FRAGMENTATION IS EFFICIENTLY LEARNABLE BY QUANTUM NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

In certain classes of physical quantum systems, the exponentially large state space "fragments" into many low-dimensional, dynamically disconnected subspaces. We introduce a learning problem known as *fragment classification*, where given a quantum state input, one is interested in classifying to which subspace the state belongs. We prove that solving this learning problem is efficient on a quantum computer when the fragmentation phenomenon satisfies certain conditions. Furthermore, we give evidence supporting the classical hardness of this task by demonstrating that known dequantization techniques fail for the fragment classification problem. Consequently, this work provides a rare example of a physically motivated quantum machine learning task that is both efficient for quantum computers to perform and admits no known classical dequantization.

1 INTRODUCTION

Quantum algorithms are known to outperform the best known classical algorithms on many computational problems exhibiting some sort of algebraic structure (Shor, 1999; Gyurik et al., 2024; Berry et al., 2024; Harrow et al., 2009). A significant research direction today is focused on trying to understand whether the unique advantages offered by quantum computers can help in machine learning tasks and whether *quantum neural networks* (QNNs) can be superior to classical methods. Classical neural networks are known to behave like Gaussian processes in the asymptotic limit (Neal, 1996), which provides theoretical guarantees on their efficient trainability in a wide range of applications. On the other hand, QNNs are known to have poorly-behaved loss landscapes. They are often dominated by *barren plateaus* (McClean et al., 2018; Cerezo et al., 2021; Ragone et al., 2024), which generally prevent gradients from being estimated efficiently, and *poor local minima*, which generally prevent gradient descent from reaching the global optimum (Anschuetz, 2022; Anschuetz & Kiani, 2022; You et al., 2022). While there do exist settings where QNNs are efficient to train via gradient descent—such as when they have many symmetries (Meyer et al., 2023; Schatzki et al., 2024; Nguyen et al., 2024; Gibbs et al., 2024; Anschuetz et al., 2023b; West et al., 2024; Anschuetz & Gao, 2024)—in these settings there often also exist efficient classical simulation algorithms that render the use of a quantum computer for the task unnecessary outside of potentially an initial data acquisition phase (Anschuetz et al., 2023a; Goh et al., 2025; Cerezo et al., 2025).

Recent work (Anschuetz, 2025) formulated the Jordan Algebraic Wishart System (JAWS) framework, which is the first full theoretical characterization of QNN loss landscapes that provides the conditions under which they can become efficiently trainable using gradient descent. In this work, we use these tools to formally demonstrate a specific setting where QNNs *can* in fact be efficiently trained to solve a physically motivated problem for which no efficient classical simulation algorithm is known. Our approach is to showcase a setting where the QNN has a high-dimensional symmetry group but, crucially, the end-user does not know what that group is and—to the best of our knowledge—has no efficient strategy for learning it. At a high level, this mirrors the mechanism underlying other well-known quantum-classical separations, such as the abelian hidden subgroup problem (Kitaev, 1995).

The rest of our paper is organized as follows. First, in Section 2 we give background on the physical mechanism underlying our learning task known as Hilbert space fragmentation, as well as the general quantum learning technique we will be utilizing here. Second, in Section 3 we demonstrate our main results: we construct a QNN architecture, show that it satisfies the conditions required for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

efficient trainability, and then discuss the conjectured classical hardness of the learning task. In Section 4, we outline a scenario in which our construction could be potentially used in an experimental setting. Finally, we conclude in Section 5.

2 PRELIMINARIES

2.1 FRAGMENTATION AND THE SCHUR BASIS

The problem that we study in this work is that of classifying states in systems which exhibit *Hilbert space fragmentation* (Pai et al., 2019; Sala et al., 2020; Khemani et al., 2020), a well-studied barrier to ergodicity in quantum many-body systems conceptually similar to the better-known quantum many-body scarring phenomenon (Bernien et al., 2017; Turner et al., 2018; Moudgalya & Motrunich, 2024). The mechanism underlying Hilbert space fragmentation has been recently mathematically characterized (Moudgalya & Motrunich, 2022), and we review this characterization here.

We let L be the problem size—physically known as the *system size*—which we will later take to be large. For each L, one is given a set of Hermitian operators $\mathcal{S}$ known as *Hamiltonians* acting on a Hilbert space $\mathcal{H}$ with a tensor product structure $\mathcal{H} = \bigotimes_{i=1}^{L} \mathcal{H}_i$. $\mathcal{S}$ is constructed such that all Hamiltonians $\boldsymbol{H} \in \mathcal{S}$ take the form:

$$\boldsymbol{H} = \sum_{i=1}^{m} c_i \boldsymbol{h}_i, \tag{1}$$

where all $c_i \in \mathbb{R}$ and the $\boldsymbol{h}_i$ are *local operators*, i.e., $\boldsymbol{h}_i$ acts nontrivially only on an L-independent number of tensor product factors $\mathcal{H}_i$. Furthermore, m only grows at most polynomially quickly with L.

Let $\mathcal{A} = \langle \boldsymbol{h}_1, \ldots, \boldsymbol{h}_m \rangle$ be the associative algebra generated by these operators under addition and matrix multiplication. Let $\mathcal{C}$ be the *commutant algebra*, consisting of all operators in $\mathcal{L}(\mathcal{H})$ that commute with all elements of $\mathcal{A}$. By the Von Neumann bicommutant theorem (Landsman, 1998), we know that $\mathcal{A}$ and $\mathcal{C}$ are each other's centralizers, and we can decompose the Hilbert space as:

$$\mathcal{H} = \bigoplus_{\lambda=1}^{\Lambda} \mathcal{H}_\lambda^{(\mathcal{A})} \otimes \mathcal{H}_\lambda^{(\mathcal{C})}, \tag{2}$$

where each $\mathcal{H}_\lambda^{(\mathcal{A})}$ is an irreducible representation of $\mathcal{A}$ and each $\mathcal{H}_\lambda^{(\mathcal{C})}$ is an irreducible representation of $\mathcal{C}$. Let $N_\lambda = \dim\left(\mathcal{H}_\lambda^{(\mathcal{A})}\right), N'_\lambda = \dim\left(\mathcal{H}_\lambda^{(\mathcal{C})}\right)$. In this decomposition, each $\mathcal{H}_\lambda^{(\mathcal{A})}$ is called a *Krylov subspace* of the system, preserved by the action of Hamiltonians constructed from the generators $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_m$. Each of these subspaces has N'_λ degenerate copies. The representation in $\mathcal{H}$ of any operator $A \in \mathcal{A}$ can be written (in a slight abuse of notation identifying the operator with its representation) in the form:

$$\boldsymbol{A} = \bigoplus_{\lambda=1}^{\Lambda} \boldsymbol{A}^\lambda \otimes \mathrm{Id}_{N'_\lambda}. \tag{3}$$

That is, any such operator acts as the identity on the "multiplicity labels" in $\mathcal{H}_\lambda^{(\mathcal{C})}$, meaning that it acts identically on each of the N'_λ degenerate Krylov subspaces isomorphic to $\mathcal{H}_\lambda^{(\mathcal{A})}$.

The phenomenon of *Hilbert space fragmentation* is said to occur when the total multiplicity, $\sum_\lambda N'_\lambda$, is exponential in the system size L (Moudgalya & Motrunich, 2022). We are interested in the stronger condition where the dimension of $\mathcal{A}$ is polynomial in the system size. If a system does not have this property, we may be able to restrict to a subspace that does: we pick Λ' as a function of L to be such that

$$N = \sum_{\lambda=1}^{\Lambda'} N_\lambda = \mathrm{poly}(L). \tag{4}$$

Note that this Λ' may be a constant or it may be equal to Λ depending on the system in question and the degree of degeneracy in the Krylov subspaces. We additionally make an assumption that

$N_\lambda = \omega(1)$ for all $\lambda \in [\Lambda']$ - that is, the dimension of all sectors must grow as a function of L. Now, we can define the space

$$\mathcal{H}' = \bigoplus_{\lambda=1}^{\Lambda'} \mathcal{H}_\lambda^{(\mathcal{A})} \otimes \mathcal{H}_\lambda^{(\mathcal{C})}. \tag{5}$$

The representation in $\mathcal{H}'$ of any operator $A \in \mathcal{A}$ can then be written as

$$\boldsymbol{A} = \bigoplus_{\lambda=1}^{\Lambda'} \boldsymbol{A}^\lambda \otimes \mathrm{Id}_{N'_\lambda}. \tag{6}$$

While this construction may seem ad hoc and technical, surprisingly the phenomenon of Hilbert space fragmentation occurs in many quantum systems. It underlies the ergodicity breaking of the Temperley–Lieb model (Batchelor & Barber, 1990), the t–J_z model (Zhang et al., 1997), and the Pair-Flip model (Moudgalya & Motrunich, 2022).

One can also define an orthonormal basis of $\mathcal{H}$ respecting the decomposition of $\mathcal{H}$ into Krylov subspaces (Moudgalya & Motrunich, 2022). In analogy with the terminology for permutation-invariant quantum systems (Bacon et al., 2006), we can construct a *Schur basis* for $\mathcal{H}'$ as a set of states $|\lambda, q_\lambda, p_\lambda\rangle = |\lambda, q_\lambda\rangle \otimes |\lambda, p_\lambda\rangle$ such that the states $|\lambda, q_\lambda\rangle$ form an orthonormal basis for $\mathcal{H}_\lambda^{(\mathcal{A})}$ and the states $|\lambda, p_\lambda\rangle$ form an orthonormal basis for $\mathcal{H}_\lambda^{(\mathcal{C})}$.

2.2 The Classification Task

Given a quantum state that is a basis state promised to be in the Schur basis, we would like to perform some kind of quantum measurement to determine which state exactly it is, ignoring the degeneracy. In some special settings—such as when the system is permutation invariant, and one is interested in the associated Schur basis of that Krylov decomposition—this can be done efficiently using a quantum algorithm known as the Schur transform (Bacon et al., 2006). However, generally, we know only the generators $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_m$. Instead, we hope in effect to *learn* the Schur transform associated with $\mathcal{S}$, assuming we are also given copies of Schur basis states $|\lambda, q_\lambda, p_\lambda\rangle$. Specifically, we want to learn a transformation $\boldsymbol{U}$ such that

$$\mathrm{Tr}_{\mathcal{H}'}\left[\boldsymbol{U}^\dagger(|\lambda, q_\lambda, p_\lambda\rangle\langle\lambda, q_\lambda, p_\lambda| \otimes |\boldsymbol{0}\rangle\langle\boldsymbol{0}|)\boldsymbol{U}\right] \approx |\boldsymbol{\lambda}, \boldsymbol{q_\lambda}\rangle\langle\boldsymbol{\lambda}, \boldsymbol{q_\lambda}|. \tag{7}$$

Here, we have added an ancillary register of n_a qubits with Hilbert space of dimension $N_a = 2^{n_a}$ and initialized it in the computational basis state $|\boldsymbol{0}\rangle$. We also use the notation $|\boldsymbol{\lambda}, \boldsymbol{q_\lambda}\rangle$ to label bit strings encoded in the computational basis labeling λ and q_λ. The unitary U will be implemented as a quantum circuit that should write a bit string label $\boldsymbol{\lambda}, \boldsymbol{q_\lambda}$ to this ancillary register in the computational basis. Note that here we are ambivalent to the multiplicity label p_λ; there are exponentially many such labels, and we cannot hope to learn them efficiently. Instead, we also require that our network *generalizes* its classification of any $|\lambda, q_\lambda, p_\lambda\rangle$ given training examples of $|\lambda, q_\lambda, p'_\lambda\rangle$ for p_λ not necessarily equal to p'_λ. We call this problem *fragment classification*.

Definition 1 (Fragment Classification)**.** *Suppose that we have a dataset*

$$\mathcal{D} = \{(|x\rangle, \boldsymbol{O}_x)\}_{x\in[M]}, \tag{8}$$

of size $M \leq N$ where each $|x\rangle$ is of the form $|\lambda, q_\lambda, p_\lambda\rangle \otimes |0\rangle$, and the corresponding $\boldsymbol{O}_x$ is of the form

$$\boldsymbol{O}_x = -\mathrm{Id}_{\mathcal{H}'} \otimes |\boldsymbol{\lambda}, \boldsymbol{q_\lambda}\rangle\langle\boldsymbol{\lambda}, \boldsymbol{q_\lambda}|, \tag{9}$$

where $\boldsymbol{\lambda}, \boldsymbol{q_\lambda}$ are bit string representations of λ, q_λ on the ancillary register of $n_a = \lceil \log_2 M \rceil$ qubits. Let $N_a = 2^{n_a}$ be the dimension of the Hilbert space of the ancillary register. We say that a quantum learning algorithm solves the fragment classification problem if, given the dataset $\mathcal{D}$, for any $\epsilon > 0$ it can construct in $\mathrm{poly}(L, 1/\epsilon)$ time a quantum circuit representing a unitary $\boldsymbol{U}$ acting on $\mathcal{H}' \otimes \mathbb{C}^{N_a}$, such that for any Schur basis state $|\lambda, q_\lambda, p_\lambda\rangle$ (including those p_λ for which the state is not part of $\mathcal{D}$), we have

$$\langle\boldsymbol{\lambda}, \boldsymbol{q_\lambda}| \,\mathrm{Tr}_{\mathcal{H}'}\left[\boldsymbol{U}^\dagger(|\lambda, q_\lambda, p_\lambda\rangle\langle\lambda, q_\lambda, p_\lambda| \otimes |\boldsymbol{0}\rangle\langle\boldsymbol{0}|)\boldsymbol{U}\right] |\boldsymbol{\lambda}, \boldsymbol{q_\lambda}\rangle \geq 1 - \epsilon. \tag{10}$$

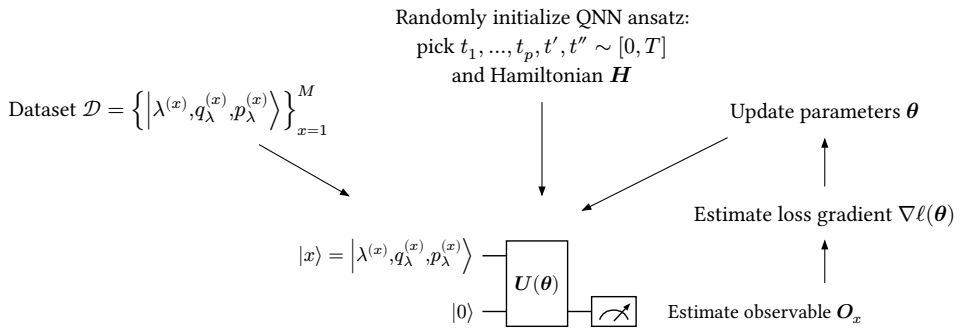


Figure 1: Diagram of our QNN training process. We are given a dataset of Schur basis states and randomly sample a QNN ansatz architecture and an initial vector of parameters (see Section 3.1). We then compile the QNN into a quantum circuit and repeatedly run on the input states to estimate the gradient of the loss function, which is then used to adjust the parameters.

Remark. *In a typical situation where fragment classification might be used, it could be difficult to obtain a dataset containing multiple orthogonal Schur basis states from a given sector λ labeled with (λ, q_λ). A more realistic scenario is one in which only λ is learned. This can be done in our formulation of the problem by simply removing the $\boldsymbol{q_\lambda}$ from the bitstring representation. Note that in this case, the model will still generalize across different p_λ (but not across different q_λ).*

2.3 Variational Quantum Algorithms

The quantum learning algorithm we will later construct will be a *variational quantum algorithm* (Peruzzo et al., 2014). Here, one hopes to learn a unitary $\boldsymbol{U}$ by tuning parameters of a variational ansatz constructed using exponentials of local operators. The variational parameters of the ansatz are then tuned using a gradient-based optimizer to minimize some loss function.

Specializing to our setting, $\boldsymbol{U}(\boldsymbol{\theta})$ will be constructed out of matrix exponentials of Hamiltonians constructed from the given generators $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_m$, and is parameterized by $\boldsymbol{\theta} \in \mathbb{R}^p$. We use gradient descent to find optimal values of the parameters to minimize the loss function:

$$\ell(\boldsymbol{\theta}) = \frac{1}{M} \sum_{x=1}^{M} \langle x | \, \boldsymbol{U}(\boldsymbol{\theta}) \boldsymbol{O}_x \boldsymbol{U}(\boldsymbol{\theta})^\dagger \, | x \rangle \,, \tag{11}$$

where each $|x\rangle$ is of the form $|\lambda, q_\lambda, p_\lambda\rangle \otimes |\boldsymbol{0}\rangle$ and the corresponding *objective observable* is

$$\boldsymbol{O}_x = -\operatorname{Id}_{\mathcal{H}'} \otimes |\boldsymbol{\lambda}, \boldsymbol{q_\lambda}\rangle\langle \boldsymbol{\lambda}, \boldsymbol{q_\lambda}|. \tag{12}$$

The training process is illustrated in Figure 1. We will give full details of the construction of $\boldsymbol{U}(\boldsymbol{\theta})$ in Section 3.1, and demonstrate that $\ell(\boldsymbol{\theta})$ can be efficiently optimized in Section 3.2.

2.4 Trainability Conditions

To formally prove that QNNs can efficiently solve this problem, we need to demonstrate the two main conditions needed for a QNN to be efficiently trainable. The following are standard criteria of QNN trainability in the literature (Anschuetz, 2025; Cerezo et al., 2025):

(1) **The absence of barren plateaus**. In the asymptotic limit of large system size, the distribution of the loss landscape over the random initialization of the QNN architecture converges to one where the derivatives only decay polynomially with the system size, thus allowing the gradient to be efficiently estimated on a quantum computer in polynomial time. We formally prove this in Theorem 1.

(2) **The absence of poor local minima**. When the number of parameters p is sufficiently large, the loss landscape enters an *overparameterized regime* where all spurious local minima disappear. We follow the approach of Larocca et al. (2023) and show the Hessian is not full-rank in order to demonstrate a lack of poor local minima. The value of p needed

for this to occur is only polynomially large in the system size, making the QNN in the overparameterized phase efficiently implementable on a quantum computer. We formally prove this in Proposition 2. We then apply the Kac-Rice formula (Adler & Taylor, 2007) to the loss function modified by a small perturbation, demonstrating the absence of poor local minima in Theorem 2.

3 Main Results

3.1 The QNN Ansatz Construction

Suppose that we have a dataset

$$\mathcal{D} = \{(|x\rangle, \boldsymbol{O}_x)\}_{x\in[M]}, \tag{13}$$

of size $M \leq N$ as in Definition 1. We assume we are given at most one example x for any given λ, q_λ pair - we will see in Proposition 1 that we do not require an example for each p_λ. Let $N_a = 2^{n_a}$ be the dimension of the Hilbert space of the ancillary register. Let

$$\mathcal{H}^* = \mathcal{H}' \otimes \mathbb{C}^{N_a} = \bigoplus_{\lambda=1}^{\Lambda'} \mathcal{H}_\lambda^* \otimes \mathcal{H}_\lambda^{(\mathcal{C})}, \tag{14}$$

where we define $\mathcal{H}_\lambda^* = \mathcal{H}_\lambda^{(\mathcal{A})} \otimes \mathbb{C}^{N_a}$, of dimension $N_\lambda^* = N_\lambda N_a$.

