Analyzing Convergence in Quantum Neural Networks: Deviations from Neural Tangent Kernels

Xuchen You [1] **Shouvanik Chakrabarti** [2] **Boyang Chen** [3] **Xiaodi Wu** [1]

Abstract

A quantum neural network (QNN) is a parameterized mapping efficiently implementable on near-term Noisy Intermediate-Scale Quantum (NISQ) computers. It can be used for supervised learning when combined with classical gradient-based optimizers. Despite the existing empirical and theoretical investigations, the convergence of QNN training is not fully understood. Inspired by the success of the neural tangent kernels (NTKs) in probing into the dynamics of classical neural networks, a recent line of works proposes to study over-parameterized QNNs by examining a quantum version of tangent kernels. In this work, we study the dynamics of QNNs and show that contrary to popular belief it is qualitatively different from that of any kernel regression: due to the unitarity of quantum operations, there is a non-negligible deviation from the tangent kernel regression derived at the random initialization. As a result of the deviation, we prove the at-most sub-linear convergence for QNNs with Pauli measurements, which is beyond the explanatory power of any kernel regression dynamics. We then present the actual dynamics of QNNs in the limit of over-parameterization. The new dynamics capture the change of convergence rate during training, and implies that the range of measurements is crucial to the fast QNN convergence.

1. Introduction

Analogous to the classical logic gates, quantum gates are the basic building blocks for quantum computing. A variational quantum circuit (also referred to as an ansatz) is composed of parameterized quantum gates. A quantum neural network (QNN) is nothing but an instantiation of learning with parametric models using variational quantum circuits and quantum measurements: A p-parameter d-dimensional QNN for a dataset $\{\mathbf{x}_i, y_i\}$ is specified by an encoding $\mathbf{x}_i \mapsto \boldsymbol{\rho}_i$ of the feature vectors into quantum states in an underlying d-dimensional Hilbert space $\mathcal{H}$, a variational circuit $\mathbf{U}(\boldsymbol{\theta})$ with real parameters $\boldsymbol{\theta} \in \mathbb{R}^p$, and a quantum measurement $\mathbf{M}_0$. The predicted output $\hat{y}_i$ is obtained by measuring $\mathbf{M}_0$ on the output $\mathbf{U}(\boldsymbol{\theta})\boldsymbol{\rho}_i\mathbf{U}^\dagger(\boldsymbol{\theta})$. Like deep neural networks, the parameters $\boldsymbol{\theta}$ in the variational circuits are optimized by gradient-based methods to minimize an objective function that measures the misalignments of the predicted outputs and the ground truth labels.

With the recent development of quantum technology, the near-term Noisy Intermediate-Scale Quantum (NISQ) (Preskill, 2018) computer has become an important platform for demonstrating quantum advantage with practical applications. As a hybrid of classical optimizers and quantum representations, QNNs is a promising candidate for demonstrating such advantage on quantum computers available to us in the near future: quantum machine learning models are proved to have a margin over the classical counterparts in terms of the expressive power due the to the exponentially large Hilbert space of quantum states (Huang et al., 2021; Anschuetz, 2022). On the other hand by delegating the optimization procedures to classical computers, the hybrid method requires significantly less quantum resources, which is crucial for readily available quantum computers with limited coherence time and error correction. There have been proposals of QNNs (Dunjko & Briegel, 2018; Schuld & Killoran, 2019) for classification (Farhi et al., 2020; Romero et al., 2017) and generative learning (Lloyd & Weedbrook, 2018; Zoufal et al., 2019; Chakrabarti et al., 2019).

Despite their potential there are challenges in the practical deployment of QNNs. Most notably, the optimization problem for training QNNs can be highly non-convex. The landscape of QNN training may be swarmed with spurious local minima and saddle points that can trap gradient-based optimization methods (You & Wu, 2021; Anschuetz & Kiani, 2022). QNNs with large dimensions also suffer

[1]Department of Computer Science, University of Maryland, College Park, United States [2]Global Technology Applied Research, J. P. Morgan Chase & Co. [3]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China. Correspondence to: Xuchen You <xyou@umd.edu>, Xiaodi Wu <xwu@cs.umd.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

from a phenomenon called the *barren plateau* (McClean et al., 2018), where the gradients of the parameters vanish at random intializations, making convergence slow even in a trap-free landscape. These difficulties in training QNNs, together with the challenge of classically simulating QNNs at a decent scale, calls for a theoretical understanding of the convergence of QNNs.

Neural Tangent Kernels Many of the theoretical difficulties in understanding QNNs have also been encountered in the study of classical deep neural networks: despite the landscape of neural networks being non-convex and susceptible to spurious local minima and saddle points, it has been empirically observed that the training errors decays exponentially in the training time (Livni et al., 2014; Arora et al., 2019) in the highly *over-parameterized* regime with sufficiently many number of trainable parameters. This phenomenon is theoretically explained by connecting the training dynamics of neural networks to the kernel regression: the kernel regression model generalizes the linear regression by equipping the linear model with non-linear feature maps. Given a training set $\{\mathbf{x}_j, y_j\}_{j=1}^m \subset \mathcal{X} \times \mathcal{Y}$ and a non-linear feature map $\phi : \mathcal{X} \to \mathcal{X}'$ mapping the features to a potentially high-dimensional feature space $\mathcal{X}'$. The kernel regression solves for the optimal weight $\mathbf{w}$ that minimizes the mean-square loss $\frac{1}{2m}\sum_{j=1}^m (\mathbf{w}^T\phi(\mathbf{x}_j) - y_j)^2$. The name of kernel regression stems from the fact that the optimal hypothesis $\mathbf{w}$ depends on the high-dimensional feature vectors $\{\phi(\mathbf{x}_j)\}_{j=1}^m$ through a $m \times m$ *kernel* matrix $\mathbf{K}$, such that $K_{ij} = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$. The kernel regression enjoys a linear convergence (i.e. the mean square loss decaying exponentially over time) when $\mathbf{K}$ is positive definite.

The kernel matrix associated with a neural network is determined by tracking how the predictions for each training sample evolve jointly at random initialization. The study of the neural network convergence then reduces to characterizing the corresponding kernel matrices (the neural tangent kernel, or the NTK). In addition to the convergence results, NTK also serves as a tool for studying other aspect of neural networks including generalization (Canatar et al., 2021; Chen et al., 2020) and stability (Bietti & Mairal, 2019).

The key observation that justifies the study of neural networks with neural tangent kernels, is that the NTK becomes a constant (over time) during training in the limit of infinite layer widths. This has been theoretically established starting with the analysis of wide fully-connected neural networks (Jacot et al., 2018; Arora et al., 2019; Chizat et al., 2019) and later generalized to a variety of architectures (e.g. Allen-Zhu et al. (2019)).

Quantum NTKs Inspired by the success of NTKs, recent years have witnessed multiple works attempting to associate over-parameterized QNNs to kernel regression. Along the line there are two types of studies. The first category investigates and compares the properties of the "quantum" kernel induced by the quantum encoding of classical features, where K_{ij} associated with the i-th and j-th feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ equals $\mathrm{tr}(\boldsymbol{\rho}_\mathbf{i}\boldsymbol{\rho}_\mathbf{j})$ with $\boldsymbol{\rho}_i$ and $\boldsymbol{\rho}_j$ being the quantum state encodings, without referring to the dynamics of training (Schuld & Killoran, 2019; Huang et al., 2021; Liu et al., 2022b). The second category seeks to directly establish the quantum version of NTK for QNNs by examining the evolution of the model predictions at random initialization, which is the recipe for calculating the classical NTK in Arora et al. (2019): Shirai et al. (2021) empirically evaluates the direct training of the quantum NTK instead of the original QNN formulation. On the other hand, by analyzing the time derivative of the quantum NTK at initialization, Liu et al. (2022a) conjectures that in the limit of over-parameterization, the quantum NTK is a constant over time and therefore the dynamics reduces to a kernel regression.

Despite recent efforts, a rigorous answer remains evasive whether the quantum NTK is a constant during training for over-parameterized QNNs. We show that the answer to this question is indeed, surprisingly negative: as a result of the unitarity of quantum circuits, there is a finite change in the conjectured quantum NTK as the training error decreases, even in the the limit of over-parameterization.

Contributions In this work, we focus on QNNs equipped with the mean square loss, trained using gradient flow, following Arora et al. (2019). In Section 3, we show that, despite the formal resemblance to kernel regression dynamics, the over-parameterized QNN does not follow the dynamics of *any* kernel regression due to the unitarity: for the widely-considered setting of classifications with Pauli measurements, we show that the objective function at time t decays at most as a polynomial function of $1/t$ (Theorem 3.2). This contradicts the dynamics of any kernel regression with a positive definite kernel, which exhibits convergence with $L(t) \leq L(0)\exp(-ct)$ for some positive constant c. We also identify the true asymptotic dynamics of QNN training as regression with a time-varying Gram matrix $\mathbf{K}_{\mathsf{asym}}$ (Lemma 4.1), and show rigorously that the real dynamics concentrates to the asymptotic one in the limit $p \to \infty$ (Theorem 4.2). This reduces the problem of investigating QNN convergence to studying the convergence of the asymptotic dynamics governed by $\mathbf{K}_{\mathsf{asym}}$.

We also consider a model of QNNs where the final measurement is post-processed by a linear scaling. In this setting, we provide a complete analysis of the convergence of the asymptotic dynamics in the case of 1 training sample (Corollary 4.3), and provide further theoretical evidence of convergence in the neighborhood of most global minima when the number of samples $m > 1$ (Theorem 4.4). These theoretical

evidences are supplemented with an empirical study that demonstrates in generality, the convergence of the asymptotic dynamics when $m \geq 1$. Coupled with our proof of convergence, these form the strongest concrete evidences of the convergence of training for over-parameterized QNNs.

Connections to previous works Our result extends the existing literature on QNN landscapes (e.g. Anschuetz (2022); Russell et al. (2017)) and looks into the training dynamics, which allows us to characterize the rate of convergence and to show how the range of the measurements affects the convergence to global minima. The dynamics for over-parameterized QNNs proposed by us can be reconciled with the existing calculations of quantum NTK as follows: in the regime of over-parameterization, the QNN dynamics coincides with the quantum NTK dynamics conjectured in Liu et al. (2022a) at random initialization; yet it deviates from quantum NTK dynamics during training, and the deviation does not vanish in the limit of $p \to \infty$.

2. Preliminaries

Empirical risk minimization (ERM) A supervised learning problem is specified by a joint distribution $\mathcal{D}$ over the feature space $\mathcal{X}$ and the label space $\mathcal{Y}$, and a family $\mathcal{F}$ of mappings from $\mathcal{X}$ to $\mathcal{Y}$ (i.e. the hypothesis set). The goal is to find an $f \in \mathcal{F}$ that well predicts the label y given the feature $\mathbf{x}$ in expectation, for pairs of $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ drawn $i.i.d.$ from the distribution $\mathcal{D}$.

Given a training set $\mathcal{S} = \{\mathbf{x}_j, y_j\}_{j=1}^m$ composed of m pairs of features and labels, we search for the optimal $f \in \mathcal{F}$ by the *empirical risk minimization* (ERM): let ℓ be a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, ERM finds an $f \in \mathcal{F}$ that minimizes the average loss: $\min_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \ell(\hat{y}_i, y_i)$, where $\hat{y}_i = f(\mathbf{x}_i)$. We focus on the common choice of the *square loss* $\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$.

Classical neural networks A popular choice of the hypothesis set $\mathcal{F}$ in modern-day machine learning is the *classical neural networks*. A vanilla version of the L-layer feed-forward neural network takes the form $f(x; \mathbf{W}_1, \cdots, \mathbf{W}_L) = \mathbf{W}_L \sigma(\cdots \mathbf{W}_2 \sigma(\mathbf{W}_1 \sigma(x)) \cdots)$, where $\sigma(\cdot)$ is a non-linear activation function, and for all $l \in [L]$, $\mathbf{W}_l \in \mathbb{R}^{d_l \times d_{l-1}}$ is the weights in the l-th layer, with $d_L = 1$ and d_0 the same as the dimension of the feature space $\mathcal{X}$. It has been shown that, in the limit $\min_{l=1}^{L-1} d_l \to \infty$, the training of neural networks with square loss is close to kernel learning, and therefore enjoys a linear convergence rate (Jacot et al., 2018; Arora et al., 2019; Allen-Zhu et al., 2019; Oymak & Soltanolkotabi, 2020).

Quantum neural networks Quantum neural networks is a family of parameterized hypothesis set analogous to its classical counterpart. At a high level, it has the layered-structure like a classical neural network. At each layer, a linear transformation acts on the output from the last layer. A quantum neural network is different from its classical counterpart in the following three aspects.

(1) Quantum states as inputs A d-dimensional quantum state is represented by a *density matrix* $\boldsymbol{\rho}$, which is a positive semidefinite $d \times d$ Hermitian with trace 1. A state is said to be pure if $\boldsymbol{\rho}$ is rank-1. Pure states can therefore be equivalently represented by a state vector $\mathbf{v}$ such that $\boldsymbol{\rho} = \mathbf{v}\mathbf{v}^\dagger$. The inputs to QNNs are quantum states. They can either be drawn as samples from a quantum-physical problem or be the encodings of classical feature vectors.

(2) Parameterization In classical neural networks, each layer is composed of a linear transformation and a non-linear activation, and the matrix associated with the linear transformation can be directly optimized at each entry. In QNNs, the entries of each linear transformation can not be directly manipulated. Instead we update parameters in a variational ansatz to update the linear transformations. More concretely, a general p-parameter ansatz $\mathbf{U}(\boldsymbol{\theta})$ in a d-dimensional Hilbert space can be specified by a set of $d \times d$ unitaries $\{\mathbf{U}_0, \mathbf{U}_1, \cdots, \mathbf{U}_p\}$ and a set of non-zero $d \times d$ Hermitians $\{\mathbf{H}^{(1)}, \mathbf{H}^{(2)}, \cdots, \mathbf{H}^{(p)}\}$ as

$$\mathbf{U}_p \exp(-i\theta_p \mathbf{H}^{(p)}) \mathbf{U}_{p-1} \exp(-i\theta_{p-1} \mathbf{H}^{(p-1)}) \cdot \cdots \cdot \exp(-i\theta_2 \mathbf{H}^{(2)}) \mathbf{U}_1 \exp(-i\theta_1 \mathbf{H}^{(1)}) \mathbf{U}_0. \quad (1)$$

Without loss of generality, we assume that $\mathrm{tr}(\mathbf{H}^{(l)}) = 0$. This is because adding a Hermitian proportional to $\mathbf{I}$ on the generator $\mathbf{H}^{(l)}$ does not change the density matrix of the output states. Notice that most p-parameter ansatze $\mathbf{U} : \mathbb{R}^p \to \mathbb{C}^{d \times d}$ can be expressed as Equation 1. One exception may be the anastz design with intermediate measurements (e.g. Cong et al. (2019)). In Section 4, we will also consider the periodic anastz:

Definition 2.1 (Periodic ansatz)**.** A d-dimensional p-parameter periodic anasatz $\mathbf{U}(\boldsymbol{\theta})$ is defined as

$$\mathbf{U}_p \exp(-i\theta_p \mathbf{H}) \cdot \cdots \cdot \mathbf{U}_1 \exp(-i\theta_1 \mathbf{H}) \mathbf{U}_0, \quad (2)$$

where $\mathbf{U}_l$ are sampled $i.i.d.$ with respect to the Haar measure over the special unitary group $SU(d)$, and $\mathbf{H}$ is a non-zero trace-0 Hermitian.

Up to a unitary transformation, the periodic ansatz is equivalent to an ansatz in Line (1) where $\{\mathbf{H}^{(l)}\}_{l=1}^p$ sampled as $\mathbf{V}_l \mathbf{H} \mathbf{V}_l^\dagger$ with $\mathbf{V}_l$ being haar random $d \times d$ unitary matrices. Similar ansatze have been considered in McClean et al. (2018); Anschuetz (2022); You & Wu (2021); You et al. (2022).

(3) Readout with measurements Contrary to classical neural networks, the readout from a QNN requires performing quantum *measurements*. A measurement is specified by a Hermitian $\mathbf{M}$. The outcome of measuring a quantum state $\boldsymbol{\rho}$ with a measurement $\mathbf{M}$ is $\mathrm{tr}(\boldsymbol{\rho}\mathbf{M})$, which is a linear function of $\boldsymbol{\rho}$. A common choice is the Pauli measurement: Pauli matrices are 2×2 Hermitians that are also unitary. The Pauli measurements are tensor products of Pauli matrices, featuring eigenvalues of ± 1.

A common choice is the Pauli measurement: Pauli matrices are 2×2 Hermitians that are also unitary:

$$\sigma_X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \sigma_Y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \sigma_Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

The Pauli measurements are tensor products of Pauli matrices, featuring eigenvalues of ± 1.

ERM of quantum neural network. We focus on quantum neural networks equipped with the mean-square loss. Solving the ERM for a dataset $\mathcal{S} := \{(\boldsymbol{\rho}_j, y_j)\}_{j=1}^m \subseteq (\mathbb{C}^{d\times d} \times \mathbb{R})^m$ involves optimizing the objective function $\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) := \frac{1}{2m}\sum_{j=1}^m \left(\hat{y}_j(\boldsymbol{\theta}) - y_j\right)^2$, where $\hat{y}_j(\boldsymbol{\theta}) = \mathrm{tr}(\boldsymbol{\rho}_j \mathbf{U}^\dagger(\boldsymbol{\theta})\mathbf{M}_0\mathbf{U}(\boldsymbol{\theta}))$ for all $j \in [m]$ with $\mathbf{M}_0$ being the quantum measurement and $\mathbf{U}(\boldsymbol{\theta})$ being the variational ansatz. Typically, a QNN is trained by optimizing the ERM objective function by gradient descent: at the t-th iteration, the parameters are updated as $\boldsymbol{\theta}(t+1) \leftarrow \boldsymbol{\theta}(t) - \eta \nabla L(\boldsymbol{\theta}(t))$, where η is the learning rate; for sufficiently small η, the dynamics of gradient descent reduces to that of the gradient flow: $d\boldsymbol{\theta}(t)/dt = -\eta \nabla L(\boldsymbol{\theta}(t))$. Here we focus on the gradient flow setting following Arora et al. (2019).

