G \to \mathrm{U}(N)$, a linear operator $H : \mathbb{C}^N \to \mathbb{C}^N$ is invariant under $R(G)$ if $\forall g \in G$:

$$R(g)\, H R(g)^\dagger = H. \tag{4.1}$$

Note that any invariant operator is also an *equivariant operator* [46] in the sense that it commutes with the representation of the group.

Any operator can be projected onto the symmetric subspace induced by $R(G)$ using the twirling superoperator Re_R (more commonly known as the Reynold's operator in invariant theory) [145, 146], which maps any operator onto the set of equivariant

operators:

$$\mathrm{Re}_R(M) = \frac{1}{|G|} \sum_{g \in G} R(g) M R(g)^\dagger . \tag{4.2}$$

Invariant subspaces of a larger Hilbert space can be identified by performing an isotypic decomposition of the representation of a group. As an example, in the case of systems invariant under permutations of the qudits, the Schur decomposition maps the computational basis into blocks of invariant subspaces. We graphically visualize this phenomenon in Figure 4-1(b) and provide further details in Appendix C.1.

Throughout this study, runtime complexities are denoted as a function of the matrix multiplication exponent ω. The best known upper bound is currently $\omega = 2.37188$ [147], which implies asymptotic runtimes of $\mathrm{O}(n^{\omega+\alpha})$ for any $\alpha > 0$ for stably performing common linear algebraic routines including eigendecomposition, singular value decomposition, and matrix inversion [148].

4.3 Algorithms for General Symmetry Groups

In this Section, we discuss the general problem of finding the ground state energy, ground state, and performing time evolution under a Hamiltonian H on a finite-dimensional Hilbert space which is invariant under some representation R of a symmetry group G. Consider the *-subalgebra of operators invariant under R to which H belongs. We think of this subalgebra as a standalone *-algebra X, such that the embedding of X into the full operator algebra defines a representation A of X.

The practical relevance of these considerations is when the size of the total Hilbert space grows exponentially with some scaling parameter n. The paradigmatic example is the Hilbert space of n qubits. If there are enough symmetries, it can happen that the dimension $N(n)$ of X only grows polynomially with n, in which case many properties can be calculated efficiently [137]. This restriction of X to a lower-dimensional subspace may more generally happen beyond systems symmetric in the sense of Definition 4.1. Due to this, for now we focus explicitly on X and A, rather than on G and its representation R; we will discuss the connection of our results to G and R

more specifically at the end of this Section.

For the various algorithms we now consider, we will assume that different properties of X and A are known. For the algorithm for finding the ground state energy of H given in Theorem 4.2, we will assume that the structure constants of X in some preferred basis are known. In a slight abuse of notation, we will refer to those structure constants as $X_k^{i,j}$, where i, j, and k label basis elements. We note that the structure constants can frequently be efficiently obtained from the generators of an algebra, for example in the case of Lie subalgebras [137, 149].

In the course of proving Theorem 4.3, we give an algorithm for finding the ground state of H. Every finite-dimensional *-algebra is isomorphic to a direct sum of irreducible blocks, and every representation is isomorphic to a direct sum of irreducible representations. That is, there is a block-diagonal orthonormal basis $|\lambda, q_\lambda, p_\lambda\rangle$ of the vector space acted upon by A, where λ labels an irrep of X, q_λ labels a basis vector internal to λ, and p_λ labels a basis vector in the multiplicity vector space of λ; this is the basis in which we compute the ground state of H (for some arbitrary and fixed dimension label $p_{\lambda 0}$). To prove our theorem, we assume knowledge of the matrix elements:

$$F^{i,\lambda}_{q_\lambda, q'_\lambda} \equiv \langle \lambda, q_\lambda, p_{\lambda 0}| \, A_i \, |\lambda, q'_\lambda, p_{\lambda 0}\rangle \, . \tag{4.3}$$

Finally, for Theorem 4.4, we assume the knowledge of a *symmetric transform operator* implementable on a quantum computer, i.e. an isometry V_{STO} such that:

$$V_{\mathrm{STO}} \, |\lambda, q_\lambda, p_\lambda\rangle = |\boldsymbol{\lambda}\rangle \, |\boldsymbol{q_\lambda}\rangle \, |\boldsymbol{p_\lambda}\rangle \, . \tag{4.4}$$

Here, $|\boldsymbol{\lambda}\rangle \, , |\boldsymbol{q_\lambda}\rangle \, , |\boldsymbol{p_\lambda}\rangle$ are bitstring encodings of $\lambda, q_\lambda, p_\lambda$, respectively, in the computational basis, labeling the $\boldsymbol{\lambda}$ register, $\boldsymbol{q}$ register, and $\boldsymbol{p}$ register, respectively. An example of such an operator is the *Schur transform* [150, 151], described in more detail in the Supplementary Information.

In all three theorems, we assume we are given the Hamiltonian $H \in A\,(X)$ as

$h \in X$ expressed in the preferred basis, such that

$$H = \sum_i h_i A_i. \tag{4.5}$$

We now state our main results. First, we give a simple construction of a classical algorithm for finding the ground state energy of some representation of a Hamiltonian obeying the given symmetries.

Theorem 4.2 (Finding the ground state energy of symmetric Hamiltonians)**.** *Consider a subalgebra X of dimension N, and assume that the structure constants of X in some preferred basis are known as discussed above. Let $H \in A(X)$ be a Hamiltonian given in the preferred basis as in Equation* (4.5)*. Then the ground state energy of H can be found in time $\mathrm{O}\left(N^{\omega}\right)$.*

Proof. Consider the operator with indices:

$$\hat{h}_k^j \equiv \sum_i h_i X_k^{i,j}, \tag{4.6}$$

which is nothing but the regular representation of h for the algebra X. Then we have that their ground state energies are equal:

$$\mathrm{GSE}\left(H\right) = \mathrm{GSE}\left(\hat{h}\right). \tag{4.7}$$

This is because the regular representation is faithful, and the ground state energy of an operator is the same in any faithful representation. Since X has dimension N, the ground state energy of $\hat{h}$ can be found in time $\mathrm{O}\left(N^{\omega}\right)$. □

An advantage of this algorithm is that the only necessary information are the structure constants of X; no knowledge of the irreps of X is needed. However, due to this we have poor scaling with the number of irreps n_λ, as the direct sum structure of X is not necessarily known. Another disadvantage of this approach is that it only gives the ground state energy, rather than the ground state itself (in a representation that is not the regular representation).

We now focus on the case when we are interested in finding the ground state of some representation of such a Hamiltonian, in a basis where the action of the representation is known.

Theorem 4.3 (Finding the ground state of symmetric Hamiltonians)**.** *Consider a subalgebra $A(X)$, and assume that the matrix elements $F^{i,\lambda}_{q_\lambda,q'_\lambda}$ are known as discussed above. Then the ground state energy and ground state of H in the $|\lambda, q_\lambda, p_{\lambda 0}\rangle$ basis can be found in time $\mathrm{O}\left(n_\lambda n_q^\omega\right)$, where n_λ are the number of irreps of X and n_q the maximum irrep dimension.*

Proof. For each λ, consider the operator with indices:

$$\hat{h}^\lambda_{q_\lambda,q'_\lambda} \equiv \sum_i h_i F^{i,\lambda}_{q_\lambda,q'_\lambda}. \tag{4.8}$$

Note that in the $|\lambda, q_\lambda, p_\lambda\rangle$ basis, H has a block diagonal form. Furthermore, as p_λ labels isomorphic copies of irreps, we can find the ground state by fixing $p_{\lambda 0}$ WLOG. Namely, the ground state energy is given by:

$$\mathrm{GSE}\left(H\right) = \min_\lambda \mathrm{GSE}\left(\hat{h}^\lambda\right), \tag{4.9}$$

where:

$$\mathrm{GSE}\left(\hat{h}^\lambda\right) \equiv \min_{|\psi\rangle} \langle\psi| \hat{h}^\lambda |\psi\rangle. \tag{4.10}$$

Furthermore, let

$$\lambda_{\min} \equiv \mathrm{argmin}_\lambda \, \mathrm{GSE}\left(\hat{h}^\lambda\right) \tag{4.11}$$

and

$$|\psi^*\rangle \equiv \mathrm{argmin}_{|\psi\rangle} \langle\psi| \hat{h}^{\lambda_{\min}} |\psi\rangle. \tag{4.12}$$

Then, for any p,

$$|\lambda_{\min}, \psi^*, p\rangle \tag{4.13}$$

is a ground state in the $|\lambda, q_\lambda, p_\lambda\rangle$ basis. The dimension of $\hat{h}^\lambda$ is $\dim_X(\lambda) \times \dim_X(\lambda)$, and thus calculating $|\psi^*\rangle$ will take time $\mathrm{O}\left(\dim_X(\lambda)^\omega\right)$. In total, finding the ground

state of H in the $|\lambda, q_\lambda, p_{\lambda 0}\rangle$ basis takes time $\mathrm{O}\left(n_\lambda \dim_X(\lambda)^\omega\right) = \mathrm{O}\left(n_\lambda n_q^\omega\right)$. □

We now show that the dynamics of an initial state under equivariant unitaries can be classically simulated even if $\rho \neq A(X)$. The given procedure is fully classical if the initial state is given as a classical shadows description of the state; if the input is given as a quantum state, we show that performing classical shadow measurements is efficient and then reduces the algorithm to the purely classical setting. This generalizes a similar approach taken in Reference [152] in the case of particle number symmetry.

Theorem 4.4 (Simulating equivariant dynamics)**.** *Let*

$$O = \sum_i o_i A_i \tag{4.14}$$

be a projective measurement and

$$U = \sum_i u_i A_i \tag{4.15}$$

a unitary operator. Assume the matrix elements $F^{i,\lambda}_{q_\lambda, q'_\lambda}$ as described previously are known. Assume also the existence of a symmetry transform operator V_{STO} with depth v. Then,

$$\ell(\rho) = \mathrm{tr}\left(OU\rho U^\dagger\right) \tag{4.16}$$

can be estimated to additive error ϵ with probability $1-\delta$ via $\tilde{\mathrm{O}}\left(|O|_\infty^2 \epsilon^{-2} n_\lambda^2 n_q^2 \log\left(\delta^{-1}\right)\right)$ calls to a quantum computer each of depth $v+1$, up to an additional time $\mathrm{O}\left(n_\lambda n_q^\omega\right)$ in classical processing. Here, n_λ are the number of irreps of X and n_q the maximum irrep dimension.

Proof. Let $\tilde{O}_{\boldsymbol{p_0}}, \tilde{U}_{\boldsymbol{p_0}}$ be projections of $V_{\mathrm{STO}} O V_{\mathrm{STO}}^\dagger, V_{\mathrm{STO}} U V_{\mathrm{STO}}^\dagger$ onto some particular $\boldsymbol{p}$. Classically, we can calculate:

$$M_{\boldsymbol{p_0}} = \tilde{U}_{\boldsymbol{p_0}}^\dagger \tilde{O}_{\boldsymbol{p_0}} \tilde{U}_{\boldsymbol{p_0}} \tag{4.17}$$

in time $\mathrm{O}\left(n_\lambda n_q^\omega\right)$, as it is given by the matrix multiplication of n_λ blocks each of size at most $n_q \times n_q$. $\tilde{\mathrm{O}}\left(|O|_\infty^2 \epsilon^{-2} n_\lambda^2 n_q^2 \log\left(\delta^{-1}\right)\right)$ random Pauli measurements of the state

$$\tilde{\rho} = \mathrm{tr}_{\boldsymbol{p}}\left(V_{\mathrm{STO}} \rho V_{\mathrm{STO}}^\dagger\right) \tag{4.18}$$

then suffice to estimate the expectation:

$$\tilde{\ell}\left(\rho\right) = \mathrm{tr}\left(M_{\boldsymbol{p_0}} \tilde{\rho}\right) \tag{4.19}$$

to additive error ϵ with probability at least $1-\delta$ using classical shadows [82]. Finally, observe that:

$$\begin{aligned} \ell\left(\rho\right) &= \mathrm{tr}\left(OU\rho U^\dagger\right) \\ &= \mathrm{tr}\left(\tilde{O}_{\boldsymbol{p_0}} \tilde{U}_{\boldsymbol{p_0}} \, \mathrm{tr}_{\boldsymbol{p}}\left(V_{\mathrm{STO}} \rho V_{\mathrm{STO}}^\dagger\right) \tilde{U}_{\boldsymbol{p_0}}^\dagger\right) \\ &= \tilde{\ell}\left(\rho\right). \end{aligned} \tag{4.20}$$

□

Note that in principle, the sample complexity of this procedure can potentially be improved to $\tilde{\mathrm{O}}(|O|_\infty^2 \epsilon^{-2} n_\lambda n_q^2 \log\left(\delta^{-1}\right))$ as $\tilde{\rho}$ only has $n_\lambda n_q^2$ degrees of freedom. However, the block diagonal structure over irreps is lost when transforming to the bitstring encoding $|\boldsymbol{\lambda}\rangle |\boldsymbol{q}\rangle |\boldsymbol{p}\rangle$ via V_{STO}, and thus we arrive at the sample complexity given.

In the above considerations, the group G and representation R do not directly enter. In practice, however, we might want to start with those two. The irreps of X are in one-to-one correspondence with those of R. By a simple corollary of the von Neumann bicommutant theorem, the dimensions of the irreps of X are the multiplicities of the irreps of R. We thus have that:

$$\dim\left(X\right) = \sum_\lambda \dim_X\left(\lambda\right)^2 = \sum_\lambda \mathrm{mult}_R\left(\lambda\right)^2. \tag{4.21}$$

Thus, the problems discussed above become classically tractable if the number n_λ of irreps of G with nonzero multiplicity in R, as well as the maximum multiplicity n_q of

an irrep λ in R, both grow only polynomially with n.

4.4 Permutation Invariance on Qubits

We now discuss such an example of a symmetry group with low-multiplicity irreps. Namely, we will apply the previously described procedures to the case where G is given by S_n and R is the representation on n qubits acting by permutations

$$R(\pi)\left|i_1\right\rangle \otimes \left|i_2\right\rangle \otimes \cdots \otimes \left|i_n\right\rangle = \left|i_{\pi^{-1}1}\right\rangle \otimes \left|i_{\pi^{-1}2}\right\rangle \otimes \cdots \otimes \left|i_{\pi^{-1}1}\right\rangle . \tag{4.22}$$

A straightforward basis for the algebra of invariant operators can be obtained by applying the Reynold's operator in Equation (4.2) to the Pauli basis. Normalizing such that all operators A_i are sums of unit norm Pauli terms, we obtain:

$$A_{\boldsymbol{i}} = \frac{1}{i_1!i_x!i_y!i_z!} \sum_{\pi \in \mathrm{S}_n} R(\pi) \left(\sigma_1^{\otimes i_1} \otimes \sigma_x^{\otimes i_x} \otimes \sigma_y^{\otimes i_y} \otimes \sigma_z^{\otimes i_z}\right) R^{-1}(\pi) \tag{4.23}$$

for every 4-tuple of positive integers

$$\boldsymbol{i} = (i_1, i_x, i_y, i_z) : i_1 + i_x + i_y + i_z = n, \tag{4.24}$$

where σ denote Pauli operators, and $\sigma_1 = \mathrm{id}_2$. In the remainder of this Section we assume that the systems, dynamics, and measurements being studied are given in this basis. This is a natural setting for, for instance, quantum machine learning when qubit permutation invariance is known to be present [47]. However, one could imagine systems which "secretly" obey permutation invariance (or are secretly symmetric under some other large symmetry group) and we are, for instance, only given oracular access to matrix elements. We give no classical algorithms under such an input model, and this may be a scenario where a quantum advantage still exists.

The dimension of the algebra X is of order $\mathrm{O}(n^3)$, and the previously stated theorems can be applied, reducing the naive ground state algorithm for permutation-invariant Hamiltonians on n qubits from an exponential to a polynomial runtime in

n. This is formalized below.

Corollary 4.5. *The ground state energy of a permutation-symmetric Hamiltonian on n qubits, given as h_i in the basis of symmetrized Pauli monomials above, can be computed in time $\mathrm{O}\left(n^{3\omega}\right)$ via Theorem 4.2.*

Proof. All that is needed for applying Theorem 4.2 are the structure constants of the algebra X, which are computed in Appendix C.2 (see Lemma C.1). □

One can similarly find the ground state of such an H efficiently as well. The output of the classical algorithm is a classical description of the state which can be efficiently constructed on a quantum computer via the Schur transform [151].

Corollary 4.6. *The ground state and ground state energy of a permutation-invariant Hamiltonian on n qubits, given as h_i in the basis of symmetrized Pauli monomials above, can be computed in time $\mathrm{O}\left(n^{\omega+1}\right)$ via Theorem 4.3.*

Proof. To apply Theorem 4.3, we must know the action of $A\left(X\right)$ on nontrivial eigenvectors of its projectors onto irreps. These eigenvectors are just the Schur basis [150]; we discuss this basis in more detail in Appendix C, where we also explicitly give analytical expressions for matrix elements of $A\left(X\right)$ (see Lemma C.2). It is then easy to see that $\dim_X\left(\lambda\right) = \mathrm{O}\left(n\right)$, and also that the number of irreps with nonzero multiplicity is $\mathrm{O}\left(n\right)$. From Theorem 4.3, we immediately see that this gives an $\mathrm{O}\left(n^{\omega+1}\right)$-time algorithm for computing the ground state of S_n-equivariant Hamiltonians in the Schur basis. □

Remark 4.7. Though the structure constants $X_{\boldsymbol{k}}^{\boldsymbol{i},\boldsymbol{j}}$ and matrix elements $F_{q_\lambda,q'_\lambda}^{\boldsymbol{i},\lambda}$ for the completely symmetrized Pauli representation are problem independent, it is important to note that runtimes for evaluating the analytical expressions can be expensive polynomials in n that may matter in practice. Namely, we give expressions for the structure constants that take a total time $\mathrm{O}\left(n^{15}\right)$ and matrix elements that take a total time $\mathrm{O}\left(n^{10}\right)$ to evaluate numerically. We leave more efficient evaluations of these to future work.

Finally, we consider an application of Theorem 4.4 to the symmetric group case. We note that the Schur transform on n qubits can be implemented up to an accuracy ϵ in time $\tilde{\mathrm{O}}\left(n \operatorname{poly}\log\left(\epsilon^{-1}\right)\right)$ [150], giving an efficient (approximate) implementation of V_{STO}. As a specific application of this result, we now consider a learning problem for which a variational quantum algorithm was given in Reference [47]. We emphasize that here, just as in Theorem 4.4, we do not require that the input states ρ_i respect the symmetries of the model.

Corollary 4.8 (Efficient classical simulation of permutation-invariant models)**.** *Consider a binary classification problem with labels $y_i \in \{-1, 1\}$ and empirical loss*

$$\hat{\mathcal{L}}\left(\boldsymbol{\theta}\right) = -\frac{1}{M}\sum_{i=1}^{M} y_i \ell_{\boldsymbol{\theta}}\left(\rho_i\right), \tag{4.25}$$

where $\ell_{\boldsymbol{\theta}}\left(\rho_i\right)$ is as in Equation (4.16) with a $\boldsymbol{\theta}$-dependent U. $\hat{\mathcal{L}}$ can be estimated to additive error ϵ at P points in time

$$\tilde{\mathrm{O}}\left(M\left|O\right|_{\infty}^{2}\epsilon^{-2}n^{5}\log\left(\frac{P}{\delta}\right) + MPn^{\omega+1}\right) \tag{4.26}$$

with total probability of success at least $1-\delta$.

Proof. This follows immediately from Theorem 4.4 with $\delta \to \frac{\delta}{P}$ by the union bound. □

Corollary 4.8 implies that the loss of these models can be estimated completely classically when the states ρ_i are given as certain classical shadows descriptions; in Appendix C.4, we also show that this procedure is efficient when the ρ_i have efficient matrix product state descriptions, even if they do not respect the symmetries of the model. As a point of comparison, consider the runtime of using a variational quantum algorithm to perform this binary classification task. Assume the variational circuits are of depth $\Omega\left(n^3\right)$ as required in Theorem 3 of Reference [47] to ensure convergence. Then—taking $\Omega\left(\left|O\right|_{\infty}^{2}\epsilon^{-2}\right)$ samples for each measurement to achieve an overall shot noise of $\mathrm{O}\left(\epsilon\right)$—this yields an overall runtime of $\Omega\left(MP\left|O\right|_{\infty}^{2}\epsilon^{-2}n^{3}\right)$.

For P sufficiently large, compare this to the time $\mathrm{O}\left(MPn^{\omega+1}\right)$ algorithm found for the classical algorithm where, even if quantum states are given as input, a classical shadow representation can be measured in quantum depth only $\mathrm{O}\left(n \operatorname{poly}\log\left(\epsilon^{-1}\right)\right)$. Unlike the quantum algorithm, this algorithm can be parallelized over irreps (i.e. over n_λ) easily, giving an effective runtime $\mathrm{O}\left(MPn^{\omega}\right) = \mathrm{o}\left(MPn^3\right)$. Even for P small, given many QPUs capable of running depth $\sim n$ quantum circuits, the classical algorithm parallelizes more effectively than the quantum algorithm as the required shadow tomography can be parallelized over shots.

4.5 Conclusion

We have specified a general framework for classically simulating highly symmetric quantum systems. Specializing to the symmetric group, we showed that these techniques yield an efficient classical algorithm for finding the ground state of quantum systems obeying an S_n symmetry, evaluating dynamics, and simulating S_n-equivariant quantum machine learning models. We hope that this framework sets the foundations for the future study of classical characterizations of symmetric quantum systems.

Chapter 5

Interpretable Quantum Advantage in Neural Sequence Learning

The results of this chapter were featured in Reference [52], work done in collaboration with Hong-Ye Hu, Jin-Long Huang, and Xun Gao.

5.1 Introduction

In Chapter 4, we showed how certain approaches to construct efficiently trainable quantum machine learning (QML) architectures based on symmetries yield efficient classical simulation algorithms. This begs the natural question: is there *any* regime where a trainable QML architecture exists, yet also yields a provable quantum advantage over classical architectures?

Previous results [56, 58–60, 153] have studied quantum advantages in machine learning classical data, though they rely on complexity theoretic assumptions. Not only do these architectures suffer from barren plateaus [63–65, 97] and poor local minima [49, 50, 77] due to their universal nature, it is unclear what realistic classical data sets one should expect a separation to hold in practice since the proofs of separation are abstract. Because of these concerns, it has become increasingly clear that quantum models should be carefully constructed to fit the task at hand. Above all else, the *interpretability* of any expressivity separation achieved by a QML

model has become increasingly important. Interpretability reveals which features of quantum mechanics yield more expressive models compared to classical models and, armed with this knowledge, allows one to find classes of problems where a practical quantum advantage on real data is potentially achievable.

Wishing to construct a model with an interpretable quantum advantage, we here focus on *sequence-to-sequence* learning tasks [84], and consider a quantization of *linear recurrent neural networks* (LRNNs) [48]. Classical LRNNs are recurrent neural networks with only linear activation functions. Such models can equivalently be considered a classical dynamical system governed by quadratic Hamiltonian evolution in the canonical variables $(\boldsymbol{q}, \boldsymbol{p})$. By lifting these canonical variables to operators $(\hat{\boldsymbol{q}}, \hat{\boldsymbol{p}})$ that satisfy the canonical commutation relations (in units where $\hbar = \frac{1}{2}$):

$$[\hat{q}_j, \hat{p}_k] = \frac{\mathrm{i}}{2}\delta_{jk}, \tag{5.1}$$

we arrive at a continuous variable (CV) quantum model where time evolution on an eigenstate of the canonical operators is performed under a quadratic Hamiltonian. To measure properties of the state of the system, the most natural choice is to perform *homodyne measurement*; that is, measure linear combinations of the canonical operators $\hat{q}_j$ and $\hat{p}_k$. This yields a quantum generative model where all operations are Gaussian. However, as all operations, initial states, and measurements are Gaussian, there are efficient Wigner function based simulations of sampling from such a system [85]. In other words, such models on n modes are equivalent to deep belief networks [86]—a class of commonly used classical models—with $2n$ latent variables.

Instead, we extend this model slightly further by allowing for measurements of the canonical operators *modulo* 2π, beginning in an eigenstate of periodic functions of the canonical operators [94, 95]. We call this introduced class of models *contextual recurrent neural networks* (CRNNs). Our main result is that CRNNs are more memory efficient at expressing certain distributions than essentially all trainable classical sequence models—independent of their internal, latent representations—even though CRNNs are not universal for CV quantum computation. Concretely, we show uncon-

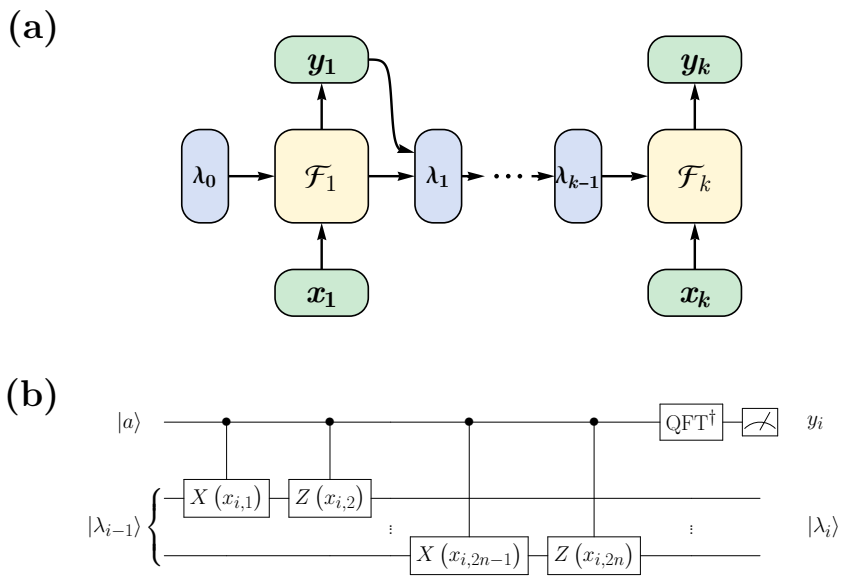


Figure 5-1: (a) An online neural sequence model. The model autoregressively takes input tokens $\boldsymbol{x}_i$ and outputs decoded tokens $\boldsymbol{y}_i$ with the map $\mathcal{F}_i$. The model also has an unobserved internal memory with state $\boldsymbol{\lambda}_{i-1} \in L$ that $\mathcal{F}_i$ can depend on. When the model is quantized to a CRNN, the n-dimensional space of λ_i is promoted to the Hilbert space of n qumode states $|\lambda_i\rangle$. (b) An implementation of a phase estimation circuit for CV Pauli operators, which forms the recurrent cell of the CRNN we use to prove our separations. Here, $|a\rangle$ is a fixed ancilla state and "QFT" represents the quantum Fourier transform. Formally, if $|a\rangle$ is a GKP state, this circuit allows for infinite precision measurements. In practice, $|a\rangle$ can be a tensor product of a constant number of qubit $|+\rangle$ states for finite precision phase estimation.

ditionally that there exists a class of CRNNs with $\mathrm{O}\left(n\right)$ qumodes that can express certain distributions that no "reasonable" (which we later describe) classical model is able to represent without an $\Omega\left(n^2\right)$-dimensional latent space. Though this is only a quadratic separation in memory, the time complexity of inference for classical models is typically superlinear in the model size [48, 87–89], suggesting a superquadratic time separation. As we show a memory (rather than a time) separation, our results also potentially point to a practical *generalization* advantage for CRNNs, as smaller models tend to generalize better than larger models due to formalized versions of Occam's razor [154].

Moreover, we are able to show directly that this quantum advantage is due to quantum contextuality [90–94] present in our quantum model. Previously, quantum

contextuality was known to be the resource for the expressive power of a certain class of quantized Bayesian networks [96]. Our results show that this resource can be used to separate quantum models even from neural networks, which can be exponentially more efficient than generic Bayesian networks. Intuitively, quantum contextuality is the statement that quantum measurement results depend on which measurements were previously performed, even if the measurements in question commute. In other words, quantum contextuality is the statement that the measurement of quantum observables cannot be thought of as the revealing of preexisting classical values for the observables. Here, we give a proof of the intuition that reasonable classical models cannot get around the need to "memorize" the measurement context of given observables, which is what yields the quadratic memory separation between the quantum and classical models.

Qualitatively, quantum contextuality is similar to the linguistic contextuality present in sentences. Namely, the meaning of a given word in a sentence depends heavily on other words in the sentence, and without this context has no fixed, single meaning. Inspired by this, we also test our constructed model against state-of-the-art classical models on a real-world translation task to investigate whether the ability of the quantum model to store information in its measurement context yields a practical quantum advantage in modeling the long-time correlations present in typical sequential data sets. In particular, we evaluate the performance of an LRNN [155], an RNN with gated recurrent units (GRU RNN) [88], a Transformer [89], a Gaussian model, and our introduced contextual model on a standard Spanish-to-English data set [156]. We show that our introduced contextual model achieves better translation performance compared to all other models at each model size we consider. This separation holds even when the online models are constrained to have a similar (and where possible, the same) number of trainable parameters in each recurrent cell.

Our methods provide a novel strategy for designing QML models: through the quantization of simple classical machine learning models with some minimal quantum extension. Though such models are most likely unable to outperform state-of-the-art classical machine learning models on *all* tasks, the intuition gleaned from the simplic-

ity of the quantum models gives guidance as to which problems the quantum models may outperform classical models on. Furthermore, the simplicity of the quantum models may circumvent the recent deluge of untrainability results of general quantum models [49, 50, 63–65, 77, 97] as it is known that there exists a trade-off between the trainability of such architectures and their generality [157]. Finally, as such models are restricted in their allowed operations, they are more amenable to experimental implementation than completely generic quantum models.

5.2 Classical and Quantum Neural Sequence Learning

5.2.1 Classical Sequence Learning

Sequence-to-sequence or *sequence* learning [84] is the approximation of some given conditional distribution $p(\boldsymbol{y} \mid \boldsymbol{x})$ with a model distribution $q(\boldsymbol{y} \mid \boldsymbol{x})$. This framework encompasses sentence translation tasks [84], speech recognition [158], image captioning [159], and many more practical problems.

Sequence modeling today is typically performed using neural network based generative models, or *neural sequence models*. Generally, these models are parameterized functions that take as input the sequence $\boldsymbol{x}$ and output a sample from the conditional distribution $q(\boldsymbol{y} \mid \boldsymbol{x})$. The parameters of these functions are trained to minimize an appropriate loss function, such as the (forward) empirical cross entropy:

$$\hat{H}(p, q) = -\frac{1}{|\mathcal{T}|} \sum_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{T}} p(\boldsymbol{y} \mid \boldsymbol{x}) \log\left(q(\boldsymbol{y} \mid \boldsymbol{x})\right), \tag{5.2}$$

where $\mathcal{T} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$ are samples from $p(\boldsymbol{x}, \boldsymbol{y})$. The backward empirical cross entropy is similarly defined, with $p \leftrightarrow q$. Note that a model with support on an incorrect translation (i.e. $q \neq 0$, $p = 0$) yields an infinite backward cross entropy, and a model failing to have support on a correct translation (i.e. $p \neq 0$, $q = 0$) yields an infinite forward cross entropy.

To maintain a resource scaling independent of the input sequence length, neural sequence models usually fall into one of two classes: *online sequence models* (also known as *autoregressive sequence models*) [48, 87, 88], or *encoder-decoder models* (which include state-of-the-art sequence learning architectures, such as Transformers) [84, 89]. We focus on online models here, and discuss encoder-decoder models in more detail in Appendix D.1.

In online models, input tokens $\boldsymbol{x}_i$ are translated in sequence to output tokens $\boldsymbol{y}_i$ via functions $\mathcal{F}_i$. An unobserved internal memory (or *latent space*) L shared between time steps allows the model to represent long-range correlations in the data. A diagram of the general form of online models is given in Figure 5-1(a). Generally, there are no restrictions on the forms of $\mathcal{F}_i$, though most neural sequence models are composed of simple smooth (or almost everywhere smooth) functions out of training considerations [48, 87–89]. Here, we generalize from the typical smoothness constraints and consider *locally Lipschitz* maps.

Assuming that the codomain of $\mathcal{F}_i$ is $\mathbb{R}^m$, all maps that are almost everywhere differentiable with locally bounded Jacobian norm are locally Lipschitz [160]. Realistically, then, locally Lipschitz models can be thought of as all models trainable using gradient based methods. Equivalently, they can be thought of as models not infinitely sensitive to infinitesimal changes in their inputs. This includes all models with standard nonlinearities, including those with rectified linear unit (ReLU), hyperbolic tangent, and sigmoid activation functions. Note that this condition is much weaker than a *globally Lipschitz* constraint. We give a formal definition of local Lipschitzness in Appendix D.1.

Though neural networks are often described as functions of real-valued inputs, in practice they are implemented at finite precision. We emphasize that where we analytically consider such networks here—such as in Section 5.3—we consider the formal description of neural networks, which assumes infinite precision. Our numerical experiments in Section 5.4, however, give evidence that our analytic results also hold in the finite precision regime. We discuss this in more detail in Appendix D.3.

$X_1(\alpha)$	$X_2(\alpha)$	$X_1(\alpha)^\dagger X_2(\alpha)^\dagger$
$X_1(\alpha)^\dagger Z_2\left(\frac{\pi}{2\alpha}\right)^\dagger$	$Z_1\left(\frac{\pi}{2\alpha}\right)^\dagger X_2(\alpha)^\dagger$	$-X_1(\alpha) Z_1\left(\frac{\pi}{2\alpha}\right) X_2(\alpha) Z_2\left(\frac{\pi}{2\alpha}\right)$
$Z_2\left(\frac{\pi}{2\alpha}\right)$	$Z_1\left(\frac{\pi}{2\alpha}\right)$	$Z_1\left(\frac{\pi}{2\alpha}\right)^\dagger Z_2\left(\frac{\pi}{2\alpha}\right)^\dagger$

Table 5.1: An example of CV quantum contextuality using a Mermin–Peres magic square [93], with CV Pauli operators $X_i(a), Z_i(a)$ generated by $-2\mathrm{i}a\hat{p}_i, 2\mathrm{i}a\hat{q}_i$, respectively. For any real $\alpha \neq 0$, all operators in each row and column commute. Additionally, the product of each row and column is the identity operator, except for the final column, which gives -1. Thus, definite classical values cannot be assigned to each operator without yielding a contradiction.

5.2.2 Contextual Recurrent Neural Networks

We now consider a quantization of a simple online model. Generally, online models can be interpreted as a classical dynamical process, where queries $\boldsymbol{x}_i$ are made to a physical system described by the latent state $\boldsymbol{\lambda}_{i-1}$, yielding a result $\boldsymbol{y}_i$ and transforming the latent state $\boldsymbol{\lambda}_{i-1} \mapsto \boldsymbol{\lambda}_i$ (see Figure 5-1(a)). For *linear recurrent neural networks* (LRNNs), this can be interpreted as the physical process of querying properties of an underlying system described by $\boldsymbol{\lambda}_i$ undergoing Hamiltonian evolution under a quadratic Hamiltonian; this can be seen straightforwardly from Hamilton's equations and the linearity of the model. When quantizing the canonical position and momentum variables to operators satisfying the canonical commutation relations, such a model can then be interpreted as performing sequential measurements on a system undergoing evolution via *Gaussian operations*. When these measurements are restricted to homodyne measurements and all inputs are Gaussian states, this process can be simulated classically with memory linear in the number of modes of the Gaussian system [85]. We minimally extend this, and allow for *non-Gaussian measurements*. In particular, we are here interested in measuring via phase estimation the CV analogs of the Pauli operators [161] (in units where $\hbar = \frac{1}{2}$):

$$X_i(a) = \mathrm{e}^{-2\mathrm{i}a\hat{p}_i}, \qquad Z_i(a) = \mathrm{e}^{2\mathrm{i}a\hat{q}_i}. \tag{5.3}$$

We also promote the initial state of the network to a *Gottesman–Kitaev–Preskill (GKP) state* [95], which is an eigenstate of CV Pauli operators. We call a recurrent online model beginning in a GKP state, with cell that takes as input $\boldsymbol{x}_i$ a description of a CV Pauli operator and returns its measurement result $\boldsymbol{y}_i$, a *contextual recurrent neural network* (CRNN).

This measurement can formally be performed at infinite precision using Gaussian operations and homodyne measurement with fixed ancilla GKP states [95, 162]. A circuit description of this is given in Figure 5-1(b), where $|a\rangle$ is a uniform superposition over squeezed states $|s\rangle$ with $\hat{q}\,|s\rangle = q\,|s\rangle$, where $q \equiv 0 \pmod{2\pi}$. When performed sequentially on an initial GKP state, these measurements are what we consider when we compare in Section 5.3 CRNNs against the infinite precision classical neural networks described in Section 5.2.1. In this scenario, the model is not universal for CV quantum computation, even when additional Gaussian operations within the latent space are added [163]. Counterintuitively, when the initial state is the vacuum state or a finitely squeezed GKP state, the model *is* universal [163, 164]; this suggests a potential superpolynomial advantage in the expressive power and the time complexity of inference when implemented at finite precision. We discuss this in more detail in Appendix D.3.

Just as in the classical case, one can consider a finite precision approximation of these measurements. In this scenario, phase estimation using ancilla qubits can be performed for each measurement [165]. We discuss proposals for the experimental implementation of such a measurement in Appendix D.3. In general, parameterized Gaussian operations can be included within each recurrent cell to yield a *trainable CRNN*. This is a special case of the CV neural networks considered in Reference [26], which also considered the training of such networks. For our expressivity separations, however, we consider the fixed CRNN instance given in Figure 5-1(b).

For our purposes, these measurements are important as CV Pauli operators exhibit *quantum contextuality* [94], in complete analogy with the contextuality present in qubit Pauli operators [93]. Quantum contextuality is the statement that no definite classical values can be assigned to quantum operators, even when the measured

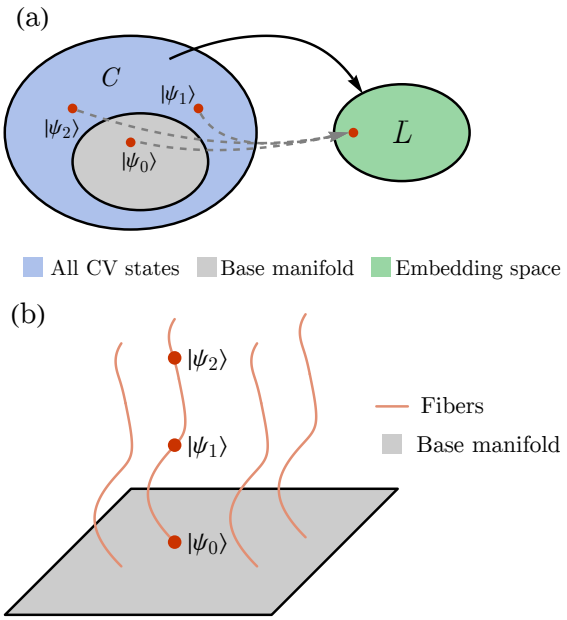


Figure 5-2: (a) A schematic of the classical model when the dimension of the latent space L (green) is less than $\frac{n(n-3)}{2}$, where n is the number of modes in the stabilizer measurement translation task. Here, "$|\psi_i\rangle$" labels the input sequence that is composed of the stabilizers of $|\psi_i\rangle$; we discuss this labeling in more detail in our proof sketch of Theorem 5.2. We show that when $\dim(L) < \frac{n(n-3)}{2}$, in the neighborhood of some input, only a subspace of inputs (gray) of the same dimension as L are mapped injectively. (b) A sketch of the space of inputs, with fibers locally induced by the model. The base manifold is mapped injectively to L. All points on a fiber (e.g. $|\psi_1\rangle$, $|\psi_2\rangle$) map to the same point as their base point (e.g. $|\psi_0\rangle$) in L. When the dimension of the fiber is large enough, we show that these states have contextual stabilizers. We then show that this implies that the states have a single-shot distinguishing measurement sequence.

operators in any given measurement scenario commute. For an example of this phenomenon, see Table 5.1; there is no consistent assignment of classical values to each operator in the Table for any real $\alpha \neq 0$.

5.2.3 Stabilizer Measurement Translation

We now focus on a classical sequence learning task that is naturally performed by the introduced CRNN. In particular, we consider the (k, n) *stabilizer measurement translation task*, parameterized by k and n. Leaving the formal definition for Ap-

pendix D.2, we give an informal definition here. We use the terminology of Figure 5-1 for clarity.

Definition 5.1 ((k, n) stabilizer measurement translation task, informal)**.** Given a k long sequence of classical descriptions $\boldsymbol{x}_i$ of CV Pauli operators on n modes, output a sequence of measurement outcomes y_i that is consistent with measuring these operators sequentially on a fixed GKP state $|\lambda_0\rangle$.

As described in Section 5.2.2, such measurement sequences can display nontrivial correlations due to quantum contextuality. Note that this task is distinct from the measurement of position and momentum operators. Here, we require the measurement of linear combinations of position and momentum operators *modulo* 2π, as we are measuring the phases of operators generated by position and momentum. This can be done using the CRNN cell described in Section 5.2.2. We consider in Appendix D.2 a slight generalization of this task, though here we consider Definition 5.1 with its fixed GKP initial state for simplicity.

5.3 Bounds on Stabilizer Measurement Translation

We now give statements and proof sketches of our main results, which are lower bounds on the performance of classical models in performing the stabilizer measurement translation task described in Section 5.2.3. This will give an expressivity separation between the classical and quantum sequence models.

For discrete models, quantum contextuality was the key resource for showing a separation in expressivity between classical and quantum models [96]. Using different proof techniques, we here show that quantum contextuality is also the resource giving the separation between continuous classical and quantum models with infinite dimensional Hilbert spaces. To do this, we specialize to two classes of models: online neural sequence models, and encoder-decoder models (which include state-of-the-art models such as seq2seq models [84] and Transformers [89]). Here, we focus on the memory separation between CRNNs and classical online neural sequence models, and

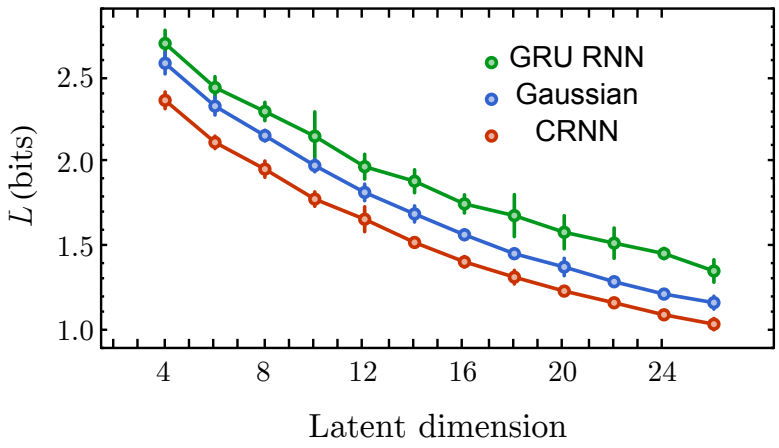


Figure 5-3: The forward empirical cross entropy (L) on a test set for a Spanish-to-English translation task as a function of the model dimension n for GRU RNNs, Gaussian RNNs, and CRNNs. The models are constrained such that the Gaussian and CRNN models have an identical number of parameters. The recurrent cells of the GRU RNN and quantum models have a number of parameters within 2.5% of each other at $n = 26$. Error bars denote the standard deviations of the loss over five independent training runs.

discuss a similar separation against encoder-decoder models in Appendix D.2. We also there formulate a general statement on the classical efficiency of simulating CV Pauli measurements on an initial GKP state, similar in spirit to the fact that the Gottesman–Knill theorem [166, 167] is optimal for qubit stabilizer simulation [168].

Our main result can be informally stated as the following Theorem (with the full statement and proof left to Appendix D.2). Note that, as discussed in Section 5.2.3, a CRNN can perform the stabilizer measurement translation task with n qumodes of memory.

Theorem 5.2 (Online stabilizer measurement translation memory lower bound, informal statement of Theorem D.2)**.** *Consider a locally Lipschitz online model with latent space L. If $\dim(L) < \frac{n(n-3)}{2}$, this model cannot achieve a finite backward cross entropy on the $(n+2, n)$ stabilizer measurement translation task.*

Proof sketch. The strategy of our proof is to show that, when the dimension of L is less than $\frac{n(n-3)}{2}$, the model must map an embedded submanifold K of the space of the first n inputs to the same point in L; in other words, the model loses the ability to

Input	"Debemos limpiar la cocina."
Truth	"We must clean up the kitchen."
CRNN	"We must clean the kitchen."
GRU	"We have to turn the right address."
Input	"Admití que estaba equivocada."
Truth	"I admitted that I was wrong."
CRNN	"I was wrong to say that."
GRU	"They had a thing to be true."
Input	"¿Cual es el lugar más bonito del mundo?"
Truth	"What's the most beautiful place in the world?"
CRNN	"What's the world largest place?"
GRU	"What's the best of is in?"
Input	"La caja es pesada."
Truth	"The box is heavy."
CRNN	"The box is heavy."
GRU	"My box is."

Table 5.2: Random samples of translation results for $n = 26$ models.

distinguish between inputs in K. The nontrivial aspect of this proof is to demonstrate that such a K exists, where distinct points in K yield different translations. As the model is unable to distinguish between points in K, this then demonstrates that the model will get a translation incorrect, corresponding to an infinite backward cross entropy on the stabilizer measurement translation task. This is equivalent to demonstrating that the quantum mechanical processes being described by points in K yield quantum states that are single-shot distinguishable.

To demonstrate that such a K exists, we use the local Lipschitzness of the model and the constant rank theorem [169]. This then implies that the map $\mathcal{F}$—given by the n-fold composition of the $\mathcal{F}_i$ as shown in Figure 5-1(a)—locally induces a fiber bundle on the input space (as shown in Figure 5-2), where $\mathcal{F}$ can be considered a projection onto the base manifold of this induced fiber bundle. We slightly abuse notation in the remainder of this proof sketch, and conflate the first n inputs (and their associated outputs) with the quantum state that would arise from this measurement sequence; this same labeling is used in Figure 5-2.

We consider a fiber of this fiber bundle, with the goal of proving that there exist points in this fiber that are single-shot distinguishable. We show in Appendix D.2

that the dimension of this fiber is large enough such that there exist three states in this fiber with stabilizers that exhibit quantum contextuality. We claim (and prove in Appendix D.2) that due to the presence of quantum contextuality in these stabilizers, these states have a distinguishing measurement sequence of length two. When performing this distinguishing measurement sequence, then, the model must give the incorrect measurement results for one of the three states, giving the lower bound on classical simulation. It is easy to see from Equation (5.2) that this yields both an infinite backward cross entropy when these sequences are in the data set $\mathcal{T}$.

As a simple example of this phenomenon, assume that three states $|\psi_1\rangle, |\psi_2\rangle, |\psi_3\rangle$ with classical representations in the same fiber are respectively stabilized by the rows of Table 5.1 for some real $\alpha \neq 0$. As $|\psi_3\rangle$ is stabilized by $Z_1 \left(\frac{\pi}{2\alpha}\right)^\dagger Z_2 \left(\frac{\pi}{2\alpha}\right)^\dagger$, upon measuring this operator, the measurement result is constrained to be 1 for the simulation to be accurate. The post-measurement state of $|\psi_1\rangle$ is then stabilized by $X_1(\alpha) X_2(\alpha)$ and $Z_1 \left(\frac{\pi}{2\alpha}\right)^\dagger Z_2 \left(\frac{\pi}{2\alpha}\right)^\dagger$. In particular, it is also stabilized by $X_1(\alpha) Z_1 \left(\frac{\pi}{2\alpha}\right)^\dagger X_2(\alpha) Z_2 \left(\frac{\pi}{2\alpha}\right)^\dagger$. Conversely, the post-measurement state of $|\psi_2\rangle$ is stabilized by $-X_1(\alpha) Z_1 \left(\frac{\pi}{2\alpha}\right)^\dagger X_2(\alpha) Z_2 \left(\frac{\pi}{2\alpha}\right)^\dagger$. Thus, then measuring $X_1(\alpha) Z_1 \left(\frac{\pi}{2\alpha}\right)^\dagger X_2(\alpha) Z_2 \left(\frac{\pi}{2\alpha}\right)^\dagger$ gives an incorrect translation for one of these states. This measurement sequence is single-shot, as only a single copy of the state being measured is used. □

Our results show that there is a general n versus $\Omega(n^2)$ bound in the memory requirements of contextual and classical models performing the stabilizer measurement translation task. In practice, this can yield an even greater separation in time complexity for given implementations of these models, as the time complexity of inference using classical models is typically superlinear in the model size. We discuss this in more detail in Appendix D.5.3.

5.4 Numerical Experiments

We now showcase the practical benefit of finding an interpretable advantage in the expressivity of our quantum model. Namely, it is able to give us intuition as to which data sets—beyond the constructed data set used in our proof—a CRNN may outper-

form classical machine learning models on. As previously discussed, the contextuality present in quantum operators behaves qualitatively similar to the linguistic contextuality present in language. That is, words can have one of many meanings, and their exact definition only becomes apparent when considering their context in a sequence. This is important for translation tasks, where different meanings of a single word in one language have different translations in other languages. We here investigate whether the ability of the quantum model to store information in its measurement context yields a practical quantum advantage in modeling the long-time correlations present in a typical sequential data set.

To explore this intuition, we consider the application of a CRNN on a standard Spanish-to-English translation data set [156], with trainable Gaussian interactions within each recurrent cell. We also consider the performance of GRUs [88] in a seq2seq learning framework [84], and Gaussian models (with Gaussian measurements). Details of our numerical simulations for all of the models we consider are given in Appendix D.5, along with details of the architectures. We also discuss in Appendices D.4 and D.5 a $\Theta(n^2)$ memory classical simulation of CRNNs with n latent modes on a restricted space of Gaussian operations, which is what we use in our numerical simulations. The Gaussian and contextual models are constrained to have exactly the same number of trainable parameters, and each recurrent cell of the GRU has a parameter count within 2.5% of those of the quantum models at the largest model size considered. In Figure 5-3, we plot the final training performance of all of our models over five independent training runs. It is easy to see that the contextual model outperforms all models under consideration in forward empirical cross entropy at a wide range of model dimensions n. It is also apparent that training CRNNs is no more difficult than training GRUs, justifying our intuition that restricted quantum machine learning models should be more trainable than generic models [49, 50, 63–65, 77, 97]. Random samples of translation results after training are shown in Table 5.2.