Now, we construct a randomized QNN ansatz parameterized by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in [0, 2\pi)^p$ as follows:

$$\boldsymbol{U}(\boldsymbol{\theta}) = e^{\mathbf{i}\boldsymbol{H}t''} \left(\prod_{i=1}^{p} e^{-\mathbf{i}\boldsymbol{H}t_i} e^{\mathbf{i}\boldsymbol{A}\theta_i} e^{\mathbf{i}\boldsymbol{H}t_i} \right) e^{-\mathbf{i}\boldsymbol{H}t'}, \tag{15}$$

where we select $t_1, \ldots, t_p, t', t''$ i.i.d. uniformly from $[0, T]$ for some fixed T, we let $\boldsymbol{H}$ be a Hamiltonian on the system and the ancillary register:

$$\boldsymbol{H} = \sum_{i=1}^{m} c_i \boldsymbol{h}_i \otimes \left(\sum_{j=1}^{n_a} c'_{i,j,x} \sigma_j^x + c'_{i,j,y} \sigma_j^y + c'_{i,j,z} \sigma_j^z \right), \tag{16}$$

and we let $\boldsymbol{A}$ be a local operator block-diagonal in the Schur basis, which we assume without loss of generality is positive semidefinite to simplify our analysis—we can just take it to be a large multiple of the identity added to $\boldsymbol{h}_1 \otimes \mathrm{Id}_{N_a}$.

We assume that coefficients in Equation (16) are chosen such that the Hamiltonian obeys the *full Eigenstate Thermalization Hypothesis* (ETH) (Foini & Kurchan, 2019) in each Krylov subspace. The full ETH is an ansatz for higher-order time-averaged correlation functions of local observables, and has been shown to imply a quantitative scrambling behavior (Fava et al., 2025). More formally, we assume the following condition holds in each Krylov subspace:

Hypothesis 1 (Full ETH, free cumulant definition, from Fava et al. (2025, Eq. (5)))**.** *We say a sequence of Hamiltonians $\boldsymbol{H}$ describing a size-L system and acting on an N-dimensional Hilbert space obeys the full ETH if the following is true. For any k, there exists some $t_k = O(\mathrm{poly}(L, k))$ such that choosing t uniformly from $[0, t_k]$ and defining the free cumulant κ_{2k} with respect to this distribution, for any local operators $\boldsymbol{A}_i(t)$ time-evolved under $\boldsymbol{H}$ and any local operators $\boldsymbol{B}_i$:*

$$\kappa_{2k}[\boldsymbol{A}_1(t), \boldsymbol{B}_1, \boldsymbol{A}_2(t), \boldsymbol{B}_2, \ldots, \boldsymbol{A}_k(t), \boldsymbol{B}_k] = O(1/N). \tag{17}$$

Remark. *It is widely believed that the full ETH generally holds in physical systems in the absence of symmetries. While we make this assumption to make the calculations for the gradient distribution more tractable, we expect the ETH condition to be a "worst case" for barren plateaus: if the variational ansatz does not uniformly mix over the accessible Hilbert space and is biased towards a smaller subspace, we would expect the gradients of the loss landscape to vanish less quickly.*

Observe that we can block-diagonalize all operators in Equation (15), writing them as a sum over the irreducible representations (algebraic sectors) of $\mathcal{A}$. We identify the operators with their representations in $\mathcal{H}^*$:

$$\boldsymbol{U}(\boldsymbol{\theta}) = \bigoplus_{\lambda=1}^{\Lambda'} \boldsymbol{U}^\lambda(\boldsymbol{\theta}), \tag{18}$$

where

$$\boldsymbol{U}^{\lambda}(\boldsymbol{\theta}) = e^{\mathbf{i}\boldsymbol{H}^{\lambda}t''}\left(\prod_{i=1}^{p} e^{-\mathbf{i}\boldsymbol{H}^{\lambda}t_i}e^{\mathbf{i}\boldsymbol{A}^{\lambda}\theta_i}e^{\mathbf{i}\boldsymbol{H}^{\lambda}t_i}\right)e^{-\mathbf{i}\boldsymbol{H}^{\lambda}t'}. \tag{19}$$

Define the loss function

$$\ell(\boldsymbol{\theta}) = \frac{1}{M}\sum_{x=1}^{M}\ell_x(\boldsymbol{\theta}) = \frac{1}{M}\sum_{x=1}^{M}\langle x|\,\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{O}_x\boldsymbol{U}(\boldsymbol{\theta})^{\dagger}\,|x\rangle\,. \tag{20}$$

We can also decompose this by sector. Note that since each input state is restricted to one of the sectors, we can write

$$\ell(\boldsymbol{\theta}) = \sum_{\lambda=1}^{\Lambda'}\sum_{x=1}^{M_\lambda}\ell_x^{\lambda}(\boldsymbol{\theta}) = \frac{1}{M}\sum_{\lambda=1}^{\Lambda'}\sum_{x=1}^{M_\lambda}\langle x|\,\boldsymbol{U}^{\lambda}(\boldsymbol{\theta})\boldsymbol{O}_x^{\lambda}\boldsymbol{U}^{\lambda}(\boldsymbol{\theta})^{\dagger}\,|x\rangle\,. \tag{21}$$

Note that the indexing of x is separate for each λ, so we have slightly abused notation here and are letting each x in the above equation corresponds to some q_λ, where M_λ is the number of dataset entries in the λ block.

We initially randomly choose the QNN ansatz by sampling $t_1, \ldots, t_p, t', t''$ and choosing the Hamiltonian $\boldsymbol{H}$. Then, these become fixed, and only the parameterized parts of the form $e^{\mathbf{i}\boldsymbol{A}\theta_i}$ will be changed during training as we adjust the parameters θ_i. To analyze the gradient of the loss function, we will consider its distribution at any particular value of $\boldsymbol{\theta}$ over the random choice of the QNN ansatz. We perform a sequence of reductions, first reducing the distribution to the case where we only have to consider $\boldsymbol{\theta} = 0$ and then writing the loss derivatives in terms of Gaussian-distributed random variables in the asymptotic limit, allowing us to prove the result about the variance of the gradient.

3.2 Efficient Trainability

First, we establish the prior claim that indeed the training of our QNN ansatz does not require a data set of Schur basis states over all multiplicity labels.

Proposition 1. *If we take the (exponential-size) dataset $\mathcal{D}'$ consisting of every Schur basis state $|\lambda, q_\lambda, p_\lambda\rangle$ for λ, q_λ present in $\mathcal{D}$, the loss $\ell_{\mathcal{D}'}(\boldsymbol{\theta})$ with respect to that dataset is upper-bounded as $\ell_{\mathcal{D}'}(\boldsymbol{\theta}) \leq \max_{\lambda,x}\ell_x^{\lambda}(\boldsymbol{\theta})$.*

Proof. This follows from the fact that every layer of the QNN is constructed from matrix exponentials of elements of $\mathcal{A}$. Thus, conjugating by $\boldsymbol{U}(\boldsymbol{\theta})$ will act in the same way on each degenerate Krylov subspace, which means that the loss contribution from the new datapoints for a given λ, q_λ will be exactly the same as that for the point in the original dataset. So, we can write

$$\ell_{\mathcal{D}'}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{D}'|}\sum_{\lambda=1}^{\Lambda'}N'_\lambda\sum_{x=1}^{M_\lambda}\ell_x^{\lambda}(\boldsymbol{\theta}) \leq \frac{\max_{\lambda,x}\ell_x^{\lambda}(\boldsymbol{\theta})}{|\mathcal{D}'|}\sum_{\lambda=1}^{\Lambda'}N'_\lambda M_\lambda \leq \max_{\lambda,x}\ell_x^{\lambda}(\boldsymbol{\theta}). \tag{22}$$

□

Now, we want to analyze the derivatives of this loss function.

Lemma 1. *At $\boldsymbol{\theta} = 0$, we have that*

$$\partial_i\ell(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=0} = \frac{\mathbf{i}}{M}\sum_{\lambda=1}^{\Lambda'}\sum_{x=1}^{M_\lambda}\langle x|\,e^{\mathbf{i}\boldsymbol{H}^{\lambda}t''}\left[e^{-\mathbf{i}\boldsymbol{H}^{\lambda}t_i}\boldsymbol{A}^{\lambda}e^{\mathbf{i}\boldsymbol{H}^{\lambda}t_i},\ e^{-\mathbf{i}\boldsymbol{H}^{\lambda}t'}\boldsymbol{O}_x^{\lambda}e^{\mathbf{i}\boldsymbol{H}^{\lambda}t'}\right]e^{-\mathbf{i}\boldsymbol{H}^{\lambda}t''}\,|x\rangle\,. \tag{23}$$

Now, we want to analyze the distribution of the gradient of the loss function. To do this, we consider joint distributions of individual terms that contribute to the expression for the gradient.

Using Lemma 5, we can perform a reduction where we approximate the distribution of the first derivatives by taking each $\boldsymbol{g}_i^{\lambda}, \boldsymbol{g}'^{\lambda}, \boldsymbol{g}''^{\lambda}$ to be Haar-random.

Lemma 2. *We can approximate, at any value of $\boldsymbol{\theta}$,*

$$\partial_i \ell_x^\lambda(\boldsymbol{\theta}) \rightsquigarrow \frac{\mathbf{i}}{M} \langle x| \boldsymbol{g}''^{\lambda\dagger} \left[\boldsymbol{g}_i^\lambda \boldsymbol{A}^\lambda \boldsymbol{g}_i^{\lambda\dagger},\ \boldsymbol{g}'^\lambda \boldsymbol{O}_x^\lambda \boldsymbol{g}'^{\lambda\dagger} \right] \boldsymbol{g}''^\lambda |x\rangle , \tag{24}$$

where $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_p, \boldsymbol{g'}, \boldsymbol{g''} \sim$ Haar up to an $o(1)$ error in the Lévy-Prokhorov metric (Definition 2), if we choose a sufficiently large T and we assume that $N_{\min} = \min_\lambda N_\lambda = \omega(1)$. Furthermore, this approximation is valid for the joint distribution of any two *gradient contributions $\partial_i \ell_x^\lambda(\boldsymbol{\theta})$ and $\partial_j \ell_y^{\lambda'}(\boldsymbol{\theta})$: their joint distribution can be approximated up to an $o(1)$ Lévy-Prokhorov error by replacing the Hamiltonian time evolution with Haar-random matrices.*

Lemma 3. *We can approximate, at any value of $\boldsymbol{\theta}$,*

$$\partial_i \ell_x^\lambda(\boldsymbol{\theta}) \rightsquigarrow \frac{\mathbf{i}}{M N_\lambda^{*2}} \langle x| \boldsymbol{g}''^{\lambda\dagger} \left[\widetilde{\boldsymbol{g}}_i^\lambda \boldsymbol{A}^\lambda \widetilde{\boldsymbol{g}}_i^{\lambda\dagger},\ \widetilde{\boldsymbol{g}}'^\lambda \boldsymbol{O}_x^\lambda \widetilde{\boldsymbol{g}}'^{\lambda\dagger} \right] \boldsymbol{g}''^\lambda |x\rangle , \tag{25}$$

where $\widetilde{\boldsymbol{g}}_1^\lambda, \ldots, \widetilde{\boldsymbol{g}}_p^\lambda, \widetilde{\boldsymbol{g}}'^\lambda$ have i.i.d. standard Gaussian entries, up to an $o(1)$ error in the Lévy-Prokhorov metric, and the same approximation is valid for the joint distribution of any two loss contributions $\partial_i \ell_x^\lambda(\boldsymbol{\theta})$ and $\partial_j \ell_y^{\lambda'}(\boldsymbol{\theta})$.

We know by construction that $\boldsymbol{A}$ is positive semidefinite. We can diagonalize it, writing each block as

$$\boldsymbol{A}^\lambda = \sum_{\mu=1}^{N_\lambda^*} a_\mu^\lambda |\mu\rangle\langle\mu| \tag{26}$$

Now, we want to perform an additional reduction to replace $\boldsymbol{A}$ with a *semi-isotropic* $\widetilde{\boldsymbol{A}}$, which we define as:

$$\widetilde{\boldsymbol{A}}^\lambda = \frac{\operatorname{Tr}(\boldsymbol{A}^\lambda)}{r_A^\lambda} \sum_{\mu=1}^{r_A^\lambda} |\mu\rangle\langle\mu|, \tag{27}$$

where

$$r_A^\lambda = \left\lceil \frac{\operatorname{Tr}(\boldsymbol{A}^\lambda)^2}{\operatorname{Tr}((\boldsymbol{A}^\lambda)^2)} \right\rceil . \tag{28}$$

That is, we want to replace $\boldsymbol{A}^\lambda$ with a matrix all of whose nonzero eigenvalues are the same, preserving the average eigenvalue of $\boldsymbol{A}^\lambda$.

Observe that

$$\operatorname{Tr}\left(\widetilde{\boldsymbol{A}}^\lambda\right) = \operatorname{Tr}\left(\boldsymbol{A}^\lambda\right) \tag{29}$$

and (up to a small incurred by rounding r_A^λ to an integer),

$$\operatorname{Tr}\left((\widetilde{\boldsymbol{A}}^\lambda)^2\right) = \frac{\operatorname{Tr}(\boldsymbol{A}^\lambda)^2}{r_A^\lambda} = \operatorname{Tr}((\boldsymbol{A}^\lambda)^2). \tag{30}$$

Lemma 4. *Replacing $\boldsymbol{A}$ with $\widetilde{\boldsymbol{A}}$, we can write*

$$\partial_i \ell_x^\lambda(\boldsymbol{\theta}) \rightsquigarrow \hat{\ell}_{x;i}^\lambda := \frac{\mathbf{i}}{M N_\lambda^{*2}} \langle x| \boldsymbol{g}''^{\lambda\dagger} \left[\widetilde{\boldsymbol{g}}_i^\lambda \widetilde{\boldsymbol{A}}^\lambda \widetilde{\boldsymbol{g}}_i^{\lambda\dagger},\ \widetilde{\boldsymbol{g}}'^\lambda \boldsymbol{O}_x^\lambda \widetilde{\boldsymbol{g}}'^{\lambda\dagger} \right] \boldsymbol{g}''^\lambda |x\rangle , \tag{31}$$

up to an $o(1)$ error in the Lévy-Prokhorov metric, assuming that $N_{\min} = \min_\lambda N_\lambda = \omega(1)$. The same is true for joint distributions of two derivative contributions as before.

Now that we have performed these reductions, we can, up to a vanishing error in Lévy-Prokhorov metric, approximate the distribution of the gradient of the loss function as some $\widehat{\nabla \ell}$ for which we can find a variance lower bound.

Theorem 1. *At any value of $\boldsymbol{\theta} \in [0, 2\pi)^p$,*

$$\mathbb{E}\left[\left\| \widehat{\nabla \ell} \right\|^2 \right] = \sum_{\lambda=1}^{\Lambda'} \frac{2 M_\lambda (N_a - 1) \operatorname{Tr}((\boldsymbol{A}^\lambda)^2) p}{M^2 N_a^4 N_\lambda^2} = \Omega\left(\frac{1}{\operatorname{poly}(L)} \right), \tag{32}$$

where

$$\left(\widehat{\nabla\ell}\right)_i := \sum_{\lambda=1}^{\Lambda'} \sum_{x=1}^{M_\lambda} \hat{\ell}^{\lambda}_{x;i}. \tag{33}$$

By proving Theorem 1, we have formally established one of the two main claims: the absence of barren plateaus. Since the gradient of the loss function only scales inversely polynomially with L and the quantum circuit for our QNN ansatz is polynomial-size, it is efficient to estimate gradients on a quantum computer by performing repeated runs and measurements with varied parameters, and thus gradient descent can be efficiently performed.

We now argue that the model enters the overparameterized regime, as in Larocca et al. (2023), by showing that the Hessian of the loss landscape stops being full-rank when the number of parameters p is only polynomial in the system size.

Proposition 2. *For some $p = O(\mathrm{poly}(L))$, the Hessian matrix $\nabla^2\ell$ at any value of $\boldsymbol{\theta} \in [0, 2\pi)^p$ (where $\langle i| \left(\nabla^2\ell\right) |j\rangle = \partial_i\partial_j\ell(\boldsymbol{\theta})$) does not have full rank.*

This rank deficiency allows us to constrain the structure of local optima through the use of the Kac–Rice formula (Adler & Taylor, 2007), as was done in previous work on machine learning loss landscapes (Choromanska et al., 2015; Anschuetz, 2022; Anschuetz & Kiani, 2022; Anschuetz, 2025). Informally, the Kac–Rice formula relates the joint distribution of the gradient and the Hessian to the expected density of local optima of a random field when that random field is a Morse function with high probability; when the expected Hessian vanishes, there are no spurious local optima.

To make this connection more precise, we consider the joint distribution of $\left(\widehat{\nabla\ell}, \nabla^2\ell\right)$, where $\widehat{\nabla\ell}$ is as we computed in Lemma 4. As ℓ is not a Morse function w.h.p.—a condition necessary for the Kac–Rice formula to hold—we add an order-ϵ perturbation to the loss function and consider the scaling behavior of the density as $\epsilon \to 0^+$. We show in this limit that the density of local minima above the global minimum vanishes. Thus, we have the following:

Theorem 2 (No spurious local minima, informal version of Theorem 3)**.** *For some $p = O(\mathrm{poly}(L))$, the expected number of local minima at loss value greater than the value of the global minimum is zero.*

The formal proofs and statements of the results in this section can be found in Appendix A.

3.3 Conjectured Classical Hardness

There exist many classical algorithms for efficiently learning properties of physical systems in the presence of symmetries (Anschuetz et al., 2023a; Goh et al., 2025; Cerezo et al., 2025). Fundamentally, the strategies of these algorithms make use of the fact that, when the algebraic structure of a learning problem is known a priori, classical algorithms can leverage this knowledge to circumvent the exponential scaling of naive classical simulation methods.

Our setting differs from these prior directions in one important way: we have no prior knowledge of the algebraic structure of the system (indeed, this is the property we are hoping to learn). While we have no formal reduction connecting this problem to standard complexity theoretic conjectures, at a high level, this resembles other problems conjectured to be difficult for classical algorithms, including the hidden subgroup problem (Kitaev, 1995) and the state hidden subgroup problem (Bouland et al., 2025).

