EXPLORING CHANNEL DISTINGUISHABILITY IN LOCAL NEIGHBORHOODS OF THE MODEL SPACE IN QUANTUM NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

With the increasing interest in Quantum Machine Learning, Quantum Neural Networks (QNNs) have emerged and gained significant attention. These models have, however, been shown to be notoriously difficult to train, which we hypothesize is partially due to the architectures, called ansatzes, that are hardly studied at this point. Therefore, in this paper, we take a step back and analyze ansatzes. We initially consider their expressivity, i.e., the space of operations they are able to express, and show that the closeness to being a 2-design, the primarily used measure, fails at capturing this property. Hence, we look for alternative ways to characterize ansatzes, unrelated to expressivity, by considering the local neighborhood of the model space, in particular, analyzing model distinguishability upon small perturbation of parameters. We derive an upper bound on their distinguishability, showcasing that QNNs using the Hardware Efficient Ansatz with few parameters are hardly discriminable upon update. Our numerical experiments support our bounds and further indicate that there is a significant degree of variability, which stresses the need for warm-starting or clever initialization. Altogether, our work provides an ansatz-centric perspective on training dynamics and difficulties in QNNs, ultimately suggesting that iterative training of small quantum models may not be effective, which contrasts their initial motivation.

1 INTRODUCTION

With the increasing computational requirements of state-of-the-art ML models, the limitations of classical hardware, as theorized by Moore's law (Shalf, 2020), are becoming increasingly prevalent (Schulz et al., 2021; Reed et al., 2022). Therefore, alternative computing paradigms are heavily studied and the immense potential of Quantum Computing, combined with recent advances in quantum hardware, have made Quantum Machine Learning (QML) a promising candidate.

Within that framework, Quantum Neural Networks (QNNs) are heavily studied, as they are suitable for near-term hardware. Besides initial successes, however, the *Barren Plateau* phenomenon has been proven under various circumstances, meaning that gradients vanish exponentially in the number of qubits, leading to essentially flat loss landscapes that affect trainability (Cerezo et al., 2021a).

In light of these issues, we hypothesize that the architectures, called *ansatzes*, are part of the problem. QNN ansatzes consist of fixed and trainable gates, however, due to hardware limitations, currently mainly so-called Hardware Efficient Ansatzes (HEAs) (Kandala et al., 2017) are used, meaning they are built primarily based on hardware constraints. Therefore, we hypothesize that executing the circuit does not lead to a meaningful feature representation.

Currently, it is unclear how individual operations affect the model space, or, when including data, the data representation. There are no standardized ansatzes for Machine Learning (ML) and, due to the many degrees of freedom, ansatz tuning poses an optimization problem with intractable search space. Further, there is little research on the ansatzes themselves, i.e., independently of data and loss, and it is non-obvious how feature extraction works in a QNN.

In particular, in classical ML, a vanilla CNN is ultimately designed to consider local features in lower layers, before considering global features in deeper layers, independently of the dataset. Transform-

ers (Vaswani et al., 2017) work similarly, however, the trainable one-qubit and fixed two-qubit gates that are used in QNNs do not allow drawing intuitive conclusions about how features are extracted.

Therefore, the goal of this paper is to analyze and characterize QNN architectures. We aim to take a first step at exploring ansatzes and the model space, initially by considering the expressivity of ansatzes. Using tools from Quantum Information Theory and extensive numerics, we later analyze how the quantum circuit, i.e., the applied operation, changes upon parameter update, thus, providing insights into the model space beyond ansatz expressivity. Our contributions are as follows.

- Based on literature, we argue that *closeness to a 2-design*, a widely used measure for ansatz expressivity, does not adequately represent the possible set of unitary operations.
- To investigate the model space, we study how an ansatz changes upon parameter update for different HEAs. In particular, we introduce a measure for the distinguishability of an ansatz upon parameter update and provide an upper bound. It shows that many parameters are required to be able to distinguish the two operations, contrasting with the initial framework.
- We provide evidence that HEAs are hardly distinguishable upon parameter update in (1) random perturbations and (2) during training, even in early stages, with larger updates.
- The observed behavior has remarkable similarities to Barren Plateaus, thus, insinuating that the architectures, independently of the data or loss used, may be flawed and play a fundamental role in the trainability problems observed today.

The paper is structured as follows. In Section 2, we provide background information on QNNs, existing expressivity measures and related work[1]. In Section 3, we discuss the closeness to being a 2-design as a measure of expressivity and Section 4 introduces a measure of distinguishability of quantum channels. We elaborate on our experimental setup in Section 5, and present our results in Section 6. Finally, Section 7 discusses implications of our results and Section 8 concludes the paper.

2 Background

2.1 Quantum Neural Networks

The small systems and the noise in today's Noisy Intermediate-Scale Quantum (NISQ) technology (Preskill, 2018), prevents execution of large-scale circuits. Consequently, so-called Variational Quantum Circuits (VQC) are applied, which are hybrid models with classical and quantum parts.

Simply put, a QNN consists of (1) an input state $|\psi_{\boldsymbol{X}}\rangle$, (2) an ansatz $U(\boldsymbol{\vartheta})$, (3) a loss function, and (4) an optimizer. The features are encoded onto the quantum state by means of a unitary feature encoding method $V(\boldsymbol{X})$, and the initial state is obtained as $|\psi_{\boldsymbol{X}}\rangle = V(\boldsymbol{X})\,|0\rangle^{\otimes n}$, with n as the number of qubits. HEAs consists of L layers of trainable one-qubit rotation gates $V_i(\vartheta_i)$ and fixed two-qubit entangling gates W_i, with $i \in [1, L]$, expressed as Equation 1. A measurement is taken in the end, which is done through a Hamiltonian observable $\hat{O}$, and the result is expressed as the expectation value of the ansatz with respect to $|\psi_{\boldsymbol{X}}\rangle$ and $\hat{O}$, as shown in Equation 2.

$$U(\boldsymbol{\vartheta}) = \prod_{i=1}^{L} V_i(\vartheta_i)W_i = V_L(\vartheta_L)W_L \dots V_1(\vartheta_1)W_1 \tag{1}$$

$$f(\boldsymbol{X}, \boldsymbol{\vartheta}) = \langle\psi_{\boldsymbol{X}}|\, U^\dagger(\boldsymbol{\vartheta})\hat{O}U(\boldsymbol{\vartheta})\,|\psi_{\boldsymbol{X}}\rangle \tag{2}$$

The loss is calculated with a classical loss function (e.g., the mean-squared error) on the predicted value (the expectation value of the QNN) and the target value, and passed to the classical optimizer. The optimizer computes the parameter updates and the circuit is again executed on the quantum computer, until some stopping criterion is fulfilled. As the optimization is done classically, this approach has the advantage of allowing a shallow quantum circuit, which limits the propagation of errors during execution in light of the significant amount of noise in today's hardware.

[1]For those unfamiliar with quantum computing, we refer to Appendix A.1 for a brief overview.

While HEAs are shallow, they are highly expressive (Holmes et al., 2022), which is a main promise of QML stemming from the significantly larger Hilbert space. Recent works, however, uncovered that loss landscapes of today's QNNs are plagued by Barren Plateaus (BPs), meaning exponentially vanishing gradients in problem size, leading to untrainable models (McClean et al., 2018).

2.2 Ansatz expressivity

We acknowledge a missing clear definition of expressivity in QML for now. It is sensible, therefore, to transfer it from classical ML, where model expressivity is the range of functions a model can compute (Raghu et al., 2017). Hence, it is a function of architecture A, input x and parameters $\boldsymbol{\vartheta}$; $F_A(x;\boldsymbol{\vartheta})$. On the contrary, our work lies in data-agnostic characterization of models. To this end, we will only consider *ansatz expressivity*; the set of functions generated solely by an ansatz itself.

The most commonly used method to evaluate ansatz expressivity is to compare the probability distribution of the unitaries of the ansatz to the *Haar measure*[2], the uniform distribution over all unitaries $\mathcal{U}(d)$ (Cerezo et al., 2021a). If the probability distribution of the ansatz follows the Haar measure up to the t-th moment, it is considered a *t-design*. Their construction requires exponential time (Harrow & Low, 2009), however, approximate t-designs can be built efficiently (Dankert et al., 2009).

Closeness to a 2-design as a measure of expressivity for ansatzes was first defined for the $|0\rangle$ input state by Sim et al. (2019). In Holmes et al. (2022), it is expanded to a specific input state (i.e., with encoded data) and the measurement operator, as shown in Equation 3. There, an ansatz is viewed as an ensemble of unitary transformations $\mathbb{U} = \{U(\boldsymbol{\vartheta}), \forall \boldsymbol{\vartheta} \in \boldsymbol{\Theta}\}$ over all possible parameters $\boldsymbol{\Theta}$.

$$A_{\mathbb{U}}^{(t)}(\cdot) = \int_{\mathcal{U}(d)} d\mu(V) V^{\otimes t}(\cdot)(V^\dagger)^{\otimes t} - \int_{\mathbb{U}} dU U^{\otimes t}(\cdot)(U^\dagger)^{\otimes t} \quad \forall\, V \in \mathcal{U}(d) \tag{3}$$

2.3 Related work

BPs have been proven for deep (McClean et al., 2018) and shallow circuits with global measurement (Cerezo et al., 2021b), expressivity (Holmes et al., 2022; McClean et al., 2018), noise (Wang et al., 2021), entanglement (Ortiz Marrero et al., 2021) and it has been linked to cost concentration and narrow gorges (Arrasmith et al., 2022). Further, even shallow models without BPs have only a small fraction of local minima close to the global minimum (Anschuetz & Kiani, 2022)[3].

The problem has been linked to the curse of dimensionality in Cerezo et al. (2024), therefore, limiting the accessible Hilbert space was proposed. However, they showed that such models can be simulated classically, given an initial data acquisition phase on a quantum computer. This was extended to a trade-off between trainability and dequantizability in Gil-Fuster et al. (2024).

