

An Analytic Theory of Quantum Imaginary Time Evolution

Min Chen,¹ Bingzhi Zhang,^{2,3} Quntao Zhuang,^{2,3} and Junyu Liu¹

¹*Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260, USA*

²*Ming Hsieh Department of Electrical and Computer Engineering,
University of Southern California, Los Angeles, CA 90089, USA*

³*Department of Physics and Astronomy, University of Southern California, Los Angeles, CA 90089, USA*

(Dated: October 28, 2025)

Quantum imaginary time evolution (QITE) algorithm is one of the most promising variational quantum algorithms (VQAs), bridging the current era of Noisy Intermediate-Scale Quantum devices and the future of fully fault-tolerant quantum computing. Although practical demonstrations of QITE and its potential advantages over the general VQA trained with vanilla gradient descent (GD) in certain tasks have been reported, a first-principle, theoretical understanding of QITE remains limited. Here, we aim to develop an analytic theory for the dynamics of QITE. First, we show that QITE can be interpreted as a form of a general VQA trained with Quantum Natural Gradient Descent (QNGD), where the inverse quantum Fisher information matrix serves as the learning-rate tensor. This equivalence is established not only at the level of gradient update rules, but also through the action principle: the variational principle can be directly connected to the geometric geodesic distance in the quantum Fisher information metric, up to an integration constant. Second, for wide quantum neural networks, we employ the quantum neural tangent kernel framework to construct an analytic model for QITE. We prove that QITE always converges faster than GD-based VQA, though this advantage is suppressed by the exponential growth of Hilbert space dimension. This helps explain certain experimental results in quantum computational chemistry. Our theory encompasses linear, quadratic, and more general loss functions. We validate the analytic results through numerical simulations. Our findings establish a theoretical foundation for QITE dynamics and provide analytic insights for the first-principle design of variational quantum algorithms.

I. INTRODUCTION

Current quantum devices [1, 2] operate with a limited number of qubits and are subject to significant noise. While challenges such as restricted qubit coherence remain, variational quantum algorithms (VQAs) have emerged as promising approaches [3–9]. In particular, quantum imaginary time evolution (QITE) algorithm [10–12] stands out by enabling efficient ground state convergence and state preparation through emulating non-unitary imaginary time dynamics in quantum chemistry and condensed matter systems. Beyond that, it can also be applied to general learning and optimization problems by following its variational principle [12] such as the “McLachlan variational principle” which determines the optimal evolution by minimizing the instantaneous difference between the exact and variational time derivatives, often leading to faster convergence. These properties have led to improved performance in diverse applications, including state preparation [12, 13] and simulations of quantum many-body systems [14, 15].

While there has been evidence that QITE exhibits experimental advantages [16], the lack of analytic studies hinders a deeper understanding of its training dynamics and constrains the potential of designing new quantum algorithms [17]. Notably, while substantial analytic progress has been made for general VQAs or quantum neural networks (QNNs) trained with vanilla gradient descent (GD), including studies on trainability [18–20], expressivity [21–25], generalization [9, 26–28], and convergence [29–31], a comparable systematic analytic framework for QITE is still lacking. This motivates us to develop a systematic analytic framework for QITE, aiming to establish a first-principle understanding of its training dynamics.

In this work, we begin by developing the analytic theory via interpreting QITE as a general VQA trained with Quantum Natural Gradient Descent (QNGD) [32] (QNGD-based VQAs) where the inverse quantum Fisher information matrix serves as the learning-rate tensor. This interpretation is based on the action principle where we relate the variational principle of QITE with the geometric geodesic distance in the quantum Fisher information metric. Our contribution proceeds along two complementary directions: (i) we show the equivalence between the objective of QITE and that of QNGD-based VQAs, and (ii) we formulate an action principle for QNGD-based VQAs and prove the direct equivalence with the variational principle of QITE, both up to an integration constant under the continuous-time limit. This also explains why these two seemingly different processes yield identical parameter update rules [32]. To pursue generality, we go beyond the scheme with linear loss function by formulating the analysis with general loss functions, along with the extension of the variational principles.

We further develop the analytic theory for the dynamics of QITE from the natural geometric perspective of QNGD, allowing us to directly compare the dynamics with general VQAs trained with GD (GD-based VQAs), focusing on a wide QNN. The results are based on the Quantum Neural Tangent Kernel (QNTK) frameworks [29, 31, 33–36]. We calculate the dynamics with respect to residual training error for sufficiently random quantum circuits modeled as unitary k -designs [37–40]. For sufficiently large number of variational parameters modelled as overparameterized regime [29], analytic solutions characterizing the convergence relation between QITE and GD-based VQAs in terms of the residual training error are established [29]. Notably, prior studies [29–31] have predominantly tackled quadratic loss without show-

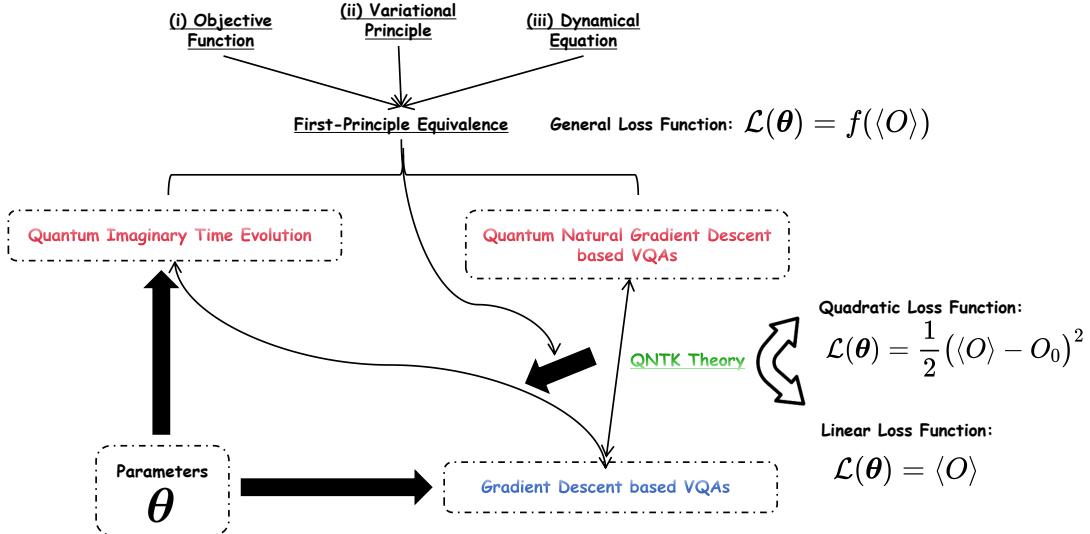


FIG. 1. Overview of Main Results. Firstly, we uncover and establish a first-principle equivalence between QITE and QNGD-based VQAs by deriving that the objective functions, the variational principles and the dynamical equations are identical up to an integration constant with general loss function in continuous-time limit. Secondly, we focus on quadratic and linear loss, and leverage QNTK theory to derive a first-principle model to characterize the training dynamics of QITE in the regimes of interest.

ing any convergence speed-up in linear loss, where we argue that it is limited as QITE is often applied to problems such as ground state energy estimation [41–44] which naturally involve linear objectives. Therefore, we target both the quadratic and linear loss here. Specifically, our analysis with quadratic loss function focuses on lazy training regime [29, 30], where the QNTK K becomes a constant at late time. Meanwhile, for the ground state search problem naturally formulated with a linear loss, we show that the dynamics of both K and the residual training error ϵ decay exponentially with their dynamics driven by the relative dQNTK λ . We demonstrate that the out theories offer a geometric perspective to analytically explain its convergence advantages in the studied regimes, and show that QITE with either type of loss function can present convergence advantages than GD-based VQAs, though these advantages are suppressed by the exponential growth of Hilbert space dimension with number of qubits. These theories help explain certain experimental results in quantum computational chemistry [16]. We verify these results with numerical studies.

In summary, the workflow of this study is outlined in Fig. 1. We first introduce the analytic theories, including the action-principle equivalence and the analytic model, with detailed derivations provided in the Appendix. We then show numerical simulations for validating the convergence alignments, and outline the methods we adopt. Finally, we discuss broader implications and potential future directions related with this work.

II. RESULTS

We consider a universal VQA or QNN that prepares a normalized state $|\psi(\theta)\rangle$ with circuit parameters θ . The learning

objective is a differentiable function of an observable expectation,

$$\mathcal{L}(\theta) = f(\langle O \rangle), \quad \langle O \rangle := \langle \psi(\theta) | O | \psi(\theta) \rangle, \quad (1)$$

where O is a Hermitian observable and $f : \mathbb{R} \rightarrow \mathbb{R}$ is differentiable. This covers common energy-based VQAs by taking $O = H$ and $f(x) = x$ or any monotone f . Below we define (i) *quantum imaginary time evolution (QITE)* and (ii) *the general VQAs trained with QNGD* (termed *QNGD-based VQAs*) such that we can establish the equivalence relation later.

QITE. QITE is widely adopted for approximating the non-unitary imaginary time dynamics.

$$\frac{d|\psi(\tau)\rangle}{d\tau} = -O|\psi(\tau)\rangle, \quad (2)$$

where $\tau = it$ defines the imaginary time, and the $|\psi(\tau)\rangle$ defines the evolved quantum state. In practice, one restricts the dynamics to a parametrized variational manifold $|\psi(\theta)\rangle$, and the goal is to determine $\theta(\tau)$ that best approximates the true imaginary time evolution. A principled way is given by McLachlan’s variational principle, which is of our focus with the reasons illustrated in *Appendix VID*. It states that the exact imaginary time derivative should be projected onto the tangent space of the variational manifold by requiring that the residual norm follows:

$$\delta \left\| \left(\frac{\partial}{\partial \tau} + O - E_\tau \right) |\psi(\theta(\tau))\rangle \right\| = 0. \quad (3)$$

This leads to equations that determines the infinitesimal parameter update $\dot{\theta}$ such that the variational trajectory remains as close as possible to the exact flow. In a discrete time setting, instead of matching derivatives, one considers a finite step $\Delta\tau$ of imaginary time evolution. The evolved state can be written

as $e^{-O\Delta\tau}\psi(\boldsymbol{\theta})$, which in general does not lie within the variational manifold. Therefore, the projected QITE addresses this by choosing the updated variational state $\psi(\boldsymbol{\theta} + \Delta\boldsymbol{\theta})$ that maximizes its fidelity with the evolved state, namely the approximation objective of QITE is defined below:

$$\Delta\boldsymbol{\theta}_{\text{QITE}} = \arg \max_{\Delta\boldsymbol{\theta}} |\langle \psi(\boldsymbol{\theta}) | e^{-\Delta\tau O} | \psi(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) \rangle|^2. \quad (4)$$

In the small step limit $\Delta\tau \rightarrow 0$, this discrete time fidelity maximization is equivalent to McLachlan's residual minimization, so that projected QITE can be interpreted as a finite-step realization of McLachlan's variational principle.

QNGD-based VQAs: QNGD-based VQAs seek an update $\Delta\boldsymbol{\theta}$ such that the loss $\mathcal{L}(\boldsymbol{\theta} + \Delta\boldsymbol{\theta})$ is minimized where $\boldsymbol{\theta}$ denotes the QNN parameters, while constraining the induced change in quantum state fidelity. To facilitate a unified geometric understanding of QITE, we reformulate QNGD-based VQAs in the setting of general differentiable loss functions $f(\langle O \rangle)$, where f is an arbitrary differentiable function. Let $|\psi(\boldsymbol{\theta})\rangle$ denotes the quantum state prepared by a parameterized circuit, then instead of constraining $\|\Delta\boldsymbol{\theta}\|$ in Euclidean space, QNGD equips the geometric manifold with the fidelity distance between $|\psi(\boldsymbol{\theta})\rangle$ and $|\psi(\boldsymbol{\theta} + \Delta\boldsymbol{\theta})\rangle$ defined as [45]:

$$d_f(\psi(\boldsymbol{\theta}), \psi(\boldsymbol{\theta} + \Delta\boldsymbol{\theta})) := 1 - |\langle \psi(\boldsymbol{\theta}) | \psi(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) \rangle|^2. \quad (5)$$

In the limit of $\Delta\boldsymbol{\theta} \rightarrow 0$, the fidelity distance reduces to the squared line element ds^2 on the quantum state manifold [32]:

$$\begin{aligned} d_f &\approx ds^2 = \frac{1}{4} \sum_{\ell_1, \ell_2} \mathcal{F}_{\ell_1 \ell_2}(\boldsymbol{\theta}) \Delta\boldsymbol{\theta}_{\ell_1} \Delta\boldsymbol{\theta}_{\ell_2}, \\ &= \sum_{\ell_1, \ell_2} g_{\ell_1 \ell_2}(\boldsymbol{\theta}) \Delta\boldsymbol{\theta}_{\ell_1} \Delta\boldsymbol{\theta}_{\ell_2}, \end{aligned} \quad (6)$$

where $\mathcal{F}(\boldsymbol{\theta})$ is the quantum Fisher information matrix (QFIM). It is also connected to the Fubini–Study metric tensor g by $\mathcal{F}_{\ell_1 \ell_2}(\boldsymbol{\theta}) = 4g_{\ell_1, \ell_2}(\boldsymbol{\theta}) = 4 \operatorname{Re}[G_{\ell_1 \ell_2}(\boldsymbol{\theta})]$, with G to be the quantum geometric tensor (QGT) [46] defined as follows:

$$\begin{aligned} G_{\ell_1 \ell_2}(\boldsymbol{\theta}) &:= \left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_1}} \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_2}} \right\rangle \\ &- \left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_1}} \middle| \psi(\boldsymbol{\theta}) \right\rangle \left\langle \psi(\boldsymbol{\theta}) \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_2}} \right\rangle. \end{aligned} \quad (7)$$

Applying the Lagrangian formulation, the approximation objective of QNGD-based VQAs becomes:

$$\begin{aligned} \Delta\boldsymbol{\theta}_{\text{QNGD}} &= \\ &\underset{\Delta\boldsymbol{\theta}}{\operatorname{argmin}} \left(\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \cdot \Delta\boldsymbol{\theta} + \frac{1}{4} \lambda \sum_{\ell_1, \ell_2} \mathcal{F}_{\ell_1 \ell_2}(\boldsymbol{\theta}) \Delta\boldsymbol{\theta}_{\ell_1} \Delta\boldsymbol{\theta}_{\ell_2} \right), \end{aligned}$$

where $\lambda > 0$ is the Lagrange multiplier. Full derivations are provided in Appendix VIIA.

A. A First-principle Equivalence Between QITE and QNGD-based VQAs

We first analyze the linear loss function defined as a linear observable expectation $\mathcal{L}(\boldsymbol{\theta}) = \langle O \rangle$.

Theorem II.1 (The Objective Equivalence of QITE and QNGD-based VQAs in the Continuous-Time Limit). *In the infinitesimal-step limit $\eta \rightarrow 0$ with $\mathcal{L}(\boldsymbol{\theta}) = \langle O \rangle$, the objective function of QNGD-based VQAs becomes equivalent to that of QITE.*

Sketch of Proof. Following Stokes *et al.* [32], we expand the fidelity between the imaginary-time evolved state $e^{-\hat{O}\Delta\tau}\psi(\boldsymbol{\theta})$ and the variationally updated state $\psi(\boldsymbol{\theta} + \Delta\boldsymbol{\theta})$ to second order. Letting $\delta := \frac{d\boldsymbol{\theta}}{d\tau}$ and taking the continuous-time limit $\Delta\tau \rightarrow 0$, then QITE's objective reduces to:

$$\frac{\partial \langle \psi(\boldsymbol{\theta}) | O | \psi(\boldsymbol{\theta}) \rangle}{\partial \boldsymbol{\theta}_{\ell_1}} \delta_{\ell_1} + \operatorname{Re}[G_{\ell_1 \ell_2}(\boldsymbol{\theta})] \delta_{\ell_1} \delta_{\ell_2}, \quad (8)$$

which matches the objective of QNGD-based VQAs in Eq. (8) for $\lambda = 1$, thereby proving the equivalence up to an overall scaling factor. See detailed derivations in Appendix VII B.

Furthermore, to establish a first-principle equivalence, we now turn to the continuous-time variational formulation of QNGD-based VQAs and compare it with QITE with the McLachlan principle. We first formalize the variational principle of QNGD-based VQAs with general loss function below:

Theorem II.2 (Variational Principle of QNGD-based VQAs with General Loss Function). *Let $|\psi(\boldsymbol{\theta})\rangle$ be a parameterized quantum state, and let the general loss be a differentiable function of the observable expectation defined as $\mathcal{L}(\boldsymbol{\theta}) = f(\langle O \rangle)$, where O is a Hermitian observable. Then the variational principle underlying QNGD-based VQAs is given by*

$$\delta \left[f'(\langle O \rangle_{\boldsymbol{\theta}}) \nabla \langle O \rangle_{\boldsymbol{\theta}}^{\top} \Delta\boldsymbol{\theta} + \frac{1}{2\eta} \Delta\boldsymbol{\theta}^{\top} \mathcal{F} \Delta\boldsymbol{\theta} \right] = 0, \quad (9)$$

where $\nabla \langle O \rangle_{\boldsymbol{\theta}}$ is the gradient of the expectation value $\langle O \rangle$ with respect to the parameters $\boldsymbol{\theta}$, $\eta > 0$ is the learning rate, and \mathcal{F} is the QFIM, and δ indicates the first variation with respect to $\Delta\boldsymbol{\theta}$. This variational principle defines the variational functional for QNGD-based VQAs with general loss function:

$$\mathcal{J}_{\text{General}}[\Delta\boldsymbol{\theta}] = f'(\langle O \rangle_{\boldsymbol{\theta}}) \nabla \langle O \rangle_{\boldsymbol{\theta}}^{\top} \Delta\boldsymbol{\theta} + \frac{1}{2\eta} \Delta\boldsymbol{\theta}^{\top} \mathcal{F} \Delta\boldsymbol{\theta}. \quad (10)$$

Detailed formalism can be found in Appendix VII C. To establish the equivalence step by step, we first give our finding with linear loss function:

Theorem II.3 (Variational Principle Equivalence between QNGD-based VQAs and QITE with Linear Loss Function). *Let $|\psi(\boldsymbol{\theta})\rangle$ be a normalized parameterized quantum state, and let the linear loss function be an observable expectation*

$\mathcal{L}(\boldsymbol{\theta}) = \langle O \rangle$. In the continuous-time limit $\eta \rightarrow 0$, the variational functional of QNGD-based VQAs $\mathcal{J}_{\text{Linear}}$ and QITE $\mathcal{D}_{\text{Linear}}$ become

$$\begin{aligned} \mathcal{J}_{\text{Linear}}[\boldsymbol{\theta}(\tau)] &= \\ \int \left(\sum_{\ell_1} \text{Re}(\langle \partial_{\ell_1} \psi | O | \psi \rangle) \dot{\boldsymbol{\theta}}_{\ell_1} + \frac{1}{2} \sum_{\ell_1, \ell_2} F_{\ell_1 \ell_2} \dot{\boldsymbol{\theta}}_{\ell_1} \dot{\boldsymbol{\theta}}_{\ell_2} \right) d\tau, \end{aligned} \quad (11)$$

$$\mathcal{D}_{\text{Linear}}[\boldsymbol{\theta}(\tau)] = \|(\partial_{\tau} + O - E_{\tau}) |\psi(\boldsymbol{\theta}(\tau))\rangle\|^2, E_{\tau} = \langle O \rangle_{\boldsymbol{\theta}(\tau)}. \quad (12)$$

We interchangeably use $\partial_{\tau}\psi$ to denote $\partial_{\tau}|\psi(\boldsymbol{\theta}(\tau))\rangle = \sum_{\ell} \frac{\partial|\psi(\boldsymbol{\theta})\rangle}{\partial\theta_{\ell}} \frac{d\theta_{\ell}}{d\tau}$. Then, up to an additive constant and an overall scaling factor that do not affect the variational dynamics,

$$\mathcal{D}_{\text{Linear}} \propto \mathcal{J}_{\text{Linear}} + \text{constant}. \quad (13)$$

Therefore, QNGD-based VQAs and QITE induce the same dynamics in the linear loss setting.

Sketch of proof. Utilizing the derivative chain rule $\partial_{\tau}|\psi\rangle = \sum_{\ell_1} \partial_{\theta_{\ell_1}}|\psi\rangle \dot{\boldsymbol{\theta}}_{\ell_1}$, the first term in $\mathcal{J}_{\text{Linear}}$ can be reduced to $\text{Re}(\langle \partial_{\tau}\psi | O | \psi \rangle)$. While for the second term we use the normalization condition $\partial_{\tau}\langle\psi|\psi\rangle = 0$ and have $\sum_{\ell_1, \ell_2} F_{\ell_1 \ell_2} \dot{\boldsymbol{\theta}}_{\ell_1} \dot{\boldsymbol{\theta}}_{\ell_2} = \langle \partial_{\tau}\psi | \partial_{\tau}\psi \rangle$. Therefore we can rewrite $\mathcal{J}_{\text{Linear}}$ as $\mathcal{J}_{\text{Linear}} = \int (\text{Re}(\partial_{\tau}\psi | O | \psi) + \frac{1}{2} \langle \partial_{\tau}\psi | \partial_{\tau}\psi \rangle) d\tau$.