The classical algorithms in the line of work mentioned above take as input the structure coefficients $c^k_{i,j}$ of the system's Lie or Jordan algebra, where the action of the algebra's operation $\circ$ (i.e. a commutator or anticommutator) on a basis $\{b_i\}_i$ of the algebra is given by

$$b_i \circ b_j = \sum_k c^k_{i,j} b_k. \tag{34}$$

In our setting, we are given the algebra's generators $\{\boldsymbol{h}_i\}_i$. To obtain a basis and compute the structure coefficients, one would need to take nested group operations of the form $\boldsymbol{h}_{i_1} \circ (\boldsymbol{h}_{i_2} \circ (\cdots \circ \boldsymbol{h}_{i_l}) \dots)$, where the depth may be as large as the dimension of the algebra. If we

have some sparse classical representation of the generators (for instance, Pauli strings), we cannot guarantee that these nested group operations will have such an efficient representation: in general, their size will blow up exponentially in the depth.

As we have demonstrated that this problem can be efficiently solved using quantum neural networks, this establishes one of the few known settings (outside of sampling-based tasks (Huang et al., 2024)) where a quantum machine learning task has no known dequantization.

4 APPLICATIONS

The problem we presented as a showcase for our QNN construction is physically motivated, and we hope that it may be of practical interest. Recent experimental work has demonstrated examples of Hilbert space fragmentation (Bluvstein et al., 2021; Adler et al., 2024), where the fragmentation phenomenon was not studied algebraically but rather by observing the dynamics of different states when performing a quench. For instance, Adler et al. (2024) observes different relaxation dynamics between "dimer" and "chequerboard" states, which belong to different Krylov sectors λ. The authors perform Hamiltonian evolution under different times in order to generate a collection of states in each of the sectors. Such an experimental setup could be used to create a dataset for training our QNN model.

Now, suppose that some other experiment prepares some unknown state on this system and one wishes to classify which sector it lies in. It should be possible to perform reverse time evolution $e^{\mathrm{i}Ht}$ on this state to obtain some known initial state (e.g. the "dimer" or "chequerboard" states), but the time t is not known. Furthermore, the state may be in a subspace corresponding to a different label p_λ, so it could be difficult to perform the correct measurement to extract the necessary information from the state without a characterization of the system's algebraic structure. But if the state is given to the trained QNN model, it would perform the classification correctly as long as it is provided with training data from the sectors of interest.

5 CONCLUSION

In this work, we demonstrated a physically motivated problem setting where (1) quantum neural networks efficiently solve the problem under a realistic set of assumptions and (2) there are no known existing classical algorithms for simulating these networks, despite the substantial recent progress in methods for dequantizing quantum learning algorithms (Anschuetz et al., 2023a; Goh et al., 2025; Cerezo et al., 2025). At a high level, our results followed by finding a setting where the quantum network was able to take advantage of algebraic structure present in a problem, but for which this structure is not given a convenient classical description for a classical simulator to take advantage of. We invite others to test our results by attempting to develop classical simulation methods in this more restricted setting, as empirical tests of the tractability of this task have implications for not only the potential utility of the algorithm presented here, but also more generally for determining unknown algebraic properties of symmetric quantum systems.

While we examine one specific setting with this property, we believe this strategy for finding physically motivated problems for quantum learning algorithms can be more generally leveraged. We hope this work motivates future studies of settings where quantum neural networks can take advantage of symmetries that are inaccessible to classical computers.

AI USE STATEMENT

In this work, we used generative AI tools to help with understanding relevant technical concepts, to survey related work on QNN loss landscapes, and assist with certain routine calculations and steps such as some of the bounds on the Lévy-Prokhorov metric in Appendix B. We have checked the correctness of these calculations. AI tools were *not* used to generate proofs, develop theoretical models, design research methodologies, or write the text of the paper. The code for the numerics in Appendix C was written without AI assistance, apart from some queries about the usage of library functions. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

We provide formal proofs of our main results in Appendix A, with additional details on error bounds in Appendix B. We also provide small-scale numerical experiments in Appendix C. The code used to run these experiments is available at https://anonymous.4open.science/r/fragmentation-qnn-anonymized-8415/qnn.ipynb.

References

Daniel Adler, David Wei, Melissa Will, Kritsana Srakaew, Suchita Agrawal, Pascal Weckesser, Roderich Moessner, Frank Pollmann, Immanuel Bloch, and Johannes Zeiher. Observation of hilbert space fragmentation and fractonic excitations in 2d. *Nature*, 636(8041):80–85, Dec 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-08188-0. URL https://doi.org/10.1038/s41586-024-08188-0.

Robert J. Adler and Jonathan E. Taylor. *Random Fields on Manifolds*, pp. 301–330. Springer New York, New York, NY, 2007. ISBN 978-0-387-48116-6. doi: 10.1007/978-0-387-48116-6_12. URL https://doi.org/10.1007/978-0-387-48116-6_12.

Eric R. Anschuetz. A unified theory of quantum neural network loss landscapes. In *International Conference on Learning Representations*. OpenReview, 2025. URL https://openreview.net/forum?id=fv8TTt9srF.

Eric R. Anschuetz and Xun Gao. Arbitrary polynomial separations in trainable quantum machine learning, 2024.

Eric R. Anschuetz and Bobak T. Kiani. Quantum variational algorithms are swamped with traps. *Nat. Commun.*, 13:7760, 2022. doi: 10.1038/s41467-022-35364-5.

Eric R. Anschuetz, Andreas Bauer, Bobak T. Kiani, and Seth Lloyd. Efficient classical algorithms for simulating symmetric quantum systems. *Quantum*, 7:1189, November 2023a. ISSN 2521-327X. doi: 10.22331/q-2023-11-28-1189. URL http://dx.doi.org/10.22331/q-2023-11-28-1189.

Eric R. Anschuetz, Hong-Ye Hu, Jin-Long Huang, and Xun Gao. Interpretable quantum advantage in neural sequence learning. *PRX Quantum*, 4:020338, 6 2023b. doi: 10.1103/PRXQuantum.4.020338.

Eric Ricardo Anschuetz. Critical points in quantum generative models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=2f1z55GVQN.

Dave Bacon, Isaac L. Chuang, and Aram W. Harrow. Efficient quantum circuits for schur and clebsch-gordan transforms. *Phys. Rev. Lett.*, 97:170502, October 2006. doi: 10.1103/PhysRevLett.97.170502.

M T Batchelor and M N Barber. Spin-s quantum chains and temperley-lieb algebras. *Journal of Physics A: Mathematical and General*, 23(1):L15, jan 1990. doi: 10.1088/0305-4470/23/1/004. URL https://doi.org/10.1088/0305-4470/23/1/004.

Istvan Berkes and Walter Philipp. Approximation Thorems for Independent and Weakly Dependent Random Vectors. *The Annals of Probability*, 7(1):29 – 54, 1979. doi: 10.1214/aop/1176995146. URL https://doi.org/10.1214/aop/1176995146.

Hannes Bernien, Sylvain Schwartz, Alexander Keesling, Harry Levine, Ahmed Omran, Hannes Pichler, Soonwon Choi, Alexander S Zibrov, Manuel Endres, Markus Greiner, et al. Probing many-body dynamics on a 51-atom quantum simulator. *Nature*, 551(7682):579–584, 2017. doi: 10.1038/nature24622.

Dominic W Berry, Yuan Su, Casper Gyurik, Robbie King, Joao Basso, Alexander Del Toro Barba, Abhishek Rajput, Nathan Wiebe, Vedran Dunjko, and Ryan Babbush. Analyzing prospects for quantum advantage in topological data analysis. *PRX Quantum*, 5(1):010319, 2024.

D. Bluvstein, A. Omran, H. Levine, A. Keesling, G. Semeghini, S. Ebadi, T. T. Wang, A. A. Michailidis, N. Maskara, W. W. Ho, S. Choi, M. Serbyn, M. Greiner, V. Vuletić, and M. D. Lukin. Controlling quantum many-body dynamics in driven rydberg atom arrays. *Science*, 371 (6536):1355–1359, 2021. doi: 10.1126/science.abg2530. URL https://www.science.org/doi/abs/10.1126/science.abg2530.

Adam Bouland, Tudor Giurgică-Tiron, and John Wright. The state hidden subgroup problem and an efficient algorithm for locating unentanglement. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, STOC '25, pp. 463–470, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715105. doi: 10.1145/3717823.3718118.

M. Cerezo, Martin Larocca, Diego García-Martín, N. L. Diaz, Paolo Braccia, Enrico Fontana, Manuel S. Rudolph, Pablo Bermejo, Aroosa Ijaz, Supanut Thanasilp, Eric R. Anschuetz, and Zoë Holmes. Does provable absence of barren plateaus imply classical simulability? *Nat. Commun.*, 16(1):7907, 2025. doi: 10.1038/s41467-025-63099-6.

Marco Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nat. Commun.*, 12(1):1791–1802, 2021. doi: 10.1038/s41467-021-21728-w.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gerard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In Guy Lebanon and S. V. N. Vishwanathan (eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 192–204, San Diego, USA, 2015. PMLR. URL http://proceedings.mlr.press/v38/choromanska15.html.

Michele Fava, Jorge Kurchan, and Silvia Pappalardi. Designs via free probability. *Phys. Rev. X*, 15:011031, Feb 2025. doi: 10.1103/PhysRevX.15.011031. URL https://link.aps.org/doi/10.1103/PhysRevX.15.011031.

Laura Foini and Jorge Kurchan. Eigenstate thermalization hypothesis and out of time order correlators. *Phys. Rev. E*, 99:042139, April 2019. doi: 10.1103/PhysRevE.99.042139.

Joe Gibbs, Zoë Holmes, Matthias C. Caro, Nicholas Ezzell, Hsin-Yuan Huang, Lukasz Cincio, Andrew T. Sornborger, and Patrick J. Coles. Dynamical simulation via quantum machine learning with provable generalization. *Phys. Rev. Res.*, 6:013241, March 2024.

Matthew L. Goh, Martin Larocca, Lukasz Cincio, M. Cerezo, and Frédéric Sauvage. Lie-algebraic classical simulations for quantum computing. *Physical Review Research*, 7(3), September 2025. ISSN 2643-1564. doi: 10.1103/3y65-f5w6. URL http://dx.doi.org/10.1103/3y65-f5w6.

Casper Gyurik, Alexander Schmidhuber, Robbie King, Vedran Dunjko, and Ryu Hayakawa. Quantum computing and persistence in topological data analysis. *arXiv preprint arXiv:2410.21258*, 2024.

Aram W Harrow, Avinatan Hassidim, and Seth Lloyd. Quantum algorithm for linear systems of equations. *Physical review letters*, 103(15):150502, 2009.

Hsin-Yuan Huang, Yunchao Liu, Michael Broughton, Isaac Kim, Anurag Anshu, Zeph Landau, and Jarrod R. McClean. Learning shallow quantum circuits. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, STOC 2024, pp. 1343–1351, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703836. doi: 10.1145/3618260.3649722.

Tiefeng Jiang. The entries of haar-invariant matrices from the classical compact groups. *Journal of Theoretical Probability*, 23(4):1227–1243, Dec 2010. ISSN 1572-9230. doi: 10.1007/s10959-009-0241-7. URL https://doi.org/10.1007/s10959-009-0241-7.

Vedika Khemani, Michael Hermele, and Rahul Nandkishore. Localization from Hilbert space shattering: From theory to physical realizations. *Phys. Rev. B*, 101:174204, May 2020. doi: 10.1103/PhysRevB.101.174204.

A. Yu. Kitaev. Quantum measurements and the Abelian stabilizer problem, 1995.

N. P. Landsman. Lecture notes on C*-algebras, Hilbert C*-modules, and quantum mechanics. *arXiv e-prints*, art. math-ph/9807030, July 1998. doi: 10.48550/arXiv.math-ph/9807030.

Martín Larocca, Nathan Ju, Diego García-Martín, Patrick J. Coles, and Marco Cerezo. Theory of overparametrization in quantum neural networks. *Nature Computational Science*, 3(6):542–551, Jun 2023. ISSN 2662-8457. doi: 10.1038/s43588-023-00467-6. URL https://doi.org/10.1038/s43588-023-00467-6.

Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nat. Commun.*, 9(1):4812, 2018. doi: 10.1038/s41467-018-07090-4.

Johannes Jakob Meyer, Marian Mularski, Elies Gil-Fuster, Antonio Anna Mele, Francesco Arzani, Alissa Wilms, and Jens Eisert. Exploiting symmetry in variational quantum machine learning. *PRX Quantum*, 4:010328, 3 2023. doi: 10.1103/PRXQuantum.4.010328.

Sanjay Moudgalya and Olexei I. Motrunich. Hilbert space fragmentation and commutant algebras. *Phys. Rev. X*, 12:011050, March 2022. doi: 10.1103/PhysRevX.12.011050.

Sanjay Moudgalya and Olexei I. Motrunich. Exhaustive characterization of quantum many-body scars using commutant algebras. *Physical Review X*, 14(4), December 2024. ISSN 2160-3308. doi: 10.1103/physrevx.14.041069. URL http://dx.doi.org/10.1103/PhysRevX.14.041069.

Radford M. Neal. *Priors for Infinite Networks*, pp. 29–53. Springer New York, New York, NY, 1996. ISBN 978-1-4612-0745-0. doi: 10.1007/978-1-4612-0745-0_2. URL https://doi.org/10.1007/978-1-4612-0745-0_2.

Quynh T. Nguyen, Louis Schatzki, Paolo Braccia, Michael Ragone, Patrick J. Coles, Frédéric Sauvage, Martín Larocca, and M. Cerezo. Theory for equivariant quantum neural networks. *PRX Quantum*, 5:020328, May 2024. doi: 10.1103/PRXQuantum.5.020328.

Shriya Pai, Michael Pretko, and Rahul M. Nandkishore. Localization in fractonic random circuits. *Phys. Rev. X*, 9:021003, April 2019. doi: 10.1103/PhysRevX.9.021003.

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. doi: 10.1038/ncomms5213.

Michael Ragone, Bojko N Bakalov, Frédéric Sauvage, Alexander F Kemper, Carlos Ortiz Marrero, Martín Larocca, and M Cerezo. A Lie algebraic theory of barren plateaus for deep parameterized quantum circuits. *Nat. Commun.*, 15(1):7172, 2024. doi: 10.1038/s41467-024-49909-3.

Pablo Sala, Tibor Rakovszky, Ruben Verresen, Michael Knap, and Frank Pollmann. Ergodicity breaking arising from Hilbert space fragmentation in dipole-conserving hamiltonians. *Phys. Rev. X*, 10:011047, February 2020. doi: 10.1103/PhysRevX.10.011047.

Louis Schatzki, Martin Larocca, Quynh T Nguyen, Frederic Sauvage, and Marco Cerezo. Theoretical guarantees for permutation-equivariant quantum neural networks. *npj Quantum Inf.*, 10(1):12, 2024. doi: 10.1038/s41534-024-00804-1.

Peter W Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM review*, 41(2):303–332, 1999.

V. Strassen. The Existence of Probability Measures with Given Marginals. *The Annals of Mathematical Statistics*, 36(2):423 – 439, 1965. doi: 10.1214/aoms/1177700153. URL https://doi.org/10.1214/aoms/1177700153.

Christopher J Turner, Alexios A Michailidis, Dmitry A Abanin, Maksym Serbyn, and Zlatko Papić. Weak ergodicity breaking from quantum many-body scars. *Nat. Phys.*, 14(7):745–749, 2018. doi: 10.1038/s41567-018-0137-5.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Dan Voiculescu. Limit laws for random matrices and free products. *Inventiones mathematicae*, 104 (1):201–220, Dec 1991. ISSN 1432-1297. doi: 10.1007/BF01245072. URL https://doi.org/10.1007/BF01245072.

Maxwell T. West, Jamie Heredge, Martin Sevior, and Muhammad Usman. Provably trainable rotationally equivariant quantum machine learning. *PRX Quantum*, 5:030320, July 2024. doi: 10.1103/PRXQuantum.5.030320.

Xuchen You, Shouvanik Chakrabarti, and Xiaodi Wu. A convergence theory for over-parameterized variational quantum eigensolvers, 2022.

Shu Zhang, Michael Karbach, Gerhard Müller, and Joachim Stolze. Charge and spin dynamics in the one-dimensional t-j_z and t-j models. *Phys. Rev. B*, 55:6491–6503, Mar 1997. doi: 10.1103/PhysRevB.55.6491. URL https://link.aps.org/doi/10.1103/PhysRevB.55.6491.