Rate of convergence In the optimization literature, the rate of convergence describes how fast an iterative algorithm approaches an (approximate) solution. For a general function L with variables $\boldsymbol{\theta}$, let $\boldsymbol{\theta}(t)$ be the solution maintained at the time step t and $\boldsymbol{\theta}^\star$ be the optimal solution. The algorithm is said to be converging *exponentially fast* or at a *linear rate* if $L(\boldsymbol{\theta}(t)) - L(\boldsymbol{\theta}^\star) \le \alpha \exp(-ct)$ for some constants c and α. In contrast, algorithms with the sub-optimal gap $L(\boldsymbol{\theta}(t)) - L(\boldsymbol{\theta}^\star)$ decreasing slower than exponential are said to be converging with a *sublinear* rate (e.g. $L(\boldsymbol{\theta}(t)) - L(\boldsymbol{\theta}^\star)$ decaying with t as a polynomial of $1/t$). We will mainly consider the setting where $L(\boldsymbol{\theta}^\star) = 0$ (i.e. the *realizable* case) with continuous time t.

Other notations We use $\|\cdot\|_{\mathsf{op}}$, $\|\cdot\|_F$ and $\|\cdot\|_{\mathrm{tr}}$ to denote the operator norm (i.e. the largest eigenvalue in terms of the absolute values), Frobenius norm and the trace norm of matrices; we use $\|\cdot\|_p$ to denote the p-norm of vectors, with the subscript omitted for $p = 2$. We use $\mathrm{tr}(\cdot)$ to denote the trace operation.

3. Deviations of QNN Dynamics from NTK

Consider a regression model on an m-sample training set: for all $j \in [m]$, let y_j and $\hat{y}_j$ be the label and the model prediction of the j-th sample. The *residual* vector $\mathbf{r}$ is a m-dimensional vector with $r_j := y_j - \hat{y}_j$. The dynamics of the kernel regression is signatured by the first-order linear dynamics of the residual vectors: let $\mathbf{w}$ be the learned model parameter, and let $\phi(\cdot)$ be the fixed non-linear map. Recall that the kernel regression minimizes $L(\mathbf{w}) = \frac{1}{2m}\sum_{j=1}^m (\mathbf{w}^T\phi(\mathbf{x}_j) - y_j)^2$ for a training set $\mathcal{S} = \{(\mathbf{x}_j, y_j)\}_{j=1}^m$, and the gradient with respect to $\mathbf{w}$ is $\frac{1}{m}\sum_{j=1}^m (\mathbf{w}^T\phi(\mathbf{x}_j) - y_j)\phi(\mathbf{x}_j) = -\frac{1}{m}\sum_{j=1}^m r_j\phi(\mathbf{x}_j)$. Under the gradient flow with learning rate η, the weight $\mathbf{w}$ updates as $\frac{d\mathbf{w}}{dt} = \frac{\eta}{m}\sum_{j=1}^m r_j\phi(\mathbf{x}_j)$, and the i-th entry of the residual vector updates as $dr_i/dt = -\phi(\mathbf{x}_i)^T\frac{d\mathbf{w}}{dt} = -\frac{\eta}{m}\sum_{j=1}^m \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)r_j$, or more succinctly $d\mathbf{r}/dt = -\frac{\eta}{m}\mathbf{K}\mathbf{r}$ with $\mathbf{K}$ being the kernel/Gram matrix defined as $K_{ij} = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ (see also Arora et al. (2019)). Notice that the kernel matrix $\mathbf{K}$ is a constant of time and is independent of the weight $\mathbf{w}$ or the labels.

Dynamics of residual vectors We start by characterizing the dynamics of the residual vectors for the general form of p-parameter QNNs and highlight the limitation of viewing the over-parameterized QNNs as kernel regressions. Similar to the kernel regression, $\frac{dr_j}{dt} = -\frac{d\hat{y}_j}{dt} = -\mathrm{tr}(\boldsymbol{\rho}_j \frac{d}{dt}\mathbf{U}^\dagger(\boldsymbol{\theta}(t))\mathbf{M}_0\mathbf{U}(\boldsymbol{\theta}(t)))$ in QNNs. We derive the following dynamics of $\mathbf{r}$ by tracking the parameterized measurement $\mathbf{M}(\boldsymbol{\theta}) = \mathbf{U}^\dagger(\boldsymbol{\theta})\mathbf{M}_0\mathbf{U}(\boldsymbol{\theta})$ as a function of time t.

Lemma 3.1 (Dynamics of the residual vector)**.** *Consider a QNN instance with an ansatz $\mathbf{U}(\boldsymbol{\theta})$ defined as in Line (1), a training dataset $\mathcal{S} = \{(\boldsymbol{\rho}_j, y_j)\}_{j=1}^m$, and a measurement $\mathbf{M}_0$. Under the gradient flow for the objective function $L(\boldsymbol{\theta}) = \frac{1}{2m}\sum_{j=1}^m \left(\mathrm{tr}(\boldsymbol{\rho}_j\mathbf{U}^\dagger(\boldsymbol{\theta})\mathbf{M}_0\mathbf{U}(\boldsymbol{\theta})) - y_j\right)^2$ with learning rate η, the residual vector $\mathbf{r}$ satisfies the differential equation*

$$\frac{d\mathbf{r}(\boldsymbol{\theta}(t))}{dt} = -\frac{\eta}{m}\mathbf{K}(\mathbf{M}(\boldsymbol{\theta}(t)))\mathbf{r}(\boldsymbol{\theta}(t)), \tag{3}$$

where $\mathbf{K}$ is a positive semi-definite matrix-valued function of the parameterized measurement. The (i,j)-th element of $\mathbf{K}$ is defined as

$$\sum_{l=1}^p \left(\mathrm{tr}\left(\mathrm{i}[\mathbf{M}(\boldsymbol{\theta}(t)), \boldsymbol{\rho}_i]\mathbf{H}_l\right)\mathrm{tr}\left(\mathrm{i}[\mathbf{M}(\boldsymbol{\theta}(t)), \boldsymbol{\rho}_j]\mathbf{H}_l\right)\right). \tag{4}$$

Here $\mathbf{H}_l := \mathbf{U}_0^\dagger\mathbf{U}_{1:l-1}^\dagger(\boldsymbol{\theta})\mathbf{H}^{(l)}\mathbf{U}_{1:l-1}(\boldsymbol{\theta})\mathbf{U}_0$, is a function of $\boldsymbol{\theta}$ with $\mathbf{U}_{1:r}(\boldsymbol{\theta})$ being the shorthand for $\mathbf{U}_r\exp(-i\theta_r\mathbf{H}^{(r)})\cdots\mathbf{U}_1\exp(-i\theta_1\mathbf{H}^{(1)})$.

While Equation (3) takes a similar form to that of the kernel regression, the matrix $\mathbf{K}$ is *dependent* on the parameterized

measurement $\mathbf{M}(\boldsymbol{\theta})$. This is a consequence of the unitarity: consider an alternative parameterization, where the objective function $\mathbf{L}(\mathbf{M}) = \frac{1}{2m}\sum_{j=1}^{m}\left(\operatorname{tr}(\boldsymbol{\rho}_j\mathbf{M}) - y_j\right)^2$ is optimized over all Hermitian matrices $\mathbf{M}$. It can be easily verified that the corresponding dynamics is exactly the kernel regression with $K_{ij} = \operatorname{tr}(\boldsymbol{\rho}_i\boldsymbol{\rho}_j)$.

Due to the unitarity of the evolution of quantum states, the spectrum of eigenvalues of the parameterized measurement $\mathbf{M}(\boldsymbol{\theta})$ is required to remain the same throughout training. In the proof of Lemma 3.1 (deferred to Section A.1 in the appendix), we see that the derivative of $\mathbf{M}(\boldsymbol{\theta})$ takes the form of a linear combination of commutators $i[\mathbf{A}, \mathbf{M}(\boldsymbol{\theta})]$ for some Hermitian $\mathbf{A}$. As a result, the traces of the k-th matrix powers $\operatorname{tr}(\mathbf{M}^k(\boldsymbol{\theta}))$ are constants of time for any integer k, since $d\operatorname{tr}(\mathbf{M}^k(\boldsymbol{\theta}))/dt = k\operatorname{tr}(\mathbf{M}^{k-1}(\boldsymbol{\theta})d\mathbf{M}(\boldsymbol{\theta})/dt) = k\operatorname{tr}(\mathbf{M}^{k-1}(\boldsymbol{\theta})i[\mathbf{A}, \mathbf{M}(\boldsymbol{\theta})]) = 0$ for any Hermitian $\mathbf{A}$. The spectrum of eigenvalues remains unchanged because the coefficients of the characteristic polynomials of $\mathbf{M}(\boldsymbol{\theta})$ is completely determined by the traces of matrix powers. On the contrary, the eigenvalues are in general not preserved for $\mathbf{M}$ evolving under the kernel regression.

Another consequence of the unitarity constraint is that a QNN can not make predictions outside the range of the eigenvalues of $\mathbf{M}_0$, while for the kernel regression with a strictly positive definite kernel, the model can (over-)fit training sets with arbitrary label assignments. Here we further show that the unitarity is pronounced in a typical QNN instance where the predictions are within the range of the measurement.

Sublinear convergence in QNNs One of the most common choices for designing QNNs is to use a (tensor product of) Pauli matrices as the measurement (see e.g. Farhi et al. (2020); Dunjko & Briegel (2018)). Such a choice features a measurement $\mathbf{M}_0$ with eigenvalues $\{\pm 1\}$ and trace zero. Here we show that in the setting of supervised learning on pure states with Pauli measurements, the (neural tangent) kernel regression is insufficient to capture the convergence of QNN training. For the kernel regression with a positive definite kernel $\mathbf{K}$, the objective function L can be expressed as $\frac{1}{2m}\sum_{j=1}^{m}(\hat{y}_j - y_j)^2 = \frac{1}{2m}\mathbf{r}^T\mathbf{r}$; under the kernel dynamics of $\frac{d\mathbf{r}}{dt} = -\frac{\eta}{m}\mathbf{K}\mathbf{r}$, it is easy to verify that $\frac{d\ln L}{dt} = -\frac{2\eta}{m}\frac{\mathbf{r}^T\mathbf{K}\mathbf{r}}{\mathbf{r}^T\mathbf{r}} \le -\frac{2\eta}{m}\lambda_{\min}(\mathbf{K})$ with $\lambda_{\min}(\mathbf{K})$ being the smallest eigenvalue of $\mathbf{K}$. This indicates that L decays at a linear rate, i.e. $L(T) \le L(0)\exp(-\frac{2\eta}{m}\lambda_{\min}(\mathbf{K})T)$. In contrast, we show that the rate of convergence of the QNN dynamics *must* be sublinear, slower than the linear convergence rate predicted by the kernel regression model with a positive definite kernel.

Theorem 3.2 (No faster than sublinear convergence)**.** *Consider a QNN instance with a training set $\mathcal{S} = \{(\boldsymbol{\rho}_j, y_j)\}$ such that $\boldsymbol{\rho}_j$ are pure states and $y_j \in \{\pm 1\}$, and a measurement $\mathbf{M}_0$ with eigenvalues in $\{\pm 1\}$. Under the gradient flow for the objective function $L(\boldsymbol{\theta}) = \frac{1}{2m}\sum_{j=1}^{m}\operatorname{tr}(\boldsymbol{\rho}_j\mathbf{M}(\boldsymbol{\theta}) - y_j)^2$, for any ansatz $\mathbf{U}(\boldsymbol{\theta})$ defined in Line (1), L converges to zero at most at a sublinear convergence rate. More concretely, for $\mathbf{U}(\boldsymbol{\theta})$ generated by $\{\mathbf{H}^{(l)}\}_{l=1}^{p}$, let η be the learning rate and m be the sample size, the objective function at time t:*

$$L(\boldsymbol{\theta}(t)) \ge 1/(c_0 + c_1 t)^2. \tag{5}$$

Here the constant $c_0 = 1/\sqrt{L(\boldsymbol{\theta}(0))}$ depends on the objective function at initialization, and $c_1 = 12\eta\sum_{l=1}^{p}\left\|\mathbf{H}^{(l)}\right\|_{\mathsf{op}}^2$.

The constant c_1 in the theorem depends on the number of parameters p through $\sum_{l=1}^{p}\left\|\mathbf{H}^{(l)}\right\|_{\mathsf{op}}^2$ if the operator norm of $\mathbf{H}^{(l)}$ is a constant of p. We can get rid of the dependency on p by scaling the learning rate η or changing the time scale, which does not affect the sublinearity of convergence.

By expressing the objective function $L(\boldsymbol{\theta}(t))$ as $\frac{1}{2m}\mathbf{r}(\boldsymbol{\theta}(t))^T\mathbf{r}(\boldsymbol{\theta}(t))$, Lemma 3.1 indicates that the decay of $\frac{dL(\boldsymbol{\theta}(t))}{dt}$ is lower-bounded by $\frac{-2\eta}{m}\lambda_{\max}(\mathbf{K}(\boldsymbol{\theta}(t)))L(\boldsymbol{\theta}(t))$, where $\lambda_{\max}(\cdot)$ is the largest eigenvalue of a Hermitian matrix. The full proof of Theorem 3.2 is deferred to Section A.2, and follows from the fact that when the QNN prediction for an input state $\boldsymbol{\rho}_j$ is close to the ground truth $y_j = 1$ or -1, the diagonal entry $K_{jj}(\boldsymbol{\theta}(t))$ vanishes. As a result the largest eigenvalue $\lambda_{\max}(\mathbf{K}(\boldsymbol{\theta}(t)))$ also vanishes as the objective function $L(\boldsymbol{\theta}(t))$ approaches 0 (which is the global minima). Notice the sublinearity of convergence is independent of the system dimension d, the choices of $\{\mathbf{H}^{(l)}\}_{l=1}^{p}$ in $\mathbf{U}(\boldsymbol{\theta})$ or the number of parameters p. This means that the dynamics of QNN training is completely different from kernel regression even in the limit where d and/or $p \to \infty$.

Experiments: sublinear QNN convergence To support Theorem 3.2, we simulate the training of QNNs using $\mathbf{M}_0$ with eigenvalues ± 1. For dimension $d = 32$ and 64, we randomly sample four d-dimensional pure states that are orthogonal, with two of samples labeled $+1$ and the other two labeled -1. The training curves (plotted under the log scale) in Figure 1 flattens as L approaches 0, suggesting the rate of convergence $-d\ln L/dt$ vanishes around global minima, which is a signature of the sublinear convergence. Note that the sublinearity of convergence is independent of the number of parameters p. For gradient flow or gradient descent with sufficiently small step-size, the scaling of a constant learning rate η leads to a scaling of time t and does not fundamentally change the (sub)linearity of the convergence. For the purpose of visual comparison, we scale η with p by choosing the learning rate as $10^{-3}/p$. For more details on the experiments, please refer to Section D.

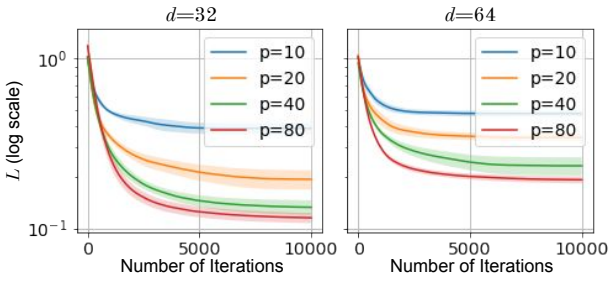

Figure 1. Sublinear convergence of QNN training. For QNNs with Pauli measurements for a classification task, the (log-scaled) training curves flatten as the number of iterations increases, indicating a sublinear convergence. The flattening of training curves remains for increasing numbers of parameters $p = 10, 20, 40, 80$. The training curves are averaged over 10 random initialization, and the error bars are the halves of standard deviations.

4. Asymptotic Dynamics of QNNs

As demonstrated in the previous section, the dynamics of the QNN training deviates from the kernel regression for any choices of the number of parameters p and the dimension d in the setting of Pauli measurements for classification. This calls for a new characterization of the QNN dynamics in the regime of over-parameterization. For a concrete definition of over-parameterization, we consider the family of the periodic ansatze in Definition 2.1, and refer to the limit of $p \to \infty$ with a fixed generating Hamiltonian $\mathbf{H}$ as the regime of over-parameterization. In this section, we derive the asymptotic dynamics of QNN training when number of parameters p in the periodic ansatze goes to infinity. We start by decomposing the dynamics of the residual $\mathbf{r}(\boldsymbol{\theta}(t))$ into a term corresponding to the asymptotic dynamics, and a term of perturbation that vanishes as $p \to \infty$. As mentioned before, in the context of the gradient flow, the choice of η is merely a scaling of the time and therefore arbitrary. For a QNN instance with m training samples and a p-parameter ansatz generated by a Hermitian $\mathbf{H}$ as defined in Line (2), we choose η to be $\frac{m}{p}\frac{d^2-1}{\mathrm{tr}(\mathbf{H}^2)}$ to facilitate the presentation:

Lemma 4.1 (Decomposition of the residual dynamics)**.** *Let $\mathcal{S}$ be a training set with m samples $\{(\boldsymbol{\rho}_j, y_j)\}_{j=1}^m$, and let $\mathbf{U}(\boldsymbol{\theta})$ be a p-parameter ansatz generated by a non-zero $\mathbf{H}$ as in Line 2. Consider a QNN instance with a training set $\mathcal{S}$, ansatz $\mathbf{U}(\boldsymbol{\theta})$ and a measurement $\mathbf{M}_0$. Under the gradient flow with $\eta = \frac{m}{p}\frac{d^2-1}{\mathrm{tr}(\mathbf{H}^2)}$, the residual vector $\mathbf{r}(t)$ as a function of time t through $\boldsymbol{\theta}(t)$ evolves as*

$$\frac{d\mathbf{r}(t)}{dt} = -(\mathbf{K}_{\mathsf{asym}}(t) + \mathbf{K}_{\mathsf{pert}}(t))\mathbf{r}(t) \tag{6}$$

where both $\mathbf{K}_{\mathsf{asym}}$ and $\mathbf{K}_{\mathsf{pert}}$ are functions of time through the parameterized measurement $\mathbf{M}(\boldsymbol{\theta}(t))$, such that

$$(\mathbf{K}_{\mathsf{asym}}(t))_{ij} := \mathrm{tr}\left(i[\mathbf{M}(t), \boldsymbol{\rho}_i]\, i[\mathbf{M}(t), \boldsymbol{\rho}_j]\right), \tag{7}$$

$$(\mathbf{K}_{\mathsf{pert}}(t))_{ij} := \mathrm{tr}\left(i[\mathbf{M}(t), \boldsymbol{\rho}_i] \otimes i[\mathbf{M}(t), \boldsymbol{\rho}_j]\Delta(t)\right). \tag{8}$$

Here $\Delta(t)$ is a $d^2 \times d^2$ Hermitian as a function of t through $\boldsymbol{\theta}(t)$.