On the other hand, expanding the QITE functional $\mathcal{D}_{\text{Linear}} = \|(\partial_{\tau} + O - E_{\tau})|\psi\rangle\|^2$ and discarding constant terms (e.g., $\langle O^2 \rangle - E_{\tau}^2$) yields the same integrand as $\mathcal{J}_{\text{Linear}}$, up to an overall scaling factor. Hence, the two variational functionals differ only by a constant scaling and additive offset, proving that they yield the same variational dynamics. Details can be found in Appendix VII D.

Now we discuss the case with general loss function. We first need to extend the variational principle of QITE as formulated below:

Theorem II.4 (Generalized McLachlan Variational Principle for QITE with general loss function). *Given a general differentiable loss function $\mathcal{L} = f(\langle \psi(\boldsymbol{\theta}) | O | \psi(\boldsymbol{\theta}) \rangle)$, the variational principle for QITE is given by:*

$$\delta \left\| \left(\frac{\partial}{\partial \tau} + f'(E_{\tau})(O - E_{\tau}) \right) |\psi(\boldsymbol{\theta}(\tau))\rangle \right\| = 0, \quad (14)$$

where $E_{\tau} = \langle O \rangle = \langle \psi(\boldsymbol{\theta}(\tau)) | O | \psi(\boldsymbol{\theta}(\tau)) \rangle$ denotes the expectation value of O .

Detailed formalism can be found in Appendix VII E.

Remark. This formulation reduces to the standard McLachlan variational principle with linear loss function $\mathcal{L}(\boldsymbol{\theta}) = \langle O \rangle$.

Accordingly, we establish the equivalence below:

Theorem II.5 (Variational Principle Equivalence of QITE and QNGD-based VQAs with General Loss Function). *Given a general differentiable loss function $\mathcal{L} = f(\langle O \rangle)$, in the continuous-time limit $\eta \rightarrow 0$, the variational functionals of QNGD-based VQAs and QITE become*

$$\mathcal{J}_{\text{general}} = \int \left[f'(\langle O \rangle) \text{Re}(\langle \partial_{\tau}\psi | O | \psi \rangle) + \frac{1}{2} \langle \partial_{\tau}\psi | \partial_{\tau}\psi \rangle \right] d\tau, \quad (15)$$

$$\mathcal{D}_{\text{general}} = \|(\partial_{\tau} + f'(\langle O \rangle)(O - \langle O \rangle)) |\psi\rangle\|^2. \quad (16)$$

Then, up to an additive constant and an overall scaling factor that do not affect the variational dynamics,

$$\mathcal{D}_{\text{Linear}} \propto \mathcal{J}_{\text{Linear}} + \text{constant}. \quad (17)$$

Therefore, QNGD-based VQAs and QITE induce the same dynamics in the general loss setting.

Detailed proofs can be found in Appendix VII F.

B. Dynamics of QITE with Quadratic Loss Function

The first-principle equivalence establishes a bridge between the parameter-space QNGD-based VQAs and the quantum state evolution of QITE with general loss function. As a result, we can analyze the training dynamics of QITE through the lens of QNGD using QNTK theory (See Section III C). In the following, we go beyond the scope of GD-based VQAs typically analyzed in previous works [29, 31, 33–36], and adopt this QNGD-informed perspective to study QITE. We formalize our regimes of interest as assumptions below:

Assumption 1 (Lazy training regime). *The variational parameters remain close to their initialization during training, resulting in small updates to the quantum state [29, 33].*

Assumption 2 (Random ansatz structure). *The parameterized unitary $U(\boldsymbol{\theta})$ is sufficiently random. Specifically, for each layer index ℓ , $U_{\ell-}$ and $U_{\ell+}$, defined in Eq. (18) are independent and match the Haar distribution up to the second moment.*

The variational quantum wavefunction [7, 42, 47–50] is defined as

$$\begin{aligned} |\psi(\boldsymbol{\theta})\rangle &= U(\boldsymbol{\theta}) |\psi_0\rangle = \prod_{\ell=1}^L W_{\ell} e^{i\theta_{\ell} X_{\ell}} |\psi_0\rangle \\ &= \prod_{\ell=1}^L W_{\ell} V_{\ell}(\boldsymbol{\theta}_{\ell}) |\psi_0\rangle = U_{\ell+} U_{\ell-} |\psi_0\rangle, \\ \text{with } U_{\ell-} &= \prod_{k=1}^{\ell-1} W_k V_k(\boldsymbol{\theta}_k), U_{\ell+} = \prod_{k=\ell}^L W_k V_k(\boldsymbol{\theta}_k), \end{aligned} \quad (18)$$

where each unitary gate is decomposed into a fixed gate W_{ℓ} and a parametrized rotation $V_{\ell}(\boldsymbol{\theta}_{\ell}) = e^{i\theta_{\ell} X_{\ell}}$ with X_{ℓ} denoting a Hermitian generator. In this work, we focus on the common case where each X_{ℓ} is a traceless operator, typically a single Pauli operator or a Pauli string of tensor products of

Pauli operators adopted in many VQA ansätze. $|\psi_0\rangle$ denotes the input state. The quadratic loss function is defined as

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2} (\langle O \rangle - O_0)^2 \equiv \frac{1}{2} \epsilon^2, \quad (19)$$

where O is the Hermitian observable and $\langle O \rangle$ is the expectation value defined in Eq.(1). O_0 is the target value and $\epsilon \equiv \langle O \rangle - O_0$ defines the residual training error [29].

When applying QITE, the ℓ_1 -th variational parameter is updated according to the difference equation below:

$$\begin{aligned} \delta\boldsymbol{\theta}_{\ell_1}(t) &\equiv \boldsymbol{\theta}_{\ell_1}(t+1) - \boldsymbol{\theta}_{\ell_1}(t) = -\eta \sum_{\ell_2} g_{\ell_1\ell_2}^+ \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_{\ell_2}}, \\ &= -\eta \epsilon(\boldsymbol{\theta}) \sum_{\ell_2} g_{\ell_1\ell_2}^+ \frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_2}}, \end{aligned} \quad (20)$$

where η is the learning rate. When η is small, by Taylor expansion to the first order in η , the time difference equation for ϵ becomes:

$$\begin{aligned} \epsilon(t+1) - \epsilon(t) \equiv \delta\epsilon &\approx \sum_{\ell} \frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell}} \delta\boldsymbol{\theta}_{\ell} \\ &= -\eta \epsilon(\boldsymbol{\theta}) \sum_{\ell_1, \ell_2} \frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_1}} g_{\ell_1\ell_2}^+ \frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_2}} \end{aligned} \quad (21)$$

We define the QNTK K_{QITE} for QITE:

Definition II.6 (Quantum Neural Tangent Kernel (QNTK) for QITE). The QNTK for QITE is defined as:

$$\begin{aligned} K_{\text{QITE}} &= \sum_{\ell_1, \ell_2} \frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_1}} g_{\ell_1\ell_2}^+ \frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_2}} \\ &= \sum_{\ell_1, \ell_2} g_{\ell_1\ell_2}^+ \frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_1}} \frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_2}} \quad (\text{by the symmetry of } g^+). \end{aligned} \quad (22)$$

It governs the decay rate of the residual optimization error $\epsilon(t)$ by:

$$\epsilon(t+1) - \epsilon(t) \equiv \delta\epsilon \approx -\eta \epsilon(\boldsymbol{\theta}) K_{\text{QITE}}. \quad (23)$$

We first model g below:

Lemma II.7. Under Assumption 2, the average of $g_{\ell_1\ell_2}$, i.e., $\overline{g_{\ell_1\ell_2}}$, turns out to be a constant dependent on the dimension of Hilbert Space (HS) :

$$\overline{g_{\ell_1\ell_2}} = \frac{N}{N+1} \delta_{\ell_1\ell_2}, \quad (24)$$

where $N = 2^n$ represents the dimension of HS, and n denotes the number of qubits. $\delta_{\ell_1\ell_2}$ represents the Kronecker delta:

$$\delta_{\ell_1\ell_2} = \begin{cases} 0 & \text{if } \ell_1 \neq \ell_2, \\ 1 & \text{if } \ell_1 = \ell_2. \end{cases} \quad (25)$$

In the large- N limit, the fluctuations in $g_{\ell_1\ell_2}$ around $\overline{g_{\ell_1\ell_2}}$ where we adopt the variance of $g_{\ell_1\ell_2}$ around its expectation $\overline{g_{\ell_1\ell_2}}$ is given by

$$\Delta g_{\ell_1\ell_2}^2 := \mathbb{E}(g_{\ell_1\ell_2}^2) - \overline{g_{\ell_1\ell_2}}^2 \approx \begin{cases} \frac{2}{N^2}, & \text{if } \ell_1 = \ell_2 = \ell, \\ \frac{1}{2N}, & \text{if } \ell_1 \neq \ell_2. \end{cases} \quad (26)$$

Numerical study for Lemma II.7 is provided in Section II D 1, along with the derivation provided in Appendix XII A and Appendix XII B. On average, in the large- N limit, g approaches an identity-like structure scaled by a dimension-dependent factor, with the fluctuations become negligible. Consequently, we approximate the average pseudoinverse $\overline{g^+(\boldsymbol{\theta})}$ using $(\overline{g})^+$, greatly simplifying the analysis.

We now derive the results regarding K_{QITE} . Reviewing the definition of QNTK for GD-based VQAs:

$$K_{\text{GD}} = \sum_{\ell} \frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell}} \frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell}}, \quad (27)$$

thus we can summarize the findings below:

Proposition II.8. Under Assumption 2, on average K_{QITE} defined in Eq. (22) and K_{GD} defined in Eq. (27) satisfy:

$$\overline{K_{\text{QITE}}} \simeq \frac{N+1}{N} \overline{K_{\text{GD}}}, \quad (28)$$

where N is the dimension of the Hilbert space. When N goes to infinity, theoretically we recover $\lim_{N \rightarrow \infty} \overline{K_{\text{QITE}}} \approx \overline{K_{\text{GD}}}$.

Furthermore, for small η , in the lazy training regime, on average the residual training error ϵ follows:

$$\epsilon(t) \approx \epsilon(0) \exp(-\eta \overline{K_{\text{GD}}}), \quad (29)$$

which implies that ϵ_{QITE} and ϵ_{GD} exhibit the following:

$$\epsilon_{\text{QITE}}(t) \approx \epsilon_{\text{GD}}(t) \cdot \exp\left(-\frac{\eta t}{N} \overline{K_{\text{GD}}}\right), \quad (30)$$

where $\overline{K_{\text{GD}}} = \frac{L \text{Tr}\{O^2\}}{N^2}$ [30] under the Assumption 1 and Assumption 2.

The numerical study for Proposition II.8 is provided in Section II D 1, with the derivation in Appendix VIII.

At the level of K itself, the difference is $\mathcal{O}(1/N)$ and hence negligible in the large- N limit. However, since the training loss evolves exponentially, even a small discrepancy in K can lead to a significant error gap over time. We define the logarithmic residual error gap as

$$\delta_{\log}(t) := \ln \epsilon_{\text{GD}}(t) - \ln \epsilon_{\text{QITE}}(t) = \frac{\eta t}{N} \overline{K_{\text{GD}}} \geq 0, \quad (31)$$

and the relative error gap between $\epsilon_{\text{GD}}(t)$ and $\epsilon_{\text{QITE}}(t)$ as

$$\delta_{\text{rel}}(t) := \frac{\epsilon_{\text{GD}}(t) - \epsilon_{\text{QITE}}(t)}{\epsilon_{\text{GD}}(t)} = 1 - e^{-\delta_{\log}(t)}. \quad (32)$$

(i) Convergence dynamics with respect to time in quadratic loss function. When $t \ll N/(\eta K_{\text{GD}})$, i.e., $\delta_{\log}(t) \ll 1$, we perform a Taylor expansion:

$$e^{-\delta_{\log}(t)} = 1 - \delta_{\log}(t) + \mathcal{O}(\delta_{\log}^2(t)), \quad (33)$$

leading to

$$\delta_{\text{rel}}(t) = \delta_{\log}(t) + \mathcal{O}(\delta_{\log}^2(t)). \quad (34)$$

Using the XXZ scaling [31],

$$\overline{K_{\text{GD}}} \sim \mathcal{O}\left(\frac{Ln}{N}\right), \quad (35)$$

we obtain

$$\delta_{\text{rel}}(t) \approx \delta_{\log}(t) = \Theta\left(\frac{Ln \eta t}{N^2}\right). \quad (36)$$

Thus, in this time scale, the relative error gap scales as $\theta(Ln \eta t / N^2)$, indicating that the two evolutions remain indistinguishable up to a vanishing error.

However, when $t = \Theta(N/(\eta K_{\text{GD}})) = \Theta(N^2/\eta Ln)$, we have $\delta_{\log}(t) = \Theta(1)$, and hence

$$\epsilon_{\text{QITE}}(t) = \epsilon_{\text{GD}}(t) \cdot e^{-\delta_{\log}(t)} \leq e^{-c} \cdot \epsilon_{\text{GD}}(t), \quad (37)$$

where $c = \Theta(1)$ is a system-size independent constant. This gives a non-vanishing relative error gap:

$$\delta_{\text{rel}}(t) = 1 - e^{-\delta_{\log}(t)} \geq 1 - e^{-c} = \Theta(1). \quad (38)$$

(ii) Asymptotic consistency. Assume $t = \mathcal{O}(N^k)$, $\eta = \mathcal{O}(N^m)$, and $L = \mathcal{O}(N^\ell)$, with constants $k, m, \ell \geq 0$. Then

$$\delta_{\log}(t) = \mathcal{O}(N^{k+m+\ell-2} \log N). \quad (39)$$

In the limit $N \rightarrow \infty$, if $k + m + \ell < 2$, then $\delta_{\log}(t) \rightarrow 0$, implying

$$\frac{\epsilon_{\text{QITE}}(t)}{\epsilon_{\text{GD}}(t)} = e^{-\delta_{\log}(t)} \rightarrow 1. \quad (40)$$

Hence, QITE and GD-based VQAs become asymptotically indistinguishable when the total scaling budget is sub-quadratic. The extra $\log N$ factor is subleading and does not affect the threshold.

C. Dynamics of QITE with Linear Loss Function

QITE is often applied in ground state search with a linear loss function defined below:

$$\mathcal{L}(\boldsymbol{\theta}) = \langle O \rangle = \langle \psi_0 | U^\dagger(\boldsymbol{\theta}) O U(\boldsymbol{\theta}) | \psi_0 \rangle. \quad (41)$$

Its convergence with the linear loss function can be described using its residual training error $\epsilon = \langle O \rangle - O_{\min}$. Therefore the difference equation for ℓ_1 -th θ is formulated as below:

$$\begin{aligned} \delta\theta_{\ell_1}(t) &\equiv \theta_{\ell_1}(t+1) - \theta_{\ell_1}(t) = -\eta \sum_{\ell_2} g_{\ell_1 \ell_2}^+ \frac{\partial \mathcal{L}}{\partial \theta_{\ell_2}} \\ &= -\eta \sum_{\ell_2} g_{\ell_1 \ell_2}^+ \frac{\partial \epsilon}{\partial \theta_{\ell_2}} \end{aligned} \quad (42)$$

With the linear loss, both K_{QITE} and ϵ_{QITE} exhibit nonlinear dynamics [31]. Thereby we go beyond the linear order expansion in Eq. (21) and introduce a higher order correction to the Taylor expansion of the time difference equation for ϵ :

$$\begin{aligned} \epsilon(t+1) - \epsilon(t) &\equiv \delta\epsilon(t) \\ &\simeq \sum_{\ell_1} \frac{\partial \epsilon}{\partial \theta_{\ell_1}} \delta\theta_{\ell_1}(t) + \frac{1}{2} \sum_{\ell_1, \ell_2} \frac{\partial^2 \epsilon}{\partial \theta_{\ell_1} \partial \theta_{\ell_2}} \delta\theta_{\ell_1} \delta\theta_{\ell_2} \\ &= -\eta \sum_{\ell_1, \ell_2} g_{\ell_1 \ell_2}^+ \frac{\partial \epsilon}{\partial \theta_{\ell_1}} \frac{\partial \epsilon}{\partial \theta_{\ell_2}} \\ &\quad + \frac{1}{2} \eta^2 \sum_{\ell_1, \ell_2, \ell_3, \ell_4} g_{\ell_1 \ell_3}^+ g_{\ell_2 \ell_4}^+ \frac{\partial^2 \epsilon}{\partial \theta_{\ell_1} \partial \theta_{\ell_2}} \frac{\partial \epsilon}{\partial \theta_{\ell_3}} \frac{\partial \epsilon}{\partial \theta_{\ell_4}}. \end{aligned} \quad (43)$$

Here we define the *quantum meta-kernel* (*dQNTK*) for QITE:

Definition II.9 (Quantum Meta-Kernel for QITE). The quantum meta-kernel μ_{QITE} associated with QITE is defined as:

$$\mu_{\text{QITE}} = \sum_{\ell_1, \ell_2, \ell_3, \ell_4} g_{\ell_1 \ell_3}^+ g_{\ell_2 \ell_4}^+ \frac{\partial^2 \epsilon}{\partial \theta_{\ell_1} \partial \theta_{\ell_2}} \frac{\partial \epsilon}{\partial \theta_{\ell_3}} \frac{\partial \epsilon}{\partial \theta_{\ell_4}}. \quad (44)$$

Then, $\delta\epsilon(t)$ reduces to:

$$\delta\epsilon(t) \simeq -\eta K_{\text{QITE}} + \frac{1}{2} \eta^2 \mu_{\text{QITE}}. \quad (45)$$

We also define the *relative dQNTK* for QITE as follows:

Definition II.10 (Relative quantum meta-kernel (*dQNTK*) for QITE). The relative *dQNTK* for QITE is defined as the ratio between $\mu_{\text{QITE}}(t)$ and $K_{\text{QITE}}(t)$, given by:

$$\lambda_{\text{QITE}}(t) = \frac{\mu_{\text{QITE}}(t)}{K_{\text{QITE}}(t)}. \quad (46)$$

This is similar to the way of defining *relative dQNTK* for GD-based VQAs $\lambda_{\text{GD}}(t)$ [31]:

$$\lambda_{\text{GD}}(t) = \frac{\mu_{\text{GD}}(t)}{K_{\text{GD}}(t)}, \quad (47)$$

where

$$\mu_{\text{GD}} = \sum_{\ell_1, \ell_2} \frac{\partial^2 \epsilon}{\partial \theta_{\ell_1} \partial \theta_{\ell_2}} \frac{\partial \epsilon}{\partial \theta_{\ell_1}} \frac{\partial \epsilon}{\partial \theta_{\ell_2}}. \quad (48)$$

Accordingly, we derive a first-order difference equation for K_{QITE} with linear loss function (See *Appendix XI* for detailed derivations):

$$\delta K_{\text{QITE}} = -2\eta\mu_{\text{QITE}} + \mathcal{O}(\eta^2). \quad (49)$$

Combining the results from Eq. (49) and Eq. (45), we can solve this as [31]:

$$2\lambda_{\text{QITE}}\epsilon_{\text{QITE}}(t) = K_{\text{QITE}}(t) \propto e^{-2\eta\lambda_{\text{QITE}}t}, \quad (50)$$

where we can see that the dynamics of both K_{QITE} and λ_{QITE} are driven by λ_{QITE} . Therefore we now formalize the findings:

Proposition II.11. Under Assumption 2, QITE and GD-based VQAs dynamics with linear loss function exhibit the following analytic structure:

First, on average λ_{QITE} and $\lambda_{\text{GD}}(t)$ satisfies:

$$\overline{\lambda_{\text{QITE}}(t)} \simeq \frac{N+1}{N} \overline{\lambda_{\text{GD}}(t)}, \quad (51)$$

where N is the dimension of HS. Based on this, the relation between K_{QITE} and K_{GD} follows:

$$K_{\text{QITE}}(t) \approx K_{\text{GD}}(t) \cdot \exp\left(-\frac{2\eta t}{N} \overline{\lambda_{\text{GD}}(t)}\right). \quad (52)$$

Consequently, the residual training error of QITE satisfies the following approximate expression:

$$\epsilon_{\text{QITE}}(t) \approx \frac{N}{2(N+1)\overline{\lambda_{\text{GD}}}} K_{\text{GD}}(t) \cdot \exp\left(-\frac{2\eta t}{N} \overline{\lambda_{\text{GD}}(t)}\right) \quad (53)$$

The numerical study for Proposition II.11 is provided in Section IID 2, with the derivation in *Appendix IX*.

A direct substitution of the XXZ results into Eq. (31) gives

$$\begin{aligned} \delta_{\log}(t) &:= \ln \epsilon_{\text{GD}}(t) - \ln \epsilon_{\text{QITE}}(t) \\ &= \ln\left(\frac{1}{2\lambda_{\text{GD}}(t)} K_{\text{GD}}(t)\right) \\ &\quad - \ln\left(\frac{N}{2(N+1)\lambda_{\text{GD}}(t)} K_{\text{GD}}(t) \cdot e^{-\frac{2\eta t}{N} \lambda_{\text{GD}}(t)}\right) \\ &= \ln\left(\frac{N+1}{N}\right) + \frac{2\eta t}{N} \lambda_{\text{GD}}(t). \end{aligned} \quad (54)$$

Analyzing the first term in the large- N limit:

$$\ln\left(\frac{N+1}{N}\right) = \ln\left(1 + \frac{1}{N}\right) = \frac{1}{N} - \frac{1}{2N^2} + \mathcal{O}\left(\frac{1}{N^3}\right),$$

which is of order $\mathcal{O}(1/N)$ and hence negligible as $N \rightarrow \infty$.