A Formal Proofs of the Main Results

Lemma 1. *At $\boldsymbol{\theta} = 0$, we have that*

$$\partial_i \ell(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=0} = \frac{\mathbf{i}}{M} \sum_{\lambda=1}^{\Lambda'} \sum_{x=1}^{M_\lambda} \langle x| \, e^{\mathbf{i}\boldsymbol{H}^\lambda t''} \left[e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i} \boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_i}, \; e^{-\mathbf{i}\boldsymbol{H}^\lambda t'} \boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'} \right] e^{-\mathbf{i}\boldsymbol{H}^\lambda t''} \, |x\rangle \, . \tag{23}$$

Proof.

$$\begin{aligned}
&\frac{\partial}{\partial \theta_i} \ell(\boldsymbol{\theta}) = \frac{\mathbf{i}}{M} \sum_{\lambda=1}^{\Lambda'} \sum_{x=1}^{M_\lambda} \frac{\partial}{\partial \theta_i} \langle x| \, \boldsymbol{U}^\lambda(\boldsymbol{\theta}) \boldsymbol{O}_x^\lambda \boldsymbol{U}^\lambda(\boldsymbol{\theta})^\dagger \, |x\rangle = \\
&= \frac{\mathbf{i}}{M} \sum_{\lambda=1}^{\Lambda'} \sum_{x=1}^{M_\lambda} \frac{\partial}{\partial \theta_i} \langle x| \, e^{\mathbf{i}\boldsymbol{H}^\lambda t''} \left(\prod_{j=1}^{p} e^{-\mathbf{i}\boldsymbol{H}^\lambda t_j} e^{\mathbf{i}\theta_j \boldsymbol{A}^\lambda} e^{\mathbf{i}\boldsymbol{H}^\lambda t_j} \right) e^{-\mathbf{i}\boldsymbol{H}^\lambda t'} \boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'} \\
&\qquad \left(\prod_{j=p}^{1} e^{-\mathbf{i}\boldsymbol{H}^\lambda t_j} e^{-\mathbf{i}\theta_j \boldsymbol{A}^\lambda} e^{\mathbf{i}\boldsymbol{H}^\lambda t_j} \right) e^{-\mathbf{i}\boldsymbol{H}^\lambda t''} \, |x\rangle = \\
&= \frac{\mathbf{i}}{M} \sum_{\lambda=1}^{\Lambda'} \sum_{x=1}^{M_\lambda} \langle x| \left(e^{\mathbf{i}\boldsymbol{H}^\lambda t''} \left(\prod_{j=1}^{p} e^{-\mathbf{i}\boldsymbol{H}^\lambda t_j} e^{\mathbf{i}\theta_j \boldsymbol{A}^\lambda} \left(\boldsymbol{A}^\lambda \right)^{\delta_{ij}} e^{\mathbf{i}\boldsymbol{H}^\lambda t_j} \right) e^{-\mathbf{i}\boldsymbol{H}^\lambda t'} \boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'} \cdot \right. \\
&\qquad \cdot \left(\prod_{j=p}^{1} e^{-\mathbf{i}\boldsymbol{H}^\lambda t_j} e^{-\mathbf{i}\theta_j \boldsymbol{A}^\lambda} e^{\mathbf{i}\boldsymbol{H}^\lambda t_j} \right) e^{-\mathbf{i}\boldsymbol{H}^\lambda t''} - \\
&\qquad - e^{\mathbf{i}\boldsymbol{H}^\lambda t''} \left(\prod_{j=1}^{p} e^{-\mathbf{i}\boldsymbol{H}^\lambda t_j} e^{\mathbf{i}\theta_j \boldsymbol{A}^\lambda} e^{\mathbf{i}\boldsymbol{H}^\lambda t_j} \right) e^{-\mathbf{i}\boldsymbol{H}^\lambda t'} \boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'} \cdot \\
&\qquad \left. \cdot \left(\prod_{j=p}^{1} e^{-\mathbf{i}\boldsymbol{H}^\lambda t_j} e^{-\mathbf{i}\theta_j \boldsymbol{A}^\lambda} \left(\boldsymbol{A}^\lambda \right)^{\delta_{ij}} e^{\mathbf{i}\boldsymbol{H}^\lambda t_j} \right) e^{-\mathbf{i}\boldsymbol{H}^\lambda t''} \right) |x\rangle \, .
\end{aligned} \tag{35}$$

At $\boldsymbol{\theta} = 0$, this simplifies to

$$\partial_i \ell(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=0} = \frac{\mathbf{i}}{M} \sum_{\lambda=1}^{\Lambda'} \sum_{x=1}^{M_\lambda} \langle x| \, e^{\mathbf{i}\boldsymbol{H}^\lambda t''} \left[e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i} \boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_i}, \; e^{-\mathbf{i}\boldsymbol{H}^\lambda t'} \boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'} \right] e^{-\mathbf{i}\boldsymbol{H}^\lambda t''} \, |x\rangle \, . \tag{36}$$

□

Lemma 5. *Consider a set of nonnegative integers* $\{k_{\lambda,x,i,1}, k_{\lambda,x,i,2} : \lambda \in [\Lambda'], x \in [M_\lambda], i \in [p]\}$. *Let*

$$k = \sum_{\lambda=1}^{\Lambda'} \sum_{x=1}^{M_\lambda} \sum_{i=1}^{p} \left(k_{\lambda,x,i,1} + k_{\lambda,x,i,2}\right). \tag{37}$$

Let $\boldsymbol{H}$ *be a Hamiltonian satisfying the full Eigenstate Thermalization Hypothesis (Hypothesis 1). Then, there exists some* $T = \mathrm{poly}(L, k)$ *such that*

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{t_1,\ldots,t_p,\\ t',t''\sim\mathcal{U}(0,T)}} \prod_{\lambda=1}^{\Lambda'} \prod_{x=1}^{M_\lambda} \prod_{i=1}^{p} \mathrm{Tr}\left(e^{-\mathbf{i}\boldsymbol{H}^\lambda t''}|x\rangle\langle x|e^{\mathbf{i}\boldsymbol{H}^\lambda t''}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_i}e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}\right)^{k_{\lambda,x,i,1}} \\
&\qquad \mathrm{Tr}\left(e^{-\mathbf{i}\boldsymbol{H}^\lambda t''}|x\rangle\langle x|e^{\mathbf{i}\boldsymbol{H}^\lambda t''}e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_i}\right)^{k_{\lambda,x,i,2}} = \\
&= O(1/N^*_{\min}) + \mathop{\mathbb{E}}_{\substack{\boldsymbol{g}_1,\ldots,\boldsymbol{g}_p,\\ \boldsymbol{g}',\boldsymbol{g}''\sim\mathrm{Haar}}} \prod_{\lambda=1}^{\Lambda'} \prod_{x=1}^{M_\lambda} \prod_{i=1}^{p} \mathrm{Tr}\left(\boldsymbol{g}''^\lambda|x\rangle\langle x|\boldsymbol{g}''^{\lambda\dagger}\boldsymbol{g}_i^\lambda\boldsymbol{A}^\lambda\boldsymbol{g}_i^{\lambda\dagger}\boldsymbol{g}'^\lambda\boldsymbol{O}_x^\lambda\boldsymbol{g}'^{\lambda\dagger}\right)^{k_{\lambda,x,i,1}} \\
&\qquad \mathrm{Tr}\left(\boldsymbol{g}''^\lambda|x\rangle\langle x|\boldsymbol{g}''^{\lambda\dagger}\boldsymbol{g}'^\lambda\boldsymbol{O}_x^\lambda\boldsymbol{g}'^{\lambda\dagger}\boldsymbol{g}_i^\lambda\boldsymbol{A}^\lambda\boldsymbol{g}_i^{\lambda\dagger}\right)^{k_{\lambda,x,i,2}},
\end{aligned} \tag{38}$$

where

$$N^*_{\min} = \min_\lambda N^*_\lambda. \tag{39}$$

Here, the Haar-random distribution is over the unitaries with the block-diagonal structure of the Krylov subspaces: equivalently, each $\boldsymbol{g}_i^\lambda, \boldsymbol{g}'^\lambda, \boldsymbol{g}''^\lambda$ *is chosen to be Haar-random in each of the sectors.*

Proof. We can justify the above claim using generally accepted physical assumptions as follows. Arguing similarly to Fava et al. (2025, Appendix C), if we assume Hypothesis 1, we can say that for a sufficiently large time T,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{t_1,\ldots,t_p,\\ t',t''\sim\mathcal{U}(0,T)}} \prod_{\lambda=1}^{\Lambda'} \prod_{x=1}^{M_\lambda} \prod_{i=1}^{p} \mathrm{Tr}\left(e^{-\mathbf{i}\boldsymbol{H}^\lambda t''}|x\rangle\langle x|e^{\mathbf{i}\boldsymbol{H}^\lambda t''}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_i}e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}\right)^{k_{\lambda,x,i,1}} \\
&\qquad \mathrm{Tr}\left(e^{-\mathbf{i}\boldsymbol{H}^\lambda t''}|x\rangle\langle x|e^{\mathbf{i}\boldsymbol{H}^\lambda t''}e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_i}\right)^{k_{\lambda,x,i,2}} = \\
&= \prod_{\lambda=1}^{\Lambda'} \Bigg(O(1/N^*_\lambda) + \\
&\quad + \prod_{x=1}^{M_\lambda} \prod_{i=1}^{p} \mathop{\mathbb{E}}_{\substack{t_1,\ldots,t_p,\\ t',t''\sim\mathcal{U}(0,T)}} \left[\mathrm{Tr}\left(e^{-\mathbf{i}\boldsymbol{H}^\lambda t''}|x\rangle\langle x|e^{\mathbf{i}\boldsymbol{H}^\lambda t''}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_i}e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}\right)\right]^{k_{\lambda,x,i,1}} \\
&\qquad \mathop{\mathbb{E}}_{\substack{t_1,\ldots,t_p,\\ t',t''\sim\mathcal{U}(0,T)}} \left[\mathrm{Tr}\left(e^{-\mathbf{i}\boldsymbol{H}^\lambda t''}|x\rangle\langle x|e^{\mathbf{i}\boldsymbol{H}^\lambda t''}e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_i}\right)\right]^{k_{\lambda,x,i,2}} \Bigg) = \\
&= O(1/N^*_{\min}) + \prod_{\lambda=1}^{\Lambda'} \prod_{x=1}^{M_\lambda} \prod_{i=1}^{p} \\
&\qquad \mathop{\mathbb{E}}_{\substack{t_1,\ldots,t_p,\\ t',t''\sim\mathcal{U}(0,T)}} \left[\mathrm{Tr}\left(e^{-\mathbf{i}\boldsymbol{H}^\lambda t''}|x\rangle\langle x|e^{\mathbf{i}\boldsymbol{H}^\lambda t''}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_i}e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}\right)\right]^{k_{\lambda,x,i,1}} \\
&\qquad \mathop{\mathbb{E}}_{\substack{t_1,\ldots,t_p,\\ t',t''\sim\mathcal{U}(0,T)}} \left[\mathrm{Tr}\left(e^{-\mathbf{i}\boldsymbol{H}^\lambda t''}|x\rangle\langle x|e^{\mathbf{i}\boldsymbol{H}^\lambda t''}e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_i}\right)\right]^{k_{\lambda,x,i,2}}.
\end{aligned} \tag{40}$$

Now, using the results of Fava et al. (2025), we can write

$$\begin{aligned}&\mathop{\mathbb{E}}_{\substack{t_1,\ldots,t_p,\\ t',t''\sim\mathcal{U}(0,T)}} \operatorname{Tr}\left(e^{-\mathbf{i}\boldsymbol{H}^\lambda t''}|x\rangle\langle x|e^{\mathbf{i}\boldsymbol{H}^\lambda t''}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_i}e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}\right) = \\ &= \frac{1}{(N_\lambda^*)^2}\operatorname{Tr}(\boldsymbol{A}^\lambda)\operatorname{Tr}(\boldsymbol{O}_x^\lambda).\end{aligned} \tag{41}$$

Now, when we replace the time evolution with the Haar-random unitaries, by free independence in the asymptotic limit (Voiculescu, 1991) we will get the same result as above up to an $O(1/N_{\min}^*)$ error. □

Lemma 2. *We can approximate, at any value of $\boldsymbol{\theta}$,*

$$\partial_i \ell_x^\lambda(\boldsymbol{\theta}) \rightsquigarrow \frac{\mathbf{i}}{M}\langle x|\,\boldsymbol{g}''^{\lambda\dagger}\left[\boldsymbol{g}_i^\lambda \boldsymbol{A}^\lambda \boldsymbol{g}_i^{\lambda\dagger},\ \boldsymbol{g}'^\lambda \boldsymbol{O}_x^\lambda \boldsymbol{g}'^{\lambda\dagger}\right]\boldsymbol{g}''^\lambda\,|x\rangle\,, \tag{24}$$

where $\boldsymbol{g}_1,\ldots,\boldsymbol{g}_p,\boldsymbol{g}',\boldsymbol{g}'' \sim \mathrm{Haar}$ up to an $o(1)$ error in the Lévy-Prokhorov metric (Definition 2), if we choose a sufficiently large T and we assume that $N_{\min} = \min_\lambda N_\lambda = \omega(1)$. Furthermore, this approximation is valid for the joint distribution of any two *gradient contributions $\partial_i \ell_x^\lambda(\boldsymbol{\theta})$ and $\partial_j \ell_y^{\lambda'}(\boldsymbol{\theta})$: their joint distribution can be approximated up to an $o(1)$ Lévy-Prokhorov error by replacing the Hamiltonian time evolution with Haar-random matrices.*

Proof. Consider the joint distribution $\mathfrak{p}$ of the terms in the expression for the $\partial_i \ell_x^\lambda$ for $\boldsymbol{\theta} = 0$, as in Lemma 1, considering the commutator expression as two separate terms. The terms are of the form

$$\langle x|\,e^{\mathbf{i}\boldsymbol{H}^\lambda t''}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_i}e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}e^{-\mathbf{i}\boldsymbol{H}^\lambda t''}\,|x\rangle \tag{42}$$

and

$$\langle x|\,e^{\mathbf{i}\boldsymbol{H}^\lambda t''}e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_i}e^{-\mathbf{i}\boldsymbol{H}^\lambda t''}\,|x\rangle\,. \tag{43}$$

and there are $d = 2$ of these terms, so the distribution is in $\mathbb{R}^2$. We can also consider the joint distribution of the terms contributing to both $\partial_i \ell_x^\lambda(\boldsymbol{\theta})$ and $\partial_j \ell_y^{\lambda'}(\boldsymbol{\theta})$, in which case the distribution will be in $\mathbb{R}^4$. Similarly, let $\mathfrak{q}$ be the joint distribution of the same terms, but with the time-evolved Hamiltonian replaced with the Haar-random matrices $\boldsymbol{g}_i,\boldsymbol{g}_j,\boldsymbol{g}',\boldsymbol{g}''$. Now, from Lemma 5, we know that we can pick $T = \mathrm{poly}(L,k)$ such that mixed moments of order k of the distributions $\mathfrak{p}$ and $\mathfrak{q}$ differ by an additive error of $\epsilon = 1/N_{\min}^*$. The random variables corresponding to the terms are bounded, and thus they are subgaussian (Vershynin, 2018) and thus their moments of order k are upper bounded by $\left(C\sqrt{k}\right)^k$ for some constant C independent of k. Now, applying Lemma 9 gives us that the distribution $\mathfrak{p}$ converges to $\mathfrak{q}$ up to an error of

$$O\left(\frac{\log\log N_{\min}}{\log N_{\min}} + \frac{\log k}{\sqrt{k}}\right) = o(1) \tag{44}$$

in the Lévy-Prokhorov metric (note that the μ term is subleading). Now, consider taking a general value of $\boldsymbol{\theta}$. Then, each of the terms becomes conjugated by some unitary matrix (see Lemma 1). By the unitary invariance property of the Haar ensemble, the mixed moments of the distribution $\mathfrak{q}$ remain the same, which means that it approximates the joint distribution of the derivative terms at any value of $\boldsymbol{\theta}$. since the Lévy-Prokhorov distance is preserved under the linear transformation of summing the terms to form the commutator expressions and is only reduced when dividing by the factor of M, the proof is complete. □

Lemma 3. *We can approximate, at any value of $\boldsymbol{\theta}$,*

$$\partial_i \ell_x^\lambda(\boldsymbol{\theta}) \rightsquigarrow \frac{\mathbf{i}}{MN_\lambda^{*2}}\langle x|\,\boldsymbol{g}''^{\lambda\dagger}\left[\widetilde{\boldsymbol{g}}_i^\lambda \boldsymbol{A}^\lambda \widetilde{\boldsymbol{g}}_i^{\lambda\dagger},\ \widetilde{\boldsymbol{g}}'^\lambda \boldsymbol{O}_x^\lambda \widetilde{\boldsymbol{g}}'^{\lambda\dagger}\right]\boldsymbol{g}''^\lambda\,|x\rangle\,, \tag{25}$$

where $\widetilde{\boldsymbol{g}}_1^\lambda,\ldots,\widetilde{\boldsymbol{g}}_p^\lambda,\widetilde{\boldsymbol{g}}'^\lambda$ have i.i.d. standard Gaussian entries, up to an $o(1)$ error in the Lévy-Prokhorov metric, and the same approximation is valid for the joint distribution of any two loss contributions $\partial_i \ell_x^\lambda(\boldsymbol{\theta})$ and $\partial_j \ell_y^{\lambda'}(\boldsymbol{\theta})$.

Proof. We first perform the reduction to Haar-random matrices as in Lemma 2. Now, we can proceed similarly to Lemma 27 of Anschuetz (2025), based on the results of Jiang (2010). From this, we incur an additional error of

$$O\left(\sqrt{\frac{\log N^*}{N^*}}\right) = o(1) \tag{45}$$

in the Lévy-Prokhorov metric. □

Lemma 4. *Replacing $\boldsymbol{A}$ with $\widetilde{\boldsymbol{A}}$, we can write*

$$\partial_i \ell_x^\lambda(\boldsymbol{\theta}) \rightsquigarrow \hat{\ell}_{x;i}^\lambda := \frac{\mathbf{i}}{MN_\lambda^{*2}} \langle x| \boldsymbol{g}''^{\lambda\dagger} \left[\widetilde{\boldsymbol{g}}_i^\lambda \widetilde{\boldsymbol{A}}^\lambda \widetilde{\boldsymbol{g}}_i^{\lambda\dagger},\ \widetilde{\boldsymbol{g}}'^\lambda \boldsymbol{O}_x^\lambda \widetilde{\boldsymbol{g}}'^{\lambda\dagger}\right] \boldsymbol{g}''^\lambda |x\rangle\,, \tag{31}$$

up to an $o(1)$ error in the Lévy-Prokhorov metric, assuming that $N_{\min} = \min_\lambda N_\lambda = \omega(1)$. The same is true for joint distributions of two derivative contributions as before.