We also compared the performance of CRNNs against classical LRNNs [155] and Transformers [89]; for the latter, we also tested the model sizes required for the quantum and classical models to achieve some fixed target in performance. We find

that CRNNs substantially outperformed both LRNNs and Transformers. We also verified that to attain a fixed target loss realized by a CRNN of size $n = 26$ a Transformer needed a memory of dimension roughly $\frac{n(n-3)}{2}$, which is the same as that proven in the setting of Theorem 5.2. We give details of these results in Appendix D.6.

5.5 Conclusion

Our results pinpoint quantum contextuality as a resource that can be used to enhance traditional machine learning models. We achieved this by constructing a sequence learning task parameterized by n that a contextual quantum model (a CRNN) of size n is able to model, yet provably no classical neural networks of size subquadratic in n can model due to their noncontextuality. To our knowledge, this is the first unconditional proof of an expressivity separation between a quantum neural network and classical neural networks on classical data. By explicitly demonstrating that quantum contextuality is the source of this advantage, we are also able to provide intuition as to which classes of problems CRNNs are able to outperform traditional machine learning models in solving. Our numerics confirm the intuition that CRNNs perform extremely well on problems exhibiting linguistic contextuality, such as the Spanish-to-English translation task we consider here.

The simple structure of CRNNs also allow (finite precision approximations of) them to be more amenable to potential experimental implementations when compared with completely general quantum architectures. In particular, all operations in the contextual model are Gaussian, up to the requirement for interactions with fixed ancilla states to perform the required non-Gaussian measurements. Furthermore, though we do not consider the effects of noise here in detail, a key component of CRNNs is intermediate measurement using ancilla GKP states; this procedure is an important building block of CV quantum error correction [95], and in future work we hope to study how the addition of fast classical feedback may allow for error correction in CRNNs. The restricted nature of the model may also circumvent the poor training landscapes of generic quantum neural networks [49, 50, 63–65, 77, 97],

though we leave further investigation of this to future work.

We believe that the specifics of the CRNN architecture can be relaxed somewhat. Due to recent results linking non-Gaussian operations to quantum contextuality [170, 171], we suspect that any non-Gaussian measurement would make a suitable replacement for the stabilizer measurements we consider here for technical reasons. We also suspect that the technical requirement that the measurements be made with infinite precision to be an artifact of the nature of our proof, which compares the quantum architecture with infinite precision classical models. We believe that in practice, performing phase estimation using ancilla qubits instead of GKP (or other non-Gaussian CV) states is all that is necessary for a practical separation. In fact, such a finite precision implementation may counterintuitively yield a *larger* quantum advantage, as our architecture implemented with a finitely squeezed initial Gaussian state is universal for CV quantum computation [163, 164]. We discuss these two points in more detail in Appendix D.3.

CRNNs demonstrate that even the quantization of a very simple class of classical architectures—here, the class of LRNNs—is able to outperform a wide range of classical models on certain tasks, even if the classical models are much more powerful than LRNNs. We leave for future work the quantization of more powerful classical architectures, which may achieve a practical quantum advantage on a wider variety of tasks than we consider here.

Chapter 6

Conclusion

In this thesis we have demonstrated generally a trade-off between two important aspects of any quantum machine learning (QML) algorithm: the efficiency of which they are trained, and their expressive power. Unfortunately, the requirement to balance these two aspects makes constructing useful QML algorithms a difficult process in general.

Luckily, there is hope. Though the separation is modest, in Chapter 5 we demonstrated the existence of a sequence learning task for which there existed a quadratic separation in the size of (trainable) quantum and classical models representing the distribution. Where does the demonstration of such a separation leave us?

In practice, a quadratic separation may not be sufficient to justify the use of the quantum model over any classical model. This is particularly true if the model is used to represent the long sequences it is expected to have an advantage over classical models on; such a task would probably require some form of quantum error correction, for which large constant factors may make impractical any quadratic time separation in the near future [172]. Perhaps this is too pessimistic for the system discussed in Chapter 5—where only Gaussian interactions, along with state preparation and measurement (SPAM) errors need to be error corrected—but ideally a larger separation would be shown.

One potential way forward is to consider generalizations of the quantized linear system of Chapter 5. Particularly, one can consider quantizing *polynomial* systems

of bounded degree. There, the analog of the graph states considered in Chapter 5 are *hypergraph states* [173]. Using the rough heuristic demonstrated throughout this thesis that training times should scale with the dimension of Hilbert space explored by the model, these QML architectures may yield a natural way to tunably balance between trainability and expressivity.

Another way to extend the results presented here would be to consider notions of trainability that are *model independent*. In particular, our untrainability results specifically looked at certain randomized classes of QML models to ensure the tractability of the proofs. This stands in stark contrast with methods used classically to prove the unlearnability of—for instance—solutions to NP-hard problems. In the classical setting, these sorts of learnability problems are tackled by looking at the solution space of the problem itself, most famously via analogs of the *overlap gap property* (OGP) [174–176]. Essentially, OGP results imply a clustering behavior in the solution space for a variety of problems that cannot be learned by so-called "stable" algorithms. Shallow instances of the quantum approximate optimization algorithm (QAOA) have been shown to fall in this class, and thus the unlearnability of e.g. Max-Cut using shallow QAOA can be derived using essentially no quantum techniques [175, 176]. Exciting recent progress in studying the low-energy states of local Hamiltonians may open the door to similar analysis for quantum problems [177], but more work remains to be done.

The field of quantum information—which lies at the intersection of mathematics, physics, and computer science—demonstrates how novel results can be found by looking where fields coexist. The results in this thesis carve out a niche where quantum algorithms, in their intersection with machine learning, may prove practically useful. Though the landscape of all possible architectures is vast, we now know where to begin the search.

Appendix A

Technical Details for Chapter 2

A.1 Variational Quantum Algorithms as Random Fields

A.1.1 Technical Exposition of Variational Quantum Algorithms

Variational quantum algorithms (VQAs) are a class of quantum generative model where one expresses the solution of some problem as the smallest eigenvalue and its corresponding eigenvector (typically called the *ground state*) of an objective Hermitian matrix H. In Section 1.2.3, we gave a brief and general overview of VQAs; we here provide more details, with some slight changes in convention such that our theorems are more easily expressed.

Given a choice of generative model—often called an *ansatz* in the quantum algorithms literature:

$$|\boldsymbol{\theta}\rangle = \prod_{i=1}^{q} U_i\left(\theta_i\right)|\psi_0\rangle \tag{A.1}$$

that for some choice $\boldsymbol{\theta}$ closely approximates the ground state of H, the solution is encoded as the minimum of the loss function

$$\tilde{F}\left(\boldsymbol{\theta}\right) = \langle\boldsymbol{\theta}| H |\boldsymbol{\theta}\rangle . \tag{A.2}$$

This loss function can be computed on a quantum computer efficiently, under some conditions on the matrix H. For simplicity of analysis, throughout this paper we will

consider the loss function

$$F(\boldsymbol{\theta}) = \frac{\langle \boldsymbol{\theta}| H |\boldsymbol{\theta}\rangle - \lambda_1}{\overline{\lambda} - \lambda_1}, \tag{A.3}$$

where λ_1 is the smallest eigenvalue of H; this has the same loss landscape as Equation (A.2), but is minimized at $F = 0$ (assuming a sufficiently expressive $|\boldsymbol{\theta}\rangle$) and is normalized by the mean eigenvalue of $H - \lambda_1$. In Equation (A.1), q is referred to as the *depth* of the circuit, and the initial vector (i.e. quantum state) $|\psi_0\rangle \in \mathbb{C}^{2^n}$ is fixed throughout the optimization procedure. Different choices of U_i constitute different choices of ansatz for the ground state of H.

Ansatz design choice generally falls in one of two categories: Hamiltonian informed ansatzes, and Hamiltonian agnostic ansatzes. Examples of Hamiltonian informed ansatzes include the chemistry-inspired unitary coupled cluster ansatz [19] and the adiabatically inspired quantum approximate optimization algorithm (QAOA) ansatz [99], known outside of the context of combinitarial optimization as the Hamiltonian variational ansatz (HVA) [178]. These ansatzes depend solely on the problem objective Hamiltonian H, and are usually physically motivated ansatzes which, in some limit, have convergence guarantees. Hamiltonian agnostic ansatzes, conversely, depend solely on the hardware the VQA is run on, and not at all on the problem objective H. This class of ansatzes includes the hardware-efficient ansatz [102]. These ansatzes are designed to eke out as much depth as possible in the objective ansatz $|\boldsymbol{\theta}\rangle$ by using U_i that can be easily implemented on the given quantum device.

Though hardware-efficient ansatzes generally can be run at larger depth q than Hamiltonian informed ansatzes, the very generic nature of the ansatz circuit means this class of ansatz is more difficult to train, often encountering barren plateaus in the optimization landscape that are difficult to escape from when q is large [63, 64, 97]. Heuristically, this can be understood as Hamiltonian agnostic objective functions being so expressive that it must explore essentially all of Hilbert space to find a local minimum, exponentially suppressing the gradients of the loss function [179].

We here consider a class of ansatzes that, like the hardware-efficient ansatz, is independent of the problem instance. In particular, we consider random parameterized

ansatzes of the form:

$$U_i \equiv \mathrm{e}^{-\mathrm{i}\theta_i Q_i} \tag{A.4}$$

for Pauli operators Q_i, where each Q_i is drawn uniformly and independently from the n-qubit Pauli operators. Throughout this paper, we will use q to denote the total number of Pauli rotations in $|\boldsymbol{\theta}\rangle$ as in Equation (A.1), p to denote the total number of independent parameters θ_i, and r_i to denote the number of Pauli rotations governed by a single independent parameter θ_i. For simplicity, we will assume $r_i = r_j \equiv r$ for all i, j, and thus take

$$r \equiv \frac{q}{p} \tag{A.5}$$

to be a natural number.

A.1.2 Mapping Variational Quantum Algorithms to Random Wishart Fields

With the background of VQAs in place, we will now show the asymptotic (weak) equivalence of VQAs with the random choice of ansatz described in Appendix A.1.1 to Wishart random fields. Throughout this section, we will consider a problem Hamiltonian H on n qubits, with ground state energy λ_1 and mean eigenvalue $\overline{\lambda}$. We also define the *degrees of freedom* parameter

$$m \equiv \frac{\|H - \lambda_1\|_*^2}{\left\|H - \overline{\lambda}\right\|_{\mathrm{F}}^2}, \tag{A.6}$$

whose interpretation will be discussed in Appendix A.1.3. Twice the degrees of freedom parameter m will turn out to govern the location of the transition from the underparameterized to the overparameterized regime (see Appendix A.3), and for physically relevant Hamiltonians is expected to be exponential in n (see Appendix A.1.3). We will also consider the Pauli decomposition of the nontrivial part of H:

$$H - \overline{\lambda} = \sum_{i=1}^{A} \alpha_i R_i, \tag{A.7}$$

where A is the number of terms in the Pauli decomposition and $\boldsymbol{\alpha}$ the Pauli coefficients.

We begin by showing the convergence of a class of randomized VQAs to a weighted sum of Wishart random fields at a rate $\gtrsim \log(n)$; the seemingly arbitrary shifts by the mean eigenvalue $\overline{\lambda}$ and the ground state energy λ_1 here will aid in future discussion, when we approximate the weighted sum of Wishart random fields with a single random field. The wide variety of assumptions will be discussed in detail in Appendix A.1.3. We demonstrate this convergence in terms of the Lévy distance $L(F_n, G_n)$, which metrizes weak convergence:

$$L(F_n, G_n) \to 0 \iff F_n, G_n \rightsquigarrow F. \tag{A.8}$$

Theorem A.1 (VQAs as RFs)**.** *Let $|\psi_0\rangle$ be an arbitrary stabilizer state (e.g. a computational basis state) on n qubits. Fix a sequence of q angles $\theta_i \in [-\pi, \pi]$ such that each θ_i is present r times in the sequence. We let $p = \frac{q}{r}$ denote the number of distinct parameters. Select an ansatz*

$$|\boldsymbol{\theta}\rangle \equiv \prod_{i=1}^{q} U_i(\boldsymbol{\theta}) |\psi_0\rangle \equiv \prod_{i=1}^{q} \mathrm{e}^{\mp \mathrm{i}\theta_i Q_i} C |\psi_0\rangle \tag{A.9}$$

by independently at random drawing each $\pm Q_i$ uniformly from the n-qubit Pauli group $\mathbb{P}_n$ and C from the n-qubit Clifford group $\mathbb{C}_n$. Consider the scaled and shifted

$$\tilde{H} \equiv \frac{H - \overline{\lambda}}{\overline{\lambda} - \lambda_1} = \frac{H - \overline{\lambda}}{2^{-n} \|H - \lambda_1\|_*}, \tag{A.10}$$

where $\|\cdot\|_*$ is the nuclear norm. Then, the random variational objective function

$$F_{VQA}(\boldsymbol{\theta}) = \frac{\langle \boldsymbol{\theta}| H |\boldsymbol{\theta}\rangle - \lambda_1}{\overline{\lambda} - \lambda_1} = \frac{\langle \boldsymbol{\theta}| H |\boldsymbol{\theta}\rangle - \lambda_1}{2^{-n} \|H - \lambda_1\|_*} \tag{A.11}$$

has first two moments exponentially close in n as $n \to \infty$ to those of

$$F_{XHX}(\boldsymbol{w}) = 2^{-n} \left(\bigotimes_{i=1}^{p} \boldsymbol{w}_i^{\intercal} \right)^{\otimes r} \cdot \boldsymbol{X} \cdot \tilde{\boldsymbol{H}} \cdot \boldsymbol{X}^{\dagger} \cdot \left(\bigotimes_{i=p}^{1} \boldsymbol{w}_i \right)^{\otimes r} + 1, \tag{A.12}$$

where $\boldsymbol{w}_i$ *are points on the circle parameterized by* θ_i *and* $\boldsymbol{X} \in \mathbb{C}^{2^q \times 2^n}$ *is a matrix of i.i.d. complex standard jointly normal random variables. Furthermore, assuming*

$$\frac{\|\boldsymbol{\alpha}\|_\infty}{\overline{\lambda} - \lambda_1} \leq f(n)^{-1} \tag{A.13}$$

for some $f(n) = \Omega(1)$*, their distributions are bounded in Lévy distance by* $\tilde{\mathrm{O}}\left(\left(\frac{\lg(A) f(n) n}{A}\right)^{-1}\right)$.

Proof. The Feynman path integral representation (i.e. the exact Taylor expansion of the matrix exponentials using the fact that Pauli operators square to the identity) of the objective function Equation (A.11) is of the form

$$F_{\mathrm{VQA}} = \sum_{\boldsymbol{\gamma}, \boldsymbol{\gamma}' \in \{0,1\}^{\times q}} w_{\boldsymbol{\gamma}'}^\dagger w_{\boldsymbol{\gamma}} \langle \psi_0 | C^\dagger Q_{\boldsymbol{\gamma}'}^\dagger \tilde{H} Q_{\boldsymbol{\gamma}} C | \psi_0 \rangle + 1, \tag{A.14}$$

where $\boldsymbol{\gamma}$ labels a term in the path integral expansion of U,

$$w_{\boldsymbol{\gamma}} \equiv \prod_{i=1}^{q} \begin{cases} \cos(\theta_i), & \text{if } \gamma_i = 0 \\ \sin(\theta_i), & \text{if } \gamma_i = 1 \end{cases} \tag{A.15}$$

is the amplitude, and

$$Q_{\boldsymbol{\gamma}} \equiv (-\mathrm{i})^{\|\boldsymbol{\gamma}\|_0} \prod_{i=1}^{q} Q_i^{\gamma_i}. \tag{A.16}$$

We can rewrite the Feynman path integral as

$$F_{\mathrm{VQA}} = \left(\bigotimes_{i=1}^{p} \boldsymbol{w}_i^\intercal\right)^{\otimes r} \cdot \tilde{\boldsymbol{X}} \cdot \tilde{\boldsymbol{H}} \cdot \tilde{\boldsymbol{X}}^\dagger \cdot \left(\bigotimes_{i=p}^{1} \boldsymbol{w}_i\right)^{\otimes r} + 1, \tag{A.17}$$

where

$$\boldsymbol{w}_i \equiv \begin{pmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{pmatrix} \tag{A.18}$$

and $\tilde{\boldsymbol{X}} \in \mathbb{C}^{2^q \times 2^n}$ is a random matrix with rows

$$\left\langle \tilde{X} \right|_{\boldsymbol{\gamma}} \equiv \langle \psi_0 | C^\dagger Q_{\boldsymbol{\gamma}}^\dagger. \tag{A.19}$$

We will proceed as follows. First, we will bound the difference in the first two moments of Equation (A.17) and its equivalent, where the rows of $\tilde{X}$ are i.i.d. Haar random, to be exponentially small in n. As Haar random vectors have first three moments matching those of random Gaussian vectors (scaled by $2^{-\frac{n}{2}}$), this gives the desired convergence through second moments. Then, we will show that the characteristic functions at x of Equation (A.17) and its i.i.d. Haar random equivalent converge exponentially quickly in n for small enough x, giving a convergence in distribution at a rate $\tilde{\Omega}\left(\frac{\lg(A)f(n)n}{A}\right)$ by [180]. Finally, convergence in distribution to Equation (A.12) will follow as the error in the relevant higher-order moments between Haar random and scaled Gaussian vectors exponentially decays in n by a generalization of Borel's lemma [181].

Obviously the first moment of Equation (A.17) matches that of the i.i.d. Haar random case; off-diagonal entries in the path integral average to zero, and the diagonal entries are correct as C is drawn from a unitary 2-design [182]. Let us now consider the second moments of the nontrivial parts of both, where we are concerned with terms of the form:

$$c_{\alpha\beta\mu\nu} = \mathbb{E}\left[\langle\psi_0| C^\dagger Q^\dagger_{\boldsymbol{\gamma}_\alpha} H Q_{\boldsymbol{\gamma}_\beta} C |\psi_0\rangle \langle\psi_0| C^\dagger Q^\dagger_{\boldsymbol{\gamma}_\mu} H Q_{\boldsymbol{\gamma}_\nu} C |\psi_0\rangle\right], \tag{A.20}$$

and how they differ from the i.i.d. Haar random equivalent

$$h_{\alpha\beta\mu\nu} = \mathbb{E}\left[\langle\psi_0| U^\dagger_\alpha H U_\beta |\psi_0\rangle \langle\psi_0| U^\dagger_\mu H U_\nu |\psi_0\rangle\right]. \tag{A.21}$$

First, assume $\alpha = \beta = \mu = \nu$; as C is drawn from a unitary 2-design [182], the terms are equal. Similarly, if

$$\boldsymbol{\gamma}_\alpha \oplus \boldsymbol{\gamma}_\beta \oplus \boldsymbol{\gamma}_\mu \oplus \boldsymbol{\gamma}_\nu \neq 0, \tag{A.22}$$

then both expectations are equal to zero; this is because $c_{\alpha\beta\mu\nu}$ must have an odd number of some Q, and $h_{\alpha\beta\mu\nu}$ an odd number of some U (or $U^\dagger$).

Let us now consider when the above conditions are not satisfied. We consider

simultaneously terms of the form

$$\left(\langle\psi_0| C^\dagger Q_{\boldsymbol{\gamma}_\alpha}^\dagger R Q_{\boldsymbol{\gamma}_\beta} C |\psi_0\rangle + \gamma_{\alpha j} \leftrightarrow \gamma_{\beta j}\right)\left(\langle\psi_0| C^\dagger Q_{\boldsymbol{\gamma}_\mu}^\dagger R' Q_{\boldsymbol{\gamma}_\nu} C |\psi_0\rangle + \gamma_{\mu j} \leftrightarrow \gamma_{\nu j}\right), \tag{A.23}$$

i.e. all terms summed where unequal components of $\boldsymbol{\gamma}_\alpha$ and $\boldsymbol{\gamma}_\beta$ (and $\boldsymbol{\gamma}_\mu$ and $\boldsymbol{\gamma}_\nu$) are swapped. Note that the parity of the permutation determines the sign of the term in Equation (A.17) (and thus in Equation (A.23)). Here, R and R' are terms in the Pauli expansion of $\tilde{H}$. Consider the largest j where $\boldsymbol{\gamma}_\alpha$ and $\boldsymbol{\gamma}_\beta$ differ; consider the sum of each pair of terms in Equation (A.23) that have component j permuted, but are equal at all $k < j$. Each pair of terms is of the form (with relative signs made explicit)

$$\begin{aligned}\langle\psi_0| C^\dagger A Q_j A' R B' B C |\psi_0\rangle - \langle\psi_0| C^\dagger A A' R B' Q_j B C |\psi_0\rangle \\ = 2\langle\psi_0| C^\dagger A Q_j A' R B' B C |\psi_0\rangle \mathbf{1}_{[Q_j, A'RB'] \neq \mathbf{0}}\end{aligned} \tag{A.24}$$

for some A, A', B, B'. For all Q_j that commute with $A'RB'$, the two terms cancel. In particular, $Q_j A'RB'$ cannot be proportional to the identity. As $\tilde{H}$ is traceless, both R and R' are also not proportional to the identity. This can be done inductively for all j where $\boldsymbol{\gamma}_\alpha$ and $\boldsymbol{\gamma}_\beta$ differ.

Consider the case where $\boldsymbol{\gamma}_\alpha + \boldsymbol{\gamma}_\beta \neq \boldsymbol{\gamma}_\mu + \boldsymbol{\gamma}_\nu$; we must have that $\boldsymbol{\gamma}_\alpha$ and $\boldsymbol{\gamma}_\beta$ have a coordinate i where they are both one, and where $\boldsymbol{\gamma}_\mu$ and $\boldsymbol{\gamma}_\nu$ are both zero (assuming Equation (A.22) is not satisfied). By Equation (A.24), WLOG we can consider the product of Pauli observables between the two Q_i as being not proportional to the identity. Then, averaging over Q_i will yield zero. This is the same as the i.i.d. Haar random case, as every term in the expansion of Equation (A.23) must have only one of some unitary when $\boldsymbol{\gamma}_\alpha + \boldsymbol{\gamma}_\beta \neq \boldsymbol{\gamma}_\mu + \boldsymbol{\gamma}_\nu$.

Finally, consider the case where $\boldsymbol{\gamma}_\alpha + \boldsymbol{\gamma}_\beta = \boldsymbol{\gamma}_\mu + \boldsymbol{\gamma}_\nu$. Under this constraint, we must have the same number of terms in each sum in Equation (A.23); we call this number of terms 2^c. In the Pauli case, every time we combine terms as in Equation (A.24) introduces an overall factor of 4, and we average only over the anticommuting Pauli operators. As the value of the expectation over C is independent of the (nonidentity)

Pauli in the expectation value, this introduces a factor of $\frac{1}{2}$ every time we combine terms. This gives

$$2^c \mathbb{E}_{C\sim\mathbb{C}_n} \left[\langle\psi_0| C^\dagger S C |\psi_0\rangle \langle\psi_0| C^\dagger S' C |\psi_0\rangle \right], \tag{A.25}$$

for some S and S' that are equal if and only if $R = R'$. Similarly, in the i.i.d. Haar random case, only products of terms with $\boldsymbol{\gamma}_\alpha = \boldsymbol{\gamma}_\mu$ and $\boldsymbol{\gamma}_\beta = \boldsymbol{\gamma}_\nu$ are homogeneous in their unitaries and give nonzero expectations, yielding

$$2^c \mathbb{E}_{U\sim\mathbb{C}_n} \left[\langle\psi_0| U_\alpha^\dagger R U_\beta |\psi_0\rangle \langle\psi_0| U_\beta^\dagger R' U_\alpha |\psi_0\rangle \right]. \tag{A.26}$$

If $R \neq R'$ (and $S \neq S'$), these are both zero. If $R = R'$ (and $S = S'$), the latter is equal to 2^{c-n} and the former to $2^{c-n}\left(1 + \mathrm{O}\left(2^{-n}\right)\right)$. Putting everything together and explicitly writing the overall factor of $\frac{2^n}{\|H-\lambda_1\|_*}$, we have that the error in the second moment is on the order of

$$\epsilon_2 = \frac{2^{2n}}{\|H - \lambda_1\|_*^2} \left(2^{-n} \sqrt{\sum_{i=1}^{A} \alpha_i^2} \right)^2, \tag{A.27}$$

where α_i are the coefficients of the Pauli expansion of $H - \overline{\lambda}$. We also have that

$$\sum_{i=1}^{A} \alpha_i^2 = 2^{-n} \left\| H - \overline{\lambda} \right\|_{\mathrm{F}}^2 = m^{-1} 2^{-n} \|H - \lambda_1\|_*^2, \tag{A.28}$$

where m is defined as in Equation (A.6). Thus,

$$\epsilon_2 = 2^{-(n+\lg(m))}. \tag{A.29}$$

Let us now consider the tth moment for $t \geq 3$. We will bound the higher moments of both models, and show that their characteristic functions have infinite radii of convergence. Then, by showing that the difference in these characteristic functions vanishes exponentially in n for all $x \geq 0$ bounded below $\frac{\lg(A) f(n) n}{A}$, we will show that

the two models converge in distribution at a rate $\tilde{\Omega}\left(\frac{\lg(A)f(n)n}{A}\right)$.

By grouping terms as in Equation (A.23), it is sufficient to only bound

$$b_t = \left(\overline{\lambda} - \lambda_1\right)^{-t} \mathbb{E}_{C\sim\mathbb{C}_n}\left[\prod_{i=1}^{t}\left(\sum_{j=1}^{A}\alpha_j \left\langle\psi_0\right| C^\dagger S_{ij} C \left|\psi_0\right\rangle\right)\right], \tag{A.30}$$

where S_{ij} is not proportional to the identity, $S_{ij} \neq S_{i'j}$ for all $i \neq i'$, and A is the number of terms in the Pauli decomposition of $\tilde{H}$. If a term in the expansion of Equation (A.30) contains two S_{ij} that anticommute, the contribution to the moment from that term is zero as $C\left|\psi_0\right\rangle$ is a stabilizer state for all C. Generally, the contribution to the moment is maximized when the S_{ij} are "maximally dependent"—that is, for d distinct S_{ij} in a term, the contribution to the moment is maximized when the S_{ij} are generated by a cardinality $\lfloor \lg(d) + 1 \rfloor$ subset of them. Thus, the contribution to the moment is bounded by $2^{-c\lfloor \lg(d)+1\rfloor n}$ for some constant c [167]. Note that this also bounds the i.i.d. Haar random case. Putting everything together and using the multinomial theorem, the tth moment of the nontrivial part of both distributions is bounded by

$$b_t \leq \sum_{\sum_i k_i = t} 2^{-c\left\lfloor \lg\left(\|\boldsymbol{k}\|_0\right)+1\right\rfloor n}\binom{t}{k_1, \ldots, k_A}\prod_{i=1}^{A}\left(\frac{\alpha_i}{\overline{\lambda} - \lambda_1}\right)^{k_i}. \tag{A.31}$$

This corresponds to the case where $S_{ij} = S_{ij'} \equiv S_i$, i.e. when there is maximal dependence between the matrix elements. Here, k_i indexes how many times S_i appears in a term in Equation (A.30), and $\|\cdot\|_0$ denotes the number of nonzero coordinates of $\cdot$. By Equation (A.13), as $t \to \infty$ for any given A and n,

$$\frac{b_t}{t!} \leq \left(1 + \mathrm{o}\left(1\right)\right) 2^{-t\lg(t) - \frac{1}{2}\lg(2\pi t) + t\lg\left(\frac{eA}{f(n)}\right) - c\lg(A)n}. \tag{A.32}$$

Thus, the Taylor series of the characteristic functions of both distributions have infinite radii of convergence, and both are completely determined by their moments. Furthermore, Equation (A.32) gives us that the difference in their characteristic functions at $0 \leq x < \frac{c\lg(A)f(n)n}{A}$ is on the order of $\exp\left(\frac{Ax}{f(n)} - c\lg\left(A\right)n\right)$ as $n \to \infty$. As

the two distributions have equal moments for $t \leq 2$, it can then be shown [180] that the error in Lévy distance between the two is $\tilde{\mathrm{O}}\left(\left(\frac{\lg(A)f(n)n}{A}\right)^{-1}\right)$. □

Now that we have shown the weak convergence of our random class of VQAs to a random field on the hypertorus, we can combine this result with a multidimensional generalization of the Welch–Satterthwaite equation [183, 184] to show that our distribution of VQAs has first two moments matching that of a Wishart hypertoroidal random field (WHRF). Once again, under further assumptions on the spectrum of H we will also bound the higher moments of the two distributions to show convergence in distribution.

Theorem A.2 (XHX RFs as WHRFs)**.** *The random field given by Equation* (A.12) *has first two moments equal to the Wishart hypertoroidal random field (WHRF)*

$$F_{WHRF}(\boldsymbol{\theta}) = m^{-1} \sum_{i_1,\ldots,i_r,i'_1,\ldots,i'_r=1}^{2^p} w_{i_1} \ldots w_{i_r} J_{i_1,\ldots,i_r,i'_1,\ldots,i'_r} w_{i'_1} \ldots w_{i'_r}, \tag{A.33}$$

where $\boldsymbol{J} \sim \mathcal{CW}_{2^q}(m, \boldsymbol{I}_{\boldsymbol{2^q}})$ *is a complex Wishart random matrix and the* effective degrees of freedom *defined in Equation* (A.6) *is formally a real number, but can be rounded to the nearest natural number with negligible error. Furthermore, assuming the largest eigenvalue of* $\tilde{H}$ *as defined in Equation* (A.10) *is at most* 2^{cn} *for some constant* c *bounded below* 1*, their distributions are bounded in Lévy distance by* $2^{-\Omega(\min(n,\lg(m)))}$.

Proof. By the unitary invariance of random matrices with Gaussian entries, by diagonalizing $\tilde{H}$ we can rewrite F_{XHX} as the random field

$$F_{\mathrm{XHX}}(\boldsymbol{w}) = \|H - \lambda_1\|_*^{-1} \left(\bigotimes_{i=1}^{p} \boldsymbol{w}_i\right)^{\otimes r} \cdot \sum_{i=1}^{2^n} \left(\left(h_i - \overline{\lambda}\right) \boldsymbol{X}_i \otimes (\boldsymbol{X}_i)^\dagger + \overline{\lambda} - \lambda_1\right) \cdot \left(\bigotimes_{i=p}^{1} \boldsymbol{w}_i\right)^{\otimes r}, \tag{A.34}$$

where $\boldsymbol{X}_i$ is the ith column of $\boldsymbol{X}$ and h_i are the eigenvalues of H. The sum over Kronecker products of columns is just the weighted sum of (at most) 2^n independent Wishart random variables, each with a single degree of freedom. It is known [185–187] that the first two moments of this weighted sum of independent Wishart random

variables is equal (up to rounding of the degrees of freedom) to that of the single Wishart random variable

$$\boldsymbol{J} \sim \mathcal{CW}_{2^q}\left(m, m^{-1}\boldsymbol{I}_{\boldsymbol{2^q}}\right), \tag{A.35}$$

where m is defined as in Equation (A.6).

Let us now consider higher moments of both distributions. A useful property of both F_{XHX} and F_{WHRF} is that they are invariant under rotations on the hypertorus $\boldsymbol{w} \mapsto \boldsymbol{O} \cdot \boldsymbol{w}$ (for real orthogonal $\boldsymbol{O} \in \mathrm{SO}\left(2\right)^{\otimes p}$) due to the invariance of the Wishart distribution under orthogonal transformations [188]. Due to this property, we will often take

$$\boldsymbol{w} = \boldsymbol{n} \equiv (1, 0, \ldots, 0)^{\intercal}, \tag{A.36}$$

i.e. perform calculations at a fixed point $\boldsymbol{\theta} = \boldsymbol{0}$ on the hypertorus. For instance, by inspection of the marginal distributions of the elements of $\boldsymbol{X} \otimes \boldsymbol{X}^{\dagger}$ and $\boldsymbol{J}$ [189, 190], we immediately see that

$$\left(\bigotimes_{i=1}^{p} \boldsymbol{w}_i\right)^{\otimes r} \cdot \left(\boldsymbol{X} \otimes \boldsymbol{X}^{\dagger}\right) \cdot \left(\bigotimes_{i=p}^{1} \boldsymbol{w}_i\right)^{\otimes r} \sim \Gamma\left(1, 1\right) \tag{A.37}$$

and

$$F_{\mathrm{WHRF}}\left(\boldsymbol{w}\right) \sim m^{-1} J_{(1,\ldots,1),(1,\ldots,1)} \sim \Gamma\left(m, m^{-1}\right); \tag{A.38}$$

here, $\Gamma\left(k, \theta\right)$ is a gamma distributed random variable with shape k and scale θ. We therefore have that the moment-generating function for $F_{\mathrm{XHX}}\left(\boldsymbol{w}\right)$ is

$$M_{\mathrm{XHX}}\left(x\right) = \mathrm{e}^{x} \prod_{i=1}^{2^n} \left(1 - \frac{h_i - \overline{\lambda}}{\left\|H - \lambda_1\right\|_*} x\right)^{-1} = \mathrm{e}^{x} \det\left(1 - 2^{-n}\tilde{H}x\right)^{-1} \tag{A.39}$$

and for $F_{\mathrm{WHRF}}\left(\boldsymbol{w}\right)$ is

$$M_{\mathrm{WHRF}}\left(x\right) = \left(1 - \frac{x}{m}\right)^{-m}. \tag{A.40}$$

Assuming the largest eigenvalue of $\tilde{H}$ is at most 2^{cn}, we see that these moment generating functions differ at any given $0 \leq x < 2^{\min((1-c)n, \lg(m))}$ by at most $\mathrm{O}\left(2^{-3(1-c)n}x^3 + m^{-3}x^3\right)$. As the two distributions have equal first and second moments, it can then be shown [180]

that the error in Lévy distance between the two is bounded by $2^{-\Omega(\min(n,\lg(m)))}$. $\square$

Combining the two theorems, we roughly see that under reasonable assumptions on the spectrum of H the random fields induced by the specific class of VQAs we consider can be approximated by WHRFs up to an error on the order of $\tilde{\mathrm{O}}\left(\left(\frac{\lg(A)f(n)n}{A}\right)^{-1} + m^{-1}\right)$ as $m, n \to \infty$.

A.1.3 Discussion of the Mapping

Let us now briefly discuss the intuition and assumptions behind the results proved in Appendix A.1.2, beginning with the random class of ansatzes we consider. Of course, in practice, VQA ansatzes are not chosen at random. Indeed, VQA ansatzes have a layered structure that precludes any independence between layers even if the layers were randomly chosen. Though this randomness assumption is strong, heuristically deep enough circuits (that are independent of the problem Hamiltonian) will still look roughly uniform over stabilizer states in the Feynman path integral expansion performed in the proof of Theorem A.1, giving qualitatively similar results. Importantly, even with this assumption, we are able to demonstrate regimes where there exist no barren plateaus yet the model still has poor local minima; see Section 2.3.3, where we discuss how barren plateaus can be demonstrated in our framework. Furthermore, though throughout this paper we consider results in expectation over this distribution of ansatzes, we find numerically in Section 2.4.1 that our analytic results seem to also hold in distribution; we therefore suspect that our analytic results in Appendix A.3 hold more generally for individual ansatzes that are independent of the problem Hamiltonian.

Given the randomized class of ansatzes, in Theorem A.1 we show that the VQA loss function is close in distribution to that of the random field given in Equation (A.12) (the "XHX" model). Intuitively, this just stems from the fact that different paths in the Feynman path integral are matrix elements in uniformly random stabilizer states. We then show that the error induced in higher moments by taking each of these paths to be independent vanishes as $n \to \infty$. To prove this formally, we rely

on the boundedness of Equation (A.13) to bound higher moments of the distribution. Luckily, in practice this bound holds; for extensive Hamiltonians, one expects $f(n) \gtrsim n$.

Theorem A.2 extends Theorem A.1 by showing that the XHX model can be written as a sum of Wishart models weighted by the (scaled and shifted) eigenvalues of H, which can then be approximated by a single Wishart model. Heuristically, one can think of complex Wishart matrices as multidimensional generalizations of the gamma distribution; then, the approximation used in Theorem A.2 is just a multidimensional generalization of the Welch–Satterthwaite approximation [183, 184]. This approximation (in both the univariate and multivariate cases) is exact in the first two moments of the distribution when the *effective degrees of freedom* m given in Equation (A.6) is allowed to be real. In practice, m is rounded to the nearest natural number, inducing a slight error in the approximation. Generally, errors in higher moments in the Welch–Satterthwaite approximation may be large when the moments of the approximated distribution is large, particularly when the coefficients of the sum can have arbitrary sign and are at different scales [183, 191]. However, for physically relevant Hamiltonians, the spectral radius is much smaller than 2^n, and the coefficients of the sum are approximately equal. We show that under such conditions, errors in the moment generating functions vanish as $m, n \to \infty$ at the given rate.

The effective degrees of freedom m as in Equation (A.6) can be interpreted as roughly a signal-to-noise ratio of the mean eigenvalue of $H - \lambda_1$, and generically is at least exponentially large in n (for small eigenvalue spacings, as is typically found in physical Hamiltonians studied with VQAs [19, 61, 99]). We show in Section 2.3.3 that m sharply dictates the variational loss landscape; for a number of independent parameters $p \geq 2m$, local minima concentrate near the global minimum. Conversely, for p bounded below $2m$, local minima concentrate far away from the global minimum. This would imply that for the class of randomized ansatz we consider here, training large instances is infeasible. However, consider an ansatz that is allowed to depend on the problem instance H, such as in the Hamiltonian variational ansatz (HVA) [178]. With a clever enough ansatz, one can in principle "reweigh" the coeffi-

cients of Equation (A.34) by having a nonuniform distribution over stabilizer states in the Feynman path integral expansion of Equation (A.17), effectively making m smaller. This would be consistent with what was numerically investigated in prior work [70] (and in Section 2.4.2), where it was shown that even for a modest number of parameters the distribution of local minima concentrate near the global minimum for the HVA. We leave further investigation in this direction for future work.

Finally, we note that all of our asymptotic equivalence results so far have been shown to converge at a rate

$$\rho \equiv \lg(A) f(n) n/A, \tag{A.41}$$

which is typically $\gtrsim \lg(n)$ for physically relevant (i.e. two-local with arbitrary range, molecular in the plane wave dual basis [192], etc.) Hamiltonians. In Appendix A.4, we give the loss landscape of WHRFs (Equation (A.33)) as $p, m \to \infty$, taking into account large deviations in p. If p grows as $\Omega(\lg(\rho))$, then in principle uncontrolled large deviations in the convergence of VQAs to WHRFs will dominate the asymptotics of the landscape (Equation (2.17)). In particular, with probability $\sim \rho^{-1}$, deviations of the eigenvalues of the Hessian on the order of the eigenvalues themselves can occur, which are then "blown up" by a factor exponentially large in p if all deviations constructively interfere. Thus, though Equation (2.17) holds for WHRFs, it does not necessarily hold for VQAs when $p = \Omega(\lg(\rho))$. If the deviations of eigenvalues of the Hessian due to the mapping from VQAs to WHRFs are roughly independent between eigenvalues, however, then these deviations are further exponentially suppressed in p, and the result holds independently of how p scales with n. We believe in practice this is what occurs, and see numerically in Section 2.4.1 that our analytic results hold well even when $p \gg n$.

A.2 The Kac–Rice Formula and its Assumptions

For completeness, we state the formal version of Lemma 2.2—with all assumptions—here. We borrow heavily from [100]. By $\boldsymbol{\nabla} f$, we mean the covariant gradient of f.

Lemma A.3 (Kac–Rice formula [100])**.** *Let M be a compact, oriented, N-dimensional C^1 manifold with C^1 Riemannian metric g. Let $B \subset \mathbb{R}^K$ be an open set such that ∂B has dimension $K-1$. Let $f : M \to \mathbb{R}^K$ be a random field on M, and let $\iota(\cdot)$ denote the index of $\cdot$. Furthermore, assume that:*

1. *All components of f, $\boldsymbol{\nabla} f$, and $\boldsymbol{\nabla}^2 f$ are almost surely continuous and have finite variances over M.*

2. *The marginal density $p_t(\boldsymbol{\nabla} f(t))$ of $\boldsymbol{\nabla} f$ at $t \in M$ is continuous at $\boldsymbol{\nabla} f = \mathbf{0}$.*

3. *The conditional densities $p_t\left(\boldsymbol{\nabla} f(t) \mid f(t), \boldsymbol{\nabla}^2 f(t)\right)$ are bounded above and continuous at $\boldsymbol{\nabla} f = \mathbf{0}$, uniformly in $t \in M$.*

4. *The conditional densities $p_t\left(\det\left(\boldsymbol{\nabla}^2 f(t)\right) \mid \boldsymbol{\nabla} f(t) = \mathbf{0}\right)$ are continuous in the neighborhood of $\det\left(\boldsymbol{\nabla}^2 f\right) = 0$ and $\boldsymbol{\nabla} f(t) = \mathbf{0}$, uniformly in $t \in M$.*

5. *The conditional densities $p_t(f(t) \mid \boldsymbol{\nabla} f(t) = \mathbf{0})$ are continuous for all f and for all $\boldsymbol{\nabla} f$ in a neighborhood of $\mathbf{0}$, uniformly in $t \in M$.*

6. *The Hessian moments are bounded, i.e.*

$$\sup_{t \in M} \max_{i,j} \mathbb{E}\left[\left|\left(\boldsymbol{\nabla}^2 f(t)\right)_{i,j}\right|^N\right] < \infty. \tag{A.42}$$

7. *The moduli of continuity with respect to (the canonical metric induced by) g of each component of f, $\boldsymbol{\nabla} f$, and $\boldsymbol{\nabla}^2 f$ all satisfy*

$$\mathbb{P}[\omega(\eta) > \epsilon] = \mathrm{o}\left(\eta^N\right) \tag{A.43}$$

for all $\epsilon > 0$ as $\eta \to 0^+$.

Then,

$$\mathbb{E}\left[\mathrm{Crt}_k^f(B)\right] = \int_M p_{\boldsymbol{\sigma}}\left(\boldsymbol{\nabla} f(\boldsymbol{\sigma}) = 0\right) \times \mathbb{E}\left[\left|\det\left(\boldsymbol{\nabla}^2 f(\boldsymbol{\sigma})\right)\right| \mathbf{1}\left\{f(\boldsymbol{\sigma}) \in B\right\} \mathbf{1}\left\{\iota\left(\boldsymbol{\nabla}^2 f(\boldsymbol{\sigma})\right) \leq k\right\} \mid \boldsymbol{\nabla} f(\boldsymbol{\sigma}) = 0\right] \mathrm{d}\boldsymbol{\sigma}\,, \tag{A.44}$$

where $\mathrm{d}\boldsymbol{\sigma}$ *is the volume element induced by* g *on* M*.*

It is obvious by Lemma A.4 that conditions 2-6 are satisfied by WHRFs given $B = (0, u)$. Furthermore, as F is a polynomial in $\{\cos(\theta_i), \sin(\theta_i)\}$, F and its derivatives are continuous for any value of the components of $m^{-1}\boldsymbol{J}$, and all have finite variance. Similarly, it is easy to see that the modulus of continuity of f and its gradients go as $J\eta^r$ as $\eta \to 0^+$, where J is the largest component of $m^{-1}\boldsymbol{J}$. As the distributions of the components of a Wishart matrix have exponential tails, the probability that $J = \Omega(\eta^{-r})$ is indeed $\mathrm{o}\left(\eta^N\right)$ and therefore all conditions are satisfied by WHRFs.

A.3 The Loss Landscape of Wishart Hypertoroidal Random Fields

A.3.1 The Joint Distribution of $F_{\mathbf{WHRF}}$ and its Derivatives

In order to utilize the Kac–Rice formula (Lemma A.3), we must calculate the joint distribution of the random field

$$F_{\mathrm{WHRF}}(\boldsymbol{\theta}) = m^{-1} \sum_{i_1,\ldots,i_r,i'_1,\ldots,i'_r=1}^{2^p} w_{i_1} \ldots w_{i_r} J_{i_1,\ldots,i_r,i'_1,\ldots,i'_r} w_{i'_1} \ldots w_{i'_r}, \tag{A.45}$$

with its derivatives. In the course of proving Theorem A.2, we already have shown that the function value is gamma distributed (see Equation (A.38)). Here, we explicitly calculate the distribution of the Hessian when given the function value and that the covariant gradient is zero, and also calculate the distribution of the gradient given

the function value. We will heavily lean on the rotational invariance property of the distribution discussed in the proof of Theorem A.2, with $\boldsymbol{n}$ once again the fixed point with all $\theta_i = 0$. Note that for the given embedding of the hypertorus into $\mathbb{R}^{2p}$, the Christoffel symbols are zero (i.e. we are considering the Euclidean hypertorus) and thus for the most part we can ignore the distinction between covariant and normal derivatives. Here, we choose local coordinates $\boldsymbol{\theta}$ such that:

$$\boldsymbol{w}_i = \begin{pmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{pmatrix}. \tag{A.46}$$

Perhaps surprisingly, we will find that conditioned on being at a critical point at a specified energy, the Hessian takes the simple form of a normalized and shifted Wishart matrix summed with a normalized GOE matrix. The gradient conditioned on the function value is similarly simple, given by independent Gaussian variables.

Lemma A.4 (Hessian and gradient distributions)**.** *The scaled Hessian $m\partial_i\partial_j F_{WHRF}(\boldsymbol{w})$ conditioned on $F_{WHRF}(\boldsymbol{w}) = x$ and $\partial_k F_{WHRF}(\boldsymbol{w}) = 0$ is distributed as*

$$m\tilde{C}_{ij}(x) = -2rmx\delta_{ij} + rW_{ij} + r\sqrt{2mx}N_{ij}, \tag{A.47}$$

where $\boldsymbol{W} \sim \mathcal{W}_p(2m, \boldsymbol{I_p})$ and $\boldsymbol{N} \sim GOE_p$ are independent. Furthermore, the scaled gradient $m\partial_k F_{WHRF}(\boldsymbol{w})$ conditioned on $F_{WHRF}(\boldsymbol{w}) = x$ is distributed as

$$m\tilde{G}_k(x) = \sqrt{2mrx}N_k, \tag{A.48}$$

where N_k are i.i.d. standard normally distributed random variables independent from all W_{ij} and N_{ij}.