From the XXZ scaling results [31], we have

$$\lambda_{\text{GD}}(t) \simeq \overline{\lambda_0} \sim \mathcal{O}(L/N). \quad (55)$$

Substituting into Eq. (54) and neglecting the constant offset, we obtain

$$\delta_{\log}(t) \approx \frac{2\eta Lt}{N^2}. \quad (56)$$

(i) Convergence dynamics with respect to time in linear loss function. When $t \ll N^2/(\eta L)$, the error gap remains small:

$$\delta_{\text{rel}}(t) = \delta_{\log}(t) + \mathcal{O}(\delta_{\log}^2(t)) \approx \frac{2\eta Lt}{N^2}. \quad (57)$$

Therefore, in this regime, GD-based VQAs and QITE are indistinguishable up to a vanishingly small relative error.

However, for $t = \Theta(N^2/\eta L)$, we have $\delta_{\log}(t) = \Theta(1)$, leading to an exponential gap:

$$\begin{aligned} \epsilon_{\text{QITE}}(t) &= \epsilon_{\text{GD}}(t) e^{-\delta_{\log}(t)} \leq e^{-c} \epsilon_{\text{GD}}(t), \\ \delta_{\text{rel}}(t) &\geq 1 - e^{-c} = \Theta(1). \end{aligned} \quad (58)$$

Thus, QITE removes an $\mathcal{O}(1)$ fraction of the error of GD-based VQAs in this regime.

(ii) Asymptotic consistency. Assume $t = \mathcal{O}(N^k)$, $\eta = \mathcal{O}(N^m)$, and $L = \mathcal{O}(N^\ell)$, with constants $k, m, \ell \geq 0$. Then

$$\delta_{\log}(t) = \frac{2\eta Lt}{N^2} = \mathcal{O}(N^{k+m+\ell-2}).$$

Hence in $N \rightarrow \infty$:

$k+m+\ell < 2 \implies \delta_{\log}(t) \rightarrow 0 \implies \epsilon_{\text{QITE}}(t)/\epsilon_{\text{GD}}(t) \rightarrow 1$, i.e., QITE and GD-based VQAs become asymptotically indistinguishable.

D. Numerical Studies

1. Numerical Studies with Quadratic Loss Function

We numerically examine the training dynamics of GD-based VQAs and QITE using the XXZ model with a quadratic loss. As shown in Fig. 2, the QNTK value $K(t)$ exhibits distinct behaviors under the two VQAs. Both K_{QITE} and K_{GD} values remain constant throughout the optimization, and K_{QITE} is larger than K_{GD} . Besides, the relation is analytically predictable by following the formula derived in Theorem II.8. Correspondingly, the training error $\epsilon(t)$ of QITE demonstrates a steeper descent compared to GD-based VQAs. Importantly, the QITE error dynamics also closely matches the analytic prediction given by $\epsilon_{\text{GD}}(t) \cdot \exp(-\frac{\eta t}{N} K_{\text{GD}})$, validating our theoretical approximation in Theorem II.8.

According to our theory, the discrepancy in $\epsilon(t)$ trajectories arises from the properties of the underlying Fubini–study metric tensor $g(t)$, particularly its trace component. To validate this, we conduct a numerical analysis of both the average trace $\text{Tr}(g)$ and the off-diagonal elements over the course of training. As shown in Fig. 2, the trace of QITE remains relatively stable and closely follows the analytic prediction. In contrast, the significant suppression of the off-diagonal terms indicates that the QITE dynamics are primarily driven by local parameter updates (the diagonal terms), with non-local correlations having only a negligible impact.

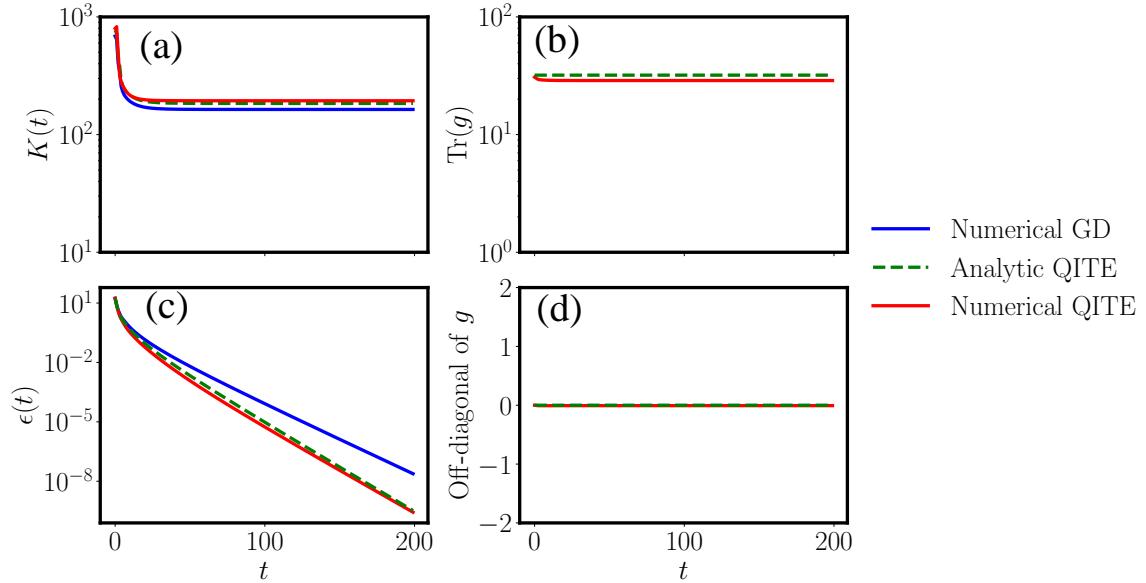


FIG. 2. Training Dynamics of GD-based VQAs and QITE with quadratic loss function. Here, in the example of XXZ model, we respectively investigate the QNTK $K(t)$, the residual error $\epsilon(t)$, the average trace and the off-diagonal terms of the Fubini-study metric tensor g . Each numerical curves are plotted by averaging over 50 times, indicating 50 initializations. We adopt HEA ansatz with 6 layers, and set the number of qubits $n = 3$. The learning rate for optimization is $\eta = 0.001$ with 200 steps. Red curves (denoted as “Numerical QITE”) represent ensemble average results of QITE. Blue curves (denoted as “Numerical GD”) represent the ensemble average numerical results of GD-based VQAs. Green dashed curves represent the analytic prediction of the dynamics of QITE. We also plot the gray lines in the plot of average trace of g , indicating 50 random samples.

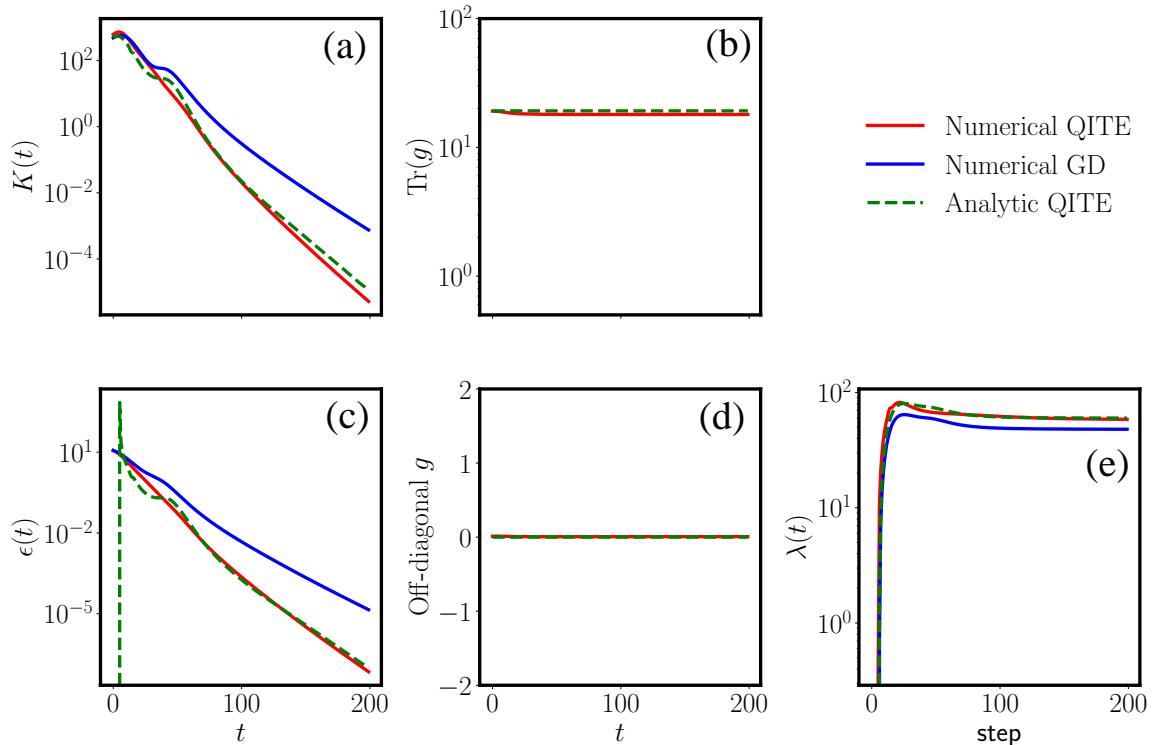


FIG. 3. Training Dynamics of QNNs in GD-based VQAs and QITE with linear loss function. Comparing with Fig 2, we additionally investigate the relative dQNTK value $\lambda(t)$ which drives the dynamics of $K(t)$ and $\epsilon(t)$. Each numerical curves are plotted by averaging over 50 times, indicating 50 initializations. The settings is identical to Fig. 2.

2. Numerical Studies with Linear Loss Function

We further investigate the training dynamics with the linear loss function defined in Eq. (41). As shown in Fig. 3, the relative dQNTK $\lambda(t)$ converges to a stable value at late time. Notably, the converged value of λ_{QITE} is consistently larger than that of λ_{GD} , and their ratio follows the analytic prediction in Theorem II.11. Correspondingly, the QNTK $K(t)$ no longer remains constant as in the quadratic loss case. Both $K_{\text{GD}}(t)$ and $K_{\text{QITE}}(t)$ exhibit exponential decay, with $K_{\text{QITE}}(t)$ decaying at a faster rate. This relative difference is again analytically predictable, as the ratio $K_{\text{QITE}}(t)/K_{\text{GD}}(t)$ follows the exponential form derived from the difference in λ values. As a consequence, the training error $\epsilon(t)$ of QITE and GD-based VQAs exhibits a similar exponential decay behavior to that of $K(t)$. These dynamics are well captured by the analytic expression as described in Theorem II.11.

We also examine the numerical properties of $g(t)$ with linear loss function. Similar to the case with quadratic loss function, the trace $\text{Tr}(g)$ of QITE remains constant over time. Additionally, the off-diagonal components of g are strongly suppressed. These structural trends in $g(t)$ directly govern the observed behavior of $\lambda(t)$, $K(t)$, and $\epsilon(t)$, and are in excellent agreement with our theoretical analysis.

III. METHODS

A. Tasks, Ansatz, and Observable

Tasks. The first task is to establish a first-principle equivalence between QITE and QNGD-based VQAs through the variational principle, encompassing linear, quadratic and more general loss functions up to an integration constant. We then focus on the cases of quadratic and linear loss functions. The tasks lie in two assumptions, formulated as Assumption 1 and Assumption 2. Specifically, the task requires developing a closed-form model that quantitatively connects the training dynamics of classical GD-based VQAs and QNGD-based VQAs, thereby characterizing the behavior of QITE. This enables a systematic comparison of the convergence behavior, curvature scaling, and training error dynamics across these two paradigms.

Ansatz. We employ a hardware-efficient ansatz (HEA) [42] architecture to parameterize the quantum state $|\psi(\theta)\rangle$ used throughout our experiments. The circuit is composed of D alternating layers of local rotations and entangling operations applied to n qubits. Specifically, in each layer, we apply a sequence of single-qubit gates $R_Y(\theta_{i,1}^{(d)})$ and $R_Z(\theta_{i,2}^{(d)})$ to every qubit i , where d indexes the circuit depth. These rotations are followed by entangling operations implemented as a brickwall-patterned array of CNOT gates acting on nearest neighbors. Each pair of R_Y and R_Z rotations contributes two trainable parameters per qubit per layer. Thus, for a circuit with depth D , the total number of trainable parameters is $L = 2nD$. The variational parameters $\theta = (\theta_1, \dots, \theta_L)$ are initialized independently at random from a uniform distribu-

tion. While the HEA ansatz is not exactly Haar-distributed, it has been observed [18, 20] that such circuits with sufficient depth can approximate Haar randomness and form approximate unitary 2-designs, especially when randomized layer permutations or Pauli basis choices are employed. Consequently, we adopt HEA to simulate random circuit behavior in our numerical experiments.

Observable. We consider optimization tasks involving a general Hermitian observable O , acting on an n -qubit system. To simplify the analysis, we often assume O to be traceless, *i.e.*, $\text{Tr}(O) = 0$, as this removes constant energy offsets that do not affect optimization dynamics. A typical traceless observable can be expanded in the Pauli basis as

$$O = \sum_{i=1}^N c_i \hat{P}_i, \quad (59)$$

where $\hat{P}_i \in \{\hat{\sigma}^x, \hat{\sigma}^y, \hat{\sigma}^z\}^{\otimes n} \setminus \{\mathbb{I}^{\otimes n}\}$ are nontrivial Pauli strings and $c_i \in \mathbb{R}$ are real coefficients. For concrete examples and exact expressions, we consider structured XXZ Hamiltonians below:

$$O_{\text{XXZ}} = - \sum_{i=1}^n [\hat{\sigma}_i^x \hat{\sigma}_{i+1}^x + \hat{\sigma}_i^y \hat{\sigma}_{i+1}^y + J (\hat{\sigma}_i^z \hat{\sigma}_{i+1}^z + \hat{\sigma}_i^z)], \quad (60)$$

where J is a tunable interaction strength.

B. Haar Random Ensemble As A Statistical Assumption

For analytic tractability, we follow previous studies [18, 20, 30] and adopt the *Haar random ensemble* to model the typical parameterized unitaries. Under random initialization, the variational circuit is modeled as being drawn from the Haar measure on the unitary group $\mathcal{U}(N)$. This modeling relies on the assumption that the circuit ensemble forms a *unitary k -design* [37–40]. Formally, an ensemble $\mathcal{E} = \{U_i\}$ is said to form a unitary k -design if for any degree- k polynomial $P(U, U^\dagger)$ in the matrix elements of U , we have

$$\mathbb{E}_{U \sim \mathcal{E}}[P(U, U^\dagger)] = \mathbb{E}_{U \sim \text{Haar}}[P(U, U^\dagger)]. \quad (61)$$

This condition ensures that the ensemble \mathcal{E} statistically mimics the Haar measure up to the k -th moment, allowing us to analytically compute quantities like

$$\overline{\frac{\partial \epsilon}{\partial \theta_\ell}} = 0, \quad \overline{g_{\ell_1 \ell_2}} = \frac{N}{N+1} \delta_{\ell_1 \ell_2}, \quad (62)$$

and other key quantities used in our analysis. For example, a unitary 2-design suffices to match the second-order statistical properties of the Haar distribution, such as the expected value of gradients and metric tensors.

C. Variational Principles and First-Principle Equivalence

Variational principles. Generally, a variational principle denotes that the evolution of a system can be characterized

as the stationary point of a *functional*, *i.e.*, a mapping that assigns to each trial function a numerical quantity. Instead of solving the governing equations directly, one introduces a parametrized family of trial functions and determines the parameters by requiring that a chosen functional is optimized with respect to the variational parameters. The choice of functional encodes the physical or mathematical structure of the problem. Typically, it could be an *action functional*, defined as the time integral of the Lagrangian, whose stationarity yields the Euler–Lagrange equations. Similarly, in optimization and machine learning scheme, analogous energy-like or loss functionals define objective landscapes to be minimized. Formally, the variational principle enforces a variational functional $\mathcal{F}[\psi]$ satisfying

$$\delta\mathcal{F}[\psi] = 0, \quad (63)$$

subject to admissible variations of ψ . Though different problems instantiate different functionals, the unifying principle is that the governing equations are recovered by requiring that \mathcal{F} is stationary under variations within the chosen family of trial functions. This formulation emphasizes that the essence of a variational method lies not in dynamical details, but in the specification of the functional whose extremum characterizes the desired variational solution [50–52].

First-principle equivalence. Two algorithms \mathcal{A} and \mathcal{B} can be regarded as *first-principle equivalent* if they induce the same continuous time flow on the variational manifold, up to a relabeling of coordinates or a rescaling of time. In this definition, “*first-principle*” emphasizes that the equivalence is established at the level of the underlying variational functional rather than at the level of discretized updates or empirical performance. In particular, if the Euler–Lagrange equations derived from the variational principles underlying \mathcal{A} and \mathcal{B} coincide, then the induced state trajectories are identical, thereby justifying the notion of first-principle equivalence. More concretely, consider two algorithms \mathcal{A} and \mathcal{B} , each formulated through a variational principle:

$$\delta\mathcal{F}_{\mathcal{A}}[\psi] = 0, \quad \delta\mathcal{F}_{\mathcal{B}}[\psi] = 0, \quad (64)$$

where $\mathcal{F}_{\mathcal{A}}$ and $\mathcal{F}_{\mathcal{B}}$ are the respective variational functionals. If the two functionals coincide, *i.e.*, $\mathcal{F}_{\mathcal{A}}[\psi] \equiv \mathcal{F}_{\mathcal{B}}[\psi]$ up to an integration constant, then the corresponding Euler–Lagrange equations are identical, leading to the same continuous-time dynamics on the variational manifold. Thus, although \mathcal{A} and \mathcal{B} may arise from different algorithmic constructions, *e.g.*, GD-based VQAs versus QITE, their underlying dynamics are identical and independent of implementation details. Therefore, it is adequate to analyze QITE from the perspective of

QNGD in the continuous-time limit because the variational flow of QNGD-based VQAs faithfully reproduces that of QITE, allowing us to apply the QNTK framework and thereby providing a principled and tractable approach to its analysis.

IV. DISCUSSION

This work goes beyond existing analytic theories of quantum learning tasks driven by GD in general QNNs. By building a first-principle equivalence, this work explains why QITE and QNGD-based VQAs are equivalent up to a constant scaling factor within continuous time limit, enabling QITE to be interpreted through QNGD’s natural geometric framework. Building on this connection, we derive the corresponding training dynamics and compare them with those of GD-based VQAs. Our analysis reveals regimes where QITE offers a convergence advantage, and further suggests that well-developed analytical frameworks, such as the QNTK theory, can be applied to advanced VQAs. This could potentially contribute to the identification and design of quantum algorithms that outperform their classical counterparts.

The theory developed in this work has several potential applications. One direction is to analyze QITEs with specific structures, *e.g.*, symmetry [35, 53], that may enhance the performance of QITE. Such potential can be harnessed to guide the design of new quantum algorithms. Another direction is to extend the analysis beyond the lazy training regime or the specific problem considered here, exploring how the dynamics may change and whether the convergence advantage persists. In addition, the notations and formulations developed in this work could be adopted for other analytical studies, such as investigations of expressivity.

ACKNOWLEDGMENT

MC and JL is supported in part by the University of Pittsburgh, School of Computing and Information, Department of Computer Science, Pitt Cyber, Pitt Momentum fund, PQI Community Collaboration Awards, John C. Mascaro Faculty Scholar in Sustainability, NASA under award number 80NSSC25M7057, and Fluor Marine Propulsion LLC (U.S. Naval Nuclear Laboratory) under award number 140449-R08. This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725. QZ and BZ acknowledge support from NSF (OMA-2326746,2350153, CCF-2240641), ONR (N00014-23-1-2296), DARPA (HR00112490453, HR0011-24-9-0362, D24AC00153-02) and AFOSR MURI FA9550-24-1-0349.

-
- [1] J. Preskill, arXiv preprint arXiv:1801.00862 (2018).
[2] J. Preskill, “Beyond nisq: The megaquop machine,” (2025).
[3] Y. Liu, Z. Chen, C. Shu, S. C. Chew, B. C. Khoo, X. Zhao, and Y. Cui, Physics of Fluids **34** (2022).
[4] M. Lubasch, J. Joo, P. Moinier, M. Kiffner, and D. Jaksch, Physical Review A **101**, 010301 (2020).
[5] D. Amaro, M. Rosenkranz, N. Fitzpatrick, K. Hirano, and M. Fiorentini, EPJ Quantum Technology **9**, 5 (2022).