Proof. Consider each of the terms contributing to the gradient as previously. Replace each $\boldsymbol{A}^\lambda$ term with $\boldsymbol{A}^\lambda - \widetilde{\boldsymbol{A}}^\lambda$, so we have terms of the form

$$\begin{aligned} &L_{x,i}^\lambda := \\ &= \frac{1}{MN_\lambda^{*2}} \left(\langle x| \widetilde{\boldsymbol{g}}_i^\lambda \left(\boldsymbol{A}^\lambda - \widetilde{\boldsymbol{A}}^\lambda\right) \widetilde{\boldsymbol{g}}_i^{\lambda\dagger} \widetilde{\boldsymbol{g}}'^\lambda \boldsymbol{O}_x^\lambda \widetilde{\boldsymbol{g}}'^{\lambda\dagger} |x\rangle - \langle x| \widetilde{\boldsymbol{g}}'^\lambda \boldsymbol{O}_x^\lambda \widetilde{\boldsymbol{g}}'^{\lambda\dagger} \widetilde{\boldsymbol{g}}_i^\lambda \left(\boldsymbol{A}^\lambda - \widetilde{\boldsymbol{A}}^\lambda\right) \widetilde{\boldsymbol{g}}_i^{\lambda\dagger} |x\rangle\right) = \\ &= \frac{1}{MN_\lambda^{*2}} \operatorname{Tr}\left(\boldsymbol{M}_x^\lambda \widetilde{\boldsymbol{g}}_i^\lambda \left(\boldsymbol{A}^\lambda - \widetilde{\boldsymbol{A}}^\lambda\right) \widetilde{\boldsymbol{g}}_i^{\lambda\dagger}\right), \end{aligned} \tag{46}$$

and

$$\boldsymbol{M}_x^\lambda = \widetilde{\boldsymbol{g}}'^\lambda \boldsymbol{O}_x^\lambda \widetilde{\boldsymbol{g}}'^{\lambda\dagger} |x\rangle\langle x| - |x\rangle\langle x| \widetilde{\boldsymbol{g}}'^\lambda \boldsymbol{O}_x^\lambda \widetilde{\boldsymbol{g}}'^{\lambda\dagger} \tag{47}$$

is an anti-Hermitian traceless operator of rank 2 and operator norm at most 2. We can therefore write it in the form

$$\boldsymbol{M}_x^\lambda = \mathbf{i}\mu_x^\lambda |\psi_x^\lambda\rangle\langle\psi_x^\lambda| - \mathbf{i}\mu_x^\lambda |\varphi_x^\lambda\rangle\langle\varphi_x^\lambda| \tag{48}$$

with $\left|\mu_x^\lambda\right|$ bounded by 2. Now, we can also write

$$\widetilde{\boldsymbol{g}}_i^\lambda \left(\boldsymbol{A}^\lambda - \widetilde{\boldsymbol{A}}^\lambda\right) \widetilde{\boldsymbol{g}}_i^{\lambda\dagger} = \sum_{k=1}^{N_\lambda^*} (a_k^\lambda - \widetilde{a}_k^\lambda) |\chi_{i,k}^\lambda\rangle\langle\chi_{i,k}^\lambda|, \tag{49}$$

where the $a_k^\lambda, \widetilde{a}_k^\lambda$ are the eigenvalues of $\boldsymbol{A}^\lambda, \widetilde{\boldsymbol{A}}^\lambda$ in non-increasing order and $\left|\chi_{i,k}^\lambda\right\rangle$ is the kth column of $\widetilde{\boldsymbol{g}}_i^\lambda$ in the basis in which $\boldsymbol{A}^\lambda$ and $\widetilde{\boldsymbol{A}}^\lambda$ are both diagonal. So now, we have that

$$L_{x,i}^\lambda = \frac{\mathbf{i}\mu_x^\lambda}{MN_\lambda^{*2}} \sum_{k=1}^{N_\lambda^*} (a_k^\lambda - \widetilde{a}_k^\lambda) \left(\left|\left\langle\chi_{i,k}^\lambda\middle|\psi_x^\lambda\right\rangle\right|^2 - \left|\left\langle\chi_{i,k}^\lambda\middle|\varphi_x^\lambda\right\rangle\right|^2\right). \tag{50}$$

Now, for any normalized vector $|\psi\rangle$, observe that $\left|\left\langle\chi_{i,k}^\lambda\middle|\psi\right\rangle\right|^2$ is a sum of squared Gaussian random variables, so $\left|\left\langle\chi_{i,k}^\lambda\middle|\psi_x^\lambda\right\rangle\right|^2 - \left|\left\langle\chi_{i,k}^\lambda\middle|\varphi_x^\lambda\right\rangle\right|^2$ is subexponential. Then, by the version of Bernstein's inequality found in Vershynin (2018), since $\left|\mu_x^\lambda\right| \leq 2$, we have that

$$\begin{aligned} &\Pr\left[\left|L_{x,i}^\lambda\right| \geq \epsilon\right] \leq \\ &\Pr\left[\left|\frac{\left|\mu_x^\lambda\right|}{MN_\lambda^{*2}} \sum_{k=1}^{N_\lambda^*} (a_k^\lambda - \widetilde{a}_k^\lambda) \left(\left|\left\langle\chi_{i,k}^\lambda\middle|\psi_x^\lambda\right\rangle\right|^2 - \left|\left\langle\chi_{i,k}^\lambda\middle|\varphi_x^\lambda\right\rangle\right|^2\right)\right| \geq \epsilon\right] \leq \\ &\leq 2\exp\left[-c\min\left(\frac{\epsilon}{\frac{2}{MN_\lambda^{*2}} \max_k \left|a_k^\lambda - \widetilde{a}_k^\lambda\right|}, \frac{\epsilon^2}{\frac{4}{M^2 N_\lambda^{*4}} \sum_k \left(a_k^\lambda - \widetilde{a}_k^\lambda\right)^2}\right)\right] \leq \\ &\leq 2\exp\left[-cMN_\lambda^{*2}\min\left(\frac{\epsilon}{2\max_k \left|a_k^\lambda - \widetilde{a}_k^\lambda\right|}, \frac{\epsilon^2}{\frac{4}{MN_\lambda^{*2}} \sum_k \left(a_k^\lambda - \widetilde{a}_k^\lambda\right)^2}\right)\right] \end{aligned} \tag{51}$$

for some constant $c > 0$. Now, observe that

$$\frac{1}{MN_\lambda^{*2}} \sum_i \left(a_k^\lambda - \widetilde{a}_k^\lambda\right)^2 = \frac{1}{MN_\lambda^{*2}} \operatorname{Tr}\left[\left(\boldsymbol{A}^\lambda - \widetilde{\boldsymbol{A}}^\lambda\right)^2\right] =$$
$$= \frac{1}{MN_\lambda^{*2}} \operatorname{Tr}\left[\left(\boldsymbol{A}^\lambda\right)^2 + \left(\widetilde{\boldsymbol{A}}^\lambda\right)^2 - \boldsymbol{A}^\lambda \widetilde{\boldsymbol{A}}^\lambda - \widetilde{\boldsymbol{A}}^\lambda \boldsymbol{A}^\lambda\right] = \frac{2}{MN_\lambda^{*2}} \left(\operatorname{Tr}((\boldsymbol{A}^\lambda)^2) - \operatorname{Tr}\left(\boldsymbol{A}^\lambda \widetilde{\boldsymbol{A}}^\lambda\right)\right). \tag{52}$$

Now, since

$$\operatorname{Tr}\left(\boldsymbol{A}^\lambda \widetilde{\boldsymbol{A}}^\lambda\right) = \frac{\operatorname{Tr}(\boldsymbol{A}^\lambda)}{r_A^\lambda} \sum_{i=1}^{r_A^\lambda} a_k^\lambda \geq \frac{\operatorname{Tr}((\boldsymbol{A}^\lambda)^2)}{\operatorname{Tr}(\boldsymbol{A}^\lambda)} \left(\operatorname{Tr}(\boldsymbol{A}^\lambda) - N_\lambda^* a_{r_A^\lambda+1}^\lambda\right), \tag{53}$$

we must have that

$$\frac{4}{MN_\lambda^{*2}} \sum_i \left(a_k^\lambda - \widetilde{a}_k^\lambda\right)^2 \leq \frac{8}{MN_\lambda^{*2}} \frac{\operatorname{Tr}((\boldsymbol{A}^\lambda)^2)}{\operatorname{Tr}(\boldsymbol{A}^\lambda)} N_\lambda^* a_{r_A^\lambda+1}^\lambda = o(1), \tag{54}$$

which holds since $\boldsymbol{A}$ is a local operator and its eigenvalues do not depend on the system size. We also have that

$$2 \max_k \left|a_k^\lambda - \widetilde{a}_k^\lambda\right| = O\left(a_1^\lambda - \frac{\operatorname{Tr}((\boldsymbol{A}^\lambda)^2)}{\operatorname{Tr}(\boldsymbol{A}^\lambda)} a_{r_A^\lambda+1}\right) = O(1). \tag{55}$$

Now, we can just apply the union bound on all terms (of which there at most 2 since we are considering the joint distribution of at most 2 contributions) to say that

$$\Pr\left[\left|L_{x,i}^\lambda\right| \geq \epsilon \vee \left|L_{x',i'}^{\lambda'}\right| \geq \epsilon\right] \leq 4 \exp\left[-cMN_{\min}^{*2} \min(\epsilon, \epsilon^2)\right]. \tag{56}$$

Now, to solve for the Ky Fan metric, which upper-bounds the Lévy-Prokhorov metric (Strassen, 1965), we need to find an upper bound on

$$\inf\left\{\epsilon > 0 : \Pr\left[\left|L_{x,i}^\lambda\right| \geq \epsilon \vee \left|L_{x',i'}^{\lambda'}\right| \geq \epsilon\right] \leq \epsilon\right\}. \tag{57}$$

For the case where $\epsilon < \epsilon^2$, equality occurs when

$$4 \exp\left(-cMN_{\min}^{*2}\epsilon\right) = \epsilon \tag{58}$$
$$\epsilon \exp\left(cMN_{\min}^{*2}\epsilon\right) = 4 \tag{59}$$
$$cMN_{\min}^{*2}\epsilon \exp\left(cMN_{\min}^{*2}\epsilon\right) = 4cMN_{\min}^{*2} \tag{60}$$
$$cMN_{\min}^{*2}\epsilon = W(4cMN_{\min}^{*2}), \tag{61}$$

where W is the Lambert W function. Since it is upper-bounded by the logarithm, an upper bound for the Ky Fan metric is

$$cMN_{\min}^{*2}\epsilon \leq \log(4cMN_{\min}^{*2})$$
$$\epsilon \leq \frac{\log(4cMN_{\min}^{*2})}{cMN_{\min}^{*2}} = O\left(\frac{\log(MN_{\min})}{MN_{\min}^2}\right). \tag{62}$$

When $\epsilon^2 < \epsilon$ (when $\epsilon < 1$), equality occurs when

$$4 \exp\left(-cMN_{\min}^{*2}\epsilon^2\right) = \epsilon \tag{63}$$
$$cMN_{\min}^{*2}\epsilon^2 = W(4cMN_{\min}^{*2}\epsilon), \tag{64}$$

so the upper bound for the Ky Fan metric (and thus the Lévy-Prokhorov metric) is

$$cMN_{\min}^{*2}\epsilon^2 \leq \log(4cMN_{\min}^{*2}\epsilon) \leq \log(4cMN_{\min}^{*2}) \tag{65}$$
$$\epsilon = O\left(\frac{\sqrt{\log(MN_{\min})}}{MN_{\min}^2}\right) = o(1). \tag{66}$$

Since $\epsilon < 1$ for large enough system sizes, this is the correct asymptotic bound. □

Theorem 1. *At any value of* $\boldsymbol{\theta} \in [0, 2\pi)^p$*,*

$$\mathbb{E}\left[\left\|\widehat{\nabla\ell}\right\|^2\right] = \sum_{\lambda=1}^{\Lambda'} \frac{2M_\lambda(N_a-1)\,\mathrm{Tr}((\boldsymbol{A}^\lambda)^2)p}{M^2 N_a^4 N_\lambda^2} = \Omega\left(\frac{1}{\mathrm{poly}(L)}\right), \tag{32}$$

where

$$\left(\widehat{\nabla\ell}\right)_i := \sum_{\lambda=1}^{\Lambda'}\sum_{x=1}^{M_\lambda} \hat{\ell}^\lambda_{x;i}. \tag{33}$$

Proof. We know by the reductions above that the contribution to the derivative with respect to θ_i from entry x in sector λ can be approximated as

$$\hat{\ell}^\lambda_{x;i} = \frac{\mathbf{i}}{MN_\lambda^{*2}} \langle x|\, \widetilde{\boldsymbol{g}}''^\lambda \left[\widetilde{\boldsymbol{g}}_i^\lambda \widetilde{\boldsymbol{A}}^\lambda \widetilde{\boldsymbol{g}}_i^{\lambda\dagger},\, \widetilde{\boldsymbol{g}}'^\lambda \boldsymbol{O}_x^\lambda \widetilde{\boldsymbol{g}}'^{\lambda\dagger}\right] \widetilde{\boldsymbol{g}}''^\lambda |x\rangle\,. \tag{67}$$

By unitary invariance, we can replace $\widetilde{\boldsymbol{g}}_i^\lambda$ with $\widetilde{\boldsymbol{g}}'^\lambda \widetilde{\boldsymbol{g}}_i^\lambda$ while maintaining the same probability distribution (note that technically, we are applying this *before* the reduction in Lemma 3 to the Haar-random matrices). But now, since

$$\begin{aligned}
&\left[\widetilde{\boldsymbol{g}}'^\lambda \widetilde{\boldsymbol{g}}_i^\lambda \widetilde{\boldsymbol{A}}^\lambda \widetilde{\boldsymbol{g}}_i^{\lambda\dagger} \widetilde{\boldsymbol{g}}'^{\lambda\dagger},\, \widetilde{\boldsymbol{g}}'^\lambda \boldsymbol{O}_x^\lambda \widetilde{\boldsymbol{g}}'^{\lambda\dagger}\right] = \\
&= \widetilde{\boldsymbol{g}}'^\lambda \widetilde{\boldsymbol{g}}_i^\lambda \widetilde{\boldsymbol{A}}^\lambda \widetilde{\boldsymbol{g}}_i^{\lambda\dagger} \boldsymbol{O}_x^\lambda \widetilde{\boldsymbol{g}}'^{\lambda\dagger} - \widetilde{\boldsymbol{g}}'^\lambda \boldsymbol{O}_x^\lambda \widetilde{\boldsymbol{g}}_i^\lambda \widetilde{\boldsymbol{A}}^\lambda \widetilde{\boldsymbol{g}}_i^{\lambda\dagger} \widetilde{\boldsymbol{g}}'^{\lambda\dagger} = \\
&= \widetilde{\boldsymbol{g}}'^\lambda \left[\widetilde{\boldsymbol{g}}_i^\lambda \widetilde{\boldsymbol{A}}^\lambda \widetilde{\boldsymbol{g}}_i^{\lambda\dagger},\, \boldsymbol{O}_x^\lambda\right] \widetilde{\boldsymbol{g}}'^{\lambda\dagger},
\end{aligned} \tag{68}$$

we can absorb $\boldsymbol{g}''^\lambda$ into $\boldsymbol{g}'^\lambda$ and write

$$\begin{aligned}
\hat{\ell}^\lambda_{x;i} &= \frac{\mathbf{i}}{MN_\lambda^{*2}} \langle x|\, \widetilde{\boldsymbol{g}}'^\lambda \left[\widetilde{\boldsymbol{g}}_i^\lambda \widetilde{\boldsymbol{A}}^\lambda \widetilde{\boldsymbol{g}}_i^{\lambda\dagger},\, \boldsymbol{O}_x^\lambda\right] \widetilde{\boldsymbol{g}}'^{\lambda\dagger} |x\rangle = \\
&= \frac{\mathbf{i}}{MN_\lambda^{*2}} \frac{\mathrm{Tr}((\boldsymbol{A}^\lambda)^2)}{\mathrm{Tr}(\boldsymbol{A}^\lambda)} \sum_{\mu=1}^{r_A^\lambda} \langle x|\, \widetilde{\boldsymbol{g}}'^\lambda \left[\widetilde{\boldsymbol{g}}_i^\lambda |\mu\rangle\langle\mu| \widetilde{\boldsymbol{g}}_i^{\lambda\dagger},\, \boldsymbol{O}_x^\lambda\right] \widetilde{\boldsymbol{g}}'^{\lambda\dagger} |x\rangle
\end{aligned} \tag{69}$$

Now, we can take a matrix $\boldsymbol{X}^\lambda$ of dimension $N_\lambda^* \times (M_\lambda + pr_A^\lambda)$ with i.i.d. Gaussian entries, such that

$$\boldsymbol{X}^\lambda |x\rangle = \widetilde{\boldsymbol{g}}'^{\lambda\dagger} |x\rangle \tag{70}$$

and

$$\boldsymbol{X}^\lambda |i,\mu\rangle = \widetilde{\boldsymbol{g}}_i^\lambda |\mu\rangle\,. \tag{71}$$

Note that we are using $\{x\}$ and $\{(i,\mu)\}$ to label the $M_\lambda + pr_A^\lambda$ columns of $\boldsymbol{X}^\lambda$.

Now, we can write

$$\begin{aligned}
\hat{\ell}^{\lambda}_{x;i} &= \frac{\mathbf{i}\,\mathrm{Tr}((\boldsymbol{A}^{\lambda})^2)}{MN_{\lambda}^{*2}\,\mathrm{Tr}(\boldsymbol{A}^{\lambda})}\sum_{\mu=1}^{r_A^{\lambda}}\langle x|\,\boldsymbol{X}^{\lambda\dagger}\left[\boldsymbol{X}^{\lambda}|i,\mu\rangle\langle i,\mu|\boldsymbol{X}^{\lambda\dagger},\boldsymbol{O}_x^{\lambda}\right]\boldsymbol{X}^{\lambda}\,|x\rangle = \\
&= \frac{\mathbf{i}\,\mathrm{Tr}((\boldsymbol{A}^{\lambda})^2)}{MN_{\lambda}^{*2}\,\mathrm{Tr}(\boldsymbol{A}^{\lambda})}\sum_{\mu=1}^{r_A^{\lambda}}\Big(\langle x|^{\lambda}\,\boldsymbol{X}^{\lambda\dagger}\boldsymbol{X}^{\lambda}\,|i,\mu\rangle\,\langle i,\mu|\,\boldsymbol{X}^{\lambda\dagger}\boldsymbol{O}_x^{\lambda}\boldsymbol{X}^{\lambda}\,|x\rangle - \\
&\qquad\qquad - \langle x|\,\boldsymbol{X}^{\lambda\dagger}\boldsymbol{O}_x^{\lambda}\boldsymbol{X}^{\lambda}\,|i,\mu\rangle\,\langle i,\mu|\,\boldsymbol{X}^{\lambda\dagger}\boldsymbol{X}^{\lambda}\,|x\rangle\Big) = \\
&= -\frac{\mathbf{i}\,\mathrm{Tr}((\boldsymbol{A}^{\lambda})^2)}{MN_{\lambda}^{*2}\,\mathrm{Tr}(\boldsymbol{A}^{\lambda})}\sum_{\mu=1}^{r_A^{\lambda}}\Big(\langle x|\,\boldsymbol{W}^{\lambda}\,|i,\mu\rangle\,\langle i,\mu|\,\boldsymbol{W}^{\lambda,x}\,|x\rangle - \langle x|\,\boldsymbol{W}^{\lambda,x}\,|i,\mu\rangle\,\langle i,\mu|\,\boldsymbol{W}^{\lambda}\,|x\rangle\Big) = \\
&= -\frac{\mathbf{i}\,\mathrm{Tr}((\boldsymbol{A}^{\lambda})^2)}{MN_{\lambda}^{*2}\,\mathrm{Tr}(\boldsymbol{A}^{\lambda})}\sum_{\mu=1}^{r_A^{\lambda}}\sum_{y=1}^{M_\lambda+1}\Big(\langle x|\,\boldsymbol{W}^{\lambda,y}\,|i,\mu\rangle\,\langle i,\mu|\,\boldsymbol{W}^{\lambda,x}\,|x\rangle - \\
&\qquad\qquad - \langle x|\,\boldsymbol{W}^{\lambda,x}\,|i,\mu\rangle\,\langle i,\mu|\,\boldsymbol{W}^{\lambda,y}\,|x\rangle\Big) = \\
&= -\frac{\mathbf{i}\,\mathrm{Tr}((\boldsymbol{A}^{\lambda})^2)}{MN_{\lambda}^{*2}\,\mathrm{Tr}(\boldsymbol{A}^{\lambda})}\sum_{\mu=1}^{r_A^{\lambda}}\sum_{y=1}^{M_\lambda+1}(-2i)\,\mathrm{Im}\left(\langle x|\,\boldsymbol{W}^{\lambda,x}\,|i,\mu\rangle\,\langle i,\mu|\,\boldsymbol{W}^{\lambda,y}\,|x\rangle\right) = \\
&= -\frac{2\,\mathrm{Tr}((\boldsymbol{A}^{\lambda})^2)}{MN_{\lambda}^{*2}\,\mathrm{Tr}(\boldsymbol{A}^{\lambda})}\sum_{\mu=1}^{r_A^{\lambda}}\sum_{y=1}^{M_\lambda+1}\mathrm{Im}\left(\langle x|\,\boldsymbol{W}^{\lambda,x}\,|i,\mu\rangle\,\langle i,\mu|\,\boldsymbol{W}^{\lambda,y}\,|x\rangle\right).
\end{aligned} \tag{72}$$

Here, we define $\boldsymbol{W}^{\lambda,x} \coloneqq -\boldsymbol{X}^{\lambda\dagger}\boldsymbol{O}_x^{\lambda}\boldsymbol{X}^{\lambda}$ for $1 \le x \le M_\lambda$ (recall that $\boldsymbol{O}_x^{\lambda}$ is negative semidefinite), and we define

$$\boldsymbol{W}^{\lambda,M_\lambda+1} \coloneqq \boldsymbol{W}^{\lambda} - \sum_{x=1}^{M_\lambda}\boldsymbol{W}^{\lambda,x} \tag{73}$$

and $\boldsymbol{W}^{\lambda} \coloneqq \boldsymbol{X}^{\lambda\dagger}\boldsymbol{X}^{\lambda}$.