Proof. Without loss of generality we take $\boldsymbol{w} = \boldsymbol{n}$. Let $\boldsymbol{i} \in \{1, 2\}^{\times p}$ be the vector with the ith component equal to 2 and all others equal to 1, $(\boldsymbol{i}, \boldsymbol{j})$ similar with both the ith and jth component, and $\boldsymbol{b}$ the vector with all components equal to 1. Taking

derivatives explicitly yields

$$m\partial_i F_{\mathrm{WHRF}}(\boldsymbol{n}) = 2\,\mathrm{Re}\left\{J_{(\boldsymbol{i},\boldsymbol{b},\ldots,\boldsymbol{b}),(\boldsymbol{b},\ldots,\boldsymbol{b})}\right\} + \ldots + 2\,\mathrm{Re}\left\{J_{(\boldsymbol{b},\ldots,\boldsymbol{i}),(\boldsymbol{b},\ldots,\boldsymbol{b})}\right\} \tag{A.49}$$

and

$$\begin{aligned} m\partial_i\partial_j F_{\mathrm{WHRF}}(\boldsymbol{n}) &= -2r\delta_{ij}J_{(\boldsymbol{b},\ldots,\boldsymbol{b}),(\boldsymbol{b},\ldots,\boldsymbol{b})} \\ &+ 2\,\mathrm{Re}\left\{J_{(\boldsymbol{i},\boldsymbol{b},\ldots,\boldsymbol{b}),(\boldsymbol{j},\boldsymbol{b},\ldots,\boldsymbol{b})}\right\} + 2\,\mathrm{Re}\left\{J_{(\boldsymbol{i},\boldsymbol{b},\ldots,\boldsymbol{b}),(\boldsymbol{b},\boldsymbol{j},\ldots,\boldsymbol{b})}\right\} + \ldots + 2\,\mathrm{Re}\left\{J_{(\boldsymbol{b},\ldots,\boldsymbol{b},\boldsymbol{i}),(\boldsymbol{b},\ldots,\boldsymbol{b},\boldsymbol{j})}\right\} \\ &+ 2\,\mathrm{Re}\left\{J_{((\boldsymbol{i},\boldsymbol{j}),\boldsymbol{b},\ldots,\boldsymbol{b}),(\boldsymbol{b},\ldots,\boldsymbol{b})}\right\} + 2\,\mathrm{Re}\left\{J_{(\boldsymbol{i},\boldsymbol{j},\ldots,\boldsymbol{b}),(\boldsymbol{b},\ldots,\boldsymbol{b})}\right\} + \ldots + 2\,\mathrm{Re}\left\{J_{(\boldsymbol{b},\ldots,\boldsymbol{b},(\boldsymbol{i},\boldsymbol{j})),(\boldsymbol{b},\ldots,\boldsymbol{b})}\right\}. \end{aligned} \tag{A.50}$$

As $\boldsymbol{J}$ is a Wishart matrix with identity scale matrix, it can be written as $\boldsymbol{X}\cdot\boldsymbol{X}^\dagger$ for $\boldsymbol{X}$ a $2^q \times m$ matrix with i.i.d. standard complex normal entries. By performing an LQ decomposition of $\boldsymbol{X}$, one can then by inspection determine the distributions of the entries of $\boldsymbol{J}$ [189, 190]. For ease of notation, we let $\tau : \{1,2\}^{\times q} \to \{1,\ldots,2^q\}$ be a mapping between representations of the indices of J, with the convention $\tau((\boldsymbol{b},\ldots,\boldsymbol{b})) = 1$. We then find (taking $\tau((\boldsymbol{i},\ldots,2)) < \tau((\boldsymbol{j},\ldots,2))$ WLOG) that

$$2J_{(\boldsymbol{b},\ldots,\boldsymbol{b}),(\boldsymbol{b},\ldots,\boldsymbol{b})} = 2mF_{\mathrm{WHRF}}(\boldsymbol{n}), \tag{A.51}$$

$$2\,\mathrm{Re}\left\{J_{(\boldsymbol{i},\ldots,\boldsymbol{b}),(\boldsymbol{b},\ldots,\boldsymbol{b})}\right\} = \sqrt{2mF_{\mathrm{WHRF}}(\boldsymbol{n})}M_{(\boldsymbol{b},\ldots,\boldsymbol{b}),(\boldsymbol{i},\ldots,\boldsymbol{b})}, \tag{A.52}$$

$$2\,\mathrm{Re}\left\{J_{(\boldsymbol{i},\boldsymbol{j},\ldots,\boldsymbol{b}),(\boldsymbol{b},\ldots,\boldsymbol{b})}\right\} = \sqrt{2mF_{\mathrm{WHRF}}(\boldsymbol{n})}M_{(\boldsymbol{b},\ldots,\boldsymbol{b}),(\boldsymbol{i},\boldsymbol{j},\ldots,\boldsymbol{b})}; \tag{A.53}$$

and, for $\tau((\boldsymbol{i},\ldots,\boldsymbol{b})) \leq m$,

$$\begin{aligned} 2\,\mathrm{Re}\left\{J_{(\boldsymbol{i},\ldots,\boldsymbol{b}),(\boldsymbol{j},\ldots,\boldsymbol{b})}\right\} &= \sqrt{2\Gamma_{(\boldsymbol{i},\ldots,\boldsymbol{b})}}M_{(\boldsymbol{i},\ldots,\boldsymbol{b}),(\boldsymbol{j},\ldots,\boldsymbol{b})} \\ &+ \sum_{\mu=1}^{\tau((\boldsymbol{i},\ldots,\boldsymbol{b}))-1} M_{\tau^{-1}(\mu),(\boldsymbol{i},\ldots,\boldsymbol{b})}M_{\tau^{-1}(\mu),(\boldsymbol{j},\ldots,\boldsymbol{b})} + \sum_{\mu=1}^{\tau((\boldsymbol{i},\ldots,\boldsymbol{b}))-1} \tilde{M}_{\tau^{-1}(\mu),(\boldsymbol{i},\ldots,\boldsymbol{b})}\tilde{M}_{\tau^{-1}(\mu),(\boldsymbol{j},\ldots,\boldsymbol{b})} \end{aligned} \tag{A.54}$$

and otherwise

$$2\,\mathrm{Re}\left\{J_{(\boldsymbol{i},\ldots,\boldsymbol{b}),(\boldsymbol{j},\ldots,\boldsymbol{b})}\right\} = \sum_{\mu=1}^{m} M_{\tau^{-1}(\mu),(\boldsymbol{i},\ldots,\boldsymbol{b})} M_{\tau^{-1}(\mu),(\boldsymbol{j},\ldots,\boldsymbol{b})} + \sum_{\mu=1}^{m} \tilde{M}_{\tau^{-1}(\mu),(\boldsymbol{i},\ldots,\boldsymbol{b})} \tilde{M}_{\tau^{-1}(\mu),(\boldsymbol{j},\ldots,\boldsymbol{b})}. \tag{A.55}$$

Here, $\boldsymbol{M}$ and $\tilde{\boldsymbol{M}}$ are symmetric with off-diagonal entries i.i.d. drawn from the standard normal distribution, and $\Gamma_{\tau^{-1}(\mu)} \equiv M^2_{\tau^{-1}(\mu),\tau^{-1}(\mu)}$ has entries i.i.d. drawn from $\Gamma\left(m-\mu+1,1\right)$. Note that each $\sqrt{2\Gamma}$ is chi-square distributed with $2\left(m-\mu+1\right)$ degrees of freedom; therefore, Equation (A.54) and Equation (A.55) can be considered as elements of a real Wishart matrix $\tilde{\boldsymbol{W}}$ with $2m$ degrees of freedom. Also, note that Equation (A.54) and Equation (A.55) are independent of $\partial_k F_{\mathrm{WHRF}}\left(\boldsymbol{n}\right)$ when conditioned on $F_{\mathrm{WHRF}}\left(\boldsymbol{n}\right) \equiv x = 0$. If $x \neq 0$, the condition $\partial_k F_{\mathrm{WHRF}}\left(\boldsymbol{n}\right) = 0$ is equivalent to taking each sum over the elements of $\boldsymbol{M}$ from $\mu = 2$ instead of $\mu = 1$, which is equivalent to taking the convention $\tau\left((\boldsymbol{b},\ldots,\boldsymbol{b})\right) = 2^q$ and shifting the indices of $\boldsymbol{M}$ and $\tilde{\boldsymbol{M}}$. Therefore, the (scaled) Hessian conditioned on $F_{\mathrm{WHRF}}\left(\boldsymbol{n}\right) = x$ and $\partial_k F_{\mathrm{WHRF}}\left(\boldsymbol{n}\right) = 0$ is distributed as

$$m\tilde{C}_{ij}\left(x\right) = -2rmx\delta_{ij} + \left(\boldsymbol{O}\cdot\tilde{\boldsymbol{W}}\cdot\boldsymbol{O}^{\intercal}\right)_{ij} + r\sqrt{2mx}N_{ij}; \tag{A.56}$$

here, $\boldsymbol{N} \sim GOE_p$ (with the convention that diagonal entries $\sim \mathcal{N}\left(0,2\right)$ and off-diagonal entries $\sim \mathcal{N}\left(0,1\right)$), and $\boldsymbol{O}$ is a matrix such that $O_{i\mu} = 1$ if and only if $\tau^{-1}\left(\mu\right)$ is of the form $(\boldsymbol{b},\ldots,\boldsymbol{i},\ldots,\boldsymbol{b})$, and is otherwise equal to 0. The invariance of the Wishart distribution under orthogonal transformations and partitioning [188, 189] leads to the final result. □

A.3.2 The Exact Distribution of Critical Points

Given the joint distribution of F_{WHRF}, its gradient, and its Hessian, we are now equipped to calculate the expected number of critical points of a given index k using the Kac–Rice formula (Lemma A.3).

Theorem A.5 (Distribution of critical points in WHRFs)**.** *Let*

$$\mu_{\boldsymbol{C}(x)} = \frac{1}{p}\sum_{i=1}^{p}\delta\left(\lambda_i^{\boldsymbol{C}(x)}\right) \tag{A.57}$$

be the empirical spectral measure of the random matrix

$$\boldsymbol{C}\left(x\right) = \frac{r}{m}\left(\boldsymbol{W} + \sqrt{2mx}\boldsymbol{N}\right), \tag{A.58}$$

where $\boldsymbol{W} \sim \mathcal{W}_p\left(2m, \boldsymbol{I_p}\right)$ *and* $\boldsymbol{N} \sim GOE_p$ *are independent and* $\lambda_i^{\boldsymbol{C}}\left(x\right)$ *is the* i*th smallest eigenvalue of* $\boldsymbol{C}\left(x\right)$*. Then, the distribution of the expected number of critical points of index* k *at an energy* $E > 0$ *of* F_{WHRF} *is given by*

$$\begin{aligned}&\mathbb{E}\left[\mathrm{Crt}_k\left(E\right)\right]\\ &= \left(\frac{\pi}{r}\right)^{\frac{p}{2}}\Gamma\left(m\right)^{-1}m^{(1+\gamma)m}\mathbb{E}_{\boldsymbol{C}(E)}\left[\mathrm{e}^{p\int\ln(|\lambda-2rE|)\mathrm{d}\mu_{\boldsymbol{C}(E)}}\mathbf{1}\left\{\lambda_{k+1}^{\boldsymbol{C}(E)} \geq 2rE\right\}\right]E^{(1-\gamma)m-1}\mathrm{e}^{-mE},\end{aligned} \tag{A.59}$$

where

$$\gamma = \frac{p}{2m}. \tag{A.60}$$

Proof. As discussed in Appendix A.2, the assumptions of the Kac–Rice formula (i.e. Lemma A.3) are satisfied. Furthermore, due to the invariance of the Wishart distribution with respect to rotations on the hypertorus [188, 189], we can integrate out the volume element independently; the volume of $\left(S^1\right)^{\times p}$ is

$$\int_{\left(S^1\right)^{\times p}}\mathrm{d}\boldsymbol{w} = \left(2\pi\right)^p. \tag{A.61}$$

Additionally, we have from Lemma A.4 that the probability density of the gradient vector being zero at any $\boldsymbol{w}$ conditioned on $H_{\mathrm{WHRF}}\left(\boldsymbol{w}\right) = x$ is

$$p_{\boldsymbol{w}}\left(\boldsymbol{\nabla}H_{\mathrm{WHRF}}\left(\boldsymbol{w}\right) = 0 \mid H_{\mathrm{WHRF}}\left(\boldsymbol{w}\right) = x\right) = \left(\frac{4\pi rx}{m}\right)^{-\frac{p}{2}}. \tag{A.62}$$

Taking the expectation over x via Equation (A.38) and using the Hessian distrribution from Lemma A.4, we have from Lemma 2.2 that

$$\begin{aligned}\mathbb{E}\left[\mathrm{Crt}_k\left(B=(0,E)\right)\right] &= \left(\frac{\pi}{r}\right)^{\frac{p}{2}}\Gamma\left(m\right)^{-1}m^{(1+\gamma)m} \\ &\times \int_0^E \mathbb{E}_{\boldsymbol{C}(x)}\left[\mathrm{e}^{p\int\ln(|\lambda-2rx|)\mathrm{d}\mu_{\boldsymbol{C}}(x)}\mathbf{1}\left\{\lambda_{k+1}^{\boldsymbol{C}(x)}\geq 2rx\right\}\right]x^{(1-\gamma)m-1}\mathrm{e}^{-mx}\,\mathrm{d}x\,.\end{aligned} \tag{A.63}$$

Taking the derivative of this cumulative distribution with respect to E yields the final result. □

A.4 Logarithmic Asymptotics via Free Probability Theory

Though Equation (A.59) is exact, it is difficult to use in practice. Luckily, we are able to use a surprising fact about the eigenvalue distributions of Wishart and GOE matrices; asymptotically, the empirical spectral distributions of these matrices weakly converge to fixed distributions. Concretely, in the limit $p\to\infty$ where $\gamma=\frac{p}{2m}$ is held constant, the eigenvalue distribution of $\boldsymbol{W}/2m$ where $\boldsymbol{W}\sim\mathcal{W}_p\left(2m,\boldsymbol{I_p}\right)$ weakly converges to the *Marchenko–Pastur distribution* [193]:

$$\mathrm{d}\mu_{\mathrm{M.P.}}=\left(1-\gamma^{-1}\right)\mathbf{1}\left\{\gamma>1\right\}\delta\left(\lambda\right)\mathrm{d}\lambda+\frac{1}{2\pi\gamma\lambda}\sqrt{\left((1+\sqrt{\gamma})^2-\lambda\right)\left(\lambda-(1-\sqrt{\gamma})^2\right)}\,\mathrm{d}\lambda\,. \tag{A.64}$$

Similarly, the eigenvalue distribution of $\boldsymbol{N}/\sqrt{p}$ where $\boldsymbol{N}\sim\mathrm{GOE}_p$ weakly converges to the *Wigner semicircle distribution* [194]:

$$\mathrm{d}\mu_{\mathrm{s.c.}}=\frac{1}{2\pi}\sqrt{4-\lambda^2}\,\mathrm{d}\lambda\,. \tag{A.65}$$

Furthermore, by using free probability theory one can find the asymptotic distribution of eigenvalues for a weighted sum of these matrices, given their eigenbases are in "generic position" with respect to each other. We now give a brief review of free

probability theory—at least in the context of random matrix theory—here. Later, we will also briefly review large deviations theory, which we use to bound the probability of large deviations from the weak convergence of the eigenvalue distributions of Wishart and GOE matrices. Note that, as we are unable to control large deviations in Theorem A.1, in principle large deviations in the weak convergence of VQAs to WHRFs could dominate the large deviations in WHRFs; however, as discussed in Appendix A.1.3, this provably does not occur at shallow enough depths with respect to n, and there are reasons to believe it does not occur even at large depths (which we additionally give numerical evidence for in Section 2.4).

We begin by reviewing the techniques in free probability theory and large deviations theory that we use in studying the asymptotic behavior of Equation (A.59).

A.4.1 Free Probability Theory

Free probability theory is the study of noncommutative random variables. Specializing to random matrix theory on $N \times N$ matrices, we define the unital linear functional

$$\phi(X) \equiv \frac{1}{N}\mathbb{E}[\mathrm{tr}(X)] \tag{A.66}$$

as the free analog of the expectation. Note that the eigenvalues of a matrix A are completely constrained by the trace of powers A^k—therefore, one can study the average distribution of the eigenvalues of a random matrix A via the moments $\phi\left(A^k\right)$. *Free independence* (or *freeness*) is a generalization of the notion of independence in commutative probability theory to free probability theory. In the context of random matrix theory, two $N \times N$ random matrices A and B are said to be freely independent if the mixed moments are identically zero; that is,

$$\phi\left(\left(A^{m_1} - \phi\left(A^{m_1}\right)\right)\left(B^{n_1} - \phi\left(B^{n_1}\right)\right)\ldots\left(A^{m_k} - \phi\left(A^{m_k}\right)\right)\left(B^{n_k} - \phi\left(B^{n_k}\right)\right)\right) = 0 \tag{A.67}$$

for all $n_i, m_i \in \mathbb{N}$. Roughly, the free independence of two random matrices means that their eigenbases are in "generic position" from one another.

Taking the analogy with commutative probability theory further, the analog of the moment-generating function associated with the distribution of a random variable is the *Stieltjes transform* of the measure μ:

$$G_\mu(z) = \int \frac{\mathrm{d}\mu(t)}{z-t}, \tag{A.68}$$

which can be inverted via the Stieltjes inversion formula:

$$\mathrm{d}\mu(t) = -\frac{1}{\pi} \lim_{\epsilon \to 0^+} \mathrm{Im}\left\{G_\mu(t + \mathrm{i}\epsilon)\right\} \mathrm{d}t\,. \tag{A.69}$$

Similarly, the free analog of the cumulant-generating function is the *R-transform*, which can be defined via the Stieltjes transform as the solution to the implicit equation:

$$\mathcal{R}_\mu(G_\mu(z)) + \frac{1}{G_\mu(z)} = z. \tag{A.70}$$

The R-transform is important in that, if two random variables A and B are freely independent with probability measures μ_A and μ_B respectively, the probability measure μ_{A+B} of $A+B$ satisfies

$$\mathcal{R}_{\mu_{A+B}} = \mathcal{R}_{\mu_A} + \mathcal{R}_{\mu_B}. \tag{A.71}$$

This can be interpreted as the free analog of the additivity of cumulants for commutative random variables. The probability measure μ_{A+B} is called the *free convolution* of μ_A and μ_B, and is denoted using the notation

$$\mu_{A+B} = \mu_A \boxplus \mu_B. \tag{A.72}$$

Thus, given the probability distributions of two free random variables A and B, there is a prescription for determining the probability distribution of their sum by taking their free convolution, just as the convolution in commutative probability theory describes the distribution of the sum of random variables.

A.4.2 Large Deviations Theory

In order to bound the probability of large deviations from the weak convergence of the eigenvalue distribution of $\boldsymbol{C}$ to its asymptotic limit we will use results from large deviations theory, which we briefly review here. A sequence of measures $\{\mu_n\}$ is said to satisfy a large deviation principle in the limit $n \to \infty$ with speed $s(n)$ and lower semicontinuous rate function I with codomain $[0, \infty]$ if and only if [195]

$$-\inf_{x \in \Gamma^\circ} I(x) \leq \lim \inf_{n \to \infty} \frac{1}{s(n)} \ln\left(\mu_n(\Gamma)\right) \leq \lim \sup_{n \to \infty} \frac{1}{s(n)} \ln\left(\mu_n(\Gamma)\right) \leq -\inf_{x \in \overline{\Gamma}} I(x) \tag{A.73}$$

for all Borel measurable sets Γ that all μ_n are defined on. Here, $\overline{\Gamma}$ denotes the closure of Γ and Γ° the interior of Γ. The rate function I is said to be *good* if all level sets of I are compact. Large deviations theory will be useful for us to bound the probabilities of large deviations of the empirical spectral distribution of $\mu_{\boldsymbol{C}(x)}$ as $p \to \infty$, and show that they do not contribute to leading order in the (logarithmic) asymptotic distribution of critical points. We do this using Varadhan's lemma, which we state now.

Lemma A.6 (Varadhan's lemma [195])**.** *Suppose* $\{\mu_n\}$ *satisfies a large deviation principle with speed* $s(n)$ *and good rate function* I *and let* ϕ *be a real-valued continuous function. Further assume either the* tail condition

$$\lim_{M \to \infty} \lim \sup_{n \to \infty} \frac{1}{s(n)} \ln\left(\mathbb{E}_{X_n \sim \mu_n}\left[\mathrm{e}^{s(n)\phi(X_n)} \mathbf{1}\left\{\phi(X_n) \geq M\right\}\right]\right) = -\infty, \tag{A.74}$$

or the moment condition *for some* $\gamma > 1$

$$\lim \sup_{n \to \infty} \frac{1}{s(n)} \ln\left(\mathbb{E}_{X_n \sim \mu_n}\left[\mathrm{e}^{\gamma s(n)\phi(X_n)}\right]\right) < \infty. \tag{A.75}$$

Then,

$$\lim_{n \to \infty} \frac{1}{s(n)} \ln\left(\mathbb{E}_{X_n \sim \mu_n}\left[\mathrm{e}^{s(n)\phi(X_n)}\right]\right) = \sup_x \left(\phi(x) - I(x)\right). \tag{A.76}$$

A.4.3 Logarithmic Asymptotics of the Distribution of Critical Points

Equipped with these mathematical tools, we prove our first result on the asymptotic behavior of $\mu_{\boldsymbol{C}(x)}$, which is present in the expectation of Equation (A.59).

Lemma A.7 (Asymptotic behavior of $\mu_{\boldsymbol{C}(x)}$)**.** *Define $G_x^*(z)$ as the implicit solution of the equation*

$$8r^3\gamma^2 x G_x^*(z)^3 - 2r\gamma(z+2rx)G_x^*(z)^2 + (z-2r(1-\gamma))G_x^*(z) - 1 = 0 \tag{A.77}$$

with the smallest imaginary part. Define

$$\mathrm{d}\mu_x^* \equiv -\frac{1}{\pi}\operatorname{Im}\{G_x^*\}\,\mathrm{d}\lambda\,. \tag{A.78}$$

Let $p, m \to \infty$ as $\gamma = \frac{p}{2m}$ is held constant. Then, the empirical spectral measure $\mu_{\boldsymbol{C}(x)}$ satisfies a large deviation principle as $p \to \infty$ with speed p^2 with good rate function uniquely minimized at μ_x^* with a value of 0*.*

Proof. The empirical spectral measure of the random matrix $\boldsymbol{N}/\sqrt{p}$ satisfies a large deviation principle at a scale p^2, with good rate function minimized by Wigner's semicircle law [196]. Similarly, the empirical spectral measure of the random matrix $\boldsymbol{W}/2m$ satisfies a large deviation principle at a scale p^2, with good rate function minimized by the Marchenko–Pastur distribution [197]. As the R-transform of the empirical spectral distribution of $\boldsymbol{A}$ satisfies the scaling property

$$\mathcal{R}_{a\boldsymbol{A}}(z) = a\mathcal{R}_{\boldsymbol{A}}(az)\,, \tag{A.79}$$

the R-transform of the empirical spectral distribution of the weighted GOE term of $\boldsymbol{C}$ is of the form

$$\mathcal{R}_{\mathrm{GOE}}(z) = 4r^2\gamma x z \tag{A.80}$$

and the R-transform of the weighted Wishart term is of the form

$$\mathcal{R}_{\text{Wishart}}(z) = \frac{2r}{1 - 2r\gamma z}. \tag{A.81}$$

By the asymptotic freeness of independent GOE and Wishart matrices [198], $\mu_{\boldsymbol{C}(x)}$ converges weakly to the fixed measure μ^* with R-transform

$$\mathcal{R}_x(z) = \mathcal{R}_{\text{Wishart}}(z) + \mathcal{R}_{\text{GOE}}(z). \tag{A.82}$$

Equation A.77 and Equation (A.78) now follow from inverting the R-transform $\mathcal{R}_x$ via Equation (A.70) and Equation (A.69), respectively.

We now consider large deviations in the weak convergence $\mu_{\boldsymbol{C}(x)} \rightsquigarrow \mu_x^*$. Conditioning on the empirical spectral distribution of $\boldsymbol{W}/2m$ and using the "strongest growth wins" principle [195], we have that $\mu_{\boldsymbol{C}(x)}$ satisfies a large deviation principle with speed p^2 with rate function given by

$$I(\mu) = \inf_{\mu_{\boldsymbol{W}/2m}} \left(J_{\mu_{\boldsymbol{W}/2m}}(\mu) + K\left(\mu_{\boldsymbol{W}/2m}\right) \right); \tag{A.83}$$

here, K is the rate function governing convergence of the empirical spectral distribution of the Wishart ensemble [197] and $J_{\mu_{\boldsymbol{W}/2m}}$ is the rate function governing convergence of the empirical spectral distribution of a fixed matrix with asymptotic eigenvalue distribution $\mu_{\boldsymbol{W}/2m}$ summed with a GOE matrix [199]. This sum is obviously uniquely minimized by $\mu = \mu^*$, when $I(\mu^*) = 0$. □

Now, we examine the asymptotic behavior of the smallest eigenvalue $\lambda_1^{\boldsymbol{C}(x)}$ of $\boldsymbol{C}(x)$. Unlike the empirical spectral measure $\mu_{\boldsymbol{C}(x)}$ which satisfies a large deviation principle at a speed p^2, we will see that this eigenvalue satisfies a large deviation principle at a speed p, with deviations at this speed to the left of the asymptotic value $\lambda_{x,1}^*$.

Lemma A.8 (Asymptotic behavior of $\lambda_1^{\boldsymbol{C}(x)}$)**.** *Let $\lambda_{x,1}^*$ be the infimum of the support of μ_x^* as defined in Equation* (A.78)*. Then, the smallest eigenvalue $\lambda_1^{\boldsymbol{C}(x)}$ of $\boldsymbol{C}(x)$*

satisfies a large deviation principle with speed p with good rate function that is infinite at $y > \lambda_{x,1}^*$ and is uniquely minimized at $y = \lambda_{x,1}^*$ with a value of 0.

Proof. The limiting smallest eigenvalues $\lambda_1^{\boldsymbol{W}}, \lambda_1^{\boldsymbol{N}}$ of $\boldsymbol{W}/2m$ and $\boldsymbol{N}/\sqrt{p}$ both satisfy large deviation principles with speed p that are infinite for λ_1 in the bulk of their respective limiting empirical spectral distributions [200, 201]. As in the proof of Lemma A.7, we condition on large deviations of these eigenvalues [195] and therefore have that the rate function governing $\lambda_1^{\boldsymbol{C}(x)}$ is

$$I(y) = \inf_{\lambda_1^{\boldsymbol{W}}, \lambda_1^{\boldsymbol{N}}} \left(J_{\lambda_1^{\boldsymbol{W}}, \lambda_1^{\boldsymbol{N}}}(y) + K\left(\lambda_1^{\boldsymbol{W}}\right) + L\left(\lambda_1^{\boldsymbol{N}}\right) \right); \tag{A.84}$$

here, K is the rate function governing the convergence of $\lambda_1^{\boldsymbol{W}}$, L that of $\lambda_1^{\boldsymbol{N}}$, and J that of the smallest eigenvalue $\boldsymbol{C}(x)$ conditioned on the eigenvalue distributions of $\boldsymbol{W}$ and $\boldsymbol{N}$. Using known results on the large deviations of the smallest eigenvalue of the sum of two matrices with fixed eigenvalues (i.e. $J_{\lambda_1^{\boldsymbol{W}}, \lambda_1^{\boldsymbol{N}}}$) [202], we see that $I(y)$ is infinite for $y > \lambda_{x,1}^*$ and is uniquely minimized at $y = \lambda_{x,1}^*$ with a value of 0. □

Using Lemmas A.7 and A.8, we can prove the following logarithmic asymptotics on the expectation term in Equation (A.59). We will find that neither the large deviations in the convergence $\mu_{\boldsymbol{C}(x)}$ or $\lambda_1^{\boldsymbol{C}(x)}$ will contribute to leading order in the logarithmic asymptotics of $\mathrm{Crt}_k(E)$, as at a speed p the only large deviations are $\lambda_1^{\boldsymbol{C}(x)} \leq \lambda_{x,1}^*$ which are dominated by $\lambda_1^{\boldsymbol{C}(x)} = \lambda_{x,1}^*$ in the expectation.

Lemma A.9 (Logarithmic asymptotics of the determinant)**.** *Let $\mathrm{d}\mu_E^*$ be the spectral measure given in Equation* (A.78)*, with $\lambda_{E,1}^*$ the infimum of its support. Let $p, m \gg 1$ with $\frac{p}{2m} = \gamma = \mathrm{O}(1)$. Then,*

$$\begin{aligned} &\frac{1}{p} \ln\left(\mathbb{E}_{\boldsymbol{C}(E)} \left[\mathrm{e}^{p \int \ln(|\lambda - 2rE|) \mathrm{d}\mu_{\boldsymbol{C}(E)}} \mathbf{1}\left\{ \lambda_{k+1}^{\boldsymbol{C}(E)} \geq 2rE \right\} \right] \right) \\ &= \int \ln\left(\mathbf{1}\left\{ \lambda_{E,1}^* \geq 2rE \right\} |\lambda - 2rE| \right) \mathrm{d}\mu_E^* + \mathrm{o}(1). \end{aligned} \tag{A.85}$$

Proof. As $\mu_{\boldsymbol{C}(E)}$ satisfies a large deviation principle with speed p^2 with rate function minimized at μ_E^* by Lemma A.7, and as $\mathbf{1}\left\{ \lambda_1^{\boldsymbol{C}(E)} \geq 2rE \right\} \leq 1$, we have that the tail

condition of Varadhan's lemma at speed p is satisfied [195] and therefore

$$\begin{aligned}\lim_{p\to\infty}\frac{1}{p}\ln\left(\mathbb{E}_{\boldsymbol{C}(E)}\left[\mathrm{e}^{p\int\ln(|\lambda-2rE|)\mathrm{d}\mu_{\boldsymbol{C}(E)}}\mathbf{1}\left\{\lambda_1^{\boldsymbol{C}(E)}\geq 2rE\right\}\right]\right)\\ =\sup_{\lambda\in\mathbb{R}}\left(\int\ln\left(\mathbf{1}\left\{\lambda\geq 2rE\right\}|\lambda-2rE|\right)\mathrm{d}\mu_E^*-I\left(\lambda\right)\right).\end{aligned}\tag{A.86}$$

Here, I is as in Equation (A.84). The supremum over λ is obviously achieved when $\lambda=\lambda_{E,1}^*$ by the properties of I discussed in Lemma A.8, giving the leading order term in Equation (A.85). The result being exact in the $p\to\infty$ limit gives the subleading $\mathrm{o}\left(1\right)$. □

Using Lemma A.9, we can therefore finally calculate the logarithmic asymptotic distribution of local minima of a WHRF.

Theorem A.10 (Logarithmic asymptotics of the local minima distribution)**.** *Let* $\mathrm{d}\mu_E^*$ *be the spectral measure given in Equation* (A.78)*, with* $\lambda_{E,1}^*$ *the infimum of its support. Let* $p,m\gg 1$ *with* $\frac{p}{2m}=\gamma=\mathrm{O}\left(1\right)$*. Then, the expected distribution of local minima of* F_{WHRF} *at a fixed energy* $E>0$ *is given by*

$$\begin{aligned}\frac{1}{p}\ln\left(\mathbb{E}\left[\mathrm{Crt}_0\left(E\right)\right]\right)=\frac{1}{2}\ln\left(\frac{\pi q}{2\gamma}\right)+\frac{1}{2\gamma}\left(1-E\right)+\frac{1}{2}\left(\gamma^{-1}-1\right)\ln\left(E\right)\\ +\int\ln\left(\left|\frac{\lambda}{r}-2E\right|\mathbf{1}\left\{\frac{\lambda_{E,1}^*}{r}\geq 2E\right\}\right)\mathrm{d}\mu_E^*+\mathrm{o}\left(1\right).\end{aligned}\tag{A.87}$$

Proof. The result follows directly from applying Lemma A.9 to Theorem A.5. □

Note that, though we only prove the asymptotic distribution of local minima in Theorem A.10, we expect similar theorems to also hold for critical points of constant index k (taking $\lambda_{E,1}^*\mapsto\lambda_{E,k}^*$ in the integrand). The only difference in the derivation is the exact form of the large deviations of the kth smallest eigenvalue of $\boldsymbol{C}\left(x\right)$. This is similar to the case in Gaussian hyperspherical random fields [73].

A.5 Details of the Numerical Simulations

We now give further details on the numerical simulations performed in Section 2.4. We performed all simulations via Qiskit [203], and used standard gradient descent (via the method of finite differences) to optimize the VQA loss function

$$F(\boldsymbol{\theta}) = \langle \boldsymbol{\theta} | H_{\boldsymbol{T},\boldsymbol{U}} | \boldsymbol{\theta} \rangle \tag{A.88}$$

until convergence. $H_{T,U}$ is the 1D n site *spinless Fermi–Hubbard Hamiltonian* [101]

$$H_{\boldsymbol{T},\boldsymbol{U}} = -\sum_{i=1}^{n-1} T_i \left(c_i^\dagger c_{i+1} + c_{i+1}^\dagger c_i \right) + \sum_{i=1}^{n-1} U_i c_i^\dagger c_i c_{i+1}^\dagger c_{i+1}, \tag{A.89}$$

where c is the fermionic annihilation operator. T_i and U_i are i.i.d. normally distributed in order to break translational invariance. In our simulations, these random variables were centered at $T = 1$ and $U = 2$, respectively, and each had a variance of 10^{-2}.

Our implementation of gradient descent used a learning rate of 0.05 and a momentum of 0.9, and halted when either the function value improved by no more than 10^{-5} or after 10^6 iterations, whichever came first. We initialized each instance at a uniformly random point in parameter space, with each parameter initialized within $[-2\pi, 2\pi]$.

To estimate the empirical distribution of local minima for the studied instances of the varitional quantum eigensolver (VQE) [19], we repeated this procedure 52 times, using a new ansatz and uniformly random starting point for each training instance. We also verified numerically that m as defined in Equation (A.6) is at least on the order of 2^n for $H_{1,2}$, though we directly used Equation (A.6) when computing γ. In all plotted instances, we normalize the energy scale by a factor of c_{VQA}, where

$$c_{\mathrm{VQA}} = \overline{\lambda} - \lambda_1; \tag{A.90}$$

this is just the overall factor of Equation (2.7). These units are such that the mean eigenvalue of $H - \lambda_1$ in the subspace of interest is at $E = 1$.

In Section 2.4.2, we tested our analytic results against a Hamiltonian informed ansatz. Specifically, we used the *Hamiltonian variational ansatz* (HVA) [178]. For the Fermi–Hubbard Hamiltonian of Equation (A.89), each HVA layer is of the form

$$U_i^{\boldsymbol{T},\boldsymbol{U}}\left(\boldsymbol{\theta}_i\right) = \mathrm{e}^{-\mathrm{i}\theta i,2H_{T,U,\mathrm{odd}}}\mathrm{e}^{-\mathrm{i}\theta i,2H_{T,U,\mathrm{even}}}\mathrm{e}^{-\mathrm{i}\theta_{i,1}H_{T,U,\mathrm{Coulomb}}}. \tag{A.91}$$

Here, $H_{\boldsymbol{T},\boldsymbol{U},\mathrm{Coulomb}}$ is composed of the terms proportional to U_i in $H_{\boldsymbol{T},\boldsymbol{U}}$, $H_{\boldsymbol{T},\boldsymbol{U},\mathrm{even}}$ the hopping terms on even links, and $H_{\boldsymbol{T},\boldsymbol{U},\mathrm{odd}}$ the hopping terms on odd links. We took the starting state $|\psi_0\rangle$ to be the computational basis state $|1\rangle$ on the first $\frac{n}{2}$ qubits and $|0\rangle$ on the other $\frac{n}{2}$ qubits. To observe the effects of scaling the number of independent parameters p, we overparameterize our ansatz at a fixed overall depth by fixing the total number of ansatz layers $U_i^{\boldsymbol{T},\boldsymbol{U}}$ to be 6, but introduce extra parameters to govern each evolution. For instance, for a multiplicative factor $f = 2$, we double the number of parameters by splitting into a sum of two terms

$$H_{\boldsymbol{T},\boldsymbol{U},\mathrm{Coulomb}} = H^{(1)}_{\boldsymbol{T},\boldsymbol{U},\mathrm{Coulomb}} + H^{(2)}_{\boldsymbol{T},\boldsymbol{U},\mathrm{Coulomb}}, \tag{A.92}$$

$$H_{\boldsymbol{T},\boldsymbol{U},\mathrm{even}} = H^{(1)}_{\boldsymbol{T},\boldsymbol{U},\mathrm{even}} + H^{(2)}_{\boldsymbol{T},\boldsymbol{U},\mathrm{even}}, \tag{A.93}$$

$$H_{\boldsymbol{T},\boldsymbol{U},\mathrm{odd}} = H^{(1)}_{\boldsymbol{T},\boldsymbol{U},\mathrm{odd}} + H^{(2)}_{\boldsymbol{T},\boldsymbol{U},\mathrm{odd}}, \tag{A.94}$$

and parameterize the evolution under each term separately. For $f = 1$, this ansatz preserves the fermion number of the initial state; thus, for these simulations we calculate m in this $\frac{n}{2}$-fermion subspace. For large f, this parameterization breaks the fermion number conservation of the ansatz, but still preserves the parity of the fermion number. In practice, then, the γ we compute should be considered an upper bound on the true γ, strengthening our empirical results.

Appendix B

Technical Details for Chapter 3

B.1 Training Error Dominates in the Optimization of Variational Quantum Algorithms

The variational quantum eigensolver [19] is purely a problem of optimization and may appear unrelated to the challenges in learning via variational algorithms; however, by decomposing the error of a learning algorithm into key terms using well-established methods [204], we will show that variational learning algorithms essentially face the same optimization task and its associated challenges. In both cases, the hardness of learning or optimizing with variational circuits manifests itself in the challenges of optimizing over a cost landscape riddled with traps (or other barriers to optimization).

We restrict ourselves here to the supervised learning framework of empirical risk minimization, where our goal is to learn a space of input and output pairs $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$ drawn from a distribution $P(\boldsymbol{x}, \boldsymbol{y})$. Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$, we quantify how well our function performs by considering the expected risk $\mathcal{R}$:

$$\mathcal{R}(f) = \mathbb{E}_{\boldsymbol{x} \sim P(\boldsymbol{x})} \left[\ell(f(\boldsymbol{x}), f^*(\boldsymbol{x})) \right], \tag{B.1}$$

where the expectation above is taken with respect to $P(\boldsymbol{x}, \boldsymbol{y})$. To benchmark performance, we compare to the "optimal" or target function f^* which is the minimizer of

the risk:

$$f^*(\boldsymbol{x}) = \text{argmin}_{\hat{\boldsymbol{y}} \in \mathcal{Y}} \mathbb{E}_{\boldsymbol{y} \sim P(\boldsymbol{y}|\boldsymbol{x})} \left[\ell(\boldsymbol{y}, \hat{\boldsymbol{y}}) \right]. \tag{B.2}$$

To perform learning, we search for a function $\hat{f} \in \mathcal{F}$ in the function class $\mathcal{F}$ (for instance, the set of functions expressed by quantum neural networks). The expected risk $\mathcal{R}(f)$ is not something one can calculate since it requires access to the full probability distribution of the data. Instead, one minimizes the empirical risk $\hat{\mathcal{R}}(f)$ (often named the training error) over a given training data set $\mathcal{D}$ of size N consisting of pairs $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$:

$$\hat{\mathcal{R}}(f) = \sum_{i=1}^{N} \ell\left(f(\boldsymbol{x}_i), f^*(\boldsymbol{x}_i)\right). \tag{B.3}$$

Note that we use the hat in $\hat{\mathcal{R}}$ and $\hat{f}$ to denote the expected risk measure and function that one actually has access to during training or optimization. Given the above, one can bound the expected risk of any function $\hat{f}$ as a decomposition below [204]:

$$\mathbb{E}\left[\mathcal{R}\left(\hat{f}\right) - \mathcal{R}(f^*)\right] \leq \underbrace{\min_{f \in \mathcal{F}} \mathcal{R}(f) - \mathcal{R}(f^*)}_{\text{approximation error}} + \underbrace{2\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\mathcal{R}(f) - \hat{\mathcal{R}}(f)\right|\right]}_{\text{generalization error}} + \underbrace{\mathbb{E}\left[\hat{\mathcal{R}}\left(\hat{f}\right) - \min_{f \in \mathcal{F}} \hat{\mathcal{R}}(f)\right]}_{\text{optimization error}}, \tag{B.4}$$

where the expectation above is taken with respect to the distribution over data sets or training sets. The proof of this statement follows by a careful, yet straightforward, application of additions/subtractions with corresponding bounds [204].

Proof. Let $\hat{f}_{\mathcal{F}} = \text{argmin}_{f \in \mathcal{F}} \hat{\mathcal{R}}(f)$ and $f_{\mathcal{F}} = \text{argmin}_{f \in \mathcal{F}} \mathcal{R}(f)$. Then, by adding and subtracting quantities, we obtain the following result:

$$\begin{aligned} \mathbb{E}\left[\mathcal{R}\left(\hat{f}\right) - \mathcal{R}(f^*)\right] = \mathbb{E}\Big[&\mathcal{R}\left(\hat{f}\right) - \mathcal{R}(f^*) \\ &+ \hat{\mathcal{R}}\left(\hat{f}_{\mathcal{F}}\right) - \hat{\mathcal{R}}\left(\hat{f}_{\mathcal{F}}\right) \\ &+ \mathcal{R}(f_{\mathcal{F}}) - \mathcal{R}(f_{\mathcal{F}}) + \hat{\mathcal{R}}(f_{\mathcal{F}}) - \hat{\mathcal{R}}(f_{\mathcal{F}}) \\ &+ \hat{\mathcal{R}}\left(\hat{f}\right) - \hat{\mathcal{R}}\left(\hat{f}\right)\Big]. \end{aligned} \tag{B.5}$$

We reorder the above as follows and note their relation to the main statement:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}\left(\hat{f}\right) - \mathcal{R}\left(f^*\right)\right] =& \mathbb{E}\left[\mathcal{R}\left(f_{\mathcal{F}}\right) - \mathcal{R}\left(f^*\right)\right] && \text{(approximation error)} && \text{(B.6)}\\
&+ \mathbb{E}\left[\mathcal{R}\left(\hat{f}\right) - \hat{\mathcal{R}}\left(\hat{f}\right)\right] && \text{(generalization error)} && \text{(B.7)}\\
&+ \mathbb{E}\left[\hat{\mathcal{R}}\left(f_{\mathcal{F}}\right) - \mathcal{R}\left(f_{\mathcal{F}}\right)\right] && \text{(generalization error)} && \text{(B.8)}\\
&+ \mathbb{E}\left[\hat{\mathcal{R}}\left(\hat{f}_{\mathcal{F}}\right) - \hat{\mathcal{R}}\left(f_{\mathcal{F}}\right)\right] && (\leq 0 \text{ since } \hat{f}_{\mathcal{F}} \text{ minimizes } \hat{\mathcal{R}}) && \\
& && && \text{(B.9)}\\
&+ \mathbb{E}\left[\hat{\mathcal{R}}\left(\hat{f}\right) - \hat{\mathcal{R}}\left(\hat{f}_{\mathcal{F}}\right)\right] && \text{(optimization error).} && \text{(B.10)}
\end{aligned}$$

For the quantities in the generalization error, we have since $\hat{f}, f_{\mathcal{F}} \in \mathcal{F}$:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}\left(\hat{f}\right) - \hat{\mathcal{R}}\left(\hat{f}\right)\right] &\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathcal{R}\left(f\right) - \hat{\mathcal{R}}\left(f\right)\right|\right]\\
\mathbb{E}\left[\hat{\mathcal{R}}\left(f_{\mathcal{F}}\right) - \mathcal{R}\left(f_{\mathcal{F}}\right)\right] &\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathcal{R}\left(f\right) - \hat{\mathcal{R}}\left(f\right)\right|\right].
\end{aligned} \tag{B.11}$$

Plugging these into Equation (B.5) and noting as before that $\mathbb{E}\left[\hat{\mathcal{R}}\left(\hat{f}_{\mathcal{F}}\right) - \hat{\mathcal{R}}\left(f_{\mathcal{F}}\right)\right] \leq 0$, we arrive at the desired result. □

In the context of quantum variational algorithms, each of these has the following properties:

- The **approximation error** quantifies how well the most optimal function in the hypothesis class $\mathcal{F}$ can fit the function. In variational settings, the approximation error is typically bounded by assuming the target function is generated from a nice class of functions (e.g. shallow circuits) or arguing either analytically or theoretically that a given ansatz can (approximately) express the target function [58, 179, 205, 206].

- The **generalization error** quantifies the statistical error that arises from having a finite data set and is typically insignificant in quantum variational algorithms where circuit complexity is limited with regards to the number of training samples. More precisely, for data sets of size m, previous work [207, 208] bound

the generalization error as $\tilde{\mathrm{O}}\left(\sqrt{|G|/m}\right)$ where $|G|$ is the number of trainable gates. In contrast, generalization error in heavily overparameterized classical neural network models are challenging to bound and it is still an open question why deep learning models generalize so well [209, 210].

- The **optimization error** measures how well one is able to reduce the empirical risk. Issues with optimization such as poor local minima and barren plateaus arise here. Note that there is a distinct difference between quantum and classical deep learning here. With classical deep neural networks, this quantity is typically negligible since neural networks are overparameterized with respect to the data set size and can fit random data arbitrarily well [209, 211]. Furthermore, due to efficient means of calculating gradients with bit-level precision, classical machine learning algorithms perform optimization over parameters far more efficiently than quantum variational algorithms. In quantum variational models, overparameterization with respect to the Hilbert space dimension is generally needed to arbitrarily fit data [49, 69, 70]. Since the Hilbert space dimension grows exponentially with the number of qubits, such overparameterization becomes prohibitive rather rapidly.

In summary, the approximation error and generalization error can be bounded efficiently with sufficient data so failures in learning are typically related to optimization over the empirical risk. As an aside, this is loosely analogous to the classical setting of learning polynomial size Boolean circuits which is strongly conjectured to be hard since the space of Boolean functions is challenging to search over [212].

Finally, we would like to stress that the decomposition of the excess risk performed in this section is neither unique nor necessarily tight. The decomposition can be performed in various other ways depending on the quantities one would like to bound. We chose the decomposition here to relate errors in quantum machine learning algorithms to their classical counterparts and to highlight the challenges one may face when attempting to provably learn a target function class.

B.2 Statistical Query Framework: Background and Additional Details

The statistical query (SQ) framework was introduced nearly 25 years ago to analyze the hardness of learning problems [111]. This framework restricts algorithms to a series of noisy queries, and hardness results are stated in terms of the number of queries needed to learn a given class of functions. Since there are various different ways of defining the statistical query model—including a recent quantum oracular version proposed in Reference [213]—let us first review some of the various models considered in prior work.

1. **Classical statistical query model**: Introduced by Reference [111], this was the first statistical query model introduced. For a given distribution D of inputs over an input space X and target concept $c : X \to \{-1, +1\}$, one can make a statistical query $\mathrm{SQ}(q, \tau)$, by providing a threshold $\tau \in \mathbb{R}^+$ and a query function $q : X \times \{-1, +1\} \to \{-1, +1\}$. The query returns a value in the range:

$$\mathbb{E}_{x\sim D}\left[q\left(x, c\left(x\right)\right) - \tau\right] \leq \mathrm{SQ}\left(q, \tau\right) \leq \mathbb{E}_{x\sim D}\left[q\left(x, c\left(x\right)\right) + \tau\right]. \tag{B.12}$$

2. **Correlational statistical query model**: The query is the same as before, except now, one queries correlations $\mathrm{CSQ}(q, \tau)$ only by providing a threshold $\tau \in \mathbb{R}^+$ and a query function $h : X \to \{-1, +1\}$. The query returns a value in the range:

$$\mathbb{E}_{x\sim D}\left[h\left(x\right) c\left(x\right) - \tau\right] \leq \mathrm{CSQ}\left(h, \tau\right) \leq \mathbb{E}_{x\sim D}\left[h\left(x\right) c\left(x\right) + \tau\right]. \tag{B.13}$$

   This model is strictly less powerful than the standard statistical query model since one can perform a correlational statistical query with a standard statistical query [104].

3. **Quantum statistical query model**: This is a statistical query model with quantum samples [213]. Here, we are restricted to target (classical) Boolean

functions $c : \{0,1\}^n \to \{-1,+1\}$. A quantum statistical query $\mathrm{Qstat}(\tau, M)$ is provided with a threshold $\tau \in \mathbb{R}^+$ and an observable or Hamiltonian $M \in (\mathbb{C}^2)^{n+1} \times (\mathbb{C}^2)^{n+1}$ satisfying $\|M\|_\infty \leq 1$ and returns a number in the range:

$$\langle\psi_c| M |\psi_c\rangle - \tau \leq \mathrm{Qstat}(M, \tau) \leq \langle\psi_c| M |\psi_c\rangle + \tau, \tag{B.14}$$

where $|\psi_c\rangle = \sum_{\boldsymbol{x}\in\{0,1\}^n} \sqrt{D(\boldsymbol{x})} |\boldsymbol{x}\rangle |c(\boldsymbol{x})\rangle$. This model is useful to analyze the hardness of learning classical Boolean functions when given the extra power of querying the classical function in superposition. Our work considers learning quantum data and thus does not fit into the framework of this SQ model.

The SQ learning setting is related to the probably approximately correct (PAC) setting of learning theory [214] in that if an algorithm can learn a given function class in the SQ learning setting under any input distribution, then that function class is also PAC learnable [109, 111]. Two very recent works have studied the SQ hardness of learning data generated by quantum circuits. First, Reference [119] analyze the hardness of learning the output distribution of clifford circuits and stabilizer states showing that these distributions are hard to learn using classical Boolean SQ oracles. Nevertheless, when given samples from the Boolean hypercube of the distribution, they provide an efficient algorithm based on linear regression to determine the stabilizer state underlying the distribution. Such a result is similar to classic results in Reference [111] showing that parity functions are hard to learn using only SQ oracle calls but easy when performing linear regression with enough samples. Second, Reference [118] shows that learning stabilizer states is hard in an SQ setting where queries are made over two-outcome POVMs. Their results show that learning stabilizer states in such a setting is as hard as learning the function class of parity with noise in the standard Boolean setting. Our results expand the set of quantum functions that are hard to learn in SQ settings and relate such hardness results to the variational setting.

B.2.1 Quantum Statistical Query Models

Variational quantum algorithms are inherently noisy due to unavoidable sources such as the need for sampling outputs, or potentially correctable sources such as gate errors and state preparation noise. In such noisy settings, the statistical query (SQ) model provides a useful framework for quantifying the complexity of learning a class of functions by considering how many query calls to a noisy oracle are needed to learn any function in that class. We consider two forms of statistical queries which relate to learning a target Hamiltonian or a target unitary, both of which result in exponential hardness results for learning simple variational classes of data:

Definition B.1 (Quantum correlational statistical query (qCSQ))**.** Assume there is a target observable M that we would like to learn on some distribution over states $\mathcal{D}$. Applying the correlational SQ model to the quantum setting, we define the query $\mathrm{qCSQ}\left(O,\tau\right)$ which takes in a bounded observable O with $\left\|O\right\|_\infty \leq 1$ and a tolerance τ and returns a value in the range:

$$\mathbb{E}_{\rho\sim\mathcal{D}}\left[\mathrm{tr}(O\rho)\,\mathrm{tr}(M\rho)-\tau\right] \leq \mathrm{qCSQ}(O,\tau) \leq \mathbb{E}_{\rho\sim\mathcal{D}}\left[\mathrm{tr}(O\rho)\,\mathrm{tr}(M\rho)+\tau\right]. \tag{B.15}$$

Definition B.2 (Quantum unitary statistical query (qUSQ))**.** In the unitary compilation setting, one aims to learn a target unitary transformation U_* over a distribution $\mathcal{D}$ of input/output pairs of that unitary transformation. Here, the oracle $\mathrm{qUSQ}(V,\tau)$ takes in a unitary matrix V and a tolerance τ and returns a value in the range:

$$\mathbb{E}_{\rho\sim\mathcal{D}}\left[\mathrm{Re}\left\{\mathrm{tr}\left(U_*^\dagger V\rho\right)\right\}-\tau\right] \leq \mathrm{qUSQ}\left(V,\tau\right) \leq \mathbb{E}_{\rho\sim\mathcal{D}}\left[\mathrm{Re}\left\{\mathrm{tr}\left(U_*^\dagger V\rho\right)\right\}+\tau\right]. \tag{B.16}$$

Importantly, if $\mathcal{D}$ is a 1-design over n qubit states, then the above can be simplified using the formula $\mathbb{E}_{\rho\sim\mathcal{D}}\left[\mathrm{Re}\left\{\mathrm{tr}\big(U_*^\dagger V\rho\big)\right\}\right] = 2^{-n}\,\mathrm{Re}\left\{\mathrm{tr}\left(U_*^\dagger V\right)\right\}$ (see proof in Appendix B.3). Queries to qUSQ are related to performing a Hadamard test [215], also a common subroutine in variational algorithms [216].