Besides that, parameter initialization strategies have been proposed, with some even providing better bounds on gradient magnitudes (Rad et al., 2022; Wang et al., 2023; Kulshrestha & Safro, 2022; Grant et al., 2019), however, the results of Herbst et al. (2024) imply that it is not the type of statistical distribution, but rather the used parameter ranges that could lead to better starting points.

Further, expressivity is commonly discussed to prove a quantum advantage with respect to classical models. The closeness to being a 2-design is the most widely employed measure, however, others have been proposed. In particular, Equation 3 is related to the Frame Potential (Gross et al., 2007; Roberts & Yoshida, 2017; Nakaji & Yamamoto, 2021). Another approach is to analyze the entangling capability as in Sim et al. (2019), i.e., the entanglement of the produced states. Other measures are the Covering Number from Du et al. (2022), analyzing Fourier coefficients (Schuld et al., 2021; Caro et al., 2021), or the Effective Dimension from Abbas et al. (2021). Further, one can consider *unitary t-designs*, i.e., consider the operations without the input state[4]. We employ the closeness to being a 2-design in our study, as it is frequently used, due to its intuitive interpretation. Beyond that, due to computational complexity, there is little research on properties of the architectures for now.

[2] We refer to Appendix A.3 for a short introduction to unitary groups and the Haar measure.

[3] For further information, we refer to Larocca et al. (2024), who published an excellent review on the topic.

[4] Unitary designs and the difference between unitary and state t-designs are discussed in Appendix A.3.

Moreover, the field of QML has seen a rise in works on efficient design of ansatzes. Recent works, such as Leone et al. (2024), have actively highlighted that well-defined ansatzes are a crucial step for large-scale deployment. In the field of ML, such approaches particularly include the field of Geometric Quantum Machine Learning (GQML) (Ragone et al., 2023; Larocca et al., 2022; Tüysüz et al., 2024; Wiersema et al., 2024), or adaptive ansatzes (Bilkis et al., 2023).

Further, the practical utility of the HEA has already been questioned and studied by Leone et al. (2024). They find that problems with product input states and data that satisfies a volume law of entanglement should avoid using such ansatzes, whereas area law of entanglement data leads to optimization landscapes without BPs. Moreover, the necessity of understanding and characterizing ansatzes has been discussed in Zhang et al. (2024), as well as in Larocca et al. (2023).

Changes in quantum channel or loss landscape upon perturbing parameters has not yet been studied for QNNs. However, perturbation analysis and sensitivity analysis have emerged as tools in classical ML. Examples are perturbing the input for explainability (Ivanovs et al., 2021; Pizarroso et al., 2022; Fel et al., 2023), or checking for stability (Testa et al., 2024). Similarly, weights can be perturbed for trainability (Wen et al., 2018), or to increase robustness and generalization (Wu et al., 2020; He et al., 2019; Dumford & Scheirer, 2020). To the best of our knowledge, there is no research on perturbing parameters to check how the underlying function changes, as, due to the stochasticity of quantum computing, this research question seems to be much more relevant in the quantum realm.

3 2-DESIGN AS A MEASURE OF EXPRESSIVITY

While 2-designs have turned out to be sufficient for many quantum information processing protocols (Holmes et al., 2022; Harrow & Low, 2009), in this section, we summarize and connect results from the literature to argue that closeness to a 2-design is an inadequate measure for model expressivity (i.e., the ability to sufficiently approximate the unitary group to a high degree).

For the argument we will use the so-called *Welch Bounds* which, loosely put, are inequalities related to evenly distributing a finite number of unit vectors in a vector space. This inequality was addressed in Welch (1974), where a lower bound on the inner-product between unit vectors was provided. Intuitively, a smaller lower bound corresponds to more evenly spread vectors in the vector space, i.e., a more uniform distribution.

Quantum **state t-designs** reduce integrals of polynomials over all quantum states to averages over a discrete set (cf. Equation 19). These are probability distributions over pure quantum states that replicate properties of the uniform (Haar) measure on the quantum states up to the t-th moment. We measure expressivity of 2-design based architectures by comparing (generalized) Welch bound (Scott, 2008) values for state 2-designs against its t-design counterparts, with $t > 2$[5]. We state the inequality for a state t-design based architecture:

State t-design Welch bound: The choice of the polynomial function for state t-designs is presented in Appendix A.4. We state the following inequality for a state in a d-dimensional Hilbert space, $|\psi_i\rangle \in \mathbb{C}^d$ (cf. Appendix B.2 for the derivation):

$$\sum_{1\leq j,k\leq n} |\langle\psi_j|\psi_k\rangle|^{2t} \geq \frac{n^2}{\binom{d+t-1}{t}} = \frac{n^2}{\frac{(d+t-1)(d+t-2).....d}{t!}} \quad \forall t \in \mathbb{N}. \tag{4}$$

It is clear that for an increasing t, the term in the denominator $\frac{(d+t-1)(d+t-2).....d}{t!}$ increases rapidly (as compared to Equation 24 for 2-designs), yielding a vanishing overlap between the state vector summands. This makes the inequality in Equation 4 more stringent for increasing t values. This indicates that higher-order designs ($t > 2$) approximating the Haar measure up to the t-th order impose stronger constraints, requiring greater separation between a sequence of state vectors $\{\psi_1, \psi_2, ..., \psi_d\}$ to maintain equality in the bounds. This means that 2-design models have far fewer *degrees of freedom* in building complex quantum circuits. On the contrary, t-designs ($t > 2$) can more efficiently construct higher-order representations capturing more complex interactions between quantum states. Thus, evaluating Welch bounds for 2-designs corroborates its inadequacy for constituting an expressive architecture.

[5] As mentioned in Appendix A.3, unitary t-designs induce state t-designs when acting on fixed input states. Thus, our data-agnostic approach is also valid when considering state t-designs in the following.

An ansatz forming a 2-design is not expressive as following the Haar measure only up to the second moment constrains how *spread-out* (statistical spread) the unitaries must be, however, does not allow drawing conclusions about how densely it covers $\mathcal{U}(N)$. If one were to define expressivity as the capability of a model to represent all possible set of unitaries, it would be required to go way beyond second moments (for, e.g., Kurtosis, off-diagonal correlations etc) to ensure the necessary weights are captured where there is sufficiently dense population. As the moment operator requires tensor products proportional to the moments, this becomes computationally infeasible very quickly. It is not clear how to overcome the curse of dimensionality while proposing a measure that captures expressivity adequately; in fact, most tools from QIT suffer from this very phenomenon, speaking to the difficulty of designing such a metric. Further, there already exists a line of non-unitary VQA research (e.g., Cong et al. (2019); Deshpande et al. (2024)), that needs to be considered as well.

As expressivity is currently the most common metric used to describe the ansatzes, our work employs an alternative approach to analyze the model space, *unrelated to expressivity*, but highly relevant for trainability and training dynamics. Therefore, we focus on local neighborhoods of the model space, in particular, we study how the QNN changes upon perturbation of parameters. This allows analyzing training dynamics and tracking the extent to which the QNN changes during training.

4 Channel Sensitivity

Comparison between quantum channels is an important topic in Quantum Information Theory. For consistency with respect to, e.g., highly entangling operations, however, operator norms need to be stable under tensor product. Therefore, the diamond norm was proposed in Kitaev (1997) as follows.

Definition 1 *The **diamond norm** of a superoperator T is defined as (Harrow & Low, 2009; Kitaev et al., 2002).*

$$\|T\|_\diamond = \sup_d \|T \otimes id_d\|_\infty = \sup_d \sup_{X \neq 0} \frac{\|(T \otimes id_d)X\|_1}{\|X\|_1} \tag{5}$$

Here, $\otimes$ denotes the tensor product and id_d the identity channel on d dimensions, with $d \leq 2^n$, and n as the number of qubits. The $\|X\|_1$ in this context is the *Trace norm* or *Schatten 1-norm*, which is defined as $\|X\|_1 = Tr\sqrt{X^\dagger X}$. The diamond norm is defined for any superoperator T, wherein, superoperators are defined as the set of linear maps acting on a vector space of linear operators. For a valid quantum channel $\mathcal{E}_1$ (a completely positive and trace preserving (CPTP) matrix), the diamond norm is strictly upper bounded by 1 and lower bounded by 0. When using the diamond norm to measure the distance between two CPTP quantum channels $\mathcal{E}_1$ and $\mathcal{E}_2$, the channels are subtracted. The resulting matrix is not necessarily CPTP, resulting in a revised upper bound of 2.

Operationally, the diamond norm measures the maximum distinguishability between the output states of the two maps under any input state, i.e., for any input state we will be *less or equally able to distinguish* the output. A value of 0, means that they are indistinguishable, whereas a value of 2 means they are perfectly distinguishable. Computing the diamond norm is non-trivial, however, can be reduced to a semi-definite program (Watrous, 2009; 2013). We apply the diamond norm to evaluate the distinguishability of a parametrized unitary $U(\boldsymbol{\vartheta})$, representing a quantum channel, from the unitary $U(\boldsymbol{\vartheta} + \boldsymbol{\delta})$, where $\boldsymbol{\delta}$ is a perturbation of $\boldsymbol{\vartheta}$, as follows.

Definition 2 *We define the **channel sensitivity** as the distinguishability of the quantum channel from itself upon small perturbation of $\boldsymbol{\vartheta}$ by $\boldsymbol{\delta}$.*

$$cs_U(\boldsymbol{\vartheta}, \boldsymbol{\delta}) = \|U(\boldsymbol{\vartheta}) - U(\boldsymbol{\vartheta} + \boldsymbol{\delta})\|_\diamond \tag{6}$$

We can provide an upper bound on the channel sensitivity through Taylor expansion, which is shown in Equation 7. Therefore, we establish a direct dependence of the distinguishability of the channels to the sum of changes applied[6]. The bound holds if the Hermitian generators of the trainable gates

[6]For the precise mathematical derivation, we refer to Appendix C.