Now, we can write this in terms of the entries of $\boldsymbol{X}^{\lambda,x}$ which we take to be an $r_x \times (M_\lambda + pr_A^{\lambda})$ submatrix of $\boldsymbol{X}$. Here for $1 \le x \le M_\lambda$ we have $r_x = \mathrm{rank}(\boldsymbol{O}_x^{\lambda}) = N_\lambda$, and $r_{M_\lambda+1} = N_\lambda^* - M_\lambda N_\lambda$.

$$\hat{\ell}^{\lambda}_{x;i} = -\frac{2\,\mathrm{Tr}((\boldsymbol{A}^{\lambda})^2)}{MN_{\lambda}^{*2}\,\mathrm{Tr}(\boldsymbol{A}^{\lambda})}\sum_{\mu=1}^{r_A^{\lambda}}\sum_{y\neq x}\mathrm{Im}\left(\sum_{j=1}^{r_x}\left(X^{\lambda,x}_{j,x}\right)^{*}X^{\lambda,x}_{j,(i,\mu)}\sum_{k=1}^{r_y}\left(X^{\lambda,y}_{k,(i,\mu)}\right)^{*}X^{\lambda,y}_{k,x}\right). \tag{74}$$

Now, observe that for $\lambda \neq \lambda'$ or $x \neq x'$, we must have by a symmetry argument that

$$\mathbb{E}\left[(\hat{\ell}^{\lambda}_{x;i})(\hat{\ell}^{\lambda'}_{x';i})\right] = 0, \tag{75}$$

since each term in the expansion of this will have a factor that is a Gaussian random variable independent from the rest. Also recall that this is a valid approximation for the corresponding term in the actual loss since we ensured in our reductions that the joint distribution of two contributions such as this converges in distribution to our approximation.

Now,

$$\begin{aligned}
&\mathbb{E}\left[(\hat{\ell}_{x;i})^2\right] = \\
&= \left(\frac{2\operatorname{Tr}((\boldsymbol{A}^\lambda)^2)}{MN_\lambda^{*2}\operatorname{Tr}(\boldsymbol{A}^\lambda)}\right)^2 \sum_{\mu=1}^{r_A^\lambda}\sum_{\substack{y=1\\y\neq x}}^{M_\lambda+1} \mathbb{E}\left[\operatorname{Im}\left(\sum_{j=1}^{r_x}\left(X_{j,x}^{\lambda,x}\right)^* X_{j,(i,\mu)}^{\lambda,x}\sum_{k=1}^{r_y}\left(X_{k,(i,\mu)}^{\lambda,y}\right)^* X_{k,x}^{\lambda,y}\right)^2\right] = \\
&= \frac{1}{2}\left(\frac{2\operatorname{Tr}((\boldsymbol{A}^\lambda)^2)}{MN_\lambda^{*2}\operatorname{Tr}(\boldsymbol{A}^\lambda)}\right)^2 \sum_{\mu=1}^{r_A^\lambda}\sum_{\substack{y=1\\y\neq x}}^{M_\lambda+1} \mathbb{E}\left[\left|\sum_{j=1}^{r_x}\left(X_{j,x}^{\lambda,x}\right)^* X_{j,(i,\mu)}^{\lambda,x}\sum_{k=1}^{r_y}\left(X_{k,(i,\mu)}^{\lambda,y}\right)^* X_{k,x}^{\lambda,y}\right|^2\right].
\end{aligned} \tag{76}$$

Note that all cross terms in the above calculation get eliminated due to symmetry in the Gaussian distributions of the unmatched matrix elements.

Now, observe that

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{j=1}^{r_x}\left(X_{j,x}^{\lambda,x}\right)^* X_{j,(i,\mu)}^{\lambda,x}\sum_{k=1}^{r_y}\left(X_{k,(i,\mu)}^{\lambda,y}\right)^* X_{k,x}^{\lambda,y}\right|^2\right] = \\
&= \mathbb{E}\left[\left|\sum_{j=1}^{r_x}\left(X_{j,x}^{\lambda,x}\right)^* X_{j,(i,\mu)}^{\lambda,x}\right|^2\right]\mathbb{E}\left[\left|\sum_{k=1}^{r_y}\left(X_{k,(i,\mu)}^{\lambda,y}\right)^* X_{k,x}^{\lambda,y}\right|^2\right] = \\
&= \mathbb{E}\left[\sum_{j=1}^{r_x}\left|X_{j,x}^{x,\lambda}\right|^2\left|X_{j,(i,\mu)}^{\lambda,x}\right|^2\right]\mathbb{E}\left[\sum_{k=1}^{r_y}\left|X_{k,(i,\mu)}^{\lambda,y}\right|^2\left|X_{k,x}^{\lambda,y}\right|^2\right] = r_x r_y = N_\lambda r_y.
\end{aligned} \tag{77}$$

Then, since

$$\sum_{\substack{y=1\\y\neq x}}^{N_\lambda+1} r_y = N_\lambda^* - N_\lambda = N_\lambda(N_a-1), \tag{78}$$

we have that

$$\begin{aligned}
&\mathbb{E}\left[(\hat{\ell}_{x;i})^2\right] = \frac{1}{2}\left(\frac{2\operatorname{Tr}((\boldsymbol{A}^\lambda)^2)}{MN_\lambda^{*2}\operatorname{Tr}(\boldsymbol{A}^\lambda)}\right)^2 r_A^\lambda(N_a-1)N_\lambda^2 = \\
&= \frac{2\operatorname{Tr}((\boldsymbol{A}^\lambda)^2)^2}{M^2N_\lambda^{*4}\operatorname{Tr}(\boldsymbol{A}^\lambda)^2}\frac{\operatorname{Tr}(\boldsymbol{A}^\lambda)^2}{\operatorname{Tr}((\boldsymbol{A}^\lambda)^2)}(N_a-1)N_\lambda^2 = \frac{2(N_a-1)\operatorname{Tr}((\boldsymbol{A}^\lambda)^2)}{M^2N_a^4N_\lambda^2}
\end{aligned} \tag{79}$$

Thus,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\nabla\ell}\right\|^2\right] = \sum_{i=1}^{p}\mathbb{E}\left[(\hat{\ell}_i)^2\right] = \sum_{i=1}^{p}\sum_{\lambda=1}^{\Lambda'}\sum_{x=1}^{M_\lambda}\mathbb{E}\left[(\hat{\ell}_{x;i})^2\right] = \\
&= \sum_{\lambda=1}^{\Lambda'}\frac{2M_\lambda(N_a-1)\operatorname{Tr}((\boldsymbol{A}^\lambda)^2)p}{M^2N_a^4N_\lambda^2} = \Omega\left(\frac{1}{\operatorname{poly}(L)}\right).
\end{aligned} \tag{80}$$

□

Proposition 2. *For some $p = O(\operatorname{poly}(L))$, the Hessian matrix $\nabla^2\ell$ at any value of $\boldsymbol{\theta} \in [0, 2\pi)^p$ (where $\langle i|\left(\nabla^2\ell\right)|j\rangle = \partial_i\partial_j\ell(\boldsymbol{\theta})$) does not have full rank.*

Proof. We can write the second derivative of the loss function as:

$$
\begin{aligned}
\partial_i\partial_j\ell(\boldsymbol{\theta}) &= \frac{1}{M}\sum_{\lambda=1}^{\Lambda'}\sum_{x=1}^{M_\lambda}\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\left\langle x\right|\boldsymbol{U}^\lambda(\boldsymbol{\theta})\boldsymbol{O}_x^\lambda\boldsymbol{U}^\lambda(\boldsymbol{\theta})^\dagger\left|x\right\rangle = \\
&= -\frac{1}{M}\sum_{\lambda=1}^{\Lambda'}\sum_{x=1}^{M_\lambda}\left\langle x\right|e^{\mathbf{i}\boldsymbol{H}^\lambda t''}\Bigg(\left(\prod_{k=1}^{p}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_k}e^{\theta_k\boldsymbol{A}^\lambda}\left(\boldsymbol{A}^\lambda\right)^{\delta_{ik}+\delta_{jk}}e^{\mathbf{i}\boldsymbol{H}^\lambda t_k}\right)e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}\cdot \\
&\qquad\cdot\left(\prod_{k=p}^{1}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_k}e^{-\mathbf{i}\theta_k\boldsymbol{A}^\lambda}e^{\mathbf{i}\boldsymbol{H}^\lambda t_k}\right)- \\
&\quad-\left(\prod_{k=1}^{p}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_k}e^{\theta_k\boldsymbol{A}^\lambda}\left(\boldsymbol{A}^\lambda\right)^{\delta_{ik}}e^{\mathbf{i}\boldsymbol{H}^\lambda t_k}\right)e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}\cdot \\
&\qquad\cdot\left(\prod_{k=p}^{1}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_k}e^{-\mathbf{i}\theta_k\boldsymbol{A}^\lambda}\left(\boldsymbol{A}^\lambda\right)^{\delta_{jk}}e^{\mathbf{i}\boldsymbol{H}^\lambda t_k}\right)- \\
&\quad-\left(\prod_{k=1}^{p}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_k}e^{\theta_k\boldsymbol{A}^\lambda}\left(\boldsymbol{A}^\lambda\right)^{\delta_{jk}}e^{\mathbf{i}\boldsymbol{H}^\lambda t_k}\right)e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}\cdot \\
&\qquad\cdot\left(\prod_{k=p}^{1}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_k}e^{-\mathbf{i}\theta_k\boldsymbol{A}^\lambda}\left(\boldsymbol{A}^\lambda\right)^{\delta_{ik}}e^{\mathbf{i}\boldsymbol{H}^\lambda t_k}\right)+ \\
&\quad+\left(\prod_{k=1}^{p}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_k}e^{\theta_k\boldsymbol{A}^\lambda}e^{\mathbf{i}\boldsymbol{H}^\lambda t_k}\right)e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}\cdot \\
&\qquad\cdot\left(\prod_{k=p}^{1}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_k}e^{-\mathbf{i}\theta_k\boldsymbol{A}^\lambda}\left(\boldsymbol{A}^\lambda\right)^{\delta_{ik}+\delta_{jk}}e^{\mathbf{i}\boldsymbol{H}^\lambda t_k}\right)\Bigg)e^{-\mathbf{i}\boldsymbol{H}^\lambda t''}\left|x\right\rangle.
\end{aligned}
\tag{81}
$$

Now, for every i, let

$$
\boldsymbol{V}_i \coloneqq \prod_{k=1}^{i}e^{-\mathbf{i}\boldsymbol{H}^\lambda t_k}e^{\theta_k\boldsymbol{A}^\lambda}e^{\mathbf{i}\boldsymbol{H}^\lambda t_k}. \tag{82}
$$

Then, assuming $i \geq j$ (without loss of generality, since $\partial_i\partial_j\ell(\boldsymbol{\theta}) = \partial_j\partial_i\ell(\boldsymbol{\theta})$), we can rewrite the above as follows:

$$
\begin{aligned}
\partial_i\partial_j\ell(\boldsymbol{\theta}) &= -\frac{1}{M}\sum_{\lambda=1}^{\Lambda'}\sum_{x=1}^{M_\lambda}\left\langle x\right|e^{\mathbf{i}\boldsymbol{H}^\lambda t''}\cdot \\
&\qquad\cdot\Big(\boldsymbol{V}_j e^{-\mathbf{i}\boldsymbol{H}^\lambda t_j}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_j}\boldsymbol{V}_j^\dagger\boldsymbol{V}_i e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{V}_i^\dagger\boldsymbol{V}_p e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{V}_p^\dagger- \\
&\qquad-\boldsymbol{V}_i e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{V}_i^\dagger\boldsymbol{V}_p e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{V}_p^\dagger\boldsymbol{V}_j e^{-\mathbf{i}\boldsymbol{H}^\lambda t_j}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_j}\boldsymbol{V}_j^\dagger- \\
&\qquad-\boldsymbol{V}_j e^{-\mathbf{i}\boldsymbol{H}^\lambda t_j}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_j}\boldsymbol{V}_j^\dagger\boldsymbol{V}_p e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{V}_p^\dagger\boldsymbol{V}_i e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{V}_i^\dagger+ \\
&\qquad+\boldsymbol{V}_p e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{V}_p^\dagger\boldsymbol{V}_i e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{V}_i^\dagger\boldsymbol{V}_j e^{-\mathbf{i}\boldsymbol{H}^\lambda t_j}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_j}\boldsymbol{V}_j^\dagger\Big)\cdot \\
&\qquad\cdot e^{-\mathbf{i}\boldsymbol{H}^\lambda t''}\left|x\right\rangle = \\
&= -\frac{1}{M}\sum_{\lambda=1}^{\Lambda'}\sum_{x=1}^{M_\lambda}\left\langle x\right|e^{\mathbf{i}\boldsymbol{H}^\lambda t''}\cdot \\
&\quad\cdot\left[\boldsymbol{V}_j e^{-\mathbf{i}\boldsymbol{H}^\lambda t_j}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_j}\boldsymbol{V}_j^\dagger,\ \left[\boldsymbol{V}_i e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{A}^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t_i}\boldsymbol{V}_i^\dagger,\ \boldsymbol{V}_p e^{-\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'}\boldsymbol{V}_p^\dagger\right]\right]\cdot \\
&\quad\cdot e^{-\mathbf{i}\boldsymbol{H}^\lambda t''}\left|x\right\rangle.
\end{aligned}
\tag{83}
$$

Now, analogously to Theorem 1, we let $\boldsymbol{X}^\lambda$ be a matrix of dimension $N_\lambda^* \times (M_\lambda + pN_\lambda^*)$ (note that we are no longer doing the semi-isotropic reduction on $\boldsymbol{A}$) such that:

$$\boldsymbol{X}^\lambda |x\rangle = e^{-\mathbf{i}\boldsymbol{H}^\lambda t''} |x\rangle \tag{84}$$

and

$$\boldsymbol{X}^\lambda |i,\mu\rangle = \boldsymbol{V}_i e^{-\mathbf{i}\boldsymbol{H}^\lambda t_i} |\mu\rangle . \tag{85}$$

Let $\widetilde{\boldsymbol{O}}_x^\lambda = \boldsymbol{V}_p e^{-\mathbf{i}\boldsymbol{H}^\lambda t'} \boldsymbol{O}_x^\lambda e^{\mathbf{i}\boldsymbol{H}^\lambda t'} \boldsymbol{V}_p^\dagger$, $\boldsymbol{W}^\lambda = \boldsymbol{X}^{\lambda\dagger}\boldsymbol{X}^\lambda$, and $\boldsymbol{W}^{\lambda,x} = -\boldsymbol{X}^{\lambda\dagger}\widetilde{\boldsymbol{O}}_x^\lambda \boldsymbol{X}^\lambda$. Then, we can write:

$$\begin{aligned}
&\partial_i\partial_j \ell(\boldsymbol{\theta}) = \\
&= -\frac{1}{M}\sum_{\lambda=1}^{\Lambda'}\sum_{x=1}^{M_\lambda}\sum_{\mu=1}^{N_\lambda^*}\sum_{\nu=1}^{N_\lambda^*} \\
&\qquad a_\mu^\lambda a_\nu^\lambda \langle x| \boldsymbol{X}^{\lambda\dagger}\left[\boldsymbol{X}^\lambda |j,\nu\rangle\langle j,\nu| \boldsymbol{X}^{\lambda\dagger}, \left[\boldsymbol{X}^\lambda |i,\mu\rangle\langle i,\mu| \boldsymbol{X}^{\lambda\dagger}, \widetilde{\boldsymbol{O}}_x^\lambda\right]\right]\boldsymbol{X}^\lambda |x\rangle = \\
&= \frac{1}{M}\sum_{\lambda=1}^{\Lambda'}\sum_{x=1}^{M_\lambda}\sum_{\mu=1}^{N_\lambda^*}\sum_{\nu=1}^{N_\lambda^*} a_\mu^\lambda a_\nu^\lambda \Big(\langle x| \boldsymbol{W}^\lambda |j,\nu\rangle\langle j,\nu| \boldsymbol{W}^\lambda |i,\mu\rangle\langle i,\mu| \boldsymbol{W}^{\lambda,x} |x\rangle - \\
&\qquad - \langle x| \boldsymbol{W}^\lambda |j,\nu\rangle\langle j,\nu| \boldsymbol{W}^{\lambda,x} |i,\mu\rangle\langle i,\mu| \boldsymbol{W}^\lambda |x\rangle - \\
&\qquad - \langle x| \boldsymbol{W}^\lambda |i,\mu\rangle\langle i,\mu| \boldsymbol{W}^{\lambda,x} |j,\nu\rangle\langle j,\nu| \boldsymbol{W}^\lambda |x\rangle + \\
&\qquad + \langle x| \boldsymbol{W}^{\lambda,x} |i,\mu\rangle\langle i,\mu| \boldsymbol{W}^\lambda |j,\nu\rangle\langle j,\nu| \boldsymbol{W}^\lambda |x\rangle\Big) = \\
&= -\frac{2}{M}\sum_{\lambda=1}^{\Lambda'}\sum_{x=1}^{M_\lambda}\sum_{\mu=1}^{N_\lambda^*}\sum_{\nu=1}^{N_\lambda^*} a_\mu^\lambda a_\nu^\lambda \operatorname{Re}\Big(\langle x| \boldsymbol{W}^\lambda |i,\mu\rangle\langle i,\mu| \boldsymbol{W}^{\lambda,x} |j,\nu\rangle\langle j,\nu| \boldsymbol{W}^\lambda |x\rangle - \\
&\qquad - \langle x| \boldsymbol{W}^{\lambda,x} |i,\mu\rangle\langle i,\mu| \boldsymbol{W}^\lambda |j,\nu\rangle\langle j,\nu| \boldsymbol{W}^\lambda |x\rangle\Big) = \\
&= -\frac{2}{M}\sum_{\lambda=1}^{\Lambda'}\sum_{x=1}^{M_\lambda}\sum_{\mu=1}^{N_\lambda^*}\sum_{\nu=1}^{N_\lambda^*} a_\mu^\lambda a_\nu^\lambda \operatorname{Re}\left(\boldsymbol{S}_{i,j}^{\lambda,x,\mu,\nu,1}\boldsymbol{W}_{i,j}^{\lambda,x,\mu,\nu} - \boldsymbol{S}_{i,j}^{\lambda,x,\mu,\nu,2}\boldsymbol{W}_{i,j}^{\lambda,\mu,\nu}\right),
\end{aligned} \tag{86}$$

where

$$\boldsymbol{W}_{i,j}^{\lambda,x,\mu,\nu} = \langle i,\mu| \boldsymbol{W}^{\lambda,x} |j,\nu\rangle \tag{87}$$

$$\boldsymbol{W}_{i,j}^{\lambda,\mu,\nu} = \langle i,\mu| \boldsymbol{W}^{\lambda} |j,\nu\rangle \tag{88}$$

$$\boldsymbol{S}_{i,j}^{\lambda,x,\mu,\nu,1} = \langle x| \boldsymbol{W}^\lambda |i,\mu\rangle\langle j,\nu| \boldsymbol{W}^\lambda |x\rangle \tag{89}$$

$$\boldsymbol{S}_{i,j}^{\lambda,x,\mu,\nu,2} = \langle x| \boldsymbol{W}^{\lambda,x} |i,\mu\rangle\langle j,\nu| \boldsymbol{W}^\lambda |x\rangle . \tag{90}$$