The queries above take the forms of inner products, with $\langle M_1, M_2\rangle_{\mathcal{D}} = \mathbb{E}_{\rho\sim\mathcal{D}}\left[\mathrm{tr}\left(M_1\rho\right)\mathrm{tr}\left(M_2\rho\right)\right]$ and $\langle U_1, U_2\rangle_{\mathcal{D}} = \mathbb{E}_{\rho\sim\mathcal{D}}\left[\mathrm{Re}\left\{\mathrm{tr}\left(U_1^\dagger U_2\rho\right)\right\}\right]$. The inner products also induce corre-

sponding L_2 norms: $\|M\|_{\mathcal{D}} = \sqrt{\langle M, M\rangle_{\mathcal{D}}}$. As the magnitude of this norm can change with the dimension, we introduce the quantity $C_{\max}$ to denote the maximum value a query can take for any target observable in the qCSQ model, i.e. $C_{\max} = \max_{M:\|M\|_\infty \leq 1} \|M\|_{\mathcal{D}}^2$. We quantify noise tolerances and hardness bounds with respect to $C_{\max}$ to normalize units. Note that for the qUSQ model $C_{\max} = 1$, but in the qCSQ model, $C_{\max}$ can decay with the number of qubits under for example the Haar distribution of inputs.

A statistical query algorithm learns a function class if it can output a unitary or observable that is close to any target in that class.

Definition B.3 (qCSQ / qUSQ learning of hypothesis class)**.** A given algorithm using only statistical queries to qCSQ (qUSQ) successfully learns a hypothesis class $\mathcal{H}$ consisting of observables $M, \|M\|_\infty \leq 1$ (unitaries U) up to ϵ error if it is able to output an observable O (unitary V) which is ϵ-close to the unknown target observable $M \in \mathcal{H}$ ($U \in \mathcal{H}$) in the L_2 norm, i.e., $\|M - O\|_{\mathcal{D}} \leq \epsilon$ ($\|U - V\|_{\mathcal{D}} \leq \epsilon$).

The statistical query dimension quantifies the complexity of a hypothesis class $\mathcal{H}$ and is related to the number of queries needed to learn functions drawn from a class, as summarized in Theorem B.5.

Definition B.4 (Statistical query dimension [103, 109])**.** For a distribution $\mathcal{D}$ and concept class $\mathcal{H}$ where $\|M\|_{\mathcal{D}}^2 \leq C_{\max}$ for all $M \in \mathcal{H}$, the statistical query dimension (SQ-DIM$_{\mathcal{D}}(\mathcal{H})$) is the largest positive integer d such that there exists d observables $M_1, M_2, \ldots, M_d \in \mathcal{H}$ such that for all $i \neq j$, $\left|\langle M_i, M_j\rangle_{\mathcal{D}}\right| \leq C_{\max}/d$.

Theorem B.5 (Query complexity of learning [103, 104])**.** *Given a distribution $\mathcal{D}$ on inputs and a hypothesis class $\mathcal{H}$ where $\|M\|_{\mathcal{D}}^2 \leq C_{max}$ for all $M \in \mathcal{H}$, let $d =$* SQ-DIM$_{\mathcal{D}}(\mathcal{H})$ *be the statistical query dimension of $\mathcal{H}$. Any* qCSQ *or* qUSQ *learner making queries with tolerance $C_{max}\tau$ must make at least $(d\tau^2 - 1)/2$ queries to learn $\mathcal{H}$ up to error $C_{max}\tau$.*

Proof. Since we are restricted to the weaker setting of correlational statistical queries in this study, we can reuse a simple and elegant proof from Reference [104].

Let $M_1, M_2, \ldots, M_d$ be d functions that saturate SQ-DIM$_{\mathcal{D}}(\mathcal{H})$, i.e., $\langle M_i, M_j \rangle_{\mathcal{D}} \leq 1/d$ for all $i \neq j$. Assume we apply query O and let $S = \{i \in [d] : \langle O, M_i \rangle_{\mathcal{D}} > C_{\max}\tau\}$ Then, by simple application of the Cauchy–Schwarz inequality, we have that for any query O:

$$\begin{aligned} \left\langle O, \sum_{i \in S} M_i \right\rangle_{\mathcal{D}}^2 &\leq C_{\max} \left\| \sum_{i \in S} M_i \right\|_{\mathcal{D}}^2 \\ &= C_{\max} \sum_{i,j \in S} \langle M_i, M_j \rangle_{\mathcal{D}} \\ &\leq C_{\max}^2 \left(|S| + |S|^2 / d \right). \end{aligned} \tag{B.17}$$

Note that we can also bound the quantity above from below by using the definition of S:

$$\left\langle O, \sum_{i \in S} M_i \right\rangle_{\mathcal{D}} \geq C_{\max} |S| \tau. \tag{B.18}$$

Combining the above, we have that

$$|S| \leq \frac{d}{d\tau^2 - 1}. \tag{B.19}$$

Similarly, defining $S' = \{i \in [d] : \langle O, M_i \rangle_{\mathcal{D}} < -C_{\max}\tau\}$ with correlation less than $-\tau$, we follow the steps above to also note that $|S'| \leq d/(d\tau^2 - 1)$. Altogether, we have that $|S'| + |S| \leq 2d/(d\tau^2 - 1)$, which implies that each oracle call returning 0 is inconsistent with at most $2d/(d\tau^2 - 1)$ functions. This results in the lower bound stated, as d functions must be ruled inconsistent to learn the target class. □

This result forms the basis for our resulting proofs of hardness, summarized in Table 3.1 and proved in Appendix B.3.

Analogous to work in classical machine learning [110], one can perform noisy gradient descent as a series of statistical queries. As an example, consider the task of learning a target Hamiltonian M by constructing a variational Hamiltonian $H(\boldsymbol{\theta}) = U(\boldsymbol{\theta})^\dagger H U(\boldsymbol{\theta})$ with parameterized Pauli rotations and minimizing the mean squared error between expectations of M versus $H(\boldsymbol{\theta})$ over a distribution of states $\mathcal{D}$. Our loss function is

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\rho \sim \mathcal{D}} \left[\left(\operatorname{tr}(M\rho) - \operatorname{tr}(H(\boldsymbol{\theta})\rho) \right)^2 \right]. \tag{B.20}$$

The parameter shift rule [112] provides a means to calculate the partial derivative of a function $f(\mu)$ with respect to a parameter μ applied as a parameterized quantum gate $\mathrm{e}^{-\mathrm{i}\mu G}$ by calculating the function itself at two shifted coordinates. For example, for parameterized Pauli gates ($G \in \frac{1}{2}\{Z, X, Y\}$), this takes the form:

$$\frac{\partial f(\mu)}{\partial \mu} = \frac{1}{2}\left(f\left(\mu + \frac{\pi}{2}\right) - f\left(\mu - \frac{\pi}{2}\right)\right). \tag{B.21}$$

By applying the parameter shift rule [112], we can evaluate the gradient of the loss with respect to parameter entry θ_i as

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_i} = \mathbb{E}_{\rho\sim\mathcal{D}}\left[\left(\mathrm{tr}\left(H(\boldsymbol{\theta})\rho\right) - \mathrm{tr}\left(M\rho\right)\right)\left(\mathrm{tr}\left(H\left(\boldsymbol{\theta}^{+}\right)\rho\right) - \mathrm{tr}\left(H\left(\boldsymbol{\theta}^{-}\right)\rho\right)\right)\right], \tag{B.22}$$

where $\boldsymbol{\theta}^{+}$ and $\boldsymbol{\theta}^{-}$ are the values of the parameters shifted at the ith entry according to the parameter shift rule for the gradient. The quantity:

$$\mathbb{E}_{\rho\sim\mathcal{D}}\left[\mathrm{tr}\left(H(\boldsymbol{\theta})\rho\right)\left(\mathrm{tr}\left(H\left(\boldsymbol{\theta}^{+}\right)\rho\right) - \mathrm{tr}\left(H\left(\boldsymbol{\theta}^{-}\right)\rho\right)\right)\right] \tag{B.23}$$

can be directly evaluated without statistical queries, and the quantity:

$$\mathbb{E}_{\rho\sim\mathcal{D}}\left[\mathrm{tr}\left(M\rho\right)\left(\mathrm{tr}\left(H\left(\boldsymbol{\theta}^{+}\right)\rho\right) - \mathrm{tr}\left(H\left(\boldsymbol{\theta}^{-}\right)\rho\right)\right)\right] \tag{B.24}$$

can be evaluated using 2 statistical queries to qCSQ where the tolerance τ accounts for the noise in the estimate.

As a second example, this time in the unitary compiling setting of qUSQ, we can evaluate the commonly used procedure of measuring the inner product or average fidelity of n-qubit states between a target unitary U_* and a variationally chosen unitary $V(\boldsymbol{\theta})$ using statistical queries analogous to a swap test on actual quantum hardware [123, 217–219]. With slight abuse of notation, let $|\phi\rangle \sim \mathcal{D}$ denote a distribution over pure states which forms a 2-design. Then via averaging over 2-designs

(see Appendix B.3 for details), the average fidelity equals

$$\mathbb{E}_{|\phi\rangle\sim\mathcal{D}}\left[F\left(U_*\left|\phi\right\rangle, V\left(\boldsymbol{\theta}\right)\left|\phi\right\rangle\right)\right] = \mathbb{E}_{|\phi\rangle\sim\mathcal{D}}\left[\left|\left\langle\phi\right| V\left(\boldsymbol{\theta}\right)^\dagger U_*\left|\phi\right\rangle\right|^2\right] = \frac{2^{-n}\left|\mathrm{tr}\left(V\left(\boldsymbol{\theta}\right)^\dagger U_*\right)\right|^2 + 1}{2^n+1}. \tag{B.25}$$

Note that the key quantity:

$$\left|\mathrm{tr}\left(V\left(\boldsymbol{\theta}\right)^\dagger U_*\right)\right|^2 = \mathrm{Re}\left\{\mathrm{tr}\left(V\left(\boldsymbol{\theta}\right)^\dagger U_*\right)\right\}^2 + \mathrm{Re}\left\{\mathrm{i}\,\mathrm{tr}\left(V\left(\boldsymbol{\theta}\right)^\dagger U_*\right)\right\}^2 \tag{B.26}$$

can be evaluated up to a desired tolerance using statistical queries $\mathrm{qUSQ}\left(V\left(\boldsymbol{\theta}\right), \tau\right)$ and $\mathrm{qUSQ}\left(\mathrm{i}V\left(\boldsymbol{\theta}\right), \tau\right)$.

Our proofs expand on recent research showing hardness results in the SQ setting for certain quantum machine learning problems. More specifically, recent results that have shown that certain fundamental and rather simple classes of quantum "functions" are hard to learn in the SQ setting. Namely, (classical) output distributions of locally constructed quantum states [119] and the set of Clifford circuits [118] are hard to learn given properly chosen statistical query oracles. Following these results, we show that simple classes of functions generated by variational circuits are also exponentially difficult to learn in the SQ settings we consider. We also directly connect the statistical query setting to actual optimization algorithms that are used in practice for variational optimization. Our results indicate that training algorithms must be carefully constructed to avoid these poor lower bounds.

B.2.2 Limitations of Hardness Results in the SQ Framework

Though the SQ framework is a useful tool for analyzing the hardness of learning a class of functions in noisy settings, there are a few caveats and limitations of any hardness results proven in the SQ setting:

- The statistical query model inherently requires noise in the form of the tolerance τ. Furthermore, the guarantees of learning must handle worst case noise scenarios where the noise acts adversarially on the statistical query. Though

quantum variational algorithms are inherently noisy, this noise typically does not arise in an adversarial nature.

- The statistical query model places bounds on learning classes of functions using optimizers that query this SQ model and is not directly related to issues of loss landscapes since there is no loss landscape to actually optimize. Nevertheless, since (noisy) calculations of gradients and loss function values are themselves examples of statistical queries, any issues with optimizing over a loss landscape will also arise in performing the optimizer through a series of statistical queries.

- Learning every function in a class $\mathcal{C}$ can be restrictive, and in practice, one may only really want to learn a given function or a small set of functions. In fact, it can be shown that even the class of functions generated by shallow neural networks is hard to learn in the SQ setting [105, 110, 220–222]; nevertheless, neural networks are very successful at learning specific functions such as the classification of real-world images [42].

- Specific to the settings considered here, our hardness results were obtained in the correlational SQ setting by constructing a family of orthogonal functions drawn from a given function class. We chose this setting for its close relation to the algorithms used in practice for performing optimization over variational parameters. However, as mentioned in the main text, the correlational SQ setting is strictly weaker than the more general SQ setting, and separations between SQ and correlational SQ results have been made in prior work [223, 224].

B.3 Proofs of Statistical Query Results

Throughout this section, we make use of standard formulas from Weingarten calculus to integrate over Haar measure or t-designs [65, 225, 226]. Let $|I_m^n\rangle$ denote n copies

of the unnormalized maximally entangled state on a Hilbert space of dimension m:

$$|I_m^n\rangle = \sum_{i_1,i_2,\ldots,i_n=1}^{m} |i_1, i_2, \ldots, i_n\rangle\, |i_1, i_2, \ldots, i_n\rangle\,. \tag{B.27}$$

For $n = 2$, let $|S_m^2\rangle$ denote the same unnormalized state as above with a SWAP operation applied to the second register:

$$\begin{aligned} \left|S_m^2\right\rangle &= (I \otimes \mathrm{SWAP}) \left|I_m^2\right\rangle \\ &= \sum_{i_1,i_2=1}^{m} |i_1, i_2\rangle\, |i_2, i_1\rangle\,. \end{aligned} \tag{B.28}$$

The following hold over a distribution $\mathcal{D}$ that is a 2-design over the unitary matrices of dimension m:

$$\mathbb{E}_{U\sim\mathcal{D}}\left[U \otimes \bar{U}\right] = \frac{1}{m}\left|I_m^1\right\rangle\left\langle I_m^1\right|, \tag{B.29}$$

$$\mathbb{E}_{U\sim\mathcal{D}}\left[U \otimes U \otimes \bar{U} \otimes \bar{U}\right] = \frac{1}{m^2-1}\left(\left|I_m^2\right\rangle\left\langle I_m^2\right| + \left|S_m^2\right\rangle\left\langle S_m^2\right|\right) - \frac{1}{m(m^2-1)}\left(\left|I_m^2\right\rangle\left\langle S_m^2\right| + \left|S_m^2\right\rangle\left\langle I_m^2\right|\right), \tag{B.30}$$

where $\bar{U}$ denotes the matrix with entries that are the complex conjugate of entries of U.

As a simple example of applying the techniques above, we show that for unitaries U_* and V of dimension d^n (e.g., $d = 2$ for qubits and n is the number of qubits), $\mathbb{E}_{\rho\sim\mathcal{D}}\left[\mathrm{Re}\left\{\mathrm{tr}\left(U_*^\dagger V\rho\right)\right\}\right] = d^{-n}\,\mathrm{Re}\left\{\mathrm{tr}\left(U_*^\dagger V\right)\right\}$ whenever $\mathcal{D}$ forms a 1-design. This is a crucial formula that we use in the evaluation of statistical queries to qUSQ.

Lemma B.6. *For any distribution $\mathcal{D}$ that is a 1-design over states of dimension d^n,*

$$\mathbb{E}_{\rho\sim\mathcal{D}}\left[\mathrm{Re}\left\{\mathrm{tr}\left(W^\dagger V\rho\right)\right\}\right] = d^{-n}\,\mathrm{Re}\left\{\mathrm{tr}\left(W^\dagger V\right)\right\}. \tag{B.31}$$

Proof. WLOG, we rewrite the equation above in terms of a distribution over pure states and with a slight abuse of notation, we let $\mathcal{D}$ also denote a distribution over

unitary matrices U that forms a 1-design:

$$\mathbb{E}_{\rho\sim\mathcal{D}}\left[\mathrm{Re}\left\{\mathrm{tr}\left(W^\dagger V\rho\right)\right\}\right]=\mathbb{E}_{U\sim\mathcal{D}}\left[\mathrm{Re}\left\{\langle 0|\,U^\dagger W^\dagger VU\,|0\rangle\right\}\right]. \tag{B.32}$$

Using Equation (B.27), we have:

$$\mathbb{E}_{U\sim\mathcal{D}}\left[\mathrm{Re}\left\{\langle 0|\,U^\dagger W^\dagger VU\,|0\rangle\right\}\right]=\mathbb{E}_{U\sim\mathcal{D}}\left[\left\langle I^1_{d^n}\right|\left((W^\dagger V)\otimes I\right)\left(U\otimes\bar{U}\right)|0\rangle\,|0\rangle\right]. \tag{B.33}$$

Applying Equations (B.29) and (B.30), we have:

$$\begin{aligned}\mathbb{E}_{U\sim\mathcal{D}}\left[\mathrm{Re}\left\{\langle 0|\,U^\dagger W^\dagger VU\,|0\rangle\right\}\right]&=\mathbb{E}_{U\sim\mathcal{D}}\left[\mathrm{Re}\left\{\left\langle I^1_{d^n}\right|\left(\left(W^\dagger V\right)\otimes I\right)\left(U\otimes\bar{U}\right)|0\rangle\,|0\rangle\right\}\right]\\&=\frac{1}{d^n}\,\mathrm{Re}\left\{\left\langle I^1_{d^n}\right|\left(\left(W^\dagger V\right)\otimes I\right)\left|I^1_{d^n}\right\rangle\left\langle I^1_{d^n}|0\right\rangle|0\rangle\right\}\\&=\frac{1}{d^n}\,\mathrm{Re}\left\{\mathrm{tr}\left(W^\dagger V\right)\right\}.\end{aligned} \tag{B.34}$$

□

B.3.1 Proofs of Statistical Query Dimensions for Variational Function Classes

Proposition B.7 (SQ dimension for $L=1$ and fixed global measurement)**.** *Given n qubits, let $\mathcal{H}$ be the concept class containing functions $f:\mathbb{C}^{2^n}\to\mathbb{R}$ consisting of single qubit rotations followed by a global Pauli Z measurement, i.e. functions of the form*

$$f\left(|\psi\rangle;U_1,U_2,\ldots,U_n\right)=\langle\psi|\left(U_1^\dagger\otimes U_2^\dagger\otimes\ldots\otimes U_n^\dagger\right)\left(Z_1\otimes Z_2\otimes\ldots Z_n\right)\left(U_1\otimes U_2\otimes\ldots\otimes U_n\right)|\psi\rangle, \tag{B.35}$$

where $|\psi\rangle$ is the input to the function and $U_1,U_2,\ldots,U_n$ are the parameterized 1-qubit rotation operations on distinct qubits. Then, the concept class $\mathcal{H}$ has SQ dimension $\text{SQ-DIM}_{\mathcal{D}}\left(\mathcal{H}\right)\geq 3^n$ *under any distribution of states that forms a 2-design.*

Proof. The simple proof of this proposition relies on the fact that all Pauli operators are pairwise orthogonal for a 2-design, i.e. given two distinct Pauli operators P_1 and

P_2 then $\mathbb{E}_{\rho\sim\mathcal{D}}\left[\operatorname{tr}\left(P_1\rho\right)\operatorname{tr}\left(P_2\rho\right)\right]=0$. Therefore, we simply show that the concept class $\mathcal{H}$ is capable of producing any Pauli string not containing the identity.

To proceed, note that we can rewrite the function class as follows:

$$f\left(\left|\psi\right\rangle;U_1,U_2,\ldots,U_n\right)=\left\langle\psi\right|\left(U_1^\dagger Z_1U_1\right)\otimes\left(U_2^\dagger Z_2U_2\right)\otimes\ldots\otimes\left(U_n^\dagger Z_nU_n\right)\left|\psi\right\rangle. \quad \text{(B.36)}$$

To obtain any arbitrary Pauli string, we simply conjugate the Z_i operator for the ith qubit by a corresponding operation. If the ith qubit of a Pauli string is equal to X, then we set $U_i=H$ or the Hadamard transform. Similarly, if the ith qubit of a Pauli string is equal to Y, then we set $U_i=H\sqrt{Z}^\dagger$. By conjugation of the individual 1-qubit operators, we thus can produce any Pauli operator in $\{X,Y,Z\}^{\otimes n}$. □

Corollary B.8. *By application of Theorem B.5, the class of functions defined in Proposition B.7 consisting of a single layer of parameterized single qubit unitary gates and a fixed global measurement on n qubits requires* $2^{\Omega(n)}$ *queries to learn for a query tolerance greater than* $3^{-\beta n}$*, where* $\beta=1/2-\Omega\left(1\right)$.

Proposition B.9 (SQ dimension for $L=\lceil\log_2\left(n\right)\rceil$, 2-qubit gates, and single Pauli Z measurement)**.** *Given n qubits, let* $\mathcal{H}$ *be the concept class containing functions* $f:\mathbb{C}^{2^n}\to\mathbb{R}$ *consisting of* $\lceil\log_2\left(n\right)\rceil$ *layers of 2-qubit gates followed by a Pauli Z measurement on a single qubit. Then, the concept class* $\mathcal{H}$ *has SQ dimension* $\text{SQ-DIM}_{\mathcal{D}}\left(\mathcal{H}\right)\geq 4^n-1$ *under any distribution of inputs that forms a 2-design.*

Proof. We will show that $\mathcal{H}$ is powerful enough to perform any nontrivial Pauli measurement (i.e., any Pauli but the identity) and hence construct at least 4^n-1 orthogonal functions. Classically, any parity function can be constructed in $\lceil\log_2\left(n\right)\rceil$ layers, and we use a similar construction here.

Without loss of generality, assume the Pauli measurement is on the first qubit. Let $U\left(\theta\right)$ represent a possible unitary that can be applied using the given hypothesis class, resulting in a final measurement of $U\left(\theta\right)^\dagger Z_1U\left(\theta\right)$ on a given input state $\left|\psi\right\rangle$. We will show that we can parameterize the circuit such that for any Pauli measurement, $P_1\otimes P_2\otimes\ldots\otimes P_n=U\left(\theta\right)^\dagger Z_1U\left(\theta\right)$ where P_i indicates the Pauli operator of qubit i (i.e., $P_i\in\{I,X,Y,Z\}$).

To construct any Pauli operator $P_1 \otimes P_2 \otimes \ldots \otimes P_n$, we follow the steps below:

1. In the first layer, apply a unitary to each qubit i which maps the computational basis to the basis of the Pauli for qubit i. In more detail, if $P_i = I$ or $P_i = Z$, then apply the identity map to keep the basis the same. If $P_i = X$, then apply the Hadamard transform and if $P_i = Y$ then apply the operation $H\sqrt{Z}^\dagger$.

2. In the lth layer, apply a specific two qubit gate to qubit pairs:

$$\left\{1, 2^{l-1}+1\right\}, \left\{2\left(2^{l-1}\right)+1, 3\left(2^{l-1}\right)+1\right\}, \left\{4\left(2^{l-1}\right)+1, 5\left(2^{l-1}\right)+1\right\}, \ldots. \tag{B.37}$$

   For a layer l and a given pair $\{i, j\}$, apply the following gate:

   - if all of $P_i, P_{i+1}, \ldots, P_{j+2^{l-1}}$ are equal to I, then apply the identity.
   - if any of $P_i, P_{i+1}, \ldots, P_{j-1}$ are not equal to I and all of $P_j, P_{j+1}, \ldots, P_{j+2^{l-1}}$ are equal to I then apply the identity as well.
   - if all of $P_i, P_{i+1}, \ldots, P_{j-1}$ are equal to I and any of $P_j, P_{j+1}, \ldots, P_{j+2^{l-1}}$ are not equal to I, then apply a SWAP gate between qubits i and j.
   - otherwise, apply the following 2-qubit gate to i and j which conjugates $Z \otimes I$ to $Z \otimes Z$:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}. \tag{B.38}$$

3. Repeat step 2 from $l = 1$ to $l = \lceil \log_2(n) \rceil$. Measuring the first qubit will measure the corresponding desired Pauli. Note, that the single qubit operations of step 1 and the 2-qubit operations of step 2 can be combined into a single 2-qubit gate thus not changing the depth.

Following the steps above, at layer l, the measurement of the first qubit corresponds to the Pauli measurement of the first 2^l qubits. Recursively applying this procedure l layers produces any arbitrary Pauli string. □

Corollary B.10. *By application of Theorem B.5, the class of functions defined in Proposition B.9 consisting of* $\lceil \log_2(n) \rceil$ *2-qubit unitary gates and a fixed measurement on a single qubit requires* $2^{\Omega(n)}$ *queries to learn for a query tolerance greater than* $4^{-\beta n}$*, where* $\beta = 1/2 - \Omega(1)$*.*

Proposition B.11 (SQ dimension for L layers, neighboring 2-local gates in one-dimensional lattice and fixed single qubit measurement)**.** *Given n qubits, let* $\mathcal{H}$ *be the concept class containing functions* $f : \mathbb{C}^{2^n} \to \mathbb{R}$ *consisting of L layers of 2-qubit unitary operations followed by a Pauli Z measurement on a single qubit (labeled qubit m), i.e. functions of the form*

$$f(|\psi\rangle; W_1, W_2, \ldots, W_L) = \langle\psi| W_1^\dagger W_2^\dagger \ldots W_L^\dagger (Z_m) W_L \ldots W_2 W_1 |\psi\rangle, \tag{B.39}$$

where $|\psi\rangle$ *is the input to the function and* $W_1, W_2, \ldots, W_L$ *are the unitary operations at each layer consisting of tensor products of 2-local unitary operators acting on neighboring qubits. Then, the concept class* $\mathcal{H}$ *has SQ dimension* $\text{SQ-DIM}_{\mathcal{D}}(\mathcal{H}) \geq 4^{\min(2L,n)} - 1$ *under any distribution of states that forms a 2-design.*

Proof. Our proof relies on the fact that with L layers, one can conjugate the fixed single qubit measurement on qubit m to produce any $2L$-qubit Pauli on the $2L$ qubits within the reverse light cone of m. We follow a proof outline similar to Proposition B.9.

Before we proceed, we assume without loss of generality, that L is odd and the first layer applies a two qubit unitary to qubit m and the preceeding qubit $m-1$. It is straightforward to extend this to the case where L is even. Therefore, qubit m is the Lth qubit in the reverse light cone of qubit m, i.e., the light cone traverses qubits $m-L$ to $m+L-1$. For the steps below, we then index the qubits from $-L$ to $L-1$ so that the numbering is relative to qubit m. To perform a given $2L$ Pauli operator $P_{-L} \otimes P_{-L+1} \otimes \ldots \otimes P_{L-1}$ in the reverse light cone of qubit m, we follow the steps below, many of which are copied from Proposition B.9.:

1. In the first layer, apply a unitary to each qubit i which maps the computational basis to the basis of the Pauli for qubit i. In more detail, if $P_i = I$ or $P_i = Z$,

then apply the identity map to keep the basis the same. If $P_i = X$, then apply the Hadamard transform and if $P_i = Y$ then apply the operation $H\sqrt{Z}^\dagger$.

2. in the Lth layer, for the 2-qubit unitary acting on qubits 0 and -1, apply the following gate:

   - if all of $P_{-L}, P_{-L+1}, \ldots, P_{-1}$ are equal to I and all of $P_1, P_2, \ldots, P_{L-1}$ are equal to I, then apply the identity.
   - if any of $P_{-L}, P_{-L+1}, \ldots, P_{-1}$ are not equal to I and and any of $P_1, P_2, \ldots, P_{L-1}$ are not equal to I, apply the following 2-qubit gate (CNOT) to i and $i+1$ which conjugates $I \otimes Z$ to $Z \otimes Z$:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{B.40}$$

   - otherwise, apply the SWAP operation between qubits 0 and -1.

3. In the lth layer for any $l \neq L$, apply a specific two qubit gate to neighboring qubit pairs $\{-L+l-1, -L+l\}$ on the edge of the reverse light cone. For simplicity, let $i = -L+1-1$ and apply the following gate to qubit pair $\{i, i+1\}$:

   - if all of $P_{-L}, P_{-L+1}, \ldots, P_i$ are equal to I, then apply the identity.
   - if any of $P_{-L}, P_{-L+1}, \ldots, P_i$ are not equal to I and $P_{i+1} = I$, then apply a SWAP between qubits i and $i+1$.
   - otherwise, apply the following 2-qubit gate (CNOT) to i and $i+1$ which conjugates $I \otimes Z$ to $Z \otimes Z$:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{B.41}$$

Similarly, for the other edge of the reverse light cone, we apply the same gates, but in "reverse" logic. Here, we apply a 2-qubit unitary to qubit pair $\{L-l-2, L-l-1\}$. For simplicity, let $i = L-l-2$ and apply the following gate to qubit pair $\{i, i+1\}$:

- if all of $P_{i+1}, P_{i+2}, \ldots, P_{L-1}$ are equal to I, then apply the identity.
- if any of $P_{i+1}, P_{i+2}, \ldots, P_{L-1}$ are not equal to I and $P_{i+1} = I$, then apply a SWAP between qubits i and $i+1$.
- otherwise, apply the following 2-qubit gate to i and $i+1$ which conjugates $Z \otimes I$ to $Z \otimes Z$:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}. \tag{B.42}$$

4. Repeat step 3 from $l = 1$ to $l = L-1$. Measuring the mth qubit will measure the corresponding desired Pauli. Note that the single qubit operations of step 1 and the 2-qubit operations of step 2 can be combined into a single 2-qubit gate thus not changing the depth.

□

Corollary B.12. *By application of Theorem B.5, the class of functions defined in Proposition B.11 consisting of L layers of neighboring 2-qubit gates and a fixed measurement on a single qubit requires $2^{\Omega(\min(2L,n))}$ queries to learn for a constant query tolerance that does not depend on L and n.*

The above can be generalized to lower bound the statistical query dimension for circuits of L layers on d-dimensional lattices as we show below. In d-dimensional lattices, since the light cone of a single qubit measurement grows at a rate of L^d for an L layers, we can prove that the statistical query dimension grows as $2^{\Omega\left(\min\left(2L, n^{1/d}\right)^d\right)}$.

Proposition B.13 (SQ dimension for L layers, neighboring 2-local gates on d-dimensional lattice and fixed single qubit measurement)**.** *Given n qubits, let $\mathcal{H}$ be the concept class*

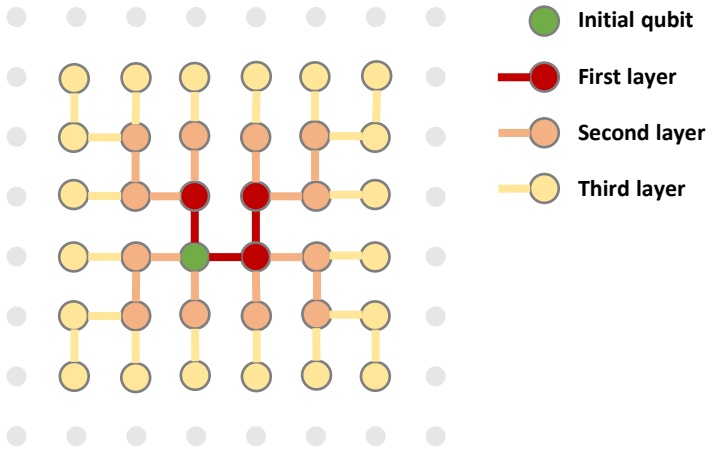


Figure B-1: Growth of the light cone for a 2-dimensional lattice, where the initial qubit is the one that is measured. The size of the lattice grows with the perimeter of the light cone for each layer which consists of local 2-qubit gates applied in each dimension. Each qubit is connected to a qubit in the edge of the light cone of the prior layer, forming a graph which is a tree rooted at the initial qubit.

containing functions $f : \mathbb{C}^{2^n} \to \mathbb{R}$ *consisting of* L *layers of* 2*-qubit unitary operations applied in each dimension followed by a Pauli* Z *measurement on a single qubit (labeled qubit* m*), i.e. functions of the form*

$$f\left(|\psi\rangle; W_1, W_2, \ldots, W_L\right) = \langle\psi| W_1^\dagger W_2^\dagger \ldots W_L^\dagger \left(Z_m\right) W_L \ldots W_2 W_1 |\psi\rangle, \tag{B.43}$$

where $|\psi\rangle$ *is the input to the function and* $W_1, W_2, \ldots, W_L$ *are the unitary operations at each layer consisting of tensor products of* 2*-local unitary operators acting along each dimension on neighboring qubits in a* d*-dimensional lattice. Then, the concept class* $\mathcal{H}$ *has SQ dimension* $\text{SQ-DIM}_{\mathcal{D}}\left(\mathcal{H}\right) = 2^{\Omega\left(\min\left(2L, n^{1/d}\right)^d\right)}$ *under any distribution of states that forms a* 2*-design.*

Proof. Our proof relies on the fact that with L layers, one can conjugate the fixed single qubit measurement on qubit m to produce any Pauli on the $\Omega\left(L^d\right)$ qubits within the reverse light cone of m. We follow a proof outline similar to Proposition B.11.

To be more precise, let us introduce some notation. To perform any Pauli mea-

surement in the reverse light cone at a given layer $l \in [L]$ indexed in reverse order, we apply gates to the perimeter of the reverse light cone at layer $l-1$. We assume there are N_l qubits in the reverse light cone at layer l and index these qubits from 1 to N_l to construct the Pauli $P_1 \otimes P_2 \otimes \ldots \otimes P_{N_l}$. Like in Proposition B.11, we grow the Pauli at each layer.

To grow the light cone and properly choose the 2-qubit gates, we construct a graph which is a tree where the parent of any qubit is the prior qubit which it was connected to in the light cone of the previous layer (see Figure B-1 for an example). The root of the tree is the qubit which is being measured. For example, at layer $l = 1$, the light cone is of size two in each dimension and the qubit being measured is the parent to the child node which it is connected to. To construct any pauli $P_1 \otimes P_2 \otimes \ldots \otimes P_{N_L}$, we follow the steps below:

1. In the lth layer, for all parent and child qubits p and c respectively connected in the tree at layer l, apply a unitary acting on qubits p and c as follows:

   - If all of the qubits that are descendants of qubit c and qubit c itself have Pauli terms that are equal to I, then apply the identity gate between qubit p and c.
   - if any of the qubits that are descendants of qubit c or qubit c itself have Pauli terms that are not equal to I and the Pauli term of qubit p is equal to I, then apply the SWAP gate between p and c.
   - otherwise, apply the following 2-qubit gate to qubits p and c which conjugates $Z \otimes I$ to $Z \otimes Z$:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}. \tag{B.44}$$

2. Repeat step 1 above from $l = 1$ to $l = L$.

3. In the first layer ($l = L$), apply a unitary to each qubit i which maps the computational basis to the basis of the Pauli for qubit i. In more detail, if

$P_i = I$ or $P_i = Z$, then apply the identity map to keep the basis the same. If $P_i = X$, then apply the Hadamard transform and if $P_i = Y$ then apply the operation $H\sqrt{Z}^\dagger$.

Following the above steps, measuring the single qubit will measure the corresponding desired Pauli. Note, that the single qubit operations of the first layer and the 2-qubit operations of that layer can be combined into a single 2-qubit gate thus not changing the depth.

In each layer, 2-qubit gates act along each dimension in some order. We can assume an ordering of the dimensions without loss of generality and assume that we apply gates along that dimension in order. After the first layer, the lattice has size 2 along each dimension. For each layer thereafter, the lattice grows by 2 qubits in each dimension (see Figure B-1 for an example). Therefore, the reverse light cone grows at a rate $\Omega\left(L^d\right)$. Since the light cone can be at most of size n (number of qubits), then the light cone is of size $2^{\Omega\left(\min\left(2L, n^{1/d}\right)^d\right)}$ for all L layers. □

Corollary B.14. *By application of Theorem B.5, the class of functions defined in Proposition B.13 consisting of L layers of neighboring 2-qubit gates and a fixed measurement on a single qubit requires $2^{\Omega\left(\min\left(2L, n^{1/d}\right)^d\right)}$ queries to learn for a query tolerance that decays no faster than $2^{\omega\left(\min\left(2L, n^{1/d}\right)^d\right)}$. For $L \ll n$, this is equal to $2^{\Omega\left(L^d\right)}$ for any constant query tolerance that does not depend on L or n.*

Proposition B.15 (SQ dimension for $L = 1$, unitary compiling, and single qubit gates)**.** *Given n qubits, let $\mathcal{H}$ be the concept class containing unitary transformations $V : \mathbb{C}^{2^n} \to \mathbb{C}^{2^n}$ consisting of single qubit rotations in a single layer*

$$V\left(|\psi\rangle, U_1, U_2, \ldots, U_n\right) = U_1 \otimes U_2 \otimes \ldots \otimes U_n |\psi\rangle, \tag{B.45}$$

where $|\psi\rangle$ is the input to the transformation and $U_1, U_2, \ldots, U_n$ are the parameterized 1-qubit operations. Then, the concept class $\mathcal{H}$ has SQ dimension $\text{SQ-DIM}_{\mathcal{D}}\left(\mathcal{H}\right) \geq 4^n$ under the qUSQ *model and any distribution $\mathcal{D}$ of inputs that is a 2-design.*

Proof. From Lemma B.6, we have that $\langle U, V\rangle_{\mathcal{D}} = 2^{-n} \operatorname{Re}\left\{\operatorname{tr}\left(U^\dagger V\right)\right\}$. With one

layer of single qubit unitary operations, any Pauli matrix can be constructed. Since $\mathrm{tr}\left(P_1 P_2\right) = 0$ for any two distinct Pauli matrices P_1 and P_2, there are at least 4^n matrices in $\mathcal{H}$ which are orthogonal under the inner product. □

Corollary B.16. *By application of Theorem B.5, the class of functions defined in Proposition B.15 consisting of a single layer of single qubit unitaries requires* $2^{\Omega(n)}$ *queries to learn for a query tolerance greater than* $4^{-\beta n}$*, where* $\beta < 1/2 - \Omega\left(1\right)$*.*

B.3.2 Swap Test via Statistical Queries

In the task of unitary compiling, one is given copies of states which are inputs and outputs of a target unitary transformation, and the goal is to learn the unitary transformation from those states. More formally, we aim to learn a unitary U_* given a distribution over inputs or a dataset of m state pairs $\left\{\left|\phi_i\right\rangle, U_* \left|\phi_i\right\rangle\right\}_{i\in[m]}$.

One means of measuring overlaps between states is via the swap test [218]. For pure states, $\left|\phi\right\rangle$ and $\left|\psi\right\rangle$, the swap test measures the fidelity $\left|\left\langle\phi|\psi\right\rangle\right|^2$. The measured register in the swap test outputs $\left|0\right\rangle$ with probability $1/2 + \left|\left\langle\phi|\psi\right\rangle\right|^2/2$ and $\left|1\right\rangle$ otherwise. As we show in Appendix B.2.1, this quantity can be calculated using queries to qUSQ. We use the helper lemma below to prove this fact.

Lemma B.17. *For any distribution* $\mathcal{D}$ *over pure states that is a* 2*-design on a Hilbert space of dimension* m*,*

$$\mathbb{E}_{\rho\sim\mathcal{D}}\left[\mathrm{tr}\left(U_*\rho U_*^\dagger V\rho V^\dagger\right)\right] = \frac{m^{-1}\left|\mathrm{tr}\left(V^\dagger U_*\right)\right|^2 + 1}{m+1}. \tag{B.46}$$

Proof. With a slight abuse of notation, we let $\mathcal{D}$ also denote a distribution over unitary matrices U that forms a 2-design:

$$\begin{aligned}\mathbb{E}_{\rho\sim\mathcal{D}}\left[\mathrm{tr}\left(U_*\rho U_*^\dagger V\rho V^\dagger\right)\right] &= \mathbb{E}_{U\sim\mathcal{D}}\left[\left\langle 0\right| U^\dagger V^\dagger U_* U\left|0\right\rangle \left\langle 0\right| U^\dagger U_*^\dagger V U\left|0\right\rangle\right] \\ &= \mathbb{E}_{U\sim\mathcal{D}}\left[\left|\left\langle 0\right| U^\dagger V^\dagger U_* U\left|0\right\rangle\right|^2\right].\end{aligned} \tag{B.47}$$

Using Equation (B.27) and Equation (B.28), we have

$$\begin{aligned}\mathbb{E}_{U\sim\mathcal{D}}\left[\left|\langle 0|\, U^\dagger V^\dagger U_* U\,|0\rangle\right|^2\right] &= \mathbb{E}_{U\sim\mathcal{D}}\left[\langle 0|\, U^\dagger V^\dagger U_* U\,|0\rangle\,\langle 0|\, U^\dagger U_*^\dagger V U\,|0\rangle\right] \\ &= \mathbb{E}_{U\sim\mathcal{D}}\left[\left\langle I_m^2\right|\left(\left(V^\dagger U_*\right)\otimes\left(U_*^\dagger V\right)\otimes I\otimes I\right)\left(U\otimes U\otimes\bar{U}\otimes\bar{U}\right)|0\rangle^{\otimes 4}\right].\end{aligned} \tag{B.48}$$

Applying Equations (B.29) and (B.30), we have that:

$$\begin{aligned}\mathbb{E}_{U\sim\mathcal{D}}\left[\left|\langle 0|\, U^\dagger V^\dagger U_* U\,|0\rangle\right|^2\right] =& \frac{1}{m^2-1}\left\langle I_m^2\right|\left(\left(V^\dagger U_*\right)\otimes\left(U_*^\dagger V\right)\otimes I\otimes I\right) \\ & \cdot\left(\left|I_m^2\right\rangle\left\langle I_m^2\right| + \left|S_m^2\right\rangle\left\langle S_m^2\right|\right)|0\rangle^{\otimes 4} \\ & - \frac{1}{m\left(m^2-1\right)}\left\langle I_m^2\right|\left(\left(V^\dagger U_*\right)\otimes\left(U_*^\dagger V\right)\otimes I\otimes I\right) \\ & \cdot\left(\left|I_m^2\right\rangle\left\langle S_m^2\right| + \left|S_m^2\right\rangle\left\langle I_m^2\right|\right)|0\rangle^{\otimes 4} \\ =& \frac{1}{m^2-1}\left(\operatorname{tr}\left(V^\dagger U_*\right)\operatorname{tr}\left(U_*^\dagger V\right) + \operatorname{tr}\left(V^\dagger U_* U_*^\dagger V\right)\right) \\ & - \frac{1}{m\left(m^2-1\right)}\left(\operatorname{tr}\left(V^\dagger U_*\right)\operatorname{tr}\left(U_*^\dagger V\right) + \operatorname{tr}\left(V^\dagger U_* U_*^\dagger V\right)\right) \\ =& \left(\frac{1}{m^2-1} - \frac{1}{m\left(m^2-1\right)}\right)\left(\left|\operatorname{tr}\left(V^\dagger U_*\right)\right|^2 + m\right) \\ =& \frac{m^{-1}\left|\operatorname{tr}\left(V^\dagger U_*\right)\right|^2 + 1}{m+1}.\end{aligned} \tag{B.49}$$

□

B.4 Shallow VQAs as Random Fields

B.4.1 Random Fields on Manifolds

Hardness results from barren plateaus or SQ models both intuitively arise from the exponential decay of quantities necessary to perform optimization. To analyze the shallow circuit setting beyond the SQ model—where such exponentially decaying quantities tend not to exist—we look toward models of variational loss landscapes as random fields on manifolds. This mirrors References [44, 45, 49] in studying the loss landscapes of machine learning models via mapping to certain random fields which

are easier to study analytically. As in Chapter 2, here we show that certain classes of variational loss functions of shallow quantum models converge in some limit to Wishart hypertoroidal random fields (WHRFs). Though in Chapter 2 we discuss (and derive) the loss landscapes of WHRFs in great detail, we give a brief review here such that our discussion is self-contained.

WHRFs in q variables are random fields on a specific tensor product embedding of the hypertorus $(S^1)^{\times q}$ in $\mathbb{R}^{2^q}$. More specifically, points on this embedding are described by the Kronecker product:

$$\boldsymbol{w} = \bigotimes_{i=1}^{q} \begin{pmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{pmatrix} \tag{B.50}$$

for angles $-\pi \leq \theta_i < \pi$. These random fields are then of the form:

$$F_{\text{WHRF}}(\boldsymbol{\theta}) = \boldsymbol{w}^{\intercal} \cdot \boldsymbol{J} \cdot \boldsymbol{w}, \tag{B.51}$$

where $\boldsymbol{J}$ is drawn from the normalized complex Wishart distribution $\mathcal{CW}_{2^q}(m, \boldsymbol{\Sigma})$ with m degrees of freedom. The complex Wishart distribution is a natural multivariate generalization of the gamma distribution, and is given by the distribution of the square of a complex Gaussian random matrix. Specifically, for $\boldsymbol{X} \in \mathbb{C}^{N \times m}$ a matrix with i.i.d. complex Gaussian columns with covariance matrix $\boldsymbol{\Sigma}$, the matrix

$$\boldsymbol{W} = \frac{1}{m} \boldsymbol{X} \cdot \boldsymbol{X}^{\dagger} \tag{B.52}$$

is normalized complex Wishart distributed with scale matrix $\boldsymbol{\Sigma}$ and m degrees of freedom. As discussed in Chapter 2, the loss landscapes of WHRFs exhibit a complexity phase transition governed by the overparameterization ratio

$$\gamma = \frac{q}{2m}, \tag{B.53}$$

where models with $\gamma \geq 1$ have local minima near the global minimum, and models with $\gamma \ll 1$ have local minima far from the global minimum. Thus, the degrees of

freedom parameter m plays a pivotal role in governing the loss landscapes of WHRFs: when q is much smaller than m, training is typically infeasible due to an abundance of "traps" in the training landscape. Our main result here is in demonstrating that even for certain shallow VQAs, the corresponding WHRF is such that $\gamma \ll 1$, and training is infeasible.

B.4.2 Shallow VQAs Converge in Distribution to WHRFs

As discussed informally in the main text, our goal is to demonstrate that certain distributions of shallow variational quantum algorithms (VQAs) weakly converge to Wishart hypertoroidal random fields (WHRFs). The distribution of local minima of WHRFs was shown in Chapter 2 to exhibit a phase transition in trainability, where underparameterized models are untrainable due to poor local minima, and overparameterized models exhibit local minima close to the global minimum (though may still be untrainable for other reasons, e.g. due to barren plateaus [63–65]).

Unlike the nonlocal ansatz case, here we are unable to show the full convergence in distribution of shallow local VQAs to WHRFs. Instead, we focus on the joint distribution of the loss function, gradient norm, and Hessian determinant, where the gradient and Hessian have been normalized by the number of parameters q in the reverse light cone of each term in the Pauli expansion of the problem Hamiltonian; by the parameter shift rule [112, 113], it is easy to see that this bounds the gradient norm and Hessian eigenvalues as q is large. The local minima results of Chapter 2 depend only on this joint distribution, and thus showing this convergence suffices for our purposes.

We now review the setup of the VQA loss functions we are considering. Throughout the course of this review, we will make various assumptions, particularly on the distribution of gates in the VQA ansatz and on the independence of various reverse light cones; we discuss these assumptions and whether or not they are reasonable in more detail at the end of this Section. As mentioned in the main text, we consider

optimizing VQAs on the problem Hamiltonian $H \neq 0$, which has Pauli decomposition:

$$H = \sum_{i=1}^{A} \alpha_i P_i. \tag{B.54}$$

WLOG, we assume here H is traceless, and that all $\alpha_i > 0$. To simplify our analysis, we will consider the case where the reverse light cone of each term $\alpha_i P_i$ in the Pauli decomposition of H is i.i.d. drawn from the same distribution of ansatzes, with the same parameter dependence. To make this more concrete, assume that the reverse light cone of each $\alpha_i P_i$ is of the form $V_i(\boldsymbol{\theta}) |\mathbf{0}\rangle$ where $\boldsymbol{\theta} \in \mathbb{R}^q$, and has support on a number $l \ll n$ of qubits. In this regime, we can scale and shift the loss landscape of the standard variational loss function

$$F(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}| H |\boldsymbol{\theta}\rangle \tag{B.55}$$

to be of the form:

$$F_{\mathrm{VQE}}(\boldsymbol{\theta}) = 1 - \lambda_0^{-1} \sum_{i=1}^{A} \alpha_i \langle \mathbf{0}| V_i(\boldsymbol{\theta})^\dagger P_i V_i(\boldsymbol{\theta}) |\mathbf{0}\rangle = 1 + \|\boldsymbol{\alpha}\|_1^{-1} \sum_{i=1}^{A} \alpha_i \langle \mathbf{0}| V_i^\dagger(\boldsymbol{\theta}) P_i V_i(\boldsymbol{\theta}) |\mathbf{0}\rangle, \tag{B.56}$$

where λ_0 is the ground state energy of H and $\boldsymbol{\alpha}$ is the vector of all α_i.

We assume that $V_i(\boldsymbol{\theta})$ is of the form:

$$V_i(\boldsymbol{\theta}) = W_i(\boldsymbol{\theta}) U_i, \tag{B.57}$$

where U_i are i.i.d. drawn from an ϵ-approximate t-design under the monomial measure on l qubits [124], where $\epsilon = \mathrm{O}(1)$. Note that in particular, though the total ansatz size n may be large, all potential scrambling of the ansatz may only happen locally, in regions of size $l \ll n$; in other words, these ansatzes are not expected to suffer from barren plateaus, particularly if $l = \mathrm{O}(\log(n))$ [64, 65]. W_i is composed of fixed parameterized rotations which we take WLOG to be of the form $R_{Y_a}(\theta_b) = \exp(-\mathrm{i}\theta_b Y_a)$ (where as previously mentioned, this parameter dependence is identical across all W_i), fixed gates, and potentially randomly chosen gates such that W_i itself

is a random field. For simplicity, we also assume that all θ_i are independent from one another (i.e. we are in the $r = 1$ regime of Chapter 2), and that each qubit in the reverse light cone has at least one parameterized gate. We also assume that the field $W_i(\boldsymbol{\theta})$ is rotationally invariant in θ_i.