Under the random initialization by sampling $\{\mathbf{U}_l\}_{l=1}^p$ i.i.d. from the haar measure over the special unitary group $SU(d)$, $\Delta(0)$ concentrates at zero as p increases. We further show that $\Delta(t) - \Delta(0)$ has a bounded operator norm decreasing with number of parameters. This allows us to associate the convergence of the over-parameterized QNN with the properties of $\mathbf{K}_{\mathsf{asym}}(t)$:

Theorem 4.2 (Linear convergence of QNN with mean-square loss)**.** *Let $\mathcal{S}$ be a training set with m samples $\{(\boldsymbol{\rho}_j, y_j)\}_{j=1}^m$, and let $\mathbf{U}(\boldsymbol{\theta})$ be a p-parameter ansatz generated by a non-zero $\mathbf{H}$ as in Line (2). Consider a QNN instance with the training set $\mathcal{S}$, ansatz $\mathbf{U}(\boldsymbol{\theta})$ and a measurement $\mathbf{M}_0$, trained by gradient flow with $\eta = \frac{m}{p}\frac{d^2-1}{\mathrm{tr}(\mathbf{H}^2)}$. Then for sufficiently large number of parameters p, if the smallest eigenvalue of $\mathbf{K}_{\mathsf{asym}}(t)$ is greater than a constant C_0, then with high probability over the random initialization of the periodic ansatz, the loss function converges to zero at a linear rate*

$$L(t) \leq L(0)\exp(-\frac{C_0 t}{2}). \tag{9}$$

We defer the proof to Section B.2. Similar to $\mathbf{r}(t)$, the evolution of $\mathbf{M}(t)$ decomposes into an asymptotic term

$$\frac{d}{dt}\mathbf{M}(t) = \sum_{j=1}^m r_j[\mathbf{M}(t), [\mathbf{M}(t), \boldsymbol{\rho}_j]] \tag{10}$$

and a perturbative term depending on $\Delta(t)$. Theorem 4.2 allows us to study the behavior of an over-parameterized QNN by simulating/characterizing the asymptotic dynamics of $\mathbf{M}(t)$, which is significantly more accessible.

Application: QNN with one training sample To demonstrate the proposed asymptotic dynamics as a tool for analyzing over-parameterized QNNs, we study the convergence of the QNN with one training sample $m = 1$. To set a separation from the regime of the sublinear convergence, consider the following setting: let $\mathbf{M}_0$ be a Pauli measurement, for any input state $\boldsymbol{\rho}$, instead of assigning $\hat{y} = \mathrm{tr}(\boldsymbol{\rho}\mathbf{U}(\boldsymbol{\theta})^\dagger \mathbf{M}_0 \mathbf{U}(\boldsymbol{\theta}))$, take $\gamma\,\mathrm{tr}(\boldsymbol{\rho}\mathbf{U}(\boldsymbol{\theta})^\dagger \mathbf{M}_0 \mathbf{U}(\boldsymbol{\theta}))$ as the prediction $\hat{y}$ at $\boldsymbol{\theta}$ for a scaling factor $\gamma > 1.0$. The γ-scaling of the measurement outcome can be viewed as a classical processing in the context of quantum information, or as an activation function (or a link function) in the context of machine learning, and is equivalent to a QNN

with measurement $\gamma \mathbf{M}_0$. The following corollary implies the convergence of 1-sample QNN for $\gamma > 1.0$ under a mild initial condition:

Corollary 4.3. *Let $\boldsymbol{\rho}$ be a d-dimensional pure state, and let y be ± 1. Consider a QNN instance with a Pauli measurement $\mathbf{M}_0$, an one-sample training set $\mathcal{S} = \{(\boldsymbol{\rho}, y)\}$ and an ansatz $\mathbf{U}(\boldsymbol{\theta})$ defined in Line (2). Assume the scaling factor $\gamma > 1.0$ and $p \to \infty$ with $\eta = \frac{d^2-1}{p \operatorname{tr}(\mathbf{H}^2)}$. Under the initial condition that the prediction at $t = 0$, $\hat{y}(0)$ is less than 1, the objective function converges linearly with*

$$L(t) \leq L(0) \exp(-C_1 t) \tag{11}$$

with the convergence rate $C_1 \geq \gamma^2 - 1$.

With a scaling factor γ and training set $\{(\boldsymbol{\rho}_j, y_j)\}_{j=1}^m$, the objective function, as a function of the parameterized measurement $\mathbf{M}(t)$, reads as: $L(\mathbf{M}(t)) = \frac{1}{2m} \sum_{j=1}^m (\gamma \operatorname{tr}(\boldsymbol{\rho}_j \mathbf{M}(t)) - y_j)^2$. As stated in Theorem 4.2, for sufficiently large number of parameters p, the convergence rate of the residual $\mathbf{r}(t)$ is determined by $\mathbf{K}_{\text{asym}}(t)$, as the asymptotic dynamics of $\mathbf{r}(t)$ reads as $\frac{d}{dt}\mathbf{r} = -\mathbf{K}_{\text{asym}}(\mathbf{M}(t))\mathbf{r}(t)$ with the chosen η. For $m = 1$, the asymptotic matrix $\mathbf{K}_{\text{asym}}$ reduces to a scalar $k(t) = -\operatorname{tr}([\gamma \mathbf{M}(t), \boldsymbol{\rho}]^2) = 2(\gamma^2 - \hat{y}(t)^2)$. $\hat{y}(t)$ approaches the label y if $k(t)$ is strictly positive, which is guaranteed for $\hat{y}(t) < \gamma$. Therefore $|\hat{y}(0)| < 1$ implies that $|\hat{y}(t)| < 1$ and $k(t) \geq 2(\gamma^2 - 1)$ for all $t > 0$.

In Figure 2 (top), we plot the training curves of one-sample QNNs with $p = 320$ and varying $\gamma = 1.2, 1.4, 2.0, 4.0, 8.0$ with the same learning rate $\eta = 1e-3/p$. As predicted in Corollary 4.3, the rate of convergence increases with the scaling factor γ. The proof of the corollary additionally implies that $k(t)$ depends on $\hat{y}(t)$: the convergence rate changes over time as the prediction $\hat{y}$ changes. Therefore, despite the linear convergence, the dynamics is different from that of kernel regression, where the kernel remains constant during training in the limit $p \to \infty$.

In Figure 2 (bottom), we plot the empirical rate of convergence $-\frac{d}{dt} \ln L(t)$ against the rate predicted by $\hat{y}$. Each data point is calculated for QNNs with different γ at different time steps by differentiating the logarithms of the training curves. The scatter plot displays an approximately linear dependency, indicating the proposed asymptotic dynamics is capable of predicting how the convergence rate changes during training, which is beyond the explanatory power of the kernel regression model. Note that the slope of the linear relation is not exactly one. This is because we choose a learning rate much smaller than η in the corollary statement to simulate the dynamics of gradient flow.

QNNs with one training sample have been considered before (e.g. (Liu et al., 2022a)), where the linear convergence has been shown under the assumption of "frozen QNTK",

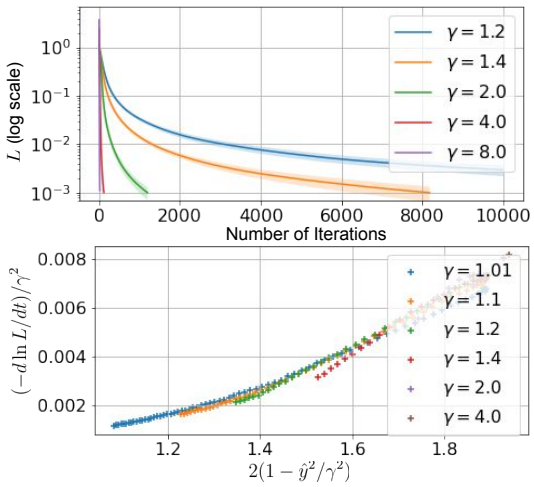

Figure 2. (Top) The training curves of one-sample QNNs with varying γ. The smallest convergence rate $-d \ln L/dt$ during training (i.e. the slope of the training curves under the log scale) increases with γ. (Bottom) The convergence rate $-d \ln L/dt|_{t=T}$ as a function of $2(\gamma^2 - \hat{y}^2(T))$ (jointly scaled by $1/\gamma^2$ for visualization) are evaluated at different time steps T for different γ. The approximately linear dependency shows that the proposed dynamics captures the QNN convergence beyond the explanatory power of the kernel regressions.

namely assuming $\mathbf{K}$, the time derivative of the log residual remains almost constant throughout training. In the corollary above, we provide an end-to-end proof for the one-sample linear convergence without assuming a frozen $\mathbf{K}$. In fact, we observe that in our setting $\mathbf{K} = 2(\gamma^2 - \hat{y}(t))$ changes with $\hat{y}(t)$ (see also Figure 2) and is therefore not frozen.

QNN convergence for $m > 1$ To characterize the convergence of QNNs with $m > 1$, we seek to empirically study the asymptotic dynamics in Line (10). According to Theorem 4.2, the (linear) rate of convergence is lower-bounded by the smallest eigenvalue of $\mathbf{K}_{\text{asym}}(t)$, up to an constant scaling. In Figure 3, we simulate the asymptotic dynamics with various combinations of (γ, d, m), and evaluate the smallest eigenvalue of $\mathbf{K}_{\text{asym}}(t)$ throughout the dynamics (Figure 3, details deferred to Section D). For sufficiently large dimension d, the smallest eigenvalue of $\mathbf{K}_{\text{asym}}$ depends on the ratio between the number of samples and the system dimension m/d and is proportional to the square of the scaling factor γ^2.

Empirically, we observe that the smallest convergence rates for training QNNs are obtained near the global minima (See Figure 6 in the appendix), suggesting the bottleneck of convergence occurs when L is small.

We now give theoretical evidence that, at most of the global minima, the eigenvalues of $\mathbf{K}_{\text{asym}}$ are lower bounded by

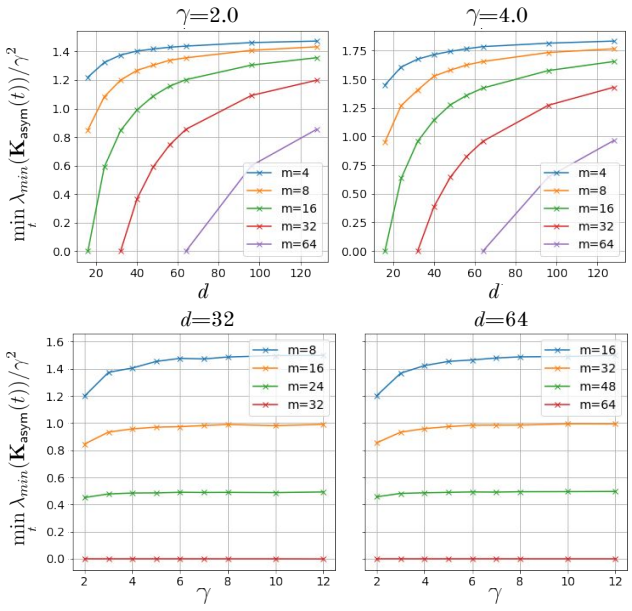

Figure 3. The smallest eigenvalue of $\mathbf{K}_{\text{asym}}$ for the asymptotic dynamics with varying system dimension d, scaling factor γ and number of training samples m. For sufficiently large d, the smallest eigenvalue depends on the ratio m/d and is proportional to the square of the scaling factor γ^2.

$2\gamma^2(1 - 1/\gamma^2 - O(m^2/d))$, suggesting a linear convergence in the neighborhood of these minima. To make this notion precise, we define the uniform measure over global minima as follows: consider a set of pure input states $\{\boldsymbol{\rho}_j = \mathbf{v}_j\mathbf{v}_j^\dagger\}_{j=1}^m$ that are mutually orthogonal (i.e. $\mathbf{v}_i^\dagger\mathbf{v}_j = 0$ if $i \neq j$). For a large dimension d, the global minima of the asymptotic dynamics is achieved when the objective function is 0. Let $\mathbf{u}_j(t)$ (resp. $\mathbf{w}_j(t)$) denote the components of $\mathbf{v}_j$ projected to the positive (resp. negative) subspace of the measurement $\mathbf{M}(t)$ at the global minima. Recall that for a γ-scaled QNN with a Pauli measurement, the predictions $\hat{y}(t) = \gamma\,\mathrm{tr}(\rho_j\mathbf{M}(t)) = \gamma(\mathbf{u}_j^\dagger(t)\mathbf{u}_j(t) - \mathbf{w}_j^\dagger(t)\mathbf{w}_j(t))$. At the global minima, we have $\mathbf{u}_j(t) = \frac{1}{2}(1 \pm 1/\gamma)\hat{\mathbf{u}}_j(t)$ for some unit vector $\hat{\mathbf{u}}_j(t)$ for the j-th training sample with label ± 1. On the other hand, given a set of unit vectors $\{\hat{\mathbf{u}}_j\}_{j=1}^m$ in the positive subspace, there is a corresponding set of $\{\mathbf{u}_j(t)\}_{j=1}^m$ and $\{\mathbf{w}_j(t)\}_{j=1}^m$ such that $L = 0$ for sufficiently large d. By uniformly and independently sampling a set of unit vectors $\{\hat{\mathbf{u}}_j\}_{j=1}^m$ from the $d/2$-dimensional subspace associated with the positive eigenvalues of $\mathbf{M}(t)$, we induce a uniform distribution over all the global minima. The next theorem characterizes $\mathbf{K}_{\text{asym}}$ under such an induced uniform distribution over all the global minima:

Theorem 4.4. *Let $\mathcal{S} = \{(\boldsymbol{\rho}_j, y_j)\}_{j=1}^m$ be a training set with orthogonal pure states $\{\boldsymbol{\rho}_j\}_{j=1}^m$ and equal number of positive and negative labels $y_j \in \{\pm 1\}$. Consider the smallest eigenvalue λ_g of $\mathbf{K}_{\text{asym}}$ at the global minima of the asymptotic dynamics of an over-parameterized QNN with the training set $\mathcal{S}$, scaling factor γ and system dimension d. With probability $\geq 1 - \delta$ over the uniform measure over all the global minima*

$$\lambda_g \geq 2\gamma^2\left(1 - \frac{1}{\gamma^2} - C_2 \max\left\{\frac{m^2}{d}, \frac{m}{d}\log\frac{2}{\delta}\right\}\right), \qquad (12)$$

which is strictly positive for large $\gamma > 1$ and $d = \Omega(\text{poly}(m))$. Here C_2 is a positive constant.

We defer the proof of Theorem 4.4 to Section C in the appendix. A similar notion of a uniform measure over global minima was also used in Canatar et al. (2021). Notice that the uniformness is dependent on the parameterization of the global minima, and the uniform measure over all the global minima is not necessarily the measure induced by random initialization and gradient-based training. Therefore Theorem 4.4 is not a rigorous depiction of the distribution of convergence rate for a randomly-initialized over-parameterized QNN. Yet the prediction of the theorem aligns well with the empirical observations in Figure 3 and suggests that by scaling the QNN measurements, a faster convergence can be achieved: In Figure 4, we simulate p-parameter QNNs with dimension $d = 32$ and 64 with a scaling factor $\gamma = 4.0$ using the same setup as in Figure 1. The training early stops when the average $L(t)$ over the random seeds is less than 1×10^{-2}. In contrast to Figure 1, the convergence rate $-d\ln L/dt$ does not vanish as $L \to 0$, suggesting a simple (constant) scaling of the measurement outcome can lead to convergence within much fewer number of iterations.

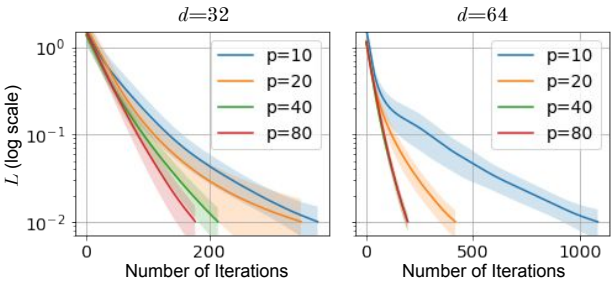

Figure 4. Training curves of QNNs with $\gamma = 4.0$ for learning a 4-sample dataset with labels ± 1. For $p = 10, 20, 40, 80$, the rate of convergence is greater than 0 as $L \to 0$, and it takes less than 1000 iterations for L in most of the instances to convergence below 1×10^{-2}. In contrast, in Figure 1, $L > 1 \times 10^{-1}$ after 10000 iterations despite the increasing number of parameters.

Another implication of Theorem 4.4 is the deviation of QNN dynamics from any kernel regressions. By straightforward calculation, the normalized matrix $\mathbf{K}_{\text{asym}}(0)/\gamma^2$ at the random initialization is independent of the choices of γ. In contrast, the typical value of λ_g/γ^2 in Theorem 4.4

is dependent on γ^2, suggesting non-negligible changes in the matrix $\mathbf{K}_{\mathsf{asym}}(t)$ governing the dynamics of $\mathbf{r}$ for finite scaling factors γ. Such phenomenon is empirically verified in Figure 5 in the appendix.

