RIEMANNIAN OPTIMIZATION OF VARIATIONAL QUANTUM ALGORITHMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantum Natural Gradient (QNG) is widely considered the theoretically optimal optimizer for variational quantum circuits. However, QNG can become unstable as variational quantum circuits become deeper and noisier, yet the geometric origin of this instability remains poorly understood. We uncover a noise-induced Riemannian spike (NIRS): under local Markovian noise, the pullback metric can vanish exponentially faster than the Euclidean gradient, causing metric inversion to amplify the natural-gradient update exponentially with circuit depth. This reveals a previously overlooked failure mechanism in which noise can destabilize optimization through metric inversion, rather than through gradient suppression alone. In parallel, we introduce Dually Flat Stratified Quantum Natural Gradient (DFS-QNG), a structured Riemannian optimizer that exploits commuting parameter strata. By restricting each update to a commuting submanifold, DFS-QNG obtains exact blockwise geodesic updates while replacing costly global metric inversion with smaller blockwise systems. We further establish geodesic smoothness and convergence guarantees for the resulting Riemannian optimization framework. Experiments on noisy and noiseless VQE and QAOA circuits demonstrate improved stability and faster convergence than standard QNG and Adam, together with substantially reduced metric computation.

1 INTRODUCTION

Variational quantum algorithms is one of the most important near-term application of quantum hardware Peruzzo et al. (2014); Farhi et al. (2014); Cerezo et al. (2021). A parametrized quantum circuit prepares a trial state, while a classical optimizer updates the parameters to minimize an objective. The parameter update is classical, the state preparation is quantum. However, updating these parameters does not follow a path in flat Euclidean space $\mathbb{R}^P$. The ansatz maps parameters to quantum states on an immersed submanifold of projective Hilbert space $\mathbb{CP}^{d-1}$, inheriting a Riemannian geometry from the Fubini–Study metric. If the space of states is curved, standard Euclidean gradient descent does not reflect the natural geometry of the problem.

The Quantum Natural Gradient (QNG) Stokes et al. (2020) was introduced to resolve this. Drawing on Amari's information geometry of classical statistical models Amari (1998), QNG replaces the Euclidean inner product with the Fubini–Study metric pulled back through the variational ansatz, and computes the steepest-descent direction on the resulting quantum statistical manifold. This approach possesses several desirable geometric properties: QNG is invariant under circuit reparametrization, information-geometrically accurate, and widely considered as the "theoretically optimal" optimizer for variational quantum algorithms Yamamoto (2019); Wierichs et al. (2020); Koczor & Benjamin (2022); Cerezo et al. (2021). Empirically, however, the picture is more fractured. QNG reliably outperforms vanilla gradient descent on ansätze of moderate depth Wierichs et al. (2020); Yamamoto (2019); Wang et al. (2023), but on comprehensive benchmarks spanning diverse VQE and QML problem classes, first-order adaptive Euclidean methods, such as Adam Kingma & Ba (2014), frequently match or surpass QNG in both convergence efficiency and solution quality Lockwood (2022); Jones et al. (2025); Vu et al. (2025); Spall (1998). More troubling is a failure mode that emerges at scale: in deep circuits, QNG becomes sensitive to the step size, exhibiting oscillatory or divergent behavior even with learning rates well within the stable regime of first-order methods. Researchers have responded by defaulting to Euclidean optimizers, inflating the Tikhonov regular-

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

ization ϵ of the metric inverse, or simply restricting circuit depth Fitzek et al. (2024); Halla (2025b). While it is widely known that the metric becomes ill-conditioned in deeper circuits, there is still a lack of understanding as to why and when this ill-conditioning emerges, and why Riemannian updates become unstable in practice.

Furthermore, the geometric properties of the quantum statistical manifold itself remain poorly understood. Much work in Riemannian optimization typically relies on well-studied, highly symmetric spaces (such as Stiefel or Grassmann manifolds) where convergence is well established Edelman et al. (1998). In contrast, the manifold formed by a quantum circuit is shaped by layers of non-commuting exponentials, making its geometry more complex. As a result, prior work has lacked formal analysis on the convergence behavior of Riemannian methods on variational quantum landscapes.

Contributions.

In this paper, we make four contributions.

First, we establish a convergence theory for Riemannian optimization on variational quantum landscapes. We prove that the variational quantum manifold satisfies L_{R}-geodesic smoothness (Theorem 1), yielding convergence rate of Riemannian gradient descent to a δ-stationary point at the standard $\mathcal{O}(\mathsf{L}_{\mathrm{R}}/\delta)$ rate (Corollary 1).

Second, we identify and characterize a noise-induced Riemannian spike (NIRS). Under local Markovian noise, we prove that the pullback metric and Euclidean gradient contract at different exponential rates (Theorem 2). Consequently, inversion of the noisy metric can exponentially amplify the natural-gradient update with circuit depth. This establishes a distinct noise-induced failure mechanism for Riemannian optimization that is not captured by gradient suppression alone.

Third, we introduce Dually Flat Stratified QNG (DFS-QNG), a geometry-aware block-coordinate formulation of natural-gradient optimization. We partition circuit parameters into mutually commuting strata and prove that, under the corresponding commutativity condition (Eq. 11), each restricted submanifold has constant pullback geometry and vanishing connection coefficients (Proposition 1). Consequently, the Riemannian exponential map reduces exactly to an Euler step within each stratum. DFS-QNG therefore converts the global QNG procedure into exact Riemannian block-coordinate descent, while reducing metric computation and inversion from global P-dimensional systems to blockwise systems with complexity $\mathcal{O}(\sum_{\pi=1}^{S} l_\pi^2)$. We establish an $\mathcal{O}(\mathsf{L}_{\mathrm{R}}/\delta)$ convergence rate for the resulting method (Theorem 3).

Fourth, we connect the theoretical failure mechanisms to empirical optimization behavior. Across VQE and QAOA benchmarks, we evaluate QNG and DFS-QNG over varying circuit depths and noise conditions, measuring optimization error, convergence stability, and metric-computation cost. Our experiments demonstrate the predicted depth- and noise-dependent behavior of the Riemannian gradient and show that exploiting commuting local geometry can substantially improve optimization stability and reduce the computational burden of natural-gradient optimization.

2 Geometric Setup for Variational Quantum Optimization

2.1 The Parameterized State Space and Pullback Geometry

Let $\mathcal{H} \cong \mathbb{C}^d$ denote the Hilbert space of n qubits with dimension $d = 2^n$. Physical quantum states correspond to rays in projective space $\mathbb{CP}^{d-1}$ because global phase is unobservable.

Definition 1 (Variational Ansatz)**.** *Let $\Theta \subseteq \mathbb{R}^P$ be the parameter space. A variational ansatz prepares a trial state by applying L parameterized layers of unitaries to a reference state $|\psi_0\rangle \in \mathcal{H}$:*

$$\mathcal{U}(\boldsymbol{\theta}) = \prod_{l=1}^{L} \mathcal{U}_l(\boldsymbol{\theta}_l), \quad \mathcal{U}_l(\boldsymbol{\theta}_l) = \prod_{m=1}^{M_l} \mathcal{W}_{(l,m)}, \exp\big(-i\boldsymbol{\theta}_{(l,m)}\mathfrak{g}_{(l,m)}\big) \tag{1}$$

where $\mathcal{W}_{(l,m)}$ are fixed unitaries, $\mathfrak{g}_{(l,m)} \in \mathfrak{su}(d)$ are Hermitian generators, and $P = \sum_{l=1}^{L} M_l$ is the total parameter count. The circuit defines a smooth mapping $\Phi : \Theta \to \mathbb{CP}^{d-1}$ via $\Phi(\boldsymbol{\theta}) = [\mathcal{U}(\boldsymbol{\theta})|\psi_0\rangle]$, where $[\cdot]$ denotes equivalence up to global phase.

The natural distance between states is captured by pulling back the Fubini–Study metric of $\mathbb{CP}^{d-1}$ onto parameter space. This induced metric $g(\boldsymbol{\theta}) \in \mathbb{R}^{P\times P}$ is the real part of the Quantum Geometric Tensor (QGT):

$$g_{jk}(\boldsymbol{\theta}) = \mathrm{Re}\left[\langle\partial_j\psi(\boldsymbol{\theta})|\partial_k\psi(\boldsymbol{\theta})\rangle - \langle\partial_j\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|\partial_k\psi(\boldsymbol{\theta})\rangle\right]$$

where $|\psi(\boldsymbol{\theta})\rangle = \mathcal{U}(\boldsymbol{\theta})|\psi_0\rangle$ and $\partial_j \equiv \partial/\partial\theta_j$. When parameters are redundant or overparameterized, $g(\boldsymbol{\theta})$ can lose rank and become singular. To ensure invertibility everywhere, we use the regularized metric:

$$g^{(\epsilon)}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + \epsilon I, \quad \epsilon \geq 0 \tag{2}$$

Throughout this work, we operate on the $(\mathcal{M}, g^{(\epsilon)})$ manifold where $\mathcal{M} \subseteq \Theta$ under the ansatz in Definition 1. All optimization steps, tangent vectors, gradients, and Hessians are defined with respect to this regularized Riemannian structure.

2.2 Riemannian Updates: Connections, Geodesics, and Retractions

The objective of a variational quantum algorithm is to find parameters $\boldsymbol{\theta} \in \mathcal{M}$ that minimize the energy expectation of an observable $\hat{O}$: $\mathcal{L}(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|\hat{O}|\psi(\boldsymbol{\theta})\rangle$. On a Riemannian manifold $(\mathcal{M}, g^{(\epsilon)})$, the steepest descent direction is given by the Riemannian or natural gradient:

$$\mathrm{grad}_{g^{(\epsilon)}}\mathcal{L}(\boldsymbol{\theta}) = \left(g^{(\epsilon)}(\boldsymbol{\theta})\right)^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) \tag{3}$$

where $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) = (\partial_1\mathcal{L}, \ldots, \partial_P\mathcal{L})^\top$ is the standard Euclidean gradient. Because $\mathrm{grad}_{g^{(\epsilon)}}\mathcal{L}(\boldsymbol{\theta})$ is a vector in the tangent space $T_{\boldsymbol{\theta}}\mathcal{M}$, the space of directional velocities at $\boldsymbol{\theta}$, it must be mapped back onto the manifold $\mathcal{M}$ to update the parameters.

Definition 2 (Exponential Map and Retraction)**.** *Let $T_{\boldsymbol{\theta}}\mathcal{M} \cong \mathbb{R}^P$ denote the tangent space at $\boldsymbol{\theta}$, and let $\mathbf{v} \in T_{\boldsymbol{\theta}}\mathcal{M}$ be an update vector. The **exponential map** $\mathrm{Exp}_{\boldsymbol{\theta}} : T_{\boldsymbol{\theta}}\mathcal{M} \to \mathcal{M}$ maps $\mathbf{v}$ to $\gamma(1)$, where γ is the unique geodesic (shortest path) satisfying $\gamma(0) = \boldsymbol{\theta}$ with initial velocity $\dot{\gamma}(0) = \mathbf{v}$. A **retraction** $R_{\boldsymbol{\theta}} : T_{\boldsymbol{\theta}}\mathcal{M} \to \mathcal{M}$ is any smooth first-order approximation to the exponential map satisfying $R_{\boldsymbol{\theta}}(0) = \boldsymbol{\theta}$ and differential $\mathrm{D}R_{\boldsymbol{\theta}}(0) = \mathrm{id}_{T_{\boldsymbol{\theta}}\mathcal{M}}$.*

How an optimizer executes a step using the tangent vector $\mathbf{v} = -\eta\,\mathrm{grad}_{g^{(\epsilon)}}\mathcal{L}(\boldsymbol{\theta})$ separates Riemannian gradient descent from standard QNG (Definition 2).

3 Geodesic Smoothness of Variational Quantum Manifolds

This section addresses whether the variational quantum manifold is regular enough for Riemannian optimization theory to apply. We show that, despite its nontrivial structure, the quantum manifold is geodesically smooth, which yields the first convergence rate for Riemannian descent.

3.1 Geometric Regularity and the Quantum Obstruction

In smooth non-convex optimization, first-order Riemannian gradient descent (RGD) is guaranteed to find an approximate δ-stationary point at a rate of $\mathcal{O}(1/\delta)$ iterations tracking the squared gradient norm or the average decrease per step, provided the objective function is geodesically smooth. While this convergence guarantee is well-established on classical matrix manifolds such as the Stiefel manifold $\mathrm{St}(n,p)$, the Grassmannian $\mathrm{Gr}(n,p)$, and symmetric positive-definite cones Boumal (2023); Absil et al. (2008), the quantum statistical manifold $(\mathcal{M}, g^{(\epsilon)})$ has lacked formal convergence analysis. Prior works applying Riemannian methods to variational circuits have largely assumed this regularity Ferreira-Martins et al. (2025); Halla (2025a). However, this assumption may not necessarily be true. The quantum statistical manifold is structurally different from these well-studied classical manifolds in several respects.

On benchmark matrix manifolds, convergence rates and optimal step sizes are derived using known bounds on curvature, injectivity radius, and geodesic smoothness. Extending these guarantees to

variational quantum circuits requires addressing specific structural characteristics. Unlike homogeneous matrix spaces, the variational state map is parameterized by layers of non-commuting matrix exponentials (Definition 1). As a result, the metric is a highly nonlinear function of all circuit parameters rather than a uniform algebraic metric. In practical parameterized circuits, parameters can become redundant, causing the differential $d\Phi_{\boldsymbol{\theta}}$ to drop rank and making the unregularized pullback metric singular. The Tikhonov-regularized metric (eq. 2) ensures global invertibility, but requires bounds that hold uniformly across the transition between curved and flat directions.

3.2 Geodesic Smoothness and Convergence

The variation of the Riemannian gradient along geodesics is governed by the covariant Hessian.

$$\mathrm{Hess}_{\nabla}\mathcal{L}(X,Y) = g^{(\epsilon)}(\nabla_X \mathrm{grad}_{g^{(\epsilon)}}\mathcal{L}, Y), \quad \forall X, Y \in T_{\boldsymbol{\theta}}\mathcal{M} \tag{4}$$

An objective is geodesically L_{R}-smooth if $\|\mathrm{Hess}_{\nabla}\mathcal{L}(\boldsymbol{\theta})\|_{g^{(\epsilon)}} \leq \mathsf{L}_{\mathrm{R}}$ uniformly across $\mathcal{M}$. To bound this quantity, we decompose the covariant Hessian into coordinate derivatives and Christoffel connection terms (Lemmas 1–8). Through a sequence of Cauchy–Schwarz bounds, the generator norms factor into a rank-one outer product, collapsing the spectral norm computations and yielding a closed-form smoothness bound (Proof in Appendix D).

Theorem 1 (Geodesic Smoothness)**.** *Let $(\mathcal{M}, g^{(\epsilon)})$ be the quantum statistical manifold with the regularized metric $g^{(\epsilon)}$ (eq. 2), and let $\mathcal{L}(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|\hat{O}\,|\psi(\boldsymbol{\theta})\rangle$. Then $\mathcal{L}$ is geodesically L_{R}-smooth:*

$$\|\mathrm{Hess}_{\nabla}\mathcal{L}(\boldsymbol{\theta})\|_{g^{(\epsilon)}} \;\leq\; \mathsf{L}_{\mathrm{R}} \;=\; \tfrac{\|\hat{O}\|_2}{\epsilon}\left[4 + \tfrac{18}{\epsilon}\left(\textstyle\sum_l \sum_k \|\mathfrak{g}_{(l,k)}\|_2\right)^2\right] \textstyle\sum_l \sum_k \|\mathfrak{g}_{(l,k)}\|_2^2. \tag{5}$$

As an immediate consequence of Theorem 1, the Riemannian Descent Lemma (Boumal, 2023) yields the first provable convergence rate for Riemannian gradient descent on a quantum variational landscape.

Corollary 1 (Convergence Rate of Riemannian Gradient Descent)**.** *Under the hypotheses of Theorem 1, Riemannian gradient descent with exact geodesic steps and step size $\eta = 1/\mathsf{L}_{\mathrm{R}}$ converges to an δ-stationary point (where $\frac{1}{T}\sum_{t=0}^{T-1} \|grad_{g^{(\epsilon)}}\mathcal{L}(\boldsymbol{\theta}^{(t)})\|^2_{g^{(\epsilon)}} \leq \delta$) in at most:*

$$T = \mathcal{O}\left(\frac{\mathsf{L}_{\mathrm{R}}(\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}^*)}{\delta}\right) \quad \textit{iterations} \tag{6}$$

where $\mathcal{L}^* = \inf_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$ is the global minimum energy.

The proof of Corollary 1 is given in Appendix E. Theorem 1 and Corollary 1 establish that the quantum statistical manifold is regular enough to support standard non-convex Riemannian convergence rates. While our smoothness constant L_{R} represents a conservative worst-case bound that scales polynomially in the parameter count and ϵ^{-1}, it provides the first convergence rate for Riemannian descent on variational quantum landscapes. Future work may tighten this constant, regardless the primary theoretical takeaway is established: the underlying optimization landscape is well-behaved. The empirical breakdown of standard QNG in deep circuits must therefore stem from how updates are executed in practice.

4 Noise-Induced Riemannian Spike

Every quantum processor operating below the fault-tolerance threshold is subject to environmental decoherence Huo et al. (2025); Preskill (2018). In Euclidean optimization, noise induces barren plateaus (NIBP) that contract gradient magnitudes exponentially with circuit depth, causing optimization to stall Wang et al. (2021). While empirical benchmarks consistently report that Quantum Natural Gradient (QNG) also degrades under hardware noise Koczor & Benjamin (2022); Lockwood (2022), this failure is often conflated with Euclidean signal loss. However, the standard NIBP framework evaluates gradients strictly within the ambient coordinate chart; it characterizes the exterior derivative $d\mathcal{L}$, but leaves the Riemannian gradient unexamined. On a quantum statistical manifold, the natural gradient is obtained by contracting $d\mathcal{L}$ with the inverse metric tensor $(g^{(\epsilon)})^{-1}$,

which is itself an explicit function of the noisy state. Because the pullback metric and the cost gradient contract at fundamentally different exponential rates, the Riemannian landscape does not simply inherit the Euclidean plateau. In Theorem 2, we establish that rather than vanishing, the natural gradient *diverges* exponentially with circuit depth under weak regularization. We term this failure mode the *noise-induced Riemannian spike* (NIRS).