- [6] H. Singh, S. Majumder, and S. Mishra, *The Journal of Chemical Physics* **159** (2023).
- [7] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'brien, *Nature communications* **5**, 4213 (2014).
- [8] P. Wittek, *Quantum machine learning: what quantum computing means to data mining* (Academic Press, 2014).
- [9] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner, *Nature Computational Science* **1**, 403 (2021).
- [10] M. Motta, C. Sun, A. T. Tan, M. J. O'Rourke, E. Ye, A. J. Minnich, F. G. Brandao, and G. K.-L. Chan, *Nature Physics* **16**, 205 (2020).
- [11] X. Yuan, S. Endo, Q. Zhao, Y. Li, and S. Benjamin, “Theory of variational quantum simulation,” (2018), arXiv:1812.08767v4, 1812.08767.
- [12] S. McArdle, T. Jones, S. Endo, Y. Li, S. C. Benjamin, and X. Yuan, *npj Quantum Information* **5**, 75 (2019).
- [13] T. Tsuchimochi, Y. Ryo, S. L. Ten-No, and K. Sasasako, *Journal of Chemical Theory and Computation* **19**, 503 (2023).
- [14] S.-N. Sun, M. Motta, R. N. Tazhigulov, A. T. Tan, G. K.-L. Chan, and A. J. Minnich, *PRX Quantum* **2**, 010317 (2021).
- [15] M. J. Beach, R. G. Melko, T. Grover, and T. H. Hsieh, *Physical Review B* **100**, 094434 (2019).
- [16] S. Guo, J. Sun, H. Qian, M. Gong, Y. Zhang, F. Chen, Y. Ye, Y. Wu, S. Cao, K. Liu, *et al.*, *Nature Physics* **20**, 1240 (2024).
- [17] M. Schuld and N. Killoran, *Prx Quantum* **3**, 030101 (2022).
- [18] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, *Nature communications* **9**, 4812 (2018).
- [19] A. Pesah, M. Cerezo, S. Wang, T. Volkoff, A. T. Sornborger, and P. J. Coles, *Physical Review X* **11**, 041011 (2021).
- [20] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, *Nature communications* **12**, 1791 (2021).
- [21] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, *PRX quantum* **3**, 010313 (2022).
- [22] S. Sim, P. D. Johnson, and A. Aspuru-Guzik, *Advanced Quantum Technologies* **2**, 1900070 (2019).
- [23] K. Nakaji and N. Yamamoto, *Quantum* **5**, 434 (2021).
- [24] Y. Du, Z. Tu, X. Yuan, and D. Tao, *Physical Review Letters* **128**, 080506 (2022).
- [25] K. Sharma, M. Cerezo, Z. Holmes, L. Cincio, A. Sornborger, and P. J. Coles, *Physical Review Letters* **128**, 070501 (2022).
- [26] M. C. Caro, H.-Y. Huang, M. Cerezo, K. Sharma, A. Sornborger, L. Cincio, and P. J. Coles, *Nature communications* **13**, 4919 (2022).
- [27] H.-Y. Huang, R. Kueng, and J. Preskill, *Physical Review Letters* **126**, 190505 (2021).
- [28] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, *Nature communications* **12**, 2631 (2021).
- [29] J. Liu, F. Tacchino, J. R. Glick, L. Jiang, and A. Mezzacapo, *PRX Quantum* **3**, 030323 (2022).
- [30] J. Liu, K. Najafi, K. Sharma, F. Tacchino, L. Jiang, and A. Mezzacapo, *Physical Review Letters* **130**, 150601 (2023).
- [31] B. Zhang, J. Liu, X.-C. Wu, L. Jiang, and Q. Zhuang, *Nature Communications* **15**, 9354 (2024).
- [32] J. Stokes, J. Izaac, N. Killoran, and G. Carleo, *Quantum* **4**, 269 (2020).
- [33] J. Liu, K. Najafi, K. Sharma, F. Tacchino, L. Jiang, and A. Mezzacapo, *Phys. Rev. Lett.* **130**, 150601 (2023).
- [34] J. Liu, Z. Lin, and L. Jiang, *Machine Learning: Science and Technology* **5**, 015058 (2024).
- [35] X. Wang, J. Liu, T. Liu, Y. Luo, Y. Du, and D. Tao, *arXiv preprint arXiv:2208.14057* (2022).
- [36] L.-W. Yu, W. Li, Q. Ye, Z. Lu, Z. Han, and D.-L. Deng, *Reports on Progress in Physics* **87**, 110501 (2024).
- [37] D. A. Roberts and B. Yoshida, *Journal of High Energy Physics* **2017**, 1 (2017).
- [38] J. Cotler, N. Hunter-Jones, J. Liu, and B. Yoshida, *Journal of High Energy Physics* **2017**, 1 (2017).
- [39] J. Liu, *Physical Review D* **98**, 086026 (2018).
- [40] J. Liu, *Physical Review Research* **2**, 043164 (2020).
- [41] A. Aspuru-Guzik, A. D. Dutoi, P. J. Love, and M. Head-Gordon, *Science* **309**, 1704 (2005).
- [42] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, *nature* **549**, 242 (2017).
- [43] A. Kandala, K. Temme, A. D. Córcoles, A. Mezzacapo, J. M. Chow, and J. M. Gambetta, *Nature* **567**, 491 (2019).
- [44] L. Lehtovaara, J. Toivanen, and J. Eloranta, *Journal of Computational Physics* **221**, 148 (2007).
- [45] J. J. Meyer, *Quantum* **5**, 539 (2021).
- [46] M. Kolodrubetz, D. Sels, P. Mehta, and A. Polkovnikov, *Physics Reports* **697**, 1 (2017).
- [47] E. Farhi, J. Goldstone, and S. Gutmann, *arXiv preprint arXiv:1411.4028* (2014).
- [48] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, *New Journal of Physics* **18**, 023023 (2016).
- [49] S. McArdle, S. Endo, A. Aspuru-Guzik, S. C. Benjamin, and X. Yuan, *Reviews of Modern Physics* **92**, 015003 (2020).
- [50] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, *et al.*, *Nature Reviews Physics* **3**, 625 (2021).
- [51] T. Kibble and F. H. Berkshire, *Classical mechanics* (world scientific publishing company, 2004).
- [52] V. I. Arnol'd, *Mathematical methods of classical mechanics*, Vol. 60 (Springer Science & Business Media, 2013).
- [53] M. Larocca, P. Czarnik, K. Sharma, G. Muraleedharan, P. J. Coles, and M. Cerezo, *Quantum* **6**, 824 (2022).
- [54] J. Haegeman, J. I. Cirac, T. J. Osborne, I. Pižorn, H. Verschelde, and F. Verstraete, *Physical review letters* **107**, 070601 (2011).
- [55] Y. Li and S. C. Benjamin, *Physical Review X* **7**, 021050 (2017).
- [56] X. Yuan, S. Endo, Q. Zhao, Y. Li, and S. C. Benjamin, *Quantum* **3**, 191 (2019).
- [57] F. Verstraete, V. Murg, and J. I. Cirac, *Advances in physics* **57**, 143 (2008).
- [58] Y. Ashida, T. Shi, M. C. Banuls, J. I. Cirac, and E. Demler, *Physical Review B* **98**, 024103 (2018).
- [59] A. D. McLachlan, *Molecular Physics* **8**, 39 (1964).
- [60] M. Bukov, D. Sels, and A. Polkovnikov, *Physical Review X* **9**, 011034 (2019).
- [61] A. Shapere and F. Wilczek, *Geometric phases in physics*, Vol. 5 (World scientific, 1989).
- [62] M. Fukuda, R. König, and I. Nechita, *Journal of Physics A: Mathematical and Theoretical* **52**, 425303 (2019).
- [63] B. Collins and P. Śniady, *Communications in Mathematical Physics* **264**, 773 (2006).

CONTENTS

V. Summary of Notations	12
VI. Overview of QITE	12
A. Real Time Evolution	12
B. Imaginary Time Evolution	13
C. Differentiating the Normalized State	14
D. Variational Formulation of QITE	14
VII. Reformulate QNGD-based VQAs and Build A First-principle Equivalence with QITE	15
A. Objective Function of QNGD-based VQAs with General Loss Function	15
B. Equivalence between QITE and QNGD-based VQAs In the Objective Function	16
C. Variational Principle Formulation of QNGD-based VQAs with General Loss Function	16
D. Equivalence between QITE and QNGD-based VQAs In the Variational Principle with Linear Loss Function	17
E. Extend the Variational Principle of QITE	19
F. Equivalence between QITE and QNGD-based VQAs In the Variational Principle with General Loss Function	20
VIII. Proof for Proposition II.8	20
IX. Proof for Proposition II.11	22
X. Additional Numerical Studies	24
A. Numerical Validation of Eq. (131)	24
B. Numerical Studies with Scaling Qubits and Layers	25
XI. Time Difference Equation for $K_{\text{QITE}}(t)$	25
XII. Results with Haar Random Ensemble	29
A. Average Fubini-Study Metric Tensor g Result under Haar Random Ensemble	29
B. Fluctuations of Fubini-Study Metric Tensor g Under Haar Random Ensemble	31

V. SUMMARY OF NOTATIONS

Notations are elaborated in Table I.

For clarity, we consider the observable O to be the system Hamiltonian H throughout this work. Besides, we interchangeably use $\frac{\partial}{\partial \tau}$ and ∂_τ . We also interchangeably use $\langle O \rangle$ and E_τ .

VI. OVERVIEW OF QITE

In this section, we briefly review quantum simulation tasks [54–56], with a particular focus on both real-time and imaginary-time evolution. We then describe the variational principle underlying QITE, restricted to the pure state scenario [12]. We denote the Hermitian observable generating the dynamics as O , which replaces the standard Hamiltonian symbol H . This notation aligns with our later discussion, where the observable also defines the loss function.

A. Real Time Evolution

In real time, the evolution of a quantum state is governed by the Schrödinger equation:

$$\frac{d|\psi(t)\rangle}{dt} = -iO|\psi(t)\rangle, \quad (65)$$

where the reduced Planck constant \hbar is absorbed into the definition of O .

The corresponding unitary evolution operator is:

$$U(t) = e^{-iOt}. \quad (66)$$

TABLE I. Notations

Symbol	Description
D	Depth of QNN, <i>i.e.</i> , number of layers in the parameterized quantum circuit (PQC).
n	Number of qubits.
L	Number of variational parameters; for hardware-efficient ansatz, $L = 2nD$.
N	Hilbert space dimension, $N = 2^n$.
$\boldsymbol{\theta}$	Vector of variational parameters, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_L)$.
$ \psi(\boldsymbol{\theta})\rangle$	Normalized parameterized quantum state prepared by PQC.
$ \psi_0\rangle$	Input (initial) quantum state.
$U(\boldsymbol{\theta})$	Parameterized unitary of the ansatz circuit.
U_ℓ^-, U_ℓ^+	Partial products of the ansatz unitary before and after parameter θ_ℓ , as in Eq. (16).
$V_\ell(\boldsymbol{\theta}_\ell)$	Parameterized single-qubit rotation $e^{i\theta_\ell X_\ell}$ with Hermitian generator X_ℓ .
X_ℓ	Hermitian generator of parameter θ_ℓ , often a single Pauli or tensor product of Paulis.
O	Hermitian observable; in this work taken as the system Hamiltonian H .
H	System Hamiltonian; for the XXZ model, see O_{XXZ} .
O_0	Target value of observable O .
O_{\min}	Minimum eigenvalue of O (ground state energy).
E_τ	Expectation value of O at imaginary time τ , $E_\tau = \langle \psi(\boldsymbol{\theta}(\tau)) O \psi(\boldsymbol{\theta}(\tau)) \rangle$.
$\langle O \rangle$	Shorthand for E_τ .
$\mathcal{L}(\boldsymbol{\theta})$	Loss function.
ϵ	Residual training error.
η	Learning rate.
τ	Imaginary time variable; related to gradient descent steps via $\eta \rightarrow 0$ continuous-time limit.
∂_τ	Imaginary-time derivative; interchangeably written as $\frac{\partial}{\partial \tau}$.
$\delta_{\ell_1 \ell_2}$	Kronecker delta.
$g_{\ell_1 \ell_2}(\boldsymbol{\theta})$	Fubini–Study metric tensor (real part of the quantum geometric tensor).
$g_{\ell_1 \ell_2}^+(\boldsymbol{\theta})$	Pseudoinverse of $g_{\ell_1 \ell_2}(\boldsymbol{\theta})$.
$F_{\ell_1 \ell_2}(\boldsymbol{\theta})$	Quantum Fisher information matrix (QFIM), $F = 4g$.
K_{GD}	Quantum neural tangent kernel (QNTK) for gradient descent: $K_{\text{GD}} = \sum_\ell (\partial_\ell \epsilon)^2$.
K_{QITE}	QNTK for QITE-based optimization: $K_{\text{QITE}} = \sum_{\ell_1, \ell_2} g_{\ell_1 \ell_2}^+ \partial_{\ell_1} \epsilon \partial_{\ell_2} \epsilon$.
μ_{GD}	Quantum meta-kernel (dQNTK) for GD: $\mu_{\text{GD}} = \sum_{\ell_1, \ell_2} \partial_{\ell_1 \ell_2}^2 \epsilon \partial_{\ell_1} \epsilon \partial_{\ell_2} \epsilon$.
μ_{QITE}	Quantum meta-kernel for QITE: $\mu_{\text{QITE}} = \sum_{\ell_1, \ell_2, \ell_3, \ell_4} g_{\ell_1 \ell_3}^+ g_{\ell_2 \ell_4}^+ \partial_{\ell_1 \ell_2}^2 \epsilon \partial_{\ell_3} \epsilon \partial_{\ell_4} \epsilon$.
λ_{GD}	Relative dQNTK for GD: $\lambda_{\text{GD}} = \mu_{\text{GD}} / K_{\text{GD}}$.
λ_{QITE}	Relative dQNTK for QITE: $\lambda_{\text{QITE}} = \mu_{\text{QITE}} / K_{\text{QITE}}$.
$\delta_{\log}(t)$	Logarithmic residual error gap: $\delta_{\log}(t) = \ln \epsilon_{\text{GD}}(t) - \ln \epsilon_{\text{QITE}}(t)$.
$\delta_{\text{rel}}(t)$	Relative error gap: $\delta_{\text{rel}}(t) = 1 - e^{-\delta_{\log}(t)}$.
O_{XXZ}	XXZ Hamiltonian: $O_{XXZ} = -\sum_{i=1}^n [\sigma_i^x \sigma_{i+1}^x + \sigma_i^y \sigma_{i+1}^y + J (\sigma_i^z \sigma_{i+1}^z + \sigma_i^z)]$.
J	Coupling parameter in the XXZ model; controls anisotropy between z -axis and x/y -axis interactions.
$\sigma^x, \sigma^y, \sigma^z$	Pauli operators.
\hat{P}_i	Generic Pauli string in the expansion $O = \sum_i c_i \hat{P}_i$.
c_i	Real coefficient of Pauli string \hat{P}_i in observable O .

The normalized state at time t is then given by:

$$|\psi(t)\rangle = A(t)U(t)|\psi(0)\rangle, \quad (67)$$

where the normalization factor is

$$A(t) = \frac{1}{\sqrt{\langle \psi(0) | e^{-2iOt} | \psi(0) \rangle}}. \quad (68)$$

B. Imaginary Time Evolution

Imaginary time evolution replaces t with $\tau = it$, resulting in a non-unitary flow. The imaginary-time Schrödinger equation becomes:

$$\frac{d|\psi(\tau)\rangle}{d\tau} = -O|\psi(\tau)\rangle. \quad (69)$$

Its solution up to normalization is:

$$|\psi(\tau)\rangle = A(\tau)e^{-O\tau}|\psi(0)\rangle, \quad (70)$$

where the normalization factor is

$$A(\tau) = \frac{1}{\sqrt{\langle\psi(0)|e^{-2O\tau}|\psi(0)\rangle}}. \quad (71)$$

C. Differentiating the Normalized State

Differentiating the normalized QITE state gives:

$$\frac{d}{d\tau}|\psi(\tau)\rangle = \frac{dA}{d\tau}e^{-O\tau}|\psi(0)\rangle + A(\tau)\frac{d}{d\tau}e^{-O\tau}|\psi(0)\rangle. \quad (72)$$

Using the normalization condition $\langle\psi(\tau)|\psi(\tau)\rangle = 1$, and the fact that $A(\tau) \in \mathbb{R}$, $O^\dagger = O$, and $[e^{-O\tau}, O] = 0$, we compute the derivative of $A(\tau)$ as follows:

$$\begin{aligned} \frac{dA}{d\tau} &= \frac{d}{d\tau} (\langle\psi(0)|e^{-2O\tau}|\psi(0)\rangle)^{-1/2} \\ &= -\frac{1}{2}A^3(\tau)\langle\psi(0)|(-2O)e^{-2O\tau}|\psi(0)\rangle \\ &= A^3(\tau)\langle\psi(0)|Oe^{-2O\tau}|\psi(0)\rangle \\ &= A(\tau)\langle\psi(\tau)|O|\psi(\tau)\rangle \\ &= A(\tau)E_\tau, \end{aligned} \quad (73)$$

where $E_\tau := \langle\psi(\tau)|O|\psi(\tau)\rangle$ denotes the instantaneous energy.

Substituting into Eq. (72), we obtain the Wick-rotated Schrödinger equation:

$$\begin{aligned} \frac{d}{d\tau}|\psi(\tau)\rangle &= A(\tau)E_\tau e^{-O\tau}|\psi(0)\rangle - A(\tau)Oe^{-O\tau}|\psi(0)\rangle \\ &= (E_\tau - O)|\psi(\tau)\rangle. \end{aligned} \quad (74)$$

D. Variational Formulation of QITE

On a variational manifold [56–58], we consider a normalized parameterized trial state $|\psi(\boldsymbol{\theta}(\tau))\rangle$ with real parameters $\boldsymbol{\theta} \in \mathbb{R}^L$. The QITE evolution equation is then approximated as:

$$\sum_j \frac{\partial|\psi(\boldsymbol{\theta}(\tau))\rangle}{\partial\theta_j} \dot{\boldsymbol{\theta}}_j \approx (E_\tau - O)|\psi(\boldsymbol{\theta}(\tau))\rangle, \quad (75)$$

where $\dot{\boldsymbol{\theta}}_j := \frac{d\theta_j}{d\tau}$ denotes the imaginary-time derivative of the parameters.

This variational formulation enables efficient simulation of imaginary-time dynamics within a tractable subspace. As discussed in [56], there exist three variational principles (Dirac–Frenkel, McLachlan, and time-dependent variational principle) that are equivalent when parameters are complex. However, since we restrict to real parameters, only McLachlan’s variational principle is applicable.

McLachlan’s Variational Principle [12, 56, 59]. This principle offers a natural way to project non-unitary quantum dynamics, such as imaginary time evolution, onto a variational ansatz. Instead of requiring the trial state to follow the exact equation of motion, McLachlan’s approach minimizes the distance between the true derivative of the state and its projection within the variational manifold. Specifically, the evolution path is chosen such that the deviation $(\frac{\partial}{\partial\tau} + O - E_\tau)|\psi(\boldsymbol{\theta}(\tau))\rangle$ remains orthogonal to the tangent space of allowed variations. The condition is formally expressed as:

$$\delta \left\| \left(\frac{\partial}{\partial\tau} + O - E_\tau \right) |\psi(\boldsymbol{\theta}(\tau))\rangle \right\| = 0, \quad (76)$$

which ensures that the evolution follows the most faithful trajectory allowed by the variational parameters. This principle is particularly suitable when parameters are constrained to be real, as in many practical ansatz constructions.

VII. REFORMULATE QNGD-BASED VQAS AND BUILD A FIRST-PRINCIPLE EQUIVALENCE WITH QITE

Stokes *et al.* [32] presents that the optimizer QNGD induces same parameter update rule with QITE from a phenomenological observation, yet the reason behind remains unclear. We provide a detailed derivations to address this gap.

A. Objective Function of QNGD-based VQAs with General Loss Function

QNGD-based VQAs update their variational parameters by determining an optimal update direction $\Delta\theta$ within a local neighborhood of parameters θ , while accounting for the underlying geometry of the quantum state manifold, which is described by the Quantum Fisher Information Matrix (QFIM) [32, 45]. To generalize the optimization framework, we introduce a general loss function of the form:

$$\mathcal{L}(\theta) = f(\langle O \rangle), \quad (77)$$

where $\langle O \rangle = \langle \psi(\theta) | O | \psi(\theta) \rangle$ denotes the expected value of a given observable O , and f is a differentiable scalar-valued function. A commonly used example is the quadratic loss function:

$$\mathcal{L}(\theta) = \frac{1}{2} (\langle O \rangle - O_0)^2, \quad (78)$$

where O_0 is the target observable value.