5. Limitations and Outlook

In the setting of $m > 1$, the proof of the linear convergence of QNN training (Section 4) relies on the convergence of the asymptotic QNN dynamics as a premise. Given our empirical results, an interesting future direction might be to rigorously characterize the condition for the convergence of the asymptotic dynamics. Also we mainly consider (variants of) two-outcome measurements $\mathbf{M}_0$ with two eigensubspaces. It might be interesting to look into measurements with more complicated spectrums and see how the shapes of the spectrums affect the rates of convergence.

A QNN for learning a classical dataset is composed of three parts: a classical-to-quantum encoder, a quantum classifier and a readout measurement. Here we have mainly focused on the stage after encoding, i.e. training a QNN classifier to manipulate the density matrices containing classical information that are potentially too costly for a classically-implemented linear model. Our analysis highlights the necessity for measurement design, assuming the design of the quantum classifier mixes to the full $d \times d$ special unitary group. Our result can be combined with existing techniques of classifier designs (i.e. ansatz design) ((Ragone et al., 2022; Larocca et al., 2021; Wang et al., 2022; You et al., 2022)) by engineering the invariant subspaces, or be combined with encoder designs explored in (Huang et al., 2021; Du et al., 2022).

Acknowledgements

We thank E. Anschuetz, B. T. Kiani, J. Liu and anonymous reviewers for useful comments. This work received support from the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Accelerated Research in Quantum Computing and Quantum Algorithms Team programs, as well as the U.S. National Science Foundation grant CCF-1816695, and CCF-1942837 (CAREER).

Disclaimer

References

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/allen-zhu19a.html.

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

Anschuetz, E. R. and Kiani, B. T. Quantum variational algorithms are swamped with traps. *Nature Communications*, 13(1):1–10, 2022.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., and Wang, R. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/dbc4d84bfcfe2284ba11beffb853a8c4-Paper.pdf.

Bietti, A. and Mairal, J. On the inductive bias of neural tangent kernels. *Advances in Neural Information Processing Systems*, 32, 2019.

Canatar, A., Bordelon, B., and Pehlevan, C. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):1–12, 2021.

Chakrabarti, S., Yiming, H., Li, T., Feizi, S., and Wu, X. Quantum wasserstein generative adversarial networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Chen, Z., Cao, Y., Gu, Q., and Zhang, T. A generalized neural tangent kernel analysis for two-layer neural networks. *Advances in Neural Information Processing Systems*, 33: 13363–13373, 2020.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/ae614c557843b1df326cb29c57225459-Paper.pdf.

Collins, B. and Śniady, P. Integration with respect to the haar measure on unitary, orthogonal and symplectic group. *Communications in Mathematical Physics*, 264(3):773–795, 2006.

Cong, I., Choi, S., and Lukin, M. D. Quantum convolutional neural networks. *Nature Physics*, 15(12):1273–1278, 2019.

Du, Y., Yang, Y., Tao, D., and Hsieh, M.-H. Demystify problem-dependent power of quantum neural networks on multi-class classification. *arXiv preprint arXiv:2301.01597*, 2022.

Dunjko, V. and Briegel, H. J. Machine learning and artificial intelligence in the quantum domain: a review of recent progress. *Reports on Progress in Physics*, 81(7):074001, jun 2018. doi: 10.1088/1361-6633/aab406. URL https://doi.org/10.1088/1361-6633/aab406.

Farhi, E., Neven, H., et al. Classification with quantum neural networks on near term processors. *Quantum Review Letters*, 1(2 (2020)):10–37686, 2020.

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge university press, 2012.

Huang, H.-Y., Broughton, M., Mohseni, M., Babbush, R., Boixo, S., Neven, H., and McClean, J. R. Power of data in quantum machine learning. *Nature Communications*, 12(1), may 2021. doi: 10.1038/s41467-021-22539-9. URL https://doi.org/10.1038%2Fs41467-021-22539-9.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Larocca, M., Ju, N., García-Martín, D., Coles, P. J., and Cerezo, M. Theory of overparametrization in quantum neural networks. *arXiv preprint arXiv:2109.11676*, 2021.

Liu, J., Najafi, K., Sharma, K., Tacchino, F., Jiang, L., and Mezzacapo, A. An analytic theory for the dynamics of wide quantum neural networks. *arXiv preprint arXiv:2203.16711*, 2022a.

Liu, J., Tacchino, F., Glick, J. R., Jiang, L., and Mezzacapo, A. Representation learning via quantum neural tangent kernels. *PRX Quantum*, 3(3):030323, 2022b.

Livni, R., Shalev-Shwartz, S., and Shamir, O. On the computational efficiency of training neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/3a0772443a0739141292a5429b952fe6-Paper.pdf.

Lloyd, S. and Weedbrook, C. Quantum generative adversarial learning. *Physical Review Letters*, 121(4), jul 2018. doi: 10.1103/physrevlett.121.040502. URL https://doi.org/10.1103%2Fphysrevlett.121.040502.

McClean, J. R., Boixo, S., Smelyanskiy, V. N., Babbush, R., and Neven, H. Barren plateaus in quantum neural network training landscapes. *Nature communications*, 9 (1):1–6, 2018.

Oymak, S. and Soltanolkotabi, M. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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

Ragone, M., Braccia, P., Nguyen, Q. T., Schatzki, L., Coles, P. J., Sauvage, F., Larocca, M., and Cerezo, M. Representation theory for geometric quantum machine learning. *arXiv preprint arXiv:2210.07980*, 2022.

Romero, J., Olson, J. P., and Aspuru-Guzik, A. Quantum autoencoders for efficient compression of quantum data. *Quantum Science and Technology*, 2(4):045001, aug 2017. doi: 10.1088/2058-9565/aa8072. URL https://doi.org/10.1088/2058-9565/aa8072.

Russell, B., Rabitz, H., and Wu, R.-B. Control landscapes are almost always trap free: a geometric assessment. *Journal of Physics A: Mathematical and Theoretical*, 50(20):205302, apr 2017. doi: 10.1088/1751-8121/aa6b77. URL https://dx.doi.org/10.1088/1751-8121/aa6b77.

Schuld, M. and Killoran, N. Quantum machine learning in feature hilbert spaces. *Physical review letters*, 122(4):040504, 2019.

Shirai, N., Kubo, K., Mitarai, K., and Fujii, K. Quantum tangent kernel. *arXiv preprint arXiv:2111.02951*, 2021.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12 (4):389–434, 2012.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Wang, X., Liu, J., Liu, T., Luo, Y., Du, Y., and Tao, D. Symmetric pruning in quantum neural networks. *arXiv preprint arXiv:2208.14057*, 2022.

You, X. and Wu, X. Exponentially many local minima in quantum neural networks. In *International Conference on Machine Learning*, pp. 12144–12155. PMLR, 2021.

You, X., Chakrabarti, S., and Wu, X. A convergence theory for over-parameterized variational quantum eigensolvers. *arXiv:2205.12481*, 2022.

Zoufal, C., Lucchi, A., and Woerner, S. Quantum generative adversarial networks for learning and loading random distributions. *npj Quantum Information*, 5(1), nov 2019. doi: 10.1038/s41534-019-0223-2. URL https://doi.org/10.1038%2Fs41534-019-0223-2.

A. Proofs for Section 3

A.1. Proof of Lemma 3.1

Lemma A.1 (Dynamics of the residual vector)**.** *Consider a QNN instance with an ansatz $\mathbf{U}(\boldsymbol{\theta})$ defined as in Line (1), a training dataset $\mathcal{S} = \{(\boldsymbol{\rho}_j, y_j)\}_{j=1}^m$, and a measurement $\mathbf{M}_0$. Under the gradient flow for the objective function $L(\boldsymbol{\theta}) = \frac{1}{2m}\sum_{j=1}^m \left(\operatorname{tr}(\boldsymbol{\rho}_j \mathbf{U}^\dagger(\boldsymbol{\theta})\mathbf{M}_0\mathbf{U}(\boldsymbol{\theta})) - y_j\right)^2$ with learning rate η, the residual vector $\mathbf{r}$ satisfies the differential equation*

$$\frac{d\mathbf{r}(\boldsymbol{\theta}(t))}{dt} = -\frac{\eta}{m}\mathbf{K}(\mathbf{M}(\boldsymbol{\theta}(t)))\mathbf{r}(\boldsymbol{\theta}(t)), \tag{3}$$

where $\mathbf{K}$ is a positive semi-definite matrix-valued function of the parameterized measurement. The (i,j)-th element of $\mathbf{K}$ is defined as

$$\sum_{l=1}^{p}\left(\operatorname{tr}\left(\mathrm{i}[\mathbf{M}(\boldsymbol{\theta}(t)), \boldsymbol{\rho}_i]\mathbf{H}_l\right)\operatorname{tr}\left(\mathrm{i}[\mathbf{M}(\boldsymbol{\theta}(t)), \boldsymbol{\rho}_j]\mathbf{H}_l\right)\right). \tag{4}$$

Here $\mathbf{H}_l := \mathbf{U}_0^\dagger\mathbf{U}_{1:l-1}^\dagger(\boldsymbol{\theta})\mathbf{H}^{(l)}\mathbf{U}_{1:l-1}(\boldsymbol{\theta})\mathbf{U}_0$, is a function of $\boldsymbol{\theta}$ with $\mathbf{U}_{1:r}(\boldsymbol{\theta})$ being the shorthand for $\mathbf{U}_r\exp(-i\theta_r\mathbf{H}^{(r)})\cdots\mathbf{U}_1\exp(-i\theta_1\mathbf{H}^{(1)})$.

Proof. For succinctness, we drop the dependency on $\boldsymbol{\theta}(t)$ when there are no ambiguity. The unitary $\mathbf{U}_{r:p}(\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}_l$ for $l \geq r$:

$$\frac{\partial\mathbf{U}_{r:p}}{\partial\theta_l} = \mathbf{U}_{l:p}(\boldsymbol{\theta})(-\mathrm{i}\mathbf{H}^{(l)})\mathbf{U}_{r:l-1}(\boldsymbol{\theta}) = -\mathrm{i}\mathbf{U}_{l:p}\mathbf{H}^{(l)}\mathbf{U}_{l:p}^\dagger\mathbf{U}_{r:p}. \tag{13}$$

Therefore for all $l \in [p]$

$$\begin{aligned}
\frac{\partial\mathbf{M}(\boldsymbol{\theta}(t))}{\partial\theta_l} &= \mathbf{U}_0^\dagger\left(\frac{\partial\mathbf{U}_{1:p}}{\partial\theta_l}\right)^\dagger\mathbf{M}_0\mathbf{U}_{1:p}\mathbf{U}_0 + \mathbf{U}_0^\dagger\mathbf{U}_{1:p}^\dagger\mathbf{M}_0\frac{\partial\mathbf{U}_{1:p}}{\partial\theta_l}\mathbf{U}_0, &(14)\\
&= \mathrm{i}(\mathbf{U}_0^\dagger\mathbf{U}_{1:p}^\dagger\mathbf{U}_{l:p}\mathbf{H}^{(l)}\mathbf{U}_{l:p}^\dagger\mathbf{M}_0\mathbf{U}_{1:p}\mathbf{U}_0) - \mathrm{i}(\mathbf{U}_0^\dagger\mathbf{U}_{1:p}^\dagger\mathbf{M}_0\mathbf{U}_{l:p}\mathbf{H}^{(l)}\mathbf{U}_{l:p}^\dagger\mathbf{U}_{1:p}\mathbf{U}_0), &(15)\\
&= \mathrm{i}[\mathbf{H}_l, \mathbf{M}(\boldsymbol{\theta}(t))]. &(16)
\end{aligned}$$

By the chain rule with matrix parameters, we have

$$\frac{\partial L(\boldsymbol{\theta}(t))}{\partial\theta_l} = \operatorname{tr}\left(\nabla_{\mathbf{M}}L\frac{\partial\mathbf{M}}{\partial\theta_l}\right) = \mathrm{i}\operatorname{tr}(\nabla_{\mathbf{M}}L[\mathbf{H}_l, \mathbf{M}(\boldsymbol{\theta}(t))]). \tag{17}$$

Furthermore, due to the gradient flow dynamics,

$$\begin{aligned}
\frac{d\mathbf{M}(\boldsymbol{\theta}(t))}{dt} &= \sum_{l=1}^{p}\frac{d\theta_l}{dt}\frac{\partial\mathbf{M}(\boldsymbol{\theta}(t))}{\partial\theta_l} = -\eta\sum_{l=1}^{p}\frac{\partial L(\boldsymbol{\theta}(t))}{\partial\theta_l}\frac{\partial\mathbf{M}(\boldsymbol{\theta}(t))}{\partial\theta_l}, &(18)\\
&= \eta\sum_{l=1}^{p}\operatorname{tr}(\nabla_{\mathbf{M}}L[\mathbf{H}_l, \mathbf{M}(\boldsymbol{\theta}(t))])[\mathbf{H}_l, \mathbf{M}(\boldsymbol{\theta}(t))]. &(19)
\end{aligned}$$

By plugging in $\nabla_{\mathbf{M}}L = -\frac{1}{m}\sum_{j=1}^m r_j\boldsymbol{\rho}_j$, we show that the parameterized measurement $\mathbf{M}(\boldsymbol{\theta}) = \mathbf{U}^\dagger(\boldsymbol{\theta})\mathbf{M}_0\mathbf{U}(\boldsymbol{\theta})$ follows the dynamics

$$d\mathbf{M}(\boldsymbol{\theta}(t))/dt = \frac{\eta}{m}\sum_{l=1}^{p}\operatorname{tr}\left(\sum_{j=1}^{m}r_j\boldsymbol{\rho}_j\mathrm{i}[\mathbf{H}_l, \mathbf{M}(\boldsymbol{\theta}(t))]\right)\mathrm{i}[\mathbf{H}_l, \mathbf{M}(\boldsymbol{\theta}(t))]. \tag{20}$$

By definition $r_i := y_i - \hat{y}_i$, and

$$\frac{dr_i}{dt} = -\frac{d\,\mathrm{tr}(\boldsymbol{\rho}_i \mathbf{M}(\boldsymbol{\theta}(t)))}{dt} = -\,\mathrm{tr}\left(\boldsymbol{\rho}_i \frac{d\mathbf{M}(\boldsymbol{\theta}(t))}{dt}\right) \tag{21}$$

$$= -\frac{\eta}{m}\sum_{l=1}^{p} \mathrm{tr}\left(\sum_{j=1}^{m} r_j \boldsymbol{\rho}_j \mathrm{i}[\mathbf{H}_l, \mathbf{M}(\boldsymbol{\theta}(t))]\right) \mathrm{tr}\left(\boldsymbol{\rho}_i \mathrm{i}[\mathbf{H}_l, \mathbf{M}(\boldsymbol{\theta}(t))]\right) \tag{22}$$

$$= -\frac{\eta}{m}\sum_{j=1}^{m} r_j \Big(\mathrm{tr}\left(\boldsymbol{\rho}_i \mathrm{i}[\mathbf{H}_l, \mathbf{M}(\boldsymbol{\theta}(t))]\right) \mathrm{tr}\left(\boldsymbol{\rho}_j \mathrm{i}[\mathbf{H}_l, \mathbf{M}(\boldsymbol{\theta}(t))]\right)\Big) \tag{23}$$

$$= -\frac{\eta}{m}\sum_{j=1}^{m} r_j \Big(\mathrm{tr}\left(\mathrm{i}[\mathbf{M}(\boldsymbol{\theta}(t)), \boldsymbol{\rho}_i]\mathbf{H}_l\right) \mathrm{tr}\left(\mathrm{i}[\mathbf{M}(\boldsymbol{\theta}(t)), \boldsymbol{\rho}_j]\mathbf{H}_l\right)\Big). \tag{24}$$

The last equality is due to the cyclicity of the trace operation. Making the identification $K_{ij}(\mathbf{M}(\boldsymbol{\theta}(t))) = \big(\mathrm{tr}\left(\mathrm{i}[\mathbf{M}(\boldsymbol{\theta}(t)), \boldsymbol{\rho}_i]\mathbf{H}_l\right) \mathrm{tr}\left(\mathrm{i}[\mathbf{M}(\boldsymbol{\theta}(t)), \boldsymbol{\rho}_j]\mathbf{H}_l\right)\big)$, we have

$$\frac{d\mathbf{r}(\boldsymbol{\theta}(t))}{dt} = -\frac{\eta}{m}\mathbf{K}(\mathbf{M}(\boldsymbol{\theta}(t)))\mathbf{r}(\boldsymbol{\theta}(t)). \tag{25}$$

□

A.2. Proof of Theorem 3.2

Proof. The mean squared loss function $L(\boldsymbol{\theta}(t))$ can be expressed as $\frac{1}{2m}\mathbf{r}(\boldsymbol{\theta}(t))^T\mathbf{r}(\boldsymbol{\theta}(t))$. Using Lemma 3.1, the rate of convergence can be lower-bounded as

$$\frac{1}{L(\boldsymbol{\theta}(t))}\frac{dL(\boldsymbol{\theta}(t))}{dt} \tag{26}$$

$$= \frac{1}{\mathbf{r}(\boldsymbol{\theta}(t))^T\mathbf{r}(\boldsymbol{\theta}(t))}\frac{d}{dt}\mathbf{r}(\boldsymbol{\theta}(t))^T\mathbf{r}(\boldsymbol{\theta}(t)), \tag{27}$$

$$= -\frac{2\eta}{m}\cdot\frac{\mathbf{r}(\boldsymbol{\theta}(t))^T\mathbf{K}(\boldsymbol{\theta}(t))\mathbf{r}(\boldsymbol{\theta}(t))}{\mathbf{r}(\boldsymbol{\theta}(t))^T\mathbf{r}(\boldsymbol{\theta}(t))}, \tag{28}$$

$$\geq -\frac{2\eta}{m}\lambda_{\max}(\mathbf{K}(\boldsymbol{\theta}(t))). \tag{29}$$