We now give the formal statement and proof of the loss landscapes of local, shallow VQAs. First, the formal statement:

Theorem B.18 (Approximately locally scrambled variational loss functions converge to WHRFs)**.** *Let $p_{VQE,\boldsymbol{\theta}}$ be the joint distribution of the loss function of Equation* (B.56)*, its gradient norm, and the determinant of its Hessian at $\boldsymbol{\theta}$, where the gradient and Hessian are normalized by q. Let $p_{WHRF,\boldsymbol{\theta}}$ be the same for the WHRF:*

$$F_{WHRF}(\boldsymbol{\theta}) = m^{-1} \sum_{i,j=1}^{2^l} w_i J_{i,j} w_j \tag{B.58}$$

with $m = \frac{\|\boldsymbol{\alpha}\|_1^2}{\|\boldsymbol{\alpha}\|_2^2} 2^{l-1}$ degrees of freedom, where $\boldsymbol{J} \sim \mathcal{CW}_{2^l}(m, \boldsymbol{I}_{\boldsymbol{2^l}})$. Here, $\boldsymbol{w}$ are points on the hypertorus $(S^1)^{\times l}$ parameterized by $\tilde{\boldsymbol{\theta}}$, where $\tilde{\theta}_i$ is the sum of all θ_j on qubit i. We then have that $p_{VQE,\boldsymbol{\theta}}$ weakly converges to $p_{WHRF,\boldsymbol{\theta}}$, up to an error $\tilde{\mathrm{O}}\left(\mathrm{poly}\left(\frac{1}{t} + \epsilon + \exp(-l)\right)\right)$ in Lévy–Prokhorov distance.

As we previously mentioned, for technical reasons, we only prove the convergence of the joint distribution of the loss and certain functions of its first two derivatives. We emphasize once more that this does not affect our final conclusions, as all results on the local minima distribution of WHRFs given in Chapter 2 depend only on this joint distribution.

To prove Theorem B.18, we begin by showing that, up to terms that go to zero polynomially quickly as $\epsilon \to 0, t \to \infty$, one can WLOG consider ansatzes of the form of Equation (B.56) that are explicitly Haar random within each reverse light cone of size l.

Lemma B.19 (Approximate local scrambling bound on the loss function and its derivatives)**.** *Let $p_{VQE,\boldsymbol{\theta}}$ be the joint distribution described in Theorem B.18. Let $p_{Haar,\boldsymbol{\theta}}$ be the same, for U_i taken to be i.i.d. Haar random. We then have that $p_{VQE,\boldsymbol{\theta}}$*

weakly converges to $p_{Haar,\boldsymbol{\theta}}$, up to an error $\tilde{\mathrm{O}}\left(\mathrm{poly}\left(\frac{1}{t}+\epsilon\right)\right)$ in Lévy–Prokhorov distance.

Proof. Let $\phi_{\mathrm{VQE}}\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right)$ be the joint characteristic function of $p_{\mathrm{VQE},\boldsymbol{\theta}}$, and similarly $\phi_{\mathrm{Haar}}\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right)$. Since U_i are assumed to be i.i.d. ϵ-approximate t-designs under the monomial measure, for any moments $M_{\mathrm{VQE},\boldsymbol{\theta}}, M_{\mathrm{Haar},\boldsymbol{\theta}}$ of degree s of $p_{\mathrm{VQE},\boldsymbol{\theta}}, p_{\mathrm{Haar},\boldsymbol{\theta}}$, respectively, we have that:

$$\left|M_{\mathrm{VQE},\boldsymbol{\theta}} - M_{\mathrm{Haar},\boldsymbol{\theta}}\right| = \mathrm{O}\left(\epsilon \mathbf{1}\left[s \leq t\right] + \mathbf{1}\left[s > t\right]\right). \tag{B.59}$$

In particular, for all T sublinear in t,

$$\left|\phi_{\mathrm{VQE}}\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right) - \phi_{\mathrm{Haar}}\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right)\right| = \mathrm{O}\left(\epsilon\, \mathrm{poly}\left(T\right) + \frac{\left(3T\right)^t}{t!}\right) \tag{B.60}$$

for all $\boldsymbol{x}$ with $\left\|\boldsymbol{x}\right\|_\infty \leq T$. Similar inequalities hold for the partial derivatives of the joint characteristic functions. Therefore, there exists some $T = \Omega\left(\mathrm{poly}\left(\min\left(t, \frac{1}{\epsilon}\right)\right)\right)$ such that the second bound of Theorem 4 of [227] (with $m = \log\left(T\right)$) on the Lévy–Prokhorov distance is $\tilde{\mathrm{O}}\left(\mathrm{poly}\left(\frac{1}{t}+\epsilon\right)\right)$. □

Until now, we have considered ansatzes with generic parameter dependence. We now show that up to terms vanishing exponentially quickly in the reverse light cone size l, we can consider a canonical ansatz form WLOG.

Lemma B.20 (Canonical form for Hamiltonian agnostic variational loss functions)**.** *Let $p_{Haar,\boldsymbol{\theta}}$ be the joint distribution described in Lemma B.19. Let $p_{can,\boldsymbol{\theta}}$ be the same for the variational loss function*

$$F_{can}\left(\boldsymbol{\theta}\right) = \left\|\boldsymbol{\alpha}\right\|_1^{-1} \sum_{i=1}^{A} \alpha_i \left\langle \mathbf{0}\right| R\left(\boldsymbol{\theta}\right)^\dagger U_i^\dagger P_i U_i R\left(\boldsymbol{\theta}\right) \left|\mathbf{0}\right\rangle + 1, \tag{B.61}$$

where $R\left(\boldsymbol{\theta}\right)$ is the product of the parameterized rotations of Equation (B.57)*. We then have that $p_{Haar,\boldsymbol{\theta}}$ weakly converges to $p_{can,\boldsymbol{\theta}}$, up to an error $\tilde{\mathrm{O}}\left(\mathrm{poly}\exp\left(-l\right)\right)$ in Lévy–Prokhorov distance.*

Proof. Let us consider (generally mixed) moments involving random variables of the form:

$$K_{ij}\left(\boldsymbol{\theta_j}\right) = \left\langle \mathbf{0}\right| U_i^\dagger W_i\left(\boldsymbol{\theta_j}\right)^\dagger P_i W_i\left(\boldsymbol{\theta_j}\right) U_i \left|\mathbf{0}\right\rangle - \left\langle \mathbf{0}\right| \tilde{U}_{ij}^\dagger U_i^\dagger W_i\left(\boldsymbol{\theta_j}\right)^\dagger P_i W_i\left(\boldsymbol{\theta_j}\right) U_i \tilde{U}_{ij} \left|\mathbf{0}\right\rangle, \tag{B.62}$$

where $U_i, \tilde{U}_{ij}$ are i.i.d. Haar random on l qubits. By the asymptotic free independence of Haar random matrices from constant matrices, and the fact that

$$\operatorname{tr}\left(W_i\left(\boldsymbol{\theta_j}\right) P_i W_i\left(\boldsymbol{\theta_j}\right)^\dagger\right) = \operatorname{tr}\left(\left|\mathbf{0}\right\rangle\left\langle \mathbf{0}\right| - \tilde{U}_{ij}\left|\mathbf{0}\right\rangle\left\langle \mathbf{0}\right| \tilde{U}_{ij}^\dagger\right) = 0, \tag{B.63}$$

we have that any such moment is on the order of $\mathrm{O}\left(\operatorname{poly}\exp\left(-l\right)\right)$ [228]. In particular, it is easy to see that up to an error in Lévy–Prokhorov distance on this order, one can WLOG take $p_{\mathrm{can},\boldsymbol{\theta}}$ as if the gradient and Hessian components had i.i.d. $\boldsymbol{U_{ij}}$ rather than U_i—for instance, this follows identically to the proof of Lemma B.19 with $\epsilon = \mathrm{O}\left(\operatorname{poly}\exp\left(-l\right)\right)$. The result then follows from the unitary invariance of the Haar measure. □

We are now able to prove Theorem B.18, following essentially the same procedure as proving Theorem A.2.

Proof. By Lemmas B.19 and B.20, $p_{\mathrm{VQE},\boldsymbol{\theta}}$ weakly converges to $p_{\mathrm{Haar},\boldsymbol{\theta}}$ up to an error $\tilde{\mathrm{O}}\left(\operatorname{poly}\left(\frac{1}{t} + \epsilon + \exp\left(-l\right)\right)\right)$ in Lévy–Prokhorov distance. By Corollary 1 of Reference [181], this then proves weak convergence of $p_{\mathrm{VQE},\boldsymbol{\theta}}$ to the corresponding joint distribution of a weighted sum of WHRFs each with 2^{l-1} degrees of freedom, up to an additional error in Lévy–Prokhorov distance exponentially small in l. Weak convergence to $p_{\mathrm{WHRF},\boldsymbol{\theta}}$ then follows from a trivial generalization of Theorem A.2. □

Scope of results We now comment on the applicability of the results of Chapter 2 on the local minima distribution of WHRFs when Theorem B.18 holds. All analysis of the local minima distribution of WHRFs in Chapter 2 depends only on the joint distribution $p_{\mathrm{WHRF},\boldsymbol{\theta}}$, up to a change in normalization of the gradient and Hessian by l rather than q that does not contribute to the logarithmic asymptotics

(i.e. Theorem A.10) when $q \log(q) = \mathrm{o}(m)$. Thus, in the discussion of the main text, we take this as an extra assumption. Furthermore, we note that the analysis in the main text holds only up to shifts on the order of $\tilde{\mathrm{O}}\left(\mathrm{poly}\left(\frac{1}{t} + \epsilon + \exp(-l)\right)\right)$ in the joint distribution $p_{\mathrm{WHRF},\boldsymbol{\theta}}$, due to the rate of convergence of Theorem B.18. However, shifts on this order do not affect the conclusions of Chapter 2 for sufficiently large constant ϵ^{-1}, t. For completeness, we summarize this discussion and known results on the loss landscapes of WHRFs with the following Corollary:

Corollary B.21 (Shallow, local VQAs have poor loss landscapes)**.** *Let F_{VQE} be a local VQA loss function of the form of Equation* (B.56)*. Assume all coefficients α_i of the Pauli decomposition of H are $\Theta(1)$, and*

$$l \log(n) + q \log(q) = \mathrm{o}\left(2^l A\right). \tag{B.64}$$

Then $p_{VQE,\boldsymbol{\theta}}$ weakly converges to $p_{WHRF,\boldsymbol{\theta}}$ as in Theorem B.18, where the associated WHRF has a fraction superpolynomially small in n of local minima within any constant additive error of the ground state energy.

Proof. The result follows immediately by applying Theorem A.10 to Theorem B.18. □

Assumptions Let us now discuss in more detail the assumptions made in the course of proving Theorem B.18. First, we assume that at least some part of the ansatz circuit scrambles some local region around any measured observable; that is, we assume that the ansatz locally is an ϵ-approximate t design for sufficiently large ϵ^{-1}, t. It is known that shallow, local circuits dimensions exhibit this property, when 2-local Haar random gates are applied [124]; thus, in a practical sense, our results assume that the local gates in any distribution of ansatzes under consideration are approximately Haar random. This is a typical model of Hamiltonian agnostic ansatzes, where the ansatz is chosen independently from the problem Hamiltonian H; see for instance the discussion in the main text and the references therein. The inapplicability of this assumption to Hamiltonian informed ansatzes—particularly for highly symmetric problems—is

discussed in more detail in Section 3.5, where we review models that may not suffer from the poor trainability properties we show here.

Our other major assumption is the independence of the $V_i(\boldsymbol{\theta})$ (up to the repeated use of parameters). Of course, in practice this is almost never true, as otherwise variational optimization would proceed via optimizing each reverse light cone independently. However, given a problem Hamiltonian H and a shallow ansatz, one can consider a subset of Pauli operators in the Pauli decomposition of H such that their reverse light cones do not overlap. There is little reason to believe that the loss landscape of this simplified problem should be any more difficult to optimize over than the full problem. We therefore suspect that this assumption is little more than a technical requirement. A similar generalization one could consider is taking the parameters of each V_i to being almost entirely independent of one another (though not entirely independent, as one could then optimize each subproblem independently, and n would no longer an accurate measure of the size of the problem). However, in this regime we expect the "effective" overparameterization ratio γ to go as $\left(\frac{l}{2^l}\right)^A$, as the problem essentially reduces to simultaneously optimizing A loss functions. For $A \sim n$, for instance, this decays exponentially in n, and thus we believe that models of this form are also not trainable.

B.5 Additional Numerical Experiments

B.5.1 Teacher Student Learning With Checkerboard Ansatzes

One particular challenge with quantum variational learning is that an overparameterized model needs more parameters than the dimension of the quantum input state (exponential in the number of qubits), whereas classically, overparameterization with respect to the size of the data set typically suffices [44, 229, 230]. To illustrate this phenomenon, we consider learning states generated by random shallow checkerboard circuits (denoted the teacher circuit) using checkerboard circuits of the same or more depth (denoted the student circuit). The data set used to train the circuit consist of

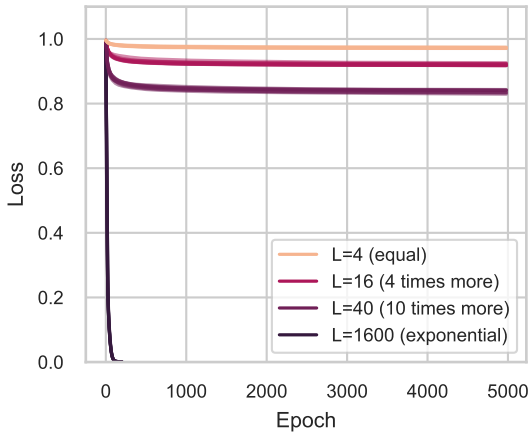


Figure B-2: Exponential depth is needed to overparamaterize a model to successfully learn a random circuit of the same form. Here, for each student circuit depth denoted by L, 10 randomly initialized 8 qubit student circuits are trained to learn a random $L = 4$ layer teacher circuit drawn from the same ansatz and parameter distribution.

512 pairs of inputs randomly drawn from computational basis states with their corresponding output state taken from applying the input state to the teacher circuit. We use the loss $\ell\left(|\psi\rangle, |\phi\rangle\right) = 1 - |\langle\psi|\phi\rangle|^2$ to measure the success of learning. Note that, though this is a global loss metric, gradients are analytically calculated to precision sufficient enough to obtain accurate values of the gradients for the relatively small number of qubits considered here.

As shown in Figure B-2, exponential depth (and number of parameters) is needed to always successfully learn the data generated by a shallow checkerboard circuit of 4 layers. We considered ansatzes only over 8 qubits, which is small enough to be able to feasibly overparameterize the models in our simulations. For fewer qubits and shallower circuits, we found that learning with equal numbers of qubits and layers was sometimes successful; but unsurprisingly, as we show in the main text, learning becomes much harder as qubits are added.

B.5.2 Random VQE Model

Here, we empirically analyze the performance of a layer-wise optimizer trained on the random VQE task in the main text. In Figure B-3, we train an 11 qubit ansatz using

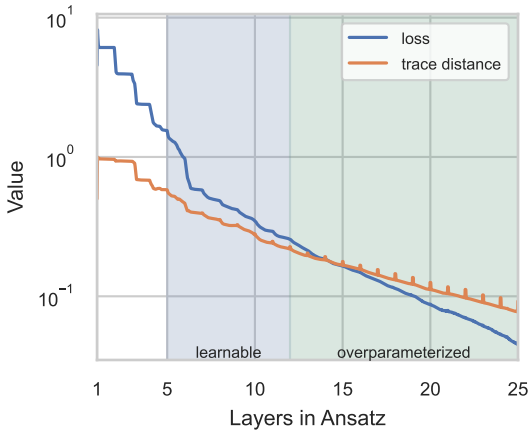


Figure B-3: When optimizing in a layer-wise fashion, the VQE algorithm converges to a local minimum at each layer until the overparameterized regime where the loss function steadily decreases regardless of the number of layers. Even in the learnable regime where the checkerboard ansatz is capable of expressing the global minima, the ansatz is still unable to find the correct parameters for this global minimum. Bumps in the loss function appear due to small instabilities in training immediately after adding a layer.

a layer-wise optimizer [231, 232], which initially trains a single layer of the ansatz and adds layers after every 5000 steps to continually add expressiveness. The target Hamiltonian H_t here has 4 layers of perturbations applied to it. Although layer-wise optimizers can avoid issues with barren plateaus [231], our numerical findings clearly show that this does not guarantee the algorithm will avoid traps in the landscape. After 5 layers, the ansatz has enough parameters to capably express the global optimum (denoted by the label "learnable"), but nevertheless stalls in optimizing to the ground state. Not until there are at least 12 layers, enough to overparameterize the ansatz with respect to the Hilbert space dimension, does learning smoothly converge to the globally optimal solution.

B.5.3 XYZ Hamiltonian Model

All of the numerical experiments performed elsewhere studied settings where the optimization was performed to numerical precision, two qubit gates were fully parameterized, and the existence of a global minimum at zero loss was guaranteed.

The analysis there focused on answering the question of whether convergence to the global minimum is empirically likely to be observed. To study a more realistic setting where such favorable conditions cannot be guaranteed but there still exists hope of some good convergence properties, we turn now to the problem of trying to variationally obtain the ground state of an approximately translationally invariant Heisenberg XYZ Hamiltonian [127]. Similar Hamiltonians have been studied and analyzed in previous works related to VQE [178, 233]. We perform experiments both with and without Gaussian noise added to the gradients to account for shot noise on a quantum computer.

The particular target Hamiltonian we aim to optimize is one where qubits are placed on a 2-dimensional grid and interaction terms take place between neighboring qubits. The Hamiltonian takes the form

$$H = \sum_i Z_i + \sum_{\langle i,j \rangle} \alpha_{ij} Z_i \otimes Z_j + \sum_{\langle i,j \rangle} \beta_{ij} \left(X_i \otimes X_j + 0.66 Y_i \otimes Y_j \right), \tag{B.65}$$

where $\langle i, j \rangle$ sums over the neighboring qubits i and j in the grid and α_{ij} and β_{ij} are random numbers drawn from the normal distribution with standard deviations set to 0.25 and means set to 1 and 3, respectively.

As shown in Table B.1, finding the ground state of the XYZ hamiltonian is in general challenging using the ansatz considered. For few layers, the ansatz is not expressible enough to find the target and converges to a poor critical point. For many layers, the VQE algorithm tends to converge to a better optimum, but issues with barren plateaus can begin to arise as indicated by the comparison in performance with assuming infinite shots versus finite shots.

B.6 Details of Numerical Experiments

All experiments were performed in Python using the PyTorch [234] package to perform automatic differentiation. Computation was performed on Nvidia RTX™ A6000 GPUs. Important hyperparameters for the experiments are listed in Table B.2. Un-

		energy error			trace distance		
	grid size	3×2	5×2	7×2	3×2	5×2	7×2
layers	shots						
3	10000	0.459	0.512	0.504	0.983	0.999	1.000
	400	0.456	0.501	0.392	0.983	0.999	1.000
	∞	0.466	0.512	0.395	0.981	0.999	1.000
9	10000	0.269	0.358	0.351	0.750	0.965	0.998
	400	0.343	0.434	0.386	0.845	0.991	0.994
	∞	0.245	0.350	0.344	0.659	0.924	0.993
15	10000	0.104	0.293	0.303	0.428	0.894	0.997
	400	0.180	0.356	0.318	0.577	0.965	0.987
	∞	0.054	0.244	0.251	0.293	0.842	0.968
21	10000	0.008	0.201	0.214	0.162	0.799	0.982
	400	0.043	0.277	0.247	0.269	0.882	0.984
	∞	0.011	0.178	0.162	0.151	0.747	0.933
27	10000	0.009	0.177	0.200	0.152	0.752	0.976
	400	0.034	0.254	0.251	0.244	0.882	0.976
	∞	0.010	0.122	0.129	0.136	0.663	0.948

Table B.1: Error in energy from the ground state (normalized by the magnitude of the ground state energy), and trace distance from the ground state, of a VQE optimizing the Heisenberg XYZ model. Results are averaged across 12 random initializations of the experiment for each entry in the table. Note the poor performance of VQE, particularly at the larger problem sizes.

less otherwise stated, all gradients were calculated using analytic formulas for automatic differentation with computer precision (32 bit floating point). Therefore, issues with decaying gradients and barren plateaus do not appear in these simulations for the relatively small number of qubits considered. Gradient based optimization was performed using vanilla gradient descent or the Adam optimizer [128], a popular and effective algorithm for training deep neural networks. We tested other optimizers as well and found no noticeable difference in performance. The processed data generated and analyzed for this study—as well as the code—are available at `https://github.com/bkiani/Beyond-Barren-Plateaus` and Reference [235].

Loss surface plot To generate this plot, we chart the loss landscape at initialization of training in the teacher-student setup of the main text for the 14 qubit QCNN circuit. The teacher and student circuit were both initialized as described in

Ansatz	Experiment	# Parameters	Optimizer	Learning Rate
QCNN	Teacher-Student	$16\lceil \log_2(n) \rceil = \mathrm{O}(\log(n))$	Adam	0.001
Checkerboard	Teacher-Student	$32L\left\lfloor \frac{n}{2} \right\rfloor = \mathrm{O}(nL)$	Adam	0.001 (underparameterized) 0.0001 (overparameterized)
	Random VQE (GD)	$128\left\lfloor \frac{n}{2} \right\rfloor = \mathrm{O}(n)$	vanilla GD	0.01
	Random VQE (Adam)	$128\left\lfloor \frac{n}{2} \right\rfloor = \mathrm{O}(n)$	Adam	0.003
	Adaptive VQE	$160L = \mathrm{O}(L)$	Adam	0.002 (5% reduction each layer)
XYZ ansatz	XYZ Hamiltonian VQE	$7\left\lfloor \frac{L}{3} \right\rfloor = \mathrm{O}(L)$	Adam	0.007 (halved every 1000 steps)

Table B.2: List of parameter counts, optimizers, and learning rates for the various ansatzes and experiments. L denotes the number of layers and n the number of qubits.

Appendix B.6.1.

The loss is plotted along two normalized directions of the parameter landscape. Normalization is applied individually to the 3 filters of the 14 qubit QCNN. We loosely follow the "filter-wise" normalization strategy of Reference [236], where we first generate a random direction by drawing a value for each parameter from an i.i.d. standard normal distribution. Then, we divide values for the parameters in a given layer by the Frobenius norm of the matrix for the corresponding layer.

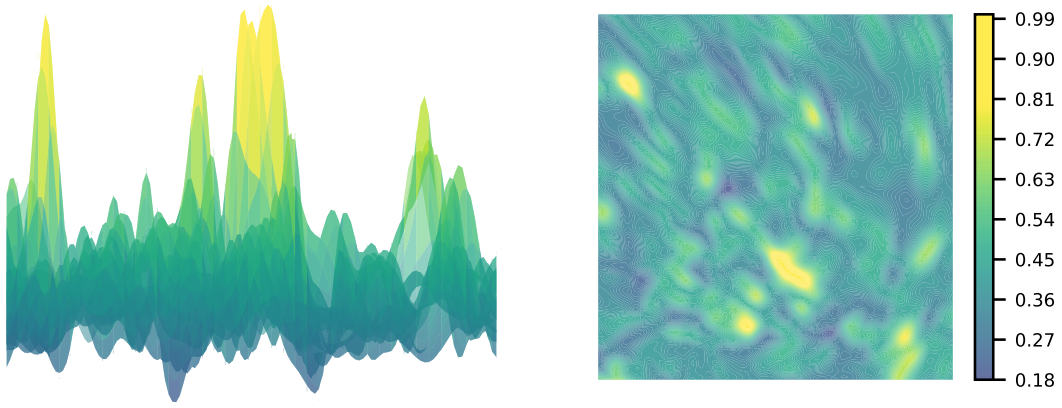


Figure B-4: Loss landscape of the QCNN experiment replicated from the main text, except the initialization of the student circuit is randomly chosen. Here, the global minimum is likely far away and the landscape also appears "bumpy"; all local minima in the region considered here are far from the global optimum.

In the loss surface plot of the main text, we plot the mean squared error loss for the teacher-student task for a batch size of 128 randomly chosen computational basis states. The legend in the plot is shown relative to the maximum value of the

loss in the range considered. A value of 0 here corresponds to the loss at the global minimum. The middle of the plot corresponds to the exact parameters of the teacher circuit, and hence, is a global minimum. This setting is, in a sense, an optimistic setting since initialization is near a global minimum. For comparison, we include in Figure B-4 an example of a loss surface where the student circuit is not initialized near the parameters of the teacher circuit. As is evident in this setting, no longer is there a global minimum in the parameter region considered, and the landscape also appears to be filled with traps.

B.6.1 QCNN Experiments

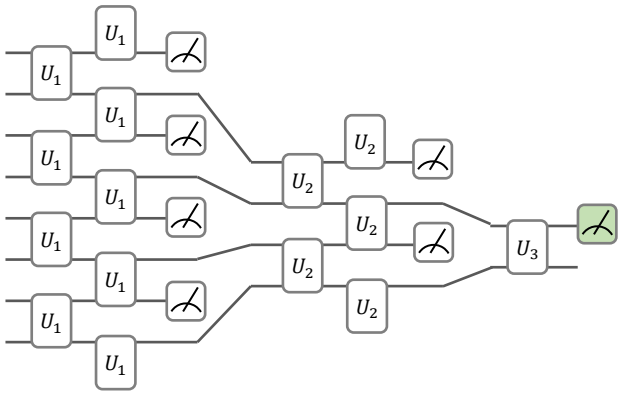


Figure B-5: Layers of shared 2-local unitary transformations are applied followed by measurement of every other qubit. Gates at the edge of the circuit above are applied in a cyclic fashion (i.e. the top and bottom qubit interact). The measurement colored in green is the measurement outcome whose probability we aim to predict in the teacher-student setup. Generically for n qubits, this ansatz has depth $\lceil \log_2(n) \rceil$. During training, the 2-local unitaries are fully parameterized for our simulations.

The quantum convolutional neural network (QCNN) is an ansatz originally proposed in Reference [108]. This ansatz features parameter sharing across gates in a single layer. The form of this circuit is provided in Figure B-5. In our experiments, we use the same form of the 2-local ansatz as in Reference [108] and also studied

in Reference [126]. Between convolutional layers, we include no controlled unitary operations based on the measurement outcomes. In learning settings, we fully parameterize the 2-local unitaries in the skew Hermitian basis of the unitary Lie algebra. To achieve this, we train directly over parameter entries of a matrix M and apply e^{H}, where $H = M - M^{\dagger}$, to perform the resulting unitary transformation. Entries of the matrix M were initialized i.i.d. from a standard normal distribution.

For the teacher-student experiments in the main text, we aim to predict the outcome of the final green measurement depicted in Figure B-5 for 512 randomly chosen computational basis states. For n qubits, the QCNN ansatz for both the teacher and student circuits have $16\left\lceil \log_2(n) \right\rceil$ parameters which is a relatively small number compared to the dimension of the Hilbert space. All networks were trained for 5000 epochs and a learning rate of 0.001 using the Adam optimizer.

B.6.2 Checkerboard Ansatz

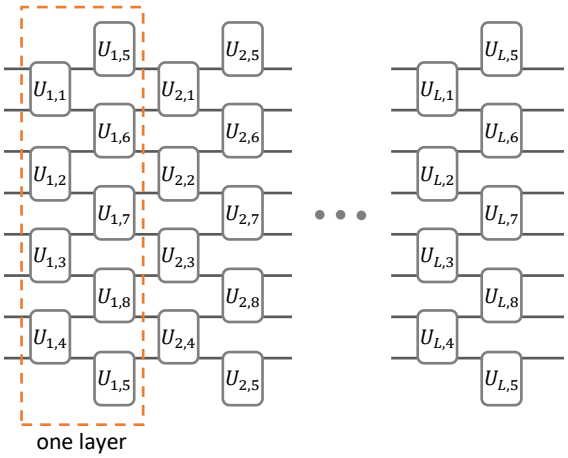


Figure B-6: Gates at the edge of the circuit above are applied in a cyclic fashion (i.e. the top and bottom qubit interact). Generically for n qubits, this ansatz has $32L\lfloor n/2 \rfloor$ parameters. During training, the 2-local unitaries are fully parameterized for our simulations.

The checkerboard circuit applies gates in a one-dimensional lattice as shown in Figure B-6. As in the QCNN experiments, we train directly over parameter entries

of a matrix M and apply e^H, where $H = M - M^\dagger$, to perform the resulting unitary transformation. Since the exponential map from the Lie algebra is surjective onto the unitary group, this parameterization is capable of expressing any unitary matrix. Entries of the matrix M were initialized i.i.d. from a standard normal distribution.

For the teacher-student simulations of the main text, we train networks over 512 randomly chosen computational basis states which is more than the dimension of the Hilbert space and enough information to recover the full unitary transformation. Optimization was performed using the Adam optimizer and a batch size of 128. Networks were trained for 5000 epochs and training was stopped if the loss fell below 0.001 which only occured for the overparameterized setting. We observed that for fewer than 8 qubits, training was successful with very small probability in the underparameterized setting.

B.6.3 VQE Experiments on Random Hamiltonians

For all of our VQE experiments, the target Hamiltonian H_t was constructed by conjugating a local Hamiltonian of n qubits equal to $\sum_{i=1}^n Z_i$ with alternating layers of products of 2-qubit unitaries U_1 and U_2. That is, H_t takes the form below as copied from the main text:

$$H_t = \left(U_2^\dagger U_1^\dagger\right)^L \left[\sum_{i=1}^n Z_i\right] (U_1 U_2)^L + nI. \tag{B.66}$$

U_1 and U_2 are the tensor product of 2-qubit unitaries which for n even take the form:

$$\begin{aligned} U_1 &= U_1^{(1,2)} \otimes U_1^{(3,4)} \otimes \ldots \otimes U_1^{(n-1,n)} \\ U_2 &= U_2^{(2,3)} \otimes U_2^{(4,5)} \otimes \ldots \otimes U_2^{(n,n+1)}, \end{aligned} \tag{B.67}$$

where superscripts above indicate the pair of qubits each 2-local unitary acts on and indexing is taken modulo n. Each 2-local unitary is drawn from the distribution e^H, where $H = G - G^\dagger$, and each G is a 4×4 matrix with entries drawn i.i.d. from a random normal distribution. Trained unitaries in the checkerboard ansatz are also initialized in this fashion. Optimization is then performed directly on the entries

of the matrix in the Lie algebra which form a complete basis for all of the 2-local unitaries.

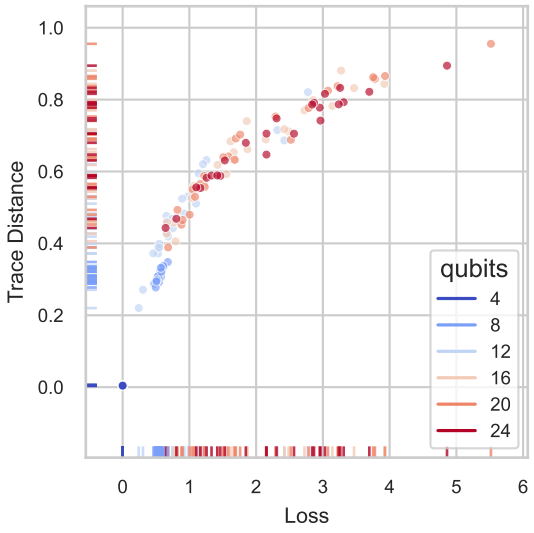


Figure B-7: Scatter plot showing the values of the loss and trace distance of the final VQE state after 30000 steps of optimization using the Adam optimizer shows that the algorithm converges to poorer local minima as the number of qubits grows. Setting is replicated from the VQE loss experiment from the main text, with the sole change of the optimizer from gradient descent to Adam.

In the VQE loss experiment of the main text, each VQE instance was optimized for 30000 steps using a vanilla gradient descent optimizer with a learning rate of 0.01. For completeness, we replicate this plot with the Adam optimizer in Figure B-7 and unsurprisingly observe similar convergence results. All calculations were performed to computer precision, which provides a best-case setting for optimization via real quantum hardware, since gradients and loss function values would have to be calculated using less precise sampling methods on actual quantum computers. In layer-wise VQE experiment of the main text, optimization is performed using an adaptive VQE algorithm similar to the one in Reference [232]. Here, a checkerboard ansatz is initialized as a single layer and optimization is performed layer-wise. We set $n = 11$ and small enough such that it is computationally feasible to overparameterize the ansatz. Each 5000 steps of optimization, a layer is added to the ansatz and initialized to the identity mapping. Each additional layer adds 160 trainable parameters to the

ansatz. After each layer is added, the learning rate is multiplied by 0.95 to make the training more stable with more parameters. At each point in time, all parameters of the ansatz across all layers are trained. For aesthetic purposes and to see the course of training without significant jumps in the plot, we plot a moving average of the values across 10 sequential datapoints in the main text.

B.6.4 VQE experiments on XYZ Hamiltonian

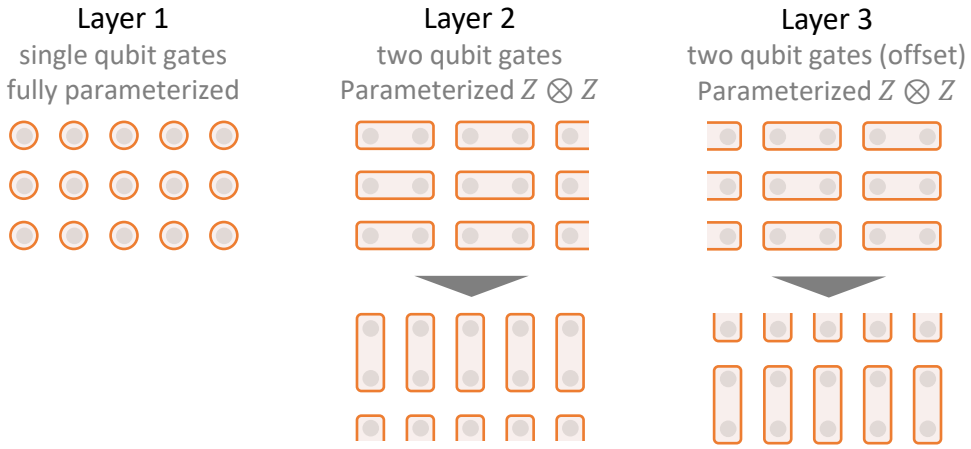


Figure B-8: Here, alternating blocks of three layers are composed onto each other. The first layer in each block is a fully parameterized single qubit gate. The next two layers are parameterized Pauli $Z \otimes Z$ terms to connect all neighboring qubits. Parameters are shared across a layer to better address the near translationally invariance in the model.

For the XYZ Hamiltonian model empirically analyzed in Appendix B.5.3, we implemented a gate-based ansatz which is fully parameterized in the single qubit gates and parameterized only with Pauli $Z \otimes Z$ terms for two qubit gates. The form of the ansatz is depicted in Figure B-8. Since the Hamiltonian of the model is approximately translationally invariant in both directions, we implemented sharing of parameters across a layer. Parameters were initialized as random normal variables. Each instance was optimized using the Adam optimizer [128] for 5000 steps. The learning rate was initially set to 0.007 and halved every 1000 steps. For calculations of the trace distance to the ground state, the ground state of the Hamiltonian was

calculated by performing an eigendecomposition of the complete Hamiltonian. To account for shot noise, random centered Gaussian noise with standard deviation equal to $1/\sqrt{\#\text{ shots}}$ was added to gradients with respect to the parameters.

B.7 Untrainability Beyond Gradient Descent

One may wish to avoid local minima by changing the loss function or performing more advanced versions of gradient-based optimizers. Here, we gives heuristic reasons why these two adjustments will likely not fix any issues of untrainability.

First, we examine changes in the loss function. This is commonly done to avoid barren plateaus and make gradients easier to compute. Let us assume that $\mathcal{L}(\boldsymbol{\theta})$ is our original loss function (as a function of the parameters $\boldsymbol{\theta}$), which is changed to a new loss function $\tilde{\mathcal{L}}(\boldsymbol{\theta})$. Typically, $\tilde{\mathcal{L}}(\boldsymbol{\theta})$ is chosen so that it upper and lower bounds $\mathcal{L}(\boldsymbol{\theta})$, i.e. $C\tilde{\mathcal{L}}(\boldsymbol{\theta}) \leq \mathcal{L}(\boldsymbol{\theta}) \leq D\tilde{\mathcal{L}}(\boldsymbol{\theta})$ for some constants C, D. This guarantees convergence in both metrics when changing the loss function and is the case for e.g. local versions of the inner product and the quantum earth mover's (EM) distance [121, 237]. Now, let us assume that every continuous path from a local minimum at $\boldsymbol{\theta_l}$ to the global minimum $\boldsymbol{\theta^*}$ must increase the loss function by a factor $M > D/C$, i.e. there exists a point in the path that has value at least $M\mathcal{L}(\boldsymbol{\theta_l})$. Then, in the new loss function $\tilde{\mathcal{L}}(\boldsymbol{\theta_l}) \leq \mathcal{L}(\boldsymbol{\theta_l})/C$. Furthermore, at some point in any continuous path, $\mathcal{L}(\boldsymbol{\theta}) > M\mathcal{L}(\boldsymbol{\theta_l})$ which implies that at that point $\tilde{\mathcal{L}}(\boldsymbol{\theta}) \geq \mathcal{L}(\boldsymbol{\theta})/D > M\mathcal{L}(\boldsymbol{\theta_l})/D = \mathcal{L}(\boldsymbol{\theta_l})/C$. Thus, $\boldsymbol{\theta_l}$ is not within a convex region around the global optimum. This may be too restrictive of an assumption since local minima can often be very shallow, but it also seems to be backed by experiments.

Second, we consider changing the optimization algorithm to a second order optimization algorithm such as in Reference [238]. These algorithms perform gradient descent by applying a transformation to the gradient of the form:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mu \boldsymbol{\Sigma}^{+} \boldsymbol{\nabla}_{\boldsymbol{\theta}_t} \mathcal{L}(\boldsymbol{\theta}_t), \tag{B.68}$$

where $\boldsymbol{\Sigma}$ incorporates our second order term, e.g. the Hessian or Fubini–Study metric tensor, and $\boldsymbol{\Sigma}^+$ is its pseudoinverse. Clearly, in the above, this does not allow one to escape a local minima. Setting $\boldsymbol{\theta}_t = \boldsymbol{\theta}^*$ above sets the gradient term to zero, and one again one is stuck in a local minimum.

Though other training methods exist, it is not clear *a priori* why they should succeed. For example, training in a layer-wise fashion also does not work as References [98, 121, 231] show. Finally, note that the above methods can be very effective at alleviating barren plateaus. In fact, changes in metric and second order optimization methods are often precisely designed to fix this issue. Nevertheless, these methods only provably converge to the global optimum in convex or close to convex settings, which is not the case for essentially all variational quantum models.

Appendix C

Technical Details for Chapter 4

C.1 The Schur Basis

In the presence of permutation invariance, the action of operations can be fully understood by analyzing a much smaller subspace of the larger Hilbert space. To precisely understand the form of that subspace, we turn to the Schur–Weyl decomposition of n qubits into subspaces corresponding to irreducible representations of the symmetric and unitary groups labeled by Young diagrams. Schur–Weyl duality offers a means to perform this decomposition by considering the natural representations of the permutation group and n-fold unitary group acting on n qubits [150, 239]. To describe the Schur basis and the resulting Schur transform, first we note the natural action of a permutation operation $R(\pi)$ acting on qubits:

$$R(\pi)\left|i_1\right\rangle \otimes \left|i_2\right\rangle \otimes \cdots \otimes \left|i_n\right\rangle = \left|i_{\pi^{-1}1}\right\rangle \otimes \left|i_{\pi^{-1}2}\right\rangle \otimes \cdots \otimes \left|i_{\pi^{-1}1}\right\rangle \tag{C.1}$$

as in the main text.

Similarly, a unitary $U \in \mathrm{U}(2)$ acting as the n-fold product $Q(U)$ takes the form

$$Q(U)\left|i_1\right\rangle \otimes \left|i_2\right\rangle \otimes \cdots \otimes \left|i_n\right\rangle = U\left|i_1\right\rangle \otimes U\left|i_2\right\rangle \otimes \cdots \otimes U\left|i_n\right\rangle. \tag{C.2}$$

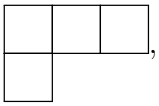
Schur–Weyl duality takes advantage of the fact that $Q(\cdot)$ and $R(\cdot)$ are each others'

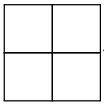
commutants, stating that the subspace of $(\mathbb{C}^2)^{\otimes n}$ decomposes as

$$Q(U)\,R(\pi) \cong \bigoplus_{\lambda} \rho_\lambda(U) \otimes \sigma_\lambda(\pi), \tag{C.3}$$

where λ runs over the set of partitions of n into at most two elements, and $\rho_\lambda(\cdot)$ and $\sigma_\lambda(\cdot)$ are irreducible representations of the unitary group $\mathrm{U}(2)$ and the symmetric group S_n, respectively. Note that irreps of both of these groups are indexed by partitions. More generally, for the space $\left(\mathbb{C}^d\right)^{\otimes n}$ of n qudits of dimension d, the λ would span over partitions of n into at most d elements. Partitions can equivalently be enumerated by Young diagrams. For example for the setting of 4 qubits, we have the 3 Young diagrams below that appear in the decomposition above:

$\lambda = (4,0)$:  ,

$\lambda = (3,1)$:  ,

$\lambda = (2,2)$:  .

A consequence of the above is that there exists a basis indexed by $|\lambda, q_\lambda, p_\lambda\rangle$ called the *Schur basis* where the actions of $Q(\cdot)$ and $R(\cdot)$ are separated [150]:

$$Q(U)\,|\lambda, q_\lambda, p_\lambda\rangle = \rho_\lambda(U)\,|\lambda, q_\lambda, p_\lambda\rangle, \tag{C.4}$$

$$R(\pi)\,|\lambda, q_\lambda, p_\lambda\rangle = \sigma_\lambda(\pi)\,|\lambda, q_\lambda, p_\lambda\rangle, \tag{C.5}$$

where we have implicitly projected onto the subspace indexed by λ. Here, $\rho_\lambda(U)$ and $\sigma_\lambda(\pi)$ act only on the q_λ and p_λ space, respectively. $\rho_\lambda(U)$ and $\sigma_\lambda(\pi)$ are respectively the linear transformations corresponding to the irreducible representations of $\mathrm{U}(2)$ and S_n for the irreducible representation indexed by λ. The above also presents a useful fact about permutation invariance. Namely, such an operation will act in-

variantly on the permutation register $|p_\lambda\rangle$ thus significantly reducing the degrees of freedom of a problem. The Schur transform U_{Sch} is a unitary transformation that acts as a change of basis from the computational to the Schur basis described above. The Schur transform can be efficiently implemented on a quantum computer running in time $\mathrm{O}\left(n \operatorname{poly}\left(d, \log\left(n\right), 1/\epsilon\right)\right)$ for error ϵ on qudit systems of dimension d [150]. We follow the notation of Reference [150]:

$$|\lambda, q_\lambda, p_\lambda\rangle = \sum_{i_1,i_2,\ldots,i_n=0}^{d-1} [U_{\text{Sch}}]_{i_1,i_2,\ldots,i_n}^{\lambda,q_\lambda,p_\lambda} |i_1\rangle |i_2\rangle \cdots |i_n\rangle . \tag{C.6}$$

As noted in the main text, the total degrees of freedom reduces to $\binom{n+3}{3}$ in settings with permutation invariance. To see this, note that the dimension of the $|q_\lambda\rangle$ register for a partition (a, b) is equal to $a - b + 1$. Therefore, we have

$$\text{DOF} = \sum_{k=0}^{\lfloor n/2 \rfloor} \left(2k + 1 + n - 2\left\lfloor \frac{n}{2} \right\rfloor\right)^2 = \binom{n+3}{3} \tag{C.7}$$

degrees of freedom. A similar calculation can be performed via a stars-and-bars counting argument. The above is also enumerated by the tetrahedral numbers [240].

To expand and manipulate individual basis states indexed by the $|q_\lambda\rangle$ register, one can use the Young symmetrizer Π_{p_λ} to project onto an explicit basis for each λ [150, 241]. Here, p_λ is a particular Young tableau for the Young diagram λ. The Young symmetrizer projects onto a subspace isomorphic to the subspace spanned by $|q_\lambda\rangle$:

$$\Pi_{p_\lambda} = \frac{\dim\left(\lambda\right)}{n!} \left(\sum_{c \in \text{Col}(p_\lambda)} \operatorname{sgn}\left(c\right) R\left(c\right)\right) \left(\sum_{r \in \text{Row}(p_\lambda)} R\left(r\right)\right), \tag{C.8}$$

where $\text{Row}\left(p_\lambda\right)$ and $\text{Col}\left(p_\lambda\right)$ are the set of permutations which permute integers within only rows and columns of the Young tableau p_λ, respectively [150, 241]. An example of the basis found via application of the Young symmetrizer is shown in Figure C-1. Throughout our study, we consider the Young tableau formed by filling entries in order first column-wise and then row-wise to be the “canonical" basis that

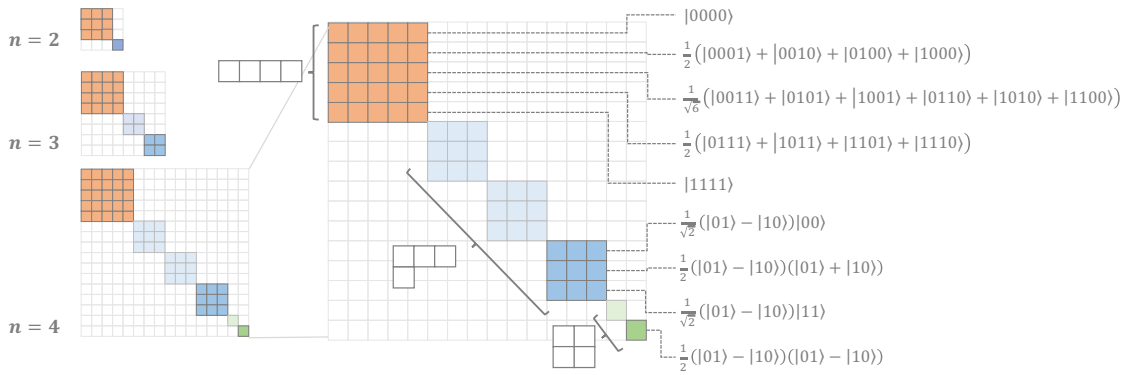


Figure C-1: Graphical depiction of Schur decomposition for $n = 4$ qubits. There are three Young diagrams of at most two rows for 4 qubits. Due to the presence of permutation invariance, we can restrict attention to the darker colored subspaces which correspond to a single subspace over the multiplicity of the permutation irreps. To project onto this darker colored subspace, we use the Young symmetrizer (Equation (C.8)).

we study. As an example, for 4 qubits, there are the following Young tableaus in our "canonical" basis:

$$\begin{array}{|c|c|c|c|}\hline 1&2&3&4\\\hline\end{array}, \qquad \begin{array}{|c|c|c|}\hline 1&3&4\\\hline 2\\\cline{1-1}\end{array}, \qquad \begin{array}{|c|c|}\hline 1&3\\\hline 2&4\\\hline\end{array}. \tag{C.9}$$

C.2 Structure Coefficients of X for Qubit Permutation Invariance

In this Appendix we evaluate the structure constants of the algebra X of operators symmetric under the action of permutation operators on qubits.

Lemma C.1. *The structure coefficients $X_{\boldsymbol{k}}^{\boldsymbol{i,j}}$ of the completely symmetrized Pauli representation are given by:*

$$X_{\boldsymbol{k}}^{\boldsymbol{i,j}} = \sum_{\substack{\{f_{ab}\}_{a,b\in\{1,x,y,z\}} \\ /Equation(\text{C.11}),Equation(\text{C.12})}} \frac{k_1!}{f_{11}!f_{xx}!f_{yy}!f_{zz}!}\frac{k_x!}{f_{1x}!f_{x1}!f_{yz}!f_{zy}!}\frac{k_y!}{f_{1y}!f_{y1}!f_{xz}!f_{zx}!}\frac{k_z!}{f_{1z}!f_{z1}!f_{xy}!f_{yx}!} \times \mathrm{i}^{f_{xy}+f_{yz}+f_{zx}}\,(-\mathrm{i})^{f_{yx}+f_{xz}+f_{zy}} \tag{C.10}$$

where the variables in the sum are non-negative integers subject to the constraints

$$\sum_{a\in\{1,x,y,z\}} f_{ab} = j_b, \qquad \sum_{b\in\{1,x,y,z\}} f_{ab} = i_a, \tag{C.11}$$

and

$$\begin{aligned} f_{11} + f_{xx} + f_{yy} + f_{zz} &\equiv k_1, \\ f_{1x} + f_{x1} + f_{yz} + f_{zy} &= k_x, \\ f_{1y} + f_{y1} + f_{xz} + f_{zx} &= k_y, \\ f_{1z} + f_{z1} + f_{xy} + f_{yx} &= k_z. \end{aligned} \tag{C.12}$$

Proof. For calculating the structure constants X, we first note that

$$A_{\boldsymbol{j}} = \frac{1}{i_1!i_x!i_y!i_z!} \sum_{\pi\in \mathrm{S}_n} R(\pi) \left(\sigma_1^{\otimes i_1} \otimes \sigma_x^{\otimes i_x} \otimes \sigma_y^{\otimes i_y} \otimes \sigma_z^{\otimes i_z}\right) R^{-1}(\pi) = \sum_{p_{\boldsymbol{i}}\in P_{\boldsymbol{i}}} p_{\boldsymbol{i}}, \tag{C.13}$$

where $P_{\boldsymbol{i}}$ is the set of Pauli words with i_a times σ_a for $a \in \{1, x, y, z\}$. Now we evaluate the product:

$$A_{\boldsymbol{i}} \cdot A_{\boldsymbol{j}} = \sum_{p_{\boldsymbol{i}}\in P_{\boldsymbol{i}}, p_{\boldsymbol{j}}\in P_{\boldsymbol{j}}} p_{\boldsymbol{i}} p_{\boldsymbol{j}} = \sum_{\boldsymbol{k}, p_{\boldsymbol{k}}\in P_{\boldsymbol{k}}} \sum_{\substack{p_{\boldsymbol{i}}\in P_{\boldsymbol{i}}, p_{\boldsymbol{j}}\in P_{\boldsymbol{j}}: \\ p_{\boldsymbol{i}}p_{\boldsymbol{j}} = \alpha_{p_{\boldsymbol{i}},p_{\boldsymbol{j}},p_{\boldsymbol{k}}} \cdot p_{\boldsymbol{k}}}} \alpha_{p_{\boldsymbol{i}},p_{\boldsymbol{j}},p_{\boldsymbol{k}}} p_{\boldsymbol{k}}. \tag{C.14}$$

This is a sum over products of exponentionally many Pauli words. The idea to evaluate this is that many of the summands have equal value, so it suffices to sum over a few different values multiplied by the number of summands with that value.