In gradient descent, optimization is performed under Euclidean geometry, where the update direction is constrained by an ℓ_2 norm:

$$\|\Delta\theta\| \leq \varepsilon, \quad \text{with } \Delta\theta = \epsilon\nu, \quad (79)$$

where ν is an arbitrary unit vector and $\epsilon > 0$ is a small scalar step size. However, in quantum variational algorithms, such parameter displacements may not reflect the true distance between quantum states. Instead, QNGD regularizes updates using the fidelity distance between quantum states:

$$d_f(|\psi(\theta)\rangle, |\psi(\theta + \Delta\theta)\rangle) = 1 - |\langle \psi(\theta) | \psi(\theta + \Delta\theta) \rangle|^2. \quad (80)$$

In the infinitesimal limit $\Delta\theta \rightarrow 0$, the fidelity distance becomes the squared line element on the Hilbert space manifold:

$$ds^2 = \sum_{\ell_1, \ell_2} g_{\ell_1 \ell_2}(\theta) \Delta\theta_{\ell_1} \Delta\theta_{\ell_2} = \frac{1}{4} \sum_{\ell_1, \ell_2} \mathcal{F}_{\ell_1 \ell_2}(\theta) \Delta\theta_{\ell_1} \Delta\theta_{\ell_2}, \quad (81)$$

where $g_{\ell_1 \ell_2}(\theta)$ is the Fubini–Study metric tensor, defined as:

$$g_{\ell_1 \ell_2}(\theta) = \text{Re} \left[\left\langle \frac{\partial \psi(\theta)}{\partial \theta_{\ell_1}} \middle| \frac{\partial \psi(\theta)}{\partial \theta_{\ell_2}} \right\rangle - \left\langle \frac{\partial \psi(\theta)}{\partial \theta_{\ell_1}} \middle| \psi(\theta) \right\rangle \left\langle \psi(\theta) \middle| \frac{\partial \psi(\theta)}{\partial \theta_{\ell_2}} \right\rangle \right], \quad (82)$$

and $\mathcal{F}(\theta) = 4g(\theta)$ is the QFIM given by:

$$\mathcal{F}_{\ell_1 \ell_2}(\theta) = 4 \text{Re} \left[\left\langle \frac{\partial \psi(\theta)}{\partial \theta_{\ell_1}} \middle| \frac{\partial \psi(\theta)}{\partial \theta_{\ell_2}} \right\rangle - \left\langle \frac{\partial \psi(\theta)}{\partial \theta_{\ell_1}} \middle| \psi(\theta) \right\rangle \left\langle \psi(\theta) \middle| \frac{\partial \psi(\theta)}{\partial \theta_{\ell_2}} \right\rangle \right]. \quad (83)$$

With fidelity-based regularization, the update direction is formulated as the solution to a constrained optimization problem:

$$\Delta\theta^* = \arg \min_{\Delta\theta \text{ s.t. } d_f(\psi(\theta), \psi(\theta + \Delta\theta))=c} \mathcal{L}(\theta + \Delta\theta), \quad (84)$$

where $c > 0$ is a fixed fidelity threshold. Reformulating this as a Lagrangian and applying a first-order Taylor expansion of \mathcal{L} , we obtain:

$$\begin{aligned} \Delta\theta^* &= \arg \min_{\Delta\theta} \left[\mathcal{L}(\theta) + \nabla_\theta \mathcal{L}(\theta) \cdot \Delta\theta + \frac{\lambda}{4} \sum_{\ell_1, \ell_2} \mathcal{F}_{\ell_1 \ell_2}(\theta) \Delta\theta_{\ell_1} \Delta\theta_{\ell_2} - \lambda c \right] \\ &\equiv \arg \min_{\Delta\theta} \nabla_\theta \mathcal{L}(\theta) \cdot \Delta\theta + \frac{\lambda}{4} \sum_{\ell_1, \ell_2} \mathcal{F}_{\ell_1 \ell_2}(\theta) \Delta\theta_{\ell_1} \Delta\theta_{\ell_2}, \end{aligned} \quad (85)$$

where λ is the Lagrange multiplier associated with the fidelity constraint. Substituting the definition of $\mathcal{F}_{\ell_1 \ell_2}$, we obtain the explicit objective of QNGD-based VQAs:

$$\begin{aligned} \Delta\theta^* = \arg \min_{\Delta\theta} & \nabla_{\theta} \mathcal{L}(\theta) \cdot \Delta\theta \\ & + \lambda \sum_{\ell_1, \ell_2} \operatorname{Re} \left[\left\langle \frac{\partial \psi(\theta)}{\partial \theta_{\ell_1}} \middle| \frac{\partial \psi(\theta)}{\partial \theta_{\ell_2}} \right\rangle - \left\langle \frac{\partial \psi(\theta)}{\partial \theta_{\ell_1}} \middle| \psi(\theta) \right\rangle \left\langle \psi(\theta) \middle| \frac{\partial \psi(\theta)}{\partial \theta_{\ell_2}} \right\rangle \right] \Delta\theta_{\ell_1} \Delta\theta_{\ell_2}. \end{aligned} \quad (86)$$

B. Equivalence between QITE and QNGD-based VQAs In the Objective Function

In this section, we demonstrate that projected QITE shares an equivalent objective function with QNGD-based VQAs in the continuous-time limit. Specifically, projected QITE seeks to variationally approximate the imaginary-time evolved state $e^{-O\Delta\tau}\psi(\theta(\tau))$ using a parametrized ansatz $\psi(\theta + \Delta\theta)$, by maximizing their fidelity:

$$\arg \max_{\Delta\theta \in \mathbb{R}^d} |\langle e^{-O\Delta\tau}\psi_\theta, \psi_{\theta+\Delta\theta} \rangle|^2 \equiv \arg \min_{\Delta\theta \in \mathbb{R}^d} |1 - \langle e^{-O\Delta\tau}\psi_\theta, \psi_{\theta+\Delta\theta} \rangle|^2. \quad (87)$$

Assuming small $\Delta\tau$ and $\Delta\theta$, we follow the expansion technique introduced in Stokes *et al.* [32]. Letting $\bar{\psi}_\theta := e^{-O\Delta\tau}\psi_\theta$, we perform a second-order Taylor expansion and obtain:

$$\begin{aligned} \arg \min_{\Delta\theta \in \mathbb{R}^d} (1 - |\langle \bar{\psi}_\theta, \psi_{\theta+\Delta\theta} \rangle|^2) &= \arg \min_{\Delta\theta \in \mathbb{R}^d} (1 - |\langle \bar{\psi}_\theta, \psi(\theta) \rangle|^2) \\ &\quad + \left[\left\langle \frac{\partial \psi(\theta)}{\partial \theta_{\ell_1}}, O\psi(\theta) \right\rangle + \left\langle O\psi(\theta), \frac{\partial \psi(\theta)}{\partial \theta_{\ell_1}} \right\rangle \right] \Delta\theta_{\ell_1} \Delta\tau \\ &\quad + \operatorname{Re} [G_{\ell_1 \ell_2}(\theta)] \Delta\theta_{\ell_1} \Delta\theta_{\ell_2}, \end{aligned} \quad (88)$$

where $G_{\ell_1 \ell_2}(\theta)$ is defined in Eq. (7). Discarding constant terms and reorganizing the expression, we have:

$$\begin{aligned} &\arg \min_{\Delta\theta \in \mathbb{R}^d} \left[\left\langle \frac{\partial \psi(\theta)}{\partial \theta_{\ell_1}}, O\psi(\theta) \right\rangle + \left\langle O\psi(\theta), \frac{\partial \psi(\theta)}{\partial \theta_{\ell_1}} \right\rangle \right] \Delta\theta_{\ell_1} \Delta\tau \\ &\quad + \operatorname{Re} [G_{\ell_1 \ell_2}(\theta)] \Delta\theta_{\ell_1} \Delta\theta_{\ell_2} \\ &= \arg \min_{\Delta\theta \in \mathbb{R}^d} (\Delta\tau)^2 \left\{ \left[\left\langle \frac{\partial \psi(\theta)}{\partial \theta_{\ell_1}}, O\psi(\theta) \right\rangle + \left\langle O\psi(\theta), \frac{\partial \psi(\theta)}{\partial \theta_{\ell_1}} \right\rangle \right] \frac{\Delta\theta_{\ell_1}}{\Delta\tau} \right. \\ &\quad \left. + \operatorname{Re} [G_{\ell_1 \ell_2}(\theta)] \frac{\Delta\theta_{\ell_1}}{\Delta\tau} \frac{\Delta\theta_{\ell_2}}{\Delta\tau} \right\}. \end{aligned} \quad (89)$$

To take the continuous-time limit, we define the instantaneous update direction $\delta := \frac{d\theta}{d\tau}$. Substituting $\Delta\theta = \delta \cdot \Delta\tau$ into the expression above and letting $\Delta\tau \rightarrow 0$, the objective becomes:

$$\arg \min_{\delta \in \mathbb{R}^d} \frac{\partial \langle \psi(\theta(\tau)) | O | \psi(\theta(\tau)) \rangle}{\partial \theta_{\ell_1}} \delta_{\ell_1} + \sum_{\ell_1, \ell_2 \in [d]} \operatorname{Re} [G_{\ell_1 \ell_2}(\theta)] \delta_{\ell_1} \delta_{\ell_2}. \quad (90)$$

This final expression is identical in form to the objective of QNGD-based VQAs in Eq. (86) when the loss function is chosen as a linear expectation value $\mathcal{L}(\theta) = \langle O \rangle$, and the regularization parameter λ is set to 1. This equivalence reveals a deep connection between projected QITE and QNGD-based VQAs in the continuous-time regime, unifying them under a shared optimization principle rooted in the geometry of quantum state space.

C. Variational Principle Formulation of QNGD-based VQAs with General Loss Function

To facilitate a first-principle comparison with QITE, we formalize the variational principle of QNGD-based VQAs. Using a first-order Taylor approximation, the change in the loss function \mathcal{L} under a small update $\Delta\theta$ is:

$$\Delta\mathcal{L} \approx \nabla \mathcal{L}^\top \Delta\theta, \quad (91)$$

where $\nabla \mathcal{L}$ represents the gradient. To respect the geometric structure of the quantum state manifold, QNGD-based VQAs introduce a curvature penalty via the Fubini–Study metric [32, 46, 60, 61]:

$$\|\Delta\theta\|_{\mathcal{F}}^2 = \Delta\theta^\top \mathcal{F} \Delta\theta, \quad (92)$$

To unify optimization and geometry, we define a variational functional:

$$\mathcal{S}[\Delta\theta] = \nabla\mathcal{L}^\top\Delta\theta + \frac{1}{2\eta}\Delta\theta^\top\mathcal{F}\Delta\theta, \quad (93)$$

where $\eta > 0$ is a regularization parameter controlling step size. The first term captures loss descent, while the second penalizes large displacements in the quantum state space.

The variational principle of QNGD-based VQAs then requires the first variation of this functional to vanish:

$$\delta \left[\nabla\mathcal{L}^\top\Delta\theta + \frac{1}{2\eta}\Delta\theta^\top\mathcal{F}\Delta\theta \right] = 0, \quad (94)$$

Now, Consider a general differentiable loss function $\mathcal{L}(\theta) = f(\langle O \rangle)$, where O is a Hermitian observable,

$$\Delta\mathcal{L} = f'(\langle O \rangle_\theta)\Delta\langle O \rangle_\theta, \quad (95)$$

and the first-order change of the expectation value is

$$\Delta\langle O \rangle_\theta \approx \nabla\langle O \rangle_\theta^\top\Delta\theta, \quad (96)$$

which yields the variational principle of QNGD-based VQAs with general loss function:

$$\delta \left[f'(\langle O \rangle_\theta)\nabla\langle O \rangle_\theta^\top\Delta\theta + \frac{1}{2\eta}\Delta\theta^\top\mathcal{F}\Delta\theta \right] = \delta[\mathcal{J}_{\text{General}}] = 0, \quad (97)$$

where we define $\mathcal{J}_{\text{General}}$ as the variational functional of QNGD-based VQAs with general loss function.

D. Equivalence between QITE and QNGD-based VQAs In the Variational Principle with Linear Loss Function

We firstly connect the two variational principles with linear loss function. As a review, the variational principle of QNGD-based VQAs focuses on parameter changes $\Delta\theta$, while QITE's variational principle directly constrains the quantum state's time evolution $\partial_\tau|\psi(\theta(\tau))\rangle$. The connection between the two can be established by the effect of parameter changes on quantum state evolution:

A parameter change $\Delta\theta$ leads to a change in the quantum state:

$$|\psi(\theta + \Delta\theta)\rangle \approx |\psi(\theta)\rangle + \sum_{\ell_1} \partial_{\theta_{\ell_1}} |\psi(\theta)\rangle \Delta\theta_{\ell_1}. \quad (98)$$

In the continuous-time limit $\eta \rightarrow 0$, the rate of change of parameters $\dot{\theta} = \frac{d\theta}{d\tau}$ corresponds to the time derivative of the quantum state:

$$\partial_\tau|\psi(\theta(\tau))\rangle = \sum_{\ell_1} \partial_{\theta_{\ell_1}} |\psi\rangle \dot{\theta}_{\ell_1}. \quad (99)$$

Accordingly, we can convert the variational principle of QNGD-based VQAs into the form including quantum state evolution. Reviewing Eq. (94), when with linear loss function:

$$\mathcal{L}(\theta) = \langle O \rangle = \langle \psi(\theta) | O | \psi(\theta) \rangle, \quad (100)$$

the gradient is computed as:

$$\nabla_{\ell_1}\mathcal{L} = \frac{\partial\mathcal{L}}{\partial\theta_{\ell_1}} = \langle \partial_{\ell_1}\psi | O | \psi \rangle + \langle \psi | O | \partial_{\ell_1}\psi \rangle. \quad (101)$$

Exploiting Hermitian symmetry, we have:

$$\begin{aligned} \nabla_{\ell_1}\mathcal{L} &= \langle \partial_{\ell_1}\psi | O | \psi \rangle + (\langle \partial_{\ell_1}\psi | O | \psi \rangle)^* \\ &= 2\operatorname{Re}(\langle \partial_{\ell_1}\psi | O | \psi \rangle). \end{aligned} \quad (102)$$

The factor of 2 can often be absorbed into the learning rate, leading to the standard gradient expression:

$$\nabla_{\ell_1} \mathcal{L} = \operatorname{Re} (\langle \partial_{\ell_1} \psi | O | \psi \rangle). \quad (103)$$

Using Eq. (103) and $\Delta \boldsymbol{\theta} = \dot{\boldsymbol{\theta}} \eta$, the variational principle under the continuous-time limit becomes:

$$\delta \left[\eta \sum_{\ell_1} \operatorname{Re} (\langle \partial_{\ell_1} \psi | O | \psi \rangle) \dot{\boldsymbol{\theta}}_{\ell_1} + \frac{\eta}{2} \sum_{\ell_1, \ell_2} F_{\ell_1 \ell_2} \dot{\boldsymbol{\theta}}_{\ell_1} \dot{\boldsymbol{\theta}}_{\ell_2} \right] = \delta [\mathcal{S}_{\text{inst}}] = 0, \quad (104)$$

where $\mathcal{S}_{\text{inst}}$ is the instantaneous variational functional.

As $\eta \rightarrow 0$, the sum over time steps converges to an integral over τ with $\eta \rightarrow d\tau$. The total variation is then:

$$\delta \int \left(\sum_{\ell_1} \operatorname{Re} (\langle \partial_{\ell_1} \psi | O | \psi \rangle) \dot{\boldsymbol{\theta}}_{\ell_1} + \frac{1}{2} \sum_{\ell_1, \ell_2} F_{\ell_1 \ell_2} \dot{\boldsymbol{\theta}}_{\ell_1} \dot{\boldsymbol{\theta}}_{\ell_2} \right) d\tau = \delta [\mathcal{J}_{\text{Linear}}] = 0, \quad (105)$$

where we define $\mathcal{J}_{\text{Linear}}$ as the continuous-time action for QNGD-based VQAs with linear loss function.

Substituting Eq. (99) into the first term:

$$\begin{aligned} \sum_{\ell_1} \operatorname{Re} (\langle \partial_{\ell_1} \psi | O | \psi \rangle) \dot{\boldsymbol{\theta}}_{\ell_1} &= \operatorname{Re} \left(\sum_{\ell_1} \langle \partial_{\ell_1} \psi | O | \psi \rangle \dot{\boldsymbol{\theta}}_{\ell_1} \right) \\ &= \operatorname{Re} (\langle \partial_{\tau} \psi | O | \psi \rangle). \end{aligned} \quad (106)$$

Remark. This transformation projects the parameter gradient onto the quantum state's evolution direction.

Substituting Eq. (99) into the second term:

$$\begin{aligned} \sum_{\ell_1, \ell_2} F_{\ell_1 \ell_2} \dot{\boldsymbol{\theta}}_{\ell_1} \dot{\boldsymbol{\theta}}_{\ell_2} &= \operatorname{Re} \left(\sum_{\ell_1, \ell_2} [\langle \partial_{\ell_1} \psi | \partial_{\ell_2} \psi \rangle - \langle \partial_{\ell_1} \psi | \psi \rangle \langle \psi | \partial_{\ell_2} \psi \rangle] \dot{\boldsymbol{\theta}}_{\ell_1} \dot{\boldsymbol{\theta}}_{\ell_2} \right) \\ &= \operatorname{Re} (\langle \partial_{\tau} \psi | \partial_{\tau} \psi \rangle - \langle \partial_{\tau} \psi | \psi \rangle \langle \psi | \partial_{\tau} \psi \rangle). \end{aligned} \quad (107)$$

From the normalization condition $\partial_{\tau} \langle \psi | \psi \rangle = 0$, we derive:

$$\langle \partial_{\tau} \psi | \psi \rangle + \langle \psi | \partial_{\tau} \psi \rangle = 0, \text{ i.e., } \operatorname{Re} (\langle \psi | \partial_{\tau} \psi \rangle) = 0. \quad (108)$$

This simplifies the metric to:

$$\sum_{\ell_1, \ell_2} F_{\ell_1 \ell_2} \dot{\boldsymbol{\theta}}_{\ell_1} \dot{\boldsymbol{\theta}}_{\ell_2} = \langle \partial_{\tau} \psi | \partial_{\tau} \psi \rangle. \quad (109)$$

Therefore, the variational principle is reformulated as:

$$\delta \int \left(\operatorname{Re} (\langle \partial_{\tau} \psi | O | \psi \rangle) + \frac{1}{2} \langle \partial_{\tau} \psi | \partial_{\tau} \psi \rangle \right) d\tau = \delta [\mathcal{J}_{\text{Linear}}] = 0, \quad (110)$$

Meanwhile, according to Appendix VI, QITE's variational principle is associated with the following instantaneous variational functional :

$$\left\| \left(\frac{\partial}{\partial \tau} + O - E_{\tau} \right) |\psi(\boldsymbol{\theta}(\tau))\rangle \right\|^2, \quad (111)$$

Expanding this gives:

$$\begin{aligned} &\|(\partial_{\tau} + O - E_{\tau}) |\psi\rangle\|^2 \\ &= \langle \partial_{\tau} \psi | \partial_{\tau} \psi \rangle + \langle \psi | (O - E_{\tau})^2 | \psi \rangle \\ &\quad + 2 \operatorname{Re} (\langle \partial_{\tau} \psi | (O - E_{\tau}) | \psi \rangle) \\ &= \langle \partial_{\tau} \psi | \partial_{\tau} \psi \rangle + \langle \psi | O^2 | \psi \rangle - E_{\tau}^2 + 2 \operatorname{Re} (\langle \partial_{\tau} \psi | O | \psi \rangle) \end{aligned} \quad (112)$$

Ignoring constant terms (*i.e.*, E_τ^2 and $\langle O^2 \rangle$ if O is fixed) and dividing by an overall constant factor (which does not affect the variational dynamics), the variational principle can be reduced to:

$$\delta \int \left(\text{Re}(\langle \partial_\tau \psi | O | \psi \rangle) + \frac{1}{2} \langle \partial_\tau \psi | \partial_\tau \psi \rangle \right) d\tau = \delta [\mathcal{D}_{Linear}] = 0, \quad (113)$$

which is exactly the same as QNGD-based VQAs' continuous-time variational principle with linear loss function in Eq. (110).

In sum, when $\mathcal{L}(\boldsymbol{\theta}) = \langle O \rangle_{\boldsymbol{\theta}}$, $E_\tau = \mathcal{L}(\boldsymbol{\theta}(\tau))$, and if O is a stationary operator, higher-order terms $\langle O^2 \rangle - E_\tau^2$ can be treated as constants in the variational principle (or canceled by normalization), and thus do not affect the resulting dynamics. Thus, the two variational functionals are related as:

$$\mathcal{D}_{Linear} \propto \mathcal{J}_{Linear} + \text{constant}, \quad (114)$$

i.e., the two variational functionals are equivalent up to a constant and a scaling factor, hence their variational principles are equivalent and lead to the same dynamics.

Remark. The QNGD-based VQAs' variational principle implicitly optimizes the quantum state evolution path through a balance between geometric penalty in the parameter space and the rate of observable expectation decay, while QITE's variational principle explicitly constrains quantum state evolution to approximate imaginary time dynamics. When given a linear loss function, *i.e.*, an observable expectation, both variational principles reduce to the same problem in the continuous-time limit.

E. Extend the Variational Principle of QITE

Similar to the above section, we expand $\left\| \left(\frac{\partial}{\partial \tau} + O - E_\tau \right) |\psi(\boldsymbol{\theta}(\tau))\rangle \right\|^2$, which gives:

$$\begin{aligned} & \left\| \left(\frac{\partial}{\partial \tau} + O - E_\tau \right) |\psi(\boldsymbol{\theta}(\tau))\rangle \right\|^2 \\ &= \left(\left(\frac{\partial}{\partial \tau} + O - E_\tau \right) |\psi(\boldsymbol{\theta}(\tau))\rangle \right)^\dagger \left(\frac{\partial}{\partial \tau} + O - E_\tau \right) |\psi(\boldsymbol{\theta}(\tau))\rangle \\ &= \sum_{\ell_1, \ell_2} \frac{\partial \langle \psi(\boldsymbol{\theta}(\tau)) |}{\partial \boldsymbol{\theta}_{\ell_1}} \frac{\partial |\psi(\boldsymbol{\theta}(\tau))\rangle}{\partial \boldsymbol{\theta}_{\ell_2}} \dot{\boldsymbol{\theta}}_{\ell_1} \dot{\boldsymbol{\theta}}_{\ell_2} + \sum_{\ell_1} \frac{\partial \langle \psi(\boldsymbol{\theta}(\tau)) |}{\partial \boldsymbol{\theta}_{\ell_1}} (O - E_\tau) |\psi(\boldsymbol{\theta}(\tau))\rangle \dot{\boldsymbol{\theta}}_{\ell_1} \\ &\quad + \sum_{\ell_1} \langle \psi(\boldsymbol{\theta}(\tau)) | (O - E_\tau) \frac{\partial |\psi(\boldsymbol{\theta}(\tau))\rangle}{\partial \boldsymbol{\theta}_{\ell_1}} \dot{\boldsymbol{\theta}}_{\ell_1} + \langle \psi(\boldsymbol{\theta}(\tau)) | (O - E_\tau)^2 |\psi(\boldsymbol{\theta}(\tau))\rangle. \end{aligned} \quad (115)$$

Different loss function types only change $\sum_{\ell_1} \frac{\partial \langle \psi(\boldsymbol{\theta}(\tau)) |}{\partial \boldsymbol{\theta}_{\ell_1}} (O - E_\tau) |\psi(\boldsymbol{\theta}(\tau))\rangle \dot{\boldsymbol{\theta}}_{\ell_1} + \sum_{\ell_1} \langle \psi(\boldsymbol{\theta}(\tau)) | (O - E_\tau) \frac{\partial |\psi(\boldsymbol{\theta}(\tau))\rangle}{\partial \boldsymbol{\theta}_{\ell_1}} \dot{\boldsymbol{\theta}}_{\ell_1}$, as analyzed in Appendix VII D. Now we extend the principle to both quadratic loss function and general loss function.