The positive semi-definiteness of $\mathbf{K}(\boldsymbol{\theta}(t))$ suggests that $\lambda_{\max}(\mathbf{K}(\boldsymbol{\theta}(t))) \leq \mathrm{tr}(\mathbf{K}(\boldsymbol{\theta}(t)))$. We now proceed to bound $\mathrm{tr}(\mathbf{K}(\boldsymbol{\theta}(t)))$. Since the eigenvalues of $\mathbf{M}_0$ and $\mathbf{M}(\boldsymbol{\theta})$ all lie in $\{\pm 1\}$, $\mathbf{M}(\boldsymbol{\theta}(t))$ decomposes into the difference of to projections, $\boldsymbol{\Pi}_+(\boldsymbol{\theta}(t))$ and $\boldsymbol{\Pi}_-(\boldsymbol{\theta}(t))$, projecting onto the subspaces associated with eigenvalues of $+1$ and -1 respectively. When $\hat{y}_j$ approaches y_j, the input state $\boldsymbol{\rho}_j$ lies almost completely in one of the eigen-subspaces, leading to a vanishing commutator $\mathrm{i}[\mathbf{M}(\boldsymbol{\theta}(t)), \boldsymbol{\rho}_j]$ such that $K_{jj}(\boldsymbol{\theta}(t))$ approaches zero:

Let $\mathbf{v}_j$ be the statevector representation of the pure state $\boldsymbol{\rho}_j$, such that $\boldsymbol{\rho}_j = \mathbf{v}_j\mathbf{v}_j^\dagger$. Vector $\mathbf{v}_j$ decomposes into the components within the positive and negative eigen-subspaces of $\mathbf{M}(\boldsymbol{\theta}(t))$: $\mathbf{v}_j = \mathbf{u}_j(\boldsymbol{\theta}(t)) + \mathbf{w}_j(\boldsymbol{\theta}(t))$, where $\mathbf{u}_j(\boldsymbol{\theta}(t)) = \boldsymbol{\Pi}_+(\boldsymbol{\theta}(t))\mathbf{v}_j$ and $\mathbf{w}_j(\boldsymbol{\theta}(t)) = \boldsymbol{\Pi}_-(\boldsymbol{\theta}(t))\mathbf{v}_j$. In the following we omit the arguments $\boldsymbol{\theta}(t)$ in $\mathbf{u}_j$ and $\mathbf{v}_j$ for succinctness, but the time dependence is to be implicitly understood. The commutator between the parameterized measurement and the input state can be written as $[\mathbf{M}(\boldsymbol{\theta}(t)), \boldsymbol{\rho}_j] = 2(\mathbf{u}_j\mathbf{w}_j^\dagger - \mathbf{w}_j\mathbf{u}_j^\dagger)$. Therefore

$$|\,\mathrm{tr}(\mathrm{i}[\mathbf{M}, \boldsymbol{\rho}_j]\mathbf{H}_l)| \leq 4\,\|\mathbf{H}_l\|_{\mathsf{op}}\,\|\mathbf{u}_j\|\,\|\mathbf{w}_j\|\,. \tag{30}$$

Assume without loss of generality that the j-th label y_j is $+1$. Then $\|\mathbf{u}_j\|^2 + \|\mathbf{w}_j\|^2 = \|\mathbf{v}_j\|^2 = 1$ by definition, and $\|\mathbf{u}_j\|^2 - \|\mathbf{w}_j\|^2 = \mathrm{tr}(\mathbf{M}(\boldsymbol{\theta}(t))\boldsymbol{\rho}_j) = y_j - r_j = 1 - r_j$. Then $\|\mathbf{w}_j\|^2 = |r_j|/2$, and $\|\mathbf{u}_j\|^2\|\mathbf{w}_j\|^2 = (1 - |r_j|/2)|r_j|/2$.

Therefore we have,

$$K_{jj}(\boldsymbol{\theta}(t)) = \sum_{l=1}^{p} \mathrm{tr}^2(\mathrm{i}[\mathbf{M}(\boldsymbol{\theta}(t)), \boldsymbol{\rho}_j]\mathbf{H}_l) \tag{31}$$

$$\leq 16 \sum_{l=1}^{p} \|\mathbf{H}_l\|_{\mathsf{op}}^2 \frac{|r_j|}{2}(1 - \frac{|r_j|}{2}) \tag{32}$$

$$\leq 16 \sum_{l=1}^{p} \|\mathbf{H}_l\|_{\mathsf{op}}^2 \frac{|r_j|}{2} \tag{33}$$

As a result

$$\frac{1}{L(\boldsymbol{\theta}(t))} \frac{dL(\boldsymbol{\theta}(t))}{dt} \tag{34}$$

$$\geq -\frac{2\eta}{m} \mathrm{tr}(\mathbf{K}(\boldsymbol{\theta}(t))) \geq -\frac{2\eta}{m} \sum_{i=1}^{m} K_{ii} \tag{35}$$

$$\geq -\frac{16\eta}{m} \sum_{l=1}^{p} \|\mathbf{H}_l\|_{\mathsf{op}}^2 \sum_{i=1}^{m} |r_j| \tag{36}$$

$$\geq -16\sqrt{2}\eta \sum_{l=1}^{p} \|\mathbf{H}_l\|_{\mathsf{op}}^2 \sqrt{L(\boldsymbol{\theta}(t))} \tag{37}$$

$$= -16\sqrt{2}\eta \sum_{l=1}^{p} \left\|\mathbf{H}^{(l)}\right\|_{\mathsf{op}}^2 \sqrt{L(\boldsymbol{\theta}(t))}. \tag{38}$$

Here we use the fact that $\sum_{j=1}^{m} |r_j| \leq \sqrt{m}\sqrt{\mathbf{r}(\boldsymbol{\theta}(t))^T \mathbf{r}(\boldsymbol{\theta}(t))} = \sqrt{2m^2 L(\boldsymbol{\theta}(t))}$.

The theorem statement follows directly by integrating the inequality above:

$$L(\boldsymbol{\theta}(t))^{-\frac{3}{2}} dL(\boldsymbol{\theta}(t)) \geq -24\eta \sum_{l=1}^{p} \left\|\mathbf{H}^{(l)}\right\|_{\mathsf{op}}^2 dt \tag{39}$$

$$\implies -2d(L(\boldsymbol{\theta}(t))^{-1/2}) \geq -24\eta \sum_{l=1}^{p} \left\|\mathbf{H}^{(l)}\right\|_{\mathsf{op}}^2 dt \tag{40}$$

$$\implies L(\boldsymbol{\theta}(T))^{-\frac{1}{2}} - L(\boldsymbol{\theta}(0))^{-\frac{1}{2}} \leq 12\eta \sum_{l=1}^{p} \left\|\mathbf{H}^{(l)}\right\|_{\mathsf{op}}^2 T \tag{41}$$

$$\implies L(\boldsymbol{\theta}(T))^{-\frac{1}{2}} - c_0 \leq c_1 T \tag{42}$$

□

Note that the same "at most sublinear convergence" holds for a measurement $\mathbf{M}_0$ such that $\mathbf{M}_0 = \mathbf{\Pi}_+ - \mathbf{\Pi}_-$ and $\mathbf{\Pi}_+ + \mathbf{\Pi}_- + \mathbf{\Pi}_0 = \mathbf{I}$ for some non-zero projection $\mathbf{\Pi}_0$. The proof still holds with the following modification: define

$s_j := \|\mathbf{u}_j\|^2 + \|\mathbf{w}_j\|^2 \leq 1$, we have

$$\begin{aligned}
&\|\mathbf{u}_j\|^2 \cdot \|\mathbf{w}_j\|^2 \\
=&(\frac{s_j - y_j + r_j}{2})(\frac{s_j + y_j - r_j}{2}) \\
=&\frac{s_j^2 - (y_j - r_j)^2}{4} \\
\leq&\frac{1 - (y_j - r_j)^2}{4} \\
=&\frac{(1 - y_j)^2 + 2y_j r_j - r_j^2}{4} \\
=&\frac{y_j r_j}{2} - \frac{r_j^2}{4} \leq \frac{y_j r_j}{2} \\
=&\frac{|r_j|}{2}
\end{aligned}$$

The last equality follows from the fact that $r_j \geq 0$ (resp. $r_j \leq 0$) for $y_j = 1$ (resp. $y_j = -1$).

B. Proofs for Asymptotic Dynamics

B.1. Proof of Lemma 4.1

Lemma B.1 (Decomposition of the residual dynamics)**.** *Let $\mathcal{S}$ be a training set with m samples $\{(\boldsymbol{\rho}_j, y_j)\}_{j=1}^m$, and let $\mathbf{U}(\boldsymbol{\theta})$ be a p-parameter ansatz generated by a non-zero $\mathbf{H}$ as in Line 2. Consider a QNN instance with a training set $\mathcal{S}$, ansatz $\mathbf{U}(\boldsymbol{\theta})$ and a measurement $\mathbf{M}_0$. Under the gradient flow with $\eta = \frac{m}{p}\frac{d^2-1}{\mathrm{tr}(\mathbf{H}^2)}$, the residual vector $\mathbf{r}(t)$ as a function of time t through $\boldsymbol{\theta}(t)$ evolves as*

$$\frac{d\mathbf{r}(t)}{dt} = -(\mathbf{K}_{\mathsf{asym}}(t) + \mathbf{K}_{\mathsf{pert}}(t))\mathbf{r}(t) \tag{6}$$

where both $\mathbf{K}_{\mathsf{asym}}$ and $\mathbf{K}_{\mathsf{pert}}$ are functions of time through the parameterized measurement $\mathbf{M}(\boldsymbol{\theta}(t))$, such that

$$(\mathbf{K}_{\mathsf{asym}}(t))_{ij} := \mathrm{tr}\left(i[\mathbf{M}(t), \boldsymbol{\rho}_i]\, i[\mathbf{M}(t), \boldsymbol{\rho}_j]\right), \tag{7}$$

$$(\mathbf{K}_{\mathsf{pert}}(t))_{ij} := \mathrm{tr}\left(i[\mathbf{M}(t), \boldsymbol{\rho}_i] \otimes i[\mathbf{M}(t), \boldsymbol{\rho}_j]\Delta(t)\right). \tag{8}$$

Here $\Delta(t)$ is a $d^2 \times d^2$ Hermitian as a function of t through $\boldsymbol{\theta}(t)$.

Throughout the proof, we make use of the following notations. Let $\mathcal{H}$ be a d-dimensional Hilbert space, and let $\{\mathbf{e}_a\}_{a\in[d]}$ be a basis of $\mathcal{H}$. We use $\mathbf{I}_{d\times d}$ denote the identity matrix $\sum_{a\in[d]} \mathbf{e}_a\mathbf{e}_a^\dagger$. We use $\otimes$ for kronecker products on vectors, matrices and Hilbert spaces. For the $d^2 \times d^2$-dimensional product space $\mathcal{H} \otimes \mathcal{H}$, let $\mathbf{W}_{d^2\times d^2}$ denote the swap matrix $\sum_{a,b\in[d]} \mathbf{e}_a\mathbf{e}_b^\dagger \otimes \mathbf{e}_b\mathbf{e}_a^\dagger$.

We will also make use of the well-known integration formula with respect to the haar measure over d-dimensional unitaries (see e.g. Collins & Śniady (2006) for more details).

Proof. As proven in Lemma 3.1, we track the dynamics of the parameterized measurement $\mathbf{M}(\boldsymbol{\theta})$:

$$\frac{d\mathbf{M}(\boldsymbol{\theta})}{dt} = \sum_{l=1}^{p} \frac{d\theta_l}{dt} \cdot \frac{\partial \mathbf{M}(\boldsymbol{\theta})}{\partial \theta_l} \tag{43}$$

$$= \sum_{l=1}^{p} (-\eta) \operatorname{tr}\left(i[\mathbf{H}_l, \mathbf{M}(\boldsymbol{\theta})] \nabla_{\mathbf{M}} L\right) i[\mathbf{H}_l, \mathbf{M}(\boldsymbol{\theta})] \tag{44}$$

$$= \sum_{l=1}^{p} \eta \operatorname{tr}\left(i[\nabla_{\mathbf{M}} L, \mathbf{M}(\boldsymbol{\theta})] \mathbf{H}_l\right) i[\mathbf{H}_l, \mathbf{M}(\boldsymbol{\theta})] \tag{45}$$

$$= \sum_{l=1}^{p} \eta i[\operatorname{tr}\left(i[\nabla_{\mathbf{M}} L, \mathbf{M}(\boldsymbol{\theta})] \mathbf{H}_l\right) \mathbf{H}_l, \mathbf{M}(\boldsymbol{\theta})] \tag{46}$$

$$= \sum_{l=1}^{p} \eta i[\operatorname{tr}_1\left((i[\nabla_{\mathbf{M}} L, \mathbf{M}(\boldsymbol{\theta})] \otimes \mathbf{I})(\mathbf{H}_l \otimes \mathbf{H}_l)\right), \mathbf{M}(\boldsymbol{\theta})]. \tag{47}$$

Here $\operatorname{tr}_1(\cdot)$ is the partial trace: Given the product of two Hilbert spaces $\mathcal{H}_1 \otimes \mathcal{H}_2$, the partial trace on the first Hilbert space is a linear mapping such that

$$\operatorname{tr}_1\left(\mathbf{A} \otimes \mathbf{B}\right) = \operatorname{tr}(\mathbf{A})\mathbf{B}$$

for any Hermitians $\mathbf{A}$ and $\mathbf{B}$ on the spaces $\mathcal{H}_1$ and $\mathcal{H}_2$. By linearity,

$$\operatorname{tr}_1\left(\sum_l \mathbf{A}_l \otimes \mathbf{B}_l\right) = \sum_l \operatorname{tr}(\mathbf{A}_l)\mathbf{B}_l$$

for any Hermitians $\{\mathbf{A}_l\}$ and $\{\mathbf{B}_l\}$ on the spaces $\mathcal{H}_1$ and $\mathcal{H}_2$.

Let $Z(\mathbf{H}, d)$ denote the ratio $\frac{\operatorname{tr}(\mathbf{H}^2)}{d^2-1}$, the learning rate η can be expressed as $\frac{m}{pZ(\mathbf{H},d)}$. Let $\mathbf{Y}(\boldsymbol{\theta}(t))$ denote the normalized $d^2 \times d^2$-complex matrix $\frac{1}{pZ(\mathbf{H},d)} \sum_{l=1}^{p} \mathbf{H}_l \otimes \mathbf{H}_l$ for $\mathbf{H}_l$ defined in Lemma 3.1 and let $\mathbf{Y}^\star$ denote $\mathbf{W}_{d^2 \times d^2} - \frac{1}{d}\mathbf{I}_{d^2 \times d^2}$, the asymptotic version of $\mathbf{Y}$. We can accordingly decompose the dynamics into the asymptotic dynamics and the deviation (perturbation) from the asymptotic dynamics:

$$\frac{d\mathbf{M}(\boldsymbol{\theta})}{dt} = (\eta p Z(\mathbf{H}, d)) i[\operatorname{tr}_1\left((i[\nabla_{\mathbf{M}} L, \mathbf{M}(\boldsymbol{\theta})] \otimes \mathbf{I})\mathbf{Y}\right), \mathbf{M}(\boldsymbol{\theta})] \tag{48}$$

$$= (\eta p Z(\mathbf{H}, d)) i[\operatorname{tr}_1\left((i[\nabla_{\mathbf{M}} L, \mathbf{M}(\boldsymbol{\theta})] \otimes \mathbf{I})\mathbf{Y}^\star\right), \mathbf{M}(\boldsymbol{\theta})] \tag{49}$$

$$+ (\eta p Z(\mathbf{H}, d)) i[\operatorname{tr}_1\left((i[\nabla_{\mathbf{M}} L, \mathbf{M}(\boldsymbol{\theta})] \otimes \mathbf{I})(\mathbf{Y}(\boldsymbol{\theta}(t)) - \mathbf{Y}^\star)\right), \mathbf{M}(\boldsymbol{\theta})] \tag{50}$$

$$= (\eta p Z(\mathbf{H}, d)) i[(i[\nabla_{\mathbf{M}} L, \mathbf{M}(\boldsymbol{\theta})], \mathbf{M}(\boldsymbol{\theta})] \tag{51}$$

$$+ (\eta p Z(\mathbf{H}, d)) i[\operatorname{tr}_1\left((i[\nabla_{\mathbf{M}} L, \mathbf{M}(\boldsymbol{\theta})] \otimes \mathbf{I})(\mathbf{Y}(\boldsymbol{\theta}(t)) - \mathbf{Y}^\star)\right), \mathbf{M}(\boldsymbol{\theta})] \tag{52}$$

$$= - (\eta p Z(\mathbf{H}, d))[\mathbf{M}(\boldsymbol{\theta}), [\mathbf{M}(\boldsymbol{\theta}), \nabla_{\mathbf{M}} L]] \tag{53}$$

$$- (\eta p Z(\mathbf{H}, d))[\mathbf{M}(\boldsymbol{\theta}), \operatorname{tr}_1\left(([\mathbf{M}(\boldsymbol{\theta}), \nabla_{\mathbf{M}} L] \otimes \mathbf{I})(\mathbf{Y}(\boldsymbol{\theta}(t)) - \mathbf{Y}^\star)\right)] \tag{54}$$

Plugging in that $\nabla_{\mathbf{M}} L(\mathbf{M}(\boldsymbol{\theta})) = -\frac{1}{m}\sum_{i=1}^{m} r_i \boldsymbol{\rho}_i$ with the residual $r_i := y_i - \hat{y}_i = \operatorname{tr}(\mathbf{M}(\boldsymbol{\theta})\rho_i) - y_i$:

$$\frac{d\mathbf{M}(\boldsymbol{\theta})}{dt} = \sum_{j=1}^{m} r_j [\mathbf{M}(\boldsymbol{\theta}), [\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_j]] \tag{55}$$

$$+ \sum_{j=1}^{m} r_j [\mathbf{M}(\boldsymbol{\theta}), \operatorname{tr}_1\left(([\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_j] \otimes \mathbf{I})(\mathbf{Y}(\boldsymbol{\theta}(t)) - \mathbf{Y}^\star)\right)] \tag{56}$$

Trace after multiplying $\boldsymbol{\rho}_i$ on both sides:

$$\frac{dr_i}{dt} = -\operatorname{tr}(\boldsymbol{\rho}_i \frac{d\mathbf{M}(\boldsymbol{\theta})}{dt}) = -\sum_{j=1}^{m} r_j \operatorname{tr}\left(\boldsymbol{\rho}_i [\mathbf{M}(\boldsymbol{\theta}), [\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_j]]\right) \tag{57}$$

$$- \sum_{j=1}^{m} r_j \operatorname{tr}\left(\boldsymbol{\rho}_i [\mathbf{M}(\boldsymbol{\theta}), \operatorname{tr}_1\left(([\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_j] \otimes \mathbf{I})(\mathbf{Y}(\boldsymbol{\theta}(t)) - \mathbf{Y}^\star)\right)]\right) \tag{58}$$