So, we can write the Hessian as

$$\nabla^2\ell = -\frac{2}{M}\sum_{\lambda=1}^{\Lambda'}\sum_{x=1}^{M_\lambda}\sum_{\mu=1}^{N_\lambda^*}\sum_{\nu=1}^{N_\lambda^*} a_\mu^\lambda a_\nu^\lambda \operatorname{Re}\left(\boldsymbol{S}^{\lambda,x,\mu,\nu,1}\odot\boldsymbol{W}^{\lambda,x,\mu,\nu} - \boldsymbol{S}^{\lambda,x,\mu,\nu,2}\odot\boldsymbol{W}^{\lambda,\mu,\nu}\right). \tag{91}$$

Now, we know that

$$\operatorname{rank}(\boldsymbol{W}^{\lambda,x,\mu,\nu}) \le \operatorname{rank}(\boldsymbol{W}^{\lambda,\mu,\nu}) \le \operatorname{rank}(\boldsymbol{W}^\lambda) \le N_\lambda^*. \tag{92}$$

Additionally, it is clear that $\boldsymbol{S}^{\lambda,x,\mu,\nu,1}$ and $\boldsymbol{S}^{\lambda,x,\mu,\nu,2}$ both have rank 1: their entries are products of a term only depending on i and a term only depending on j, so they can be written as outer products of two vectors. Now, it is known that the rank of a Hadamard product of two matrices is upper-bounded by the product of their ranks, which means that

$$\operatorname{rank}\left(\nabla^2\ell\right) \le \sum_{\lambda=1}^{\Lambda'}\sum_{x=1}^{M_\lambda}\sum_{\mu=1}^{N_\lambda^*}\sum_{\nu=1}^{N_\lambda^*} 2N_\lambda^* = \sum_{\lambda=1}^{\Lambda'} 2M_\lambda (N_\lambda^*)^3 \le 2MN^{*3} = 2MN_a^3N^3. \tag{93}$$

Thus, if we choose $p \geq \operatorname{rank}\left(\nabla^2 \ell\right)$, it must be the case that $\nabla^2 \ell$, which is a $p \times p$ matrix, does not have full rank. □

Theorem 3 (No spurious local minima, formal version of Theorem 2)**.** *For some $p = O(\operatorname{poly}(L))$, define:*

$$\tilde{\ell}(\boldsymbol{\theta}) \coloneqq \ell(\boldsymbol{\theta}) + \xi(\boldsymbol{\theta}) \tag{94}$$

where we have defined the perturbation:

$$\xi\left(\boldsymbol{\theta}\right) \coloneqq \epsilon \sum_{i=1}^{p} \left(a_i \cos(\theta_i) + b_i \sin(\theta_i)\right), \tag{95}$$

where $\epsilon > 0$, $\boldsymbol{\theta} \in [0, 2\pi)^p$, and a_i, b_i are i.i.d. standard Gaussian (independent from the coefficients in the QNN ansatz). the expected number of critical points $\tilde{\ell}$ above the global minimum approaches 0 as $\epsilon \to 0^+$.

Proof. We choose p sufficiently large as in Proposition 2 so that the Hessian becomes singular. Now, consider the perturbation term $\xi : [0, 2\pi)^p \to \mathbb{R}$ as defined in the theorem statement. Observe that ξ is a smooth function on the p-torus and its distribution is rotationally invariant. Furthermore, the second derivative of ξ with respect to any θ_i is nonzero with high probability, and the marginal probability density of the gradient of ξ at any $\boldsymbol{\theta}$ is continuous. If we consider $\tilde{\ell}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \xi(\boldsymbol{\theta})$, we can now check the following conditions for applying the Kac-Rice formula (Adler & Taylor, 2007), as stated in Anschuetz (2022).

(1) All components of $\tilde{\ell}, \nabla\tilde{\ell}, \nabla^2\tilde{\ell}$ are almost surely continuous and have finite variances over $[0, 2\pi)^p$.

(2) At any $\boldsymbol{\theta} \in [0, 2\pi)^p$, the marginal probability density $\mathfrak{p}\left(\nabla\tilde{\ell}(\boldsymbol{\theta})\right)$ is continuous at $\nabla\tilde{\ell}(\boldsymbol{\theta}) = 0$.

(3) The conditional density of $\nabla\tilde{\ell}(\boldsymbol{\theta})$ conditioned on $\ell(\boldsymbol{\theta}), \nabla^2\tilde{\ell}(\boldsymbol{\theta})$ is bounded above and continuous at $\nabla\tilde{\ell}(\boldsymbol{\theta}) = 0$, uniformly in $\boldsymbol{\theta}$.

(4) The conditional density of $\det\left(\nabla^2\tilde{\ell}(\boldsymbol{\theta})\right)$ conditioned on $\nabla\tilde{\ell}(\boldsymbol{\theta}) = 0$ is continuous at $\det\left(\nabla^2\tilde{\ell}(\boldsymbol{\theta})\right) = 0$, uniformly in $\boldsymbol{\theta}$.

(5) The conditional density of $\tilde{\ell}(\boldsymbol{\theta})$ conditioned on $\nabla\tilde{\ell}(\boldsymbol{\theta}) = 0$ is continuous for values of $\tilde{\ell}(\boldsymbol{\theta})$ and $\nabla\tilde{\ell}(\boldsymbol{\theta})$ in a neighborhood of 0, uniformly in $\boldsymbol{\theta}$.

(6) The pth moments of the entries of the Hessian are bounded, uniformly in $\boldsymbol{\theta}$.

(7) The moduli of continuity ω of the components of $\tilde{\ell}(\boldsymbol{\theta}), \nabla\tilde{\ell}(\boldsymbol{\theta}), \nabla^2\tilde{\ell}(\boldsymbol{\theta})$ satisfy $\Pr\left[\omega(\eta) > \delta\right] = o(\eta^p)$ for all $\delta > 0$ as $\eta \to 0^+$.

Anschuetz (2022) demonstrated that these assumptions hold for variational loss landscapes; for us, the main difference is ensuring conditions (4) and (5), since we are looking at the overparameterized regime where the original Hessian is singular.

(1) follows by the construction of the QNN ansatz, (2) and (3) can be seen by considering the approximation to the distribution of the gradient in Theorem 1. (4) is ensured by adding the perturbation term $\xi(\boldsymbol{\theta})$ to the loss, which makes the Hessian non-singular with high probability. Similarly, for condition (5), the perturbation term breaks any symmetry in critical manifolds with high probability, thus making the conditional density of $\tilde{\ell}(\boldsymbol{\theta})$ continuous. (6) holds since the second derivatives are uniformly bounded. (7) holds by the argument in Anschuetz (2022).

We can now apply the Kac-Rice formula, which says that the expected number of local minima with loss values greater than the optimum ℓ_* is given by

$$\int_{[0,2\pi)^p} \mathfrak{p}\left(\nabla\tilde{\ell}(\boldsymbol{\theta}) = 0\right) \mathbb{E}\left[\left|\det\left(\nabla^2\tilde{\ell}(\boldsymbol{\theta})\right)\right| \mathbb{1}\left[\tilde{\ell}(\boldsymbol{\theta}) > \ell_* \wedge \nabla^2\tilde{\ell}(\boldsymbol{\theta}) \succeq 0\right] \mid \nabla\tilde{\ell}(\boldsymbol{\theta}) = 0\right] \mathrm{d}\boldsymbol{\theta}. \tag{96}$$

Since ϵ can be made arbitrarily small and the determinant is a continuous function of the entries of the Hessian, the expectation of the determinant in the integrand above can also be made arbitrarily small. Since the density of the gradient at zero is bounded, and the gradient is not affected by the perturbation to leading order in ϵ, the probability density term in the integrand remains bounded. Thus, the statement follows. □

B Details on the Lévy-Prokhorov Metric and Error Bounds

Definition 2 (Lévy-Prokhorov Metric)**.** *Consider two probability measures F, G over $\mathbb{R}^d$, we define the Lévy-Prokhorov distance between F and G relative to a norm $\|\cdot\|$ on $\mathbb{R}^d$ as*

$$\pi(F, G) = \inf\{\epsilon > 0 : F(A) < G(A^\epsilon) + \epsilon \ \forall \ A\}, \tag{97}$$

where A ranges over the Borel sets in $\mathbb{R}^d$ and A^ϵ is an ϵ-neighborhood of A in the given norm.

Lemma 6. $\pi(F, G) = \pi(G, F)$.

Proof. Suppose that for all Borel sets A, we have $F(A) < G(A^\epsilon) + \epsilon$. Then,

$$F(A^c) < G((A^c)^\epsilon) + \epsilon. \tag{98}$$

For $\epsilon > 0$, define $A^{-\epsilon}$ to be the set of all points x such that the open ϵ-ball centered at x is contained in A. Then, observe that

$$\overline{(A^c)^\epsilon} = \overline{(A^{-\epsilon})^c}, \tag{99}$$

which means that we can write

$$1 - F(A) < 1 - G(A^{-\epsilon}) + \epsilon \tag{100}$$

$$G(A^{-\epsilon}) < F(A) + \epsilon \tag{101}$$

Now, since for all A, we have $A \subseteq (A^\epsilon)^{-\epsilon}$, so we can substitute

$$G(A) \leq G((A^\epsilon)^{-\epsilon}) < F(A^\epsilon) + \epsilon. \tag{102}$$

Thus, $\pi(F, G) = \pi(G, F)$. □

Definition 3 (Convolution)**.** *If φ, γ are the probability densities of the distributions F, G, then the probability density of the convolution $F * G$ is defined to be*

$$\rho(t) = \int_{\mathbb{R}^d} \varphi(x)\gamma(t - x)\,\mathrm{d}x. \tag{103}$$

Definition 4 (Characteristic Function)**.** *The characteristic function of a distribution F with density φ is defined as*

$$f(t) = \mathbb{E}[e^{\mathbf{i}tX}] = \int_{\mathbb{R}} e^{\mathbf{i}tx}\varphi(x)\,\mathrm{d}x, \tag{104}$$

where in the above $X \sim F$ is a random variable.

Lemma 7. *Let F, G, and H be probability distributions on $\mathbb{R}^d$, and let $r > 0$. Then, the following inequality holds:*

$$\pi(F, G) \leq \pi(F * H, G * H) + 2\max\{r, H(\|x\| \geq r)\}, \tag{105}$$

Proof. Let $X \sim F, Y \sim H$ be independent random variables. Let $A, K \subseteq \mathbb{R}^d$. We know that

$$\begin{aligned} &(F * H)(A) = \Pr[X + Y \in A] \geq \\ &\geq \Pr[X + k \in A \ \forall k \in K] \Pr[Y \in K] \geq \\ &\geq \Pr[X + k \in A \ \forall k \in K] + \Pr[Y \in K] - 1 = \\ &= F(\{x : x + K \subseteq A\}) + H(K) - 1. \end{aligned} \tag{106}$$

Thus, letting K be the open ball of radius ϵ, we have that

$$(F * H)(A^\epsilon) \geq F(A) + H(\|x\| < \epsilon) - 1 = F(A) - H(\|x\| \geq \epsilon). \tag{107}$$

Note that we use the shorthand notation $H(\|x\| < \epsilon)$ to mean $H(\{x : \|x\| < \epsilon)\}$.

Now, we can also argue that

$$\begin{aligned}&(F*H)(A) = \Pr[X+Y\in A] \leq \\ &\leq \Pr[X\in A-K]\Pr[Y\in K] + \Pr[Y\in K^c] \leq \\ &\leq \Pr[X\in A-K] + \Pr[Y\in K^c] = F(A-K) + H(K^c).\end{aligned} \tag{108}$$

This gives us the inequality

$$(F*H)(A) \leq F(A^\epsilon) + H(\|x\|\geq\epsilon). \tag{109}$$

The same holds when replacing F with G. Now, if we let $L = \pi(F*H, G*H)$, then we must have by the above and by the definition of the Lévy-Prokhorov metric that

$$\begin{aligned}&F(A) \leq (F*H)(A^r) + H(\|x\|\geq r) \leq \\ &\leq (G*H)(A^{r+L}) + H(\|x\|\geq r) + L \leq \\ &\leq G(A^{2r+L}) + 2H(\|x\|\geq r) + L \leq \\ &\leq G(A^{L+2\max\{r, H(\|x\|\geq r)\}}) + 2H(\|x\|\geq r) + L + 2\max\{r, H(\|x\|\geq r)\},\end{aligned} \tag{110}$$

which proves that

$$\pi(F,G) \leq \pi(F*H, G*H) + 2\max\{r, H(\|x\|\geq r)\}. \tag{111}$$

□

Lemma 8. *(This lemma is adapted from Lemma 2.2 of Berkes & Philipp (1979)). Let F and G be probability distributions on $\mathbb{R}^d$ with characteristic functions f and g, respectively. Then, for some sufficiently large $T = O(d)$, we have that the Lévy-Prokhorov metric relative to the Euclidean norm is bounded as*

$$\pi_2(F,G) < \left(\frac{T}{\pi}\right)^d \int_{\|u\|_2\leq T} |f(u)-g(u)|\,\mathrm{d}u + F(\{x : \|x\|_2 \geq T/2\}) + 16d\frac{\log T}{T}. \tag{112}$$

Proof. Let H be a distribution on $\mathbb{R}^d$ with density $\nu(x)$ and characteristic function $h \in L^1$. Let $F_1 = F*H$ and $G_1 = G*H$. Let φ and γ be the probability densities of F_1 and G_1, respectively. Their characteristic functions are $f_1 = fh$ and $g_1 = gh$. Now, using the definition of convolution and applying the inverse Fourier transform, we have

$$\begin{aligned}&|\varphi(x)-\gamma(x)| = (2\pi)^{-d}\left|\int_{\mathbb{R}^d} e^{-\mathbf{i}\langle u,x\rangle}(f_1(u)-g_1(u))\,\mathrm{d}u\right| \leq \\ &\leq (2\pi)^{-d}\int_{\mathbb{R}^d} |(f(u)-g(u))|\,|h(u)|\,\mathrm{d}u \leq \\ &\leq (2\pi)^{-d}\left(\int_{\|u\|_2\leq T} |(f(u)-g(u))|\,\mathrm{d}u + 2\int_{\|u\|_2\geq T}|h(u)|\,\mathrm{d}u\right),\end{aligned} \tag{113}$$

where T is any real number. Then, for any Borel set $B \in \mathbb{R}^d$,

$$\begin{aligned}&F_1(B) - G_1(B) \leq F_1(B\cap\{\|x\|_2\leq T\}) - G_1(B\cap\{\|x\|_2\leq T\}) + F_1(\|x\|_2\geq T) \leq \\ &\leq \int_{\|x\|_2\leq T} |\varphi(x)-\gamma(x)|\,\mathrm{d}x + F(\|x\|_2\geq T/2) + H(\|x\|_2\geq T/2) \leq \\ &\leq \left(\frac{T}{\pi}\right)^d\left(\int_{\|u\|_2\leq T}|(f(u)-g(u))|\,\mathrm{d}u + 2\int_{\|u\|_2\geq T}|h(u)|\,\mathrm{d}u\right) + \\ &+ F(\|x\|_2\geq T/2) + H(\|x\|_2\geq T/2).\end{aligned} \tag{114}$$

Since for all B

$$F_1(B) = G_1(B) + (F_1(B)-G_1(B)) \leq G_1(B^{(F_1(B)-G_1(B))}) + (F_1(B)-G_1(B)), \tag{115}$$

we have that

$$\pi(F_1,G_1) \leq |F_1(B)-G_1(B)|\,. \tag{116}$$

We also know from Lemma 7 that for any $r > 0$,

$$\pi(F, G) \leq \pi(F_1, G_1) + 2\max\{r, H(\|x\|_2 \geq r)\}. \tag{117}$$

Now, if we choose $r \leq T/2$, we can put this together to get

$$\begin{aligned}\pi(F, G) \leq \left(\frac{T}{\pi}\right)^d \left(\int_{\|u\|_2 \leq T} |(f(u) - g(u))| \, \mathrm{d}u + 2\int_{\|u\|_2 \geq T} |h(u)| \, \mathrm{d}u\right) + \\ + F(\|x\|_2 \geq T/2) + 3\max\{r, H(\|x\|_2 \geq r)\}.\end{aligned} \tag{118}$$

Now, let $\sigma = 3d^{1/2}T^{-1}\log^{1/2} T$ and $r = 5dT^{-1}\log T$. Let H be a distribution with probability density

$$\nu(x) = (2\pi\sigma^2)^{-d/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^d x_j^2\right). \tag{119}$$

Then,

$$\begin{aligned}h(u) &= \int_{\mathbb{R}^d} e^{\mathbf{i}\langle u, x\rangle}\nu(x) \, \mathrm{d}x = \\ &= (2\pi\sigma^2)^{-d/2}\prod_{j=1}^d \int_{\mathbb{R}} e^{\mathbf{i}u_j x - \frac{1}{2\sigma^2}x^2} \, \mathrm{d}x = \\ &= (2\pi\sigma^2)^{-d/2}\prod_{j=1}^d \int_{\mathbb{R}} e^{-\frac{1}{2\sigma^2}(x^2 - 2\mathbf{i}\sigma^2 u_j x - \sigma^4 u_j^2 + \sigma^4 u_j^2)} \, \mathrm{d}x = \\ &= (2\pi\sigma^2)^{-d/2}\prod_{j=1}^d \int_{\mathbb{R}} e^{-\frac{1}{2\sigma^2}(x - \mathbf{i}\sigma^2 u_j)^2 - \frac{1}{2}\sigma^2 u_j^2} \, \mathrm{d}x = \\ &= (2\pi\sigma^2)^{-d/2}\exp\left(-\frac{1}{2}\sigma^2\sum_{j=1}^d u_j^2\right)\left(\int_{\mathbb{R}} e^{-\frac{1}{2\sigma^2}x^2} \, \mathrm{d}x\right)^d = \\ &= \exp\left(-\frac{1}{2}\sigma^2\sum_{j=1}^d u_j^2\right).\end{aligned} \tag{120}$$

Note that we are able to remove the $\mathbf{i}\sigma^2 u_j$ term in the exponent in the integrand, because the integrand is an entire function and so its integral around a rectangle with one side on the real axis and one side shifted by $-\mathbf{i}\sigma^2 u_j$ must be zero. Since the the contribution of the other two sides of the rectangle goes to zero, the value of the integral remains the same when moved to the real axis.