Quadratic Loss Extension. For convenience, we do not consider an $\frac{1}{x}$ scaling for the loss types beyond linear loss function. The quadratic loss function we consider here is $\mathcal{L} = (\langle \psi | O | \psi \rangle)^2$. Then the gradient becomes:

$$\begin{aligned} \frac{\partial \langle \psi | O | \psi \rangle^2}{\partial \boldsymbol{\theta}_{\ell_1}} &= \langle \psi | O | \psi \rangle \cdot \frac{\partial \langle \psi | O | \psi \rangle}{\partial \boldsymbol{\theta}_{\ell_1}} \\ &= 2E_\tau \left[\frac{\partial \langle \psi |}{\partial \boldsymbol{\theta}_{\ell_1}} (O - E_\tau) |\psi\rangle + \langle \psi | (O - E_\tau) \frac{\partial |\psi\rangle}{\partial \boldsymbol{\theta}_{\ell_1}} + \frac{\partial \langle \psi |}{\partial \boldsymbol{\theta}_{\ell_1}} E_\tau |\psi\rangle + \langle \psi | E_\tau \frac{\partial |\psi\rangle}{\partial \boldsymbol{\theta}_{\ell_1}} \right] \\ &= 2E_\tau \left[\frac{\partial \langle \psi |}{\partial \boldsymbol{\theta}_{\ell_1}} (O - E_\tau) |\psi\rangle + \langle \psi | (O - E_\tau) \frac{\partial |\psi\rangle}{\partial \boldsymbol{\theta}_{\ell_1}} \right]. \end{aligned} \quad (116)$$

where the last equality holds due to the normalization condition $\langle \psi | \psi \rangle = 1$ and E_τ is a scalar.

Substituting into the variational condition, we obtain:

$$\delta \left\| \left(\frac{\partial}{\partial \tau} + E_\tau (O - E_\tau) \right) |\psi(\tau)\rangle \right\| = 0. \quad (117)$$

General Loss Extension. Let $\mathcal{L} = f(\langle \psi | O | \psi \rangle)$. The chain rule yields:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_{\ell_1}} &= f'(E_\tau) \cdot \frac{\partial \langle \psi | O | \psi \rangle}{\partial \boldsymbol{\theta}_{\ell_1}} \\ &= f'(E_\tau) \cdot \left[\frac{\partial \langle \psi |}{\partial \boldsymbol{\theta}_{\ell_1}} (O - E_\tau) |\psi\rangle + \langle \psi | (O - E_\tau) \frac{\partial |\psi\rangle}{\partial \boldsymbol{\theta}_{\ell_1}} \right]. \end{aligned} \quad (118)$$

Thus, the generalized McLachlan variational principle becomes:

$$\delta \left\| \left(\frac{\partial}{\partial \tau} + f'(E_\tau)(O - E_\tau) \right) |\psi(\tau)\rangle \right\| = 0. \quad (119)$$

Remark. This generalized variational principle retains the fidelity to imaginary-time dynamics while enabling flexible loss definitions. When $f(E_\tau) = E_\tau$, we recover the standard McLachlan variational formulation for QITE.

F. Equivalence between QITE and QNGD-based VQAs In the Variational Principle with General Loss Function

According to Appendix [VII C](#) and [VII D](#), for a general differentiable loss function $f(\langle O \rangle)$, the continuous-time limit ($\eta \rightarrow 0$) of QNGD-based VQAs yields the following variational functional:

$$\mathcal{J}_{\text{general}} = \int \left[f'(\langle O \rangle) \operatorname{Re} (\langle \partial_\tau \psi | O | \psi \rangle) + \frac{1}{2} \langle \partial_\tau \psi | \partial_\tau \psi \rangle \right] d\tau, \quad (120)$$

where $\langle O \rangle = \langle \psi(\theta) | O | \psi(\theta) \rangle$, and O denotes the observable.

From Appendix [VII E](#), for the QITE with general loss function, the variational functional is given by:

$$\mathcal{D}_{\text{general}} = \| (\partial_\tau + f'(\langle O \rangle)(O - \langle O \rangle)) |\psi\rangle \|^2. \quad (121)$$

Expanding Eq. (121), we obtain:

$$\begin{aligned} \mathcal{D}_{\text{general}} &= \langle \partial_\tau \psi | \partial_\tau \psi \rangle + f'(\langle O \rangle)^2 \langle (O - \langle O \rangle)^2 \rangle \\ &\quad + 2f'(\langle O \rangle) \operatorname{Re} (\langle \partial_\tau \psi | O - \langle O \rangle | \psi \rangle). \end{aligned} \quad (122)$$

Since $\langle \psi | \psi \rangle = 1$, we similarly divide by an overall constant factor (which does not affect the variational dynamics), thus the variational functional reduces to:

$$\mathcal{D}_{\text{general}} = \frac{1}{2} \langle \partial_\tau \psi | \partial_\tau \psi \rangle + f'(\langle O \rangle) \operatorname{Re} (\langle \partial_\tau \psi | O | \psi \rangle) + \text{constant}, \quad (123)$$

where the constant term $f'(\langle O \rangle)^2 \langle (O - \langle O \rangle)^2 \rangle$ does not influence the dynamics.

Comparing with Eq. (120), we observe:

$$\mathcal{D}_{\text{general}} \propto \mathcal{J}_{\text{general}} + \text{const}, \quad (124)$$

which shows that minimizing $\mathcal{D}_{\text{general}}$ is equivalent to minimizing $\mathcal{J}_{\text{general}}$, up to an overall scaling and additive constant. Since these factors do not affect the optimization trajectory, the two variational principles are equivalent with general loss function.

Remark. This equivalence reveals that under general differentiable loss functions $f(\langle O \rangle)$, QNGD-based VQAs and QITE share the same variational principles. The scalar factor $f'(\langle O \rangle)$ modulates the imaginary time evolution rate without altering the optimal descent direction. Hence, the equivalence between parameter-space natural gradient descent and quantum-state imaginary time evolution persists beyond the linear loss case.

VIII. PROOF FOR PROPOSITION [II.8](#)

We present a simplified, semi-rigorous proof for Proposition [II.8](#).

Firstly, we prove that

$$\overline{K_{\text{QITE}}} = \frac{N+1}{N} \overline{K_{\text{GD}}} \quad (125)$$

holds under Assumption 2, without considering an explicit analysis of high-order fluctuations. We begin by expanding the expectation:

$$\overline{K_{\text{QITE}}} = \overline{\sum_{\ell_1, \ell_2} \frac{\partial \epsilon}{\partial \theta_{\ell_1}} g_{\ell_1 \ell_2}^+ \frac{\partial \epsilon}{\partial \theta_{\ell_2}}} = \sum_{\ell_1, \ell_2} \overline{\frac{\partial \epsilon}{\partial \theta_{\ell_1}} g_{\ell_1 \ell_2}^+} \frac{\partial \epsilon}{\partial \theta_{\ell_2}}. \quad (126)$$

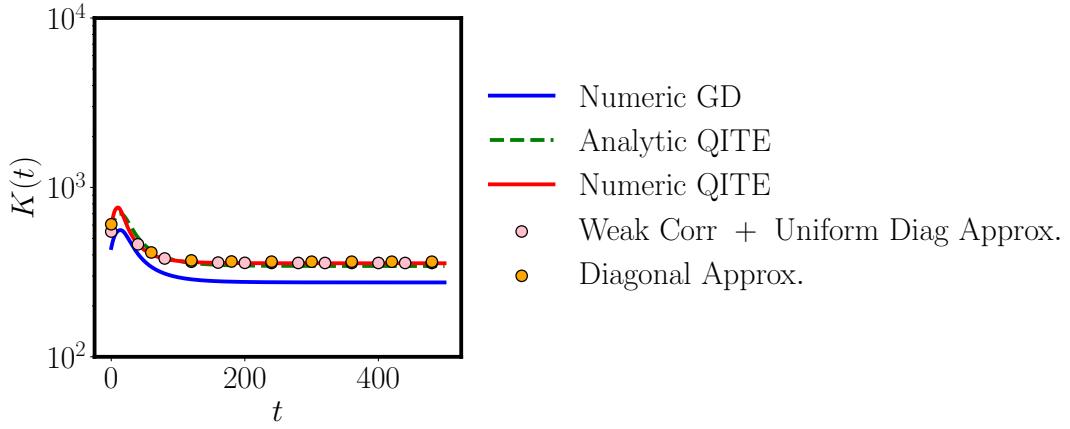


FIG. 4. The diagonal approximation (yellow circles) is highlighted at selected steps to evaluate its agreement with the full K_{QITE} . The agreement confirms that diagonal elements dominate the kernel structure, supporting the validity of the approximation in practice.

From Lemma II.7, the expected metric tensor is diagonal:

$$\overline{g_{\ell_1 \ell_2}} = \frac{N}{N+1} \delta_{\ell_1 \ell_2}, \quad (127)$$

and the entrywise variance is given by:

$$\text{Var}(g_{\ell_1 \ell_2}) = \begin{cases} \frac{2}{N^2}, & \text{if } \ell_1 = \ell_2, \\ \frac{1}{2N}, & \text{if } \ell_1 \neq \ell_2. \end{cases} \quad (128)$$

Together with the results under Haar random ensemble:

$$\begin{aligned} \overline{\frac{\partial \epsilon}{\partial \theta_\ell}} &= 0, \\ \overline{\frac{\partial \epsilon}{\partial \theta_{\ell_1}} \frac{\partial \epsilon}{\partial \theta_{\ell_2}}} &= \delta_{\ell_1 \ell_2} \cdot \overline{\frac{\partial \epsilon}{\partial \theta_\ell} \frac{\partial \epsilon}{\partial \theta_\ell}}, \end{aligned} \quad (129)$$

we can make a diagonal approximation (justified by Fig. 4 that denotes “Diagonal Approx.”) as follows:

$$\begin{aligned} \overline{K_{\text{QITE}}} &= \sum_{\ell_1, \ell_2} \overline{\frac{\partial \epsilon}{\partial \theta_{\ell_1}} \cdot g_{\ell_1 \ell_2}^+ \cdot \frac{\partial \epsilon}{\partial \theta_{\ell_2}}} \\ &\stackrel{\text{(diagonal approx.)}}{\approx} \sum_{\ell} \overline{\left(\frac{\partial \epsilon}{\partial \theta_\ell} \right)^2 \cdot g_{\ell \ell}^+}. \end{aligned} \quad (130)$$

Since the fluctuations are small in the large- N limit, we can also approximate (see analysis in Section II B):

$$\overline{g_{\ell \ell}^+} \approx (\bar{g}^{-1})_{\ell \ell} = \frac{N+1}{N}. \quad (131)$$

Remark. Although $\overline{A^+} \neq (\overline{A})^{-1}$ in general for a general matrix A , the deviation is suppressed by the small variance of $g_{\ell_1 \ell_2}$, therefore we assume the approximation holds for g .

Therefore, since we have $\overline{g_{\ell \ell}^+} = \frac{N+1}{N}$ (a constant independent of ℓ), we make below approximation (justified by Fig. 4 that denotes “Weak Corr + Uniform Diag Approx.”):

$$\overline{K_{\text{QITE}}} \approx \frac{N+1}{N} \left(\sum_{\ell} \overline{\frac{\partial \epsilon}{\partial \theta_{\ell}} \frac{\partial \epsilon}{\partial \theta_{\ell}}} \right) \quad (132)$$

where we assume weak correlations between g^+ and the $\left(\frac{\partial \epsilon}{\partial \theta_{\ell}}\right)^2$ while uniform diagonals.
Thus we reach:

$$\overline{K_{\text{QITE}}} \approx \frac{N+1}{N} \overline{K_{\text{GD}}}. \quad (133)$$

Now, we derive the relation regarding ϵ . Using Eq. (29), we express the residual training error for both GD and QITE:

$$\begin{aligned} \epsilon_{\text{GD}}(t) &= \epsilon(0) \exp(-\eta \overline{K_{\text{GD}}} t), \\ \epsilon_{\text{QITE}}(t) &= \epsilon(0) \exp(-\eta \overline{K_{\text{QITE}}} t). \end{aligned} \quad (134)$$

Therefore,

$$\epsilon_{\text{QITE}}(t) = \epsilon(0) \exp\left(-\eta \cdot \frac{N+1}{N} \cdot \overline{K_{\text{GD}}} \cdot t\right). \quad (135)$$

Dividing both error expressions, we obtain:

$$\frac{\epsilon_{\text{QITE}}(t)}{\epsilon_{\text{GD}}(t)} = \exp\left(-\eta \overline{K_{\text{GD}}} t \cdot \left(\frac{N+1}{N} - 1\right)\right) \quad (136)$$

$$= \exp\left(-\frac{\eta t}{N} \overline{K_{\text{GD}}}\right). \quad (137)$$

i.e.,

$$\epsilon_{\text{QITE}}(t) \approx \epsilon_{\text{GD}}(t) \cdot \exp\left(-\frac{\eta t}{N} \overline{K_{\text{GD}}}\right). \quad (138)$$

IX. PROOF FOR PROPOSITION II.11

We present a simplified, semi-rigorous proof for Proposition II.8.

Firstly, we prove that

$$\overline{\lambda_{\text{QITE}}(t)} = \frac{N+1}{N} \overline{\lambda_{\text{GD}}(t)} \quad (139)$$

holds under Assumption 2, without considering an explicit analysis of high-order fluctuations.
Recall the definitions:

$$\lambda_{\text{QITE}}(t) = \frac{\mu_{\text{QITE}}(t)}{K_{\text{QITE}}(t)}, \quad \lambda_{\text{GD}}(t) = \frac{\mu_{\text{GD}}(t)}{K_{\text{GD}}(t)}, \quad (140)$$

where

$$\mu_{\text{QITE}} = \sum_{\ell_1, \ell_2, \ell_3, \ell_4} g_{\ell_1 \ell_3}^+ g_{\ell_2 \ell_4}^+ \frac{\partial^2 \epsilon}{\partial \theta_{\ell_1} \partial \theta_{\ell_2}} \frac{\partial \epsilon}{\partial \theta_{\ell_3}} \frac{\partial \epsilon}{\partial \theta_{\ell_4}}, \quad (141)$$

and

$$\mu_{\text{GD}} = \sum_{\ell_1, \ell_2} \frac{\partial^2 \epsilon}{\partial \theta_{\ell_1} \partial \theta_{\ell_2}} \frac{\partial \epsilon}{\partial \theta_{\ell_1}} \frac{\partial \epsilon}{\partial \theta_{\ell_2}}. \quad (142)$$

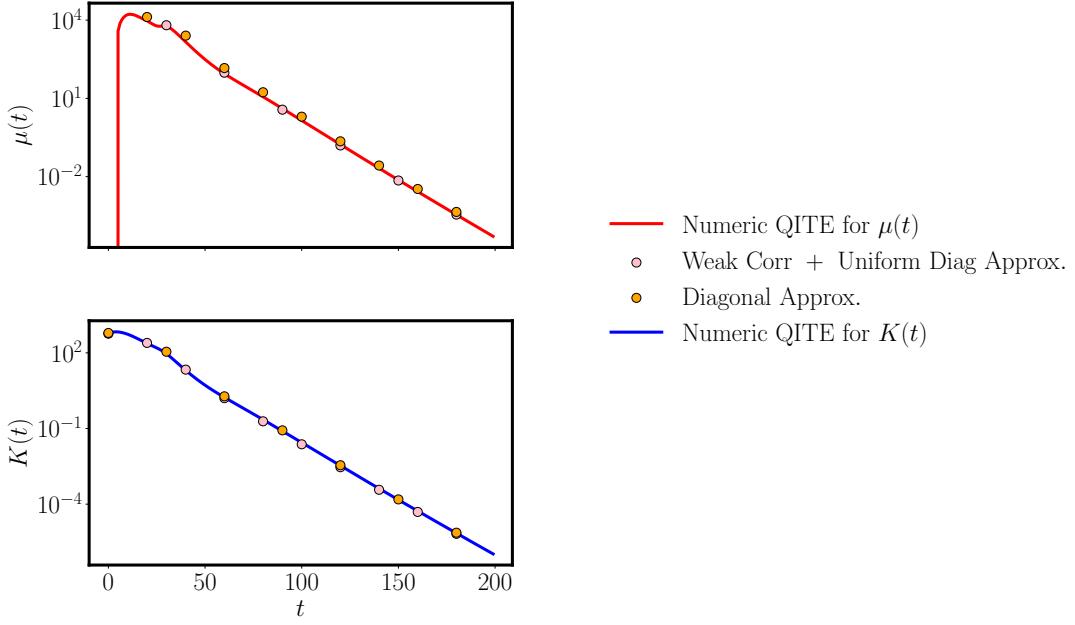


FIG. 5. The diagonal approximation (yellow circles) is highlighted at selected steps to evaluate its agreement with the full K_{QITE} . The agreement confirms that diagonal elements dominate the kernel structure, supporting the validity of the approximation in practice.

Under the Haar random initialization, we have:

$$\begin{aligned} \overline{\frac{\partial \epsilon}{\partial \theta_\ell}} &= 0, \\ \overline{\frac{\partial \epsilon}{\partial \theta_{\ell_3}} \frac{\partial \epsilon}{\partial \theta_{\ell_4}}} &= \delta_{\ell_3 \ell_4} \cdot \overline{\left(\frac{\partial \epsilon}{\partial \theta_\ell} \right)^2}, \\ \overline{\frac{\partial^2 \epsilon}{\partial \theta_{\ell_1} \partial \theta_{\ell_2}}} &= \delta_{\ell_1 \ell_2} \cdot \overline{\frac{\partial^2 \epsilon}{\partial \theta_\ell^2}}. \end{aligned} \quad (143)$$

We similarly apply diagonal approximation to both metric tensors $g_{\ell_1 \ell_3}^+ \approx \delta_{\ell_1 \ell_3} g_{\ell_1 \ell_1}^+$ and $g_{\ell_2 \ell_4}^+ \approx \delta_{\ell_2 \ell_4} g_{\ell_2 \ell_2}^+$, yielding (justified by Fig. 5 that denotes “Diagonal Approx.”):

$$\begin{aligned} \overline{\mu_{\text{QITE}}} &= \overline{\sum_{\ell_1, \ell_2, \ell_3, \ell_4} g_{\ell_1 \ell_3}^+ g_{\ell_2 \ell_4}^+ \frac{\partial^2 \epsilon}{\partial \theta_{\ell_1} \partial \theta_{\ell_2}} \frac{\partial \epsilon}{\partial \theta_{\ell_3}} \frac{\partial \epsilon}{\partial \theta_{\ell_4}}} \\ &\approx \sum_{\ell_1, \ell_2} g_{\ell_1 \ell_1}^+ g_{\ell_2 \ell_2}^+ \overline{\frac{\partial^2 \epsilon}{\partial \theta_{\ell_1} \partial \theta_{\ell_2}} \frac{\partial \epsilon}{\partial \theta_{\ell_1}} \frac{\partial \epsilon}{\partial \theta_{\ell_2}}} \end{aligned} \quad (144)$$

Assuming small fluctuations of $g_{\ell \ell}^+$, we similarly approximate

$$\overline{(g_{\ell \ell}^+)^2} \approx \overline{(g_{\ell \ell}^+)^2} = \left(\frac{N+1}{N} \right)^2 \quad (145)$$

(justified by Fig. 5 that denotes “Weak Corr + Uniform Diag Approx.”), leading to:

$$\overline{\mu_{\text{QITE}}(t)} \approx \left(\frac{N+1}{N} \right)^2 \overline{\sum_{\ell_1, \ell_2} \frac{\partial^2 \epsilon}{\partial \theta_{\ell_1} \partial \theta_{\ell_2}} \frac{\partial \epsilon}{\partial \theta_{\ell_1}} \frac{\partial \epsilon}{\partial \theta_{\ell_2}}}. \quad (146)$$

Meanwhile, similar to *Appendix VIII*, we have (justified by Fig. 5 that denotes “*Diagonal Approx.*” and “*Weak Corr + Uniform Diag Approx.*”):

$$\overline{K_{\text{QITE}}} \approx \frac{N+1}{N} \left(\sum_{\ell} \overline{\frac{\partial \epsilon}{\partial \theta_{\ell}} \frac{\partial \epsilon}{\partial \theta_{\ell}}} \right) \quad (147)$$

Combining numerator and denominator:

$$\overline{\lambda_{\text{QITE}}(t)} = \frac{\overline{\mu_{\text{QITE}}(t)}}{\overline{K_{\text{QITE}}(t)}} \approx \frac{N+1}{N} \overline{\lambda_{\text{GD}}(t)}. \quad (148)$$

□

Dynamics of $K_{\text{QITE}}(t)$ and $\epsilon_{\text{QITE}}(t)$. We assume initial conditions $K_{\text{QITE}}(0) = K_{\text{GD}}(0)$. Given that both GD and QITE satisfy:

$$2\bar{\lambda}\epsilon(t) = K(t) \propto e^{-2\eta\bar{\lambda}t}, \quad (149)$$

For GD-based VQAs, the decay dynamics are:

$$K_{\text{GD}}(t) = K_{\text{GD}}(0)e^{-2\eta\bar{\lambda}_{\text{GD}}t}. \quad (150)$$

For QITE, with decay rate $\bar{\lambda}_{\text{QITE}} = \frac{N+1}{N}\bar{\lambda}_{\text{GD}}$:

$$\begin{aligned} K_{\text{QITE}}(t) &= K_{\text{QITE}}(0)e^{-2\eta\bar{\lambda}_{\text{QITE}}t} \\ &= \frac{K_{\text{GD}}(t)}{e^{-2\eta\bar{\lambda}_{\text{GD}}t}} \cdot e^{-2\eta\frac{N+1}{N}\bar{\lambda}_{\text{GD}}t} \\ &= K_{\text{GD}}(t) \cdot e^{-2\eta\frac{1}{N}\bar{\lambda}_{\text{GD}}t}. \end{aligned} \quad (151)$$

In terms of ϵ_{QITE} , the error for QITE is:

$$\epsilon_{\text{QITE}}(t) = \frac{K_{\text{QITE}}(t)}{2\bar{\lambda}_{\text{QITE}}}. \quad (152)$$

Substitute $\bar{\lambda}_{\text{QITE}} = \frac{N+1}{N}\bar{\lambda}_{\text{GD}}$ and the expression for $K_{\text{QITE}}(t)$:

$$\epsilon_{\text{QITE}}(t) = \frac{N}{2(N+1)\bar{\lambda}_{\text{GD}}} \left[K_{\text{GD}}(t) \cdot e^{-2\eta\frac{1}{N}\bar{\lambda}_{\text{GD}}t} \right]. \quad (153)$$

X. ADDITIONAL NUMERICAL STUDIES

A. Numerical Validation of Eq. (131)

To validate the used approximation

$$\overline{g_{\ell\ell}^+} \approx (\overline{g}^{-1})_{\ell\ell} = \frac{N+1}{N}, \quad (154)$$

we perform a numerical comparison of the trace of both sides across training steps. Here, g^+ denotes g ’s pseudoinverse. The approximation suggests that on average, the diagonal elements of g^+ can be estimated using the diagonal of the inverse of the averaged matrix \overline{g} . Specifically, we compare the trace of the average pseudoinverse, $\text{Tr}[g^+]$, with the trace of the inverse of the average metric tensor, $\text{Tr}[(\overline{g})^{-1}]$. These two quantities are evaluated across training steps and averaged over random initializations. As shown in Fig. 6, the two traces closely match throughout training and align well with the analytic prediction $\text{Tr} \approx \frac{N+1}{N} \cdot L$, thus supporting the validity of the diagonal approximation under Haar-random assumptions.