Table 1: Employed ansatzes

COMPONENT	CONFIGURATIONS
Parameterized	[R_X, R_Y, R_Z, $R_X R_Y$, $R_Y R_Z$, $R_X R_Z$, $R_X R_Y R_Z$]
Entangling	[CNOT, CZ, CNOT CZ]

are unitary as well, which is relevant for HEAs, as the trainable gates are exponentiations of the X, Y, and Z Pauli gates. The precision of the bound depends on the magnitude of $\boldsymbol{\delta}$, however, considering that the models are trained iteratively in small steps, we expect the bound to hold.

$$\begin{aligned} \|U(\boldsymbol{\vartheta}) - U(\boldsymbol{\vartheta}+\boldsymbol{\delta})\|_\diamond &= \|U(\boldsymbol{\vartheta}) - (U(\boldsymbol{\vartheta}) + \sum_{j=1}^{dim(\boldsymbol{\delta})} \delta_j \frac{\partial U(\boldsymbol{\vartheta})}{\partial \vartheta_j} + O(\boldsymbol{\delta}^2))\|_\diamond \\ &\approx \| - \sum_{j=1}^{dim(\boldsymbol{\delta})} \delta_j \frac{\partial U(\boldsymbol{\vartheta})}{\partial \vartheta_j} \|_\diamond \leq \frac{\sum_{j=1}^{dim(\boldsymbol{\delta})} |\delta_j|}{2} \end{aligned} \tag{7}$$

Unless the feature encoding is trained as well, the upper bound on distinguishability cannot be improved by including data. Intuitively, due to the calculation of the diamond being based on the maximum distinguishability from any input state, it includes any possible unitary data encoding as well. Mathematically, this property follows from the unitary invariance of the Schatten 1-norm.

Our bound shows that (assuming small parameters updates), there is a direct relationship between the number of parameters and the distinguishability of the operation. In particular, for small models, which are used extensively at the moment, it can be expected that subsequent unitaries during training are hardly distinguishable, hindering effective training. This is, to the best of our knowledge, the first result attempting to establish a connection between ansatz and trainability issues. To strengthen our results, we want to, in the following, support our bounds through numerical experiments.

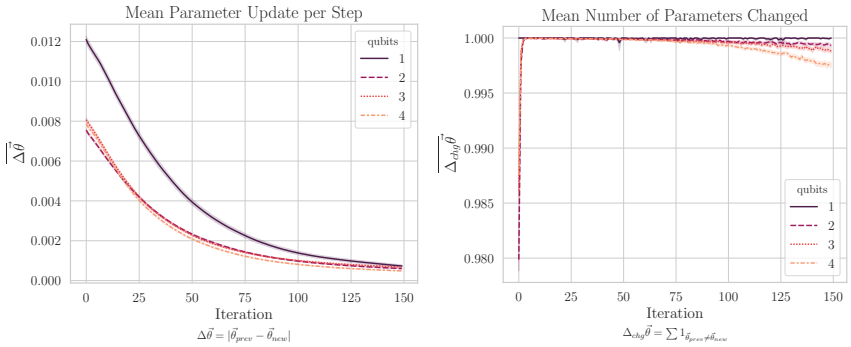
5 Experimental setup

Ansatzes We consider layered architectures in our experiments, each consisting of trainable *parameterized* and fixed *entangling* components. For the first, we use rotations around the x- (R_X), y- (R_Y), and z-axis (R_Z) of the Bloch sphere, and the controlled-NOT (CNOT) and controlled-Z (CZ) gates for the latter[7], constituting a widely-used gate set for HEAs. For a list of parameterized and entangling components, we refer to Table 1, and we run our experiments using all combinations.

We do not permute the operators, due to considerations on the expressivity per qubit. That is, if all rotations are applied, all permutations of $R_X(\alpha)R_Y(\beta)R_Z(\gamma)$, with $\alpha, \beta, \gamma \in [0, 2\pi]$ span the $SU(2)$, the set of unitary operations on one qubit. Therefore, permuting the operations does not affect expressivity per layer. Similarly, two rotation operations per layer explore a subspace of the $SU(2)$ owing to the Lie-algebraic closure relations, e.g., $R_X(\alpha)R_Y(\beta)$ spans the upper half of the Bloch sphere, as does $R_Y(\alpha)R_X(\beta)$. Any permutation allows the same operations per qubit, which is why we deem all permutations as equal.

Perturbation For the experiments, we uniformly sample parameters in range $[0, 2\pi]$ and choose a random 95% of them to perturb their values by $\pm t\%$ (in our case $[0.1, 0.5, 1]$). These values were in no way chosen randomly, rather, by training QNNs and collecting summary statistics on parameter updates, we found that these are the ranges where updates are performed. Then, we evaluate the channel sensitivity with the two obtained unitaries. We take 100 samples per parameter involved in the circuit and consider the distribution of the channel sensitivity we obtain for further analysis.

[7] We refer to Appendix A.1 for an overview of the operations.



(a) Mean absolute change (b) Amount of parameters changed

Figure 1: Training parameter updates

Training Moreover, we want to compare the channel sensitivity for random perturbations and training runs. Therefore, we train binary classification models on the two largest classes of the wine and breast cancer datasets and use PCA to reduce the features to 2^n, with n as the number of qubits. Then, the features are normalized and encoded into the amplitudes of the input state. The measurement is taken in the z-basis of the first qubit, and we use the mean squared error as the loss.

To account for favorable parameter initializations, we train each architecture 50 times. We use 150 iterations, and monitor convergence and parameter changes. We use the Adam optimizer (Kingma & Ba, 2015) from Pennylane with default parameters (step size: 0.01, beta 1: 0.9, beta 2: 0.99, epsilon: 1e-08). Further, we calculate the channel sensitivity in every iteration, to get an operationally meaningful analysis of how much the model changes. This also allows evaluating whether the model changes more when updated based on the loss function, rather than by just randomly permuting parameters. We display aggregated results using confidence intervals with a 95% confidence level.

Qubits Due to the computational complexity involved in solving the diamond norm, we run our experiments from one to four qubits and one to five layers. Unfortunately, due to matrix dimensions, it is challenging to scale the diamond norm any further. Nonetheless, the results are relevant, considering that they (1) provide a first step at analyzing small scale architectures, and (2) have direct implications for training dynamics that can be expected in the NISQ era.

Today's QNNs use shallow circuits with a depth in $O(\log\, n)$ and a local cost function, as they can be shown to be trainable (Cerezo et al., 2021b). Even if favorable loss landscapes were found for deeper circuits, adding layers results in noisier loss estimates, prohibiting effective learning. Moreover, scaling in number of qubits is limited due to quantum hardware limitations, or, when the models are simulated classically, to the curse of dimensionality. Therefore, many recent works proposing QML to solve a particular problem use few qubits, e.g., Blance & Spannowsky (2021), with two qubits, or Yano et al. (2020) with three and four qubits.

Further, our main goal is to take initial steps in quantifying and characterizing QNN model behavior. As has been mentioned before, due to the power of classical ML, we are unlikely to find an advantage in QML in the near term (Schuld & Killoran, 2022). Therefore, it would be important to focus on more fundamental questions, such as "how can we exploit quantum mechanics for ML purposes?", to explore the potential of quantum computing for data processing. In that sense, understanding how quantum channels change upon parameter updates is a foundational topic in ansatz design.

Framework Our experiments are implemented using the Python programming language. We use Pennylane (Bergholm et al., 2022) for obtaining the unitary transformations and QuTIP (Johansson et al., 2012; 2013) for calculating the diamond norm. Unfortunately, a diamond norm computation between two operators $\|A - B\|_\diamond$ with a finitely-large overlap $A^\dagger B \approx \mathbb{I}$, yields numerical instabilities with the QuTIP package. To account for these numerical issues, the operator can be multiplied with a global phase, which fixes a gauge (cf. *Gauge-Fixing*, in Appendix F).

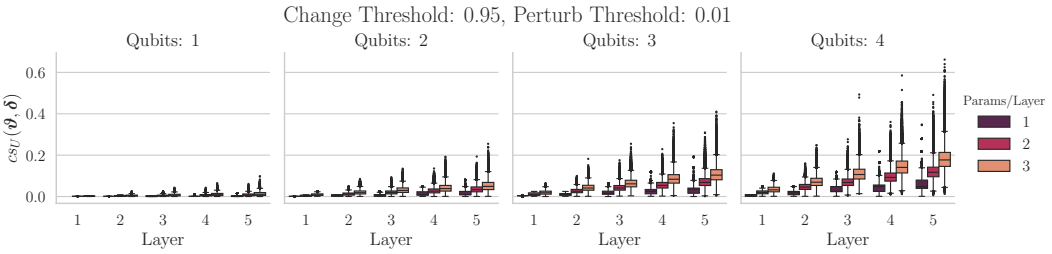
6 EXPERIMENTAL RESULTS

We run extensive numerical experiments to validate our bounds. First, we randomly perturb parameters, then we compare them to the channel sensitivity we observe during actual training runs.