The lemma follows directly from rearranging: for the first term,

$$-\sum_{j=1}^{m} r_j \operatorname{tr}(\boldsymbol{\rho}_i[\mathbf{M}(\boldsymbol{\theta}), [\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_j]]) \tag{59}$$

$$= -\sum_{j=1}^{m} r_j \operatorname{tr}([\boldsymbol{\rho}_i, \mathbf{M}(\boldsymbol{\theta})][\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_j]) \tag{60}$$

$$= -\sum_{j=1}^{m} r_j \operatorname{tr}(i[\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_i] i[\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_j]). \tag{61}$$

For the second term,

$$-\sum_{j=1}^{m} r_j \operatorname{tr}\left(\boldsymbol{\rho}_i[\mathbf{M}(\boldsymbol{\theta}), \operatorname{tr}_1\left(([\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_j] \otimes \mathbf{I})(\mathbf{Y}(\boldsymbol{\theta}(t)) - \mathbf{Y}^\star)\right)]\right) \tag{62}$$

$$= -\sum_{j=1}^{m} r_j \operatorname{tr}\left(i[\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_i] \operatorname{tr}_1\left((i[\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_j] \otimes \mathbf{I})(\mathbf{Y}(\boldsymbol{\theta}(t)) - \mathbf{Y}^\star)\right)\right) \tag{63}$$

$$= -\sum_{j=1}^{m} r_j \operatorname{tr}\left((\mathbf{I} \otimes i[\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_i])(i[\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_j] \otimes \mathbf{I})(\mathbf{Y}(\boldsymbol{\theta}(t)) - \mathbf{Y}^\star)\right) \tag{64}$$

$$= -\sum_{j=1}^{m} r_j \operatorname{tr}\left((i[\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_j] \otimes i[\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_i])(\mathbf{Y}(\boldsymbol{\theta}(t)) - \mathbf{Y}^\star)\right) \tag{65}$$

$$= -\sum_{j=1}^{m} r_j \operatorname{tr}\left((i[\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_i] \otimes i[\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_j])(\mathbf{Y}(\boldsymbol{\theta}(t)) - \mathbf{Y}^\star)\right) \tag{66}$$

The last equality follows from the fact that $\mathbf{Y}$ and $\mathbf{Y}^\star$ are invariant under the swapping of spaces. The lemma follows by identifying the matrix $\Delta(t)$ with $\mathbf{Y}(\boldsymbol{\theta}(t)) - \mathbf{Y}^\star$. □

B.2. Proof of Theorem 4.2

Theorem 4.2 (Linear convergence of QNN with mean-square loss)**.** *Let $\mathcal{S}$ be a training set with m samples $\{(\boldsymbol{\rho}_j, y_j)\}_{j=1}^m$, and let $\mathbf{U}(\boldsymbol{\theta})$ be a p-parameter ansatz generated by a non-zero $\mathbf{H}$ as in Line (2). Consider a QNN instance with the training set $\mathcal{S}$, ansatz $\mathbf{U}(\boldsymbol{\theta})$ and a measurement $\mathbf{M}_0$, trained by gradient flow with $\eta = \frac{m}{p}\frac{d^2-1}{\operatorname{tr}(\mathbf{H}^2)}$. Then for sufficiently large number of parameters p, if the smallest eigenvalue of $\mathbf{K}_{\mathsf{asym}}(t)$ is greater than a constant C_0, then with high probability over the random initialization of the periodic ansatz, the loss function converges to zero at a linear rate*

$$L(t) \leq L(0) \exp(-\frac{C_0 t}{2}). \tag{9}$$

Proof. In Lemma 4.1, we decompose the QNN dynamics into the asymptotic term and the perturbation term depending on $\Delta(t) = \mathbf{Y}(\boldsymbol{\theta}(t)) - \mathbf{Y}^\star$. We now show that the use of the terms "asymptotic" and "perturbation" are exact, by showing that $\mathbf{Y}(\boldsymbol{\theta}(t)) - \mathbf{Y}^\star$ vanishes as $p \to \infty$. We make use of the characterization of a similarly-defined quantity in (You et al., 2022), restated as Lemma B.2 and B.4, such that for sufficiently large p, $\|\mathbf{Y}(\boldsymbol{\theta}(t)) - \mathbf{Y}^\star\|_{\mathsf{op}}$ vanishes for all t with high probability over the randomness in $\{\mathbf{U}_l\}_{l=0}^p$. Recall that the perturbation term $\mathbf{K}_{\mathsf{pert}}$ is defined as

$$(\mathbf{K}_{\mathsf{pert}}(t))_{ij} := \operatorname{tr}\left((i[\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_i] \otimes i[\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\rho}_j])(\mathbf{Y}(\boldsymbol{\theta}(t)) - \mathbf{Y}^\star)\right). \tag{67}$$

By choosing sufficiently large p, we have $\|\mathbf{K}_{\mathsf{pert}}(t)\|_{\mathsf{op}} \leq C_0/10$ and therefore the loss function converging to zero at a rate $\geq C_0/2$. □

Lemma B.2 (Concentration at initialization, adapted from Lemma 3.4 in (You et al., 2022))**.** *Over the randomness of ansatz initialization (i.e. for $\{\mathbf{U}_l\}_{l=1}^p$ sampled i.i.d. with respect to the Haar measure), for any initial $\boldsymbol{\theta}(0)$, with probability $1-\delta$:*

$$\|\mathbf{Y}(\boldsymbol{\theta}(0)) - \mathbf{Y}^\star\|_{\mathsf{op}} \leq \frac{1}{\sqrt{p}} \cdot \frac{2\|\mathbf{H}\|_{\mathsf{op}}^2}{Z} \sqrt{2 \log \frac{d^2}{\delta}}. \tag{68}$$

Proof. Define

$$\mathbf{X}_l := \frac{1}{Z(\mathbf{H},d)}\left(\mathbf{U}_{0:l-1}(\boldsymbol{\theta}(0))^\dagger \mathbf{H}\mathbf{U}_{0:l-1}^\dagger(\boldsymbol{\theta}(0))\right)^{\otimes 2} - \mathbf{Y}^\star. \tag{69}$$

By straight-forward calculation (e.g. using results in Collins & Śniady (2006)) we know that X_l is centered (i.e $\mathbb{E}[X_l] = 0$). The set $\{\mathbf{X}_l\}$ can be viewed as independent random matrices as the Haar random unitary removes all the correlation. The matrix on the left-hand side can therefore be expressed as the arithmetic average of p independent random matrices. The square of $\mathbf{X}_l$ is bounded in operator norm:

$$\left\|\mathbf{X}_l^2\right\|_{\mathsf{op}} = \left\|\mathbf{X}_l\right\|_{\mathsf{op}}^2 \leq \left(\frac{\left\|\mathbf{H}\right\|_{\mathsf{op}}^2}{Z} + \frac{d+1}{d}\right)^2 \leq \left(\frac{2\left\|\mathbf{H}\right\|_{\mathsf{op}}^2}{Z(\mathbf{H},d)}\right)^2 \tag{70}$$

where the second inequality follows from the fact that the ratio $g_1 = \left\|\mathbf{H}\right\|_{\mathsf{op}}^2 / \operatorname{tr}(\mathbf{H}^2)$ satisfies that $1 \geq g_1 \geq 1/d$. By Hoeffding's inequality((Tropp, 2012), Thm 1.3), with probability $\geq 1-\delta$,

$$\left\|\mathbf{Y}(\boldsymbol{\theta}(0)) - \mathbf{Y}^\star\right\|_{\mathsf{op}} \leq \frac{1}{\sqrt{p}} \cdot \frac{2\left\|\mathbf{H}\right\|_{\mathsf{op}}^2}{Z(\mathbf{H},d)} \sqrt{\log \frac{2d^2}{\delta}}. \tag{71}$$

□

As we pointed out in the main body, a vanishing perturbation term at initialization is not sufficient to guarantee the term remain perturbative throughout the training. We now show in Lemma B.4 that $\mathbf{Y}(\boldsymbol{\theta}(t)) - \mathbf{Y}^\star$ remain small during training by showing $\mathbf{Y}(\boldsymbol{\theta}(t)) - \mathbf{Y}((0))$ vanishes in the limit $p \to \infty$. But before that, we show that, while the QNN predictions changes much during training, the change in the parameters measured in ℓ_2- or ℓ_∞-norm ($\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(0)\|_2$ or $\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(0)\|_\infty$) vanishes as $p \to \infty$ during the training of QNN:

Lemma B.3 (Slow-varying θ in QNNs)**.** *Suppose that under learning rate $\eta = \frac{m}{pZ(\mathbf{H},d)}$, for all $0 \leq t \leq T$, the loss function $L(\boldsymbol{\theta}(t)) \leq L(\boldsymbol{\theta}(0))\exp(-at)$ for some constant a, then for all $0 \leq t_1, t_2 \leq T$:*

$$\left\|\boldsymbol{\theta}(t_2) - \boldsymbol{\theta}(t_1)\right\|_\infty \leq \frac{1}{p}\frac{\sqrt{2}m\left\|\mathbf{H}\right\|_F\left\|\mathbf{M}\right\|_F\sqrt{L(\boldsymbol{\theta}(0))}}{Z}|t_1 - t_2|, \tag{72}$$

$$\left\|\boldsymbol{\theta}(t_2) - \boldsymbol{\theta}(t_1)\right\|_2 \leq \frac{1}{\sqrt{p}}\frac{\sqrt{2}m\left\|\mathbf{H}\right\|_F\left\|\mathbf{M}\right\|_F\sqrt{L(\boldsymbol{\theta}(0))}}{Z}|t_1 - t_2|. \tag{73}$$

Proof. We first bound the absolute value of the derivative $\frac{d\theta_l}{dt}$:

$$\left|\frac{d\theta_l}{dt}\right| = \eta\left|\frac{\partial L}{\partial \theta_l}\right| = \frac{\eta}{2m}\left|\sum_{i=1}^m r_i \operatorname{tr}(i[\mathbf{M}(\boldsymbol{\theta}), \mathbf{H}_l]\boldsymbol{\rho}_i)\right|. \tag{74}$$

Plugging in $\eta = \frac{m}{pZ}$, we have

$$\left|\frac{d\theta_l}{dt}\right| = \frac{1}{2pZ}\left|\sum_{i=1}^m r_i \operatorname{tr}(i[\mathbf{M}(\boldsymbol{\theta}), \mathbf{H}_l]\boldsymbol{\rho}_i)\right| = \frac{1}{2pZ}\left|\langle \mathbf{r}, \mathbf{a}\rangle\right|, \tag{75}$$

where the vector $\mathbf{a}$ is defined such that $a_j = \mathrm{tr}(i[\mathbf{M}(\boldsymbol{\theta}), \mathbf{H}_l]\boldsymbol{\rho}_j)$ for $j \in [m]$. The ℓ_2-norm of $\mathbf{a}$

$$\|\mathbf{a}\|_2^2 = \sum_{j=1}^{m} \mathrm{tr}^2(i[\mathbf{M}, \mathbf{H}_l]\boldsymbol{\rho}_j) \tag{76}$$

$$= \mathrm{tr}\left((i[\mathbf{M}, \mathbf{H}_l])^{\otimes 2} \sum_{j=1}^{m} \boldsymbol{\rho_j}^{\otimes 2}\right) \tag{77}$$

$$\leq \left\|(i[\mathbf{M}, \mathbf{H}_l])^{\otimes 2}\right\|_F \left\|\sum_{j=1}^{m} \boldsymbol{\rho_j}^{\otimes 2}\right\|_F \tag{78}$$

$$\leq \|i[\mathbf{M}, \mathbf{H}_l]\|_F^2 \left\|\sum_{j=1}^{m} \boldsymbol{\rho_j}^{\otimes 2}\right\|_F \tag{79}$$

$$\leq (2\|\mathbf{M}\|_F \|\mathbf{H}_l\|_F)^2 \sum_{j=1}^{m} \left\|\boldsymbol{\rho_j}^{\otimes 2}\right\|_F \tag{80}$$

$$\leq (2\|\mathbf{M}\|_F \|\mathbf{H}\|_F \sqrt{m})^2. \tag{81}$$

Therefore we can bound $|\frac{d\theta_l}{dt}|$ as

$$|\frac{d\theta_l}{dt}| \leq \frac{1}{2pZ}\|\mathbf{r}\|_2 \|\mathbf{a}\|_2 \tag{82}$$

$$\leq \frac{1}{2pZ}\sqrt{2mL(\boldsymbol{\theta}(t))} \cdot 2\|\mathbf{M}\|_F \|\mathbf{H}\|_F \sqrt{m} \tag{83}$$

$$= \frac{1}{p}\frac{\sqrt{2}m\|\mathbf{M}\|_F \|\mathbf{H}\|_F}{Z}\sqrt{L(\boldsymbol{\theta}(t))} \tag{84}$$

$$\leq \frac{1}{p}\frac{\sqrt{2}m\|\mathbf{M}\|_F \|\mathbf{H}\|_F}{Z}\sqrt{L(\boldsymbol{\theta}(0))}\exp(-at/2) \tag{85}$$

Hence for all $l \in [p]$:

$$|\theta_l(t_2) - \theta_l(t_1)| = |\int_{t_1}^{t_2} dt d\theta_l(t)/dt| \leq \int_{t_1}^{t_2} dt |d\theta_l(t)/dt| \tag{86}$$

$$\leq \int_{t_1}^{t_2} dt \frac{1}{p}\frac{\sqrt{2}m\|\mathbf{M}\|_F \|\mathbf{H}\|_F}{Z}\sqrt{L(\boldsymbol{\theta}(0))}\exp(-at/2) \tag{87}$$

$$\leq \frac{2}{a} \cdot \frac{1}{p}\frac{\sqrt{2}m\|\mathbf{M}\|_F \|\mathbf{H}\|_F}{Z}\sqrt{L(\boldsymbol{\theta}(0))}|\exp(-at_1/2) - \exp(-at_2/2)| \tag{88}$$

$$\leq \frac{1}{p}\frac{\sqrt{2}m\|\mathbf{M}\|_F \|\mathbf{H}\|_F}{Z}\sqrt{L(\boldsymbol{\theta}(0))}|t_1 - t_2| \tag{89}$$

$$\tag{90}$$

The bounds on the ℓ_2- and ℓ_∞-norm follows from direct computation. □

We are now ready to show $\mathbf{Y}(t_2) - \mathbf{Y}(t_1)$ vanishes as $p \to \infty$:

Lemma B.4 (Concentration during training, adapted from Lemma 3.5 in (You et al., 2022))**.** *Suppose that under learning rate $\eta = \frac{m}{pZ(\mathbf{H},d)}$, for all $0 \leq t \leq T$, the loss function $L(\boldsymbol{\theta}(t))$ decreases as $L(\boldsymbol{\theta}(0))\exp(-at)$ then with probability $\geq 1-\delta$, for all $0 \leq t \leq T$: $\|\mathbf{Y}(\boldsymbol{\theta}(t)) - \mathbf{Y}(\boldsymbol{\theta}(0))\|_{\mathsf{op}} \leq C_3 \cdot \frac{T}{\sqrt{p}}$, where C_3 is a constant of T and p.*

Proof. To bound the supremum of the matrix-valued random field, we use an adapted version of the Dudley's inequality:

Claim 1 (Dudley's inequality for matrix-valued random fields, adapted from Theorem 8.1.6 in High-dimensional probability (Vershynin, 2018).). Let $\boldsymbol{\mathcal{R}}$ be a metric space equipped with a metric $\mathbf{d}(\cdot,\cdot)$, and $\mathbf{X} : \boldsymbol{\mathcal{R}} \mapsto \mathbb{R}^{D\times D}$

with subgaussian increments i.e. it satisfies $\mathbf{Pr}[\|\mathbf{X}(r_1) - \mathbf{X}(r_2)\|_{\mathsf{op}} > t] \leq 2D\exp\left(-\frac{t^2}{C_\sigma^2 \mathbf{d}(r_1,r_2)^2}\right)$. Then with probability at least $1 - 2D\exp(-u^2)$ for any subset $\mathcal{S} \subseteq \boldsymbol{\mathcal{R}}$: $\sup_{(r_1,r_2)\in\mathcal{S}} \|\mathbf{X}(r_1) - \mathbf{X}(r_2)\|_{\mathsf{op}} \leq C \cdot C_\sigma \left[\int_0^{\mathrm{diam}(\mathcal{S})} \sqrt{\mathcal{N}(\mathcal{S},\mathbf{d},\epsilon)}\, d\epsilon + u \cdot \mathrm{diam}(\mathcal{S})\right]$ for some constant C, where $\mathcal{N}(\mathcal{S},\mathbf{d},\epsilon)$ is the metric entropy defined as the logarithm of the ϵ-covering number of $\mathcal{S}$ using metric d.