For every summand, define the subsets of qubits L_{ab} for $a, b \in \{1, x, y, z\}$,

$$L_{ab} \equiv \{l : (p_{\boldsymbol{i}})_l = \sigma_a, (p_{\boldsymbol{j}})_l = \sigma_b, 0 \le l < n\}, \tag{C.15}$$

and let

$$f_{ab} \equiv |L_{ab}| \tag{C.16}$$

be the numbers of elements in those subsets. Since every Pauli operator i_l at a qubit l is paired with some other Pauli operator j_l, f_{ab} fulfill the constraints in Equa-

tion (C.11). The multiplication algebra of Pauli operators directly implies Equation (C.12).

Let us now count how many Pauli words there are in the sum for a fixed set of numbers f_{ab} and a fixed resulting Pauli word $p_{\boldsymbol{k}}$. Every triple $p_{\boldsymbol{i}}$, $p_{\boldsymbol{j}}$, $p_{\boldsymbol{k}}$ corresponds to a decomposition of each $\boldsymbol{k}_c$-element set of qubits $\{l : (p_{\boldsymbol{k}})_l = \sigma_c\}$ for $c \in \{1, x, y, z\}$ into four subsets L_{ab} for the four different combinations $a, b \in \{1, x, y, z\}$ with $\sigma_a \sigma_b \propto \sigma_c$ under the Pauli algebra. For each c, the number of decompositions into the corresponding four subsets is given by

$$\frac{\boldsymbol{k}_c!}{\prod_{a,b:\sigma_a\sigma_b\propto\sigma_c} f_{ab}!}. \tag{C.17}$$

In total, the number of decompositions into four subsets for different c is given by

$$\frac{k_1!}{f_{11}!f_{xx}!f_{yy}!f_{zz}!}\frac{k_x!}{f_{1x}!f_{x1}!f_{yz}!f_{zy}!}\frac{k_y!}{f_{1y}!f_{y1}!f_{xz}!f_{zx}!}\frac{k_z!}{f_{1z}!f_{z1}!f_{xy}!f_{yx}!}. \tag{C.18}$$

Finally, the prefactor $\alpha_{p_{\boldsymbol{i}},p_{\boldsymbol{j}},p_{\boldsymbol{k}}}$ in Equation (C.14) only depends on the f_{ab}. Using the Pauli algebra,

$$\begin{aligned} &\sigma_x\sigma_y = \mathrm{i}\sigma_z \quad \sigma_y\sigma_z = i\sigma_x \quad \sigma_z\sigma_x = \mathrm{i}\sigma_y \\ &\sigma_y\sigma_x = -\mathrm{i}\sigma_z \quad \sigma_z\sigma_y = -\mathrm{i}\sigma_x \quad \sigma_x\sigma_z = -\mathrm{i}\sigma_y, \end{aligned} \tag{C.19}$$

it is given by

$$\alpha_{p_{\boldsymbol{i}},p_{\boldsymbol{j}},p_{\boldsymbol{k}}} = \mathrm{i}^{f_{xy}+f_{yz}+f_{zx}}\,(-\mathrm{i})^{f_{yx}+f_{xz}+f_{zy}}. \tag{C.20}$$

Using Equation (C.18) and Equation (C.20) in Equation (C.14) directly yields Equation (C.10). □

Let us quickly discuss the complexity of the computation of $X_{\boldsymbol{k}}^{\boldsymbol{i},\boldsymbol{j}}$. In the summation of Equation (C.10), we sum over 16 variables within a range of the order n, so if we naively evaluate the sum, we already obtain a polynomial runtime $\mathrm{O}(n^{16})$. However, due to the constraint Equation (C.11), we can reduce the summation to only 9 variables f_{ab} with $a, b \in \{x, y, z\}$. Equation (C.12) poses another three independent constraints, reducing the summation to 6 variables. Thus, an individual entry $X_{\boldsymbol{k}}^{\boldsymbol{i},\boldsymbol{j}}$ can be calculated in $\mathrm{O}(n^6)$ runtime, whereas all $\mathrm{O}(n^9)$ coefficients together take

runtime $\mathrm{O}\left(n^{15}\right)$.

Note that this is only the runtime for a naive evaluation of the sum in Equation (C.10). It seems likely that the runtime $\mathrm{O}\left(n^{6}\right)$ for the evaluation of a single coefficient can be reduced to a smaller exponent. We will leave this open to further investigation.

C.3 Irrep Basis of A for Qubit Permutation Invariance

In this section, we compute the matrix elements $F^{\boldsymbol{i},\lambda}_{q_\lambda,q'_\lambda}$ from the main text for the case of S_n action on n qubits by permutation. To this end, we first find the irrep basis $|\lambda, q_\lambda, p_{\lambda 0}\rangle$ where $p_{\lambda 0}$ is a standard choice of Young tableau, and then consider the representation A in this basis.

Lemma C.2. *Recalling that λ is given by a Young diagram, we choose $p_{\lambda 0}$ to be the standard Young tableaux for that diagram, with numbers increasing first in the column direction and then in row direction, as shown in Equation* (C.9)*. Then the tensor components $F^{\boldsymbol{i},\lambda}_{q_\lambda,q'_\lambda}$ discussed in the main text for the completely symmetrized Pauli representation are given by:*

$$F^{\boldsymbol{i},\lambda}_{q_\lambda,q'_\lambda} = \sum_{\substack{f_{11},f_{xx},f_{yy},f_{zz},\\ g_{010},g_{111},g_{0x1},g_{1x0},\\ g_{0y1},g_{1y0},g_{0z0},g_{1z1}\\ /Equation\ (\mathrm{C}.22)}} \frac{1}{\sqrt{\binom{n-2\lambda_1}{q_\lambda}\binom{n-2\lambda_1}{q'_\lambda}}} \mathrm{i}^{2f_{xx}+2f_{yy}+2f_{zz}+2g_{1z1}-g_{0y1}+g_{1y0}} \cdot \frac{\lambda_1!\,(n-2\lambda_1)!}{f_{11}!f_{xx}!f_{yy}!f_{zz}!g_{010}!g_{111}!g_{0x1}!g_{1x0}!g_{0y1}!g_{1y0}!g_{0z0}!g_{1z1}!}, \tag{C.21}$$

where the sum is over a set of 12 non-negative integers fulfilling the constraints

$$\begin{aligned}
g_{010} + g_{0z0} + g_{0x1} + g_{0y1} &= n - 2\lambda_1 - q_\lambda, \\
g_{010} + g_{0z0} + g_{1x0} + g_{1y0} &= n - 2\lambda_1 - q'_\lambda, \\
g_{111} + g_{1z1} + g_{1x0} + g_{1y0} &= q_\lambda, \\
g_{111} + g_{1z1} + g_{0x1} + g_{0y1} &= q'_\lambda, \\
2f_{11} + g_{010} + g_{111} &= i_1, \\
2f_{xx} + g_{0x1} + g_{1x0} &= i_x, \\
2f_{yy} + g_{0y1} + g_{1y0} &= i_y, \\
2f_{zz} + g_{0z0} + g_{1z1} &= i_z
\end{aligned} \tag{C.22}$$

and λ_1 is the length of the second row of λ.

Proof. Following the previous section, we can project onto the space with an S_n irrep λ and a fixed multiplicity label $p_{\lambda 0}$ using the Young symmetrizer in Equation (C.8). Acting with the Young symmetrizer on a computational basis state yields a superposition of basis states with the same number of 0s and 1s. Let us write $\lambda = (\lambda_0, \lambda_1)$ for the lengths of the first and second row of λ. Then we see that applying the Young symmetrizer yields 0 unless the number of 1s is between λ_1 and λ_0. This is because the row symmetrizer does not change the number of 1s, and the antisymmetrizer on λ_1 length-2 columns yields 0 if any columns are 00 or 11. Thus, the irrep basis states can be obtained by applying the Young symmetrizer to states with $\lambda_1 + q_\lambda$ ones, where $0 \leq q_\lambda \leq n - 2\lambda_1$. Specifically, we can use

$$|\lambda, p_{\lambda 0}, q_\lambda\rangle = \Pi_{\lambda : p_{\lambda 0}} |x_{q_\lambda}\rangle , \tag{C.23}$$

with

$$|x_{q_\lambda}\rangle \equiv |01\rangle^{\otimes \lambda_1} \otimes |0\rangle^{\otimes n - 2\lambda_1 - q_\lambda} \otimes |1\rangle^{\otimes q_\lambda} . \tag{C.24}$$

Let us first evaluate

$$\sum_{r\in\mathrm{Row}(p_{\lambda 0})} R(r)\,|x_{q_\lambda}\rangle = \Pi_{r\to c}\left(\left|\Sigma^{q_\lambda}_{\lambda_0}\right\rangle \otimes |1\rangle^{\otimes\lambda_1}\right), \tag{C.25}$$

where $|\Sigma^y_x\rangle$ denotes the equal-weight superposition of all computation basis states on x qubits with $x-y$ zeros and y ones, which (up to normalization) is also known as *Dicke state* on x qubits [242, 243]. $\Pi_{r\to c}$ denotes the permutation of qubits needed to obtain the "column-standard" Young tableau $p_{\lambda 0}$ from an analogous "row-standard" Young tableau where the numbers first increase in the row direction and then in column direction. In other words, if we think of the qubits being associated to the tiles of the Young diagram λ, then the qubits in the first row are in state $\left|\Sigma^{q_\lambda}_{\lambda_0}\right\rangle$, and the qubits in the second row are in state $|1\rangle^{\otimes\lambda_1}$.

Next, for a two-row standard Young tableau $p_{\lambda 0}$, we have

$$\sum_{c\in\mathrm{Col}(p_{\lambda 0})} \mathrm{sgn}(c)\,R(c) = (\mathrm{id}_2 - \tau)^{\otimes\lambda_1}\otimes \mathrm{id}_2^{\otimes n-2\lambda_1} = (|\Psi\rangle\langle\Psi|)^{\otimes\lambda_1}\otimes \mathrm{id}_2^{\otimes n-2\lambda_1}, \tag{C.26}$$

where $|\Psi\rangle$ is the 2-qubit singlet state

$$|\Psi\rangle = \frac{1}{\sqrt{2}}\left(|01\rangle - |10\rangle\right), \tag{C.27}$$

and τ denotes the SWAP operator acting on two qubits. The qubits in the second row of λ in the state of Equation (C.25) are fixed to $|1\rangle$, so applying $|\Psi\rangle\langle\Psi|$ to each of the first λ_1 columns has the same effect as applying $|\Psi\rangle\langle\Psi|\,(|0\rangle\langle 0|\otimes \mathrm{id}_2)$. Applying $|0\rangle\langle 0|$ to the first λ_1 qubits of $\left|\Sigma^{q_\lambda}_{\lambda_0}\right\rangle$ yields $|0\rangle^{\otimes\lambda_1}\otimes\left|\Sigma^{q_\lambda}_{\lambda_0-\lambda_1}\right\rangle$. Thus, we find:

$$|\lambda, p_{\lambda 0}, q_\lambda\rangle = \Pi_{\lambda:p_{\lambda 0}}\,|x_{q_\lambda}\rangle = |\Psi\rangle^{\otimes\lambda_1}\otimes\left|\Sigma^{q_\lambda}_{n-2\lambda_1}\right\rangle. \tag{C.28}$$

Now, we are ready to evaluate

$$\begin{aligned} F^{\boldsymbol{i},\lambda}_{q_\lambda,q'_\lambda} &\equiv \langle \lambda, q_\lambda, p_{\lambda 0}| \, A_{\boldsymbol{i}} \, |\lambda, q'_\lambda, p_{\lambda 0}\rangle \\ &= \left(\langle \Psi|^{\otimes \lambda_1} \otimes \left\langle \Sigma^{q_\lambda}_{n-2\lambda_1} \right| \right) \left(\sum_{p_{\boldsymbol{i}} \in P_{\boldsymbol{i}}} p_{\boldsymbol{i}} \right) \left(|\Psi\rangle^{\otimes \lambda_1} \otimes \left| \Sigma^{q'_\lambda}_{n-2\lambda_1} \right\rangle \right) \\ &= \frac{1}{\sqrt{\binom{n-2\lambda_1}{q_\lambda}\binom{n-2\lambda_1}{q'_\lambda}}} \left(\sum_{s \in S^{q_\lambda}_{n-2\lambda_1}} \langle \Psi|^{\otimes \lambda_1} \otimes \langle s| \right) \left(\sum_{p_{\boldsymbol{i}} \in P_{\boldsymbol{i}}} p_{\boldsymbol{i}} \right) \left(\sum_{s' \in S^{q'_\lambda}_{n-2\lambda_1}} |\Psi\rangle^{\otimes \lambda_1} \otimes |s'\rangle \right), \end{aligned} \tag{C.29}$$

where we used S^x_y to denote the set of bitstrings of length y with exactly x ones. This is a sum over (more than) exponentially many terms. Similarly to the previous Appendix, it can be evaluated efficiently by realizing that many summands have equal value. Thus, we instead sum over the different possible values multiplied with the number of summands with that value, which can be counted using combinatorics. Each summand is an overlap of two product states with a product operator in between. More precisely, we have a product of first λ_1 two-qubit overlaps, and then $n - 2\lambda_1$ single-qubit overlaps.

For each summand in Equation (C.29), let us denote by L_{ab} with $a, b \in \{1, x, y, z\}$ the subset of two-qubit pairs:

$$L_{ab} \equiv \{(2l, 2l+1) : (p_{\boldsymbol{i}})_{2l} = \sigma_a, (p_{\boldsymbol{i}})_{2l+1} = \sigma_b, 0 \leq l < \lambda_1\}, \tag{C.30}$$

and let us write $f_{ab} = |L_{ab}|$ for the number of elements in those subsets. The according overlap

$$\langle \Psi| \, (\sigma_a \otimes \sigma_b) \, |\Psi\rangle \tag{C.31}$$

is 0 if $a \neq b$, so we only need to consider subsets where $a = b$. The number of summands for given numbers f_{aa} is the number of decompositions of the first λ_1 qubit pairs into the four subsets L_{aa} with $a \in \{1, x, y, z\}$, which equals

$$\frac{\lambda_1!}{f_{11}! f_{xx}! f_{yy}! f_{zz}!}. \tag{C.32}$$

The value which the overlap on the first λ_1 qubit pairs contributes to each summand only depends on the numbers f_{aa}. The overlap in Equation (C.31) is given by 1 if $a = b = 1$, and -1 if $a = b$ otherwise. Thus, the overall contribution to each summand is

$$(-1)^{f_{xx}+f_{yy}+f_{zz}}\,. \tag{C.33}$$

Next, let us consider the $n - 2\lambda_1$ single-qubit overlaps. For each summand in Equation (C.29), let us denote by K_{iaj} for $i, j \in \{0, 1\}$ and $a \in \{1, x, y, z\}$ the subset of the last $n - 2\lambda_1$ qubits

$$K_{iaj} \equiv \{l : (p_{\boldsymbol{i}})_{2\lambda_1+l} = \sigma_a, s_l = i, s'_l = j, 0 \leq l < n - 2\lambda_1\}, \tag{C.34}$$

and let us write $g_{iaj} = |K_{iaj}|$ for the number of elements in those subsets. The according overlap

$$\langle i|\, \sigma_a\, |j\rangle \tag{C.35}$$

is only nonzero if $i = j$ for $a \in \{1, z\}$ and $i \neq j$ for $a \in \{x, y\}$, so we can restrict to summands where only those 8 subsets are non-empty. The number of summands for given numbers g_{iaj} is the number of decompositions of the set of the last $n - 2\lambda_1$ qubits into the 8 subsets K_{iaj}, and is thus given by

$$\frac{(n - 2\lambda_1)!}{g_{010}!g_{111}!g_{0x1}!g_{1x0}!g_{0y1}!g_{1y0}!g_{0z0}!g_{1z1}!} \tag{C.36}$$

The contribution of the overlap on the last $n - 2\lambda_1$ qubits to each summand only depends on the numbers g_{iaj}. The single-qubit overlap in Equation (C.35) evaluates to 1 for g_{010}, g_{111}, g_{0X1}, g_{1X0} and g_{0z0}, -1 for g_{1z1}, i for g_{0y1}, and $-$i for g_{1y0}. Thus the overall contribution to each summand is

$$(-1)^{g_{1z1}}\,(-\mathrm{i})^{g_{0y1}}\,\mathrm{i}^{g_{1y0}}. \tag{C.37}$$

Overall, the number of summands for given f_{aa} and g_{iaj} is the product of Equation (C.32) and Equation (C.36), and the value of each summand is given by the

product of Equation (C.33) and Equation (C.37). Plugging this into Equation (C.29) yields Equation (C.21). The constraints in Equation (C.22) are explained as follows. The first four constraints are due to the fact that the number of zeros and ones in s and s' is determined by q_λ and q'_λ, respectively. The last four constraints correspond to the fact that the number of Pauli operators σ_1, σ_x, σ_y, σ_z in $p_{\boldsymbol{i}}$ is given by i_1, i_x, i_y, and i_z, respectively. □

In a similar fashion to the previous Appendix, we can easily evaluate the runtime this method achieves in calculating all of the matrix elements. Note that each component is a sum over four independent variables due to the constraints, yielding a runtime of $\mathrm{O}(n^4)$. Taking into account the $\mathrm{O}(n^6)$ tensor components of F yields the final runtime of $\mathrm{O}(n^{10})$. Once again, it seems likely that the $\mathrm{O}(n^4)$ runtime for a single tensor component can be reduced to a smaller exponent. We will leave this open to further investigation.

C.4 End-to-End Classical Simulation From Tensor Networks

We here consider a slight variant of Corollary 8 where the inputs are given as classical matrix product state (MPS) descriptions rather than as quantum states. From Reference [244], we have that $\langle m_1, m_2, m_3, m_4|\lambda, q_\lambda, p_\lambda\rangle \langle \lambda, q_\lambda, p_\lambda|$ has an efficient MPS description, where $|m_1, m_2, m_3, m_4\rangle$ is a computational basis state; a four-qubit (i.e., $n = 4$) example is given in Figure C-2, where we have used the Clebsch–Gordan coefficients $C^{\lambda,q_\lambda}_{p_1,m_1;p_2,m_2}$. p_i indices in Figure C-2 are discarded for clarity where they are trivial. Note that these Clebsch–Gordan coefficients can be classically computed efficiently up to p bits of precision (i.e., up to additive error exponentially small in p) in time $\mathrm{poly}(n, p)$ by the Racah formula [245]. The indices associated with m_i can then be efficiently contracted with an efficient MPS description of an initial state—even if it is not permutation invariant on qubits—and the p_i indices efficiently traced out to efficiently yield matrix elements of $\tilde{\rho}$ as defined in Equation (4.18).

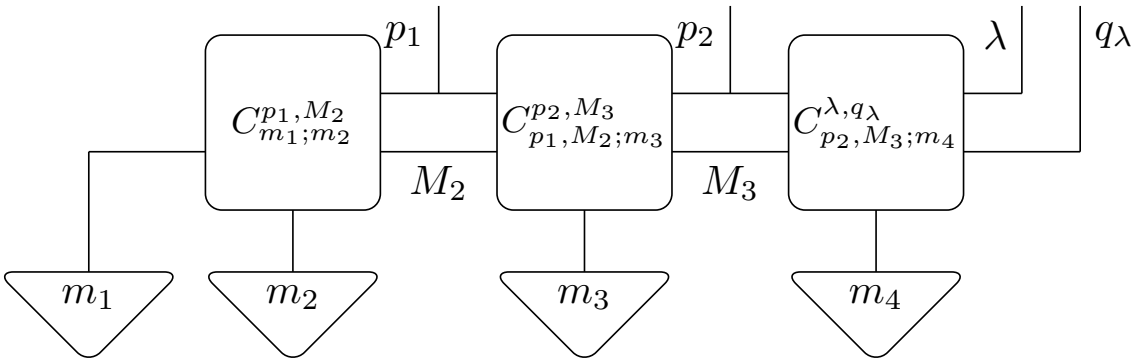


Figure C-2: A four-qubit example of the MPS giving $\langle m_1, m_2, m_3, m_4|\lambda, q_\lambda, p_\lambda\rangle \langle \lambda, q_\lambda, p_\lambda|$.

Appendix D

Technical Details for Chapter 5

D.1 Background on Sequence Learning

D.1.1 Sequence Learning

Sequence-to-sequence or *sequence* learning [84] is the approximation of some given conditional distribution $p(\boldsymbol{y} \mid \boldsymbol{x})$ with a model distribution $q(\boldsymbol{y} \mid \boldsymbol{x})$. This framework encompasses sentence translation tasks [84], speech recognition [158], image captioning [159], and many more practical problems.

The training and evaluation of sequence-to-sequence models is most often performed on the *forward (conditional) cross entropy*:

$$H(p, q) = -\int \mathrm{d}\boldsymbol{x} \int \mathrm{d}\boldsymbol{y}\, p(\boldsymbol{x}, \boldsymbol{y}) \log\left(q(\boldsymbol{y} \mid \boldsymbol{x})\right). \tag{D.1}$$

Here, "forward" indicates the ordering of the arguments of H; the *backward cross entropy* is given by $p \leftrightarrow q$. Given a finite test set $\mathcal{T} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$ of M points sampled from $p(\boldsymbol{x}, \boldsymbol{y})$, we can also define the forward *empirical* cross entropy

$$\hat{H}(p, q) = -\frac{1}{M} \sum_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{T}} p(\boldsymbol{y} \mid \boldsymbol{x}) \log\left(q(\boldsymbol{y} \mid \boldsymbol{x})\right), \tag{D.2}$$

with the backward empirical cross entropy once again given by $p \leftrightarrow q$.

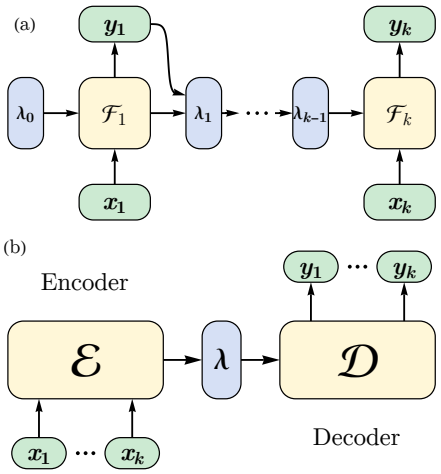


Figure D-1: (a) An online neural sequence model. The model autoregressively takes input tokens $\boldsymbol{x}_i$, and outputs decoded tokens $\boldsymbol{y}_i$, with map $\mathcal{F}_i$. The model also has an unobserved internal memory with state $\boldsymbol{\lambda}_i \in L$ after decoding token i that $\mathcal{F}_{i+1}$ can depend on. (b) A general encoder-decoder model. $\mathcal{E}$ encodes the input $\boldsymbol{x}$ into some latent representation $\boldsymbol{\lambda} \in L$. A decoder $\mathcal{D}$ then outputs the decoded sequence $\boldsymbol{y}$.

Historically, sequence learning was performed using Bayesian networks such as hidden Markov models [246, 247]. However, in recent years, the performance of these models have been eclipsed by neural network based models.

D.1.2 Neural Sequence Models

Sequence modeling today is typically performed using neural network based generative models, or *neural sequence models*. Generally, these models are parameterized functions that take as input the sequence $\boldsymbol{x}$ and output a sample from the conditional distribution $p(\boldsymbol{y} \mid \boldsymbol{x})$; the parameters of these functions are trained to minimize an appropriate loss function, such as the empirical cross entropy of Equation (D.2).

To maintain a resource scaling independent of the input sequence length, neural sequence models usually are one of two classes: *online sequence models* (also known as *autoregressive sequence models*) [48, 87, 88], or *encoder-decoder models* [84, 89]. Examples of both are given in Figure D-1. In the former class of models, input tokens $\boldsymbol{x}_i$ are translated in sequence to output tokens $\boldsymbol{y}_i$ via functions $\mathcal{F}_i$. An unobserved internal memory (or *latent space*) shared between time steps allows the model to

represent long-range correlations in the data.

In the latter of these models, an encoder $\mathcal{E}$ maps the input sequence $\boldsymbol{x}$ to a latent space representation $\boldsymbol{\lambda} \in L$; then, a decoder $\mathcal{D}$ transforms this representation to the output sentence $\boldsymbol{y}$. The advantage of encoder-decoder models over generic representations of $p(\boldsymbol{y} \mid \boldsymbol{x})$ is the improved time complexity when considering a lower dimensional representation $\boldsymbol{\lambda}$ of $\boldsymbol{x}$. When the encoder map is trivial (i.e. when L is congruent to the input space and $\mathcal{E}$ is the identity), then no compression occurs, and the model is equivalent to a general representation of $p(\boldsymbol{y} \mid \boldsymbol{x})$ given by $\mathcal{D}$.

Generally, there are no restrictions on the forms of $\mathcal{F}_i$, $\mathcal{E}$, or $\mathcal{D}$, though most neural sequence models are composed of simple smooth (or almost everywhere smooth) functions out of training considerations [48, 87–89]. Here, we generalize from the typical smoothness constraints and consider *locally Lipschitz* maps. A function $\mathcal{F} : K \to L$ for metric spaces K and L is locally Lipschitz if

$$d_L(\mathcal{F}(\boldsymbol{x}), \mathcal{F}(\boldsymbol{x'})) \leq C_{\boldsymbol{x}} d_K(\boldsymbol{x}, \boldsymbol{x'}), \tag{D.3}$$

where $C_{\boldsymbol{x}}$ is constant in some neighborhood of $\boldsymbol{x}$. Here, d_S is the distance function on the metric space S.

All practical neural sequence models are locally Lipschitz. Indeed, assuming $L = \mathbb{R}^m$, all maps that are almost everywhere differentiable with locally bounded Jacobian norm are locally Lipschitz [160]. Realistically, then, locally Lipschitz models can be thought of as all models trainable using gradient based methods; equivalently, they can be thought of as models trainable via methods not arbitrarily sensitive to local noise $\boldsymbol{x}_i \mapsto \boldsymbol{x}_i + \boldsymbol{\epsilon}$.

D.2 Proofs of Expressivity Separations

Before giving proofs of expressivity separations between our quantum model and classical models, we first give a formal definition of the translation task we will prove a separation on: namely, (k, n) *stabilizer measurement translation*, parameterized by

k and n. Note that the technical description of the task described here is a certain limit of the construction presented in the main text. First, we take $k \to k-n$, as our task is only explicitly defined when the sequence length is at least n. For instance, the $(n+2, n)$ stabilizer measurement translation task as presented in the main text will, here, be referred to as the $(2, n)$ stabilizer measurement translation task. We make this change as the formal definition of this task as presented here is undefined when $k+n < n$. Second, we now set the first n measurements to be infinite precision Gaussian measurements, i.e. measurements of linear combinations of position and momentum operators. These measurements are a limit of the periodic measurements we consider in the main text, with infinitely large periods.

To be more explicit, we consider an input language given by $n+k$-long sequences of linear combinations of position and momentum operators on n modes. Specifically, input sentences are composed of words which are of the form of rows of:

$$\boldsymbol{x} = \begin{pmatrix} s^q_{1,1} & \dots & s^q_{1,n} & s^p_{1,1} & \dots & s^p_{1,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ s^q_{n+k,1} & \dots & s^q_{n+k,n} & s^p_{n+k,1} & \dots & s^p_{n+k,n} \end{pmatrix}. \tag{D.4}$$

The first n rows describe the sequential measurement of each operator

$$\hat{s}_i = \sum_{j=1}^{n} s^q_{ij} \hat{q}_j + \sum_{j=1}^{n} s^p_{ij} \hat{p}_j \tag{D.5}$$

when beginning in some given fixed state $|\psi_0\rangle$ on n modes that is either a GKP state [95] or an infinitely squeezed Gaussian state, which maintains the nonuniversality of the model [163]. The final k rows describe the sequential measurement of each operator

$$\hat{s}_i = \exp\left(\mathrm{i} \sum_{j=1}^{n} s^q_{ij} \hat{q}_j + \mathrm{i} \sum_{j=1}^{n} s^p_{ij} \hat{p}_j \right) \tag{D.6}$$

via e.g. phase estimation, as shown in the main text. Note that the measurement of $\hat{s}_i$ is *not* equivalent to the measurement of its generator. For instance, given states

such that:

$$\hat{q}\,|\psi_1\rangle = 0, \qquad \hat{q}\,|\psi_2\rangle = 2\pi, \tag{D.7}$$

one has

$$(\exp(\mathrm{i}\hat{q}) - 1)\,|\psi_1\rangle = (\exp(\mathrm{i}\hat{q}) - 1)\,|\psi_2\rangle = 0. \tag{D.8}$$

A translation $\boldsymbol{y}$ of $\boldsymbol{x}$ is considered correct if it is of the form

$$\boldsymbol{b} = \begin{pmatrix} m_1 \\ \vdots \\ m_{n+k} \end{pmatrix}, \tag{D.9}$$

where the measurement outcomes m_i are consistent with those of quantum mechanics. To prove our separations, we will consider input sentences that exhibit quantum contextuality.

D.2.1 Expressivity Separation for Online Models

We now show that locally Lipschitz online with latent space dimension less than $\frac{n(n-3)}{2}$ can stabilizer measurement translate. Our general proof strategy is as follows:

1. We first define a potentially random online learning model with locally Lipschitz cell maps $\mathcal{F}_i^{\boldsymbol{r}}(\boldsymbol{s}_i, \boldsymbol{\lambda}_{i-1}) = (m_i, \boldsymbol{\lambda}_i)$, where we use m_i to indicate the measurement result when measuring the nullifier described by $\boldsymbol{s}_i$, and $\boldsymbol{r}$ is a random vector shared between all $\mathcal{F}_i$. We assume for any $\boldsymbol{r}$ that $\mathcal{F}_i^{\boldsymbol{r}}$ is deterministic. Let $\mathcal{F}^{\boldsymbol{r}}(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n) = \boldsymbol{\lambda}_n$ be the n-fold composition of $\mathcal{F}_i^{\boldsymbol{r}}$ on some fixed initial $\boldsymbol{\lambda_0}$ (where any m_i is implicit, as each m_i is fully determined by $\boldsymbol{r}$ and $\boldsymbol{s}_i$). Note that once $\boldsymbol{r}$ is specified, $\mathcal{F}_{\boldsymbol{r}}$ is a deterministic function. Due to this, in the following, we take the $\boldsymbol{r}$ dependence to be implicit.

2. We assume the dimension of $\boldsymbol{\lambda}_i$ is less than $\frac{n(n-3)}{2}$. We prove that then, the described online model will give a wrong measurement outcome on the final two measurement results m_{n+1}, m_{n+2}. In the following, we will refer to this as *the theorem statement.*

3. To prove that *the theorem statement* is true, we show that it is true for a subspace K of inputs which describe *CV graph states*, where the associated graphs have no self-loops. It is easy to see that the dimension of such a space is $\frac{n(n-1)}{2}$, by studying the space of adjacency matrices. We let $\mathcal{F}|_K$ be $\mathcal{F}$ restricted to this space K. Let $\boldsymbol{B}$ be coordinates of K, as defined in Equation (D.26). We assume the Jacobian of this map achieve its maximal rank at $\boldsymbol{B} = \boldsymbol{0}$, which describes a set of measurements yielding the position squeezed state $|\boldsymbol{0}\rangle_{\hat{\boldsymbol{q}}}$. By doing so, we guarantee the robustness of the Jacobian rank in a neighborhood of $\boldsymbol{B} = \boldsymbol{0}$. The assumption that $\boldsymbol{B} = \boldsymbol{0}$ is a point of maximal rank is taken WLOG, as there exist Gaussian operations that transform the $\boldsymbol{B}$ at which the Jacobian of $\mathcal{F}|_K$ attains its maximal rank to $\boldsymbol{B} = \boldsymbol{0}$.

4. As the rank is constant in the neighborhood of $\boldsymbol{B} = \boldsymbol{0}$, by the constant rank theorem [169], $\mathcal{F}_{\boldsymbol{r}}|_K$ induces a fiber bundle structure in the neighborhood of $\boldsymbol{B} = \boldsymbol{0}$. That is, $\mathcal{F}_{\boldsymbol{r}}|_K$ is a projection of fibers in a neighborhood of $\boldsymbol{B} = \boldsymbol{0}$ to their base points. This means that the model is unable to distinguish between points that share a fiber in this neighborhood.

5. We then show that when $\dim(\boldsymbol{\lambda}_n) < \dim(K) - n$, there exist $\boldsymbol{B'}, \boldsymbol{B''}$ on the fiber with base point $\boldsymbol{B} = \boldsymbol{0}$ such that $\boldsymbol{B}, \boldsymbol{B'}, \boldsymbol{B''}$ describe distinct states. We show in Lemma D.1 that these states have stabilizers which share quantum contextuality, yielding distinguishing one-shot measurement sequences. As the model is unable to distinguish between $\boldsymbol{B}, \boldsymbol{B'}, \boldsymbol{B''}$, this implies that there exist $\boldsymbol{s}_{n+1}, \boldsymbol{s}_{n+2}$ describing this distinguishing measurement sequence such that the model returns the wrong measurement result for at least one of $\boldsymbol{B}, \boldsymbol{B'}, \boldsymbol{B''}$ with certainty. This proves *the theorem statement*. We also use this general proof strategy when considering encoder-decoder models in Theorem D.4, up to some minor details.

With our proof strategy now clear, we now proceed to prove the details. First, we prove our lemma demonstrating that indeed, the stabilizer operators we consider exhibit quantum contextuality. We also show that this contextuality induces a one-shot

distinguishing measurement sequence on the states stabilized by the given operators.

Lemma D.1 (CV graph state stabilizers exhibit quantum contextuality)**.** *Consider* $|\mathbf{0}\rangle_{\hat{\boldsymbol{q}}}$*, the state nullified by all* $\hat{q}_i$*. Consider two states* $|\psi_1\rangle$ *and* $|\psi_2\rangle$ *that are CV graph states (up to arbitrary phases on their stabilizers) with no loops with distinct (modulo* π*) adjacency matrices. There exist operators that stabilize these three states that exhibit quantum contextuality. Furthermore, there exists a distinguishing measurement given by one of the stabilizers of* $|\mathbf{0}\rangle_{\hat{\boldsymbol{q}}}$ *that maps* $|\psi_1\rangle$ *and* $|\psi_2\rangle$ *to orthogonal post-measurement states when the measurement result is* 1*; in other words, there exists a distinguishing measurement sequence of length two that distinguishes these three states.*

Proof. As $|\psi_1\rangle$ and $|\psi_2\rangle$ are CV graph states with distinct adjacency matrices (modulo π), they must differ (modulo π) in the edges $\boldsymbol{e}_i$ touching some vertex i. That is, $|\psi_1\rangle$ is stabilized by some:

$$s_i' = \mathrm{e}^{2\mathrm{i}\theta'} X_i(1)\, \boldsymbol{Z}(\boldsymbol{e}_i'), \tag{D.10}$$

and $|\psi_2\rangle$ by some:

$$s_i'' = \mathrm{e}^{2\mathrm{i}\theta''} X_i(1)\, \boldsymbol{Z}(\boldsymbol{e}_i''), \tag{D.11}$$

where $\boldsymbol{e}_i'$ and $\boldsymbol{e}_i''$ differ (modulo π) in some element indexed by $j \neq i$ (as there are no loops in either graph), and where $2\theta', 2\theta''$ are phases. Here, $\boldsymbol{Z}(\cdot)$ is defined as the tensor product:

$$\boldsymbol{Z}(\boldsymbol{v}) = \bigotimes_i Z_i(v_i). \tag{D.12}$$

An example diagram of the entries of $\boldsymbol{e}_i', \boldsymbol{e}_i''$ is given in Figure D-2. By the symmetry of CV graph state adjacency matrices, we thus have that the former is also stabilized by

$$s_j' = \mathrm{e}^{2\mathrm{i}\phi'} X_j(1)\, \boldsymbol{Z}\left(\boldsymbol{e}_j'\right) \tag{D.13}$$

and the latter by

$$s_j'' = \mathrm{e}^{2\mathrm{i}\phi''} X_j(1)\, \boldsymbol{Z}\left(\boldsymbol{e}_j''\right), \tag{D.14}$$

where $\boldsymbol{e}_j'$ and $\boldsymbol{e}_j''$ differ (modulo π) in some element indexed by i, and where $2\phi', 2\phi''$

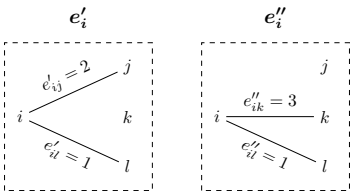


Figure D-2: An example of two graphs, with edges leaving vertex i given by $\boldsymbol{e}'_i$ and $\boldsymbol{e}''_i$, respectively. As the two graphs differ, they must differ by an edge; indeed, they differ in the weights of both $\langle i,j\rangle$ and $\langle i,k\rangle$.

are phases. Note in particular that we have the commutation relations

$$[s'_i, s''_i] = 0, \qquad \left[s'_j, s''_j\right] = 0, \qquad \left[s'_i, s'_j\right] = 0, \qquad \left[s''_i, s''_j\right] = 0, \tag{D.15}$$

$$s'_i s''_j = \mathrm{e}^{2\mathrm{i}\zeta} s''_j s'_i, \qquad s'_j s''_i = \mathrm{e}^{2\mathrm{i}\zeta} s''_i s'_j, \qquad s'_i s''^{\dagger}_j = \mathrm{e}^{-2\mathrm{i}\zeta} s''^{\dagger}_j s'_i, \qquad s'_j s''^{\dagger}_i = \mathrm{e}^{-2\mathrm{i}\zeta} s''^{\dagger}_i s'_j, \tag{D.16}$$

where

$$\zeta \equiv e'_{ij} - e''_{ij} = e'_{ji} - e''_{ji} \neq 0 \pmod{\pi}. \tag{D.17}$$

As $\zeta \neq 0 \pmod{\pi}$, there exists some $\alpha \in \mathbb{R}^*_+$ such that

$$\left\{s'^{\alpha}_i, s''^{\alpha}_j\right\} = 0, \qquad \left\{s'^{\alpha}_j, s''^{\alpha}_i\right\} = 0, \tag{D.18}$$

$$\left\{s'^{\alpha}_i, s''^{\dagger\alpha}_j\right\} = 0, \qquad \left\{s'^{\alpha}_j, s''^{\dagger\alpha}_i\right\} = 0. \tag{D.19}$$

To save on notation, we redefine all stabilizers to be given by their α power, i.e. $s^\alpha \to s$ (and similarly redefine the phases $\theta', \theta'', \phi', \phi''$ by their scaling by α). As $|\psi_1\rangle, |\psi_2\rangle$ are CV graph states, these rescaled operators are still stabilizers. We also define:

$$s_i \equiv \mathrm{e}^{-2\mathrm{i}(\theta'-\theta'')} s'_i s''^{\dagger}_i, \tag{D.20}$$

$$s_j \equiv \mathrm{e}^{-2\mathrm{i}(\phi'-\phi'')} s'_j s''^{\dagger}_j, \tag{D.21}$$

which are stabilizers of $|\mathbf{0}\rangle_{\hat{\boldsymbol{q}}}$. Thus, we have constructed nine observables with constraints satisfying those of a Mermin–Peres magic square (see Table D.1 for an exam-

s_i'	s_j'	$s_i'^\dagger s_j'^\dagger$
$s_i''^\dagger$	$s_j''^\dagger$	$s_i'' s_j''$
$s_i^\dagger$	$s_j^\dagger$	$s_i s_j$

Table D.1: The Mermin–Peres magic square of stabilizers of states mapping to the same latent space under a locally Lipschitz map (with $\theta' = \theta'' = \phi' = \phi'' = 0$ for simplicity). Stabilizers of $|\psi_1\rangle$, $|\psi_2\rangle$, and $|\mathbf{0}\rangle_{\hat{\boldsymbol{q}}}$ make up the three rows. All observables in each row and column commute. Furthermore, the product of observables in each row and column is the identity, except for the third column, which gives minus the identity. See the main text for a special case of this magic square.

ple), a well-known proof of quantum contextuality [93].

Consider now the post-measurement states $|\psi_1'\rangle, |\psi_2'\rangle$ of $|\psi_1\rangle, |\psi_2\rangle$, respectively, when $s_i s_j$ is measured to be 1. By Table D.1, $|\psi_1'\rangle$ is stabilized by $s_i' s_j'$ and $s_i s_j$; furthermore, $|\psi_2''\rangle$ is stabilized by $s_i'' s_j''$ and $s_i s_j$, and therefore is also stabilized by

$$s_i s_j s_i'' s_j'' = -\mathrm{e}^{-2\mathrm{i}(\theta' - \theta'' + \phi' - \phi'')} s_i' s_j'. \tag{D.22}$$

If

$$\theta' - \theta'' + \phi' - \phi'' \neq \frac{\pi}{2} \pmod{\pi}, \tag{D.23}$$

we have that $|\psi_1'\rangle$ and $|\psi_2'\rangle$ are orthogonal. If it is congruent to $\frac{\pi}{2} \pmod{\pi}$, then either $\theta' - \theta'' \neq 0 \pmod{\pi}$ or $\phi' - \phi'' \neq 0 \pmod{\pi}$ (or both). Assume the latter WLOG (the former case is the same with $\phi \to \theta$ and $j \to i$), and instead consider the post-measurement states $|\psi_1'\rangle, |\psi_2'\rangle$ of $|\psi_1\rangle, |\psi_2\rangle$, respectively, when s_j is measured to be 1. By Table D.1, $|\psi_1'\rangle$ is stabilized by s_j' and s_j; furthermore, $|\psi_2''\rangle$ is stabilized by s_j'' and s_j, and therefore is also stabilized by

$$s_j s_j'' = \mathrm{e}^{-2\mathrm{i}(\phi' - \phi'')} s_j'. \tag{D.24}$$

Thus, again, $|\psi_1'\rangle$ and $|\psi_2'\rangle$ are orthogonal. □

We now consider the locally Lipschitz online learner, with structure given by Figure D-1(a). We assume that the learner is deterministic, and discuss the extension to randomized models at the end of this Section.

Online neural sequence models at time step i map an input token $\boldsymbol{x}_i$ and a latent vector $\boldsymbol{\lambda}_{i-1}$ to an output token m_i and a new latent vector $\boldsymbol{\lambda}_i$. After n steps, then, we can consider the online model as a locally Lipschitz map:

$$\mathcal{F}^{\boldsymbol{r}} : \left(\mathbb{R}^{2n}\right)^n \to L \times \mathbb{R}^n, \tag{D.25}$$

where L is a locally Lipschitz latent manifold and $\boldsymbol{r}$ is a random vector such that, for any $\boldsymbol{r}$, $\mathcal{F}^{\boldsymbol{r}}$ is deterministic. This consideration of $\mathcal{F}^{\boldsymbol{r}}$ as a deterministic function of a random $\boldsymbol{r}$ is typically the implementation of stochastic learners, such as generative adversarial networks (GANs) [248] and flow-based models [249]. This also includes implementations of stochastic simulation algorithms such as Wigner function simulation [85]. As for each $\boldsymbol{r}$, there exists an input sequence such that any classical model with $\dim(L) < \frac{n(n-3)}{2}$ deterministically outputs a measurement sequence inconsistent with quantum mechanics, our results still hold for these classes of random models. Due to this, in the following, we take the $\boldsymbol{r}$ dependence to be implicit.

With the preliminaries in place, we now prove our expressivity separation.

Theorem D.2 (Online stabilizer measurement translation lower bound)**.** *Consider an online model with locally Lipschitz latent manifold L and locally Lipschitz map $\mathcal{F}$ as described in Equation* (D.25)*. If $\dim(L) < \frac{n(n-3)}{2}$, this model cannot achieve a finite backward empirical cross entropy on the $(2, n)$ stabilizer measurement translation task.*

Proof. Consider $K \subset (\mathbb{R}^{2n})^n$, with elements of the form:

$$\boldsymbol{Q} = \left(\begin{array}{c|c} \boldsymbol{B} + \boldsymbol{H}/2 & \|\boldsymbol{B}\|_{\mathrm{F}} \boldsymbol{I_n} \end{array} \right); \tag{D.26}$$

here, $\|\boldsymbol{B}\|_{\mathrm{F}}$ is the Frobenius norm of $\boldsymbol{B}$, and $\boldsymbol{B}$ is an $n \times n$ hollow (zero diagonal elements) symmetric matrix with entries bounded to be $\left[-\frac{1}{4}, \frac{1}{4}\right]$. $\boldsymbol{H}$ is the fixed $n \times n$

symmetric hollow matrix of ones, i.e.

$$\boldsymbol{H} = \begin{pmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{pmatrix}. \tag{D.27}$$

It is obvious from this construction that K is an $\frac{n(n-1)}{2}$-dimensional embedding of a compact subspace of hollow symmetric matrices (with bounded entries and norm) $\boldsymbol{B}$. Note that the states described by the measurement scenarios of points in K are exactly CV graph states without loops (with bounded weight edges, as K is compact) and, depending on the measurement results, perhaps overall phases on the stabilizers. To see this, note that the symmetric constraint on $\boldsymbol{B}$ ensures that the symplectic product of any two rows of $\boldsymbol{Q}$ is zero; furthermore, the final n columns of $\boldsymbol{Q}$ are linearly independent for all $\boldsymbol{B} \neq \boldsymbol{0}$, and the first n columns for $\boldsymbol{B} = \boldsymbol{0}$ due to the shift by $\boldsymbol{H}$. Thus, all points in K are full row rank, and the rows of $\boldsymbol{Q}$ completely determine the CV stabilizer state, up to phases given by the measurement results of these operators. Furthermore, as $\boldsymbol{B}$ and $\boldsymbol{H}$ are hollow, the CV graph state $\boldsymbol{Q}$ describes has no loops. Also note that, up to independent rescalings of the rows of $\boldsymbol{Q}$, different $\boldsymbol{Q}$ correspond to different graph states. We assume WLOG that the Jacobian of $\mathcal{F}$ attains its maximum rank at $\boldsymbol{B} = \boldsymbol{0}$ (that is, the squeezed state $|\boldsymbol{0}\rangle_{\hat{\boldsymbol{q}}}$); this can always be done by implicitly transforming the basis of inputs to the model (i.e. by appropriately relabeling points in $(\mathbb{R}^{2n})^{n+2}$), and then considering K as previously defined in this new basis.

We will proceed as follows. First, we will show that when $\dim(L)$ is sufficiently small, $\mathcal{F}$ must map three distinct CV graph states described by different $\boldsymbol{Q}$ to the same point in latent space. Then, we will use Lemma D.1 to show that the stabilizers of these states exhibit quantum contextuality (independent of the associated n measurement results), and give rise to a distinguishing measurement sequence. Thus, by considering measurement sequences of length $n+2$ that include these three $\boldsymbol{Q}$ and the

length two distinguishing measurement sequence, one of the final two measurement outcomes must be incorrect. This implies that there is an infinite backward empirical cross entropy on any finite set containing these three measurement sequences.

Let us begin by showing that $\mathcal{F}|_K$ (i.e. the locally Lipschitz map that is $\mathcal{F}$ restricted to K) must map three nontrivially distinct $\boldsymbol{Q}$ (i.e. three distinct CV graph states) to the same point in latent space when $\dim(L)$ is sufficiently small. By the constant rank theorem and the local Lipschitzness of $\mathcal{F}|_K$, $\mathcal{F}|_K$ is not injective for

$$\dim(L) < \dim(K) = \frac{n(n-1)}{2}. \tag{D.28}$$

In particular, in a sufficiently small neighborhood of $\boldsymbol{B} = \boldsymbol{0}$ (where the Jacobian of $\mathcal{F}|_K$ attains its maximal rank), there exist local coordinates $\tilde{\boldsymbol{x}}$ of K and L such that

$$\mathcal{F}|_K\left(\tilde{x}_1, \ldots, \tilde{x}_{\frac{n(n-1)}{2}}\right) = (\tilde{x}_1, \ldots, \tilde{x}_l, 0, \ldots, 0) \tag{D.29}$$

for some $l \leq \dim(L) < \frac{n(n-1)}{2}$ [169]. WLOG, we identify $\tilde{\boldsymbol{x}} = \boldsymbol{0}$ with $\boldsymbol{B} = \boldsymbol{0}$, which is the state infinitely squeezed in all $\hat{q}_i$. We will call C the fiber with local coordinates

$$\tilde{\boldsymbol{x}} = \left(0, \ldots, 0, \tilde{x}_{l+1}, \ldots, \tilde{x}_{\frac{n(n-1)}{2}}\right), \tag{D.30}$$

which is of dimension at least

$$\Delta \equiv \frac{n(n-1)}{2} - \dim(L) \geq 1. \tag{D.31}$$

By construction, all points in C—including $\boldsymbol{B} = \boldsymbol{0}$—map to the same point $l \in L$ under $\mathcal{F}|_K$. We now assume that $\dim(L) < \frac{n(n-3)}{2}$ such that $\Delta \geq n + 1$.