Setup and Noise Model: Following the framework of Wang et al. (2021), consider an n-qubit circuit of depth L consisting of alternating parameterized unitary layers $\mathcal{U}_l(\theta_l)$ and local depolarizing channels $\mathcal{N} = \bigotimes_{j=1}^{n} \mathcal{N}_j$. Each local channel attenuates non-identity Pauli components by a contraction factor $\kappa_{\text{err}} \in (0, 1)$. By the strict data processing inequality for the sandwiched 2-Rényi divergence Müller-Hermes & Reeb (2017); Wang et al. (2021), each noisy layer contracts the state toward the maximally mixed state I/d (with $d = 2^n$) via $D_2(\mathcal{N}(\rho) \,\|\, I/d) \leq \kappa_{\text{err}}^2 D_2(\rho \,\|\, I/d)$. Across L layers, the noisy output state $\tilde{\rho}_{\psi_{\boldsymbol{\theta}}}$ contracts exponentially to the maximally mixed state in trace distance:

$$\left\|\tilde{\rho}_{\psi_{\boldsymbol{\theta}}} - I/d\right\|_1 = \|\tilde{\Delta}_L(\boldsymbol{\theta})\|_1 \leq \sqrt{2 \ln 2}\; n^{1/2}\; \kappa_{\text{err}}^L. \tag{7}$$

where $\tilde{\rho}_{\psi_{\boldsymbol{\theta}}} = \frac{I}{d} + \tilde{\Delta}_L(\boldsymbol{\theta})$ separates the state into the parameter-independent identity background and a traceless signal $\tilde{\Delta}_L(\boldsymbol{\theta})$. Crucially, because I/d is invariant under all unitary operations, all parameter-dependent quantities including the cost gradient, the Fubini-Study metric, and the affine connection reside within the exponentially vanishing signal $\tilde{\Delta}_L(\boldsymbol{\theta})$.

Theorem 2 (Noise-Induced Riemannian Spike)**.** *Let $(\mathcal{M}, \tilde{g}^{(\epsilon)})$ be the quantum statistical manifold with regularized metric $\tilde{g}^{(\epsilon)}$ (eq. 2), subject to local depolarizing noise with per-layer contraction $\kappa_{\text{err}} \in (0, 1)$ across circuit depth L. Then, as $L \to \infty$:*

(i) Geometric Invariance: *For the unregularized metric $\tilde{g}$, despite the pullback metric contracting as $\|\tilde{g}\|_{\text{op}} \in \mathcal{O}(\kappa_{\text{err}}^{2L})$, the connection matrices $\tilde{\Gamma}_{ij}^k$ of the Levi-Civita connection remain asymptotically stable:*

$$\|\tilde{\Gamma}^k\|_{\text{op}} \in \mathcal{O}(1) \quad \textit{for all } k \in \{1, \ldots, P\}. \tag{8}$$

(ii) Riemannian Gradient Bifurcation: *For any parameter configuration $\boldsymbol{\theta}$ away from stationary points with non-vanishing Euclidean covector signal $\|d\tilde{\mathcal{L}}(\boldsymbol{\theta})\|_2 = \Omega(\kappa_{\text{err}}^L)$, the natural gradient exhibits a regularization-dependent bifurcation. Under weak regularization ($\epsilon \ll \kappa_{err}^{2L}$) the update norm experiences an exponential spike:*

$$\|grad_{\tilde{g}^{(\epsilon)}} \tilde{\mathcal{L}}(\boldsymbol{\theta})\|_2 \in \Omega(\kappa_{err}^{-L}) \quad \textit{(Noise-Induced Riemannian Spike)}. \tag{9}$$

Under heavy regularization ($\epsilon \gg \kappa_{err}^{L}$), the regularizer dominates erasing geometric information and collapsing the update into the Euclidean barren plateau:

$$\|grad_{\tilde{g}^{(\epsilon)}} \tilde{\mathcal{L}}(\theta)\|_2 \in \mathcal{O}(\kappa_{err}^{L}/\epsilon) \quad \textit{(Noise-Induced Barren Plateau)}.$$

The full proof is given in Appendix G. Theorem 2 reveals that noise degrades variational quantum optimization not by corrupting the manifold's intrinsic geometry, but by distorting the interface between the metric and the cost function. Statement (i) follows because both the Fubini–Study metric and its coordinate derivatives decay at the exact same exponential rate. In the Christoffel symbols, this shared decay cancels out algebraically, leaving the Levi-Civita connection and geodesic paths asymptotically invariant. Statement (ii) stems from a contraction rate mismatch. While the Euclidean gradient contracts as κ_{err}^L across L noisy layers, the pullback metric contracts quadratically faster as κ_{err}^{2L} because it measures the overlap of two perturbed states. Under weak metric regularization, converting the covector $d\tilde{\mathcal{L}}$ into a tangent update vector via $(\tilde{g}^{(\epsilon)})^{-1}$ amplifies the gradient by $\kappa_{\text{err}}^{-2L}$, producing an exponential net blowup that scales as κ_{err}^{-L}.

NIRS provides a physical explanation for the severe ill-conditioning of QNG in noisy regimes, revealing a fundamentally surprising failure mode in quantum optimization. While the extensive

literature on noise-induced barren plateaus establishes that decoherence universally suppresses gradient signals, the Riemannian setting can drive the exact opposite: an exponential gradient growth. Although metric regularization allows one to trade off between this Riemannian spike and a standard barren plateau, severe optimization degradation remains unavoidable in deep circuits. We confirm this inverse behavior empirically on a 2-qubit VQE benchmark under depolarizing noise using the unregularized metric (Appendix G, Figure 3), where the Euclidean and Riemannian gradient norms diverge in exact opposition as circuit depth and error rates scale.

5 DFS-QNG: Dually Flat Stratified Quantum Natural Gradient

In this section, we formalize the practical contribution of this work: Dually Flat Stratified Quantum Natural Gradient (DFS-QNG). We focus on three concrete practical issues of standard QNG that DFS-QNG is designed to improve on. First, standard QNG is computationally expensive. Evaluating the full pullback metric requires $\mathcal{O}(s \cdot P^2)$ circuit executions, where s is the shot budget. Thus, researchers commonly use diagonal versions or metric approximation methods Stokes et al. (2020); Yamamoto (2019).

Second, the regularized metric (Eq. 2) becomes ill-conditioned as circuit depth and parameter count scale van Straaten & Koczor (2021); Haug et al. (2021); Wierichs et al. (2020). Redundant parameterizations and entangling layers have been found to make the matrix inverse in the natural gradient update (Eq. 3) sensitive to numerical noise and regularization tuning Larocca et al. (2023); Haug et al. (2021). Metric ill-conditioning remains an active failure mode even in noiseless settings Wierichs et al. (2020). Third, standard QNG does not follow true geodesics. Instead, it updates parameters using a first-order Euclidean retraction (an Euler step) rather than the Riemannian exponential map (Definition 2). This introduces a local retraction error $\boldsymbol{\mathcal{E}}$ proportional to the Christoffel symbols Γ^k_{ij} of the connection:

$$\boldsymbol{\mathcal{E}}^k \;=\; \frac{\eta^2}{2} \sum_{i,j} \Gamma^k_{ij}\, \mathbf{v}^i\, \mathbf{v}^j \;+\; \mathcal{O}(\eta^3) \tag{10}$$

Over successive iterations, these retraction errors compound and cause the optimizer to drift from the true steepest descent path which is a well-known source of lag in Riemannian optimization. Convergence guarantees require the retraction to approximate the exponential map to at least second order, and the error grows with both the manifold's curvature and the step magnitude Boumal (2023); Absil et al. (2008); Boumal et al. (2019); Kasai et al. (2018). Prior attempts to follow true geodesics numerically integrate the geodesic ODEs, which incurs an $\mathcal{O}(P^3)$ per-step cost Halla (2025b).

5.1 Riemannian Block Descent: The DFS-QNG Algorithm

To address the practical bottlenecks of QNG, we first focus on eliminating retraction errors. To do so we take inspiration from prior work that utilize commuting structures in their circuits. In Commuting Block Circuits (CBC) Bowles et al. (2025), grouping commuting generators enabled parallel gradient evaluation, achieving backpropagation-like measurement scaling across circuit blocks. More recently Stabilizer-Logical Product Ansatz (SLPA) Chinzei et al. (2025) formalized this principle by demonstrating a fundamental trade-off between gradient measurement efficiency and circuit expressivity, using stabilizer symmetries to maximize the number of simultaneously measurable gradient components. Furthermore, related commuting-block formulations have been leveraged to reduce the sample complexity of evaluating the Quantum Fisher Information Matrix itself Gomez-Lurbe (2026). However, these existing frameworks utilize commutativity exclusively for circuit-side measurement efficiency. In contrast, we examine commuting structures from an optimization perspective. We find that restricting updates to commuting strata induces dually flat submanifolds where the Christoffel symbols vanish identically, rendering first-order Euler steps geodesically exact (Proposition 1).

Proposition 1 (Geodesic Exactness under Abelian Stratification)**.** *Let $\mathcal{M}$ be a parameter manifold with mutually commuting generators $[\mathfrak{g}_{(n,i)}, \mathfrak{g}_{(m,j)}] = 0$. Then $\mathcal{M}$ is dually flat in the sense of Amari (1998): the regularized pullback metric $g^{(\epsilon)}$ is coordinate-invariant ($\partial_k g^{(\epsilon)}_{ij} = 0$), and all Christoffel symbols vanish identically ($\Gamma^k_{ij} = 0$). Consequently, Riemannian geodesics are affine*

lines, the exponential map reduces to an Euler step $\mathrm{Exp}_{\boldsymbol{\theta}}(\mathbf{v}) = \boldsymbol{\theta} + \mathbf{v}$*, and the retraction defect vanishes identically (*$\boldsymbol{\mathcal{E}} = \mathbf{0}$*).*

We prove Proposition 1 in Appendix B. While Proposition 1 guarantees that QNG steps are geodesically exact on commuting manifolds, making an entire circuit commute globally destroys its utility. A circuit whose generators all commute collapses the dynamical Lie algebra (DLA) to an Abelian subalgebra, confining state trajectories to a maximal torus and eliminating the capacity to explore the full Hilbert space. We resolve this expressivity issue by taking inspiration from layer-wise learning and block-coordinate optimization Skolik et al. (2021); Grant et al. (2019); Kverne et al. (2026). In QML, layer-wise strategies are traditionally used to mitigate barren plateaus or reduce active gradients during training. Here, we utilize block partitioning for a fundamentally geometric purpose: preserving a non-Abelian global DLA while isolating locally flat optimization submanifolds. Specifically, we partition the P circuit parameters into $\mathfrak{S}$ disjoint blocks or strata, $\Theta = \bigsqcup_{\pi=1}^{\mathfrak{S}} \Theta_\pi$, requiring only intra-stratum commutativity:

$$[\mathfrak{g}_{(\pi,i)}, \mathfrak{g}_{(\pi,j)}] = 0 \quad \forall i,j \in \Theta_\pi, \quad \forall \pi \in 1,\dots,\mathfrak{S}. \tag{11}$$

DFS-QNG is detailed in Algorithm 1. Procedurally, DFS-QNG is simply block-wise (or layer-wise) QNG where the parameters updated in each step belong to a mutually commuting set. We do not claim this algorithmic setup itself as our primary novelty as layer-freezing and block scheduling are well-established. Rather, our core contribution is an optimization insight: we show that restricting updates to commuting blocks while freezing remaining parameters formally **transforms standard QNG into exact Riemannian block coordinate descent**. (Proof of Theorem 3 in Appendix C)

Theorem 3 (Convergence of DFS-QNG via Riemannian Block Descent)**.** *Let the objective* $\mathcal{L}(\boldsymbol{\theta})$ *be* L_{R}*-smooth on* $(\mathcal{M}, g^{(\epsilon)})$*. Under an Abelian stratification (Eq. 11), each stratum update corresponds to exact geodesic transport on a dually flat submanifold (Proposition 1). DFS-QNG converges monotonically to an* δ*-stationary point in*

$$T = \mathcal{O}\left(\frac{\mathsf{L}_{\mathrm{R}}}{\delta}\right) \quad \textit{iterations}.$$

DFS-QNG is practically lightweight. Because each stratum metric g_π is small ($l_\pi \ll P$), classical inversion is numerically stable and quantum measurement overhead drops substantially. DFS-QNG also reduces the per-step metric inversion and measurement overhead from $\mathcal{O}(P^2)$ to $\mathcal{O}\left(\sum_{\pi=1}^{\mathfrak{S}} l_\pi^2\right)$, where $l_\pi = |\Theta_\pi|$. Crucially, this commuting structure imposes no architectural burden: standard Hardware-Efficient Ansätze (alternating rotation and entangler layers) and QAOA circuits (alternating cost and mixer unitaries) natively satisfy intra-stratum commutativity out of the box.

Algorithm 1 DFS-QNG (Riemannian Block Descent)

Require: Parameters $\boldsymbol{\theta}^{(0)} \in \Theta \subseteq \mathbb{R}^P$; Abelian stratification $\{\Theta_\pi\}_{\pi=1}^{\mathfrak{S}}$ satisfying equation 11; shift s; step size η; horizon T.

for $t = 0, 1, \dots, T-1$ **do**
  for $\pi = 1, 2, \dots, \mathfrak{S}$ **do**
    Extract $\boldsymbol{\theta}_\pi^{(t)}$; freeze $\boldsymbol{\theta}_{\setminus\pi}^{(t)}$ ▷ restrict to $\mathcal{S}_\pi$
    $[d_\pi \mathcal{L}]_k \leftarrow \big(\mathcal{L}(\boldsymbol{\theta}^{(t)} + s\,\boldsymbol{e}_k) - \mathcal{L}(\boldsymbol{\theta}^{(t)} - s\,\boldsymbol{e}_k)\big)/2\sin s|_{k\in\Theta_\pi}$ ▷ parameter-shift rule
    $g_\pi \leftarrow \mathrm{Re}\big[\langle\partial_{\pi,i}\psi_{\boldsymbol{\theta}}|\partial_{\pi,j}\psi_{\boldsymbol{\theta}}\rangle - \langle\partial_{\pi,i}\psi_{\boldsymbol{\theta}}|\psi_{\boldsymbol{\theta}}\rangle\langle\psi_{\boldsymbol{\theta}}|\partial_{\pi,j}\psi_{\boldsymbol{\theta}}\rangle\big]_{i,j\in\Theta_\pi}$ ▷ stratum metric
    $\boldsymbol{\theta}_\pi^{(t+1)} \leftarrow \boldsymbol{\theta}_\pi^{(t)} - \eta\,\mathrm{grad}_{g_\pi^{(\epsilon)}}\mathcal{L}$ ▷ exact $\nabla^{(e)}$-geodesic: $\Gamma_{ij}^{(e)k}\big|_{\Theta_\pi} = 0$
    $\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}_{\setminus\pi}^{(t)} \oplus \boldsymbol{\theta}_\pi^{(t+1)}$ ▷ recombine
  end for
end for
return $\boldsymbol{\theta}^{(T)}$

5.2 Experimental Setup and Results

We benchmarked DFS-QNG on three very common variational quantum problems. Variational Quantum Eigensolver (VQE): ground-state preparation of the 1D transverse-field Ising model with Hamiltonian $\hat{O} = \sum Z_i Z_{i+1} + \sum X_i$, using a hardware-efficient ansatz partitioned into alternating strata of σ_Z and σ_X rotations. Quantum Approximate Optimization Algorithm (QAOA): MaxCut on both ring graphs and 3-regular graphs, where the circuits naturally stratify into a cost stratum of ZZ gates followed by a mixer stratum of σ_X rotations. The ring graph QAOA problem was ran in a noisy environment with depolarizing channels following every gate and a 1000 shot budget while the 3-regular graph and VQE problem were ran in noiseless settings. All problems were evaluated across $\{4, 6, 8, 10\}$ qubits and $\{2, 4, 6, 8, 10\}$ strata, with 5 independent runs per configuration. We report the Normalized Area Above Optimal (NAAO), which measures cumulative deviation from the exactly known ground-state energy over the course of training. We also report the lowest stable energy achieved by each optimizer and the number of epochs required to reach it.

The quantitative results are summarized in tables 1, 2, and 3 for the NAAO metric and tables 4, 5, and 6 for the lowest energy reached (Appendix H). The central finding is qualitative: As circuit depth increases, the optimization trajectories of standard QNG diverge, frequently failing to reach a stationary point near the ground state, whereas DFS-QNG consistently guarantees asymptotic descent (Figure 1). DFS-QNG achieves strictly lower NAAO than standard QNG at every qubit count and stratum configuration tested in both noisy and noiseless settings, with no exceptions across all three problems.

Averaged across all configurations, DFS-QNG attains a mean NAAO of 1.34, 1.26, and 2.44 on MaxCut ring, MaxCut 3-regular, and VQE respectively, compared to 2.50, 2.81, and 3.29 for QNG, reductions of 47%, 55%, and 26%. More notably, DFS-QNG also outperforms Adam Kingma & Ba (2014) on each configuration, which scores a mean NAAO of 1.93, 1.88, and 3.76 on the same benchmarks, placing DFS-QNG roughly 31%–35% lower across problems. DFS-QNG also consistently finds a lower stable energy much faster than the baselines (Tables 4–6). This is significant because it demonstrates that the Fubini–Study geometry does contain useful optimization information, the failure of standard QNG is not evidence against geometry-aware optimization, but against the global metric inversion that QNG requires to access it. DFS-QNG recovers this geometric advantage by localizing the inversion to well-conditioned commuting blocks and following block-wise accurate geodesics.