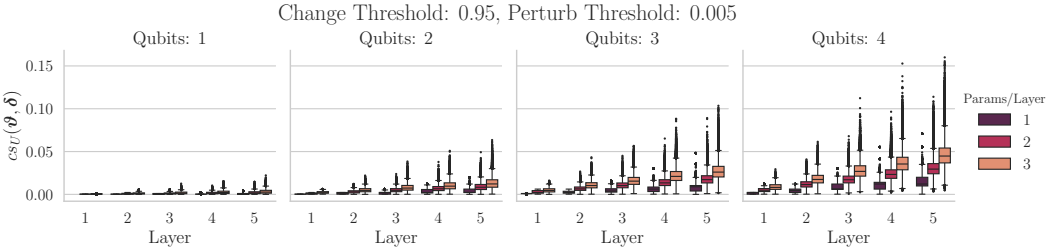
Initially, we verify the assumptions about the magnitude of parameter changes. In training runs, the mean parameter change per update in the first 10 iterations is maximum 1.4%, with a mean and standard deviation of $0.8\% \pm 0.4\%$, but it quickly decreases. We visualize this in Figure 1a, with the iterations on the x-axis, and the mean change in parameters on the y-axis. Architectures with the same number of qubits are aggregated and show very narrow confidence intervals. Further, it is also visible that changes in multi-qubit systems tend to be even smaller (less than two orders of magnitude). Moreover, we observe that almost all parameters are updated in every iteration, which is shown in Figure 1b, where the y-axis shows the percentage of parameters changed.

6.1 RANDOM PERTURBATION

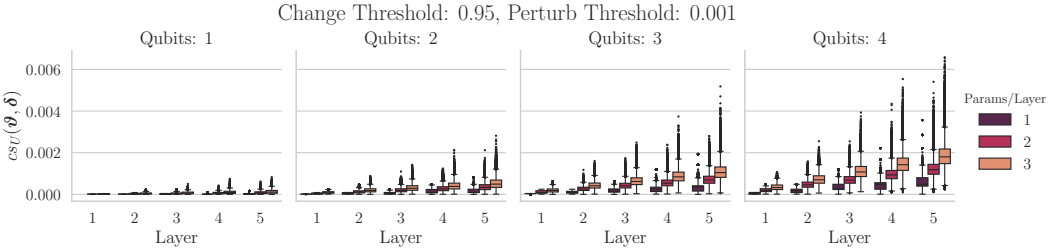
Based on these observations, we run the random perturbation experiments by changing 95% of parameters by 1%, 0.5%, and 0.1%. We visualize the results in Figure 2, showing the depth of the circuit on the x-axis and the obtained channel sensitivity on the y-axis. Per our bounds, the channel sensitivity strongly depends on the number of parameters in the circuit, hence, we visualize the results based on the number of parameters per layer.



(a) Change 95% of parameters by 1%



(b) Change 95% of parameters by 0.5%



(c) Change 95% of parameters by 0.01%

Figure 2: Channel sensitivity for random perturbations

The experiments show the same patterns as the bounds predict. In particular, fixing the qubits and parameters per layer while increasing the depth of the circuit, results in a bigger channel sensitivity. The same can be observed when fixing the number of layers and increasing the qubits. We observe

that increasing channel sensitivity is more prevalent for larger systems than for deeper circuits, e.g, architectures with 2 qubits, 5 layers and 3 parameters have a smaller channel sensitivity than architectures with 3 qubits, 5 layers and 2 parameters, despite having the same number of parameters.

The behaviour is expected for ansatzes covering large parts of the search space, as the Hilbert space of a larger system is significantly bigger, scaling as 2^n. This allows more independent search directions, therefore, potentially more changes in channel upon update. Deeper circuits can explore larger parts of the Hilbert space too, however, their expressivity stagnates when achieving overparameterization (Larocca et al., 2023). In this regime, more parameters will not result in more independent search directions, hence, the difference in channel sensitivity is expected to be smaller.

Further, it can be seen that there is a substantial degree of variability in the channel sensitivity. In the majority of experiments, the channels are hardly distinguishable, however, many outliers can be observed. This property could be particularly relevant for warm-starting (Puig et al., 2024), i.e., starting training from regions with high channel sensitivity might lead to smoother optimization.

6.2 Training updates

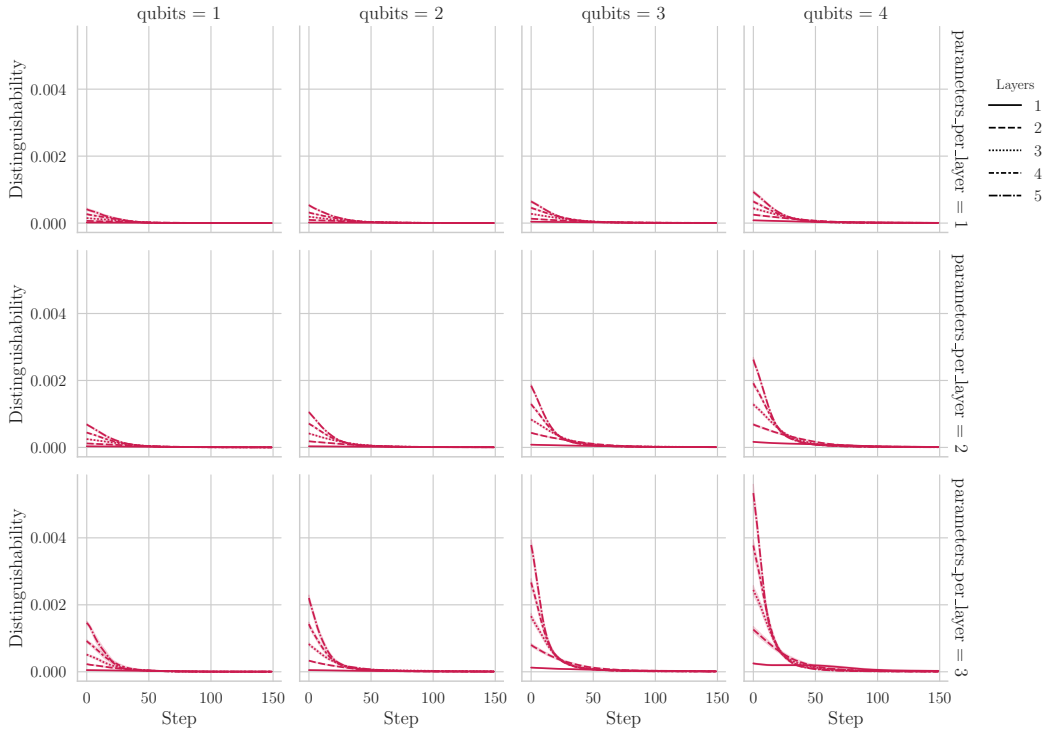


Figure 3: Channel sensitivity during training runs

Additionally, we want to consider how different the quantum channels are when updated based on loss and optimizer. In particular, it could be possible that changing the parameters during training, changes the underlying channel a lot more than perturbing in random directions. Upon analysis of our results, we can confirm the validity of our Taylor expansion even in early stages of iteration, i.e., in the $45,500$ models we trained, the bound holds for every single parameter update taken.

We plot the channel sensitivity in Figure 3 for different layers, qubits and parameters per layer. We can observe that the confidence intervals are very narrow, i.e., even though we collect data from 50 different runs for every model, the observable channel sensitivity hardly varies. Further, the channel sensitivity is substantially smaller than the bound, speaking to the severity of the issue. The discrepancy grows in particular with the number of qubits, which is visualized in Figure 4 in Appendix D. While we cannot observe the magnitude of channel sensitivity that the bound predicts, the pattern of scaling in the number of parameters can be observed. This is in particular visible in Figure 3 for

early training stages with larger parameter updates, however, it vanishes in later stages, when the updates are very small. Further, to showcase robustness with respect to the feature encoding (and different parameter updates that may result thereof), we run our experiments with angle encoding as well and verify that the magnitude of channel sensitivity does not change. We refer to Figure 5 in Appendix E for more information.

7 Discussion

Our bound establishes a direct dependence of the distinguishability of the channels during QNN training to the number of parameters. That is, the *maximum distinguishability of two output states* of the quantum channels, scales with (1) the magnitude of changes and (2) the number of parameters. As iterative updates with small learning rates are applied, which is independent of the number of parameters in the circuit, the dependence is largely dominated by the number of parameters.

This could significantly contribute to the trainability issues of today's QNNs. In particular, for reasonably sized models, the channels may be distinguishable in early training stages, but hardly distinguishable when fine-tuning the ansatz later. Comparison of our bound to the channel sensitivity obtained in training runs reveals that during training, the channels are even less distinguishable. While the channel sensitivity for random permutation is still considerably lower than our bounds would predict, a lot more variation can be observed.

This confirms our initial motivation of studying the model space independently of the loss and data, as these are global properties of ansatzes. In particular, we want to draw attention to the remarkable similarities to the BP phenomenon. It seems that the architectures significantly contribute to the untrainable models we observe today, independently of data and loss. Our findings have implications to start a more thorough investigation of the model space of the ansatzes that are currently used, and, ultimately, stress the need for good initialization and warm-starting. It is, however, non-obvious how to find such starting points without trial-and-error.

8 Conclusion

Altogether, our work provides a first step at exploring the model space of QNNs. In particular, we provide a picture of the local neighborhood of the model space of an ansatz, which suggest that the channels, at least for small scale circuits, do not differ significantly upon parameter update. Our results for random perturbations reveal that there is a substantial degree of variability of the channel sensitivity depending on the neighborhood, hence, stress the need to study the model space more thoroughly. This could reveal valuable information for parameter initialization or warm-starting.

Considering that QNNs were designed for NISQ technology, thus limited to shallow, small-scale circuits due to hardware limitations, the community will be limited to such small instances in the near-term. Therefore, our findings have direct implications on the meaningfulness of iteratively optimizing such circuits. While it is true that the models can be scaled to work for larger circuits, we have overwhelming evidence through works on BPs, that this is inherently difficult. Together with results on classical simulability and dequantizability in the absence of BPs, this work extends the evidence that we need a paradigm shift in Variational Quantum Computing or even QML altogether.

Reproducibility Statement

We provide the source code for all our experiments in an anonymized GitHub repository[8] with extensive documentation and listed dependencies. Upon acceptance, we will make the GitHub repository publicly available for convenient access.