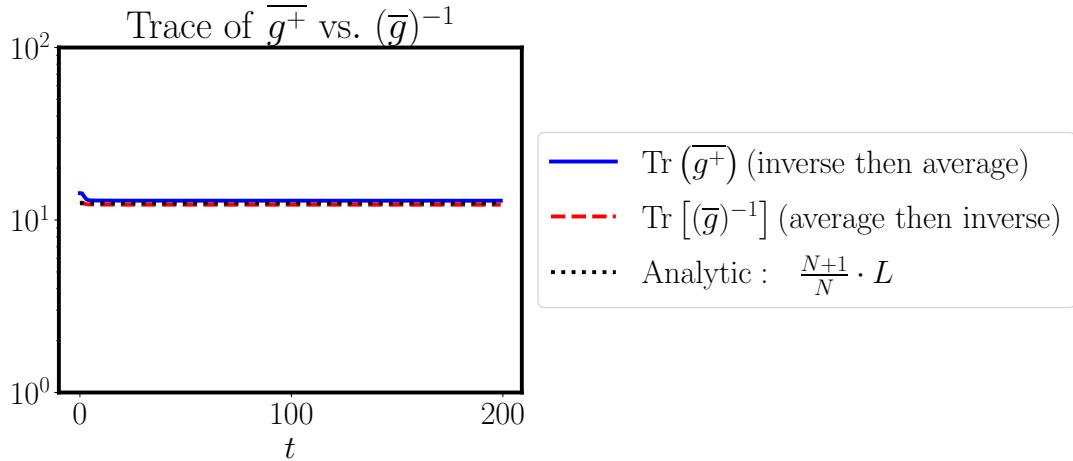


FIG. 6. Comparison between the trace of the empirical average pseudoinverse, $\text{Tr}[\bar{g}^+]$, and the trace of the inverse of the average metric tensor, $\text{Tr}[(\bar{g})^{-1}]$, over training steps. Both quantities are with number of parameters L , and averaged over random initializations. The black dashed line denotes the analytic prediction $\frac{N+1}{N} \cdot L$ under the Haar-random assumption. The numerical agreement supports the validity of the diagonal approximation $\bar{g}_{\ell\ell}^+ \approx (\bar{g}^{-1})_{\ell\ell} \approx \frac{N+1}{N}$.

B. Numerical Studies with Scaling Qubits and Layers

In this section, we examine the scaling behavior of the QNTK K and the relative QNTK λ under both quadratic and linear loss functions. Our analysis considers two complementary aspects: (i) scaling with circuit depth D at fixed qubit number n and ansatz architecture, and (ii) scaling with qubit number n under Haar-random circuit ensembles. For each loss function, we compare closed-form analytic predictions with numerical simulations. The analytic expressions for $K(D)$, $\lambda(D)$, and their n -dependence are derived under the assumption of sufficiently random circuits, modeled as unitary k -designs, yielding explicit scaling laws in the overparameterized regime. To validate these predictions, we sweep both depth and qubit number across broad ranges and extract the empirical scaling from simulations. We further report crossover behavior between shallow- and deep-depth regimes, where saturation of $K(D)$ or changes in the decay rate $\lambda(D)$ may emerge once the circuit approaches the effective mixing depth. The corresponding numerical results are summarized in Fig. 7, which confirm that the observed scaling behaviors are in close agreement with the analytic predictions.

XI. TIME DIFFERENCE EQUATION FOR $K_{\text{QITE}}(t)$

Below we provide the detailed derivation for the time difference equation for $K_{\text{QITE}}(t)$:

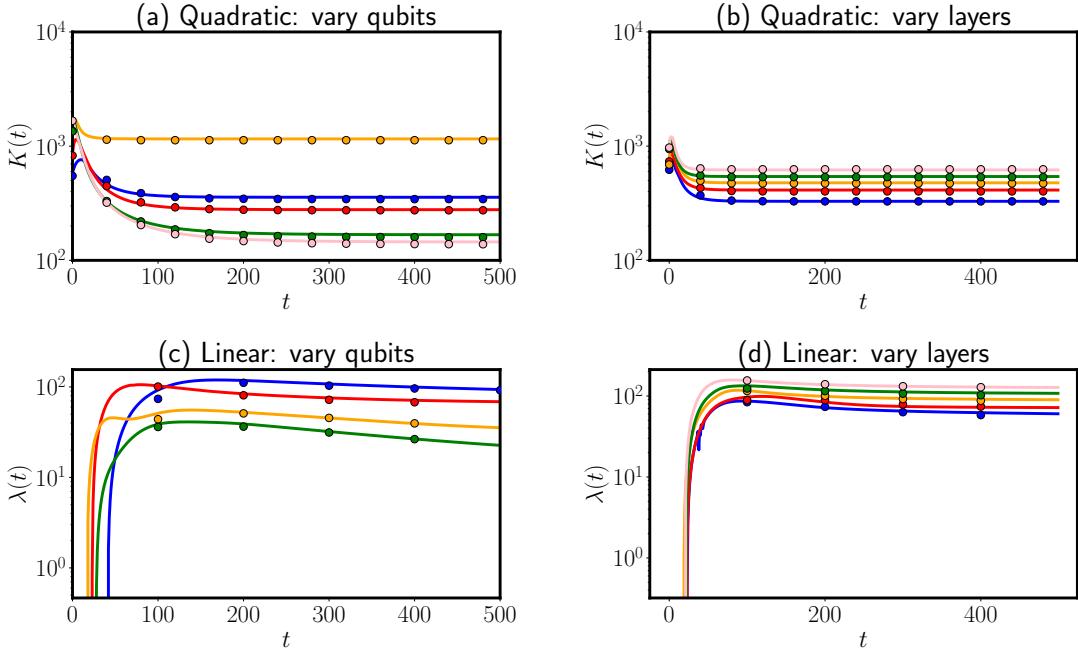


FIG. 7. Scaling behavior of QNTKK(t) and the relative dQNTK $\lambda(t)$ under quadratic and linear loss functions. Panels (a,b) show the scaling with the number of qubits $n = 2, 3, 4, 5, 6$ and the number of layers $D = 6, 7, 8, 9, 10$ for the quadratic loss, while panels (c,d) show the corresponding scaling for the linear loss. Solid lines denote numerical results averaged over 25 independent runs and circles indicate analytical predictions based on Haar random sampling. The color coding corresponds to system size: blue ($n = 2$ qubits or $D = 6$ layers), red ($n = 3$ qubits or $D = 7$ layers), orange ($n = 4$ qubits or $D = 8$ layers), green ($n = 5$ qubits or $D = 9$ layers), and pink ($n = 6$ qubits or $D = 10$ layers). Circles represent the analytic predictions, while solid lines represent the numerical results.

We can ignore each term with two δ in higher orders in η , then we have the final formula for δK as follows:

$$\delta K = \sum_{\ell_1, \ell_2} g_{\ell_1 \ell_2}^+(t) \frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_1}}(t) \delta \left(\frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_2}}(t) \right) + \sum_{\ell_1, \ell_2} \delta \left(g_{\ell_1 \ell_2}^+(t) \right) \frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_1}}(t) \frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_2}}(t) + \sum_{\ell_1, \ell_2} g_{\ell_1 \ell_2}^+(t) \delta \left(\frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_1}}(t) \right) \frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_2}}(t) \quad (156)$$

Because g is a symmetry matrix, we have:

$$\sum_{\ell_1, \ell_2} g_{\ell_1 \ell_2}^+(t) \delta \left(\frac{\partial \epsilon}{\partial \theta_{\ell_1}}(t) \right) \frac{\partial \epsilon}{\partial \theta_{\ell_2}}(t) = \sum_{\ell_2, \ell_1} g_{\ell_1 \ell_2}^+(t) \delta \left(\frac{\partial \epsilon}{\partial \theta_{\ell_2}}(t) \right) \frac{\partial \epsilon}{\partial \theta_{\ell_1}}(t) \quad (157)$$

Thus, δK can be reduced to the following formula:

$$\delta K = 2 \sum_{\ell_1, \ell_2} g_{\ell_1 \ell_2}^+(t) \delta \left(\frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_1}}(t) \right) \frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_2}}(t) + \sum_{\ell_1, \ell_2} \delta(g_{\ell_1 \ell_2}^+(t)) \frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_1}}(t) \frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_2}}(t) \quad (158)$$

With quadratic loss function, we utilize the leading order Taylor expansion on $\delta(\partial\epsilon/\partial\theta_l)$:

$$\delta \left(\frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_1}}(t) \right) = \sum_{\ell_2} \frac{\partial^2 \epsilon}{\partial \boldsymbol{\theta}_{\ell_2} \partial \boldsymbol{\theta}_{\ell_1}} \delta \boldsymbol{\theta}_{\ell_2} + \mathcal{O}(\eta^2) = -\eta \epsilon \sum_{\ell_2 \ell_3} g_{\ell_2 \ell_3}^+ \frac{\partial^2 \epsilon}{\partial \boldsymbol{\theta}_{\ell_2} \partial \boldsymbol{\theta}_{\ell_1}} \frac{\partial \epsilon}{\partial \boldsymbol{\theta}_{\ell_3}} + \mathcal{O}(\eta^2). \quad (159)$$

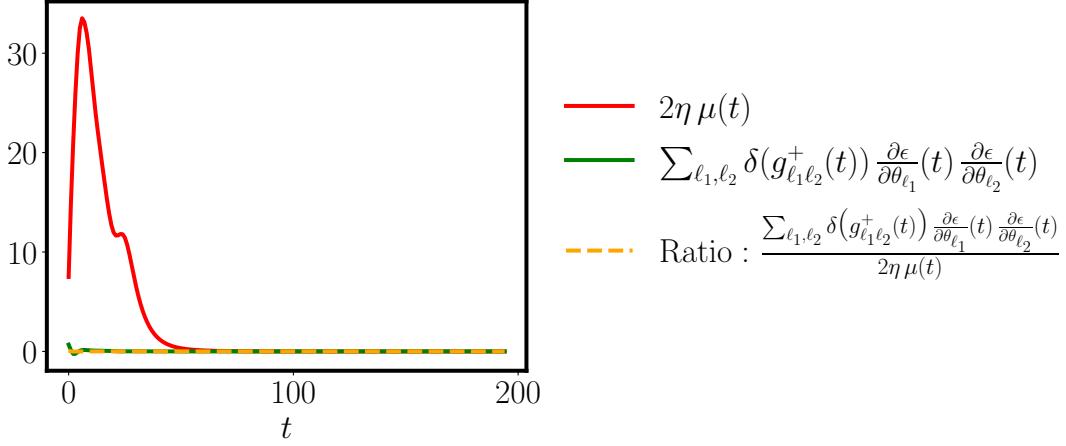


FIG. 8. Numerical verification of the relative importance between the two leading-order terms in the time difference equation of K_{QITE} . The figure plots the ratio $\frac{\sum_{\ell_1, \ell_2} \delta(g_{\ell_1 \ell_2}^+(t)) \frac{\partial \epsilon}{\partial \theta_{\ell_1}}(t) \frac{\partial \epsilon}{\partial \theta_{\ell_2}}(t)}{2\eta \mu(t)}$, confirming that the term $\sum_{\ell_1, \ell_2} \delta(g_{\ell_1 \ell_2}^+(t)) \frac{\partial \epsilon}{\partial \theta_{\ell_1}}(t) \frac{\partial \epsilon}{\partial \theta_{\ell_2}}(t)$ remains significantly smaller throughout training and can be neglected to leading order in η .

Similarly, with linear loss function, the result should be:

$$-\eta \sum_{\ell_2, \ell_3} g_{\ell_2 \ell_3}^+ \frac{\partial^2 \epsilon}{\partial \theta_{\ell_2} \partial \theta_{\ell_3}} \frac{\partial \epsilon}{\partial \theta_{\ell_3}} + \mathcal{O}(\eta^2). \quad (160)$$

Thus, the first term in δK can be rewritten using μ . With quadratic loss function, this is derived as follows:

$$\begin{aligned} 2 \sum_{\ell_1, \ell_2} g_{\ell_1 \ell_2}^+ \delta \left(\frac{\partial \epsilon}{\partial \theta_{\ell_1}}(t) \right) \frac{\partial \epsilon}{\partial \theta_{\ell_2}}(t) &= -2\eta \epsilon \sum_{\ell_1, \ell_2, \ell_3, \ell_4} g_{\ell_1 \ell_2}^+ g_{\ell_3 \ell_4}^+ \frac{\partial^2 \epsilon}{\partial \theta_{\ell_3} \partial \theta_{\ell_1}} \frac{\partial \epsilon}{\partial \theta_{\ell_4}} \frac{\partial \epsilon}{\partial \theta_{\ell_2}} + \mathcal{O}(\eta^2) \\ &= -2\eta \epsilon \mu + \mathcal{O}(\eta^2). \end{aligned} \quad (161)$$

Similarly, with linear loss function:

$$-2\eta \mu + \mathcal{O}(\eta^2). \quad (162)$$

Therefore,

$$\delta K_{\text{QITE}} = \begin{cases} -2\eta \epsilon \mu(t) + \sum_{\ell_1, \ell_2} \delta(g_{\ell_1 \ell_2}^+(t)) \frac{\partial \epsilon}{\partial \theta_{\ell_1}}(t) \frac{\partial \epsilon}{\partial \theta_{\ell_2}}(t) + \mathcal{O}(\eta^2), & (\text{Quadratic loss}) \\ -2\eta \mu(t) + \sum_{\ell_1, \ell_2} \delta(g_{\ell_1 \ell_2}^+(t)) \frac{\partial \epsilon}{\partial \theta_{\ell_1}}(t) \frac{\partial \epsilon}{\partial \theta_{\ell_2}}(t) + \mathcal{O}(\eta^2), & (\text{Linear loss}) \end{cases} \quad (163)$$

Here, we focus on linear loss function because we assume K to be constant at late time with quadratic loss function. We examine if the term $\sum_{\ell_1, \ell_2} \delta(g_{\ell_1 \ell_2}^+(t)) \frac{\partial \epsilon}{\partial \theta_{\ell_1}}(t) \frac{\partial \epsilon}{\partial \theta_{\ell_2}}(t)$ can be ignored. This is verified through a numerical study by calculating the ratio of $2\eta \mu$ and $\sum_{\ell_1, \ell_2} \delta(g_{\ell_1 \ell_2}^+(t)) \frac{\partial \epsilon}{\partial \theta_{\ell_1}}(t) \frac{\partial \epsilon}{\partial \theta_{\ell_2}}(t)$. See Fig. 8 for the details.

By examining the numerical study, we make the simplification below:

$$\delta K_{\text{QITE}} = -2\eta \mu + \mathcal{O}(\eta^2) \quad (164)$$

XII. RESULTS WITH HAAR RANDOM ENSEMBLE

As shown in the main text, the parameterized unitary $U(\boldsymbol{\theta})$ we consider is defined below:

$$U(\boldsymbol{\theta}) = \prod_{k=1}^L W_k V_k(\boldsymbol{\theta}_k) \quad (165)$$

We follow the notations in Zhang *et al.* [31]. Notably, we omit the constant factor $1/2$ in the exponent and instead define $\hat{V}_\ell(\boldsymbol{\theta}_\ell) = e^{-i\boldsymbol{\theta}_\ell \hat{X}_\ell}$ as below:

$$\hat{U}(\boldsymbol{\theta}) = \prod_{\ell=1}^L \hat{W}_\ell \hat{V}_\ell(\boldsymbol{\theta}_\ell), \quad (166)$$

with:

$$\hat{V}_\ell(\boldsymbol{\theta}_\ell) = e^{-i\boldsymbol{\theta}_\ell \hat{X}_\ell} \quad (167)$$

This choice does not affect the generality of our results, as it amounts to a simple rescaling of the parameter range. For ℓ -th parameter and specific ℓ_1, ℓ_2 -th parameters, we split $U(\boldsymbol{\theta})$:

$$U(\boldsymbol{\theta}) = U_{\ell^-} U_{\ell^+} = U_{\ell_1^-} U_{\ell_1 \rightarrow \ell_2} U_{\ell_2^+}, \quad (168)$$

where we define:

$$\begin{aligned} U_{\ell^-} &= \prod_{k=1}^{\ell-1} W_k V_k(\boldsymbol{\theta}_k) \\ U_{\ell^+} &= \prod_{k=\ell}^L W_k V_k(\boldsymbol{\theta}_k) \\ U_{\ell_1 \rightarrow \ell_2} &= \prod_{k=\ell_1}^{\ell_2-1} W_k V_k(\boldsymbol{\theta}_k) \end{aligned}$$

Indicating:

$$U_{\ell_1^-} U_{\ell_1^+} = U_{\ell_2^-} U_{\ell_2^+} = U_{\ell_1^-} U_{\ell_1 \rightarrow \ell_2} U_{\ell_2^+} \quad (169)$$

A. Average Fubini-Study Metric Tensor g Result under Haar Random Ensemble

The variational output state $|\psi(\boldsymbol{\theta})\rangle$ can be given as:

$$|\psi(\boldsymbol{\theta})\rangle = U_{\ell^+} U_{\ell^-} |\psi_0\rangle \quad (170)$$

For calculating g , we firstly calculate the quantum geometric tensor G :

$$G_{\ell_1 \ell_2}(\boldsymbol{\theta}) = \left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_1}} \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_2}} \right\rangle - \left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_1}} \middle| \psi(\boldsymbol{\theta}) \right\rangle \left\langle \psi(\boldsymbol{\theta}) \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_2}} \right\rangle. \quad (171)$$

Below we show the Haar ensemble average results for $\left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_1}} \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_2}} \right\rangle$ and $\left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_1}} \middle| \psi(\boldsymbol{\theta}) \right\rangle \left\langle \psi(\boldsymbol{\theta}) \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_2}} \right\rangle$.

We can categorize the calculations into the diagonal case and off-diagonal case for $G_{\ell_1 \ell_2}$.