Comparison with prior Riemannian methods. DFS-QNG is also the only Riemannian method that is practical at scales. Geodesic QNG (GQNG) Halla (2025b) integrates the geodesic equation at each step, incurring $\mathcal{O}(P^3)$ cost per iteration, computationally impossible for most circuits. Exact Geodesic Transport (EGT) Ferreira-Martins et al. (2025) admits closed-form geodesics but restricts the ansatz to a real projective manifold with an exponentially scaling parameterization and CNOT count of $\mathcal{O}\binom{n}{k}$, and cannot be applied to QAOA or QNN problems. Table 3 includes EGT on the VQE benchmark where it is applicable: EGT achieves strong NAAO scores, but its performance is optimizing a different circuit from QNG/DFS-QNG/Adam, as a direct comparison is impossible. In contrast, DFS-QNG operates on any stratified circuit without architectural modification, hardware-efficient ansätze, QAOA, and QNN circuits all admit natural Abelian stratifications, while requiring only the inversion of small $\ell_\pi \times \ell_\pi$ per-stratum metric blocks rather than the full $P \times P$ quantum

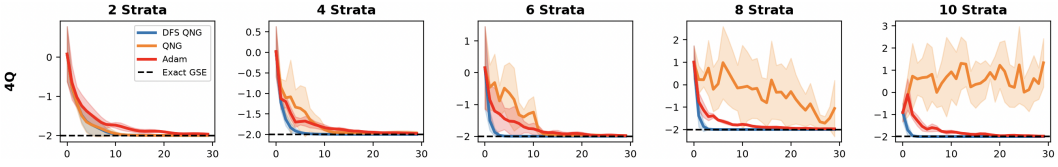


Figure 1: Convergence trajectories of DFS-QNG, QNG, and Adam on the QAOA MaxCut 3-regular graph with 2-10 strata for 4 qubits. Dashed black lines indicate the exact ground-state energy. Shaded regions denote ± 1 standard deviation over 5 runs. QNG becomes less stable as circuit grows. Full training plots in Fig. 4 for this problem.

geometric tensor, reducing both the quantum measurement cost and the classical inversion cost from $\mathcal{O}(P^2)$ to $\mathcal{O}(\sum_\pi \ell_\pi^2)$.

Table 1: Normalized Area Above Optimal (↓) on QAOA MaxCut3 graph. Budget equals total epochs (30). Best per cell in **bold**.

n	Optimizer	1 Stratum	2 Strata	3 Strata	4 Strata	5 Strata
4	DFS-QNG	**0.17**$\pm$0.07	**0.09**$\pm$0.03	**0.06**$\pm$0.04	**0.08**$\pm$0.04	**0.03**$\pm$0.02
	QNG	0.17$\pm$0.08	0.26$\pm$0.17	0.45$\pm$0.26	1.64$\pm$1.07	2.58$\pm$1.03
	Adam	0.30$\pm$0.11	0.22$\pm$0.07	0.27$\pm$0.18	0.27$\pm$0.07	0.27$\pm$0.09
6	DFS-QNG	**1.75**$\pm$0.44	**0.43**$\pm$0.19	**0.38**$\pm$0.12	**0.21**$\pm$0.06	**0.25**$\pm$0.11
	QNG	1.87$\pm$0.43	0.89$\pm$0.50	1.28$\pm$0.57	1.91$\pm$0.70	2.39$\pm$0.63
	Adam	2.15$\pm$0.31	0.88$\pm$0.20	0.63$\pm$0.13	0.56$\pm$0.11	0.81$\pm$0.23
8	DFS-QNG	**3.79**$\pm$0.42	**1.31**$\pm$0.70	**0.81**$\pm$0.30	**0.70**$\pm$0.23	**0.54**$\pm$0.08
	QNG	3.95$\pm$0.40	1.94$\pm$0.82	2.74$\pm$0.81	3.76$\pm$0.92	4.44$\pm$1.00
	Adam	4.22$\pm$0.34	2.19$\pm$0.67	1.96$\pm$0.66	1.48$\pm$0.43	1.53$\pm$0.48
10	DFS-QNG	**5.70**$\pm$0.66	**2.99**$\pm$0.74	**2.59**$\pm$1.07	**1.81**$\pm$0.48	**1.46**$\pm$0.57
	QNG	5.71$\pm$0.50	3.82$\pm$0.77	4.81$\pm$0.71	5.46$\pm$1.02	6.13$\pm$1.16
	Adam	6.06$\pm$0.69	3.95$\pm$0.53	3.79$\pm$0.88	3.39$\pm$0.76	2.57$\pm$0.68

6 Limitations and Conclusion

Our work provides a geometric account of when and why Riemannian optimization succeeds or fails on variational quantum landscapes. We establish geodesic smoothness of the variational quantum manifold and corresponding convergence guarantees, identify and rigorously characterize noise-induced Riemannian spikes (NIRS) as a distinct failure mechanism arising from the differential contraction of the noisy metric and gradient, and introduce Dually Flat Stratified QNG (DFS-QNG) to exploit local commuting structure through exact Riemannian block-coordinate updates. These results go beyond proposing an alternative optimizer: they establish that the instability of noisy natural-gradient methods has a fundamental geometric origin, and that metric inversion, not the particular optimization rule, is ultimately responsible for the exponential amplification. Our analysis therefore identifies a fundamental boundary for Riemannian optimization under noise while providing a constructive way to operate within well-conditioned regions of the manifold. Together with our theoretical convergence results and VQE/QAOA experiments, this work provides a unified framework connecting quantum-channel contraction, Riemannian geometry, and optimization stability, and establishes a foundation for designing more stable quantum optimization methods.

Our results also have several well-defined limitations. DFS-QNG does not eliminate NIRS (Theorem 2), since its stratum-wise updates necessarily retain $(\tilde{g}^{(\epsilon)})^\sharp$, whose ill-conditioning is intrinsic to the noisy metric; instead, DFS-QNG improves its impact by restricting updates to better conditioned submanifolds. The intra-stratum commutativity condition limits the architectures covered by the present guarantees, although it remains compatible with the VQE and QAOA circuits studied here and permits substantial flexibility in block construction. The geodesic smoothness bound in Theorem 1 is likewise conservative because of accumulated worst-case spectral-norm, Cauchy–Schwarz, and rank-one outer-product bounds. Its purpose is to establish manifold regularity and convergence rather than provide a practical learning-rate rule, so $\eta = 1/\mathsf{L}_{\mathrm{R}}$ should not be interpreted as an empirical prescription. Future work can build directly on this foundation by deriving tighter Hessian scalings, extending the NIRS analysis to broader noise models, and developing adaptive stratification and step-size strategies.

AI use statement

In this work, we used generative AI tools for assisting in the writing and verification of mathematical proofs. We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, designing research methodology or experiments, implementing core methods, or

interpreting results. and assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis are not applicable to this work.

Additionally, we used generative AI tools for editing the manuscript to improve readability, drafting select text segments, and creating or editing software code (specifically for debugging and writing plotting scripts).

We have reviewed all AI-assisted work. Proofs: All proofs and derivations were checked and independently verified by the authors. Code and Visualizations: All plotting scripts and debugging suggestions were inspected and ran. Text: All draft and edit suggestions were reviewed, rewritten, and fact-checked by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

Nicolas Boumal, Pierre-Antoine Absil, and Coralia Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019.

Joseph Bowles, David Wierichs, and Chae-Yeun Park. Backpropagation scaling in parameterised quantum circuits. *Quantum*, 9:1873, 2025.

Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, 2021.

Koki Chinzei, Shinichiro Yamano, Quoc Hoan Tran, Yasuhiro Endo, and Hirotaka Oshima. Trade-off between gradient measurement efficiency and expressivity in deep quantum neural networks. *npj Quantum Information*, 11(1):79, 2025.

Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm. *arXiv preprint arXiv:1411.4028*, 2014.

André J Ferreira-Martins, Renato Farias, Giancarlo Camilo, Thiago O Maciel, Allan Tosta, Ruge Lin, Abdulla Alhajri, Tobias Haug, and Leandro Aolita. Variational quantum algorithms with exact geodesic transport. *arXiv preprint arXiv:2506.17395*, 2025.

David Fitzek, Robert S Jonsson, Werner Dobrautz, and Christian Schäfer. Optimizing variational quantum algorithms with qbang: efficiently interweaving metric and momentum to navigate flat energy landscapes. *Quantum*, 8:1313, 2024.

Rafael Gomez-Lurbe. Efficient protocol to estimate the quantum fisher information matrix for commuting-block circuits: R. gómez-lurbe. *Quantum Information Processing*, 25(9):278, 2026.

Edward Grant, Leonard Wossnig, Mateusz Ostaszewski, and Marcello Benedetti. An initialization strategy for addressing barren plateaus in parametrized quantum circuits. *Quantum*, 3:214, 2019.

Mourad Halla. Modified conjugate quantum natural gradient. *EPJ Quantum Technology*, 12(1): 1–16, 2025a.

Mourad Halla. Quantum natural gradient with geodesic corrections for small shallow quantum circuits. *Physica Scripta*, 100(5):055121, 2025b.

Tobias Haug, Kishor Bharti, and MS Kim. Capacity and quantum geometry of parametrized quantum circuits. *PRX Quantum*, 2(4):040309, 2021.

Yuqian Huo, Jinbiao Wei, Christopher Kverne, Mayur Akewar, Janki Bhimani, and Tirthak Patel. Revisiting noise-adaptive transpilation in quantum computing: How much impact does it have? In *2025 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, pp. 1–9. IEEE, 2025.

Benjamin DM Jones, Lana Mineh, and Ashley Montanaro. Benchmarking a wide range of optimisers for solving the fermi–hubbard model using the variational quantum eigensolver. *Quantum Science and Technology*, 10(4):045032, 2025.

Hiroyuki Kasai, Hiroyuki Sato, and Bamdev Mishra. Riemannian stochastic recursive gradient algorithm. In *International conference on machine learning*, pp. 2516–2524. PMLR, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Bálint Koczor and Simon C Benjamin. Quantum natural gradient generalized to noisy and nonunitary circuits. *Physical Review A*, 106(6):062416, 2022.

Christopher Kverne, Mayur Akewar, Yuqian Huo, Tirthak Patel, and Janki Bhimani. Wsbd: Freezing-based optimizer for quantum neural networks. *arXiv preprint arXiv:2602.11383*, 2026.

Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J Coles, and Marco Cerezo. Theory of overparametrization in quantum neural networks. *Nature Computational Science*, 3(6):542–551, 2023.

Owen Lockwood. An empirical review of optimization techniques for quantum variational circuits. *arXiv preprint arXiv:2202.01389*, 2022.

Alexander Müller-Hermes and David Reeb. Monotonicity of the quantum relative entropy under positive maps. In *Annales Henri Poincaré*, volume 18, pp. 1777–1788. Springer, 2017.

Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J Love, Alán Aspuru-Guzik, and Jeremy L O'brien. A variational eigenvalue solver on a photonic quantum processor. *Nature communications*, 5(1):4213, 2014.

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

Andrea Skolik, Jarrod R McClean, Masoud Mohseni, Patrick Van Der Smagt, and Martin Leib. Layerwise learning for quantum neural networks. *Quantum Machine Intelligence*, 3(1):5, 2021.

James C Spall. An overview of the simultaneous perturbation method for efficient optimization. *Johns Hopkins apl technical digest*, 19(4):482–492, 1998.

James Stokes, Josh Izaac, Nathan Killoran, and Giuseppe Carleo. Quantum natural gradient. *Quantum*, 4:269, 2020.

Barnaby van Straaten and Bálint Koczor. Measurement cost of metric-aware variational quantum algorithms. *PRX Quantum*, 2(3):030324, 2021.

Tuan Hai Vu, Vu Trung Duong Le, Hoai Luan Pham, and Yasuhiko Nakashima. Benchmarking variants of the adam optimizer for quantum machine learning applications. *IEEE Open Journal of the Computer Society*, 2025.

Samson Wang, Enrico Fontana, Marco Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J Coles. Noise-induced barren plateaus in variational quantum algorithms. *Nature communications*, 12(1):6961, 2021.

Yizhi Wang, Shichuan Xue, Yaxuan Wang, Jiangfang Ding, Weixu Shi, Dongyang Wang, Yong Liu, Yingwen Liu, Xiang Fu, Guangyao Huang, et al. Experimental quantum natural gradient optimization in photonics. *Optics Letters*, 48(14):3745–3748, 2023.

David Wierichs, Christian Gogolin, and Michael Kastoryano. Avoiding local minima in variational quantum eigensolvers with the natural gradient optimizer. *Physical Review Research*, 2 (4):043246, 2020.

Naoki Yamamoto. On the natural gradient for variational quantum eigensolver. *arXiv preprint arXiv:1909.05074*, 2019.

APPENDIX

A FOUNDATIONS OF QUANTUM COMPUTING

This section outlines the standard quantum mechanical framework that underpins variational quantum algorithms. It establishes the mathematical formalism for quantum states, their evolution, and the physical realities of measurement and noise.

A.1 QUANTUM STATES AND AMPLITUDES

An n-qubit pure state is represented as a unit vector $|\psi\rangle$ living in a complex Hilbert space $\mathcal{H} \cong \mathbb{C}^{2^n}$. Unlike classical bits, which must exist in a deterministic state of 0 or 1, a quantum state can exist in a superposition of all possible computational basis states. For a single qubit, this is expressed as:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$$

The coefficients α and β are complex numbers known as probability amplitudes. The fundamental rule of quantum mechanics requires that the state vector is strictly normalized, meaning the sum of the absolute squares of its amplitudes must equal 1 (representing a 100% total probability):

$$|\alpha|^2 + |\beta|^2 = 1$$

Because global phase is physically unobservable, quantum states are formally identified with rays in a projective Hilbert space. For a single qubit, this projective geometry allows us to map any valid pure state to a point on the surface of a 3D unit sphere, known as the Bloch sphere.

A.2 QUANTUM OPERATORS: UNITARY AND HERMITIAN MATRICES

Transformations and measurements in quantum computing are driven by linear operators, which are represented as matrices. These matrices fall into two strictly defined mathematical categories:

1. Unitary Matrices (Evolution) Closed quantum systems evolve over time through the application of quantum gates, which are mathematically represented by unitary matrices $\mathcal{U}$. A matrix is unitary if its conjugate transpose $\mathcal{U}^\dagger$ is equal to its inverse $\mathcal{U}^{-1}$, meaning $\mathcal{U}^\dagger\mathcal{U} = I$. This property is vital because it ensures that the operation is reversible and preserves the l_2-norm of the state vector, guaranteeing that total probability remains 1.

2. Hermitian Matrices (Observables) When we measure a quantum system to extract information, we apply an observable. In quantum mechanics, an observable is always represented by a Hermitian matrix $\hat{O}$. A matrix is Hermitian if it is equal to its own conjugate transpose ($\hat{O} = \hat{O}^\dagger$). This mathematical property guarantees that all eigenvalues of the matrix are real numbers. This is a critical requirement, as these eigenvalues correspond to the actual, physical quantities we can record in a laboratory.

A.3 PARAMETERIZED QUANTUM CIRCUITS

In the near-term era, we manipulate quantum states using parameterized quantum circuits (ansätze). These circuits prepare a family of states via layered sequences of fixed entangling gates and parameterized single-qubit rotations.

Figure 2 is a standard representation of a hardware-efficient ansatz. The circuit begins with a state preparation or data-encoding block $\mathcal{U}_{in}(x)$, followed by repeating layers of parameterized rotations (e.g., R_x, R_z) and entangling CNOT gates, culminating in the measurement of the expected value of an observable $\hat{O}$:

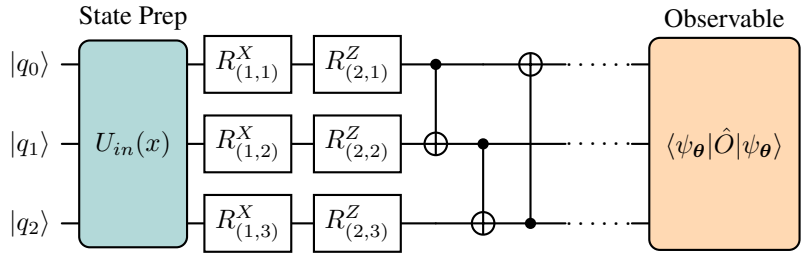


Figure 2: A standard hardware-efficient parameterized quantum circuit. This circuit follows the stratification required from Proposition 1 implicitly.

A.4 Measurement and Hardware Noise

In an ideal, noiseless scenario, the expectation value of a parameterized state with respect to an observable is deterministic. However, physical quantum hardware interacts with its surrounding environment, leading to decoherence.

This environment-induced noise transforms our idealized pure state $|\psi\rangle\langle\psi|$ into a mixed state, represented by a density matrix ρ. A common analytical model for this is the depolarizing channel. Under local depolarizing noise, the quantum state is progressively contracted layer-by-layer toward the maximally mixed state, denoted as I/d. As a state approaches this maximally mixed distribution, it loses all of its extractable quantum information, appearing indistinguishable from random thermal noise. Other common forms of noise include amplitude damping (loss of energy to the environment) and dephasing (loss of relative phase information between qubits).

B Dually Flat Stratified Quantum Natural Gradient

In this section we show how DFS-QNG eliminates retraction errors. Recall equation 10:

$$\boldsymbol{\mathcal{E}}^{(n,k)} \;=\; \frac{\eta^2}{2} \sum_{(l,i),(m,j)} \Gamma^{(n,k)}_{(l,i),(m,j)}\, \mathbf{v}^{(l,i)}\, \mathbf{v}^{(m,j)} \;+\; \mathcal{O}(\eta^3)\,.$$

where $\mathbf{v}$ is the update direction and $\Gamma^{(n,k)}_{(l,i),(m,j)}$ are the Christoffel symbols of the Levi-Civita connection. Because a functioning optimizer inherently requires non-zero parameter updates ($\mathbf{v} \neq 0$), the only way to eliminate this retraction error is to operate in a geometric regime where the Christoffel symbols themselves vanish identically ($\Gamma^{(n,k)}_{(l,i),(m,j)} = 0$).