References

Amira Abbas, David Sutter, Christa Zoufal, Aurelien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *Nature Computational Science*, 1(6):403–409, June

[8] https://anonymous.4open.science/r/distinguishability-qnn-5C41

2021. ISSN 2662-8457. doi: 10.1038/s43588-021-00084-1. URL http://dx.doi.org/10.1038/s43588-021-00084-1.

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

Andrew Arrasmith, Zoë Holmes, Marco Cerezo, and Patrick J Coles. Equivalence of quantum barren plateaus to cost concentration and narrow gorges. *Quantum Science and Technology*, 7(4): 045015, aug 2022. doi: 10.1088/2058-9565/ac7d06. URL https://dx.doi.org/10.1088/2058-9565/ac7d06.

Ingemar Bengtsson and Karol Zyczkowski. *Geometry of Quantum States: An Introduction to Quantum Entanglement*. Cambridge University Press, 2006.

Ville Bergholm, Josh Izaac, Maria Schuld, et al. Pennylane: Automatic differentiation of hybrid quantum-classical computations, 2022. URL https://arxiv.org/abs/1811.04968.

Matias Bilkis, Marco Cerezo, Guillaume Verdon, Patrick J Coles, and Lukasz Cincio. A semi-agnostic ansatz with variable structure for variational quantum algorithms. *Quantum Machine Intelligence*, 5(2):43, 2023.

Andrew Blance and Michael Spannowsky. Quantum machine learning for particle physics using a variational quantum classifier. *J. High Energ. Phys*, 2021. doi: https://doi.org/10.1007/JHEP02(2021)212.

Matthias C. Caro, Elies Gil-Fuster, Johannes Jakob Meyer, Jens Eisert, and Ryan Sweke. Encoding-dependent generalization bounds for parametrized quantum circuits. *Quantum*, 5:582, November 2021. ISSN 2521-327X. doi: 10.22331/q-2021-11-17-582. URL http://dx.doi.org/10.22331/q-2021-11-17-582.

Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, 9 2021a. ISSN 2522-5820. doi: 10.1038/s42254-021-00348-9.

Marco Cerezo, Akira Sone, Tyler Volkoff, et al. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nat Commun 12, 1791*, 2021b. doi: 10.1038/s41467-021-21728-w.

Marco Cerezo, Martin Larocca, Diego García-Martín, Nahuel L. Diaz, Paolo Braccia, Enrico Fontana, Manuel S. Rudolph, Pablo Bermejo, Aroosa Ijaz, Supanut Thanasilp, Eric R. Anschuetz, and Zoë Holmes. Does provable absence of barren plateaus imply classical simulability? or, why we need to rethink variational quantum computing, 2024.

Iris Cong, Soonwon Choi, and Mikhail D. Lukin. Quantum convolutional neural networks. *Nature Physics*, 15, 2019. doi: 10.1038/s41567-019-0648-8.

Christoph Dankert, Richard Cleve, Joseph Emerson, and Etera Livine. Exact and approximate unitary 2-designs and their application to fidelity estimation. *Phys. Rev. A*, 80:012304, Jul 2009. doi: 10.1103/PhysRevA.80.012304. URL https://link.aps.org/doi/10.1103/PhysRevA.80.012304.

Abhinav Deshpande, Marcel Hinsche, Sona Najafi, Kunal Sharma, Ryan Sweke, and Christa Zoufal. Dynamic parameterized quantum circuits: expressive and barren-plateau free, 2024. URL https://arxiv.org/abs/2411.05760.

Yuxuan Du, Zhuozhuo Tu, Xiao Yuan, and Dacheng Tao. Efficient measure for the expressivity of variational quantum algorithms. *Phys. Rev. Lett.*, 128:080506, Feb 2022. doi: 10.1103/PhysRevLett.128.080506. URL https://link.aps.org/doi/10.1103/PhysRevLett.128.080506.

Jacob Dumford and Walter Scheirer. Backdooring convolutional neural networks via targeted weight perturbations. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–9, 2020. doi: 10.1109/IJCB48548.2020.9304875.

Thomas Fel, Melanie Ducoffe, David Vigouroux, Rémi Cadène, Mikaël Capelle, Claire Nicodème, and Thomas Serre. Don't lie to me! robust and efficient explainability with verified perturbation analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16153–16163, June 2023.

Elies Gil-Fuster, Casper Gyurik, Adrián Pérez-Salinas, and Vedran Dunjko. On the relation between trainability and dequantization of variational quantum learning models, 2024. URL https://arxiv.org/abs/2406.07072.

Edward Grant, Leonard Wossnig, Mateusz Ostaszewski, and Marcello Benedetti. An initialization strategy for addressing barren plateaus in parametrized quantum circuits. *Quantum*, 3:214, 12 2019. ISSN 2521-327X. doi: 10.22331/q-2019-12-09-214.

D. Gross, K. Audenaert, and J. Eisert. Evenly distributed unitaries: On the structure of unitary designs. *Journal of Mathematical Physics*, 48(5), may 2007. ISSN 1089-7658. doi: 10.1063/1.2716992. URL http://dx.doi.org/10.1063/1.2716992.

Andrew W. Harrow and Richard A. Low. Random quantum circuits are approximate 2-designs. *Commun. Math. Phys. 291*, pp. 257–302, 2009. doi: 10.1007/s00220-009-0873-6.

Zhezhi He, Adnan Siraj Rakin, and Deliang Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Sabrina Herbst, Vincenzo De Maio, and Ivona Brandic. On optimizing hyperparameters for quantum neural networks, 2024. URL https://arxiv.org/abs/2403.18579.

S.G. Hoggar. t-designs in projective spaces. *European Journal of Combinatorics*, 3(3):233–254, 1982. ISSN 0195-6698. doi: https://doi.org/10.1016/S0195-6698(82)80035-8. URL https://www.sciencedirect.com/science/article/pii/S0195669882800358.

Zoë Holmes, Kunal Sharma, M. Cerezo, and Patrick J. Coles. Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *PRX Quantum*, 3:010313, 1 2022. doi: 10.1103/PRXQuantum.3.010313.

Roger A. Horn and Charles R. Johnson. *Norms for vectors and matrices*, pp. 257–342. Cambridge University Press, 1985.

Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234, 2021. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2021.06.030. URL https://www.sciencedirect.com/science/article/pii/S0167865521002440.

J. Robert Johansson, Paul D. Nation, and Franco Nori. Qutip: An open-source python framework for the dynamics of open quantum systems. *Computer Physics Communications*, 183(8):1760–1772, 2012. ISSN 0010-4655. doi: https://doi.org/10.1016/j.cpc.2012.02.021. URL https://www.sciencedirect.com/science/article/pii/S0010465512000835.

J. Robert Johansson, Paul D. Nation, and Franco Nori. Qutip 2: A python framework for the dynamics of open quantum systems. *Computer Physics Communications*, 184(4):1234–1240, 2013. ISSN 0010-4655. doi: https://doi.org/10.1016/j.cpc.2012.11.019. URL https://www.sciencedirect.com/science/article/pii/S0010465512003955.

A. Kandala, A. Mezzacapo, K. Temme, et al. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, pp. 242–246, 2017. doi: 10.1038/nature23879.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *3rd International Conference for Learning Representations*, 2015. doi: 10.48550/arXiv.1412.6980.

Alexei Yu Kitaev. Quantum computations: algorithms and error correction. *Russian Mathematical Surveys*, 52(6):1191, dec 1997. doi: 10.1070/RM1997v052n06ABEH002155.

Alexei Yu Kitaev, Alexander H Šen', and Mikhail N Vyalyi. *Classical and quantum computation*. Graduate studies in mathematics. American Math. Soc., Providence, RI, 2002. ISBN 0821832298.

Ankit Kulshrestha and Ilya Safro. Beinit: Avoiding barren plateaus in variational quantum algorithms. In *2022 IEEE International Conference on Quantum Computing and Engineering (QCE)*, pp. 197–203, 2022. doi: 10.1109/QCE53715.2022.00039.

Martín Larocca, Frédéric Sauvage, Faris M. Sbahi, Guillaume Verdon, Patrick J. Coles, and Marco Cerezo. Group-invariant quantum machine learning. *PRX Quantum*, 3:030341, Sep 2022. doi: 10.1103/PRXQuantum.3.030341. URL https://link.aps.org/doi/10.1103/PRXQuantum.3.030341.

Martin Larocca, Nathan Ju, Diego. García-Martín, et al. Theory of overparametrization in quantum neural networks. *Nat Comput Sci*, pp. 542 – 551, 2023. doi: 10.1038/s43588-023-00467-6.

Martin Larocca, Supanut Thanasilp, Samson Wang, Kunal Sharma, Jacob Biamonte, Patrick J. Coles, Lukasz Cincio, Jarrod R. McClean, Zoë Holmes, and Marco Cerezo. A review of barren plateaus in variational quantum computing, 2024. URL https://arxiv.org/abs/2405.00781.

Lorenzo Leone, Salvatore F.E. Oliviero, Lukasz Cincio, and Marco Cerezo. On the practical usefulness of the Hardware Efficient Ansatz. *Quantum*, 8:1395, July 2024. ISSN 2521-327X. doi: 10.22331/q-2024-07-03-1395. URL https://doi.org/10.22331/q-2024-07-03-1395.

Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9(1): 4812, 11 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07090-4.

Antonio Anna Mele. Introduction to Haar Measure Tools in Quantum Information: A Beginner's Tutorial. *Quantum*, 8:1340, May 2024. ISSN 2521-327X. doi: 10.22331/q-2024-05-08-1340. URL https://doi.org/10.22331/q-2024-05-08-1340.

Kouhei Nakaji and Naoki Yamamoto. Expressibility of the alternating layered ansatz for quantum computation. *Quantum*, 5:434, April 2021. ISSN 2521-327X. doi: 10.22331/q-2021-04-19-434. URL https://doi.org/10.22331/q-2021-04-19-434.