Firstly, we focus on the diagonal case, where $\ell = \ell_1 = \ell_2$. For this, the analytic formula for $\left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_\ell} \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_\ell} \right\rangle$ as follows:

$$\begin{aligned} \left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_\ell} \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_\ell} \right\rangle &= iU_{\ell+}X_\ell U_{\ell-}|\psi_0\rangle \\ \overline{\left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_\ell} \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_\ell} \right\rangle} &= 1 \end{aligned} \quad (172)$$

For the calculation of $\left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_\ell} \middle| \psi(\boldsymbol{\theta}) \right\rangle \left\langle \psi(\boldsymbol{\theta}) \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_\ell} \right\rangle$:

$$\begin{aligned} \overline{\left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_\ell} \middle| \psi(\boldsymbol{\theta}) \right\rangle} \cdot \overline{\left\langle \psi(\boldsymbol{\theta}) \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_\ell} \right\rangle} &= \int dU_{\ell-}dU_{\ell+}(-i)\langle\psi_0|U_{\ell-}^\dagger X_\ell U_{\ell+}^\dagger U_{\ell+}U_{\ell-}|\psi_0\rangle(i)\langle\psi_0|U_{\ell-}U_{\ell+}U_{\ell+}^\dagger X_\ell U_{\ell+}|\psi_0\rangle \\ &= \int dU_{\ell-}(-i)\langle\psi_0|U_{\ell-}^\dagger X_\ell U_{\ell-}|\psi_0\rangle(i)\langle\psi_0|U_{\ell-}^\dagger X_\ell U_{\ell-}|\psi_0\rangle \\ &= \frac{1}{N+1}, \end{aligned} \quad (173)$$

where we leverage RTNI package [62] to calculate the last equation. Therefore

$$\overline{G_{\ell\ell}} = \overline{\left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_\ell} \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_\ell} \right\rangle} - \overline{\left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_\ell} \middle| \psi(\boldsymbol{\theta}) \right\rangle} \cdot \overline{\left\langle \psi(\boldsymbol{\theta}) \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_\ell} \right\rangle} = 1 - \frac{1}{N+1} \quad (174)$$

For off-diagonal case where $\ell_1 \neq \ell_2$. Because $U(\boldsymbol{\theta}) = U_{\ell_1^-}U_{\ell_1^+} = U_{\ell_1^-}U_{\ell_1 \rightarrow \ell_2}U_{\ell_2^+}$, we have:

$$\begin{aligned} U_{\ell_1^+} &= U_{\ell_1 \rightarrow \ell_2}U_{\ell_2^+} \\ U_{\ell_1^+}^\dagger &= U_{\ell_2^+}^\dagger U_{\ell_1 \rightarrow \ell_2}^\dagger \end{aligned} \quad (175)$$

Accordingly, we can calculate the analytic formula for $\left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_1}} \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_2}} \right\rangle$ and $\left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_1}} \middle| \psi(\boldsymbol{\theta}) \right\rangle \left\langle \psi(\boldsymbol{\theta}) \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_2}} \right\rangle$ respectively, where

$$\begin{aligned} \left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_1}} \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_2}} \right\rangle &= \langle\psi_0|U_{\ell_1^-}^\dagger XU_{\ell_1^+}^\dagger U_{\ell_2^+} XU_{\ell_2^-}|\psi_0\rangle \\ &= \langle\psi_0|U_{\ell_1^-}^\dagger XU_{\ell_1 \rightarrow \ell_2}^\dagger U_{\ell_2^+}^\dagger U_{\ell_2^+} XU_{\ell_2^-}|\psi_0\rangle \\ &= \langle\psi_0|U_{\ell_1^-}^\dagger XU_{\ell_1 \rightarrow \ell_2}^\dagger XU_{\ell_2^-}|\psi_0\rangle \\ &= \text{Tr}\left\{\rho_0 U_{\ell_1^-}^\dagger XU_{\ell_1 \rightarrow \ell_2}^\dagger XU_{\ell_2^-}\right\} \end{aligned} \quad (176)$$

Therefore, the result for $\left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_1}} \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_2}} \right\rangle$ is as follows:

$$\begin{aligned} \overline{\left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_1}} \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_2}} \right\rangle} &= \int dU_{\ell_1^-}U_{\ell_1 \rightarrow \ell_2}dU_{\ell_2^-}\langle\psi_0|U_{\ell_1^-}^\dagger XU_{\ell_1 \rightarrow \ell_2}^\dagger XU_{\ell_2^-}|\psi_0\rangle \\ &= \int dU_{\ell_1^-}U_{\ell_1 \rightarrow \ell_2}dU_{\ell_2^-} \text{Tr}\left\{\rho_0 U_{\ell_1^-}^\dagger XU_{\ell_1 \rightarrow \ell_2}^\dagger XU_{\ell_2^-}\right\} \\ &= 0 \end{aligned} \quad (177)$$

Similarly, we can derive the result for $\left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_1}} \middle| \psi(\boldsymbol{\theta}) \right\rangle \cdot \left\langle \psi(\boldsymbol{\theta}) \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_2}} \right\rangle$ as below:

$$\begin{aligned}
& \left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta_{\ell_1}} \middle| \psi(\boldsymbol{\theta}) \right\rangle \cdot \left\langle \psi(\boldsymbol{\theta}) \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta_{\ell_2}} \right\rangle = \int dU_{\ell_1^-} dU_{\ell_2^+} (-i) \langle \psi_0 | U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^+}^\dagger U_{\ell_2^+} U_{\ell_1^-} | \psi_0 \rangle (i) \langle \psi_0 | U_{\ell_2^-}^\dagger U_{\ell_2^+}^\dagger U_{\ell_2^+} X_{\ell_2} U_{\ell_2^-} | \psi_0 \rangle \\
&= \int dU_{\ell_1^-} dU_{\ell_2^+} \langle \psi_0 | U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^-} | \psi_0 \rangle \langle \psi_0 | U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^-} | \psi_0 \rangle \\
&= \int dU_{\ell_1^-} dU_{\ell_2^+} \text{Tr} \left\{ \rho_0 U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^-} \rho_0 U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^-} | \psi_0 \rangle \right\} \\
&= 0
\end{aligned} \tag{178}$$

Therefore, we can have a general formula for $\overline{G_{\ell_1 \ell_2}}$:

$$\overline{G_{\ell_1 \ell_2}} = (1 - \frac{1}{N+1}) \delta_{\ell_1 \ell_2}, \quad (179)$$

where $\delta_{\ell_1 \ell_2}$ represents the Kronecker delta defined as:

$$\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases} \quad (180)$$

B. Fluctuations of Fubini-Study Metric Tensor g Under Haar Random Ensemble

In this section, we analyze the fluctuations in g around \bar{g} , i.e., $\Delta g_{\ell_1 \ell_2}^2 = \mathbb{E}(g_{\ell_1 \ell_2}^2) - \overline{g_{\ell_1 \ell_2}}^2 = \overline{g_{\ell_1 \ell_2}^2} - \overline{g_{\ell_1 \ell_2}}^2$. For the off-diagonal case where $\ell_1 \neq \ell_2$, we already have $\overline{g_{\ell_1 \ell_2}}^2 = 0$. According to:

$$\left| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta_\ell} \right\rangle = i U_{\ell^+} X_\ell U_{\ell^-} |\psi_0\rangle \quad (181)$$

For $G_{\ell_1 \ell_2}$:

$$\begin{aligned}
G_{\ell_1 \ell_2}(\boldsymbol{\theta}) &= \left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_1}} \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_2}} \right\rangle - \left\langle \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_1}} \middle| \psi(\boldsymbol{\theta}) \right\rangle \left\langle \psi(\boldsymbol{\theta}) \middle| \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{\ell_2}} \right\rangle \\
&= \left\langle \psi_0 \middle| U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^+}^\dagger U_{\ell_2^+} X_{\ell_2} U_{\ell_2^-} \middle| \psi_0 \right\rangle \\
&\quad - \left\langle \psi_0 \middle| U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^+}^\dagger U_{\ell_1^+} U_{\ell_1^-} \middle| \psi_0 \right\rangle \left\langle \psi_0 \middle| U_{\ell_2^-}^\dagger U_{\ell_2^+}^\dagger U_{\ell_2^+} X_{\ell_2} U_{\ell_2^-} \middle| \psi_0 \right\rangle \tag{182} \\
&= \left\langle \psi_0 \middle| U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^+}^\dagger U_{\ell_2^+} X_{\ell_2} U_{\ell_2^-} \middle| \psi_0 \right\rangle \\
&\quad - \left\langle \psi_0 \middle| U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^-} \middle| \psi_0 \right\rangle \left\langle \psi_0 \middle| U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^-} \middle| \psi_0 \right\rangle
\end{aligned}$$

Therefore,

$$G_{\ell_1 \ell_2}^*(\boldsymbol{\theta}) = \left\langle \psi_0 \middle| U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^+}^\dagger U_{\ell_1^+} X_{\ell_1} U_{\ell_1^-} \right| \psi_0 \right\rangle - \left\langle \psi_0 \middle| U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^-} \right| \psi_0 \right\rangle \left\langle \psi_0 \middle| U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^-} \right| \psi_0 \right\rangle. \quad (183)$$

For $g = \operatorname{Re}\{G\} = \frac{G+G^*}{2}$, we have:

$$\begin{aligned}
& \overline{\left\langle \psi_0 \middle| U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^+}^\dagger U_{\ell_2^+} X_{\ell_2} U_{\ell_2^-} \middle| \psi_0 \right\rangle} \overline{\left\langle \psi_0 \middle| U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^+}^\dagger U_{\ell_1^+} X_{\ell_1} U_{\ell_1^-} \middle| \psi_0 \right\rangle} \\
&= \int dU_{\ell_1^-} dU_{\ell_1^+} dU_{\ell_2^-} dU_{\ell_2^+} \text{Tr} \left\{ \rho_0 U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^+}^\dagger U_{\ell_2^+} X_{\ell_2} U_{\ell_2^-} \rho_0 U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^+}^\dagger U_{\ell_1^+} X_{\ell_1} U_{\ell_1^-} \right\} \\
&= \frac{1}{N}
\end{aligned} \tag{188}$$

For $\overline{\left\langle \psi_0 \middle| U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^+}^\dagger U_{\ell_2^+} X_{\ell_2} U_{\ell_2^-} \middle| \psi_0 \right\rangle} \overline{\left\langle \psi_0 \middle| U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^-} \middle| \psi_0 \right\rangle} \overline{\left\langle \psi_0 \middle| U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^-} \middle| \psi_0 \right\rangle}$,

$$\begin{aligned}
& \overline{\left\langle \psi_0 \middle| U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^+}^\dagger U_{\ell_2^+} X_{\ell_2} U_{\ell_2^-} \middle| \psi_0 \right\rangle} \overline{\left\langle \psi_0 \middle| U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^-} \middle| \psi_0 \right\rangle} \overline{\left\langle \psi_0 \middle| U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^-} \middle| \psi_0 \right\rangle} \\
&= \int dU_{\ell_1^-} dU_{\ell_1^+} dU_{\ell_2^-} dU_{\ell_2^+} \text{Tr} \left\{ \rho_0 U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^+}^\dagger U_{\ell_2^+} X_{\ell_2} U_{\ell_2^-} \rho_0 U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^-} \rho_0 U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^-} \right\} \\
&= 0.
\end{aligned} \tag{189}$$

For $\overline{\left\langle \psi_0 \middle| U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^+}^\dagger U_{\ell_1^+} X_{\ell_1} U_{\ell_1^-} \middle| \psi_0 \right\rangle} \overline{\left\langle \psi_0 \middle| U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^-} \middle| \psi_0 \right\rangle} \overline{\left\langle \psi_0 \middle| U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^-} \middle| \psi_0 \right\rangle}$,

$$\begin{aligned}
& \overline{\left\langle \psi_0 \middle| U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^+}^\dagger U_{\ell_1^+} X_{\ell_1} U_{\ell_1^-} \middle| \psi_0 \right\rangle} \overline{\left\langle \psi_0 \middle| U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^-} \middle| \psi_0 \right\rangle} \overline{\left\langle \psi_0 \middle| U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^-} \middle| \psi_0 \right\rangle} \\
&= \int dU_{\ell_1^-} dU_{\ell_1^+} dU_{\ell_2^-} dU_{\ell_2^+} \text{Tr} \left\{ \rho_0 U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^+}^\dagger U_{\ell_1^+} X_{\ell_1} U_{\ell_1^-} \rho_0 U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^-} \rho_0 U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^-} \right\} \\
&= 0.
\end{aligned} \tag{190}$$

For $\overline{\left\langle \psi_0 \middle| U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^-} \middle| \psi_0 \right\rangle}^2 \overline{\left\langle \psi_0 \middle| U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^-} \middle| \psi_0 \right\rangle}^2$,

$$\begin{aligned}
& \overline{\left\langle \psi_0 \middle| U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^-} \middle| \psi_0 \right\rangle}^2 \overline{\left\langle \psi_0 \middle| U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^-} \middle| \psi_0 \right\rangle}^2 \\
&= \int dU_{\ell_1^-} dU_{\ell_2^-} \text{Tr} \left\{ \rho_0 U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^-} \rho_0 U_{\ell_1^-}^\dagger X_{\ell_1} U_{\ell_1^-} \rho_0 U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^-} \rho_0 U_{\ell_2^-}^\dagger X_{\ell_2} U_{\ell_2^-} \right\} \\
&= \frac{1}{(N+1)^2}.
\end{aligned} \tag{191}$$

Therefore,

$$\begin{aligned}
\overline{g_{\ell_1 \ell_2}^2(\theta)} &= \frac{1}{2N} + \frac{1}{(N+1)^2}, \\
\Delta g_{\ell_1 \ell_2}^2 &= \mathbb{E}(g_{\ell_1 \ell_2}^2) - \overline{g_{\ell_1 \ell_2}^2} = \overline{g_{\ell_1 \ell_2}^2} - \overline{g_{\ell_1 \ell_2}^2} = \frac{1}{2N} + \frac{1}{(N+1)^2}.
\end{aligned} \tag{192}$$

In large-N limit,

$$\Delta g_{\ell_1 \ell_2}^2 \approx \frac{1}{2N} \tag{193}$$

When $\ell_1 = \ell_2$,

$$\begin{aligned}
& \left[g_{\ell_1 \ell_2}^2(\boldsymbol{\theta}) \right]_{\ell_1 = \ell_2 = \ell} = g_{\ell \ell}^2(\boldsymbol{\theta}) \\
&= \frac{1}{4} \left\langle \psi_0 \left| U_{\ell^-}^\dagger X_\ell U_{\ell^+}^\dagger U_{\ell^+} X_\ell U_{\ell^-} \right| \psi_0 \right\rangle^2 + \frac{1}{4} \left\langle \psi_0 \left| U_{\ell^-}^\dagger X_\ell U_{\ell^+}^\dagger U_{\ell^+} X_\ell U_{\ell^-} \right| \psi_0 \right\rangle^2 \\
&\quad + \frac{1}{2} \left\langle \psi_0 \left| U_{\ell^-}^\dagger X_\ell U_{\ell^+}^\dagger U_{\ell^+} X_\ell U_{\ell^-} \right| \psi_0 \right\rangle \left\langle \psi_0 \left| U_{\ell^-}^\dagger X_\ell U_{\ell^+}^\dagger U_{\ell^+} X_\ell U_{\ell^-} \right| \psi_0 \right\rangle \\
&\quad - \left\langle \psi_0 \left| U_{\ell^-}^\dagger X_\ell U_{\ell^+}^\dagger U_{\ell^+} X_\ell U_{\ell^-} \right| \psi_0 \right\rangle \left\langle \psi_0 \left| U_{\ell^-}^\dagger X_\ell U_{\ell^-} \right| \psi_0 \right\rangle^2 \\
&\quad - \left\langle \psi_0 \left| U_{\ell^-}^\dagger X_\ell U_{\ell^+}^\dagger U_{\ell^+} X_\ell U_{\ell^-} \right| \psi_0 \right\rangle \left\langle \psi_0 \left| U_{\ell^-}^\dagger X_\ell U_{\ell^-} \right| \psi_0 \right\rangle^2 \\
&\quad + \left\langle \psi_0 \left| U_{\ell^-}^\dagger X_\ell U_{\ell^-} \right| \psi_0 \right\rangle^4 \\
&= \frac{1}{4} \left\langle \psi_0 \left| U_{\ell^-}^\dagger X_\ell^2 U_{\ell^-} \right| \psi_0 \right\rangle^2 + \frac{1}{4} \left\langle \psi_0 \left| U_{\ell^-}^\dagger X_\ell^2 U_{\ell^-} \right| \psi_0 \right\rangle^2 \\
&\quad + \frac{1}{2} \left\langle \psi_0 \left| U_{\ell^-}^\dagger X_\ell^2 U_{\ell^-} \right| \psi_0 \right\rangle^2 - 2 \left\langle \psi_0 \left| U_{\ell^-}^\dagger X_\ell^2 U_{\ell^-} \right| \psi_0 \right\rangle \left\langle \psi_0 \left| U_{\ell^-}^\dagger X_\ell U_{\ell^-} \right| \psi_0 \right\rangle^2 \\
&\quad + \left\langle \psi_0 \left| U_{\ell^-}^\dagger X_\ell U_{\ell^-} \right| \psi_0 \right\rangle^4 \\
&= 1 - 2 \left\langle \psi_0 \left| U_{\ell^-}^\dagger X_\ell U_{\ell^-} \right| \psi_0 \right\rangle^2 + \left\langle \psi_0 \left| U_{\ell^-}^\dagger X_\ell U_{\ell^-} \right| \psi_0 \right\rangle^4
\end{aligned} \tag{194}$$

For $\overline{\langle \psi_0 | U_{\ell^-}^\dagger X_\ell U_{\ell^-} | \psi_0 \rangle^2}$,

$$\begin{aligned}
& \overline{\langle \psi_0 | U_{\ell^-}^\dagger X_\ell U_{\ell^-} | \psi_0 \rangle^2} \\
&= \int dU_{\ell^-} \left[\text{Tr} \left\{ \rho_0 U_{\ell^-}^\dagger X_\ell U_{\ell^-} \right\} \right]^2 \\
&= \frac{1}{N+1}.
\end{aligned} \tag{195}$$

For $\overline{\langle \psi_0 | U_{\ell^-}^\dagger X_\ell U_{\ell^-} | \psi_0 \rangle^4}$, we adopt Wick contraction technique as it provides an efficient method for computing high-order moments of Haar-random unitaries in the large- N limit, where matrix elements become approximately Gaussian [63]. This makes the $2k$ -point correlators decompose into $(2k-1)!!$ pairwise contractions, each contributing $1/N$, simplifying calculations compared to exact Weingarten methods [38].

$$\begin{aligned}
& \overline{\langle \psi_0 | U_{\ell-}^\dagger X_\ell U_{\ell-} | \psi_0 \rangle}^4 = \int dU_{\ell-} \left(\langle \psi_0 | U_{\ell-}^\dagger X_\ell U_{\ell-} | \psi_0 \rangle \right)^4 \\
& = \int dU_{\ell-} \prod_{a=1}^4 \langle \psi_0 | U_{\ell-}^\dagger X_\ell U_{\ell-} | \psi_0 \rangle \\
& = \int dU_{\ell-} \langle \psi_0 | U_{\ell-}^\dagger X_\ell U_{\ell-} | \psi_0 \rangle \langle \psi_0 | U_{\ell-}^\dagger X_\ell U_{\ell-} | \psi_0 \rangle \left(\langle \psi_0 | U_{\ell-}^\dagger X_\ell U_{\ell-} | \psi_0 \rangle \right)^2 \\
& = 3 \cdot \frac{1}{N} \int dU_{\ell-} \langle \psi_0 | \psi_0 \rangle \langle \psi_0 | U_{\ell-}^\dagger X_\ell U_{\ell-} | \psi_0 \rangle \left(\langle \psi_0 | U_{\ell-}^\dagger X_\ell U_{\ell-} | \psi_0 \rangle \right)^2 \\
& = 3 \cdot \frac{1}{N} \int dU_{\ell-} 1 \cdot \langle \psi_0 | U_{\ell-}^\dagger I U_{\ell-} | \psi_0 \rangle \left(\langle \psi_0 | U_{\ell-}^\dagger X_\ell U_{\ell-} | \psi_0 \rangle \right)^2 \\
& = 3 \cdot \frac{1}{N} \int dU_{\ell-} \langle \psi_0 | \psi_0 \rangle \left(\langle \psi_0 | U_{\ell-}^\dagger X_\ell U_{\ell-} | \psi_0 \rangle \right)^2 \\
& = 3 \cdot \frac{1}{N} \int dU_{\ell-} \left(\langle \psi_0 | U_{\ell-}^\dagger X_\ell U_{\ell-} | \psi_0 \rangle \right)^2 \\
& = 3 \cdot \frac{1}{N} \cdot \frac{1}{N+1} \\
& = \frac{3}{N(N+1)}. \tag{196}
\end{aligned}$$

Therefore,

$$\overline{g_{\ell\ell}^2(\theta)} = 1 - \frac{2}{N+1} + \frac{3}{N(N+1)} \tag{197}$$

Given that $\overline{g_{\ell\ell}^2(\theta)}^2 = (1 - \frac{1}{N+1})^2$, we have $\Delta g_{\ell\ell}^2$:

$$\begin{aligned}
\Delta g_{\ell\ell}^2 &= 1 - \frac{2}{N+1} + \frac{3}{N(N+1)} - \frac{N^2}{(N+1)^2} \\
&= \frac{2N+3}{N^3+2N^2+N} \tag{198}
\end{aligned}$$

In the large-N limit, we get:

$$\Delta g_{\ell\ell}^2 \approx \frac{2}{N^2} \tag{199}$$

In sum:

$$\Delta g_{\ell_1 \ell_2}^2 \approx \begin{cases} \frac{2}{N^2}, & \text{if } \ell_1 = \ell_2 = \ell, \\ \frac{1}{2N}, & \text{if } \ell_1 \neq \ell_2. \end{cases} \tag{200}$$

Therefore, the fluctuation of the elements in g is negligible when $N \rightarrow \infty$.