To understand how to achieve this, we introduce the formal definition of the Christoffel symbols. To accurately reflect the structure of a quantum circuit, we now adopt a strict $(layer, parameter)$ multi-index notation. Let a parameter be denoted by $\boldsymbol{\theta}_{l,i}$, where l indexes the circuit layer and i indexes the specific gate within that layer. The corresponding Hermitian generator is $\mathfrak{g}_{l,i}$. The Christoffel symbols in this notation are:

$$\Gamma^{(n,k)}_{(l,i),(m,j)} = \frac{1}{2} \sum_{(p,r)} (\tilde{g}^{-1})_{(n,k),(p,r)} \left(\partial_{(l,i)} \tilde{g}_{(m,j),(p,r)} + \partial_{(m,j)} \tilde{g}_{(l,i),(p,r)} - \partial_{(p,r)} \tilde{g}_{(l,i),(m,j)} \right)$$

To force $\Gamma^{(n,k)}_{(l,i),(m,j)} = 0$, we must analyze the partial derivatives of the Fubini-Study metric tensor. The derivative of a metric component $\tilde{g}_{(l,i),(m,j)}$ with respect to a third parameter $\boldsymbol{\theta}_{n,k}$ expands into three distinct inner-product terms:

$$\partial_{(n,k)}\tilde{g}_{(l,i),(m,j)} = \underbrace{\left\langle \partial^2_{(n,k),(l,i)}\psi_{\boldsymbol{\theta}}, \partial_{(m,j)}\psi_{\boldsymbol{\theta}} \right\rangle + \left\langle \partial_{(l,i)}\psi_{\boldsymbol{\theta}}, \partial^2_{(n,k),(m,j)}\psi_{\boldsymbol{\theta}} \right\rangle}_{\text{Term 1}} \tag{12}$$

$$- \underbrace{\left[\left\langle \partial^2_{(n,k),(l,i)}\psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}} \right\rangle + \left\langle \partial_{(l,i)}\psi_{\boldsymbol{\theta}}, \partial_{(n,k)}\psi_{\boldsymbol{\theta}} \right\rangle \right] \left\langle \psi_{\boldsymbol{\theta}}, \partial_{(m,j)}\psi_{\boldsymbol{\theta}} \right\rangle}_{\text{Term 2}} \tag{13}$$

$$- \underbrace{\left\langle \partial_{(l,i)}\psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}} \right\rangle \left[\left\langle \partial_{(n,k)}\psi_{\boldsymbol{\theta}}, \partial_{(m,j)}\psi_{\boldsymbol{\theta}} \right\rangle + \left\langle \psi_{\boldsymbol{\theta}}, \partial^2_{(n,k),(m,j)}\psi_{\boldsymbol{\theta}} \right\rangle \right]}_{\text{Term 3}} \tag{14}$$

Proof. We prove Proposition 1. First, we analyze the derivatives of the state vector. In a generic circuit, the derivative involves the insertion of a generator amidst surrounding unitaries. However, if we restrict our active optimization step to a single stratum where all gates in the active unitary block U commute (i.e., $AB = BA$ for all $A, B \in U$), the state derivatives simplify dramatically. The first derivative with respect to $\theta_{l,i}$ becomes:

$$\left|\partial_{(l,i)}\psi_{\boldsymbol{\theta}}\right\rangle = (-i\mathfrak{g}_{l,i})\left|\psi_{\boldsymbol{\theta}}\right\rangle$$

Similarly, the second mixed partial derivative is symmetric and straightforward:

$$\left|\partial^2_{(n,k),(l,i)}\psi_{\boldsymbol{\theta}}\right\rangle = -\mathfrak{g}_{n,k}\mathfrak{g}_{l,i}\left|\psi_{\boldsymbol{\theta}}\right\rangle$$

Substituting these exact forms into the three metric derivative terms, we can evaluate each one individually.

Term 1:

$$\begin{aligned}
&\langle \partial^2_{(n,k),(l,i)}\psi_{\boldsymbol{\theta}} | \partial_{(m,j)}\psi_{\boldsymbol{\theta}}\rangle + \langle \partial_{(l,i)}\psi_{\boldsymbol{\theta}} | \partial^2_{(n,k),(m,j)}\psi_{\boldsymbol{\theta}}\rangle \\
&= \langle \psi_{\boldsymbol{\theta}} | (-\mathfrak{g}_{n,k}\mathfrak{g}_{l,i})(-i\mathfrak{g}_{m,j})|\psi_{\boldsymbol{\theta}}\rangle + \langle \psi_{\boldsymbol{\theta}} | (i\mathfrak{g}_{l,i})(-\mathfrak{g}_{n,k}\mathfrak{g}_{m,j})|\psi_{\boldsymbol{\theta}}\rangle \\
&= i\langle \psi_{\boldsymbol{\theta}} | \mathfrak{g}_{n,k}\mathfrak{g}_{l,i}\mathfrak{g}_{m,j}|\psi_{\boldsymbol{\theta}}\rangle - i\langle \psi_{\boldsymbol{\theta}} | \mathfrak{g}_{l,i}\mathfrak{g}_{n,k}\mathfrak{g}_{m,j}|\psi_{\boldsymbol{\theta}}\rangle
\end{aligned}$$

Since all generators in the stratum commute, we have $\mathfrak{g}_{n,k}\mathfrak{g}_{l,i} = \mathfrak{g}_{l,i}\mathfrak{g}_{n,k}$. Therefore, Term 1 cleanly cancels:

$$= i\langle \mathfrak{g}_{l,i}\mathfrak{g}_{n,k}\mathfrak{g}_{m,j}\rangle - i\langle \mathfrak{g}_{l,i}\mathfrak{g}_{n,k}\mathfrak{g}_{m,j}\rangle = 0$$

Term 2: We evaluate the bracketed inner portion first:

$$\begin{aligned}
&\langle \partial^2_{(n,k),(l,i)}\psi_{\boldsymbol{\theta}} | \psi_{\boldsymbol{\theta}}\rangle + \langle \partial_{(l,i)}\psi_{\boldsymbol{\theta}} | \partial_{(n,k)}\psi_{\boldsymbol{\theta}}\rangle \\
&= \langle \psi_{\boldsymbol{\theta}} | (-\mathfrak{g}_{n,k}\mathfrak{g}_{l,i})|\psi_{\boldsymbol{\theta}}\rangle + \langle \psi_{\boldsymbol{\theta}} | (i\mathfrak{g}_{l,i})(-i\mathfrak{g}_{n,k})|\psi_{\boldsymbol{\theta}}\rangle \\
&= -\langle \psi_{\boldsymbol{\theta}} | \mathfrak{g}_{n,k}\mathfrak{g}_{l,i}|\psi_{\boldsymbol{\theta}}\rangle + \langle \psi_{\boldsymbol{\theta}} | \mathfrak{g}_{l,i}\mathfrak{g}_{n,k}|\psi_{\boldsymbol{\theta}}\rangle
\end{aligned}$$

Again, substituting the commutation relation $\mathfrak{g}_{n,k}\mathfrak{g}_{l,i} = \mathfrak{g}_{l,i}\mathfrak{g}_{n,k}$ yields:

$$= -\langle \mathfrak{g}_{l,i}\mathfrak{g}_{n,k}\rangle + \langle \mathfrak{g}_{l,i}\mathfrak{g}_{n,k}\rangle = 0$$

Because the inner bracket is zero, the entirety of Term 2 vanishes.

Term 3: We evaluate the bracketed inner portion first:

$$\begin{aligned}
&\langle \partial_{(n,k)}\psi_{\boldsymbol{\theta}} | \partial_{(m,j)}\psi_{\boldsymbol{\theta}}\rangle + \langle \psi_{\boldsymbol{\theta}} | \partial^2_{(n,k),(m,j)}\psi_{\boldsymbol{\theta}}\rangle \\
&= \langle \psi_{\boldsymbol{\theta}} | (i\mathfrak{g}_{n,k})(-i\mathfrak{g}_{m,j})|\psi_{\boldsymbol{\theta}}\rangle + \langle \psi_{\boldsymbol{\theta}} | (-\mathfrak{g}_{n,k}\mathfrak{g}_{m,j})|\psi_{\boldsymbol{\theta}}\rangle \\
&= \langle \psi_{\boldsymbol{\theta}} | \mathfrak{g}_{n,k}\mathfrak{g}_{m,j}|\psi_{\boldsymbol{\theta}}\rangle - \langle \psi_{\boldsymbol{\theta}} | \mathfrak{g}_{n,k}\mathfrak{g}_{m,j}|\psi_{\boldsymbol{\theta}}\rangle = 0
\end{aligned}$$

Because the inner bracket is zero, Term 3 also vanishes.

Since all three terms are exactly zero, $\partial_{(n,k)}\tilde{g}_{(l,i),(m,j)} = 0$ everywhere within the stratum. Consequently, the Christoffel symbols evaluate to zero ($\Gamma^{(n,k)}_{(l,i),(m,j)} = 0$), proving that the localized manifold is dually flat and the retraction error $\boldsymbol{\mathcal{E}}$ is completely eliminated. □

DUALLY FLAT SUBMANIFOLD

To circumvent the cubic curvature penalty induced by the Riemannian Hessian, we partition the variational parameters into $\mathfrak{S}$ disjoint subsets, or strata, denoted by $\Theta = \bigcup_{\pi=1}^{\mathfrak{S}} \Theta_\pi$. A stratum Θ_π is defined such that all generators associated with its parameters are mutually commuting:

$$[\mathfrak{g}_{\pi,i}, \mathfrak{g}_{\pi,j}] = 0 \quad \forall i, j \in \Theta_\pi$$

Recall from Definition 1, the global unitary evolution of the ansatz can be factorized over these strata:

$$|\psi_{\boldsymbol{\theta}}\rangle = \left(\prod_{\pi=1}^{\mathfrak{S}} \mathcal{U}_\pi(\boldsymbol{\theta}_\pi) \mathcal{W}_\pi \right) |\psi_0\rangle \quad \text{where} \quad \mathcal{U}_\pi(\boldsymbol{\theta}_\pi) = \exp\left(-i \sum_{i \in \Theta_\pi} \theta_{\pi,i} \mathfrak{g}_{\pi,i} \right)$$

Let us isolate an active block Θ_π. We can express the global state by splitting the unitary sequence into the operations preceding the block, the block itself, and the operations succeeding it:

$$|\psi_{\boldsymbol{\theta}}\rangle = \mathcal{U}_{[\mathfrak{S}:\pi+1]}\, \mathcal{U}_\pi(\boldsymbol{\theta}_\pi)\, |\psi_{\pi-1}\rangle$$

where $|\psi_{\pi-1}\rangle$ is the fixed state prior to block π, and $\mathcal{U}_{[\mathfrak{S}:\pi+1]}$ is the product of all subsequent unitaries.

B.1 BLOCK-RESTRICTED STATE DERIVATIVES

When computing the gradient and metric tensor exclusively for the parameters within the active block Θ_π, we take the derivative with respect to $\theta_{\pi,i} \in \Theta_\pi$. Because all generators within $\mathcal{U}_\pi$ commute, the derivative simplifies cleanly:

$$|\partial_{\pi,i} \psi_{\boldsymbol{\theta}}\rangle = \mathcal{U}_{[\mathfrak{S}:\pi+1]}(-i\mathfrak{g}_{\pi,i}) \mathcal{U}_\pi |\psi_{\pi-1}\rangle = \mathcal{U}_{[\mathfrak{S}:\pi+1]}(-i\mathfrak{g}_{\pi,i}) |\psi_\pi\rangle$$

where $|\psi_\pi\rangle = \mathcal{U}_\pi |\psi_{\pi-1}\rangle$ is the state immediately after the active block. Similarly, the second derivative for $i, j \in \Theta_\pi$ becomes:

$$|\partial^2_{\pi,i,\pi,j} \psi_{\boldsymbol{\theta}}\rangle = \mathcal{U}_{[\mathfrak{S}:\pi+1]}(-\mathfrak{g}_{\pi,i}\mathfrak{g}_{\pi,j}) |\psi_\pi\rangle$$

A critical property emerges when we compute the inner products required for the Fubini-Study metric $\tilde{g}_{\pi,i,\pi,j}$. The trailing unitary $\mathcal{U}_{[\mathfrak{S}:\pi+1]}$ is isometric ($\mathcal{U}^\dagger_{[\mathfrak{S}:\pi+1]} \mathcal{U}_{[\mathfrak{S}:\pi+1]} = I$) and cancels out entirely:

$$\langle \partial_{\pi,i} \psi_{\boldsymbol{\theta}} | \partial_{\pi,j} \psi_{\boldsymbol{\theta}} \rangle = \langle \psi_\pi | (i\mathfrak{g}_{\pi,i}) \mathcal{U}^\dagger_{[\mathfrak{S}:\pi+1]} \mathcal{U}_{[\mathfrak{S}:\pi+1]} (-i\mathfrak{g}_{\pi,j}) | \psi_\pi \rangle = \langle \psi_\pi | \mathfrak{g}_{\pi,i} \mathfrak{g}_{\pi,j} | \psi_\pi \rangle$$

Thus, for parameters within the same block, the metric tensor and its derivatives depend exclusively on the local state $|\psi_\pi\rangle$ and the mutually commuting generators $\mathfrak{g}_{\pi,i}, \mathfrak{g}_{\pi,j} \in \Theta_\pi$.

VANISHING CHRISTOFFEL SYMBOLS WITHIN A STRATUM

The Christoffel symbols governing the curvature of the optimization landscape are given by:

$$\Gamma^{\pi,k}_{\pi,i,\pi,j} = \frac{1}{2} \sum_r (\tilde{g}^{-1})_{\pi,k,\pi,r} \left(\partial_{\pi,i} \tilde{g}_{\pi,j,\pi,r} + \partial_{\pi,j} \tilde{g}_{\pi,i,\pi,r} - \partial_{\pi,r} \tilde{g}_{\pi,i,\pi,j} \right)$$

To achieve a dually flat landscape, we require $\Gamma^{\pi,k}_{\pi,i,\pi,j} = 0$, which necessitates that $\partial_{\pi,k} \tilde{g}_{\pi,i,\pi,j} = 0$. Expanding the derivative of the metric tensor yields three terms. Assuming $i, j, k \in \Theta_\pi$, we

substitute our block-restricted derivatives. Leveraging the commutation relation $[\mathfrak{g}_a, \mathfrak{g}_b] = 0$, we evaluate each term on the effective local state $|\psi_\pi\rangle$:

Term 1:

$$= \langle\psi_\pi|(-\mathfrak{g}_{\pi,k}\mathfrak{g}_{\pi,i})(-i\mathfrak{g}_{\pi,j})|\psi_\pi\rangle + \langle\psi_\pi|(i\mathfrak{g}_{\pi,i})(-\mathfrak{g}_{\pi,k}\mathfrak{g}_{\pi,j})|\psi_\pi\rangle$$
$$= i\langle\psi_\pi|\mathfrak{g}_{\pi,k}\mathfrak{g}_{\pi,i}\mathfrak{g}_{\pi,j}|\psi_\pi\rangle - i\langle\psi_\pi|\mathfrak{g}_{\pi,i}\mathfrak{g}_{\pi,k}\mathfrak{g}_{\pi,j}|\psi_\pi\rangle = 0$$

Term 2:

$$= [\langle\psi_\pi|(-\mathfrak{g}_{\pi,k}\mathfrak{g}_{\pi,i})|\psi_\pi\rangle + \langle\psi_\pi|(i\mathfrak{g}_{\pi,i})(-i\mathfrak{g}_{\pi,k})|\psi_\pi\rangle]\,\langle\psi_\pi| - i\mathfrak{g}_{\pi,j}|\psi_\pi\rangle$$
$$= [-\langle\psi_\pi|\mathfrak{g}_{\pi,k}\mathfrak{g}_{\pi,i}|\psi_\pi\rangle + \langle\psi_\pi|\mathfrak{g}_{\pi,i}\mathfrak{g}_{\pi,k}|\psi_\pi\rangle]\,\langle\psi_\pi| - i\mathfrak{g}_{\pi,j}|\psi_\pi\rangle = 0$$

Term 3:

$$= \langle\psi_\pi|i\mathfrak{g}_{\pi,i}|\psi_\pi\rangle\,[\langle\psi_\pi|(i\mathfrak{g}_{\pi,k})(-i\mathfrak{g}_{\pi,j})|\psi_\pi\rangle + \langle\psi_\pi|(-\mathfrak{g}_{\pi,k}\mathfrak{g}_{\pi,j})|\psi_\pi\rangle]$$
$$= \langle\psi_\pi|i\mathfrak{g}_{\pi,i}|\psi_\pi\rangle\,[\langle\psi_\pi|\mathfrak{g}_{\pi,k}\mathfrak{g}_{\pi,j}|\psi_\pi\rangle - \langle\psi_\pi|\mathfrak{g}_{\pi,k}\mathfrak{g}_{\pi,j}|\psi_\pi\rangle] = 0$$

Since Terms 1, 2, and 3 vanish identically, $\partial_{\pi,k}\tilde{g}_{\pi,i,\pi,j} = 0$ for all internal stratum parameters, confirming that the block-diagonal sub-tensor $G_{\pi\pi}$ is perfectly flat ($\Gamma^{(\pi)}_{\pi,i,\pi,j} = 0$).

C DFS-QNG Convergence Analysis

Before establishing the formal convergence rate, we highlight that DFS-QNG structurally simplifies the global optimization problem into block-wise Riemannian gradient descent. Because our stratification approach strictly confines each update step to a commuting set of generators, the local geometry reduces to a dually flat submanifold. On these strata, the Riemannian exponential map is exact and coincides with the coordinate update. Therefore, our proof hinges on applying the Riemannian Descent Lemma sequentially to each block-restricted manifold, bypassing the global cross-strata curvature terms that otherwise destabilize standard QNG.

Proof. We prove Theorem 3.

Assumptions and Setup. To analyze the convergence in the deterministic regime, we rely on the following established properties:

1. **Global Geodesic Smoothness:** The objective $\mathcal{L}(\boldsymbol{\theta})$ is L_R-smooth on the global manifold. We prove this assumption in Theorem 1.
2. **Abelian Stratification:** The parameters are partitioned into $\mathfrak{S}$ mutually disjoint commuting strata.
3. **Local Dually Flat Geodesics:** For each stratum Θ_π, the restricted submanifold is $\nabla^{(e)}$-flat, and the Riemannian exponential map coincides exactly with the coordinate update. This means our ansatz follow proposition 1.