Carlos Ortiz Marrero, Mária Kieferová, and Nathan Wiebe. Entanglement-induced barren plateaus. *PRX Quantum*, 2:040316, 10 2021. doi: 10.1103/PRXQuantum.2.040316.

Jaime Pizarroso, José Portela, and Antonio Muñoz. Neuralsens: Sensitivity analysis of neural networks. *Journal of Statistical Software*, 102(7):1–36, 2022. doi: 10.18637/jss.v102.i07. URL https://www.jstatsoft.org/index.php/jss/article/view/v102i07.

John Preskill. Quantum computing in the nisq era and beyond. *Quantum*, 2:79, 2018.

Ricard Puig, Marc Drudis, Supanut Thanasilp, and Zoë Holmes. Variational quantum simulation: a case study for understanding warm starts, 2024. URL https://arxiv.org/abs/2404.10044.

Ali Rad, Alireza Seif, and Norbert M. Linke. Surviving the barren plateau in variational quantum circuits with bayesian learning initialization, 2022.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl Dickstein. On the expressive power of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 2847–2854. JMLR.org, 2017.

Michael Ragone, Paolo Braccia, Quynh T. Nguyen, Louis Schatzki, Patrick J. Coles, Frederic Sauvage, Martin Larocca, and M. Cerezo. Representation theory for geometric quantum machine learning, 2023. URL https://arxiv.org/abs/2210.07980.

Daniel Reed, Dennis Gannon, and Jack Dongarra. Reinventing high performance computing: Challenges and opportunities, 2022.

Daniel A. Roberts and Beni Yoshida. Chaos and complexity by design. *Journal of High Energy Physics*, 2017(4), April 2017. ISSN 1029-8479. doi: 10.1007/jhep04(2017)121. URL http://dx.doi.org/10.1007/JHEP04(2017)121.

Maria Schuld and Nathan Killoran. Is quantum advantage the right goal for quantum machine learning? *PRX Quantum*, 3(3), July 2022. ISSN 2691-3399. doi: 10.1103/prxquantum.3.030101. URL http://dx.doi.org/10.1103/PRXQuantum.3.030101.

Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3), March 2021. ISSN 2469-9934. doi: 10.1103/physreva.103.032430. URL http://dx.doi.org/10.1103/PhysRevA.103.032430.

Martin Schulz, Dieter Kranzlmüller, Laura Brandon Schulz, Carsten Trinitis, and Josef Weidendorfer. On the inevitability of integrated hpc systems and how they will change hpc system operations. In *Proceedings of the 11th International Symposium on Highly Efficient Accelerators and Reconfigurable Technologies*, HEART '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385497.

A J Scott. Optimizing quantum process tomography with unitary 2-designs. *Journal of Physics A: Mathematical and Theoretical*, 41(5):055308, jan 2008. doi: 10.1088/1751-8113/41/5/055308. URL https://dx.doi.org/10.1088/1751-8113/41/5/055308.

John Shalf. The future of computing beyond moore's law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 378(2166):20190061, 2020. doi: 10.1098/rsta.2019.0061.

Sukin Sim, Peter D. Johnson, and Alán Aspuru-Guzik. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Advanced Quantum Technologies*, 2(12), October 2019. ISSN 2511-9044. doi: 10.1002/qute.201900070. URL http://dx.doi.org/10.1002/qute.201900070.

Lucia Testa, Claudio Battiloro, Stefania Sardellitti, and Sergio Barbarossa. Stability of graph convolutional neural networks through the lens of small perturbation analysis. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6865–6869, 2024. doi: 10.1109/ICASSP48485.2024.10447343.

Cenk Tüysüz, Su Yeon Chang, Maria Demidik, Karl Jansen, Sofia Vallecorsa, and Michele Grossi. Symmetry breaking in geometric quantum machine learning in the presence of noise. *PRX Quantum*, 5:030314, Jul 2024. doi: 10.1103/PRXQuantum.5.030314. URL https://link.aps.org/doi/10.1103/PRXQuantum.5.030314.

A Vaswani et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

S. Wang, E. Fontana, M. Cerezo, et al. Noise-induced barren plateaus in variational quantum algorithms. *Nat Commun 12, 6961*, 2021. doi: 10.1038/s41467-021-27045-6.

Yabo Wang, Bo Qi, Chris Ferrie, and Daoyi Dong. Trainability enhancement of parameterized quantum circuits via reduced-domain parameter initialization, 2023.

John Watrous. Semidefinite programs for completely bounded norms. *Theory of Computing*, 5(11): 217–238, 2009. doi: 10.4086/toc.2009.v005a011.

John Watrous. Simpler semidefinite programs for completely bounded norms. *Chicago Journal of Theoretical Computer Science*, 2013(8), July 2013.

John Watrous. *The Theory of Quantum Information*. Cambridge University Press, 2018.

L. Welch. Lower bounds on the maximum cross correlation of signals (corresp.). *IEEE Transactions on Information Theory*, 20(3):397–399, 1974. doi: 10.1109/TIT.1974.1055219.

Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJNpifWAb.

Roeland Wiersema, Alexander F. Kemper, Bojko N. Bakalov, and Nathan Killoran. Geometric quantum machine learning with horizontal quantum gates, 2024. URL https://arxiv.org/abs/2406.04418.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2958–2969. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1ef91c212e30e14bf125e9374262401f-Paper.pdf.

Hiroshi Yano, Yudai Suzuki, Rudy Raymond, and Naoki Yamamoto. Efficient discrete feature encoding for variational quantum classifier. In *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)*, pp. 11–21, 2020. doi: 10.1109/QCE49297.2020.00012.

Shikun Zhang, Zheng Qin, Yang Zhou, Rui Li, Chunxiao Du, and Zhisong Xiao. Single entanglement connection architecture between multi-layer bipartite hardware efficient ansatz. *New Journal of Physics*, 26(7):073042, jul 2024. doi: 10.1088/1367-2630/ad64fb. URL https://dx.doi.org/10.1088/1367-2630/ad64fb.

A Additional Background

A.1 Quantum Computing

Quantum Computing (QC) works with quantum bits (qubits). While classical bits take either the value 0 or 1, qubits are in a so-called *superposition* of the two states, meaning they are in both states at once, which is shown mathematically in Equation 8, where α and β are called *probability amplitudes*. We can do calculations with the superposed qubit, however, once we read out results, we have to take a *measurement*, where the qubit will collapse into state $|0\rangle$ with probability $|\alpha|^2$, and into state $|1\rangle$ with probability $|\beta|^2$. Due to this inherent non-determinism in QC, executions are always done multiple times.

$$|\psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle \quad \text{s.t. } |\alpha|^2 + |\beta|^2 = 1 \quad \alpha, \beta \in \mathbb{C} \tag{8}$$

A qubit is characterized by 4 numbers (the real and imaginary parts of α and β), which, when translated into polar coordinates, gives a convenient geometrical representation on the so-called *Bloch sphere*. This plays a significant part in today's QNNs, as the trainable gates are rotations of the individual qubits around the x-, y- and z-axis of this sphere.

Multiple qubits form a *quantum register*, and its associated quantum state represents 2^n states at once, where n is the number of qubits, which can make QC very powerful. A quantum state can be *entangled*, meaning that individual qubits can be correlated, such that an action on one qubit affects the ones it is entangled with as well.

Quantum states are manipulated through *unitary transformations* or *quantum gates* ($U^\dagger U = \mathbb{I}$, where $U^\dagger$ is the adjoint of U). Unitary transformations are linear and preserve vector lengths, i.e., applying a unitary to a quantum state ensures that the squared magnitudes of the probability amplitudes still sum up to one. Every unitary is generated by a *Hermitian generator* ($H = H^\dagger$), i.e., $U = e^{-iH}$. We describe the action on an initial quantum state as a so-called *unitary evolution*, which may involve one or more unitary operations (unitary matrices form a group, hence, are closed under multiplication). We also refer to this as a *quantum channel* in the following. While the term quantum channel encompasses unitary evolutions, it is a broader term, meaning it can also describe more general dynamics of quantum systems, such as decoherence or noise. It is defined as a linear map that is completely positive and trace preserving (Watrous, 2018).

Among the most fundamental operations in QC, are the so-called Pauli matrices, which are Hermitian matrices shown in Equation 9. σ_x applies a NOT (bit-flip) operation ($|0\rangle \to |1\rangle$, $|1\rangle \to |0\rangle$), σ_y changes the phase and bit ($|0\rangle \to i\,|1\rangle$, $|1\rangle \to -i\,|0\rangle$), and σ_z applies a phase-flip ($|0\rangle \to |0\rangle$, $|1\rangle \to -1\,|1\rangle$).

$$\sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \quad \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{9}$$

Taking the Pauli matrices as generators, we can rotate a qubit around the x-, y- and z-axis around the Bloch sphere, arriving at the definition of the R_X, R_Y and R_Z rotation in Equation 10.

$$\mathrm{R_X}(\theta) = e^{-i\frac{\theta}{2}\sigma_x} \quad R_Y(\theta) = e^{-i\frac{\theta}{2}\sigma_y} \quad R_Z(\theta) = e^{-i\frac{\theta}{2}\sigma_z} \tag{10}$$

Further, we define the two entangling operations we use, `CNOT` (or `CX`) and `CZ`, in Equation 11. The operators entangle two qubits i and j, where i is referred to as the *control* qubit and j is the *target* qubit. In general terms, the entangling gates apply σ_x or σ_z to the target qubit if the control is in state $|1\rangle$.

$$\mathrm{CNOT} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad \mathrm{CZ} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \tag{11}$$

A.2 Inner-product & overlap between vectors

An inner-product space is defined as a vector space V over a field $\mathbb{F}$ (typically $\mathbb{R}$ or $\mathbb{C}$) endowed with an inner-product map $\langle .|. \rangle$:

$$\langle .|. \rangle : V \times V \to \mathbb{F} \tag{12}$$

The inner-product operation is a generalization of the dot-product between vectors and measures the **overlap** or **coherence** between vectors sampled from a vector space.