To make use of *Claim 1*, we now establish the sub-gaussian increment of $\mathbf{Y}(\boldsymbol{\theta}(t))$ through the following *Claim 2* by applying *McDiarmid inequality*:

Claim 2 (Sub-gaussianity of $\mathbf{Y}$*)* $\mathbf{Pr}[\|\mathbf{Y}(\boldsymbol{\theta}) - \mathbf{Y}(\mathbf{0})\|_{\mathsf{op}} > t] \leq 2\exp\left(-\frac{-t^2 Z(\mathbf{H},d)^2}{2C_1\|\boldsymbol{\theta}\|_2^2}\right)$ for some constant C_1. Then due to the Haar distribution of the unitaries $\{\mathbf{U}_l\}_{l=0}^p$,

$$\mathbf{Pr}[\|\mathbf{Y}(\boldsymbol{\theta}_2) - \mathbf{Y}(\boldsymbol{\theta}_1)\|_{\mathsf{op}} > t] \leq 2\exp\left(-\frac{-t^2 Z(\mathbf{H},d)^2}{2C_1\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_2^2}\right). \tag{91}$$

To see that *Claim 2* is true, consider an alternative description of $\mathbf{Y}(\boldsymbol{\theta})$. Recall that $\mathbf{Y}(\boldsymbol{\theta})$ is defined as $\mathbf{Y}(\boldsymbol{\theta}) = \frac{1}{pZ(\mathbf{H},d)}\sum_{l=1}^p \mathbf{Y}_l$ with$\mathbf{Y}_l(\boldsymbol{\theta})$ being $\mathbf{H}_l^{\otimes 2}$. We consider a re-parameterization of the random variables $\mathbf{H}_l(\theta)$ by constructing random variables that are identically distributed, but are functions on a different latent probability space. Defining $\mathbf{H}_l$ as $\mathbf{U}_0^\dagger \cdots \mathbf{U}_{l-1}^\dagger \mathbf{H} \mathbf{U}_{l-1} \cdots \mathbf{U}_0$, $\mathbf{Y}$ can be rewritten as:

$$\mathbf{Y}(\boldsymbol{\theta}) = \frac{1}{pZ}\sum_{l=1}^p \left(e^{i\theta_1\mathbf{H}_1}\cdots e^{i\theta_{l-1}\mathbf{H}_{l-1}}\mathbf{H}_l e^{-i\theta_{l-1}\mathbf{H}_{l-1}}\cdots e^{-i\theta_1\mathbf{H}_1}\right)^{\otimes 2}. \tag{92}$$

By the Haar randomness of $\{\mathbf{U}_l\}_{l=1}^p$, we can view $\{\mathbf{H}_l\}_{l=1}^p$ as random Hermitians generated by $\{\mathbf{V}_l\mathbf{H}\mathbf{V}_l^\dagger\}$ for *i.i.d.* Haar random $\{\mathbf{V}_l\}_{l=1}^p$. This variable is identically distributed to $\mathbf{Y}$ and $\mathbf{Y}_l$ can be defined as each term in the sum.

We will apply the well-known McDiarmid inequality (e.g. Theorem 2.9.1 in High-dimensional probability (Vershynin, 2018)) that can be stated as follows: Consider independent random variables $X_1,\ldots,X_k \in \mathcal{X}$. Suppose a random variable $\phi\colon \mathcal{X}^k \to \mathbb{R}$ satisfies the condition that for all $1 \leq j \leq k$ and for all $x_1,\ldots,x_j,\ldots,x_k,x_j' \in \mathcal{X}$,

$$|\phi(x_1,\ldots,x_j,\ldots,x_k) - \phi(x_1,\ldots,x_j',\ldots,x_k)| \leq c_j, \tag{93}$$

then the tails of the distribution satisfy

$$\mathbf{Pr}[|\phi(X_1,\ldots,X_k) - \mathbb{E}\phi| \geq t] \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^k c_i^2}\right). \tag{94}$$

With our earlier re-parameterization we can consider $\mathbf{Y}$ and consequently $\mathbf{Y}_l$ as functions of the randomly sampled Hermitian operators $\mathbf{H}_l$. Define the variable $\mathbf{Y}^{(k)}$ as that obtained by resampling $\mathbf{H}_k$ independently, and $\mathbf{Y}_l^{(k)}$ correspondingly. Finally we define

$$\Delta^{(k)}\mathbf{Y} = \left\|(\mathbf{Y}(\boldsymbol{\theta}) - \mathbf{Y}(0)) - (\mathbf{Y}^{(k)}(\boldsymbol{\theta}) - \mathbf{Y}^{(k)}(0))\right\|_{\mathsf{op}} = \left\|\mathbf{Y}(\boldsymbol{\theta}) - \mathbf{Y}^{(k)}(\boldsymbol{\theta})\right\|_{\mathsf{op}}. \tag{95}$$

Via the triangle inequality,

$$\Delta^{(k)}\mathbf{Y} = \|\mathbf{Y}(\boldsymbol{\theta}) - \mathbf{Y}^{(k)}(\boldsymbol{\theta})\| = \frac{1}{pZ}\|\sum_{l\geq k}\mathbf{Y}_l(\boldsymbol{\theta}) - \mathbf{Y}_l^{(k)}(\boldsymbol{\theta})\| \tag{96}$$

$$\leq \frac{1}{pZ}\sum_{l\geq k}\|\mathbf{Y}_l(\boldsymbol{\theta}) - \mathbf{Y}_l^{(k)}(\boldsymbol{\theta})\|. \tag{97}$$

Then by definition,

$$\begin{aligned}
&\|\mathbf{Y}_l(\theta) - \mathbf{Y}_l^{(k)}(\boldsymbol{\theta})\| \\
=&\|(e^{i\theta_1\mathbf{H}_1}\cdots e^{i\theta_{k-1}\mathbf{H}_{k-1}})^{\otimes 2}\big((e^{i\theta_k\mathbf{H}_k}\mathbf{K}e^{-i\theta_k\mathbf{H}_k})^{\otimes 2} \\
&-(e^{i\theta_k\mathbf{H}'_k}\mathbf{K}e^{-i\theta_k\mathbf{H}'_k})^{\otimes 2}\big)(e^{-i\theta_{k-1}\mathbf{H}_{k-1}}\cdots e^{-i\theta_1\mathbf{H}_1})^{\otimes 2}\| \\
=&\|(e^{i\theta_k\mathbf{H}_k}\mathbf{K}e^{-i\theta_k\mathbf{H}_k})^{\otimes 2} - (e^{i\theta_k\mathbf{H}'_k}\mathbf{K}e^{-i\theta_k\mathbf{H}'_k})^{\otimes 2}\| \\
\leq&\|(e^{i\theta_k\mathbf{H}_k}\mathbf{K}e^{-i\theta_k\mathbf{H}_k})^{\otimes 2} - \mathbf{K}^{\otimes 2}\| + \|(e^{i\theta_k\mathbf{H}'_k}\mathbf{K}e^{-i\theta_k\mathbf{H}'_k})^{\otimes 2} - \mathbf{K}^{\otimes 2}\|.
\end{aligned}$$

where $\mathbf{K} := e^{i\theta_{k+1}\mathbf{H}_{k+1}}\cdots e^{i\theta_{l-1}\mathbf{H}_{l-1}}\mathbf{H}_l e^{-i\theta_{l-1}\mathbf{H}_{l-1}}\cdots e^{-i\theta_{k+1}\mathbf{H}_{k+1}}$. Let $\mathbf{K}(\phi)$ denote $e^{i\phi\mathbf{H}_k}\mathbf{K}e^{-i\phi\mathbf{H}_k}$, we can bound the first term on the righthand side as follows:

$$\begin{aligned}
&\|(e^{i\theta_k\mathbf{H}_k}\mathbf{K}e^{-i\theta_k\mathbf{H}_k})^{\otimes 2} - \mathbf{K}^{\otimes 2}\| \\
=&\|\mathbf{K}(\theta_k)^{\otimes 2} - \mathbf{K}(0)^{\otimes 2}\| \\
=&\|\int_0^{\theta_k} d\phi \frac{d}{d\phi}(\mathbf{K}(\phi)^{\otimes 2})\| \\
\leq&\int_0^{\theta_k} d\phi \|\frac{d}{d\phi}(\mathbf{K}(\phi)^{\otimes 2})\| \\
\leq&4|\theta_k|\|\mathbf{H}_k\|\|\mathbf{K}\|^2.
\end{aligned}$$

The last inequality follows from the fact that

$$\begin{aligned}
&\|\frac{d}{d\phi}\mathbf{K}(\phi)^{\otimes 2}\| \\
=&\|(\exp(i\phi\mathbf{H}_k))^{\otimes 2}\big([-i\mathbf{H}_k,\mathbf{K}]\otimes\mathbf{K} + \mathbf{K}\otimes[-i\mathbf{H}_k,\mathbf{K}]\big)(\exp(-i\phi\mathbf{H}_k))^{\otimes 2}\| \\
=&\|[-i\mathbf{H}_k,\mathbf{K}]\otimes\mathbf{K} + \mathbf{K}\otimes[-i\mathbf{H}_k,\mathbf{K}]\| \\
\leq&4\|\mathbf{H}_k\|\|\mathbf{K}\|^2.
\end{aligned}$$

The same reasoning holds for the term with $\mathbf{H}'_k$. Using the fact that $\|\mathbf{H}_k\| = \|\mathbf{H}'_k\| = \|\mathbf{H}\|$, and we have

$$\|\big(\mathbf{Y}_l(\boldsymbol{\theta}) - \mathbf{Y}_l(\mathbf{0})\big) - \big(\mathbf{Y}_l^{(k)}(\boldsymbol{\theta}) - \mathbf{Y}_l^{(k)}(\mathbf{0})\big)\| \leq 8|\theta_k|\|\mathbf{H}\|\|\mathbf{K}\|^2 = 8|\theta_k|\|\mathbf{H}\|^3.$$

Claim 2 follows from the direct application of McDiarmid inequality.

By Lemma B.3, $\|\boldsymbol{\theta}(t_2) - \boldsymbol{\theta}(t_1)\|_2 \leq \frac{C_L}{\sqrt{p}}|t_2 - t_1|$ with C_L being a constant with respect to p. Plugging this into *Claim 2*, we see that $\mathbf{Y}$ has sub-gaussian increments if we define the metric $\mathbf{d}(t_2, t_1) = \frac{C_L}{\sqrt{p}}\cdot|t_2 - t_1|$, thereby satisfying the conditions for *Claim 1*. Under this metric, the diameter of the interval $[0, T]$ is of order $\frac{T}{\sqrt{p}}$. Applying *Claim 1*, with $u = \sqrt{\log(2d/\delta)}$ to ensure a failure probability at most δ we have

$$\sup_{t\in[0,T]} \|\mathbf{Y}(\theta(t)) - \mathbf{Y}(\theta(0))\|_{\mathsf{op}} \leq C_3\cdot\frac{T}{\sqrt{p}}, \tag{98}$$

where C_3 is a constant of p and T and depends polynomially on other quantities including d and $\log(1/\delta)$. □

C. Proof for Theorem 4.4

In this section, we present the proof for Theorem 4.4 for characterizing the rate of convergence at global minima:

Theorem 4.4. *Let $\mathcal{S} = \{(\boldsymbol{\rho}_j, y_j)\}_{j=1}^m$ be a training set with orthogonal pure states $\{\boldsymbol{\rho}_j\}_{j=1}^m$ and equal number of positive and negative labels $y_j \in \{\pm 1\}$. Consider the smallest eigenvalue λ_g of $\mathbf{K}_{\mathsf{asym}}$ at the global minima of the asymptotic dynamics of an over-parameterized QNN with the training set $\mathcal{S}$, scaling factor γ and system dimension d. With probability $\geq 1-\delta$ over the uniform measure over all the global minima*

$$\lambda_g \geq 2\gamma^2(1 - \frac{1}{\gamma^2} - C_2\max\{\frac{m^2}{d}, \frac{m}{d}\log\frac{2}{\delta}\}), \tag{12}$$

which is strictly positive for large $\gamma > 1$ and $d = \Omega(\mathsf{poly}(m))$. Here C_2 is a positive constant.

We start by presenting a few helper lemma:

C.1. Helper Lemma for $\mathbf{K}_{\mathsf{asym}}$

Lemma C.1. *Let $\mathbf{A}, \mathbf{B}$ be $d \times d$ Hermitians. Let $\|\cdot\|_{\mathsf{op}}$ denote the operator norm of a given Hermitian and let $\circ$ denote the Hadamard product (i.e. the elementwise multiplication) of two matrices, we have*

$$\|\mathbf{A} \circ \mathbf{B}\|_{\mathsf{op}} \leq \|\mathbf{A}\|_{\mathsf{op}} \|\mathbf{B}\|_{\mathsf{op}}. \tag{99}$$

Proof. For any $d \times d$ Hermitian matrix, let $\lambda_i(\cdot)$ denote its i-th smallest eigenvalue. The Hadamard product $\mathbf{A} \circ \mathbf{B}$ is a $d \times d$ principal submatrix of the Kronecker product $\mathbf{A} \otimes \mathbf{B}$, and by the Poincaré separation theorem (see e.g. Corollary 4.3.37 in Horn & Johnson (2012)):

$$\lambda_1(\mathbf{A} \otimes \mathbf{B}) \leq \lambda_i(\mathbf{A} \circ \mathbf{B}) \leq \lambda_{d^2}(\mathbf{A} \otimes \mathbf{B}). \tag{100}$$

The statement follows from the fact that the eigenvalues of $\mathbf{A} \otimes \mathbf{B}$ take the form of $\lambda_i(\mathbf{A})\lambda_j(\mathbf{B})$ for $i, j \in [d]$. □

Lemma C.2 ($\mathbf{K}_{\mathsf{asym}}$ for asymptotic dynamics)**.** *Let $\mathcal{S}$ be a m-sample training set composed of pure states $\{\boldsymbol{\rho}_j = \mathbf{v}_j\mathbf{v}_j^\dagger\}_{j=1}^m$. Let $\mathbf{M}_0$ be a Pauli-like measurement with eigenvalues ± 1 and trace-0. Consider training a QNN with $\mathcal{S}$, measurement $\mathbf{M}_0$ and a scaling factor of γ. The positive semidefinite matrix $\mathbf{K}_{\mathsf{asym}}$ can be expressed entry-wise as*

$$(\mathbf{K}_{\mathsf{asym}})_{ij}(\mathbf{M}(t)) = 8\gamma^2 Re(\mathbf{u}_j^\dagger(t)\mathbf{u}_i(t)\mathbf{w}_i^\dagger(t)\mathbf{w}_j(t)), \tag{101}$$

where $\mathbf{u}_i(t) := \boldsymbol{\Pi}_+(t)\mathbf{v}_i$ (resp. $\mathbf{w}_i(t) := \boldsymbol{\Pi}_-(t)\mathbf{v}_i$) is the projection of $\mathbf{v}_i$ into the postive (resp. negative) subspace of $\mathbf{M}(t) = \gamma(\boldsymbol{\Pi}_+(t) - \boldsymbol{\Pi}_-(t))$. Let $\mathbf{P}(t) := (\mathbf{u}_i^\dagger(t)\mathbf{u}_j(t))_{i.j\in[m]}$ and $\mathbf{N}(t) := (\mathbf{w}_i^\dagger(t)\mathbf{w}_j(t))_{i,j\in[m]}$ be the Gram matrices of $\{\mathbf{u}_i(t)\}_{i=1}^m$ and $\{\mathbf{w}_i(t)\}_{i=1}^m$, we have:

$$\lambda_{\min}(\mathbf{K}_{\mathsf{asym}}(t)) \geq 8\gamma^2\lambda_{\min}(\mathbf{P}(t)) \min_{i\in[m]}(\mathbf{N}_{ii}(t)) \geq 8\gamma^2\lambda_{\min}(\mathbf{P}(t))\lambda_{\min}(\mathbf{N}(t)). \tag{102}$$

Proof. For succinctness, we drop the time dependency t when there are no ambiguities. Calculate the expression of $(\mathbf{K}_{\mathsf{asym}})_{ij}$ for pure states $\boldsymbol{\rho}_i = \mathbf{v}_i\mathbf{v}_i^\dagger$:

$$\begin{aligned}
(\mathbf{K}_{\mathsf{asym}}(\mathbf{M}(t)))_{ij} &= \operatorname{tr}\left(i[\mathbf{M}, \boldsymbol{\rho}_i]\ i[\mathbf{M}, \boldsymbol{\rho}_j]\right) && (103)\\
&= \operatorname{tr}\left(\mathbf{M}^2\boldsymbol{\rho}_i\boldsymbol{\rho}_j\right) + \operatorname{tr}\left(\mathbf{M}^2\boldsymbol{\rho}_j\boldsymbol{\rho}_i\right) - 2\operatorname{tr}\left(\mathbf{M}\boldsymbol{\rho}_i\mathbf{M}\boldsymbol{\rho}_j\right) && (104)\\
&= 2\gamma^2\big(\operatorname{tr}(\boldsymbol{\rho}_i\boldsymbol{\rho}_j) - \operatorname{tr}((\boldsymbol{\Pi}_+ - \boldsymbol{\Pi}_-)\boldsymbol{\rho}_i(\boldsymbol{\Pi}_+ - \boldsymbol{\Pi}_-)\boldsymbol{\rho}_j)\big) && (105)
\end{aligned}$$

Plugging in $\boldsymbol{\rho}_i = \mathbf{v}_i\mathbf{v}_i^\dagger$, we have:

$$\begin{aligned}
\frac{1}{2\gamma^2}(\mathbf{K}_{\mathsf{asym}}(\mathbf{M}(t)))_{ij} &= |\mathbf{u}_i^\dagger\mathbf{u}_j + \mathbf{w}_i^\dagger\mathbf{w}_j|^2 - |(\mathbf{u}_i + \mathbf{w}_i)^\dagger(\boldsymbol{\Pi}_+ - \boldsymbol{\Pi}_-)(\mathbf{u}_j + \mathbf{w}_j)|^2 && (106)\\
&= |\mathbf{u}_i^\dagger\mathbf{u}_j + \mathbf{w}_i^\dagger\mathbf{w}_j|^2 - |(\mathbf{u}_i + \mathbf{w}_i)^\dagger(\mathbf{u}_j - \mathbf{w}_j)|^2 && (107)\\
&= |\mathbf{u}_i^\dagger\mathbf{u}_j + \mathbf{w}_i^\dagger\mathbf{w}_j|^2 - |\mathbf{u}_i^\dagger\mathbf{u}_j - \mathbf{w}_i^\dagger\mathbf{w}_j|^2 && (108)\\
&= 2\mathbf{u}_i^\dagger\mathbf{u}_j \cdot \mathbf{w}_j^\dagger\mathbf{w}_i + 2\mathbf{u}_j^\dagger\mathbf{u}_i \cdot \mathbf{w}_i^\dagger\mathbf{w}_j && (109)\\
&= 4Re(\mathbf{u}_j^\dagger\mathbf{u}_i\mathbf{w}_i^\dagger\mathbf{w}_j), && (110)
\end{aligned}$$

or $(\mathbf{K}_{\mathsf{asym}}(\mathbf{M}(t)))_{ij} = 8\gamma^2 Re(\mathbf{u}_j^\dagger\mathbf{u}_i\mathbf{w}_i^\dagger\mathbf{w}_j)$.