Per-Stratum Descent. Because DFS-QNG freezes all other parameters while updating a given stratum, each sub-step is a strictly local operation on the submanifold defined by Θ_π. Let $\boldsymbol{\theta}^{(t,\pi)}$ denote the parameter vector at epoch t after updating strata up to π. The boundary conditions for a full epoch are $\boldsymbol{\theta}^{(t,0)} = \boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\theta}^{(t,\mathfrak{S})} = \boldsymbol{\theta}^{(t+1)}$. The update rule for the active stratum π is exactly the Riemannian exponential map on the restricted geometry:

$$\boldsymbol{\theta}^{(t,\pi)} = \mathrm{Exp}_{\boldsymbol{\theta}^{(t,\pi-1)}}\left(-\eta \mathrm{grad}_{g^{(\epsilon)}_\pi}\mathcal{L}\right)$$

Because the global objective is L_R-smooth, its restriction to any totally geodesic submanifold is at most L_R-smooth. Setting the step size to $\eta = 1/\mathsf{L}_\mathrm{R}$, we can apply the standard Riemannian Descent Lemma directly to this localized step:

$$\mathcal{L}(\boldsymbol{\theta}^{(t,\pi)}) \leq \mathcal{L}(\boldsymbol{\theta}^{(t,\pi-1)}) - \frac{1}{2\mathsf{L}_\mathrm{R}}\|\mathrm{grad}_{g^{(\epsilon)}_\pi}\mathcal{L}\|^2_{g^{(\epsilon)}_\pi}$$

Epoch-Level Descent. To find the total descent over a single full epoch of DFS-QNG, we sum the guaranteed decreases from each of the $\mathfrak{S}$ strata. By telescoping the intermediate states $\boldsymbol{\theta}^{(t,\pi)}$ from $\pi = 1$ to $\mathfrak{S}$, we obtain:

$$\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) \leq \mathcal{L}(\boldsymbol{\theta}^{(t)}) - \frac{1}{2\mathsf{L}_{\mathrm{R}}} \sum_{\pi=1}^{\mathfrak{S}} \|\mathrm{grad}_{g_\pi^{(\epsilon)}} \mathcal{L}\|^2_{g_\pi^{(\epsilon)}}$$

This inequality demonstrates that DFS-QNG is strictly monotonically decreasing, provided the step size satisfies $\eta \leq 1/\mathsf{L}_{\mathrm{R}}$.

Global Convergence Rate. To establish the standard sublinear convergence rate to a stationary point, we sum the epoch-level descent over T total epochs:

$$\sum_{t=0}^{T-1} \sum_{\pi=1}^{\mathfrak{S}} \|\mathrm{grad}_{g_\pi^{(\epsilon)}} \mathcal{L}\|^2_{g_\pi^{(\epsilon)}} \leq 2\mathsf{L}_{\mathrm{R}} \left(\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}(\boldsymbol{\theta}^{(T)}) \right) \leq 2\mathsf{L}_{\mathrm{R}} \left(\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}^* \right)$$

where $\mathcal{L}^*$ is the global minimum of the objective. Dividing by the total number of epochs T, we bound the time-averaged sum of the squared stratum gradients:

$$\frac{1}{T} \sum_{t=0}^{T-1} \sum_{\pi=1}^{\mathfrak{S}} \|\mathrm{grad}_{g_\pi^{(\epsilon)}} \mathcal{L}\|^2_{g_\pi^{(\epsilon)}} \leq \frac{2\mathsf{L}_{\mathrm{R}} \left(\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}^* \right)}{T}$$

Therefore, to achieve a δ-stationary point—where the time-averaged sum of the squared stratum gradients is less than or equal to δ—DFS-QNG requires at most:

$$T = \mathcal{O}\left(\frac{\mathsf{L}_{\mathrm{R}}}{\delta} \right)$$

iterations. □

D Riemann Hessian Bound

This section provides the formal proof for Theorem 1 (Geodesic Smoothness), establishing the foundational prerequisite for exact Riemannian gradient descent on a quantum variational landscape to behave predictably and converge. The overall strategy relies on decomposing the covariant Hessian into its coordinate and connection-correction terms, bounding each through a sequence of Cauchy-Schwarz estimates on the state derivatives and their second-order jets. By defining a canonical norm vector built directly from the generator set, we ensure every bounding operator factorizes as a rank-one outer product. This structural "rank-1 trick" collapses the complex matrix spectral norm computations into simple dot products, yielding closed-form control over the tensor without ever needing to directly integrate the highly non-linear inverse metric.

Proof. We seek to bound the spectral norm of the Riemannian Hessian, denoted as $\|\mathrm{Hess}_\nabla\|_2$. By Lemma 5, the magnitude of each element $\mathrm{Hess}_{\nabla(l,i),(m,j)}$ is bounded by:

$$|\mathrm{Hess}_{\nabla(l,i),(m,j)}| \leq 4\|\hat{O}\|_2 \|\mathfrak{g}_{(l,i)}\|_2 \|\mathfrak{g}_{(m,j)}\|_2 + \frac{18}{\epsilon} \|\hat{O}\|_2 \|\mathfrak{g}_{(l,i)}\|_2 \|\mathfrak{g}_{(m,j)}\|_2 \left(\sum_{(n,k)} \|\mathfrak{g}_{(n,k)}\|_2 \right)^2$$

Let us define a matrix B such that $B_{(l,i),(m,j)}$ corresponds to the right-hand side of the inequality above. Since:

$$|\mathrm{Hess}_{\nabla(l,i),(m,j)}| \leq B_{(l,i),(m,j)} \quad \forall (l,i),(m,j) \quad \text{where} \quad |\mathrm{Hess}_{\nabla(l,i),(m,j)}|, B_{(l,i),(m,j)} \in \mathbb{R}^+$$

it follows that $\|\mathrm{Hess}_\nabla\|_2 \leq \|B\|_2$. To analyze the structure of B, we define a vector $\mathbf{v} \in \mathbb{R}^P$ containing the ℓ_2 norms of the generators:

$$\mathbf{v} = \begin{bmatrix} \|\mathfrak{g}_{(1,1)}\|_2 \\ \vdots \\ \|\mathfrak{g}_{(L,M_L)}\|_2 \end{bmatrix}$$

Substituting this vector into the expression for $B_{(l,i),(m,j)}$, we observe a common factor of $v_{(l,i)}v_{(m,j)} = \|\mathfrak{g}_{(l,i)}\|_2\|\mathfrak{g}_{(m,j)}\|_2$ in both terms:

$$B_{(l,i),(m,j)} = \left[4\|\hat{O}\|_2 + \frac{18}{\epsilon}\|\hat{O}\|_2 \left(\sum_{(n,k)} v_{(n,k)}\right)^2\right] v_{(l,i)}v_{(m,j)}$$

Let the scalar term in the brackets be denoted by C:

$$C = \|\hat{O}\|_2 \left[4 + \frac{18}{\epsilon}\left(\sum_{(n,k)} \|\mathfrak{g}_{(n,k)}\|_2\right)^2\right]$$

We can now express the matrix B as a scaled outer product of the vector $\mathbf{v}$:

$$B = C(\mathbf{v}\mathbf{v}^T)$$

This reveals that B is a Rank-1 symmetric matrix. The spectral norm of a Rank-1 matrix of the form $c\mathbf{u}\mathbf{u}^T$ is given by $|c|\|\mathbf{u}\|_2^2$. Therefore:

$$\|B\|_2 = C\|\mathbf{v}\|_2^2 = C\left(\sum_{(n,k)} \|\mathfrak{g}_{(n,k)}\|_2^2\right)$$

Substituting the explicit form of C back into the equation yields the final bound:

$$\|\text{Hess}_\nabla\|_2 \leq B_R = \|\hat{O}\|_2 \left[4 + \frac{18}{\epsilon}\left(\sum_{(n,k)} \|\mathfrak{g}_{(n,k)}\|_2\right)^2\right]\left(\sum_{(n,k)} \|\mathfrak{g}_{(n,k)}\|_2^2\right)$$

By definition of Riemannian smoothness, we require $u^T\text{Hess}_\nabla u \leq \|\text{Hess}_\nabla\|_2\|u\|_{g^{(\epsilon)}}^2$. Since the regularized metric satisfies $\|u\|_{g^{(\epsilon)}}^2 \geq \epsilon\|u\|_2^2$, we can bound the Riemannian smoothness constant using the Euclidean spectral norm from Theorem 1: $\mathsf{L}_\text{R} \leq \frac{\|\text{Hess}_\nabla\|_2}{\epsilon}$. Thus we find the Riemann smoothness constant:

$$\mathsf{L}_\text{R} = \frac{\|\hat{O}\|_2}{\epsilon}\left[4 + \frac{18}{\epsilon}\left(\sum_{(n,k)} \|\mathfrak{g}_{(n,k)}\|_2\right)^2\right]\left(\sum_{(n,k)} \|\mathfrak{g}_{(n,k)}\|_2^2\right)$$

□

E Riemannian Gradient Descent Convergence Rate

This section provides the formal proof for Corollary 1, establishing what is, to our knowledge, the first provable convergence guarantee for Riemannian gradient descent on a quantum variational landscape. Having established the global geodesic smoothness of the manifold in Theorem 1, we can

now mathematically justify the use of the Riemannian Descent Lemma. By propagating parameter updates along true intrinsic geodesics rather than standard Euclidean retractions, we derive a strict upper bound on the objective value and guarantee sublinear convergence to a stationary point.

Proof. For an objective function $\mathcal{L}$ that is L_{R}-smooth on a Riemannian manifold, its Riemannian Hessian spectral norm is bounded everywhere by L_{R}. The Riemannian Descent Lemma guarantees that taking a step along a true geodesic $\gamma(t)$ yields the following upper bound on the objective value:

$$\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) \leq \mathcal{L}(\boldsymbol{\theta}^{(t)}) + \langle \mathrm{grad}_{g^{(\epsilon)}} \mathcal{L}(\boldsymbol{\theta}^{(t)}), \mathrm{Exp}^{-1}_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\theta}^{(t+1)}) \rangle_{g^{(\epsilon)}} + \frac{\mathsf{L}_{\mathrm{R}}}{2} \| \mathrm{Exp}^{-1}_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\theta}^{(t+1)}) \|^2_{g^{(\epsilon)}}$$

Geodesic-QNG updates the parameters directly along the geodesics of the Riemannian manifold using the exponential map:

$$\boldsymbol{\theta}^{(t+1)} = \mathrm{Exp}_{\boldsymbol{\theta}^{(t)}}(-\eta \mathrm{grad}_{g^{(\epsilon)}} \mathcal{L})$$

In Riemannian terms, the vector $\vec{\mathbf{v}} = -(g^{(\epsilon)})^{-1} d\mathcal{L}$ is exactly the negative Riemannian gradient: $-\mathrm{grad}_{g^{(\epsilon)}} \mathcal{L}$. Therefore, the update step length mapped back to the tangent space is:

$$\mathrm{Exp}^{-1}_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\theta}^{(t+1)}) = -\eta \mathrm{grad}_{g^{(\epsilon)}} \mathcal{L}(\boldsymbol{\theta}^{(t)})$$

Substitute this exact update step into our descent lemma from Step 1:

$$\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) \leq \mathcal{L}(\boldsymbol{\theta}^{(t)}) + \langle \mathrm{grad}_{g^{(\epsilon)}} \mathcal{L}, -\eta \mathrm{grad}_{g^{(\epsilon)}} \mathcal{L} \rangle_{g^{(\epsilon)}} + \frac{\mathsf{L}_{\mathrm{R}}}{2} \| - \eta \mathrm{grad}_{g^{(\epsilon)}} \mathcal{L} \|^2_{g^{(\epsilon)}}$$

Simplify the inner product and norms:

$$\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) \leq \mathcal{L}(\boldsymbol{\theta}^{(t)}) - \eta \| \mathrm{grad}_{g^{(\epsilon)}} \mathcal{L} \|^2_{g^{(\epsilon)}} + \frac{\eta^2 \mathsf{L}_{\mathrm{R}}}{2} \| \mathrm{grad}_{g^{(\epsilon)}} \mathcal{L} \|^2_{g^{(\epsilon)}}$$

$$\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) \leq \mathcal{L}(\boldsymbol{\theta}^{(t)}) - \eta \left(1 - \frac{\eta \mathsf{L}_{\mathrm{R}}}{2}\right) \| \mathrm{grad}_{g^{(\epsilon)}} \mathcal{L} \|^2_{g^{(\epsilon)}}$$

To guarantee monotonic descent and maximize the decrease in the objective function, we want to maximize the term $\eta(1 - \frac{\eta \mathsf{L}_{\mathrm{R}}}{2})$. Taking the derivative with respect to η and setting it to zero gives the optimal constant step size:

$$\eta = \frac{1}{\mathsf{L}_{\mathrm{R}}}$$

Substituting $\eta = \frac{1}{\mathsf{L}_{\mathrm{R}}}$ back into the inequality:

$$\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) \leq \mathcal{L}(\boldsymbol{\theta}^{(t)}) - \frac{1}{2\mathsf{L}_{\mathrm{R}}} \| \mathrm{grad}_{g^{(\epsilon)}} \mathcal{L} \|^2_{g^{(\epsilon)}}$$

Rearrange the inequality to isolate the gradient norm:

$$\| \mathrm{grad}_{g^{(\epsilon)}} \mathcal{L} \|^2_{g^{(\epsilon)}} \leq 2\mathsf{L}_{\mathrm{R}} (\mathcal{L}(\boldsymbol{\theta}^{(t)}) - \mathcal{L}(\boldsymbol{\theta}^{(t+1)}))$$

Summing this inequality over T iterations from $t = 0$ to $T - 1$, the intermediate terms telescope and cancel out:

$$\sum_{t=0}^{T-1} \| \mathrm{grad}_{g^{(\epsilon)}} \mathcal{L} \|^2_{g^{(\epsilon)}} \leq 2\mathsf{L}_{\mathrm{R}} (\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}(\boldsymbol{\theta}^{(T)})) \leq 2\mathsf{L}_{\mathrm{R}} (\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}^*)$$

(where $\mathcal{L}^*$ is the global minimum) . To find the average gradient norm over T iterations, we divide by T:

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\text{grad}_{g^{(\epsilon)}}\mathcal{L}\|^2_{g^{(\epsilon)}} \leq \frac{2\mathsf{L}_{\mathrm{R}}(\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}^*)}{T}$$

To achieve a δ-stationary point, where the average squared gradient norm is less than or equal to δ, we require:

$$T \geq \frac{2\mathsf{L}_{\mathrm{R}}(\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}^*)}{\delta}$$

Finally, we use the specific properties of the Quantum Natural Gradient. Theorem 1 establishes that for a regularized metric $g^{(\epsilon)}$, the Riemannian smoothness constant $\mathsf{L}_{\mathrm{R}}(\epsilon)$ scales cubically with the total parameter count P. Specifically, $\mathsf{L}_{\mathrm{R}}(\epsilon) \in \Theta(P^3)$. Substituting this scaling into our required iteration count directly yields the convergence rate of GQNG (Theorem 2):

$$T = \mathcal{O}\left(\frac{\mathsf{L}_{\mathrm{R}}}{\delta}\right)$$

□

F Useful Lemmas and Upper Bounds

This section establishes the fundamental deterministic upper bounds on the geometric objects defining the quantum statistical manifold, specifically the state derivatives, metric tensor components, connection coefficients, and both the Euclidean and Riemannian Hessians. By tightly bounding these quantities exclusively in terms of the operator norms of the observable $\hat{O}$ and the Hermitian generators $\mathfrak{g}_{(l,k)}$, we successfully decouple the underlying differential geometry of the manifold from the stochastic nature of the variational parameter landscape.

This decoupling is the mathematical engine that makes the subsequent proofs possible. It allows us to establish the global geodesic smoothness of the landscape (Theorem 1) and later find the convergence rate of both DFS-QNG and RGD.

Lemma 1. *For a variational objective of the form $\mathcal{L}(\boldsymbol{\theta}) = \langle\psi_{\boldsymbol{\theta}}|\hat{O}|\psi_{\boldsymbol{\theta}}\rangle$, the magnitude of the partial derivative with respect to coordinate $\boldsymbol{\theta}_{(l,i)}$ is strictly bounded by:*

$$|\partial_{(l,i)}\mathcal{L}| \leq 2\|\hat{O}\|_2\|\mathfrak{g}_{(l,i)}\|_2$$

Lemma 2. *The partial derivative of the regularized Fubini-Study metric element $g^{(\epsilon)}_{(l,i),(m,j)}$ with respect to parameter $\theta_{(n,k)}$ is bounded by:*

$$\left|\partial_{(n,k)}g^{(\epsilon)}_{(l,i),(m,j)}\right| \leq 6\|\mathfrak{g}_{(l,i)}\|_2\|\mathfrak{g}_{(m,j)}\|_2\|\mathfrak{g}_{(n,k)}\|_2$$

Lemma 3. *The magnitude of the Levi-Civita connection coefficients (Christoffel symbols) for the regularized metric is bounded by:*

$$|\Gamma^{(n,k)}_{(l,i),(m,j)}| \leq \frac{9}{\epsilon}\|\mathfrak{g}_{(l,i)}\|_2\|\mathfrak{g}_{(m,j)}\|_2\sum_{(p,r)}\|\mathfrak{g}_{(p,r)}\|_2$$

Lemma 4. *The elements of the Euclidean Hessian matrix $Hess$ are bounded by:*

$$|\mathit{Hess}_{(l,i),(m,j)}| \leq 4\|\hat{O}\|_2\|\mathfrak{g}_{(l,i)}\|_2\|\mathfrak{g}_{(m,j)}\|_2$$

Lemma 5. *The elements of the Riemannian Hessian matrix $Hess_{\nabla}$ are bounded by:*

$$|Hess_{\nabla(l,i),(m,j)}| \leq 4\|\hat{O}\|_2\|\mathfrak{g}_{(l,i)}\|_2\|\mathfrak{g}_{(m,j)}\|_2 + \frac{18}{\epsilon}\|\hat{O}\|_2\|\mathfrak{g}_{(l,i)}\|_2\|\mathfrak{g}_{(m,j)}\|_2 \left(\sum_{(n,k)} \|\mathfrak{g}_{(n,k)}\|_2\right)^2$$

Lemma 6. *The ℓ_2-norm of the standard Euclidean gradient (exterior derivative) $d\mathcal{L}$ is bounded by:*

$$\|d\mathcal{L}\|_2 \leq 2\|\hat{O}\|_2\sqrt{\sum_{(l,i)} \|\mathfrak{g}_{(l,i)}\|_2^2}$$

Lemma 7. *The spectral norm of the matrix $\Gamma^{(n,k)}$ associated with a fixed upper index (n,k) is bounded by:*

$$\|\Gamma^{(n,k)}\|_2 \leq \frac{9}{\epsilon}\left(\sum_{(l,i)} \|\mathfrak{g}_{(l,i)}\|_2\right)\left(\sum_{(m,j)} \|\mathfrak{g}_{(m,j)}\|_2^2\right)$$

Lemma 8. *The spectral norm of the Euclidean Hessian is bounded by:*

$$\|Hess\|_2 \leq \mathsf{L}_{\mathrm{E}} = 4\|\hat{O}\|_2 \sum_{(l,i)} \|\mathfrak{g}_{(l,i)}\|_2^2$$

LEMMA 1 PROOF

Proof. We will prove Lemma 1. Given the objective $\mathcal{L} = \langle\psi_{\boldsymbol{\theta}}|\hat{O}|\psi_{\boldsymbol{\theta}}\rangle$, we analyze the partial derivative $\partial_{(l,i)}\mathcal{L}$.