Now fix $\boldsymbol{B} = \boldsymbol{0}$ and $\boldsymbol{B'} \neq \boldsymbol{B}$ in C. As described previously, $\boldsymbol{Q}$ (and thus $\boldsymbol{B}$) completely determines a CV graph state after n measurements, up to independent rescalings of the rows of $\boldsymbol{Q}$ (and the measurement results). As the dimension of the space of points that differ (modulo π) from $\boldsymbol{B'} + \boldsymbol{H}/2$ by just a scaling factor in each row is at most n, because $\Delta \geq n + 1$ we must have that there exists another

$\boldsymbol{B''} \neq \boldsymbol{B}, \boldsymbol{B'}$ describing a distinct CV graph state. Therefore, by Lemma D.1, we have that there exists a distinguishing measurement sequence for these three states. Note that as Lemma D.1 does not depend on the phases of the CV stabilizers, the existence of this distinguishing measurement holds true regardless of what the measurement results are (i.e. independently from what the model outputs for the first n tokens in the decoded sequence). As after n tokens all three sequences map to the same point in latent space in the model, and as they share a distinguishing measurement sequence, the model must obtain an infinite backward empirical cross entropy on these three input sequences when followed by the distinguishing measurement sequence. □

D.2.2 A CV Gottesman–Knill Lower Bound

We now show that our results can be reformulated as a memory lower bound on the classical simulation of stabilizer measurement scenarios. In practice, using finite resources (i.e. at finite precision), any classical ontological model simulating $p(\boldsymbol{y} \mid \boldsymbol{x})$ can only be evaluated at a finite number of $\boldsymbol{x}$. We now show that *any* locally Lipschitz interpolation of such a model to real $\boldsymbol{x}$ cannot accurately simulate Gaussian operations on an initial GKP state. This includes, for instance, any polynomial interpolation (which always exists).

Corollary D.3 (CV Gottesman–Knill lower bound)**.** *Consider a classical ontological model $p(\boldsymbol{y} \mid \boldsymbol{x})$ with a latent space of dimension less than $\frac{n(n-3)}{2}$, simulating (k, n) stabilizer measurement translation with $k \geq 2$. Assume that this ontological model is defined at a finite number of $\boldsymbol{x}$. There exists a locally Lipschitz interpolation of this model to all $\boldsymbol{x}$. Furthermore, no locally Lipschitz interpolation of this model can faithfully perform stabilizer measurement translation at all $\boldsymbol{x}$.*

Proof. As there exists a polynomial interpolation of p, and as all polynomials of finite degree are locally Lipschitz, there exists a locally Lipschitz interpolation of this model to all $\boldsymbol{x}$. Furthermore, no locally Lipschitz interpolation of this model can faithfully perform stabilizer measurement translation by Theorem D.2, as the composition of locally Lipschitz functions is locally Lipschitz. □

D.2.3 Expressivity Separation for Encoder-Decoder Models

Though online sequence models are perhaps conceptually the simplest as they directly map input tokens to output tokens, in practice encoder-decoder models outperform them [84, 89]. We now show that no encoder-decoder model with a locally Lipschitz encoder (and an additional technical assumption) can perform stabilizer measurement translation to finite backward empirical cross entropy. The proof will be similar to that of Theorem D.2; however, as the model can see the entire input sequence at once, we do not directly have the freedom to choose the distinguishing measurement sequence as in Theorem D.2. Instead, we will require an input sequence of length quadratic in n (and our additional technical assumption) to force the distinguishing measurement sequence. Note that, as the memory of the contextual learner is independent of the sequence length, this new sequence length has no impact on the memory separation.

We consider an encoder-decoder model with structure given by Figure D-1(b). The encoder of such a model can be considered a locally Lipschitz map

$$\mathcal{E}^{\boldsymbol{r}} : \left(\mathbb{R}^{2n}\right)^{n^2} \to L \tag{D.32}$$

to some locally Lipschitz latent manifold L. As in Section D.2.1, $\boldsymbol{r}$ is a random vector such that, for any $\boldsymbol{r}$, $\mathcal{E}^{\boldsymbol{r}}$ is deterministic. We once again make the $\boldsymbol{r}$ dependence implicit in the following.

For technical reasons, we slightly change the definition of the (k, n) stabilizer measurement translation task, where now the final k measurement descriptions instead describe the measurements of the operators:

$$\hat{s}_i = \exp\left(\mathrm{i}\sum_{j=1}^{n} \frac{\mathbf{1}\left[s_{ij}^q \neq 0\right]}{s_{ij}^q}\hat{q}_j + \mathrm{i}\sum_{j=1}^{n} \frac{\mathbf{1}\left[s_{ij}^p \neq 0\right]}{s_{ij}^p}\hat{p}_j\right), \tag{D.33}$$

where we define $\frac{\mathbf{1}[x\neq 0]}{x}$ to be zero when $x = 0$. We will call this the *modified* (k, n) *stabilizer measurement translation task*. This can obviously still be performed perfectly with a CRNN of model size n, by either changing the parameters of the phase

estimation circuit used in the CRNN to be given by $\frac{\mathbf{1}[x_{ij}\neq 0]}{x_{ij}}$, or more formally by introducing a quantum circuit computing $\frac{\mathbf{1}[x_{ij}\neq 0]}{x_{ij}}$ on which these gates control.

We now discuss our additional technical assumption. Defining the subspace R of inputs as in the proof of Theorem D.4, we assume that the Jacobian of the encoder restricted to R attains its maximal rank at some point of the form $(\boldsymbol{Q}, \mathbf{0}) \in R$. A sufficient condition for this is requiring that some point of the form $(\boldsymbol{Q}, \mathbf{0})$ is not a critical point of $\mathcal{E}|_R$; this condition is satisfied by generic $\mathcal{E}$ when $\dim(L) < \frac{n(n-3)}{2}$, and also by models with encoders constrained to be submersions (such as encoders composed of linear transformations and tanh or sigmoid nonlinearities). In fact, when the encoder is constrained to be a submersion, it is easy to see from the proof of Theorem D.4 that the separation still holds on the *unmodified* (k, n) stabilizer measurement translation task, as all properties we use that hold locally then hold globally. Any one of these conditions is sufficient, and needed for our proof technique to be able to analyze any neighborhood of the non-Gaussian measurements we consider.

Theorem D.4 (Encoder-decoder stabilizer measurement translation lower bound)**.** *Consider an encoder-decoder model with locally Lipschitz latent manifold L. Let $\mathcal{E}$ be the associated locally Lipschitz encoder function, as defined in Equation* (D.32)*, and assume that the Jacobian of the map $\mathcal{E}|_R$ (where the subspace R is defined below) attains its maximal rank at some point of the form $(\boldsymbol{Q}, \mathbf{0}) \in R$. If $\dim(L) < \frac{n(n-3)}{2}$, this model cannot achieve a finite backward empirical cross entropy on the modified $(n^2 - n, n)$ stabilizer measurement translation task.*

Proof. Consider $R \equiv K \times (\mathbb{R}^{2n})^{n^2-n} \subset (\mathbb{R}^{2n})^n \times (\mathbb{R}^{2n})^{n^2-n} \cong (\mathbb{R}^{2n})^{n^2}$ with elements $(\boldsymbol{Q}, \boldsymbol{P})$ of the following form:

1. $\boldsymbol{Q} \in K$ is given by rows of matrices of the form:

$$\boldsymbol{Q} = \left(\begin{array}{c|c} \boldsymbol{B} + \boldsymbol{H}/2 & \|\boldsymbol{B}\|_{\mathrm{F}} \boldsymbol{I_n} \end{array} \right); \tag{D.34}$$

here, $\|\boldsymbol{B}\|_{\mathrm{F}}$ is the Frobenius norm of $\boldsymbol{B}$, and $\boldsymbol{B}$ is a hollow symmetric matrix with entries bounded to be $\left[-\frac{1}{4}, \frac{1}{4}\right]$. $\boldsymbol{H}$ is the fixed symmetric hollow matrix of

ones, i.e.

$$\boldsymbol{H} = \begin{pmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{pmatrix}. \tag{D.35}$$

2. $\boldsymbol{P} \in (\mathbb{R}^{2n})^{n^2-n}$ is an $n^2 - n \times 2n$ matrix that is arbitrary.

It is obvious from this construction that K is an $\frac{n(n-1)}{2}$-dimensional embedding of a compact subspace of hollow symmetric matrices (with bounded entries and norm) $\boldsymbol{B}$. We assume WLOG that the Jacobian of $\mathcal{E}|_K$ (that is, the locally Lipschitz restriction of $\mathcal{E}$ to points of the form $(\boldsymbol{Q}, \boldsymbol{0}) \in R$) attains its maximum rank at $\boldsymbol{B} = \boldsymbol{0}$ (that is, the squeezed state $|\boldsymbol{0}\rangle_{\hat{\boldsymbol{q}}}$); this can always be done by implicitly transforming the basis of the inputs to the model (i.e. by appropriately relabeling points in $(\mathbb{R}^{2n})^{n^2}$), and then considering K as previously defined in this new basis.

We now give some intuition behind points in the smooth manifold (with boundary) R. At fixed $\boldsymbol{P} = \boldsymbol{0}$, the states described by the measurement scenarios of points in R are exactly CV graph states without loops (with bounded weight edges, as K is compact) and, depending on the measurement results, perhaps overall phases on the stabilizers. To see this, note that the symmetric constraint on $\boldsymbol{B}$ ensures that the symplectic product of any two rows of $\boldsymbol{Q}$ is zero; furthermore, the final n columns of $\boldsymbol{Q}$ are linearly independent for all $\boldsymbol{B} \neq \boldsymbol{0}$, and the first n columns for $\boldsymbol{B} = \boldsymbol{0}$ due to the shift by $\boldsymbol{H}$. Thus, all points in K are full row rank, and the rows of $\boldsymbol{Q}$ completely determine the CV stabilizer state, up to phases given by the measurement results of these operators (which are not yet determined at the time of encoding). Furthermore, as $\boldsymbol{B}$ and $\boldsymbol{H}$ are hollow, the CV graph state $\boldsymbol{Q}$ describes has no loops. Also note that, up to trivial rescalings of the rows of $\boldsymbol{Q}$, different $\boldsymbol{Q}$ correspond to different graph states. At general $\boldsymbol{P}$, the state after the first n measurements is still a CV graph state completely determined by $\boldsymbol{Q}$ (up to phases from the first n measurement results); different $\boldsymbol{P}$ correspond to different (non-Gaussian) measurement scenarios given an initial CV graph state determined by $\boldsymbol{Q}$ (and the first n measurement results).

We will proceed as follows. First, we will show that as $\dim(L) < \frac{n(n-3)}{2}$, $\mathcal{E}$ must map three distinct CV graph states described by different $\boldsymbol{Q}$ to the same point in latent space when $\boldsymbol{P} = \boldsymbol{0}$. Then, we will use Lemma D.1 to show that the stabilizers of these states exhibit quantum contextuality (independent of the associated n measurement results), and give rise to a distinguishing measurement sequence. Finally, we will show that one can locally find $\boldsymbol{P} \neq \boldsymbol{0}$ mapping to the same point in latent space that contains this distinguishing measurement sequence, forcing an incorrect measurement outcome on one of these states. This gives rise to an infinite backward empirical cross entropy on this task.

Let us begin by showing that $\mathcal{E}|_R$ (i.e. the locally Lipschitz map that is $\mathcal{E}$ restricted to R) must map three nontrivially distinct $\boldsymbol{Q}$ (i.e. three distinct CV graph states) to the same point in latent space. We will consider the restriction $\mathcal{E}|_K$, which is the (locally Lipschitz) restriction of $\mathcal{E}$ to points of the form $(\boldsymbol{Q}, \boldsymbol{0}) \in R$. We will show that this map is not injective for small enough $\dim(L)$. As $\mathcal{E}|_R$ lifts to $\mathcal{E}|_K$, this will show that three distinct $\boldsymbol{Q}$ map to the same point in latent space under $\mathcal{E}|_R$. This is similar to the construction for $\mathcal{F}|_K$ in the proof Theorem D.2; we repeat it here for completeness.

By the constant rank theorem and the local Lipschitzness of $\mathcal{E}|_K$, $\mathcal{E}|_K$ is not injective for

$$\dim(L) < \dim(K) = \frac{n(n-1)}{2}. \tag{D.36}$$

In particular, in a sufficiently small neighborhood of $\boldsymbol{B} = \boldsymbol{0}$ (where the Jacobian of $\mathcal{E}|_K$ attains its maximal rank), there exist local coordinates $\tilde{\boldsymbol{x}}$ of K and L such that

$$\mathcal{E}|_K \left(\tilde{x}_1, \ldots, \tilde{x}_{\frac{n(n-1)}{2}}\right) = (\tilde{x}_1, \ldots, \tilde{x}_l, 0, \ldots, 0) \tag{D.37}$$

for some $l \leq \dim(L) < \frac{n(n-1)}{2}$ [169]. WLOG, we identify $\tilde{\boldsymbol{x}} = \boldsymbol{0}$ with $\boldsymbol{B} = \boldsymbol{0}$, which is the state infinitely squeezed in all $\hat{q}_i$. We will call C the fiber with local coordinates

$$\tilde{\boldsymbol{x}} = \left(0, \ldots, 0, \tilde{x}_{l+1}, \ldots, \tilde{x}_{\frac{n(n-1)}{2}}\right), \tag{D.38}$$

which is of dimension at least

$$\Delta \equiv \frac{n\,(n-1)}{2} - \dim\,(L) \geq 1. \tag{D.39}$$

By construction, all points in C—including $\boldsymbol{B} = \boldsymbol{0}$—map to the same point $l \in L$ under $\mathcal{E}|_K$. We now assume that $\dim\,(L) < \frac{n(n-3)}{2}$ such that $\Delta \geq n+1$.

Now fix $\boldsymbol{B} = \boldsymbol{0}$ and $\boldsymbol{B'} \neq \boldsymbol{B}$ in C. As described previously, $\boldsymbol{Q}$ (and thus $\boldsymbol{B}$) completely determines a CV graph state after n measurements, up to trivial rescalings of the rows of $\boldsymbol{Q}$ and the measurement results. As the dimension of the space of points that differ (modulo π) from $\boldsymbol{B'} + \boldsymbol{H}/2$ by just a scaling factor in each row is at most n, because $\Delta \geq n+1$ we must have that there exists another $\boldsymbol{B''} \neq \boldsymbol{B}, \boldsymbol{B'}$ describing a distinct CV graph state. Therefore, by Lemma D.1, we have that there exists a distinguishing measurement s that is a stabilizer of the state corresponding to $\boldsymbol{B} = \boldsymbol{0}$. Note that as Lemma D.1 does not depend on the phases of the CV stabilizers, the existence of this distinguishing measurement holds true regardless of what the measurement results are (i.e. independently from what the model outputs for the first n tokens in the decoded sequence). Depending on these measurement results, however, the distinguishing measurement could be one of three different measurement sequences, depending on the validity of Equation (D.23).

We have now shown that there exists $(\boldsymbol{Q}_i, \boldsymbol{0})$ for $1 \leq i \leq 3$ such that all $\boldsymbol{Q}_i$ are distinct, and that there exists a measurement of a stabilizer of $\boldsymbol{Q}_1$ such that the post-measurement states of $\boldsymbol{Q}_2$, $\boldsymbol{Q}_3$ are orthogonal. We will now show that there exist $\boldsymbol{P}_i$ such that $(\boldsymbol{Q}_i, \boldsymbol{P}_i)$ also maps to the same point in latent space, and $\boldsymbol{P}_i$ is a distinguishing measurement sequence. This will give the final separation.

First, note that one can find a $\boldsymbol{P}_i$ given an arbitrarily small bound on its norm that encodes a distinguishing measurement sequence; this is because, in the proof of Lemma D.1, one can repeatedly take $\alpha \to 3\alpha$ to yield a distinguishing measurement sequence using arbitrarily large stabilizer powers, which corresponds to arbitrarily small $\boldsymbol{P}_i$ in the modified stabilizer measurement translation task. Now, consider $\mathcal{E}|_S$, defined as the restriction of $\mathcal{E}$ to points of the form $(\boldsymbol{Q}_i, \boldsymbol{P}) \in R$ for any of the

fixed $\boldsymbol{Q}_i \in K$, where $\boldsymbol{P}$ is zero in all rows except for the last S rows. Initially, let $S = S_1 = 2$; that is, $\boldsymbol{P}$ is restricted to be all zero except for its final two rows, which are allowed to vary. Let d_{S_1} be the dimension of the image of $\mathcal{E}|_S$. If $d_{S_1} = 0$, then $\mathcal{E}$ is locally independent from the final two rows of $\boldsymbol{P}$ at $\boldsymbol{Q}_i, m_i$ when all other rows are fixed to be zero, and we are done as we can set these two rows to be the distinguishing measurement sequence. If $d_{S_1} > 0$, then consider $S = S_2 = 4$. If $d_{S_2} = d_{S_1}$, then for all local choices of the third and fourth final rows, there exists a choice of the final two rows such that $\mathcal{E}$ is constant. Then, all of the possible distinguishing measurement scenarios can be encoded into the third and fourth final rows of $\boldsymbol{P}$ (with the appropriate choice of final two rows) and map to the same point in latent space as when $\boldsymbol{P} = \boldsymbol{0}$, yielding the appropriate distinguishing measurement sequence.

If $d_{S_2} > d_{S_1}$ instead, we are able to iterate this procedure once more. As $\dim(L) < \frac{n(n-3)}{2}$, eventually this iteration will stop with $d_{S_i} = d_{S_{i-1}}$, and we have the freedom to set two rows of $\boldsymbol{P}$ to be the distinguishing measurement sequence and map to the same point in latent space as $\boldsymbol{P} = \boldsymbol{0}$.

We have thus shown that the encoder-decoder model must map three points in R that give rise to a distinguishing measurement scenario to the same point in latent space. That is, when $\dim(L) < \frac{n(n-3)}{2}$, there exists input sequences $\boldsymbol{s}_i$ that map to the same point in the latent space of the model that must give rise to orthogonal measurement results. As they are mapped to the same point in the latent space of the model, the model must output the same translation for all of them, giving at least one incorrect result. Thus, the backward empirical cross entropy when translating one of these $\boldsymbol{s}_i$ must be infinite. □

D.3 Considerations for Experimental Implementations

We have shown in Section D.2 that CRNNs are more expressive than both online and encoder-decoder sequence models, assuming a locally Lipschitz condition on the models (and an additional technical condition on the latter class of models). Though CRNNs only utilize Gaussian operations—which are believed to be much simpler to implement that universal CV quantum computing [161]—our model also utilizes non-Gaussian ancilla states to perform non-Gaussian measurements. As we formally compare our quantum model with infinite precision classical models, we have a formal requirement for GKP ancilla states with infinite homodyne precision measurements to show a separation.

Of course, in practice, classical neural sequence models are finite precision. In particular, tensor processing unit (TPU) implementations of classical neural networks often use as imprecise as 8-bit arithmetic. We therefore expect that one can circumvent the formal need for GKP states to use *qubit* ancilla states to perform phase estimation of the CV stabilizer operators to a precision matching that of classical neural networks. There are proposals for engineering the required longitudinal photon/qubit interactions using circuit quantum electrodynamics (QED) [250, 251]. As similar couplings are already used in proposals for the generation of approximate GKP states [252], this direct approach is likely more experimentally feasible. Furthermore, such a finite precision implementation may actually *gain* expressive power compared to infinite precision CRNNs. Assuming the back action of the finite precision measurement yields a finitely squeezed state in the CRNN, with this one can construct a model capable of universal CV quantum computation [163, 164]. This also holds when the initial state of the model is taken to be the vacuum state (or any other finitely squeezed Gaussian state). This suggests a potential superpolynomial advantage in the expressive power and the time complexity of inference when this model is implemented at finite precision.

If one wishes to avoid coupling to qubits, we make the bolder conjecture that

any non-Gaussian ancilla state is enough to yield a separation. Our intuition for this comes from recent work [170, 171] demonstrating that non-Gaussian operations are equivalent to the presence of quantum contextuality. As the presence of quantum contextuality is the source of the separations in our proofs, this gives evidence that one could use a more experimentally feasible non-Gaussian ancilla state than a GKP state—such as a photon subtracted state [253]—and achieve similar results.

D.4 Classical Simulation of Gaussian Operations and GKP States

We now describe the high level strategy of classically simulating both Gaussian operations applied to Gaussian states, and (restricted) Gaussian operations applied to GKP states. The former strategy will roughly follow that given in Reference [254], and the latter that given in Reference [255]. These will be the building blocks of the Gaussian RNN and contextual RNN cells we numerically test in the main text; we give the details of the full architectures, including the classical architectures, in Section D.5.

D.4.1 Gaussian States

First, we describe the simulation of Gaussian operations performed on Gaussian states. It is well known that any N mode Gaussian pure state can be created from the ground state of N harmonic oscillators with unitary operations $\mathrm{e}^{\mathrm{i}f(\hat{\boldsymbol{q}},\hat{\boldsymbol{p}})}$, where $f(\hat{\boldsymbol{q}},\hat{\boldsymbol{p}})$ are at most quadratic in terms of the quadrature operators $\hat{\boldsymbol{q}}$ and $\hat{\boldsymbol{p}}$. By stacking $\hat{\boldsymbol{q}}$ and $\hat{\boldsymbol{p}}$ together, we call $\hat{\boldsymbol{x}} = (\hat{\boldsymbol{q}},\hat{\boldsymbol{p}})^{\intercal}$ the quadrature operator. The Heisenberg evolution of the quadrature operator under a Gaussian unitary operator $\boldsymbol{R}$ is in general:

$$\hat{\boldsymbol{x}'} = \boldsymbol{R}^{\dagger}\hat{\boldsymbol{x}}\boldsymbol{R} = \boldsymbol{S}\hat{\boldsymbol{x}} + \boldsymbol{c}, \tag{D.40}$$

where $\boldsymbol{S}$ is a symplectic matrix of c-numbers, and $\boldsymbol{c} = (\boldsymbol{c_q},\boldsymbol{c_p})^{\intercal}$ is a vector of c-numbers denoting the mode center shift. There are various ways to decompose a

symplectic matrix $\boldsymbol{S}$; WLOG (as discussed in Reference [254]), we will consider symplectic matrices of the form:

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{U}^{-1/2} & \boldsymbol{0} \\ \boldsymbol{V}\boldsymbol{U}^{-1/2} & \boldsymbol{U}^{1/2} \end{pmatrix}, \tag{D.41}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are both real symmetric matrices. In addition, $\boldsymbol{U}$ is positive definite, i.e. $\boldsymbol{U} = \boldsymbol{U}^{\intercal} > 0$. If the covariance matrix of the N mode ground state is

$$\mathrm{cov}\left(\hat{\boldsymbol{x}}_0\right) = \frac{1}{2}\boldsymbol{I}, \tag{D.42}$$

then the covariance matrix of the transformed Gaussian state is:

$$\begin{aligned} \boldsymbol{\Sigma} = \mathrm{cov}\left(\hat{\boldsymbol{x}}\right) = \mathrm{cov}\left(\boldsymbol{S}\hat{\boldsymbol{x}}_0\right) &= \frac{1}{2}\left\langle \left\{ \left(\boldsymbol{S}\hat{\boldsymbol{x}}_0\right)^{\dagger}, \left(\boldsymbol{S}\hat{\boldsymbol{x}}_0\right)^{\intercal} \right\} \right\rangle \\ = \frac{1}{2}\boldsymbol{S}\boldsymbol{S}^{\intercal} &= \frac{1}{2}\begin{pmatrix} \boldsymbol{U}^{-1} & \boldsymbol{U}^{-1}\boldsymbol{V} \\ \boldsymbol{V}\boldsymbol{U}^{-1} & \boldsymbol{U} + \boldsymbol{V}\boldsymbol{U}^{-1}\boldsymbol{V} \end{pmatrix}. \end{aligned} \tag{D.43}$$

We can also write down its wavefunction in position space as:

$$\begin{aligned} \psi_{\boldsymbol{Z},\boldsymbol{c}}\left(\boldsymbol{q}\right) &= \pi^{-N/4}\left(\det\left(\boldsymbol{U}\right)\right)^{1/4}\exp\left(-\frac{1}{2}\left(\boldsymbol{q}-\boldsymbol{c}_{\boldsymbol{q}}\right)^{\intercal}\left(\boldsymbol{U}-\mathrm{i}\boldsymbol{V}\right)\left(\boldsymbol{q}-\boldsymbol{c}_{\boldsymbol{q}}\right)\right) \\ &= \pi^{-N/4}\left(\det\left(\boldsymbol{U}\right)\right)^{1/4}\exp\left(\frac{\mathrm{i}}{2}\left(\boldsymbol{q}-\boldsymbol{c}_{\boldsymbol{q}}\right)^{\intercal}\boldsymbol{Z}\left(\boldsymbol{q}-\boldsymbol{c}_{\boldsymbol{q}}\right)\right), \end{aligned} \tag{D.44}$$

where $\boldsymbol{q}$ and $\boldsymbol{c}_{\boldsymbol{q}}$ are c-number column vectors, and $\boldsymbol{Z} = \boldsymbol{V} + \mathrm{i}\boldsymbol{U}$ is a complex symmetric matrix. We can interpret $\boldsymbol{Z}$ as the adjacency matrix for an undirected graph with complex-valued edge weights. Therefore, any Gaussian pure states can be interpreted as a Gaussian graph states with complex-valued weights on the graph edge. As Gaussian states are uniquely identified by the linear combinations of position and momentum operators that nullify them [256], we can consider what the nullifiers are

for the CV graph state defined by $\boldsymbol{Z}$ and $\boldsymbol{c} = (\boldsymbol{c_q}, \boldsymbol{c_p})^\intercal$:

$$\begin{aligned}
(\hat{\boldsymbol{p}} - \boldsymbol{Z}\hat{\boldsymbol{q}} + \boldsymbol{Z}\boldsymbol{c_q} - \boldsymbol{c_p}) \left|\psi_{\boldsymbol{Z},\boldsymbol{c}}\right\rangle =& (\hat{\boldsymbol{p}} - \boldsymbol{Z}\hat{\boldsymbol{q}} + \boldsymbol{Z}\boldsymbol{c_q} - \boldsymbol{c_p}) \boldsymbol{R}_{\boldsymbol{Z},\boldsymbol{c}} \left|0\right\rangle \\
=& \boldsymbol{R}_{\boldsymbol{Z},\boldsymbol{c}} \boldsymbol{R}_{\boldsymbol{Z},\boldsymbol{c}}^\dagger (\hat{\boldsymbol{p}} - \boldsymbol{Z}\hat{\boldsymbol{q}} + \boldsymbol{Z}\boldsymbol{c_q} - \boldsymbol{c_p}) \boldsymbol{R}_{\boldsymbol{Z},\boldsymbol{c}} \left|0\right\rangle \\
=& \boldsymbol{R}_{\boldsymbol{Z},\boldsymbol{c}} \begin{pmatrix} -\boldsymbol{Z} & \boldsymbol{I} \end{pmatrix} (\boldsymbol{S}_{\boldsymbol{Z}}\hat{\boldsymbol{x}} + \boldsymbol{c}) \left|0\right\rangle + (\boldsymbol{Z}\boldsymbol{c_q} - \boldsymbol{c_p}) \boldsymbol{R}_{\boldsymbol{Z},\boldsymbol{c}} \left|0\right\rangle \\
=& \boldsymbol{R}_{\boldsymbol{Z},\boldsymbol{c}} \begin{pmatrix} -\boldsymbol{Z} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{U}^{-1/2}\hat{\boldsymbol{q}} + \boldsymbol{c_q} \\ \boldsymbol{V}\boldsymbol{U}^{-1/2}\hat{\boldsymbol{q}} + \boldsymbol{U}^{1/2}\hat{\boldsymbol{p}} + \boldsymbol{c_p} \end{pmatrix} \left|0\right\rangle \\
& + (\boldsymbol{Z}\boldsymbol{c_q} - \boldsymbol{c_p}) \boldsymbol{R}_{\boldsymbol{Z},\boldsymbol{c}} \left|0\right\rangle \\
=& \boldsymbol{R}_{\boldsymbol{Z},\boldsymbol{c}} \left(-\mathrm{i}\boldsymbol{U}^{1/2}\hat{\boldsymbol{q}} + \boldsymbol{U}^{1/2}\hat{\boldsymbol{p}} - \boldsymbol{Z}\boldsymbol{c_q} + \boldsymbol{c_p}\right) \left|0\right\rangle \\
& + (\boldsymbol{Z}\boldsymbol{c_q} - \boldsymbol{c_p}) \boldsymbol{R}_{\boldsymbol{Z},\boldsymbol{c}} \left|0\right\rangle \\
=& \boldsymbol{R}_{\boldsymbol{Z},\boldsymbol{c}} (-\boldsymbol{Z}\boldsymbol{c_q} + \boldsymbol{c_p}) \left|0\right\rangle + (\boldsymbol{Z}\boldsymbol{c_q} - \boldsymbol{c_p}) \boldsymbol{R}_{\boldsymbol{Z},\boldsymbol{c}} \left|0\right\rangle \\
=& (-\boldsymbol{Z}\boldsymbol{c_q} + \boldsymbol{c_p}) \boldsymbol{R}_{\boldsymbol{Z},\boldsymbol{c}} \left|0\right\rangle + (\boldsymbol{Z}\boldsymbol{c_q} - \boldsymbol{c_p}) \boldsymbol{R}_{\boldsymbol{Z},\boldsymbol{c}} \left|0\right\rangle \\
=& \boldsymbol{0}.
\end{aligned} \tag{D.45}$$

In the second to last line, we have use the fact that $-\boldsymbol{Z}\boldsymbol{c_q} + \boldsymbol{c}_p$ is a c-number vector, which commutes with $\boldsymbol{R}_{\boldsymbol{Z},\boldsymbol{c}}$. In addition, we also use the fact that $\boldsymbol{Z} = \boldsymbol{V} + \mathrm{i}\boldsymbol{U}$, and $(-\mathrm{i}\hat{\boldsymbol{q}} + \hat{\boldsymbol{p}}) \left|0\right\rangle = -\mathrm{i}\sqrt{2}\hat{\boldsymbol{a}} \left|0\right\rangle = 0$. Therefore, the Gaussian graph state with complex adjacency matrix $\boldsymbol{Z}$ and mode shift center $\boldsymbol{c}$ has complex nullifiers:

$$(\hat{\boldsymbol{p}} - \boldsymbol{c_p} - \boldsymbol{Z}(\hat{\boldsymbol{q}} - \boldsymbol{c_q})) \left|\psi_{\boldsymbol{Z},\boldsymbol{c}}\right\rangle = \boldsymbol{0}. \tag{D.46}$$

We now restrict to the case $\boldsymbol{V} = \boldsymbol{0}$ and $\boldsymbol{c_p} = \boldsymbol{0}$, which are the class of states we consider in our numerical experiments. Then, the nullifiers are of the form:

$$(\hat{\boldsymbol{p}} - \mathrm{i}\boldsymbol{U}\hat{\boldsymbol{q}} + \mathrm{i}\boldsymbol{U}\boldsymbol{c_q}) \left|\psi_{\boldsymbol{Z},\boldsymbol{c}}\right\rangle = \boldsymbol{0}. \tag{D.47}$$

We also restrict to symplectic transformations of the quadratures of the form:

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{W}^\intercal & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{W}^{-1} \end{pmatrix}, \tag{D.48}$$

where $\boldsymbol{W}$ is some assumed invertible matrix. We restrict to operations of this form to more efficiently allow for the classical simulation of the contextual RNN using Gaussian operations of an identical form, as discussed in Reference [255] and Section D.4.2. After performing the quantum operation described by such a symplectic matrix, the nullifier for the whole system is updated as $(\boldsymbol{W}^{-1}\hat{\boldsymbol{p}} - \mathrm{i}\boldsymbol{U}\boldsymbol{W}^{\intercal}\hat{\boldsymbol{q}} + \mathrm{i}\boldsymbol{U}\boldsymbol{c_q})\,|\psi\rangle = \boldsymbol{0}$, which is equivalent to $(\hat{\boldsymbol{p}} - \mathrm{i}\boldsymbol{W}\boldsymbol{U}\boldsymbol{W}^{\intercal}\hat{\boldsymbol{q}} + \mathrm{i}\boldsymbol{W}\boldsymbol{U}\boldsymbol{c_q})\,|\psi\rangle = \boldsymbol{0}$. Homodyne detection of the position quadrature on m qumodes is then, in general, a multivariate Gaussian random variable which is centered at $\boldsymbol{\Pi_Y}\boldsymbol{W}^{-1\intercal}\boldsymbol{c_q}$ with variance $\boldsymbol{\Pi_Y}\boldsymbol{U}^{-1}\boldsymbol{\Pi_Y}^{\intercal}$, where $\boldsymbol{\Pi_Y}$ is the projection operator onto the subspace of the m qumodes being measured. To make the training of our Gaussian models more stable—and to maintain simulability with GKP input states, as is done in Section D.4.2—we assume in our numerical simulations that there is an implicit large scaling for $\boldsymbol{U}, \boldsymbol{c_q}$ such that this variance is small. After the measurement, the hidden state at ith step is updated to a generalized graph state with adjacency matrix $\boldsymbol{\Pi_H}\boldsymbol{U}\boldsymbol{\Pi_H}^{\intercal}$ and mode shift $\boldsymbol{\Pi_H}\boldsymbol{W}^{-1\intercal}\boldsymbol{c_q}$, where $\boldsymbol{\Pi_H}$ is the projection operator onto the subspace of the latent n qumodes.

D.4.2 Gaussian Operations on GKP States

Our methods for the simulation of (restricted) Gaussian operations on GKP states are similar to the methods used by Reference [255]. We restrict to (unnormalized) states of the form:

$$|\psi\rangle = \sum_{\boldsymbol{\ell}\in\mathbb{Z}^n} |\psi_{\boldsymbol{\ell}}\rangle\,, \tag{D.49}$$

where each $|\psi_{\boldsymbol{\ell}}\rangle$ is a Gaussian state with nullifiers given by:

$$\hat{\boldsymbol{n}}_{\boldsymbol{\ell}} = \epsilon\hat{\boldsymbol{p}} - \boldsymbol{Z}\,(\hat{\boldsymbol{q}} - \boldsymbol{c_q} - \boldsymbol{L}\boldsymbol{\ell})\,. \tag{D.50}$$

We assume $\epsilon \to 0^+$, such that all $|\psi_{\boldsymbol{\ell}}\rangle$ are approximately orthogonal. Time evolution is simulated as in Section D.4.1, simultaneously for all $|\psi_{\boldsymbol{\ell}}\rangle$. As all $|\psi_{\boldsymbol{\ell}}\rangle$ are approximately orthogonal, measurement is simulated via choosing $\boldsymbol{\ell}$ uniformly at random, and performing the corresponding measurement. In principle, using many measurements

(over multiple instances of the state), one can read out $\boldsymbol{L}$ and $\boldsymbol{c_q}$; these are the measurement results we use in our numerical expirements, as described in Section D.5.2. For any given measurement on a subset of the modes, the post-measurement state on the remainder of the modes is just the uniform superposition over all $\boldsymbol{\ell}$ consistent with the resulting measurement outcome $\boldsymbol{y}$.

To make this latter observation more concrete, assume after evolution under $\boldsymbol{S}$ of the form of Equation (D.48), the nullifiers of $|\psi_{\boldsymbol{\ell}}\rangle$ are (see Section D.5):

$$\hat{\boldsymbol{n}}_{\boldsymbol{\ell}} = \epsilon\hat{\boldsymbol{p}} - \boldsymbol{W}\boldsymbol{Z}\boldsymbol{W}^{\intercal}\left(\hat{\boldsymbol{q}} - \boldsymbol{W}^{-1\intercal}\boldsymbol{c_q} - \boldsymbol{W}^{-1\intercal}\boldsymbol{L}\boldsymbol{\ell}\right), \tag{D.51}$$

We wish to find all $\boldsymbol{\ell}$ consistent with the position measurement result $\boldsymbol{y}$ (in the limit $\epsilon \to 0^{+}$); that is, all $\boldsymbol{\ell}$ such that:

$$\boldsymbol{\Pi_Y}\left(\boldsymbol{W}^{-1\intercal}\boldsymbol{c_q} + \boldsymbol{W}^{-1\intercal}\boldsymbol{L}\boldsymbol{\ell}\right) = \boldsymbol{y}, \tag{D.52}$$

where $\boldsymbol{\Pi_Y}$ is the projector onto the m mode space being measured. Let H label the space of the other n modes, and the projector onto this space $\boldsymbol{\Pi_H}$. Writing (with the assumptions that $\tilde{\boldsymbol{W}}_{\boldsymbol{YY}}$ and $\boldsymbol{W_{HH}}$ are full rank):

$$\boldsymbol{W}^{-1\intercal} = \begin{pmatrix} \tilde{\boldsymbol{W}}_{\boldsymbol{HH}} & \tilde{\boldsymbol{W}}_{\boldsymbol{HY}} \\ \tilde{\boldsymbol{W}}_{\boldsymbol{YH}} & \tilde{\boldsymbol{W}}_{\boldsymbol{YY}} \end{pmatrix}, \tag{D.53}$$

$$\boldsymbol{W}^{\intercal} = \begin{pmatrix} \boldsymbol{W_{HH}} & \boldsymbol{W_{HY}} \\ \boldsymbol{W_{YH}} & \boldsymbol{W_{YY}} \end{pmatrix}, \tag{D.54}$$

$$\boldsymbol{\ell} = \begin{pmatrix} \boldsymbol{\ell_H} \\ \boldsymbol{\ell_Y} \end{pmatrix}, \tag{D.55}$$

and assuming $\boldsymbol{L}$ (assumed full rank) is of the form:

$$\boldsymbol{L} = \begin{pmatrix} \boldsymbol{L_{HH}} & 0 \\ 0 & \boldsymbol{L_{YY}} \end{pmatrix}, \tag{D.56}$$

we find from Equation (D.52) that all $\boldsymbol{\ell}$ consistent with the measurement result satisfy:

$$\boldsymbol{L_{YY}\ell_Y} = \tilde{\boldsymbol{W}}_{\boldsymbol{YY}}^{-1}\left(\boldsymbol{y} - \boldsymbol{\Pi_Y W}^{-1\intercal}\boldsymbol{c_q} - \tilde{\boldsymbol{W}}_{\boldsymbol{YH}}\boldsymbol{L_{HH}\ell_H}\right). \tag{D.57}$$

Assuming the entries of $\boldsymbol{L_{YY}}$ are sufficiently small, up to any given machine precision the $\boldsymbol{\ell_H}$ are in one-to-one correspondence with the $\boldsymbol{\ell}$ consistent with the measurement result. Furthermore, for all such $\boldsymbol{\ell}$:

$$\begin{aligned}
\boldsymbol{\Pi_H}\left(\boldsymbol{W}^{-1\intercal}\boldsymbol{c_q} + \boldsymbol{W}^{-1\intercal}\boldsymbol{L\ell}\right) =& \boldsymbol{\Pi_H W}^{-1\intercal}\boldsymbol{c_q} + \tilde{\boldsymbol{W}}_{\boldsymbol{HH}}\boldsymbol{L_{HH}\ell_H} + \tilde{\boldsymbol{W}}_{\boldsymbol{HY}}\boldsymbol{L_{YY}\ell_Y} \\
=& \boldsymbol{\Pi_H W}^{-1\intercal}\boldsymbol{c_q} + \tilde{\boldsymbol{W}}_{\boldsymbol{HH}}\boldsymbol{L_{HH}\ell_H} \\
&+ \tilde{\boldsymbol{W}}_{\boldsymbol{HY}}\tilde{\boldsymbol{W}}_{\boldsymbol{YY}}^{-1}\left(\boldsymbol{y} - \boldsymbol{c_q} - \tilde{\boldsymbol{W}}_{\boldsymbol{YH}}\boldsymbol{L_{HH}\ell_H}\right) \\
=& \boldsymbol{\Pi_H W}^{-1\intercal}\boldsymbol{c_q} + \left(\tilde{\boldsymbol{W}}_{\boldsymbol{HH}} - \tilde{\boldsymbol{W}}_{\boldsymbol{HY}}\tilde{\boldsymbol{W}}_{\boldsymbol{YY}}^{-1}\tilde{\boldsymbol{W}}_{\boldsymbol{YH}}\right)\boldsymbol{L_{HH}\ell_H} \\
&+ \tilde{\boldsymbol{W}}_{\boldsymbol{HY}}\tilde{\boldsymbol{W}}_{\boldsymbol{YY}}^{-1}\left(\boldsymbol{y} - \boldsymbol{\Pi_Y W}^{-1\intercal}\boldsymbol{c_q}\right) \\
=& \boldsymbol{\Pi_H W}^{-1\intercal}\boldsymbol{c_q} + \boldsymbol{W}_{\boldsymbol{HH}}^{-1}\boldsymbol{L_{HH}\ell_H} \\
&+ \tilde{\boldsymbol{W}}_{\boldsymbol{HY}}\tilde{\boldsymbol{W}}_{\boldsymbol{YY}}^{-1}\left(\boldsymbol{y} - \boldsymbol{\Pi_Y W}^{-1\intercal}\boldsymbol{c_q}\right).
\end{aligned} \tag{D.58}$$

Using the appropriate displacement operator to remove the final term of Equation (D.58), then, yields the effective transformation:

$$\boldsymbol{c_q} \mapsto \boldsymbol{\Pi_H W}^{-1\intercal}\boldsymbol{c_q}, \tag{D.59}$$

$$\boldsymbol{L} \mapsto \boldsymbol{W}_{\boldsymbol{HH}}^{-1}\boldsymbol{L_{HH}}. \tag{D.60}$$

D.5 Details of the Numerical Simulations

We now discuss the details of our numerical simulations. For all models, we studied the performance in modeling a standard Spanish-to-English data set [156]. In the main text, we considered the performance over five independent training runs for each model. For each model and each n, the training set was taken to be a random sample of 80% of the data set, and the test set 20%. Each model was trained for 80 epochs, with a batch size of 64. To map the words in this data set to vectors (taken to

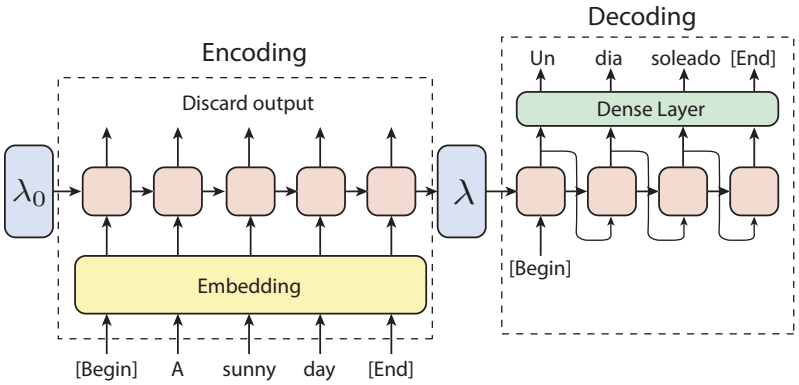


Figure D-3: An overview of the recurrent models we study. Each red box represents the recurrent cell, and the variational parameters in the recurrent cells are shared within the encoder and decoder. $\boldsymbol{\lambda_0}$ is a random fixed hidden memory vector. In the decoding phase, the output of each recurrent cell is also treated as the input for the next recurrent cell.

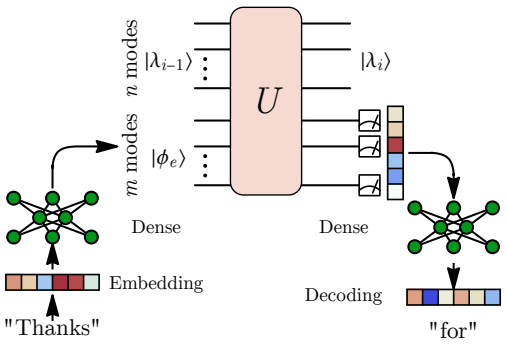


Figure D-4: One recurrent cell of the quantum recurrent architectures. Note that the only trained part of the dense network at the input of each recurrent cell are displacements in phase space acting on $|\phi_e\rangle$, in order to keep the number of trainable parameters in line with the GRU RNN at the same n. $|\phi_e\rangle$ is a Gaussian state for the Gaussian model, and a GKP state for the CRNN.

also be of dimension n, the model dimension), we used the Keras [257] implementation of word2vec ("TextVectorization") adapted to the data, with a maximum vocabulary size of 5000. The first 5000 most frequent words are mapped to distinct integers, and other words are mapped to a unique token "[Unk]." At the beginning and end of each sentence, we add unique "[Begin]" and "[End]" tokens. For the recurrent translation models, such as the GRU RNN or the CRNN, we do not need to set the length of the sentences. For the Transformer model, we set the sentence length to be 20 words. If the sentence is shorter than 20 words, the additional token "[Pad]" is added to the sentence. For the rare case when the sentence contains more than 20 words, additional words are truncated. All networks were trained using Adam [128] (with a learning rate of 10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-7}$), trained on the forward empirical cross entropy (as the backward empirical cross entropy is difficult to train on).

D.5.1 Classical Sequence Models

We studied three standard classical sequence models in our numerical experiments: an implementation of an orthogonal recurrent neural network [155], an implementation of a network using gated recurrent units (GRU) [88], and an implementation of a Transformer [89]. The first two networks were trained in a seq2seq configuration [84]; the models autoregressively map the input sequence to a latent space, which then is autoregressively decoded. For the orthogonal recurrent neural network, we used the implementation of Reference [155], with network capacity equal to the model dimension. An illustration of these architectures is given in Figure D-3.

The Transformer models we considered follow the standard construction of Reference [89]. To fairly compare against the shallow RNN cells we consider, we used a single Transformer encoder and decoder layer for each model. Our implementation used a trained positional embedding with uniform initialization, and the encoders and decoders used ReLU activations in the feedforward network layers, Glorot weight initialization, zero bias initialization, no dropout, and a single head. Each encoder and decoder layer was followed by the layer normalization implementation of Keras [257] with its default parameters. The final layer normalization of the decoder of each

Transformer we considered was followed by a dense layer with a softmax activation function.

D.5.2 Quantum Sequence Models

Based on the discussion in Section D.4, we simulated both the Gaussian RNN and the contextual RNN described in the main text. The training and architecture of these models were identical to those of the orthogonal neural network described in Section D.5.1, other than the structure of each unit cell of the recurrent network. Once again, Figure D-3 describes the overall architecture of the models, and Figure D-4 describes the recurrent cell of these models. For the Gaussian model, the simulated homodyne position measurements are what we used for our cell outputs; for the CRNN, we simulated the lattice (and position measurement) readout procedure described at the end of Section D.4.2. The pseudocode for the cells of both are given in Algorithm 1; for the Gaussian model, one takes $\boldsymbol{J}_0, \boldsymbol{K} = \boldsymbol{0}$. By considering the Gaussian RNN cell as a limit of the contextual RNN cell, and by not training $\boldsymbol{K}$ or $\boldsymbol{J}_0$ in the contextual RNN, we were able to maintain identical parameter counts for the two models. For the contextual RNN, $\boldsymbol{J}_0$ is fixed to the identity, and $\boldsymbol{K}$ is the result of an untrained dense layer $\boldsymbol{h}$ applied to the cell input (with uniform Glorot initialization and linear activation function, and biased such that $\boldsymbol{K}$ has mean the identity). Specializing to the notation of Algorithm 1: $\boldsymbol{r}$ is similar. $\boldsymbol{f}$ and $\boldsymbol{g}$ are similar, except with no bias.

D.5.3 Time Complexity

We now discuss the time complexity of implementing a CRNN, both as a quantum model implemented on a quantum computer, and as a quantum-inspired classical algorithm. On a quantum computer, each cell of the CRNN we consider in our proofs of an expressivity separation can be implemented in depth $\mathrm{O}(n)$, assuming access to the fixed ancilla state $|a\rangle$ used for non-Gaussian measurement. This further decreases to $\mathrm{O}(1)$ time if one assumes access to quantum fan-out [258]. More general Gaussian

Algorithm 1: Contextual RNN Cell

Input: `// cell inputs`
$n \times n$ latent graph adjacency matrix $\boldsymbol{A}_{i-1}$
$n \times n$ lattice $\boldsymbol{J}_{i-1}$
$n \times 1$ stabilizer phases $\boldsymbol{\alpha}_{i-1}$
$m \times 1$ input $\boldsymbol{x}_i$

`// trainable parameters`
$(n+m) \times (n+m)$ weight matrix $\boldsymbol{W}$
Dense layers $\boldsymbol{f}, \boldsymbol{g}$

`// constants`
Dense layers $\boldsymbol{h}, \boldsymbol{r}$
Projectors onto latent and input spaces, respectively, $\boldsymbol{\Pi_H}, \boldsymbol{\Pi_Y}$

Output: $n \times n$ latent graph adjacency matrix $\boldsymbol{A}_i$
$n \times n$ lattice $\boldsymbol{J}_i$
$n \times 1$ stabilizer phases $\boldsymbol{\alpha}_i$
$m \times 1$ measurement outcome $\boldsymbol{y}$

begin

$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha}_{i-1} + \boldsymbol{A}_{i-1}^{-1} \boldsymbol{f}(\boldsymbol{x}_i)$; `// perform mode shifts using a general function` f

$\boldsymbol{\beta} \leftarrow \boldsymbol{g}(\boldsymbol{x}_i)$; `// prepare the state associated with the input register`
$\boldsymbol{K} \leftarrow \boldsymbol{h}(\boldsymbol{x}_i)$;
$\boldsymbol{S} \leftarrow \boldsymbol{r}(\boldsymbol{x}_i)$;
$\boldsymbol{B} = \boldsymbol{S}\boldsymbol{S}^\intercal$; `// ensure the adjacency matrix is positive semidefinite`

$\boldsymbol{U} \leftarrow \boldsymbol{A}_{i-1} \oplus \boldsymbol{B}$; `// consider the tensor product of the latent and input states`
$\boldsymbol{\gamma} \leftarrow \boldsymbol{\alpha} \oplus \boldsymbol{\beta}$;
$\boldsymbol{L} \leftarrow \boldsymbol{J}_{i-1} \oplus \boldsymbol{K}$;

$\boldsymbol{U} \leftarrow \boldsymbol{W}\boldsymbol{U}\boldsymbol{W}^\intercal$; `// transform the graph state by performing a Gaussian operation`
$\boldsymbol{\gamma} \leftarrow \boldsymbol{W}^{-1\intercal}\boldsymbol{\gamma}$;
$\boldsymbol{L} \leftarrow \boldsymbol{W}^{-1\intercal}\boldsymbol{L}$;

$\boldsymbol{y} \leftarrow \boldsymbol{\Pi_Y}\boldsymbol{L}, \boldsymbol{\Pi_Y}\boldsymbol{\gamma}$; `// read out the lattice and stabilizer phases`

$\boldsymbol{A}_i \leftarrow \boldsymbol{\Pi_H}\boldsymbol{U}\boldsymbol{\Pi_H}^\intercal$; `// project out the measured register`
$\boldsymbol{J}_i \leftarrow (\boldsymbol{\Pi_H}\boldsymbol{W}\boldsymbol{\Pi_H}^\intercal)^{-1}\boldsymbol{J}_{i-1}$;
$\boldsymbol{\alpha}_i \leftarrow \boldsymbol{\Pi_H}\boldsymbol{\gamma}$;

operations can also be implemented in depth $\mathrm{O}(n)$ utilizing a swap network [259].