A.3 Unitary groups and the Haar Measure

We will provide key information on unitary groups and the Haar measure in this section. For an excellent tutorial on the topic, we refer to Mele (2024), which we also base this section on. We begin by defining the unitary group $\mathcal{U}(d)$ and the special unitary group $S\mathcal{U}(d)$, which are key concepts when studying quantum computing, as follows.

Definition 3 *For $d \in \mathbb{N}$, the **unitary group** $\mathcal{U}(d)$ is the group of isometries of d-dimensional complex Hilbert space $\mathbb{C}^d$. These are canonically identified with the $d \times d$ unitary matrices ($\mathbb{C}^{d\times d}$):*

$$\mathcal{U}(d) = \left\{ \mathcal{V} \in \mathbb{C}^{d\times d} | \mathcal{V} \cdot \mathcal{V}^\dagger = \mathcal{V}^\dagger \cdot \mathcal{V} = \mathbb{I}_d \right\} \tag{13}$$

The unitary group can be decomposed as $\mathcal{U}(d) = S\mathcal{U}(d) \times U(1)$. Here, the subgroup $S\mathcal{U}(d)$ is called the *special unitary* group, while $z = e^{i\theta}\ \forall z \in U(1)$ is a phase restricted on a unit-circle.

Definition 4 *For $d \in \mathbb{N}$, the **special unitary group** $S\mathcal{U}(d) \subset \mathcal{U}(d)$ are a group of $d \times d$ unitary matrices ($\mathbb{C}^{d\times d}$) that have a unit determinant:*

$$S\mathcal{U}(d) = \left\{ \mathcal{G} \in \mathcal{U}(d) |\ det(G) = 1 \right\} \tag{14}$$

The Haar measure forms a uniform probability distribution over sets of unitary matrices, in fact, it is unique for compact groups, such as the $\mathcal{U}(d)$.

Definition 5 *(Mele, 2024) We define the **Haar Measure** on the $\mathcal{U}(d)$, as the left and right invariant probability measure μ_H over the group. That is, for all integrable functions f and $\forall V \in \mathcal{U}(d)$:*

$$\int_{\mathcal{U}(d)} f(U) d\mu_H(U) = \int_{\mathcal{U}(d)} f(UV) d\mu_H(U) = \int_{\mathcal{U}(d)} f(VU) d\mu_H(U) \tag{15}$$

Thus, the Haar measure assigns an invariant volume measure to subsets of locally compact topological groups. Due to properties of a probability measure, it holds that $\int_S d\mu_H(U) \geq 0, \forall S \subseteq \mathcal{U}(d)$ and

$\int_{\mathcal{U}(d)} 1 d\mu_H(U) = 1$. It is therefore possible to set the expectation value with respect to the probability measure equal to the integral over the Haar measure as $\mathbb{E}_{U\sim\mu_H}[f(U)] = \int_{\mathcal{U}(d)} f(U) d\mu_H(U)$. This leads to the definition of the moment operator of the Haar measure.

Definition 6 *(Mele, 2024) For all operators O, the **t-th moment operator** based on probability measure μ_H is defined as follows.*

$$M_{\mu_H}^{(t)}(O) = \mathbb{E}_{U\sim\mu_H}[U^{\otimes t} O U^{\dagger\otimes t}] \tag{16}$$

Definition 7 *(Mele, 2024) A distribution v over a set of unitaries $S \subseteq \mathcal{U}(d)$ is a **unitary t-design** if and only if the following holds for all operators O.*

$$\mathbb{E}_{V\sim v}[V^{\otimes t} O V^{\dagger\otimes t}] = \mathbb{E}_{U\sim\mu_H}[U^{\otimes t} O U^{\dagger\otimes t}] \tag{17}$$

State t-designs are weaker versions of unitary t-designs in the sense that unitary t-designs induce state t-designs when they act on a fixed input state. These are defined as follows.

Definition 8 *(Mele, 2024) A **state t-design** is a probability distribution η over a set of states S if and only if the following holds.*

$$\mathbb{E}_{|\psi\rangle\sim\eta}[|\psi\rangle\langle\psi|^{\otimes k}] = \mathbb{E}_{|\psi\rangle\sim\mu_H}[|\psi\rangle\langle\psi|^{\otimes k}] \tag{18}$$

Further, we define an equivalent definition of state t-designs as follows, which will be used as well.

Definition 9 *(Hoggar, 1982) Let $\mathbb{C}^d$ denote a d-dimensional Hilbert space with orthonormal basis set $\{|n_k\rangle\}_{k=1}^d$. Owing to their redundancy under global phase transformations and normalization, the quantum states in this space correspond to points in the complex-projective space $\mathbb{CP}^{d-1}$ (Bengtsson & Zyczkowski, 2006). Thus, we define the nontrivial complex-projective t-design as a set of states $S \subsetneq \mathbb{CP}^{d-1}$, sampled according to some probability measure μ, satisfying,*

$$\mathbb{E}_{|\psi\rangle\in S} f(|\psi\rangle) = \int_{\mathbb{CP}^{d-1}} f(|\psi\rangle) d\psi. \tag{19}$$

Where, $f(|\psi\rangle)$ is a polynomial of at most degree t in its amplitudes and complex-amplitudes of $|\psi\rangle$ respectively. The canonical measure $d\psi$ defined on the set of such quantum states, is the unique unit-normalized volume measure invariant under unitary group action $\mathcal{U}(d)$.

Often it is not strictly required to have an exact t-design, but rather a distribution close to a t-design, which is termed ϵ-approximate t-design.

A.4 Choice of polynomial function for state t-designs:

The function $f(|\psi\rangle)$ is a polynomial of at most degree t. For example, 2-designs, correspond to the average value of a quadratic (second-degree polynomial) function. Hence, a common choice of state 2-design polynomial functions is the overlap (cf. Appendix A.2 for definition) between a collection of states $\{\boldsymbol{\psi_i}\}_{1\leq i\leq d}$ in the Hilbert space $|\langle\boldsymbol{\psi_j}|\boldsymbol{\psi_k}\rangle|^2$. Similarly, for an even $t > 2$, a canonical choice of the polynomial function f is $|\langle\boldsymbol{\psi_j}|\boldsymbol{\psi_k}\rangle|^{2t}$.

B Welch bounds and t-designs

Theorem 1: **(Welch bounds)** Let $n \geq d$. If $\{\boldsymbol{v_i}\}_{1\leq i\leq n}$ is a sequence of unit vectors in $\mathbb{C}^d$, then,

$$\max_{1\leq j,k\leq n, j\neq k} |\langle\boldsymbol{v}_j|\boldsymbol{v}_k\rangle|^{2t} \geq \frac{1}{n-1}\left[\frac{n}{\binom{d+t-1}{t}} - 1\right],\ \forall t \in \mathbb{N}, \tag{20a}$$

implies,

$$\sum_{1\leq j,k\leq n} |\langle \boldsymbol{v}_j|\boldsymbol{v}_k\rangle|^{2t} \geq \frac{n^2}{\binom{d+t-1}{t}} \forall t \in \mathbb{N}. \tag{20b}$$

The combinatorial (binomial) factor featuring in the denominator of Equation 20a and Equation 20b can be expanded in terms of the factorials:

$$\binom{d+t-1}{t} = \frac{(d+t-1)!}{(d-1)!\,t!} \tag{21}$$

Here ! corresponds to the factorial symbol, i.e. $n! = n(n-1)(n-2)...1$.

B.1 Simplification of Equation 20a:

Substituting Equation 21 into Equation 20a yields the following.

$$\max_{1\leq j,k\leq n, j\neq k} |\langle \boldsymbol{v}_j|\boldsymbol{v}_k\rangle|^{2t} \geq \frac{1}{n-1}\Big[\frac{n}{\frac{(d+t-1)!}{(d-1)!\,t!}} - 1\Big] = \frac{1}{n-1}\Big[\frac{n}{\frac{(d+t-1)(d+t-2)....d}{t!}} - 1\Big] \tag{22}$$

Here in the denominator, we have used the factorial property, $(d+t-1)!/(d-1)! = (d+t-1)(d+t-2)...d(d-1)!/(d-1)! = (d+t-1)(d+t-2)...d$.

B.2 Simplification of Equation 20b:

Using Equation 21 it follows that:

$$\sum_{1\leq j,k\leq n} |\langle \boldsymbol{v}_j|\boldsymbol{v}_k\rangle|^{2t} \geq \frac{n^2}{\frac{(d+t-1)(d+t-2)...d}{t!}}. \tag{23}$$

For the special case, $t = 2$, we get the inequality for **2-designs**,

$$\sum_{1\leq j,k\leq n} |\langle \boldsymbol{v}_j|\boldsymbol{v}_k\rangle|^{4} \geq \frac{n^2}{\binom{d+1}{2}} = \frac{n^2}{(d+1)d/2}. \tag{24}$$

C Channel Sensitivity Bound

C.1 Preliminaries

Before deriving the bound, we want to state some preliminaries for partial derivatives of parameterized unitaries. Equation 25 shows the partial derivative of a circuit (Holmes et al., 2022) which utilizes the convenient representation of a unitary in terms of its generator. The partial derivative essentially splits the original unitary in two parts, and we define the fractions in Equation 26.

$$\partial_j U(\boldsymbol{\vartheta}) = -\frac{i}{2} U_{j+1\to dim(\boldsymbol{\vartheta})} H_j U_{1\to j} \tag{25}$$

$$\begin{aligned} U_{l\to m}(\boldsymbol{\vartheta}) &= \prod_{j=l}^{m} V_j(\vartheta_j) W_j \\ &= \prod_{j=l}^{m} e^{-i\frac{\vartheta_j}{2}H_j} W_j \end{aligned} \tag{26}$$