Expanding the derivative yields:

$$|\partial_{(l,i)}\mathcal{L}| \leq |\langle\partial_{(l,i)}\psi_{\boldsymbol{\theta}}|\hat{O}|\psi_{\boldsymbol{\theta}}\rangle| + |\langle\psi_{\boldsymbol{\theta}}|\hat{O}|\partial_{(l,i)}\psi_{\boldsymbol{\theta}}\rangle|$$

Because $\hat{O}$ is a Hermitian operator, the first term is the complex conjugate of the second. Specifically, $\langle\partial_{(l,i)}\psi_{\boldsymbol{\theta}}|\hat{O}|\psi_{\boldsymbol{\theta}}\rangle = \langle\psi_{\boldsymbol{\theta}}|\hat{O}|\partial_{(l,i)}\psi_{\boldsymbol{\theta}}\rangle^*$. We can therefore simplify:

$$|\partial_{(l,i)}\mathcal{L}| \leq \left|2\mathrm{Re}\Big[\langle\psi_{\boldsymbol{\theta}}|\hat{O}|\partial_{(l,i)}\psi_{\boldsymbol{\theta}}\rangle\Big]\right|$$

Applying the Cauchy-Schwarz inequality and noting $\|\psi_{\boldsymbol{\theta}}\|_2 = 1$ and $\|\partial_{(l,i)}\psi_{\boldsymbol{\theta}}\|_2 \leq \|\mathfrak{g}_{(l,i)}\|_2$ we obtain

$$|\partial_{(l,i)}\mathcal{L}| \leq 2\|\hat{O}\|_2\|\mathfrak{g}_{(l,i)}\|_2$$

□

LEMMA 2 PROOF

Proof. We will prove Lemma 2. The regularized Fubini-Study Metric tensor is given by the matrix $g^{(\epsilon)}_{(l,i),(m,j)}$. We seek to bound its derivative $\partial_{(n,k)}g^{(\epsilon)}_{(l,i),(m,j)}(\boldsymbol{\theta})$. The metric is defined as:

$$g^{(\epsilon)}_{(l,i),(m,j)} = \left\langle\partial_{(l,i)}\psi_{\boldsymbol{\theta}}, \partial_{(m,j)}\psi_{\boldsymbol{\theta}}\right\rangle - \left\langle\partial_{(l,i)}\psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}}\right\rangle\left\langle\psi_{\boldsymbol{\theta}}, \partial_{(m,j)}\psi_{\boldsymbol{\theta}}\right\rangle$$

Using the product rule, the derivative with respect to $\theta_{(n,k)}$ is expanded as:

$$\partial_{(n,k)} g^{(\epsilon)}_{(l,i),(m,j)} = \underbrace{\left\langle \partial^2_{(n,k),(l,i)}\psi_{\boldsymbol{\theta}}, \partial_{(m,j)}\psi_{\boldsymbol{\theta}} \right\rangle + \left\langle \partial_{(l,i)}\psi_{\boldsymbol{\theta}}, \partial^2_{(n,k),(m,j)}\psi_{\boldsymbol{\theta}} \right\rangle}_{\text{Term 1}} \tag{15}$$

$$- \underbrace{\left[\left\langle \partial^2_{(n,k),(l,i)}\psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}} \right\rangle + \left\langle \partial_{(l,i)}\psi_{\boldsymbol{\theta}}, \partial_{(n,k)}\psi_{\boldsymbol{\theta}} \right\rangle\right] \left\langle \psi_{\boldsymbol{\theta}}, \partial_{(m,j)}\psi_{\boldsymbol{\theta}} \right\rangle}_{\text{Term 2}} \tag{16}$$

$$- \underbrace{\left\langle \partial_{(l,i)}\psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}} \right\rangle \left[\left\langle \partial_{(n,k)}\psi_{\boldsymbol{\theta}}, \partial_{(m,j)}\psi_{\boldsymbol{\theta}} \right\rangle + \left\langle \psi_{\boldsymbol{\theta}}, \partial^2_{(n,k),(m,j)}\psi_{\boldsymbol{\theta}} \right\rangle\right]}_{\text{Term 3}} \tag{17}$$

We bound each term by applying the generalized Cauchy-Schwarz inequality ($|\langle u, v\rangle| \leq \|u\|_2\|v\|_2$) and the triangle inequality directly to the derivative states. Note that the ℓ_2 norm of the state derivatives are bounded by the operator norms of their generators: $\|\partial_{(l,i)}\psi_{\boldsymbol{\theta}}\|_2 \leq \|\mathfrak{g}_{(l,i)}\|_2$ and $\|\partial^2_{(n,k),(l,i)}\psi_{\boldsymbol{\theta}}\|_2 \leq \|\mathfrak{g}_{(n,k)}\|_2\|\mathfrak{g}_{(l,i)}\|_2$, while the state itself is normalized ($\|\psi_{\boldsymbol{\theta}}\|_2 = 1$).

- **Term 1:**

$$|\text{Term 1}| \leq \left|\langle \partial^2_{(n,k),(l,i)}\psi_{\boldsymbol{\theta}}, \partial_{(m,j)}\psi_{\boldsymbol{\theta}}\rangle\right| + \left|\langle \partial_{(l,i)}\psi_{\boldsymbol{\theta}}, \partial^2_{(n,k),(m,j)}\psi_{\boldsymbol{\theta}}\rangle\right|$$

$$\leq \|\partial^2_{(n,k),(l,i)}\psi_{\boldsymbol{\theta}}\|_2\|\partial_{(m,j)}\psi_{\boldsymbol{\theta}}\|_2 + \|\partial_{(l,i)}\psi_{\boldsymbol{\theta}}\|_2\|\partial^2_{(n,k),(m,j)}\psi_{\boldsymbol{\theta}}\|_2$$

$$\leq (\|\mathfrak{g}_{(n,k)}\|_2\|\mathfrak{g}_{(l,i)}\|_2)\|\mathfrak{g}_{(m,j)}\|_2 + \|\mathfrak{g}_{(l,i)}\|_2(\|\mathfrak{g}_{(n,k)}\|_2\|\mathfrak{g}_{(m,j)}\|_2) = 2\|\mathfrak{g}_{(l,i)}\|_2\|\mathfrak{g}_{(m,j)}\|_2\|\mathfrak{g}_{(n,k)}\|_2$$

- **Term 2:**

$$|\text{Term 2}| \leq \left(\left|\langle \partial^2_{(n,k),(l,i)}\psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}}\rangle\right| + \left|\langle \partial_{(l,i)}\psi_{\boldsymbol{\theta}}, \partial_{(n,k)}\psi_{\boldsymbol{\theta}}\rangle\right|\right) \left|\langle \psi_{\boldsymbol{\theta}}, \partial_{(m,j)}\psi_{\boldsymbol{\theta}}\rangle\right|$$

$$\leq \left(\|\partial^2_{(n,k),(l,i)}\psi_{\boldsymbol{\theta}}\|_2\|\psi_{\boldsymbol{\theta}}\|_2 + \|\partial_{(l,i)}\psi_{\boldsymbol{\theta}}\|_2\|\partial_{(n,k)}\psi_{\boldsymbol{\theta}}\|_2\right) \|\psi_{\boldsymbol{\theta}}\|_2\|\partial_{(m,j)}\psi_{\boldsymbol{\theta}}\|_2$$

$$\leq (\|\mathfrak{g}_{(n,k)}\|_2\|\mathfrak{g}_{(l,i)}\|_2 \cdot 1 + \|\mathfrak{g}_{(l,i)}\|_2\|\mathfrak{g}_{(n,k)}\|_2) \cdot 1 \cdot \|\mathfrak{g}_{(m,j)}\|_2 = 2\|\mathfrak{g}_{(l,i)}\|_2\|\mathfrak{g}_{(m,j)}\|_2\|\mathfrak{g}_{(n,k)}\|_2$$

- **Term 3:**

$$|\text{Term 3}| \leq \left|\langle \partial_{(l,i)}\psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}}\rangle\right| \left(\left|\langle \partial_{(n,k)}\psi_{\boldsymbol{\theta}}, \partial_{(m,j)}\psi_{\boldsymbol{\theta}}\rangle\right| + \left|\langle \psi_{\boldsymbol{\theta}}, \partial^2_{(n,k),(m,j)}\psi_{\boldsymbol{\theta}}\rangle\right|\right)$$

$$\leq \|\partial_{(l,i)}\psi_{\boldsymbol{\theta}}\|_2\|\psi_{\boldsymbol{\theta}}\|_2 \left(\|\partial_{(n,k)}\psi_{\boldsymbol{\theta}}\|_2\|\partial_{(m,j)}\psi_{\boldsymbol{\theta}}\|_2 + \|\psi_{\boldsymbol{\theta}}\|_2\|\partial^2_{(n,k),(m,j)}\psi_{\boldsymbol{\theta}}\|_2\right)$$

$$\leq \|\mathfrak{g}_{(l,i)}\|_2 \cdot 1 \cdot (\|\mathfrak{g}_{(n,k)}\|_2\|\mathfrak{g}_{(m,j)}\|_2 + 1 \cdot \|\mathfrak{g}_{(n,k)}\|_2\|\mathfrak{g}_{(m,j)}\|_2) = 2\|\mathfrak{g}_{(l,i)}\|_2\|\mathfrak{g}_{(m,j)}\|_2\|\mathfrak{g}_{(n,k)}\|_2$$

Finally, summing the upper bounds for all three terms:

$$\left|\partial_{(n,k)} g^{(\epsilon)}_{(l,i),(m,j)}\right| \leq |\text{Term 1}| + |\text{Term 2}| + |\text{Term 3}|$$

$$\leq 2\|\mathfrak{g}_{(l,i)}\|_2\|\mathfrak{g}_{(m,j)}\|_2\|\mathfrak{g}_{(n,k)}\|_2 + 2\|\mathfrak{g}_{(l,i)}\|_2\|\mathfrak{g}_{(n,k)}\|_2\|\mathfrak{g}_{(m,j)}\|_2 + 2\|\mathfrak{g}_{(l,i)}\|_2\|\mathfrak{g}_{(n,k)}\|_2\|\mathfrak{g}_{(m,j)}\|_2$$

This yields the bound valid for general non-commuting generators:

$$\left|\partial_{(n,k)} g^{(\epsilon)}_{(l,i),(m,j)}\right| \leq 6\|\mathfrak{g}_{(l,i)}\|_2\|\mathfrak{g}_{(m,j)}\|_2\|\mathfrak{g}_{(n,k)}\|_2 \tag{18}$$

□

LEMMA 3 PROOF

Proof. We will prove Lemma 3. The Christoffel symbol is defined as:

$$\Gamma^{(n,k)}_{(l,i),(m,j)} = \frac{1}{2}\sum_{(p,r)} (g^{(\epsilon)-1})_{(n,k),(p,r)} \left(\partial_{(l,i)} g^{(\epsilon)}_{(m,j),(p,r)} + \partial_{(m,j)} g^{(\epsilon)}_{(l,i),(p,r)} - \partial_{(p,r)} g^{(\epsilon)}_{(l,i),(m,j)}\right)$$

We bound each part individually: Inverse Metric $(g^{(\epsilon)-1})_{(n,k),(p,r)}$: Since the metric is regularized by ϵ, the element-wise absolute value is bounded by the spectral norm of the inverse:

$$|(g^{(\epsilon)-1})_{(n,k),(p,r)}| \leq \frac{1}{\epsilon}$$

Metric Derivatives: By applying Lemma 2 to the metric derivatives, we obtain:

$$\left|\partial_{(c,q)} g^{(\epsilon)}_{(a,u),(b,v)}\right| \leq 6\|\mathfrak{g}_{(a,u)}\|_2 \|\mathfrak{g}_{(b,v)}\|_2 \|\mathfrak{g}_{(c,q)}\|_2$$

Substituting these give:

$$|\Gamma^{(n,k)}_{(l,i),(m,j)}| \leq \frac{1}{2}\sum_{(p,r)} \frac{1}{\epsilon}\left(18\|\mathfrak{g}_{(l,i)}\|_2 \|\mathfrak{g}_{(m,j)}\|_2 \|\mathfrak{g}_{(p,r)}\|_2\right) = \frac{9}{\epsilon}\|\mathfrak{g}_{(l,i)}\|_2 \|\mathfrak{g}_{(m,j)}\|_2 \sum_{(p,r)} \|\mathfrak{g}_{(p,r)}\|_2$$

□

LEMMA 4 PROOF

Proof. To find the Hessian elements $\text{Hess}_{(l,i),(m,j)} = \partial^2_{(l,i),(m,j)}\mathcal{L}$, we differentiate the first derivative expression that we obtained in Lemma 1:

$$\begin{aligned}
\text{Hess}_{(l,i),(m,j)} &= \partial_{(m,j)}\left[\langle \partial_{(l,i)}\psi_{\boldsymbol{\theta}}|\hat{O}|\psi_{\boldsymbol{\theta}}\rangle + \langle \psi_{\boldsymbol{\theta}}|\hat{O}|\partial_{(l,i)}\psi_{\boldsymbol{\theta}}\rangle\right] \\
&= \langle \partial^2_{(l,i),(m,j)}\psi_{\boldsymbol{\theta}}|\hat{O}|\psi_{\boldsymbol{\theta}}\rangle + \langle \partial_{(l,i)}\psi_{\boldsymbol{\theta}}|\hat{O}|\partial_{(m,j)}\psi_{\boldsymbol{\theta}}\rangle \\
&\quad + \langle \partial_{(m,j)}\psi_{\boldsymbol{\theta}}|\hat{O}|\partial_{(l,i)}\psi_{\boldsymbol{\theta}}\rangle + \langle \psi_{\boldsymbol{\theta}}|\hat{O}|\partial^2_{(l,i),(m,j)}\psi_{\boldsymbol{\theta}}\rangle
\end{aligned}$$

Grouping the terms into conjugate pairs, we obtain:

$$\text{Hess}_{(l,i),(m,j)} = 2\text{Re}\left(\langle \partial_{(m,j)}\psi_{\boldsymbol{\theta}}|\hat{O}|\partial_{(l,i)}\psi_{\boldsymbol{\theta}}\rangle + \langle \psi_{\boldsymbol{\theta}}|\hat{O}|\partial^2_{(l,i),(m,j)}\psi_{\boldsymbol{\theta}}\rangle\right) \tag{19}$$

We bound the absolute value $|\text{Hess}_{(l,i),(m,j)}|$ by applying the triangle inequality and the generalized Cauchy-Schwarz inequality $|\langle u|A|v\rangle| \leq \|u\|_2\|A\|_2\|v\|_2$:

$$\begin{aligned}
|\text{Hess}_{(l,i),(m,j)}| &\leq 2\left|\langle \partial_{(m,j)}\psi_{\boldsymbol{\theta}}|\hat{O}|\partial_{(l,i)}\psi_{\boldsymbol{\theta}}\rangle\right| + 2\left|\langle \psi_{\boldsymbol{\theta}}|\hat{O}|\partial^2_{(l,i),(m,j)}\psi_{\boldsymbol{\theta}}\rangle\right| \\
&\leq 2\|\partial_{(m,j)}\psi_{\boldsymbol{\theta}}\|_2\|\hat{O}\|_2\|\partial_{(l,i)}\psi_{\boldsymbol{\theta}}\|_2 \\
&\quad + 2\|\psi_{\boldsymbol{\theta}}\|_2\|\hat{O}\|_2\|\partial^2_{(l,i),(m,j)}\psi_{\boldsymbol{\theta}}\|_2
\end{aligned}$$

Substituting the generator norms, this simplifies to:

$$|\text{Hess}_{(l,i),(m,j)}| \leq 4\|\hat{O}\|_2\|\mathfrak{g}_{(l,i)}\|_2\|\mathfrak{g}_{(m,j)}\|_2 \tag{20}$$

□

LEMMA 5 PROOF

Proof. The Riemannian Hessian element is bounded by applying the triangle inequality to its constituent parts:

$$|\text{Hess}_{\nabla(l,i),(m,j)}| \leq \underbrace{\left|\partial^2_{(l,i),(m,j)}\mathcal{L}\right|}_{\text{Term A}} + \underbrace{\left|\sum_{(n,k)} \Gamma^{(n,k)}_{(l,i),(m,j)}\, \partial_{(n,k)}\mathcal{L}\right|}_{\text{Term B}}$$

From Lemma 4 we have a bound for term A. We analyze Term B, the Geometric Correction term:

$$C_{(l,i),(m,j)} = \sum_{(n,k)} \Gamma^{(n,k)}_{(l,i),(m,j)}\, \partial_{(n,k)}\mathcal{L}$$

We bound this by splitting it into its components:

$$|C_{(l,i),(m,j)}| \leq \sum_{(n,k)} \underbrace{|\Gamma^{(n,k)}_{(l,i),(m,j)}|}_{\text{Christoffel}} \cdot \underbrace{\left|\partial_{(n,k)}\mathcal{L}\right|}_{\text{Gradient}}$$

From Lemma 1 and Lemma 3, we have the exact bounds:

$$\left|\partial_{(n,k)}\mathcal{L}\right| \leq 2\|\hat{O}\|_2 \|\mathfrak{g}_{(n,k)}\|_2$$

$$|\Gamma^{(n,k)}_{(l,i),(m,j)}| \leq \frac{9}{\epsilon}\|\mathfrak{g}_{(l,i)}\|_2 \|\mathfrak{g}_{(m,j)}\|_2 \sum_{(p,r)} \|\mathfrak{g}_{(p,r)}\|_2$$