Examining Algorithm 1, it is easy to see that our algorithm simulates inference on a CRNN with model dimension n with time complexity $\mathrm{O}(n^\omega)$. Here, ω is the matrix multiplication exponent, with best-known bounds $2 \leq \omega < 2.37286$ [260]. Our results show that on certain tasks, such a CRNN of model dimension n performs on par with e.g. GRU RNNs with model dimension $\Omega(n^2)$, which performs inference in time $\Omega(n^4)$ due to matrix-vector multiplications present in the model [88]. Thus, our classical simulation of CRNNs may be thought of as a quantum-inspired classical model that, though it is not efficient as implementing a CRNN on a quantum computer, is asymptotically more time efficient in inference and training than typical RNNs with an n^2-dimensional latent space. Of course, our Algorithm 1 relies on matrix inversion. Though asymptotically matrix inversion takes time $\mathrm{O}(n^\omega)$, unlike matrix multiplication it has poor GPU implementations and thus often is slow in practice. We leave further investigation of the practical utility of these quantum-inspired classical models to future work.

D.6 Supplementary Numerical Results

We now provide supplemental numerical experiments, comparing CRNNs with Transformers [89] and a formulation of linear RNNs dubbed efficient unitary neural networks (EUNNs) [155].

The difficulty in comparing CRNNs and Transformers is that the effective memory of a Transformer—in the language of Figure D-1(b), the dimension n of the latent space of the model—grows with the sentence length. Thus, we fixed a trained $n = 26$ CRNN, and compare the performance of a Transformer at a variety of model dimensions against this model. This also allows us to test for the Spanish-to-English translation task whether, given a fixed performance target achieved by a CRNN of model dimension n, a Transformer is only able to achieve the same target given a memory dimension of $\frac{n(n-3)}{2}$; this is the separation we prove in Theorem D.4 on the constructed translation task we consider there. We plot these results in Figure D-5,

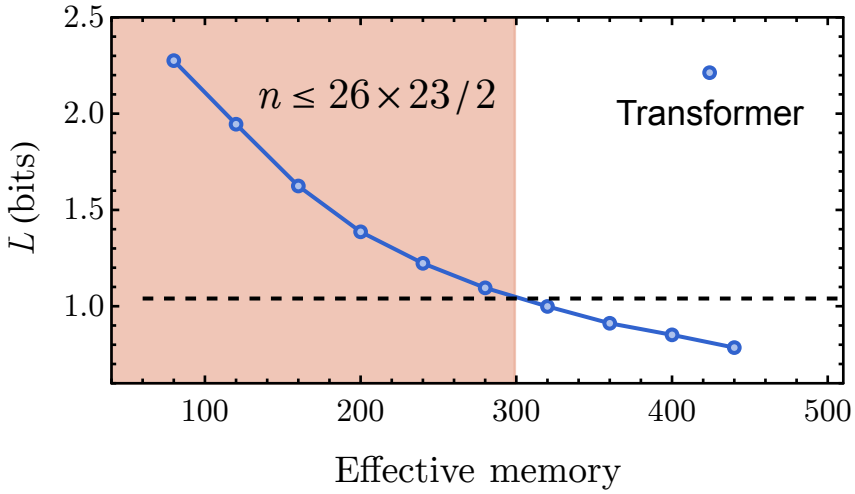


Figure D-5: The performances of an $n = 26$ CRNN model and the Transformer models. The dashed line shows the converged forward empirical cross entropy (L) for the $n = 26$ CRNN model. The dimension of the latent space (labeled "Effective memory") of a Transformer is the model dimension multiplied by the length of the (padded) input sentences. The red region labeled "$n \leq 26 \times 23/2$" is where the dimension of the memory of the Transformer is at most $\frac{n(n-3)}{2}$, where n is the model dimension of the CRNN. We find that the Transformer models with dimensions roughly equal to $\frac{n(n-3)}{2}$ achieve losses approximately equal to that of the CRNN.

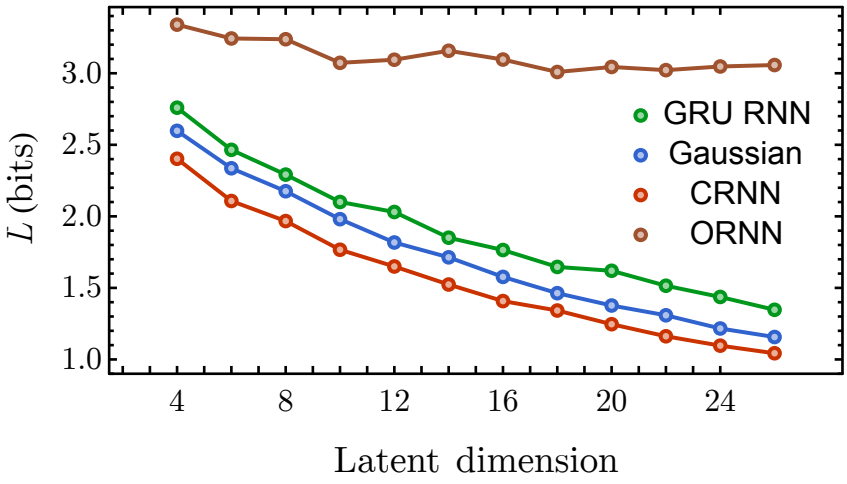


Figure D-6: The converged forward empirical cross entropy (L) as a function of the model dimensions n for ORNNs, and the online models we considered in the main text. We see that ORNNs are greatly outperformed by the other online models we consider at the given task.

where we indeed find that the performances of the Transformer models achieve that of the $n = 26$ CRNN when their memory is of dimension roughly $\frac{n(n-3)}{2}$.

We also consider the performance of EUNNs compared with CRNNs, as both are linear models. We constrain the EUNN to be real—as in our simulations of Gaussian models and CRNNs—and call the resulting model an *orthogonal recurrent neural network* (ORNN), using the implementation from Reference [155]. We see in Figure D-6 that CRNNs—and indeed, all of the online models we consider—greatly outperformed ORNNs at a variety of model dimensions.

Bibliography

[1] R. P. Feynman, Simulating physics with computers, Int. J. Theor. Phys. **21**, 467 (1982).

[2] D. S. Abrams and S. Lloyd, Simulation of many-body Fermi systems on a universal quantum computer, Phys. Rev. Lett. **79**, 2586 (1997).

[3] P. W. Shor, Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer, SIAM J. Comput. **26**, 1484 (1997).

[4] L. 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, edited by G. L. Miller (Association for Computing Machinery, New York, 1996) pp. 212–219.

[5] A. W. Harrow, A. Hassidim, and S. Lloyd, Quantum algorithm for linear systems of equations, Phys. Rev. Lett. **103**, 150502 (2009).

[6] P. W. Shor, Scheme for reducing decoherence in quantum computer memory, Phys. Rev. A **52**, R2493 (1995).

[7] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition* (Cambridge University Press, Cambridge, 2010).

[8] F. Arute *et al.*, Quantum supremacy using a programmable superconducting processor, Nature **574**, 505 (2019).

[9] H.-S. Zhong *et al.*, Quantum computational advantage using photons, Science **370**, 1460 (2020).

[10] Y. Wu *et al.*, Strong quantum computational advantage using a superconducting quantum processor, Phys. Rev. Lett. **127**, 180501 (2021).

[11] H.-S. Zhong *et al.*, Phase-programmable Gaussian boson sampling using stimulated squeezed light, Phys. Rev. Lett. **127**, 180502 (2021).

[12] Q. Zhu *et al.*, Quantum computational advantage via 60-qubit 24-cycle random circuit sampling, Sci. Bull. **67**, 240 (2022).

[13] D. Hangleiter and J. Eisert, Computational advantage of quantum random sampling (2022), arXiv:2206.04079 [quant-ph] .

[14] J. Preskill, Quantum computing in the NISQ era and beyond, Quantum **2**, 79 (2018).

[15] E. Farhi and H. Neven, Classification with quantum neural networks on near term processors (2018), arXiv:1802.06002 [quant-ph] .

[16] M. H. Amin, E. Andriyash, J. Rolfe, B. Kulchytskyy, and R. Melko, Quantum Boltzmann machine, Phys. Rev. X **8**, 021050 (2018).

[17] L. Hu, S.-H. Wu, W. Cai, Y. Ma, X. Mu, Y. Xu, H. Wang, Y. Song, D.-L. Deng, C.-L. Zou, and L. Sun, Quantum generative adversarial learning in a superconducting quantum circuit, Sci. Adv. **5**, eaav2761 (2019).

[18] E. R. Anschuetz and C. Zanoci, Near-term quantum-classical associative adversarial networks, Phys. Rev. A **100**, 052327 (2019).

[19] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, A variational eigenvalue solver on a photonic quantum processor, Nat. Commun. **5**, 4213 (2014).

[20] M. Schuld, I. Sinayskiy, and F. Petruccione, An introduction to quantum machine learning, Contemp. Phys. **56**, 172 (2015).

[21] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Quantum machine learning, Nature **549**, 195 (2017).

[22] A. Perdomo-Ortiz, M. Benedetti, J. Realpe-Gómez, and R. Biswas, Opportunities and challenges for quantum-assisted machine learning in near-term quantum computers, Quantum Sci. Technol. **3**, 030502 (2018).

[23] S. Lloyd and C. Weedbrook, Quantum generative adversarial learning, Phys. Rev. Lett. **121**, 040502 (2018).

[24] M. Schuld and N. Killoran, Quantum machine learning in feature Hilbert spaces, Phys. Rev. Lett. **122**, 040504 (2019).

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

[26] N. Killoran, T. R. Bromley, J. M. Arrazola, M. Schuld, N. Quesada, and S. Lloyd, Continuous-variable quantum neural networks, Phys. Rev. Research **1**, 033063 (2019).

[27] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, Ann. Math. Statist. **41**, 164 (1970).

[28] P. Brown, J. Cocke, S. D. Pietra, V. D. Pietra, F. Jelinek, R. Mercer, and P. Roossin, A statistical approach to language translation, in *Proceedings of the 12th Conference on Computational Linguistics - Volume 1*, COLING '88, edited by D. Vargha and E. Hajičová (Association for Computational Linguistics, Cedarville, 1988) pp. 71–76.

[29] I. Marshall, Choice of grammatical word-class without global syntactic analysis: Tagging words in the LOB corpus, Comput. Hum. **17**, 139 (1983).

[30] C. Bielza and P. Larrañaga, Bayesian networks in neuroscience: a survey, Front. Comput. Neurosci. **8**, 131 (2014).

[31] F. Cugnata, R. S. Kenett, and S. Salini, Bayesian networks in survey data: Robustness and sensitivity issues, J. Qual. Technol. **48**, 253 (2016).

[32] E. Kyrimi, S. McLachlan, K. Dube, M. R. Neves, A. Fahmi, and N. Fenton, A comprehensive scoping review of Bayesian networks in healthcare: Past, present and future, Artif. Intell. Med. **117**, 102108 (2021).

[33] C. Krapu, R. Stewart, and A. Rose, A review of Bayesian networks for spatial data, ACM Trans. Spatial Algorithms Syst. **9**, 1 (2023).

[34] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc., Ser. B, Methodol. **39**, 1 (1977).

[35] S. L. Lauritzen, The EM algorithm for graphical association models with missing data, Comput. Stat. Data Anal. **19**, 191 (1995).

[36] F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain, Psychol. Rev. **65**, 386 (1958).

[37] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning representations by back-propagating errors, Nature **323**, 533 (1986).

[38] T. Brown *et al.*, Language models are few-shot learners, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Inc., Red Hook, 2020) pp. 1877–1901.

[39] OpenAI, GPT-4 technical report (2023), arXiv:2303.08774 [cs.CL] .

[40] OpenAI *et al.*, Dota 2 with large scale deep reinforcement learning (2019), arXiv:1912.06680 [cs.LG] .

[41] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, Deep Voice 3: Scaling text-to-speech with convolutional sequence learning, in *International Conference on Learning Representations*, edited by Y. Bengio, Y. LeCun, T. Sainath, I. Murray, M. Ranzato, and O. Vinyals (OpenReview, Vancouver, 2018).

[42] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature **521**, 436 (2015).

[43] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions, J. Big Data **8**, 53 (2021).

[44] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, The loss surfaces of multilayer networks, in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, Vol. 38, edited by G. Lebanon and S. V. N. Vishwanathan (PMLR, San Diego, 2015) pp. 192–204.

[45] P. A. Chaudhari, *A Picture of the Energy Landscape of Deep Neural Networks*, Ph.D. thesis, University of California, Los Angeles (2018).

[46] J. J. Meyer, M. Mularski, E. Gil-Fuster, A. A. Mele, F. Arzani, A. Wilms, and J. Eisert, Exploiting symmetry in variational quantum machine learning (2022), arXiv:2205.06217 [quant-ph] .

[47] L. Schatzki, M. Larocca, Q. T. Nguyen, F. Sauvage, and M. Cerezo, Theoretical guarantees for permutation-equivariant quantum neural networks (2022), arXiv:2210.09974 [quant-ph] .

[48] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).

[49] E. R. Anschuetz, Critical points in quantum generative models, in *International Conference on Learning Representations*, edited by K. Hofmann, A. Rush, Y. Liu, C. Finn, Y. Choi, and M. Deisenroth (OpenReview, 2022).

[50] E. R. Anschuetz and B. T. Kiani, Quantum variational algorithms are swamped with traps, Nat. Commun. **13**, 7760 (2022).

[51] E. R. Anschuetz, A. Bauer, B. T. Kiani, and S. Lloyd, Efficient classical algorithms for simulating symmetric quantum systems (2022), arXiv:2211.16998 [quant-ph] .

[52] E. R. Anschuetz, H.-Y. Hu, J.-L. Huang, and X. Gao, Interpretable quantum advantage in neural sequence learning (2022), arXiv:2209.14353 [quant-ph] .

[53] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, Springer Series in Statistics (Springer New York, New York, 2009).

[54] R. M. Solovay, (1995), unpublished manuscript.

[55] A. Y. Kitaev, Quantum computations: algorithms and error correction, Russ. Math. Surv. **52**, 1191 (1997).

[56] X. Gao, Z.-Y. Zhang, and L.-M. Duan, A quantum machine learning algorithm based on generative models, Sci. Adv. **4**, eaat9004 (2018).

[57] N. Wiebe, A. Bocharov, P. Smolensky, M. Troyer, and K. M. Svore, Quantum language processing (2019), arXiv:1902.05162 [quant-ph] .

[58] Y. Du, M.-H. Hsieh, T. Liu, and D. Tao, Expressive power of parametrized quantum circuits, Phys. Rev. Research **2**, 033125 (2020).

[59] R. Sweke, J.-P. Seifert, D. Hangleiter, and J. Eisert, On the quantum versus classical learnability of discrete distributions, Quantum **5**, 417 (2021).

[60] Y. Liu, S. Arunachalam, and K. Temme, A rigorous and robust quantum speed-up in supervised machine learning, Nat. Phys. **17**, 1013 (2021).

[61] J. Romero, R. Babbush, J. R. McClean, C. Hempel, P. J. Love, and A. Aspuru-Guzik, Strategies for quantum computing molecular energies using the unitary coupled cluster ansatz, Quantum Sci. Technol. **4**, 014008 (2018).

[62] S. Aaronson, Read the fine print, Nat. Phys. **11**, 291 (2015).

[63] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. **9**, 4812 (2018).

[64] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Cost function dependent barren plateaus in shallow parametrized quantum circuits, Nat. Commun. **12**, 1791 (2021).

[65] J. Napp, Quantifying the barren plateau phenomenon for a model of unstructured variational ansätze (2022), arXiv:2203.06174 [quant-ph] .

[66] X. Glorot and Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, Vol. 9, edited by Y. W. Teh and M. Titterington (PMLR, Chia, 2010) pp. 249–256.

[67] M. Cerezo and P. J. Coles, Higher order derivatives of quantum neural networks with barren plateaus, Quantum Sci. Technol. **6**, 035006 (2021).

[68] A. Arrasmith, M. Cerezo, P. Czarnik, L. Cincio, and P. J. Coles, Effect of barren plateaus on gradient-free optimization, Quantum **5**, 558 (2021).

[69] B. T. Kiani, S. Lloyd, and R. Maity, Learning unitaries by gradient descent (2020), arXiv:2001.11897 [quant-ph] .

[70] R. Wiersema, C. Zhou, Y. de Sereville, J. F. Carrasquilla, Y. B. Kim, and H. Yuen, Exploring entanglement and optimization within the Hamiltonian variational ansatz, PRX Quantum **1**, 020319 (2020).

[71] J. Kim, J. Kim, and D. Rosa, Universal effectiveness of high-depth circuits in variational eigenproblems, Phys. Rev. Res. **3**, 023203 (2021).

[72] J. Kim and Y. Oz, Quantum energy landscape and circuit optimization, Phys. Rev. A **106**, 052424 (2022).

[73] A. Auffinger, G. B. Arous, and J. Černý, Random matrices and complexity of spin glasses, Commun. Pure Appl. Math. **66**, 165 (2013).

[74] E. Subag and O. Zeitouni, The extremal process of critical points of the pure p-spin spherical spin glass model, Probab. Theory Relat. Fields **168**, 773 (2016).

[75] E. Subag, The complexity of spherical p-spin models—A second moment approach, Ann. Probab. **45**, 3385 (2017).

[76] M. B. Gordy, A generalization of generalized beta distributions, Finance and Economics Discussion Series (1998).

[77] X. You and X. Wu, Exponentially many local minima in quantum neural networks, in *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 139, edited by M. Meila and T. Zhang (PMLR, 2021) pp. 12144–12155.

[78] J. Liu, K. Najafi, K. Sharma, F. Tacchino, L. Jiang, and A. Mezzacapo, An analytic theory for the dynamics of wide quantum neural networks (2022), arXiv:2203.16711 [quant-ph] .

[79] X. You, S. Chakrabarti, and X. Wu, A convergence theory for over-parameterized variational quantum eigensolvers (2022), arXiv:2205.12481 [quant-ph] .

[80] A. Deshpande, P. Niroula, O. Shtanko, A. V. Gorshkov, B. Fefferman, and M. J. Gullans, Tight bounds on the convergence of noisy random circuits to the uniform distribution, PRX Quantum **3**, 040329 (2022).

[81] I. Schur, Neue begründung der theorie der gruppencharaktere, in *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften zu Berlin*, Vol. 24, edited by A. J. G. F. Auwers, K. T. Vahlen, H. A. Diels, and H. W. G. Waldeyer (Deutsche Akademie der Wissenschaften zu Berlin, Berlin, 1905) pp. 406–432.

[82] H.-Y. Huang, R. Kueng, and J. Preskill, Predicting many properties of a quantum system from very few measurements, Nat. Phys. **16**, 1050 (2020).

[83] S. Gu, R. D. Somma, and B. Şahinoğlu, Fast-forwarding quantum evolution, Quantum **5**, 577 (2021).

[84] I. Sutskever, O. Vinyals, and Q. V. Le, Sequence to sequence learning with neural networks, in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, edited by Z. Ghahramani, M. Welling, C. Cortes, and N. Lawrence (Curran Associates, Inc., Red Hook, 2014) pp. 3104–3112.

[85] A. Mari and J. Eisert, Positive Wigner functions render classical simulation of quantum computation efficient, Phys. Rev. Lett. **109**, 230503 (2012).

[86] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman, AutoClass: A Bayesian classification system, in *Machine Learning Proceedings 1988*, edited by J. Laird (Morgan Kaufmann, San Francisco, 1988) pp. 54–64.

[87] S. Hochreiter and J. Schmidhuber, Long short-term memory, Neural Comput. **9**, 1735 (1997).

[88] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, edited by A. Moschitti, B. Pang, and W. Daelemans (Association for Computational Linguistics, Doha, 2014) pp. 1724–1734.

[89] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., Red Hook, 2017) pp. 6000–6010.

[90] A. M. Gleason, Measures on the closed subspaces of a Hilbert space, J. Math. Mech. **6**, 885 (1957).

[91] J. S. Bell, On the problem of hidden variables in quantum mechanics, Rev. Mod. Phys. **38**, 447 (1966).

[92] S. Kochen and E. P. Specker, The problem of hidden variables in quantum mechanics, J. Math. Mech. **17**, 59 (1967).

[93] N. D. Mermin, Simple unified form for the major no-hidden-variables theorems, Phys. Rev. Lett. **65**, 3373 (1990).

[94] A. R. Plastino and A. Cabello, State-independent quantum contextuality for continuous variables, Phys. Rev. A **82**, 022114 (2010).

[95] D. Gottesman, A. Kitaev, and J. Preskill, Encoding a qubit in an oscillator, Phys. Rev. A **64**, 012310 (2001).

[96] X. Gao, E. R. Anschuetz, S.-T. Wang, J. I. Cirac, and M. D. Lukin, Enhancing generative models via quantum correlations, Phys. Rev. X **12**, 021037 (2022).

[97] C. Ortiz Marrero, M. Kieferová, and N. Wiebe, Entanglement-induced barren plateaus, PRX Quantum **2**, 040316 (2021).

[98] E. Campos, A. Nasrallah, and J. Biamonte, Abrupt transitions in variational quantum circuit training, Phys. Rev. A **103**, 032607 (2021).

[99] E. Farhi, J. Goldstone, and S. Gutmann, A quantum approximate optimization algorithm (2014), arXiv:1411.4028 [quant-ph] .

[100] R. J. Adler and J. E. Taylor, Random fields on manifolds, in *Random Fields and Geometry* (Springer New York, New York, 2009) pp. 301–330.

[101] J. W. Negele and H. Orland, Chapter 8: Stochastic methods, in *Quantum Many-Particle Systems* (CRC Press, Boca Raton, 1998) pp. 400–446.

[102] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets, Nature **549**, 242 (2017).

[103] A. Blum, M. Furst, J. Jackson, M. Kearns, Y. Mansour, and S. Rudich, Weakly learning DNF and characterizing statistical query learning using Fourier analysis, in *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '94, edited by F. T. Leighton and M. Goodrich (Association for Computing Machinery, New York, 1994) pp. 253–262.

[104] B. Szörényi, Characterizing statistical query learning: Simplified notions and proofs, in *Algorithmic Learning Theory*, edited by R. Gavaldà, G. Lugosi, T. Zeugmann, and S. Zilles (Springer Berlin Heidelberg, Berlin, 2009) pp. 186–200.

[105] S. Goel, A. Gollakota, and A. Klivans, Statistical-query lower bounds via functional gradients, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates, Inc., Red Hook, 2020) pp. 2147–2158.

[106] S. Shalev-Shwartz, O. Shamir, and S. Shammah, Failures of gradient-based deep learning, in *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 70, edited by D. Precup and Y. W. Teh (PMLR, 2017) pp. 3067–3075.

[107] E. Farhi, J. Goldstone, S. Gutmann, and L. Zhou, The Quantum Approximate Optimization Algorithm and the Sherrington-Kirkpatrick model at infinite size, Quantum **6**, 759 (2022).

[108] I. Cong, S. Choi, and M. D. Lukin, Quantum convolutional neural networks, Nat. Phys. **15**, 1273 (2019).

[109] L. Reyzin, Statistical queries and statistical algorithms: Foundations and applications (2020), arXiv:2004.00557 [cs.LG] .

[110] S. Goel, A. Gollakota, Z. Jin, S. Karmalkar, and A. Klivans, Superpolynomial lower bounds for learning one-layer neural networks using gradient descent, in *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 119, edited by H. Daumé, III and A. Singh (PMLR, 2020) pp. 3587–3596.

[111] M. Kearns, Efficient noise-tolerant learning from statistical queries, J. ACM **45**, 983 (1998).

[112] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, Evaluating analytic gradients on quantum hardware, Phys. Rev. A **99**, 032331 (2019).

[113] A. Mari, T. R. Bromley, and N. Killoran, Estimating the gradient and higher-order derivatives on quantum hardware, Phys. Rev. A **103**, 012405 (2021).

[114] A. Anshu, S. Arunachalam, T. Kuwahara, and M. Soleimanifar, Sample-efficient learning of interacting quantum systems, Nat. Phys. **17**, 931 (2021).

[115] E. Bairey, I. Arad, and N. H. Lindner, Learning a local Hamiltonian from local measurements, Phys. Rev. Lett. **122**, 020504 (2019).

[116] S. Chen, S. Zhou, A. Seif, and L. Jiang, Quantum advantages for Pauli channel estimation, Phys. Rev. A **105**, 032435 (2022).

[117] H.-Y. Huang, R. Kueng, and J. Preskill, Information-theoretic bounds on quantum advantage in machine learning, Phys. Rev. Lett. **126**, 190505 (2021).

[118] A. Gollakota and D. Liang, On the hardness of PAC-learning stabilizer states with noise, Quantum **6**, 640 (2022).

[119] M. Hinsche, M. Ioannou, A. Nietner, J. Haferkamp, Y. Quek, D. Hangleiter, J.-P. Seifert, J. Eisert, and R. Sweke, Learnability of the output distributions of local quantum circuits (2021), arXiv:2110.05517 [quant-ph] .

[120] D. Wolpert and W. Macready, No free lunch theorems for optimization, IEEE Trans. Evol. Comput. **1**, 67 (1997).

[121] B. T. Kiani, G. D. Palma, M. Marvian, Z.-W. Liu, and S. Lloyd, Learning quantum data with the quantum earth mover's distance, Quantum Sci. Technol. **7**, 045002 (2022).

[122] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, Power of data in quantum machine learning, Nat. Commun. **12**, 2631 (2021).

[123] S. Khatri, R. LaRose, A. Poremba, L. Cincio, A. T. Sornborger, and P. J. Coles, Quantum-assisted quantum compiling, Quantum **3**, 140 (2019).

[124] A. Harrow and S. Mehraban, Approximate unitary t-designs by short random quantum circuits using nearest-neighbor and long-range gates (2018), arXiv:1809.06957 [quant-ph] .

[125] J. Haferkamp, Random quantum circuits are approximate unitary t-designs in depth $O\left(nt^{5+o(1)}\right)$, Quantum **6**, 795 (2022).

[126] A. Pesah, M. Cerezo, S. Wang, T. Volkoff, A. T. Sornborger, and P. J. Coles, Absence of barren plateaus in quantum convolutional neural networks, Phys. Rev. X **11**, 041011 (2021).

[127] W. Heisenberg, Zur theorie des ferromagnetismus, Z. Phys. **49**, 619 (1928).

[128] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, in *International Conference on Learning Representations*, edited by Y. Bengio, Y. LeCun, B. Kingsbury, S. Bengio, N. de Freitas, and H. Larochelle (San Diego, 2015).

[129] F. G. S. L. Brandao, M. Broughton, E. Farhi, S. Gutmann, and H. Neven, For fixed control parameters the Quantum Approximate Optimization Algorithm's objective function value concentrates for typical instances (2018), arXiv:1812.04170 [quant-ph] .

[130] M. Larocca, N. Ju, D. García-Martín, P. J. Coles, and M. Cerezo, Theory of overparametrization in quantum neural networks (2021), arXiv:2109.11676 [quant-ph] .

[131] M. Kieferová and N. Wiebe, Tomography and generative training with quantum Boltzmann machines, Phys. Rev. A **96**, 062327 (2017).

[132] E. R. Anschuetz and Y. Cao, Realizing quantum Boltzmann machines through eigenstate thermalization (2019), arXiv:1903.01359 [quant-ph] .

[133] C. Zoufal, A. Lucchi, and S. Woerner, Variational quantum Boltzmann machines, Quantum Mach. Intell. **3**, 7 (2021).

[134] H. Bethe, Zur theorie der metalle, Z. Phys. **71**, 205 (1931).

[135] M. A. Levin and X.-G. Wen, String-net condensation: A physical mechanism for topological phases, Phys. Rev. B **71**, 045110 (2005).

[136] A. Belavin, A. Polyakov, and A. Zamolodchikov, Infinite conformal symmetry in two-dimensional quantum field theory, Nucl. Phys. B **241**, 333 (1984).

[137] R. Somma, H. Barnum, G. Ortiz, and E. Knill, Efficient solvability of Hamiltonians and limits on the power of some quantum computational models, Phys. Rev. Lett. **97**, 190501 (2006).

[138] R. Zeier and T. Schulte-Herbrüggen, Symmetry principles in quantum systems theory, J. Math. Phys. **52**, 113510 (2011).

[139] G. Castelazo, Q. T. Nguyen, G. De Palma, D. Englund, S. Lloyd, and B. T. Kiani, Quantum algorithms for group convolution, cross-correlation, and equivariant transformations, Phys. Rev. A **106**, 032402 (2022).

[140] M. Larocca, F. Sauvage, F. M. Sbahi, G. Verdon, P. J. Coles, and M. Cerezo, Group-invariant quantum machine learning, PRX Quantum **3**, 030341 (2022).

[141] M. Ragone, P. Braccia, Q. T. Nguyen, L. Schatzki, P. J. Coles, F. Sauvage, M. Larocca, and M. Cerezo, Representation theory for geometric quantum machine learning (2022), arXiv:2210.07980 [quant-ph] .

[142] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, Geometric deep learning: Going beyond Euclidean data, IEEE Signal Process. Mag. **34**, 18 (2017).

[143] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, A comprehensive survey on graph neural networks, IEEE Trans. Neural Netw. Learn. Syst. **32**, 4 (2021).

[144] T. Cohen and M. Welling, Group equivariant convolutional networks, in *Proceedings of The 33rd International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 48, edited by M. F. Balcan and K. Q. Weinberger (PMLR, New York, 2016) pp. 2990–2999.

[145] P. J. Olver, *Classical Invariant Theory*, London Mathematical Society Student Texts (Cambridge University Press, Cambridge, 1999).

[146] B. Sturmfels, *Algorithms in Invariant Theory*, Texts & Monographs in Symbolic Computation (Springer Vienna, Vienna, 2008).

[147] R. Duan, H. Wu, and R. Zhou, Faster matrix multiplication via asymmetric hashing (2022), arXiv:2210.10173 [cs.DS] .

[148] J. Demmel, I. Dumitriu, and O. Holtz, Fast linear algebra is stable, Numer. Math. **108**, 59 (2007).

[149] D. Kazhdan and G. Lusztig, Affine Lie algebras and quantum groups, Int. Math. Res. Not. **1991**, 21 (1991).

[150] D. Bacon, I. L. Chuang, and A. W. Harrow, The quantum Schur transform: I. efficient qudit circuits (2006), arXiv:quant-ph/0601001 [quant-ph] .

[151] D. Bacon, I. L. Chuang, and A. W. Harrow, Efficient quantum circuits for Schur and Clebsch-Gordan transforms, Phys. Rev. Lett. **97**, 170502 (2006).

[152] G. H. Low, Classical shadows of fermions with particle number symmetry (2022), arXiv:2208.08964 [quant-ph] .

[153] B. Coyle, D. Mills, V. Danos, and E. Kashefi, The Born supremacy: quantum advantage and training of an Ising Born machine, npj Quantum Inf. **6**, 1 (2020).

[154] G. K. Dziugaite, *Revisiting Generalization for Deep Learning: PAC-Bayes, Flat Minima, and Generative Models*, Ph.D. thesis, University of Cambridge (2020).

[155] L. Jing, Y. Shen, T. Dubcek, J. Peurifoy, S. Skirlo, Y. LeCun, M. Tegmark, and M. Soljačić, Tunable efficient unitary neural networks (EUNN) and their application to RNNs, in *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 70, edited by D. Precup and Y. W. Teh (PMLR, 2017) pp. 1733–1741.

[156] C. Kelly, Tab-delimited bilingual sentence pairs (2021).

[157] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, PRX Quantum **3**, 010313 (2022).

[158] R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, and N. Jaitly, A comparison of sequence-to-sequence models for speech recognition, in *Proc. Interspeech 2017*, edited by F. Lacerda (Curran Associates, Inc., Red Hook, 2017) pp. 939–943.

[159] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, Show and tell: A neural image caption generator, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, edited by K. Barnard, H. Bischof, P. Felzenszwalb, D. Forsyth, S. Lazebnik, and J. Matas (Curran Associates, Inc., Red Hook, 2015) pp. 3156–3164.

[160] H. Federer, Rectifiability, in *Geometric Measure Theory*, edited by B. Eckmann and B. L. van der Waerden (Springer Berlin Heidelberg, Berlin, 1996) pp. 207–340.

[161] S. L. Braunstein and P. van Loock, Quantum information with continuous variables, Rev. Mod. Phys. **77**, 513 (2005).

[162] N. Liu, J. Thompson, C. Weedbrook, S. Lloyd, V. Vedral, M. Gu, and K. Modi, Power of one qumode for quantum computation, Phys. Rev. A **93**, 052304 (2016).

[163] C. Calcluth, A. Ferraro, and G. Ferrini, The vacuum provides quantum advantage to otherwise simulatable architectures (2022), arXiv:2205.09781 [quant-ph] .

[164] B. Q. Baragiola, G. Pantaleoni, R. N. Alexander, A. Karanjai, and N. C. Menicucci, All-Gaussian universality and fault tolerance with the Gottesman-Kitaev-Preskill code, Phys. Rev. Lett. **123**, 200502 (2019).

[165] M. A. Nielsen and I. L. Chuang, The quantum Fourier transform and its applications, in *Quantum Computation and Quantum Information: 10th Anniversary Edition* (Cambridge University Press, Cambridge, 2010) pp. 216–247.

[166] D. E. Gottesman, *Stabilizer Codes and Quantum Error Correction*, Ph.D. thesis, California Institute of Technology (1997).

[167] S. Aaronson and D. Gottesman, Improved simulation of stabilizer circuits, Phys. Rev. A **70**, 052328 (2004).

[168] A. Karanjai, J. J. Wallman, and S. D. Bartlett, Contextuality bounds the efficiency of classical simulation of quantum processes (2018), arXiv:1802.07744 [quant-ph] .

[169] G. J. Butler, J. G. Timourian, and C. Viger, The rank theorem for locally Lipschitz continuous functions, Can. Math. Bull. **31**, 217 (1988).

[170] R. I. Booth, U. Chabaud, and P.-E. Emeriau, Contextuality and Wigner negativity are equivalent for continuous-variable quantum measurements, Phys. Rev. Lett. **129**, 230401 (2022).

[171] J. Haferkamp and J. Bermejo-Vega, Equivalence of contextuality and Wigner function negativity in continuous-variable quantum optics (2021), arXiv:2112.14788 [quant-ph] .

[172] R. Babbush, J. R. McClean, M. Newman, C. Gidney, S. Boixo, and H. Neven, Focus beyond quadratic speedups for error-corrected quantum advantage, PRX Quantum **2**, 010103 (2021).

[173] Y. Takeuchi, A. Mantri, T. Morimae, A. Mizutani, and J. F. Fitzsimons, Resource-efficient verification of quantum computing using Serfling's bound, npj Quantum Inf. **5**, 27 (2019).

[174] D. Gamarnik, The overlap gap property: A topological barrier to optimizing over random structures, Proc. Natl. Acad. Sci. U.S.A. **118**, e2108492118 (2021).

[175] E. Farhi, D. Gamarnik, and S. Gutmann, The Quantum Approximate Optimization Algorithm needs to see the whole graph: A typical case (2020), arXiv:2004.09002 [quant-ph] .

[176] A. Anshu and T. Metger, Concentration bounds for quantum states and limitations on the QAOA from polynomial approximations (2022), arXiv:2209.02715 [quant-ph] .

[177] A. Anshu, N. P. Breuckmann, and C. Nirkhe, NLTS Hamiltonians from good quantum codes (2022), arXiv:2206.13228 [quant-ph] .

[178] D. Wecker, M. B. Hastings, and M. Troyer, Progress towards practical quantum variational algorithms, Phys. Rev. A **92**, 042303 (2015).

[179] S. Sim, P. D. Johnson, and A. Aspuru-Guzik, Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms, Adv. Quantum Technol. **2**, 1900070 (2019).

[180] N. G. Ushakov, 1. Basic properties of the characteristic functions, in *Selected Topics in Characteristic Functions* (De Gruyter, Berlin, 2011) pp. 1–66.

[181] T. Jiang, Maxima of entries of Haar distributed matrices, Probab. Theory Relat. Fields **131**, 121 (2005).

[182] D. DiVincenzo, D. Leung, and B. Terhal, Quantum data hiding, IEEE Trans. Inf. Theory **48**, 580 (2002).

[183] F. Satterthwaite, An approximate distribution of estimates of variance components, Biometrics Bull. **2**, 110 (1946).

[184] B. Welch, The generalization of 'Student's' problem when several different population variances are involved, Biometrika **34**, 28 (1947).

[185] A. Khuri, T. Mathew, and D. Nel, A test to determine closeness of multivariate Satterthwaite's approximation, J. Multivar. Anal. **51**, 201 (1994).

[186] A. I. Khuri, T. Mathew, and B. K. Sinha, Chapter 10: Multivariate mixed and random models, in *Statistical Tests for Mixed Linear Models*, Vol. 906 (John Wiley & Sons, New York, 2011) pp. 256–296.

[187] G. Pivaro, S. Kumar, G. Fraidenraich, and C. Dias, On the exact and approximate eigenvalue distribution for sum of Wishart matrices, IEEE Trans. Veh. Technol. **66**, 10537 (2017).

[188] M. L. Eaton, Chapter 8: The Wishart distribution, in *Multivariate Statistics*, Lecture Notes–Monograph Series, Vol. 53 (Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2007) pp. 302–333.

[189] M. Srivastava, Singular Wishart and multivariate beta distributions, Ann. Statist. **31**, 1537 (2003).

[190] S. Yu, J. Ryu, and K. Park, A derivation of anti-Wishart distribution, J. Multivar. Anal. **131**, 121 (2014).

[191] A. I. Khuri, The probability of a negative linear combination of independent mean squares, Biom. J. **36**, 899 (1994).

[192] R. Babbush, N. Wiebe, J. McClean, J. McClain, H. Neven, and G. K.-L. Chan, Low-depth quantum simulation of materials, Phys. Rev. X **8**, 011044 (2018).

[193] V. Marčenko and L. Pastur, Distribution of eigenvalues for some sets of random matrices, Math. USSR Sb. **1**, 457 (1967).

[194] E. P. Wigner, On the distribution of the roots of certain symmetric matrices, Ann. Math. **67**, 325 (1958).

[195] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications* (Springer Berlin Heidelberg, Berlin, 2010).

[196] G. B. Arous and A. Guionnet, Large deviations for Wigner's law and Voiculescu's non-commutative entropy, Probab. Theory Relat. Fields **108**, 517 (1997).

[197] F. Hiai and D. Petz, Eigenvalue density of the Wishart matrix and large deviations, Infin. Dimens. Anal. Quantum Probab. Relat. Top. **01**, 633 (1998).

[198] F. Hiai and D. Petz, Chapter 4: Random matrices and asymptotically free relation, in *The Semicircle Law, Free Random Variables and Entropy*, Mathematical Surveys and Monographs, Vol. 77 (American Mathematical Society, Providence, 2006) pp. 113–174.

[199] A. Guionnet and O. Zeitouni, Large deviations asymptotics for spherical integrals, J. Funct. Anal. **188**, 461 (2002).

[200] K. Johansson, Shape fluctuations and random matrices, Commun. Math. Phys. **209**, 437 (2000).

[201] G. B. Arous, A. Dembo, and A. Guionnet, Aging of spherical spin glasses, Probab. Theory Relat. Fields **120**, 1 (2001).

[202] A. Guionnet and M. Maïda, Large deviations for the largest eigenvalue of the sum of two random matrices, Electron. J. Probab. **25**, 24 pp. (2020).

[203] H. Abraham *et al.*, Qiskit: An open-source framework for quantum computing (2019).

[204] L. Bottou and O. Bousquet, The tradeoffs of large scale learning, in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, edited by J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis (Curran Associates, Inc., Red Hook, 2007) pp. 161–168.

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

[206] H. Shen, P. Zhang, Y.-Z. You, and H. Zhai, Information scrambling in quantum neural networks, Phys. Rev. Lett. **124**, 200504 (2020).

[207] M. C. Caro, H.-Y. Huang, M. Cerezo, K. Sharma, A. Sornborger, L. Cincio, and P. J. Coles, Generalization in quantum machine learning from few training data, Nat. Commun. **13**, 4919 (2022).

[208] Y. Du, Z. Tu, X. Yuan, and D. Tao, Efficient measure for the expressivity of variational quantum algorithms, Phys. Rev. Lett. **128**, 080506 (2022).

[209] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, Understanding deep learning (still) requires rethinking generalization, Commun. ACM **64**, 107 (2021).

[210] B. Neyshabur, S. Bhojanapalli, D. Mcallester, and N. Srebro, Exploring generalization in deep learning, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., Red Hook, 2017) pp. 5949–5958.

[211] H. Maennel, I. Alabdulmohsin, I. Tolstikhin, R. J. N. Baldock, O. Bousquet, S. Gelly, and D. Keysers, What do neural networks learn when trained with random labels?, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates, Inc., Red Hook, 2020) pp. 19693–19704.

[212] B. Applebaum, B. Barak, and D. Xiao, On basing lower-bounds for learning on worst-case assumptions, in *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, edited by R. Ravi (Curran Associates, Inc., Red Hook, 2008) pp. 211–220.

[213] S. Arunachalam, A. B. Grilo, and H. Yuen, Quantum statistical query learning (2020), arXiv:2002.08240 [quant-ph] .

[214] L. G. Valiant, A theory of the learnable, Commun. ACM **27**, 1134–1142 (1984).

[215] D. Aharonov, V. Jones, and Z. Landau, A polynomial quantum algorithm for approximating the Jones polynomial, in *Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '06, edited by J. Kleinberg (Association for Computing Machinery, New York, 2006) pp. 427–436.

[216] X. Wang, Z. Song, and Y. Wang, Variational quantum singular value decomposition, Quantum **5**, 483 (2021).

[217] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, Phys. Rev. A **98**, 032309 (2018).

[218] H. Buhrman, R. Cleve, J. Watrous, and R. de Wolf, Quantum fingerprinting, Phys. Rev. Lett. **87**, 167902 (2001).

[219] K. Beer, D. Bondarenko, T. Farrelly, T. J. Osborne, R. Salzmann, D. Scheiermann, and R. Wolf, Training deep quantum neural networks, Nat. Commun. **11**, 808 (2020).

[220] I. Diakonikolas, D. M. Kane, V. Kontonis, and N. Zarifis, Algorithms and SQ lower bounds for PAC learning one-hidden-layer ReLU networks, in *Proceedings of Thirty Third Conference on Learning Theory*, Proceedings of Machine

Learning Research, Vol. 125, edited by J. Abernethy and S. Agarwal (PMLR, 2020) pp. 1514–1539.

[221] S. Chen, A. Gollakota, A. Klivans, and R. Meka, Hardness of noise-free learning for two-hidden-layer neural networks, in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc., Red Hook, 2022) pp. 10709–10724.

[222] I. Diakonikolas, D. Kane, and N. Zarifis, Near-optimal SQ lower bounds for agnostically learning halfspaces and ReLUs under Gaussian marginals, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Inc., Red Hook, 2020) pp. 13586–13596.

[223] S. Chen, A. R. Klivans, and R. Meka, Learning deep ReLU networks is fixed-parameter tractable, in *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, edited by N. Vishnoi (Curran Associates, Inc., Red Hook, 2022) pp. 696–707.

[224] A. Andoni, R. Panigrahy, G. Valiant, and L. Zhang, Learning sparse polynomial functions, in *Proceedings of the 2014 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, edited by C. Chekuri (Society for Industrial and Applied Mathematics, Philadelphia, 2014) pp. 500–510.

[225] B. Collins, Moments and cumulants of polynomial random variables on unitary-groups, the Itzykson-Zuber integral, and free probability, Int. Math. Res. Not. **2003**, 953 (2003).

[226] B. Collins and P. Śniady, Integration with respect to the Haar measure on unitary, orthogonal and symplectic group, Commun. Math. Phys. **264**, 773 (2006).

[227] A. Y. Zaitsev, Estimates for the Levy-Prokhorov distance in terms of characteristic functions and some of their applications, J. Sov. Math. **27**, 3070 (1984).

[228] D. Voiculescu, Limit laws for random matrices and free products, Invent. Math. **104**, 201 (1991).

[229] Y. Li and Y. Liang, Learning overparameterized neural networks via stochastic gradient descent on structured data, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., Red Hook, 2018) pp. 8168–8177.

[230] S. Arora, S. S. Du, W. Hu, Z. Li, R. Salakhutdinov, and R. Wang, On exact computation with an infinitely wide neural net, in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, edited by

H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., Red Hook, 2019) pp. 8141–8150.

[231] A. Skolik, J. R. McClean, M. Mohseni, P. van der Smagt, and M. Leib, Layerwise learning for quantum neural networks, Quantum Mach. Intell. **3**, 5 (2021).

[232] H. R. Grimsley, S. E. Economou, E. Barnes, and N. J. Mayhall, An adaptive variational algorithm for exact molecular simulations on a quantum computer, Nat. Commun. **10**, 3007 (2019).

[233] C. Cade, L. Mineh, A. Montanaro, and S. Stanisic, Strategies for solving the Fermi-Hubbard model on near-term quantum computers, Phys. Rev. B **102**, 235122 (2020).

[234] A. Paszke *et al.*, PyTorch: An imperative style, high-performance deep learning library, in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., Red Hook, 2019).

[235] B. T. Kiani, bkiani/Beyond-Barren-Plateaus: Code for Beyond Barren Plateaus paper (2022).

[236] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, Visualizing the loss landscape of neural nets, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., Red Hook, 2018) pp. 6391–6401.

[237] G. De Palma, M. Marvian, D. Trevisan, and S. Lloyd, The quantum Wasserstein distance of order 1, IEEE Trans. Inf. Theory **67**, 6627 (2021).

[238] J. Stokes, J. Izaac, N. Killoran, and G. Carleo, Quantum Natural Gradient, Quantum **4**, 269 (2020).

[239] J.-Q. Chen, J. Ping, and F. Wang, *Group Representation Theory for Physicists*, 2nd ed. (World Scientific Publishing, Singapore, 2002).

[240] OEIS Foundation Inc., The On-Line Encyclopedia of Integer Sequences (2022), published electronically at `http://oeis.org`, Sequence A000292.

[241] W. Fulton, *Young Tableaux: With Applications to Representation Theory and Geometry*, London Mathematical Society Student Texts (Cambridge University Press, Cambridge, 1996).

[242] R. H. Dicke, Coherence in spontaneous radiation processes, Phys. Rev. **93**, 99 (1954).

[243] A. Bärtschi and S. Eidenbenz, Deterministic preparation of Dicke states, in *Fundamentals of Computation Theory*, edited by L. A. Gąsieniec, J. Jansson, and C. Levcopoulos (Springer International Publishing, Cham, 2019) pp. 126–139.

[244] V. Havlíček and S. Strelchuk, Quantum Schur sampling circuits can be strongly simulated, Phys. Rev. Lett. **121**, 060505 (2018).

[245] G. Racah, Theory of complex spectra. II, Phys. Rev. **62**, 438 (1942).

[246] J. Pearl, Bayesian networks: A model of self-activated memory for evidential reasoning, in *Proceedings of the 7th Conference of the Cognitive Science Society*, edited by R. H. Granger, Jr. and K. Eiselt (Lawrence Erlbaum Associates, Inc., Hillsdale, 1985) pp. 329–334.

[247] L. E. Baum and T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, Ann. Math. Stat. **37**, 1554 (1966).

[248] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, edited by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (Curran Associates, Inc., Red Hook, 2014) pp. 2672–2680.

[249] D. Rezende and S. Mohamed, Variational inference with normalizing flows, in *Proceedings of the 32nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 37, edited by F. Bach and D. Blei (PMLR, Lille, 2015) pp. 1530–1538.

[250] A. J. Kerman, Quantum information processing using quasiclassical electromagnetic interactions between qubits and electrical resonators, New J. Phys. **15**, 123011 (2013).

[251] A. Blais, A. L. Grimsmo, S. M. Girvin, and A. Wallraff, Circuit quantum electrodynamics, Rev. Mod. Phys. **93**, 025005 (2021).

[252] Y. Shi, C. Chamberland, and A. Cross, Fault-tolerant preparation of approximate GKP states, New J. Phys. **21**, 093007 (2019).

[253] Y.-S. Ra, A. Dufour, M. Walschaers, C. Jacquard, T. Michel, C. Fabre, and N. Treps, Non-Gaussian quantum states of a multimode light field, Nat. Phys. **16**, 144 (2020).

[254] N. C. Menicucci, S. T. Flammia, and P. van Loock, Graphical calculus for Gaussian pure states, Phys. Rev. A **83**, 042335 (2011).

[255] C. Calcluth, A. Ferraro, and G. Ferrini, Efficient simulation of Gottesman-Kitaev-Preskill states with Gaussian circuits, Quantum **6**, 867 (2022).

[256] S. D. Bartlett, B. C. Sanders, S. L. Braunstein, and K. Nemoto, Efficient classical simulation of continuous variable quantum information processes, Phys. Rev. Lett. **88**, 097904 (2002).

[257] F. Chollet *et al.*, Keras, `https://keras.io` (2015).

[258] P. Høyer and R. Špalek, Quantum fan-out is powerful, Theory Comput. **1**, 81 (2005).

[259] I. D. Kivlichan, J. McClean, N. Wiebe, C. Gidney, A. Aspuru-Guzik, G. K.-L. Chan, and R. Babbush, Quantum simulation of electronic structure with linear depth and connectivity, Phys. Rev. Lett. **120**, 110501 (2018).

[260] J. Alman and V. V. Williams, A refined laser method and faster matrix multiplication, in *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, edited by D. Marx (Society for Industrial and Applied Mathematics, Philadelphia, 2021) pp. 522–539.