C.2 Derivation of the Bound

We can now analyze the difference in diamond norm upon slight perturbation of parameters in Equation 27. We use a Taylor expansion and truncate after the first order. We arrive at Equation 28 after applying the partial derivative from Equation 25. We use the triangle inequality and homogeneity, which are both necessary conditions for matrix norms (Horn & Johnson, 1985), to arrive at Equation 29 and Equation 30 respectively. Arriving at Equation 31 is non-trivial, as, in general, it is

not guaranteed that multiplying hermitian and unitary matrices results in a unitary. However, HEAs use R_X, R_Y and R_Z rotations with generators X, Y and Z respectively, which are known to be hermitian and unitary. Since unitaries form a group that is closed under multiplication, it forms a valid quantum channel, which evaluates to 1(Watrous, 2018, Proposition 3.44).

$$\|U(\boldsymbol{\vartheta}) - U(\boldsymbol{\vartheta}+\boldsymbol{\delta})\|_\diamond = \|U(\boldsymbol{\vartheta}) - \left(U(\boldsymbol{\vartheta}) + \sum_{j=1}^{dim(\boldsymbol{\delta})} \delta_j \frac{\partial U(\boldsymbol{\vartheta})}{\partial \vartheta_j} + O(\boldsymbol{\delta}^2)\right)\|_\diamond \tag{27}$$

$$\approx \|U(\boldsymbol{\vartheta}) - (U(\boldsymbol{\vartheta}) + \sum_{j=1}^{dim(\boldsymbol{\delta})} \frac{\partial U(\boldsymbol{\vartheta})}{\partial \theta_j}\delta_j)\|_\diamond \tag{28}$$

$$= \| - \sum_{j=1}^{dim(\boldsymbol{\delta})} \frac{\partial U(\boldsymbol{\vartheta})}{\partial \theta_j}\delta_j\|_\diamond \tag{29}$$

$$= \| - \sum_{j=1}^{dim(\boldsymbol{\delta})} -\frac{i}{2} U_{j+1\to dim(\boldsymbol{\delta})} H_j U_{1\to j}\delta_j\|_\diamond \tag{30}$$

$$\leq \sum_{j=1}^{dim(\boldsymbol{\delta})} \|\frac{i}{2} U_{j+1\to dim(\boldsymbol{\delta})} H_j U_{1\to j}\delta_j\|_\diamond \tag{31}$$

$$= \sum_{j=1}^{dim(\boldsymbol{\delta})} |\frac{i}{2}\delta_j| \|U_{j+1\to dim(\boldsymbol{\delta})} H_j U_{1\to j}\|_\diamond \tag{32}$$

$$= \frac{\sum_{j=1}^{dim(\boldsymbol{\delta})} |\delta_j|}{2} \tag{33}$$

D Comparison of empirical channel sensitivity and bound

We compare the empirical channel sensitivity with the predicted bound in Figure 4. We add this figure to visualize that experimentally we can observe a large discrepancy, although the magnitude of this discrepancy makes it hard to read much more from the plot. We also would like to remark that one should be careful in drawing the conclusion that the bound is thus loose, as there might still be a point in the model space where taking such a step could achieve the bound. Our extensive experiments, however, reveal that this will likely not be the case in large areas of the model space.

E Angle Encoding

We visualize the channel sensitivity when encoding the data with angle embedding. We run these experiments to demonstrate robustness of the channel sensitivity with respect to the changed parameter updates that may occur due the different input states. While the channel sensitivity slightly varies, we can observe that its magnitude does not differ when comparing it to the results using amplitude embedding (we observe a maximum distinguishability of approximately 0.5% in both cases).

F Gauge Fixing

Considering the numerical instabilities whenever $A^\dagger B \approx \mathbb{I}$, we apply a Gauge-Fixing algorithm. In particular, we adjust the global phases of the two unitaries, which ensures numerical stability without affecting the diamond norm.

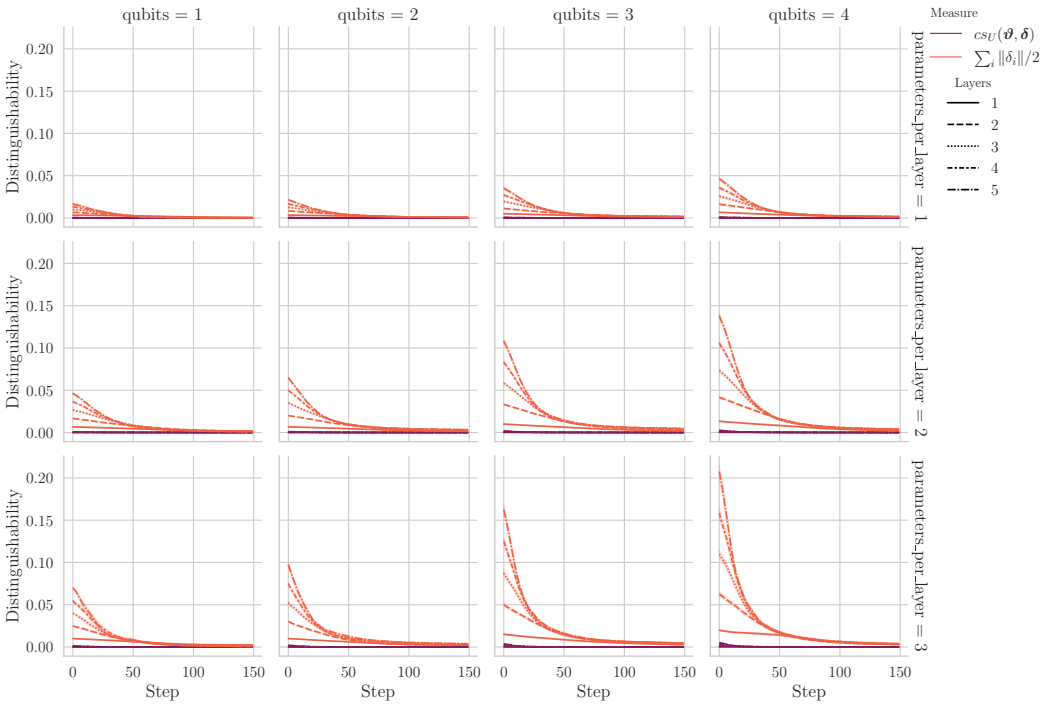


Figure 4: Comparison $\frac{\sum_j |\delta_j|}{2}$ and channel sensitivity

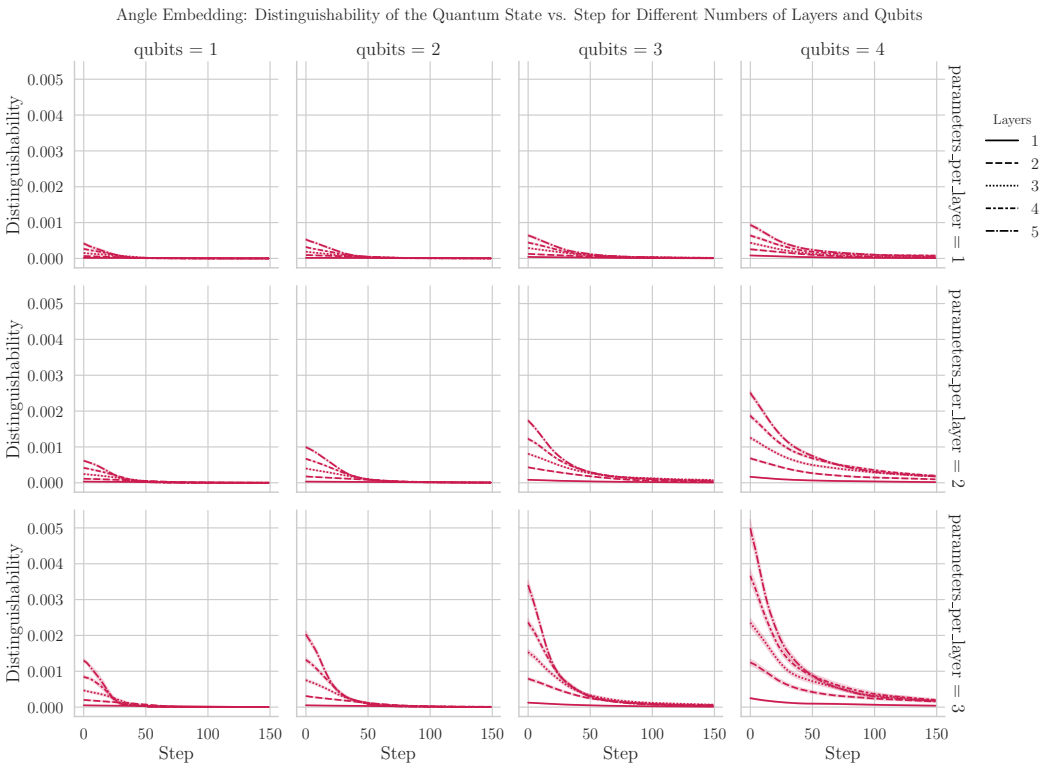


Figure 5: Channel sensitivity for angle embedding

Algorithm 1 Gauge-fixed diamond norm between difference of quantum channels

function DIAMOND NORM($\|\boldsymbol{A} - \boldsymbol{B}\|_\diamond$)
  if $\boldsymbol{A}^\dagger \boldsymbol{B} \approx \mathbb{I}$ **then**
    $\boldsymbol{z} \in \mathbb{C} \leftarrow \mathrm{Tr}(\boldsymbol{A}^\dagger \boldsymbol{B})$
    $\vartheta^* \in \mathbb{R} \leftarrow \arctan[\frac{\mathrm{Im}(\boldsymbol{z})}{\mathrm{Re}(\boldsymbol{z})}]$
    $\boldsymbol{B} \leftarrow \exp(i\vartheta^*)\boldsymbol{B}$
  end if
  return $\|\boldsymbol{A} - \boldsymbol{B}\|_\diamond$
end function