Substituting gives:

$$|C_{(l,i),(m,j)}| \leq \sum_{(n,k)} \left[\frac{18}{\epsilon}\|\hat{O}\|_2 \|\mathfrak{g}_{(n,k)}\|_2 \|\mathfrak{g}_{(l,i)}\|_2 \|\mathfrak{g}_{(m,j)}\|_2 \sum_{(p,r)} \|\mathfrak{g}_{(p,r)}\|_2\right]$$

Which can be simplified to:

$$|C_{(l,i),(m,j)}| \leq \frac{18}{\epsilon}\|\hat{O}\|_2 \|\mathfrak{g}_{(l,i)}\|_2 \|\mathfrak{g}_{(m,j)}\|_2 \sum_{(n,k)} \|\mathfrak{g}_{(n,k)}\|_2 \sum_{(p,r)} \|\mathfrak{g}_{(p,r)}\|_2$$

Thus the final bound for the Hessian becomes:

$$|\text{Hess}_{\nabla(l,i),(m,j)}| \leq 4\|\hat{O}\|_2 \|\mathfrak{g}_{(l,i)}\|_2 \|\mathfrak{g}_{(m,j)}\|_2 + \frac{18}{\epsilon}\|\hat{O}\|_2 \|\mathfrak{g}_{(l,i)}\|_2 \|\mathfrak{g}_{(m,j)}\|_2 \left(\sum_{(n,k)} \|\mathfrak{g}_{(n,k)}\|_2\right)^2$$

□

LEMMA 6 PROOF

Proof. By definition, the gradient vector's ℓ_2-norm is $\|d\mathcal{L}\|_2 = \sqrt{\sum_{(l,i)} (\partial_{(l,i)}\mathcal{L})^2}$. From Lemma 1 we have:

$$\partial_{(l,i)}\mathcal{L} = 2\text{Re}\Big[\langle\psi_{\boldsymbol{\theta}}|\hat{O}|\partial_{(l,i)}\psi_{\boldsymbol{\theta}}\rangle\Big] \quad \text{Bounded by} \quad |\partial_{(l,i)}\mathcal{L}| \leq 2\|\hat{O}\|_2 \|\mathfrak{g}_{(l,i)}\|_2$$

Which naturally gives:

$$\|d\mathcal{L}\|_2 \leq \sqrt{\sum_{(l,i)} (2\|\hat{O}\|_2 \|\mathfrak{g}_{(l,i)}\|_2)^2} = 2\|\hat{O}\|_2 \sqrt{\sum_{(l,i)} \|\mathfrak{g}_{(l,i)}\|_2^2}$$

□

LEMMA 7 PROOF

Proof. We start with the elementwise bound from Lemma 3: $|\Gamma^{(n,k)}_{(l,i),(m,j)}| \leq \frac{9}{\epsilon}\|\mathfrak{g}_{(l,i)}\|_2 \|\mathfrak{g}_{(m,j)}\|_2 \sum_{(p,r)} \|\mathfrak{g}_{(p,r)}\|_2$

To analyze the ℓ_2 norm, let us define a scalar constant C and a vector $v \in \mathbb{R}^P$ as follows:

$$C = \frac{9}{\epsilon} \sum_{(p,r)} \|\mathfrak{g}_{(p,r)}\|_2, \quad v = \begin{bmatrix} \|\mathfrak{g}_{(1,1)}\|_2 \\ \vdots \\ \|\mathfrak{g}_{(L,M_L)}\|_2 \end{bmatrix}$$

Substituting these into the elementwise bound, we observe that the matrix $\Gamma^{(n,k)}$ is entry-wise dominated by a matrix B, such that $|\Gamma^{(n,k)}_{(l,i),(m,j)}| \leq B_{(l,i),(m,j)} = C v_{(l,i)} v_{(m,j)}$. This allows us to express the dominating matrix as a scaled outer product:

$$B = C(vv^T)$$

Since B is a Rank-1 symmetric matrix with non-negative entries, its spectral norm is given by $C\|v\|_2^2$. Given that the spectral norm of $\Gamma^{(n,k)}$ is bounded by the spectral norm of its entry-wise dominator B, we obtain:

$$\|\Gamma^{(n,k)}\|_2 \leq C\|v\|_2^2 = \left(\frac{9}{\epsilon} \sum_{(l,i)} \|\mathfrak{g}_{(l,i)}\|_2\right) \left(\sum_{(m,j)} \|\mathfrak{g}_{(m,j)}\|_2^2\right)$$

□

LEMMA 8 PROOF

Proof. Following a logic identical to Lemma 7, we start with the elementwise bound from Lemma 4:

$$|\text{Hess}_{(l,i),(m,j)}| \leq 4\|\hat{O}\|_2 \|\mathfrak{g}_{(l,i)}\|_2 \|\mathfrak{g}_{(m,j)}\|_2$$

Let us define a matrix B such that $B_{(l,i),(m,j)}$ corresponds to the right-hand side of the inequality above. Since $|\text{Hess}_{(l,i),(m,j)}| \leq B_{(l,i),(m,j)}$ for all $(l,i),(m,j)$, it follows that the spectral norm of the Hessian is bounded by the spectral norm of B, such that $\|\text{Hess}\|_2 \leq \|B\|_2$. To analyze the structure of B, we define a vector $v \in \mathbb{R}^P$ containing the norms of the generators:

$$v = \begin{bmatrix} \|\mathfrak{g}_{(1,1)}\|_2 \\ \vdots \\ \|\mathfrak{g}_{(L,M_L)}\|_2 \end{bmatrix}$$

Substituting this vector into the expression for $B_{(l,i),(m,j)}$, we observe a common factor of $v_{(l,i)}v_{(m,j)} = \|\mathfrak{g}_{(l,i)}\|_2 \|\mathfrak{g}_{(m,j)}\|_2$ in every term. Let the scalar term be denoted by $C = 4\|\hat{O}\|_2$. We can now express the matrix B as a scaled outer product of the vector v:

$$B_{(l,i),(m,j)} = C v_{(l,i)} v_{(m,j)} \implies B = C(vv^T)$$

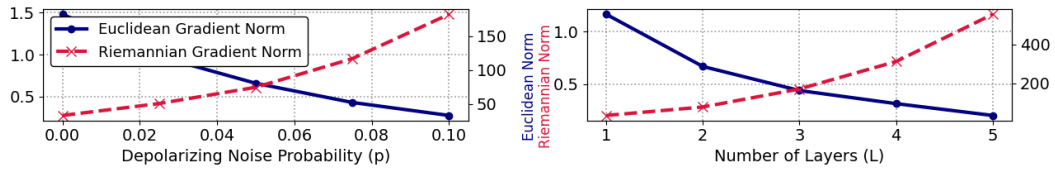


Figure 3: Euclidean vs. Riemannian gradient norms under scaling. As depolarizing noise and circuit depth increase, the gradients exhibit a strictly inverse relationship. The Euclidean gradient vanishes into a noise-induced barren plateau, while the Riemannian gradient experiences an exponential spike.

This reveals that B is a Rank-1 symmetric matrix. The spectral norm of a Rank-1 matrix of the form cuu^T is given by $|c|\|u\|_2^2$. Therefore:

$$\|B\|_2 = C\|v\|_2^2 = C\left(\sum_{(n,k)} \|\mathfrak{g}_{(n,k)}\|_2^2\right)$$

Substituting the explicit form of C back into the equation yields the final bound:

$$\|\text{Hess}\|_2 \leq 4\|\hat{O}\|_2 \sum_{(n,k)} \|\mathfrak{g}_{(n,k)}\|_2^2$$

□

G Noise Analysis: Riemannian Methods Under Local Depolarizing Noise

G.1 Noise in Variational Quantum Circuits

Every physical implementation of a quantum circuit is subject to decoherence: unwanted interactions between the system qubits and their environment that irreversibly degrade the quantum state. At the level of the density operator, decoherence is described by a completely positive, trace-preserving (CPTP) map $\mathcal{N}$ that replaces the ideal unitary evolution $\rho_{\psi_{\boldsymbol{\theta}}} \mapsto \mathcal{U}(\boldsymbol{\theta})\rho\mathcal{U}(\boldsymbol{\theta})^\dagger$ with a noisy channel $\rho_{\psi_{\boldsymbol{\theta}}} \mapsto \mathcal{N}(\mathcal{U}(\boldsymbol{\theta})\rho\mathcal{U}(\boldsymbol{\theta})^\dagger)$. The most general such map admits a Kraus representation $\mathcal{N}(\rho) = \sum_k K_k \rho_{\psi_{\boldsymbol{\theta}}} K_k^\dagger$, but in full generality the noise channel can depend on the history of all prior operations, producing temporal correlations (non-Markovian noise) that render analytical treatment of the optimization landscape essentially intractable.

Throughout this appendix, we restrict attention to *Markovian* noise: the channel applied after each circuit layer depends only on the current state, not on the history of preceding layers. This is not merely a mathematical convenience. The Markovian assumption is the standard noise model in the variational quantum algorithms literature Wang et al. (2021); Preskill (2018); Cerezo et al. (2021) and is physically well-motivated for two reasons. First, on current superconducting and trapped-ion hardware, the dominant error mechanisms, amplitude damping, dephasing, and depolarization, arise from coupling to thermal baths whose correlation times are orders of magnitude shorter than gate durations, placing the dynamics firmly in the Markovian regime Preskill (2018). Second, the Markovian framework is precisely the setting in which the noise-induced barren plateau (NIBP) results of Wang et al. (2021) are established; adopting the same model allows us to inherit their trace-norm findings and isolate the new geometric phenomena, the Riemannian spike and the noise scaling of the connection, without confounding them with non-Markovian effects that neither the NIBP framework nor any existing barren plateau analysis accounts for.

Within the Markovian class, we further specialize to *local depolarizing noise*: after each parameterized layer, an independent single-qubit depolarizing channel acts on every qubit. This is the

canonical noise model used by Wang et al. (2021) to establish the NIBP, and it is the weakest physically relevant noise model that captures the key phenomenon: layer-by-layer contraction of the quantum state toward the maximally mixed state I/d. Crucially, because entangling gates between noise layers redistribute Pauli weight across strings, individual Pauli coefficients do not decay at a uniform rate. A low-weight Pauli string can be mapped by a CNOT to a high-weight string, which is then contracted more aggressively by the subsequent noise layer. This weight redistribution means that coefficient-wise tracking is insufficient; instead, the analysis requires entropy-based contraction bounds, specifically, the strict data-processing inequality for the sandwiched 2-Rényi divergence Müller-Hermes & Reeb (2017); Wang et al. (2021), that hold for *any* circuit structure, including those with arbitrary entangling gates.

Our strategy in the sections that follow is to adopt this local depolarizing framework and derive the noise scaling of every geometric object entering the circuit, the Fubini–Study metric $\tilde{g}$, its derivatives, the Christoffel symbols $\tilde{\Gamma}^k_{ij}$, and the Riemannian gradient $\mathrm{grad}_{\tilde{g}^{(\epsilon)}}\mathcal{L}$, directly from the trace-norm bounds on the noisy state and its parameter derivatives. The resulting scalings reveal that the intrinsic differential geometry of the manifold is preserved under noise (the connection remains $\mathcal{O}(1)$), while the interface between the geometry and the objective, the musical isomorphism $(\tilde{g}^{(\epsilon)})^\sharp$, becomes exponentially ill-conditioned, producing the noise-induced Riemannian spike of Theorem 2.

G.2 Noise Model and State Representation

We represent an n-qubit quantum state in the Pauli basis. Let $\sigma_n^i \in \{I, X, Y, Z\}^{\otimes n}$ denote the i-th n-qubit Pauli string, where the index i ranges over all $4^n - 1$ strings excluding the identity $I^{\otimes n}$. Any n-qubit state ρ admits the decomposition

$$\rho \;=\; \frac{1}{2^n}\left(I^{\otimes n} + \sum_{i=1}^{4^n-1} a_i\,\sigma_n^i\right) \;=\; \frac{I}{2^n} + \Delta\,, \tag{21}$$

where $a_i = \mathrm{Tr}[\rho\,\sigma_n^i]$ are the Pauli coefficients and $\Delta = \rho - I/2^n$ is the traceless signal component. The Pauli strings $\{\sigma_n^i\}_{i=0}^{4^n-1}$ (with $\sigma_n^0 = I^{\otimes n}$) form an orthogonal basis for the space of $2^n \times 2^n$ Hermitian matrices under the Hilbert–Schmidt inner product $\langle A, B\rangle_{\mathrm{HS}} = \mathrm{Tr}[A^\dagger B]$, with $\mathrm{Tr}[\sigma_n^i\sigma_n^j] = 2^n\delta_{ij}$. The decomposition equation 21 is therefore unique, and the traceless component Δ carries all information that distinguishes ρ from the maximally mixed state. For a pure initial state $\rho_0 = |\psi_0\rangle\langle\psi_0|$, the signal $\Delta_0 = \rho_0 - I/2^n$ encodes the full information content of the state; noise progressively destroys this signal by contracting the coefficients a_i toward zero.

Layered circuit with local noise. Consider a parameterized quantum circuit of depth L, partitioned into discrete layers. Recalling the ansatz structure of equation 1, the unitary evolution at layer l is

$$\mathcal{U}_l(\boldsymbol{\theta}_l) = \prod_{m=1}^{M_l} \mathcal{W}_{(l,m)}\,\exp\!\big(-i\,\boldsymbol{\theta}_{(l,m)}\,\mathfrak{g}_{(l,m)}\big)\,,$$

We define the sequential composition of layers from a to b (with $b \geq a$) as $\mathcal{U}_{[b:a]} = \mathcal{U}_b\,\mathcal{U}_{b-1}\cdots\mathcal{U}_a$. After each unitary layer, a local Pauli noise channel $\mathcal{N} = \mathcal{N}_1 \otimes \cdots \otimes \mathcal{N}_n$ acts independently on every qubit. Each single-qubit channel $\mathcal{N}_j$ is diagonal in the Pauli basis: it maps each local Pauli operator $\sigma \in \{X, Y, Z\}$ as

$$\mathcal{N}_j(\sigma) \;=\; q_\sigma\,\sigma\,, \qquad -1 < q_X,\, q_Y,\, q_Z < 1\,, \tag{22}$$

while preserving the identity, $\mathcal{N}_j(I) = I$. The constraint $|q_\sigma| < 1$ ensures that $\mathcal{N}_j$ is a proper contraction on the traceless subspace: every non-identity Pauli component is strictly attenuated. The symmetric depolarizing channel corresponds to the special case $q_X = q_Y = q_Z = 1 - \frac{4p}{3}$, where p is the depolarizing probability, but the analysis below holds for the full asymmetric Pauli channel defined by equation 22. We characterize the noise strength by the single worst-case parameter

$$q \;=\; \max\big\{|q_X|,\,|q_Y|,\,|q_Z|\big\} \;<\; 1\,.$$

This choice ensures that every contraction bound derived below holds uniformly over all Pauli strings, regardless of their internal Pauli-type composition. The noisy output state after L layers of alternating unitaries and noise is

$$\tilde{\rho}_{\psi_{\boldsymbol{\theta}}} \;=\; \big(\mathcal{N}\circ\mathcal{U}_L\circ\mathcal{N}\circ\cdots\circ\mathcal{N}\circ\mathcal{U}_1\circ\mathcal{N}\big)(\rho_0)\,,$$

where the composition involves L unitary layers interleaved with $L+1$ applications of the noise channel (one preceding $\mathcal{U}_1$ and one following each $\mathcal{U}_l$). The noisy objective function is

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}) \;=\; \mathrm{Tr}\big[\hat{O}\,\tilde{\rho}_{\psi_{\boldsymbol{\theta}}}\big]\,. \tag{23}$$

Effect of noise on the Pauli coefficients. Acting on the full Pauli decomposition equation 21, a single application of the local noise channel maps each coefficient as

$$a_i \;\xmapsto{\mathcal{N}}\; q_X^{x(i)}\,q_Y^{y(i)}\,q_Z^{z(i)}\,a_i\,, \tag{24}$$

where $x(i)$, $y(i)$, $z(i)$ count the number of X, Y, Z factors respectively in the i-th Pauli string, and $w(i)=x(i)+y(i)+z(i)\geq 1$ is its Pauli weight (the number of non-identity tensor factors). Since each $|q_\sigma|\leq q<1$ and $w(i)\geq 1$, the contraction satisfies

$$|\tilde{a}_i| \;=\; |q_X|^{x(i)}\,|q_Y|^{y(i)}\,|q_Z|^{z(i)}\,|a_i| \;\leq\; q^{w(i)}\,|a_i| \;\leq\; q\,|a_i|\,. \tag{25}$$

Higher-weight Pauli strings are contracted more aggressively, with the suppression factor $q^{w(i)}$ decaying exponentially in the weight. This weight-dependent contraction is the mechanism by which noise selectively destroys high-weight correlations while partially preserving local (low-weight) information.

Unitary layers between noise channels preserve the Euclidean norm $\|\boldsymbol{a}\|_2$ of the full coefficient vector, since unitary conjugation $\rho\mapsto\mathcal{U}\rho\mathcal{U}^\dagger$ is an isometry of the Hilbert–Schmidt norm and therefore permutes the Pauli coefficients without changing their magnitudes. However, entangling unitaries *redistribute* the Pauli weight of individual coefficients: a CNOT gate, for instance, maps $X\otimes I\mapsto X\otimes X$, promoting a weight-1 string to weight 2. The subsequent noise layer then contracts this newly high-weight component by q^2 rather than q, producing a stronger suppression than the original weight would have incurred. This redistribution means that individual Pauli coefficients do not decay at a uniform per-layer rate, and coefficient-by-coefficient tracking across the full circuit is insufficient. The rigorous analysis therefore requires the entropy-based contraction bounds developed in the following subsection.