So now, we have that

$$\begin{aligned}&\pi(F, G) \leq \\ &\leq \left(\frac{T}{\pi}\right)^d \left(\int_{\|u\|_2 \leq T} |(f(u) - g(u))| \, \mathrm{d}u + 2\int_{\|u\|_2 \geq T} \exp\left(-\frac{1}{2}\sigma^2\sum_{j=1}^d u_j^2\right) \mathrm{d}u\right) + \\ &+ F(\|x\|_2 \geq T/2) + 3\max\left\{r, (2\pi\sigma^2)^{-d/2}\int_{\|u\|_2 \geq r} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^d u_j^2\right) \mathrm{d}u\right\}.\end{aligned} \tag{121}$$

Now, observe that for any A,

$$\begin{aligned}&\int_{\|u\|_2 \geq A} \exp\left(-\frac{1}{2}\sum_{j=1}^d u_j^2\right) \mathrm{d}u = (2\pi)^{d/2}\Pr[\chi_d^2 \geq A^2] = (2\pi)^{d/2}\Pr\left[e^{\frac{3}{8}\chi_d^2} \geq e^{\frac{3}{8}A^2}\right] \leq \\ &\leq (2\pi)^{d/2}e^{-3A^2/8}\,\mathbb{E}\left[e^{\frac{3}{8}\chi_d^2}\right] = (2\pi)^{d/2}e^{-3A^2/8}2^d.\end{aligned} \tag{122}$$

Going back to our previous expression, we can substitute $v_j = \sigma u_j$ for the second integral and $v_j = u_j/\sigma$ for the third, which gives us

$$\pi(F,G) \leq \left(\frac{T}{\pi}\right)^d \left(\int_{\|u\|_2 \leq T} |(f(u) - g(u))| \, \mathrm{d}u + \frac{2}{\sigma^d} \int_{|v| \geq \sigma T} \exp\left(-\frac{1}{2}\sum_{j=1}^d v_j^2\right) \mathrm{d}v\right) + \\ + F(\|x\|_2 \geq T/2) + 3\max\left\{r, (2\pi)^{-d/2} \int_{|v| \geq r/\sigma} \exp\left(-\frac{1}{2}\sum_{j=1}^d v_j^2\right) \mathrm{d}v\right\} = \\ = \left(\frac{T}{\pi}\right)^d \left(\int_{\|u\|_2 \leq T} |(f(u) - g(u))| \, \mathrm{d}u + \frac{2}{\sigma^d}(2\pi)^{d/2} e^{-3T^2\sigma^2/8} 2^d\right) + \\ + F(\|x\|_2 \geq T/2) + 3\max\left\{r, e^{-\frac{3r^2}{8\sigma^2}} 2^d\right\}. \tag{123}$$

Now, since

$$\frac{3r^2}{8\sigma^2} = \frac{3}{8}\frac{25d^2T^{-2}\log^2 T}{9dT^{-2}\log T} = \frac{25d\log T}{24}, \tag{124}$$

so the second term in the maximum becomes

$$T^{-\frac{25d}{24}} 2^d \tag{125}$$

which becomes subleading to

$$r = \frac{5d\log T}{T} \tag{126}$$

for large d. Also, we have that

$$\frac{2}{\sigma^d}(2\pi)^{d/2} e^{-3T^2\sigma^2/8} 2^d = \frac{2^{d+1}(2\pi)^{d/2}}{\left(3d^{1/2}T^{-1}\log^{1/2} T\right)^d} T^{-\frac{2T}{8}d}, \tag{127}$$

which becomes subleading if T is sufficiently large. Increasing the coefficient for r by one to strictly dominate the subleading terms, we arrive at the final expression:

$$\pi(F,G) \leq \left(\frac{T}{\pi}\right)^d \int_{\|u\|_2 \leq T} |(f(u) - g(u))| \, \mathrm{d}u + F(\|x\|_2 \geq T/2) + \frac{16d\log T}{T}. \tag{128}$$

□

Corollary 1. *Let F and G be two distributions on $\mathbb{R}^d$ with density functions φ and γ and characteristic functions f and g respectively. Assume that there exists some C such that*

$$\int_{\|x\|_\infty \geq C} f(x) \, \mathrm{d}x \leq \mu. \tag{129}$$

Then, there exists a universal constant K such that for all $T \geq \max(2C\sqrt{d}, Kd)$, we have that

$$\pi(F,G) \leq \left(\frac{T}{\pi}\right)^d \int_{\|u\|_\infty \leq T} |(f(u) - g(u))| \, \mathrm{d}u + \frac{16d\log T}{T} + \mu. \tag{130}$$

Proof. We need to convert from the Euclidean norm used in Lemma 8 to the infinity norm. We know that $\|u\|_\infty \leq \|u\|_2 \leq \sqrt{d}\,\|u\|_\infty$, so after performing that change, the claim immediately follows. □

Lemma 9. *(This has been adapted from Corollary 38 of Anschuetz (2025)). Let F, G be distributions on $\mathbb{R}^d$ with densities φ, γ and characteristic functions f, g. Assume that each moment of order $k' \leq k$ of F differs from that of G by an additive error of at most $\epsilon > 0$ and assume that all moments*

of order $k' > k$ are bounded by $(C\sqrt{k'})^{k'}$ for some constant C. Let $\zeta > 0$ be a sufficiently small constant. Assume that

$$\int_{\|x\|_\infty \geq \frac{1}{2\sqrt{d}} \min\left\{\frac{1-\zeta}{d}\log(\epsilon^{-1}), \frac{\sqrt{k+1}}{2eCd}\right\}} f(x)\,\mathrm{d}x \leq \mu. \tag{131}$$

Also assume that

$$d^2 = o\left(\min\left\{\frac{\log(\epsilon^{-1})}{\log\log(\epsilon^{-1})}, \frac{\sqrt{k}}{\log k}\right\}\right). \tag{132}$$

Then, the Lévy-Prokhorov distance is bounded by

$$\pi(F,G) = O\left(\frac{d^2\log\log(\epsilon^{-1})}{\log(\epsilon^{-1})} + \frac{d^2\log k}{\sqrt{k}} + \mu\right). \tag{133}$$

Proof. Let

$$T = \min\left\{\frac{1-\zeta}{d}\log(\epsilon^{-1}), \frac{\sqrt{k+1}}{2eCd}\right\}. \tag{134}$$

This, with the stated assumptions, satisfies the conditions for applying Corollary 1, which gives us that

$$\pi(F,G) \leq \widetilde{O}(T^{2d}) \sup_{\|u\|_\infty \leq T} |f(u) - g(u)| + O\left(\frac{d\log T}{T}\right) + O(\mu). \tag{135}$$

Since

$$dT \leq \frac{\sqrt{k+1}}{2eC}, \tag{136}$$

we can apply Lemma 37 of Anschuetz (2025) to say that

$$\sup_{\|u\|_\infty \leq T} |f(u) - g(u)| \leq \sup_{\|u\|_1 \leq dT} |f(u) - g(u)| \leq \epsilon e^{dT} + 2^{1-k} \leq \epsilon^\zeta + 2^{1-k}. \tag{137}$$

We also have that

$$T^{2d} = \exp(2d\log T) = \exp\left(o(\min\{\log(\epsilon^{-1}), \sqrt{k}\})\right). \tag{138}$$

So,

$$\widetilde{O}(T^{2d}) \sup_{\|u\|_\infty \leq T} |f(u) - g(u)| + O\left(\frac{d\log T}{T}\right) \leq O\left(\epsilon^{\zeta+\sqrt{k}} + e^{\sqrt{k}+(1-k)\log 2}\right). \tag{139}$$

Now, looking at the second term, we have that

$$\begin{aligned} &O\left(\frac{d\log T}{T}\right) \leq O\left(d\min\left\{\log\log(\epsilon^{-1}), \log k\right\}\max\left\{\frac{d}{\log(\epsilon^{-1})}, \frac{d}{\sqrt{k}}\right\}\right) \leq \\ &\leq O\left(\frac{d^2\log\log(\epsilon^{-1})}{\log(\epsilon^{-1})} + \frac{d^2\log k}{\sqrt{k}}\right). \end{aligned} \tag{140}$$

This second term dominates the first, which means that

$$\pi(F,G) = O\left(\frac{d^2\log\log(\epsilon^{-1})}{\log(\epsilon^{-1})} + \frac{d^2\log k}{\sqrt{k}} + \mu\right). \tag{141}$$

□

C Numerical Experiments

In this appendix, we provide small-scale numerical experiments to illustrate our main results. We train our QNN model to perform fragment classification on the Temperley-Lieb model, using simple toy datasets. It is important to note that these experiments are not intended to serve as a demonstration of a direct practical application of our work, as the model we consider already admits efficient classical simulation techniques. They are also not meant to demonstrate any specific asymptotic scaling behavior proven in our main results, as simulating the QNN model classically becomes impractical for larger system sizes. The experiments serve as illustrative examples to demonstrate QNN trainability for the fragment classification task.

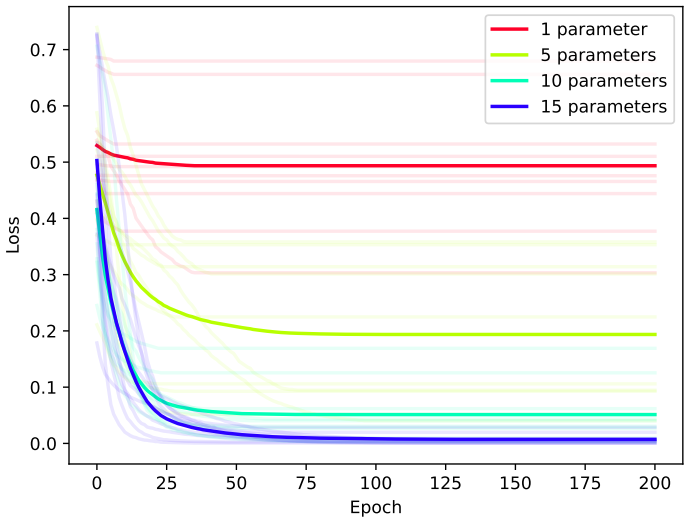


Figure 2: Training curves for QNN models with 1, 5, 10, and 15 parameters on the 4-qubit Temperley-Lieb dataset. For each number of parameters, 10 QNNs were randomly initialized and trained for 200 epochs with a learning rate of 0.1. Training was stopped if the loss was decaying slower than 5% every 5 epochs or if it reached less than 0.01. The faded lines show the individual training curves, and the solid lines show their averages. For these diagrams, the loss value is adjusted to scale from 0 to 1 instead of from -1 to 0.

C.1 QNN Training

We consider a Schur basis for the Temperley-Lieb model on 2-level quantum systems constructed in correspondence with the "dots and dimers" picture of Moudgalya & Motrunich (2022). For simplicity, we take a subset of such a basis consisting of states that can be expressed as tensor products of computational basis states and the EPR state $|\Phi^+\rangle$. We demonstrate that our QNN model can be trained to distinguish states of this form with different configurations of entangled pairs, while generalizing across any valid state on the non-entangled qubits. We consider the following 2-entry dataset on a 4-qubit system:

$$\mathcal{D}_4 = \left\{ |\Phi^+\rangle \otimes |01\rangle, |\Phi^+\rangle^{\otimes 2} \right\}, \tag{142}$$

and the following 4-entry dataset on a 6-qubit system:

$$\mathcal{D}_6 = \left\{ |\Phi^+\rangle \otimes |0101\rangle, |0101\rangle \otimes |\Phi^+\rangle, |\Phi^+\rangle^{\otimes 2} \otimes |01\rangle, |\Phi^+\rangle^{\otimes 3} \right\}. \tag{143}$$

We then construct the QNN ansatz using local generators of the form $|\Phi^+\rangle\langle\Phi^+|_{(i,i+1)}$, randomly selecting the Hamiltonian, the coefficients $t_1, \ldots, t_p, t', t''$ as in Section 3.1. The QNN is initialized with random parameters $\boldsymbol{\theta}$ and optimized by gradient descent.

The simulations were performed using the `qutip` library in Python. Since our simulations explicitly compute matrix products and exponentials of operators acting on the exponential-size Hilbert space, these methods become impractical for larger system sizes. That said, even at small system sizes our results—reported in Figures 2 to 4—support the correctness of the QNN construction and illustrate the phenomenon of the overparameterized regime: there exists a number of parameters for which the QNN will almost always reach the global optimum loss during training.

C.2 Empirical Tests of the ETH Assumption

Here we do a simple demonstration to support the claim that we can reasonably expect our Hamiltonians to satisfy the full ETH in each Krylov subspace for our analysis. For the Temperley-Lieb

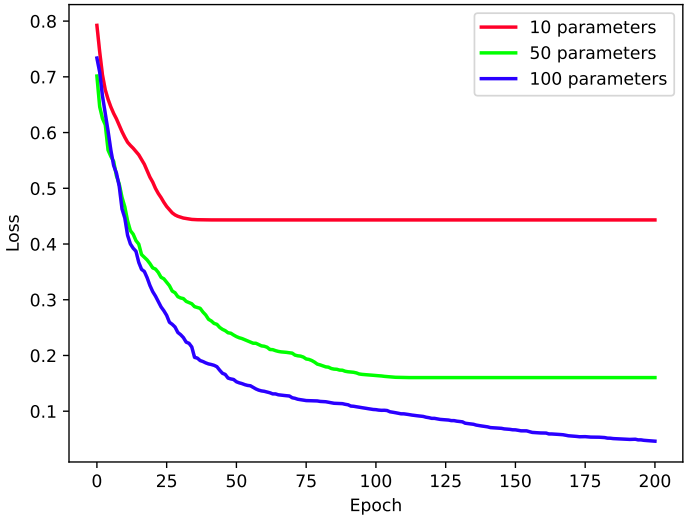


Figure 3: Training curves for QNN models with 10, 50, and 100 parameters on the 6-qubit Temperley-Lieb dataset. For each number of parameters, a QNN was randomly initialized and trained for 200 epochs with a learning rate of 1. Training was stopped if the loss was decaying slower than 5% every 5 epochs or if it reached less than 0.01. Note that the increase in the number of parameters needed compared to Figure 2 is due not just to the larger system size, but also to the increased number of dataset entries, doubling the dimension N_a.

Hamiltonian considered in the QNN training experiments above, we numerically check that

$$\mathbb{E}[\mathrm{Tr}(e^{-iHt}\left|x\right\rangle\left\langle x\right|e^{iHt}O_x)^2] - \mathbb{E}[\mathrm{Tr}(e^{-iHt}\left|x\right\rangle\left\langle x\right|e^{iHt}O_x)]^2 = O(1/N_\lambda^*), \tag{144}$$

analogously to Equation (C3) of Fava et al. (2025), which we use in our Lemma 2. Here, $|x\rangle$ is a state from some subspace labelled by λ—in this case, we take a state of the form $|\Phi^+\rangle \otimes |0101\ldots\rangle$, which is in a Krylov subspace whose dimension grows with system size. Note that for the purposes of our analysis, it is sufficient that quantities of this form decay with the system size (not necessarily as $O(1/N_\lambda^*)$).

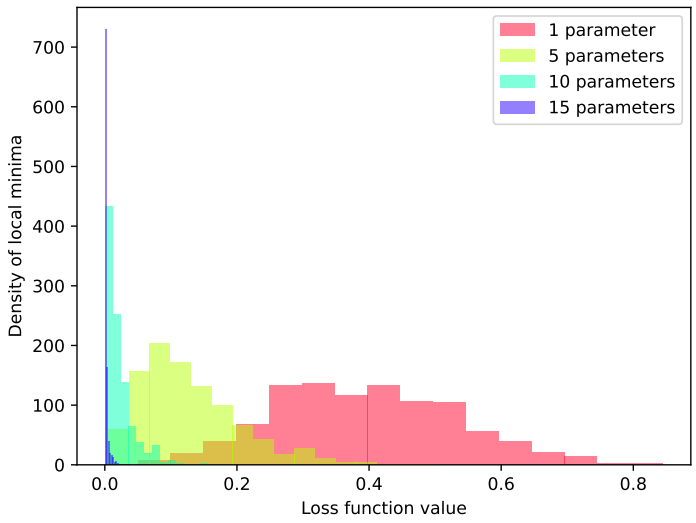


Figure 4: Distribution of local minima reached during training for the 4-qubit Temperley-Lieb dataset. For this plot, 1000 QNNs were randomly initialized for each number of parameters and trained for 200 epochs with a learning rate of 0.1. The data plotted on the x-axis is the final loss achieved during training. Since we can assume that during gradient descent the parameters converge to the "closest" local minimum, this is a proxy for the distribution of the local minima of the loss landscape. We can see that with 15 parameters, the QNN enters the overparameterized regime since the distribution of minima becomes peaked at zero.

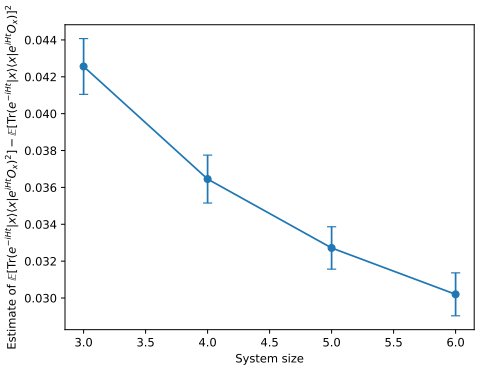


Figure 5: For each system size from 3 to 6, we sample 1000 different instances of the Temperley-Lieb Hamiltonian and time-evolve them for 100 different random times to estimate the quantity in Equation (144).