Let $\mathbf{P}(t)$ and $\mathbf{N}(t)$ be the Gram matrices for $\{\mathbf{u}_i(t)\}_{i=1}^m$ and $\{\mathbf{w}_i(t)\}_{i=1}^m$:

$$(\mathbf{P}(t))_{ij} = \mathbf{u}(t)_i^\dagger\mathbf{u}(t)_j,\ (\mathbf{N}(t))_{ij} = \mathbf{w}(t)_i^\dagger\mathbf{w}(t)_j, \tag{111}$$

the matrix $\mathbf{K}_{\mathsf{asym}}$ can be expressed as $\mathbf{K}_{\mathsf{asym}} = 4\gamma^2\mathbf{P} \circ \mathbf{N}^T + 4\gamma^2\mathbf{P}^T \circ \mathbf{N}$, where $\circ$ denotes the Hadamard product, with $\mathbf{P}$ and $\mathbf{N}$ being positive semidefinite matrices. Following a result of Schur's (e.g. see Lemma 6.5 in (Oymak & Soltanolkotabi, 2020)), we estimate the smallest eigenvalue of $\mathbf{K}_{\mathsf{asym}}$ as

$$\lambda_{\min}(\mathbf{K}_{\mathsf{asym}}(\mathbf{M}(\boldsymbol{\theta}))) \geq 8\gamma^2 \max\big(\min_{i\in[m]}(\mathbf{N}_{ii})\lambda_{\min}(\mathbf{P}), \min_{i\in[m]}(\mathbf{P}_{ii})\lambda_{\min}(\mathbf{N})\big). \tag{112}$$

□

The second statement in the limit suggests that the $\mathbf{K}_{\mathsf{asym}}$ is positive definite unless the subspaces spanned by $\mathbf{u}_j$ or $\mathbf{w}_j$ are not full rank, though we do not make use of this fact in the proof of Theorem 4.4.

C.2. Proof of Theorem 4.4

Proof. For each input state $\boldsymbol{\rho}_j = \mathbf{v}_j\mathbf{v}_j^\dagger$, let $\mathbf{u}_j$ and $\mathbf{w}_j$ denote the projection of $\mathbf{v}_j$ onto the positive and negative subspaces of the measurement. Since the measurment is updated throughout the training, $\mathbf{u}_j$ and $\mathbf{w}_j$ are functions of time. For a QNN with the scaling factor γ, the QNN prediction for the input state $\boldsymbol{\rho}_j$ at time t is $\hat{y}_j = \gamma(\mathbf{u}_j^\dagger(t)\mathbf{u}_j(t) - \mathbf{w}_j^\dagger(t)\mathbf{w}_j(t))$. Additionally by the normalization of quantum states and the orthogonality of the training sample, we have $\mathbf{u}_j^\dagger(t)\mathbf{u}_j(t) + \mathbf{w}_j^\dagger(t)\mathbf{w}_j(t) = \delta_{ij}$, where δ_{ij} is the Kronecker delta function. Combining these two conditions, we can solve that $\mathbf{u}_j^\dagger\mathbf{u}_j = \frac{1}{2}(1 \pm 1/\gamma)$ and $\mathbf{v}_j^\dagger\mathbf{v}_j = \frac{1}{2}(1 \mp 1/\gamma)$ for $y_j = \pm 1$.

By Lemma C.2, the diagonal entries $(\mathbf{K}_{\mathsf{asym}})_{jj} = 8\gamma^2 Re(\mathbf{u}_j^\dagger\mathbf{u}_j\mathbf{w}_j^\dagger\mathbf{w}_j) = 8\gamma^2 \cdot \frac{1}{2}(1 \pm 1/\gamma) \cdot \frac{1}{2}(1 \mp 1/\gamma) = 2\gamma^2(1 - 1/\gamma^2)$. Without loss of generality, assume $y_1 = y_2 = \cdots = y_{m/2} = 1$ and $y_{m/2+1} = y_{m/2+2} = \cdots y_m = -1$. Then $\mathbf{u}_j = \sqrt{\frac{1+1/\gamma}{2}}\hat{\mathbf{u}}_j$ for $1 \le j \le m/2$ and $\mathbf{u}_j = \sqrt{\frac{1-1/\gamma}{2}}\hat{\mathbf{u}}_j$ for $m/2+1 \le j \le m$. Here $\hat{\mathbf{u}}_j$ are unit vectors defined as $\mathbf{u}_j/\sqrt{\mathbf{u}_j^\dagger\mathbf{u}_j}$. For the off-diagonal entries, $(\mathbf{K}_{\mathsf{asym}})_{ij} = 8\gamma^2 Re(\mathbf{u}_i^\dagger\mathbf{u}_j\mathbf{w}_j^\dagger\mathbf{w}_i) = 8\gamma^2 Re(\mathbf{u}_i^\dagger\mathbf{u}_j \cdot (-\mathbf{u}_j^\dagger\mathbf{u}_i)) = -8\gamma^2|\mathbf{u}_i^\dagger\mathbf{u}_j|^2$. For the first equality we use the orthogonality among $\{\mathbf{v}_j\}_{j=1}^m$.

Define $m \times m$ Hermitian $\mathbf{G}$ such that $G_{ij} = \hat{\mathbf{u}}_i^\dagger\hat{\mathbf{u}}_j$ and $\mathbf{R}$ such that $R_{ij} = \frac{1}{2}(1+1/\gamma)$ for $1 \le i, j \le m/2$, $R_{ij} = \frac{1}{2}(1-1/\gamma)$ for $m/2+1 \le i, j \le m$, and $R_{ij} = \frac{1}{2}\sqrt{1-1/\gamma^2}$ for $1 \le i \le m/2, m/2+1 \le j \le m$ or $m/2+1 \le i \le m, 1 \le j \le m/2$. The off-diagonal entries can be expressed $-8\gamma^2 R_{ij}G_{ij}G_{ji}$.

Using the notations of $\mathbf{R}$ and $\mathbf{G}$, the matrix $\mathbf{K}_{\mathsf{asym}}$ at the global minima can be expressed as

$$\mathbf{K}_{\mathsf{asym}} = 2\gamma^2(1 - 1/\gamma^2)\mathbf{I} - 8\gamma^2\mathbf{R} \circ (\mathbf{G} - \mathbf{I}) \circ (\mathbf{G}^T - \mathbf{I}), \tag{113}$$

where $\mathbf{I}$ is the $m \times m$ identity matrix.

Eigenvalues of $\mathbf{R}$ Let $\mathbf{e}_1$ and $\mathbf{e}_2$ denote the unit vectors

$$\mathbf{e}_1 = \sqrt{\frac{2}{m}}(1, 1, \cdots, 1, 0, 0, \cdots, 0)^T \tag{114}$$

$$\mathbf{e}_2 = \sqrt{\frac{2}{m}}(0, 0 \cdots, 0, 1, 1, \cdots, 1)^T \tag{115}$$

that are zero in the first (last) $m/2$ entries. The matrix $\mathbf{R}$ can be written as

$$\frac{m}{2}(\frac{1}{2}(1 + 1/\gamma)\mathbf{e}_1\mathbf{e}_1^\dagger + \frac{1}{2}(1 - 1/\gamma)\mathbf{e}_2\mathbf{e}_2^\dagger + \frac{1}{2}\sqrt{1 - 1/\gamma^2}\mathbf{e}_1\mathbf{e}_2^\dagger + \frac{1}{2}\sqrt{1 - 1/\gamma^2}\mathbf{e}_2\mathbf{e}_1^\dagger) \tag{116}$$

and can be shown to have eigenvalues $(\frac{m}{2}, 0, \cdots, 0)$ by straight-forward calculation.

Eigenvalues of $\mathbf{G}$ Over the uniform measure over all the global minima, the direction vectors $\hat{\mathbf{u}}_i$ are sampled independently and uniformly from a $d/2$-dimensional (complex) sphere. By the approximate isometric properties (see e.g. Theorem 5.58 in (Vershynin, 2010)), the gram matrix $\mathbf{G}$ of $\{\hat{\mathbf{u}}_j\}_{j=1}^m$ is approximately an isometry: with probability $\ge 1 - 2\exp(-c_p t^2)$

$$\|\mathbf{G} - \mathbf{I}\|_{\mathsf{op}} \le c_m \frac{\max\{\sqrt{m}, t\}}{\sqrt{d}} \tag{117}$$

for constants c_p and c_m.

Applying Lemma C.1 to $\mathbf{R}$, $\mathbf{G} - \mathbf{I}$ and $\mathbf{G}^T - \mathbf{I}$, we have that with probability $\ge 1 - \delta$, the smallest eigenvalues of $\mathbf{K}_{\mathsf{asym}}$ at global minima is greater than or equal to

$$2\gamma^2(1 - 1/\gamma^2 - C_2 \max\{\frac{m^2}{d}, \frac{m\log(2/\delta)}{d}\}) \tag{118}$$

for some constant $C_2 > 0$. □

D. Experiments

D.1. Experiment Details

Our numerical experiments involve simulating both quantum neural networks and the asymptotic dynamics.

QNN simulation We simulate the QNN experiments using Pytorch (Paszke et al., 2019) with the periodic ansatze defined in Definition 2.1. The generating Hamiltonian $\mathbf{H}$ are chosen to be a d-dimensional diagonal matrix with $d/2$ $\sqrt{d-d^{-1}}$ and $d/2$ $-\sqrt{d-d^{-1}}$ on the diagonal (normalized such that $\mathrm{tr}(\mathbf{H^2})/(d^2-1)=1$). Each instance of the experiments is specified by the number of samples m, system dimension d, number of parameters p and the scaling factor γ. A m-sample dataset is generated by randomly sampled m orthogonal pure states $\{\mathbf{v}_i\}_{i=1}^m \in \mathbb{C}^d$ and randomly assigned half of the samples with label $+1$ and the other half label -1 (i.e. $\{y_i\}_{i=1}^m \subset \{\pm 1\}^m$).

The optimizer we use is the standard gradient descent optimizer. To simulate the dynamics of gradient flow, we choose the learning rate to be $0.001/p$ and the maximum number of epochs is set to be 10000. We run the experiments on Amazon EC2 C5 Instances.

Asymptotic dynamics simulation Theorem 4.2 allows us to examine the behavior of QNN dynamics when $p \to \infty$ by studying the asymptotic dynamics:

$$\frac{d\mathbf{M}(t)}{dt} = -\eta \sum_{j=1}^{m} r_j [\mathbf{M}(t), [\mathbf{M}(t), \boldsymbol{\rho}_j]], \quad \text{where } \forall j \in [m], r_j := \mathrm{tr}(\mathbf{M}(t)\boldsymbol{\rho}_j) - y_j. \tag{119}$$

For a QNN asymptotic dynamics with number of samples m, system dimension d and scaling factor γ, we initialize $\mathbf{M}(0)$ as

$$\gamma \mathbf{U} \begin{bmatrix} +1 & 0 & \cdots & 0 & 0 \\ 0 & +1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 0 \\ 0 & 0 & \cdots & 0 & -1 \end{bmatrix} \mathbf{U}^\dagger \tag{120}$$

with $\mathbf{U}$ being a $d \times d$ haar random unitary. Similar to the QNN simulation, the training set is chosen to be m orthogonal pure states with labels randomly sampled from $\{\pm 1\}$. The simulation of the asymptotic dynamics is run on Intel Core i7-7700HQ Processor (2.80Ghz) with 16G memory.

D.2. $\mathbf{K}_{\mathsf{asym}}$ as a Function of t

In Corollary 4.3, we see that the convergence rates for one-sample QNNs change significantly during training. Theorem 4.2 allows us further verify this observation for training sets with $m > 1$ by simulating the asymptotic dynamics.

In Figure 5, we plot the relative change of the $\mathbf{K}_{asym}(t)$ defined as

$$(\mathbf{K}_{asym}(t))_{ij} := \mathrm{tr}\left(i[\mathbf{M}(t), \boldsymbol{\rho}_i] i[\mathbf{M}(t), \boldsymbol{\rho}_j]\right). \tag{121}$$

Each of the data point is averaged over 100 random initialization of $\mathbf{M}(0)$. It is observed that $\mathbf{K}_{asym}(t)$ changes significantly ($\geq 5\%$) for each of the hyperparameters d, m and γ. Therefore we conclude that the deviation from the neural tangent kernel regression is ubiquitous in general for practical settings. Particularly it rules out the existing belief that the $d \to \infty$ alone can lead to a neural tangent kernel-like behavior in QNNs. Same is observed for over-parameterized QNNs (Figure 6)

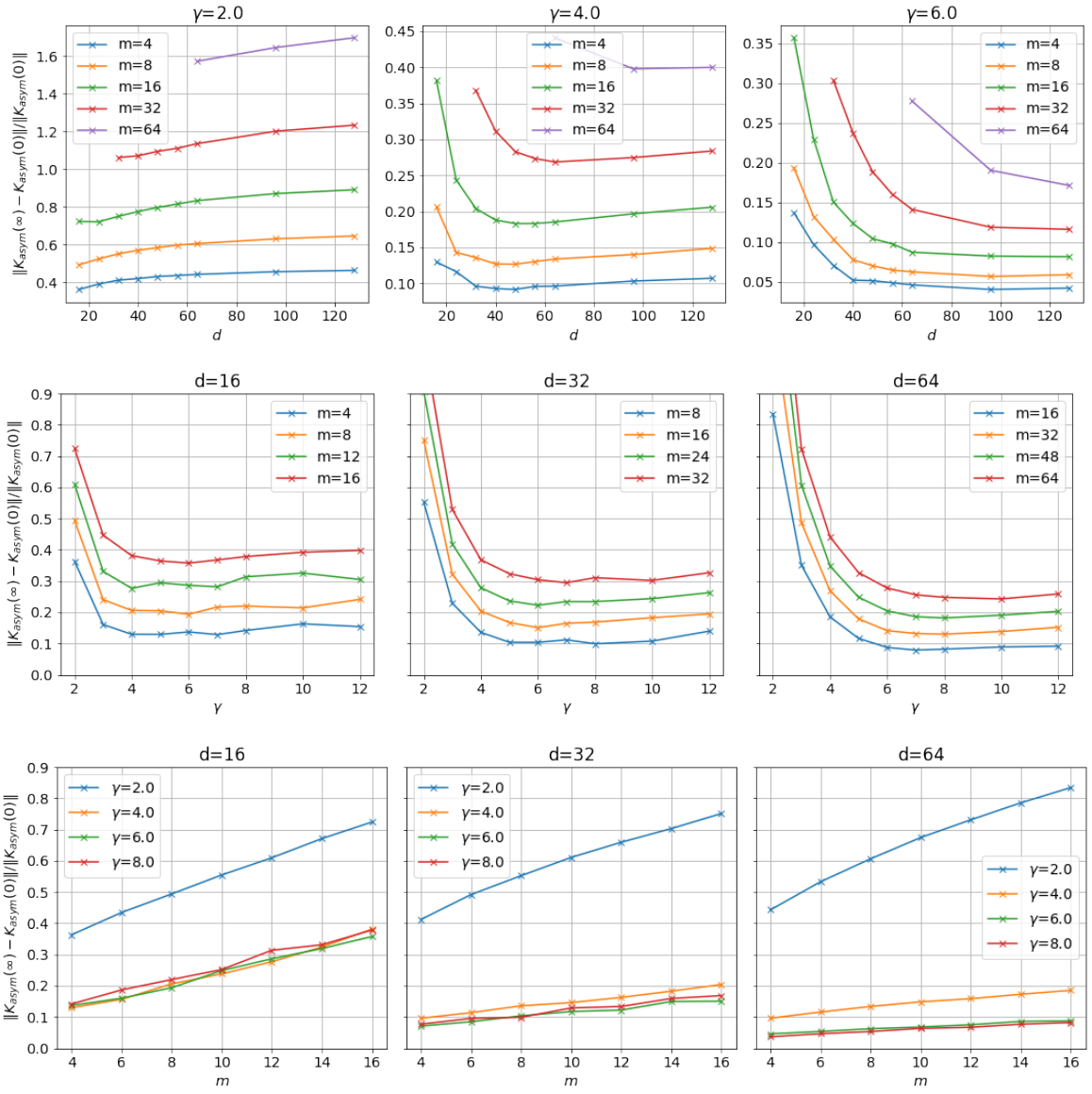

Figure 5. Relative change of $\mathbf{K}_{\text{asym}}(t)$ in the QNN asymptotic dynamics for varying system dimension d, scaling factor γ and number of training samples m. $\mathbf{K}_{asym}(t)$ changes significantly ($\geq 5\%$) throughout training.

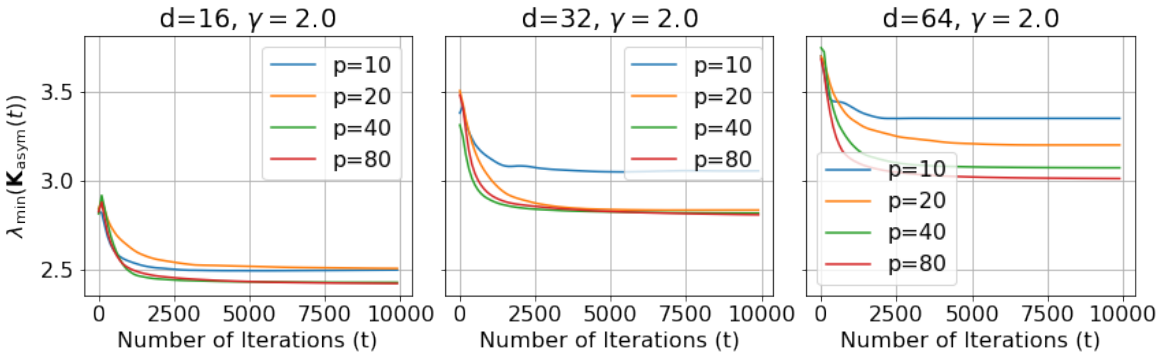

Figure 6. Change of the $\lambda_{\min}(\mathbf{K}_{\text{asym}}(t))$ during the training in QNNs with $m = 4, \gamma = 2.0$ and varying d.