Using the decomposition equation 21, the noisy output state equation G.2 can be written as

$$\tilde{\rho}_{\psi_{\boldsymbol{\theta}}} \;=\; \frac{I}{2^n}+\tilde{\Delta}_L\,, \tag{26}$$

where $\tilde{\Delta}_L$ is the surviving traceless signal. Since $I/2^n$ is invariant under both unitary conjugation and trace-preserving channels, all parameter-dependent structure, gradients, metric, and curvature, resides exclusively in $\tilde{\Delta}_L$. The derivative bounds of the following subsection quantify precisely how noise suppresses this signal; the key mechanism is that each noise layer contracts the sandwiched 2-Rényi divergence $D_2\big(\rho\,\big\|\,I/2^n\big)$ while each intermediate unitary $\mathcal{U}_l$ preserves it by unitary invariance Müller-Hermes & Reeb (2017); Wang et al. (2021), producing an exponential suppression that accumulates across the full circuit depth.

G.3 Inherited Bounds from the NIBP Framework

The following results are established by Wang et al. Wang et al. (2021) via sandwiched 2-Rényi relative entropy contraction and Pinsker's inequality. We state them here in the notation of the present paper for completeness; they hold for *any* circuit structure, including those with arbitrary entangling gates. Throughout, we write $c=1/(2\ln 2)$ for the entropy conversion constant appearing in the Pinsker bound.

Lemma 9 (State Concentration; (Wang et al., 2021, Supplementary Lemma 7))**.** *Let $\tilde{\rho}_{\psi_{\boldsymbol{\theta}}}$ denote the output state of the depth-L noisy circuit equation G.2. Then:*

$$\left\|\tilde{\rho}_{\psi_{\boldsymbol{\theta}}}-\frac{I}{2^n}\right\|_1 \;\leq\; \sqrt{2\ln 2}\;n^{1/2}\,q^{c(L+1)}\,. \tag{27}$$

Lemma 10 (Gradient Bound; (Wang et al., 2021, Theorem 1)). *The partial derivative of the noisy cost function equation 23 with respect to parameter $\boldsymbol{\theta}_{(l,k)}$ satisfies:*

$$\left|\partial_{(l,k)}\tilde{\mathcal{L}}\right| \;\in\; \mathcal{O}\Big(\|\hat{O}\|_2\,\|\mathfrak{g}_{(l,k)}\|_2\,n^{1/2}\,q^{cL}\Big)\;. \tag{28}$$

The exact prefactors involve the Pauli decomposition of the observable $\hat{O}$ and the commutator structure of the generator $\mathfrak{g}_{(l,k)}$ within its layer; we refer to (Wang et al., 2021, Theorem 1) for the full expression. For our purposes, the essential content is the exponential suppression q^{cL}, which is inherited from the entropy contraction across L noise layers.

Lemma 11 (First Derivative Trace Norm; (Wang et al., 2021, Eq. (96))). *Let $\partial_{(l,k)}\tilde{\rho}_{\psi_{\boldsymbol{\theta}}}$ denote the derivative of the noisy output state with respect to parameter $\boldsymbol{\theta}_{(l,k)}$ in layer l. Then:*

$$\left\|\partial_{(l,k)}\tilde{\rho}_{\psi_{\boldsymbol{\theta}}}\right\|_1 \;\in\; \mathcal{O}\Big(\|\mathfrak{g}_{(l,k)}\|_2\,n^{1/2}\,q^{cL}\Big)\;. \tag{29}$$

This bound follows from expressing the derivative as a commutator $[-i\mathfrak{g}_{(l,k)},\,\mathcal{U}_l\tilde{\rho}_{l-1}\mathcal{U}_l^\dagger]$ at layer l, decomposing the resulting expression via the parameter-shift rule into a linear combination of noisy states, and applying the entropy contraction bound (Lemma 9) to each constituent depth-L circuit. Crucially, the first derivative and the state itself share the same q^{cL} scaling because the differentiation inserts a single bounded generator into the unitary chain without altering the number of noise layers traversed.

To derive the geometric scalings of the metric tensor and its derivatives (Section G.5), we require one additional bound not explicitly established in Wang et al. (2021): the trace norm of the *second* derivative of the noisy state. We prove this in the following subsection, using the same entropy contraction machinery.

G.4 Second Derivative Bound

Lemma 12 (Second Derivative Trace Norm). *Let $\boldsymbol{\theta}_{(l_1,i)}$ and $\boldsymbol{\theta}_{(l_2,j)}$ be parameters located in layers l_1 and l_2 respectively (with $l_1 \le l_2$). The second derivative of the noisy output state satisfies:*

$$\left\|\partial^2_{(l_1,i),(l_2,j)}\,\tilde{\rho}_{\psi_{\boldsymbol{\theta}}}\right\|_1 \;\in\; \mathcal{O}\Big(\|\mathfrak{g}_{(l_1,i)}\|_2\,\|\mathfrak{g}_{(l_2,j)}\|_2\,n^{1/2}\,q^{cL}\Big)\;. \tag{30}$$

Proof. We proceed by explicitly constructing the second derivative from the circuit's unitary decomposition, then applying the NIBP entropy contraction to the result.

Step 1: Structure of the noisy state. Recall from equation 26 that the noisy output state decomposes as $\tilde{\rho}_{\psi_{\boldsymbol{\theta}}} = I/d + \tilde{\Delta}_L$, where $d = 2^n$ and $\tilde{\Delta}_L$ is the traceless signal component. Since I/d is invariant under both unitary conjugation and trace-preserving channels, it is independent of $\boldsymbol{\theta}$, and all derivatives act exclusively on the signal term:

$$\partial_{(l,k)}\,\tilde{\rho}_{\psi_{\boldsymbol{\theta}}} \;=\; \partial_{(l,k)}\,\tilde{\Delta}_L\,, \qquad \partial^2_{(l_1,i),(l_2,j)}\,\tilde{\rho}_{\psi_{\boldsymbol{\theta}}} \;=\; \partial^2_{(l_1,i),(l_2,j)}\,\tilde{\Delta}_L\,. \tag{31}$$

The signal $\tilde{\Delta}_L$ is the result of propagating the initial traceless component $\Delta_0 = \rho_0 - I/d$ through L layers of alternating unitary evolution and noise channels equation G.2. By the cumulative entropy contraction across all $L+1$ noise layers, its trace norm is exponentially suppressed (Lemma 9):

$$\|\tilde{\Delta}_L\|_1 \;=\; \left\|\tilde{\rho}_{\psi_{\boldsymbol{\theta}}} - I/d\right\|_1 \;\le\; \sqrt{2\ln 2}\,n^{1/2}\,q^{c(L+1)}\,. \tag{32}$$

The essential point is that the exponential suppression factor q^{cL} in the derivative bounds arises from this same contraction mechanism: differentiating with respect to a circuit parameter inserts a bounded generator into the unitary chain but does not alter the number of noise layers through which the signal propagates.

Step 2: First derivative. Consider a parameter $\boldsymbol{\theta}_{(l,k)}$ residing in layer l. In the noisy circuit composition equation G.2, only the unitary $\mathcal{U}_l$ depends on $\boldsymbol{\theta}_{(l,k)}$; the noise channels $\mathcal{N}$ and all other unitary layers are parameter-independent. Since the full composition is a chain of linear maps (the $\mathcal{N}$ and the conjugation maps $\rho \mapsto \mathcal{U}_l\rho\mathcal{U}_l^\dagger$), the derivative passes through every layer except $\mathcal{U}_l$

by linearity. At layer l, differentiating the conjugation $\mathcal{U}_l(\cdot)\mathcal{U}_l^\dagger$ with respect to $\boldsymbol{\theta}_{(l,k)}$ produces a commutator insertion:

$$\partial_{(l,k)}\Big(\mathcal{U}_l\,\rho\,\mathcal{U}_l^\dagger\Big) \;=\; -i\big[\mathfrak{g}_{(l,k)},\,\mathcal{U}_l\,\rho\,\mathcal{U}_l^\dagger\big]\,. \tag{33}$$

The full first derivative of the noisy state is therefore

$$\partial_{(l,k)}\,\tilde{\rho}_{\psi_{\boldsymbol{\theta}}} \;=\; \mathcal{N}\circ\mathcal{U}_{[L:l+1]}\circ\mathcal{N}\Big(-i\big[\mathfrak{g}_{(l,k)},\,\mathcal{U}_l\,\tilde{\rho}_{l-1}\,\mathcal{U}_l^\dagger\big]\Big)\,, \tag{34}$$

where $\tilde{\rho}_{l-1}$ denotes the (noisy) state arriving at the input of layer l, and $\mathcal{N}\circ\mathcal{U}_{[L:l+1]}\circ\mathcal{N}$ denotes the remaining post-layer-l segment of the noisy circuit. Two structural observations are essential:

1. *No noise layers are created or removed.* The differentiation is a purely algebraic operation that inserts the bounded generator $\mathfrak{g}_{(l,k)}$ as a commutator at a single point in the circuit. Every noise channel in the original composition remains in place.
2. *The result is traceless.* The commutator $[\mathfrak{g}_{(l,k)},\cdot\,]$ preserves tracelessness, and the subsequent noise channels and unitaries are trace-preserving, so $\partial_{(l,k)}\,\tilde{\rho}_{\psi_{\boldsymbol{\theta}}}$ is a traceless operator that has traversed the same sequence of noise layers as $\tilde{\Delta}_L$.

By the entropy contraction argument of Lemma 9—applied to the traceless operator propagating through the full noise sequence—the trace norm is bounded by $\big\|\partial_{(l,k)}\,\tilde{\rho}_{\psi_{\boldsymbol{\theta}}}\big\|_1 \;\in\; \mathcal{O}\big(\|\mathfrak{g}_{(l,k)}\|_2\,n^{1/2}\,q^{cL}\big)$, recovering Lemma 11. The generator norm $\|\mathfrak{g}_{(l,k)}\|_2$ enters as a prefactor because the commutator satisfies $\|[A,B]\|_1 \le 2\|A\|_2\,\|B\|_1$.

Step 3: Second derivative. Consider two parameters $\boldsymbol{\theta}_{(l,k)}$ and $\boldsymbol{\theta}_{(m,j)}$ residing in layers l and m respectively (assuming $l \le m$). Rather than performing a massive product-rule expansion, we can recognize that differentiating the noisy state simply inserts a commutator with the corresponding Hermitian generator at that specific layer. By defining Φ as the completely positive trace-preserving (CPTP) maps encompassing the alternating sequence of unitaries and noise channels before, between, and after the target layers, we can express the second derivative cleanly as a nested operation on the initial traceless signal Δ_0:

$$\partial^2_{(l,k)(m,j)}\,\tilde{\rho}_{\psi_\theta} = \Phi_{post}\left(-i\left[\mathfrak{g}_{(m,j)},\Phi_{mid}\left(-i\left[\mathfrak{g}_{(l,k)},\Phi_{pre}(\Delta_0)\right]\right)\right]\right)$$

This formulation makes the structural argument immediate. Taking higher-order derivatives merely inserts additional bounded operators into the unitary chain. It does not add, remove, or modify the $L+1$ depolarizing noise channels that the signal must traverse. Because the sandwiched 2-Rényi entropy contraction is driven exclusively by the number of noise layers encountered, the underlying exponential suppression remains perfectly intact. Applying the standard trace-norm commutator bound $\|[A,B]\|_1 \le 2\|A\|_2\|B\|_1$ to the nested structure simply extracts the operator norms of the two inserted generators as polynomial prefactors, leaving the exponential scaling untouched:

$$||\partial^2_{(l,k)(m,j)}\,\tilde{\rho}_{\psi_\theta}||_1 \in \mathcal{O}(||\mathfrak{g}_{(l,k)}||_2||\mathfrak{g}_{(m,j)}||_2 n^{1/2}q^{cL})$$

□

G.5 Geometric Scalings Under Noise and Proof of Theorem 2

With the derivative bounds in hand, we assemble the asymptotic noise scalings of the manifold's geometric tensors and prove Theorem 2. Throughout, we write $\alpha = cL$ (so that $q^\alpha = \kappa_{\mathrm{err}}^L$) and suppress polynomial prefactors $\mathcal{O}(\mathrm{poly}(n,P))$, which do not affect exponential rates.

Proof of Theorem 2. **Geometric Preliminaries.** By the Hilbert–Schmidt representation and Hölder's inequality $|\mathrm{Tr}(AB)| \le \|A\|_1\|B\|_1$, the components of the pullback metric $\tilde{g}$ and its coordinate derivatives satisfy:

$$|\tilde{g}_{(l,i),(m,j)}| \le \|\partial_{(l,i)}\tilde{\rho}_{\psi_{\boldsymbol{\theta}}}\|_1 \cdot \|\partial_{(m,j)}\tilde{\rho}_{\psi_{\boldsymbol{\theta}}}\|_1 \in \mathcal{O}(q^{2\alpha})\,, \tag{35}$$

$$|\partial_{(n,k)}\tilde{g}_{(l,i),(m,j)}| \leq \|\partial^2_{(n,k),(l,i)}\tilde{\rho}_{\psi_{\boldsymbol{\theta}}}\|_1 \|\partial_{(m,j)}\tilde{\rho}_{\psi_{\boldsymbol{\theta}}}\|_1 + \|\partial_{(l,i)}\tilde{\rho}_{\psi_{\boldsymbol{\theta}}}\|_1 \|\partial^2_{(n,k),(m,j)}\tilde{\rho}_{\psi_{\boldsymbol{\theta}}}\|_1 \in \mathcal{O}(q^{2\alpha})\,, \tag{36}$$

where both bounds follow directly from Lemmas 11 and 12. Consequently, the maximal eigenvalue of the metric satisfies $\lambda_{\max}(\tilde{g}) \leq \|\tilde{g}\|_F \leq C_g q^{2\alpha}$. For the regularized metric $\tilde{g}^{(\epsilon)} = \tilde{g} + \epsilon I$, every eigenvalue is bounded as $\epsilon \leq \lambda_i(\tilde{g}^{(\epsilon)}) \leq C_g q^{2\alpha} + \epsilon$. Inverting this positive definite matrix yields:

$$\|(\tilde{g}^{(\epsilon)})^{-1}\|_2 = \frac{1}{\lambda_{\min}(\tilde{g}^{(\epsilon)})} \leq \frac{1}{\epsilon}\,, \qquad \lambda_{\min}\big((\tilde{g}^{(\epsilon)})^{-1}\big) = \frac{1}{\lambda_{\max}(\tilde{g}^{(\epsilon)})} \geq \frac{1}{C_g q^{2\alpha} + \epsilon}\,. \tag{37}$$

In the informative weak-regularization regime ($\epsilon \ll q^{2\alpha}$), $\|(\tilde{g}^{(\epsilon)})^{-1}\|_2 \in \mathcal{O}(q^{-2\alpha})$ and $\lambda_{\min}((\tilde{g}^{(\epsilon)})^{-1}) \in \Omega(q^{-2\alpha})$.

Statement (i): Geometric Invariance. The Levi–Civita connection coefficients are $\tilde{\Gamma}^{(n,k)}_{(l,i),(m,j)} = \frac{1}{2}\sum_{(p,r)}(\tilde{g}^{(\epsilon)})^{-1}_{(n,k),(p,r)}(\partial_{(l,i)}\tilde{g}_{(m,j),(p,r)} + \partial_{(m,j)}\tilde{g}_{(l,i),(p,r)} - \partial_{(p,r)}\tilde{g}_{(l,i),(m,j)})$. Since each metric derivative contracts as $\mathcal{O}(q^{2\alpha})$ while the inverse metric scales as $\mathcal{O}(q^{-2\alpha})$, the exponential factors cancel algebraically:

$$\|\tilde{\Gamma}^{(n,k)}\|_2 \in \mathcal{O}(q^{-2\alpha}) \cdot \mathcal{O}(q^{2\alpha}) = \mathcal{O}(1)\,. \tag{38}$$

The affine connection, geodesics, and parallel transport are asymptotically preserved under noise.

Statement (ii): Riemannian Gradient Bifurcation (NIRS vs. NIBP). The natural gradient is $grad_{\tilde{g}^{(\epsilon)}}\tilde{\mathcal{L}} = -(\tilde{g}^{(\epsilon)})^{-1}d\tilde{\mathcal{L}}$. From Lemma 10, the Euclidean gradient covector contracts as $\|d\tilde{\mathcal{L}}\|_2 \leq C_{\mathcal{L}} q^{\alpha}$.

Upper Bound: By submultiplicativity of the operator norm:

$$\|grad_{\tilde{g}^{(\epsilon)}}\tilde{\mathcal{L}}\|_2 \leq \|(\tilde{g}^{(\epsilon)})^{-1}\|_2 \cdot \|d\tilde{\mathcal{L}}\|_2 \in \begin{cases} \mathcal{O}(q^{-2\alpha}) \cdot \mathcal{O}(q^{\alpha}) = \mathcal{O}(q^{-\alpha}) & \text{if } \epsilon \ll q^{2\alpha}\,, \\ \mathcal{O}(1/\epsilon) \cdot \mathcal{O}(q^{\alpha}) = \mathcal{O}(q^{\alpha}) & \text{if } \epsilon \gg q^{2\alpha}\,. \end{cases} \tag{39}$$

Lower Bound (Riemannian Spike): Because $(\tilde{g}^{(\epsilon)})^{-1}$ is symmetric positive definite, its action on any vector is bounded from below by its minimum eigenvalue:

$$\|grad_{\tilde{g}^{(\epsilon)}}\tilde{\mathcal{L}}\|_2 = \|(\tilde{g}^{(\epsilon)})^{-1}d\tilde{\mathcal{L}}\|_2 \geq \lambda_{\min}\big((\tilde{g}^{(\epsilon)})^{-1}\big)\|d\tilde{\mathcal{L}}\|_2 \geq \frac{\|d\tilde{\mathcal{L}}\|_2}{C_g q^{2\alpha} + \epsilon}\,. \tag{40}$$

For any non-stationary parameter point where the Euclidean covector preserves non-vanishing signal $\|d\tilde{\mathcal{L}}\|_2 \geq c_{\mathcal{L}} q^{\alpha}$, setting $\epsilon \ll q^{2\alpha}$ yields:

$$\|grad_{\tilde{g}^{(\epsilon)}}\tilde{\mathcal{L}}\|_2 \geq \frac{c_{\mathcal{L}} q^{\alpha}}{2C_g q^{2\alpha}} \in \Omega(q^{-\alpha})\,. \tag{41}$$

Combining the matching upper and lower bounds yields $\|grad_{\tilde{g}^{(\epsilon)}}\tilde{\mathcal{L}}\|_2 \in \Theta(q^{-\alpha}) = \Theta(\kappa_{\text{err}}^{-L})$. Under heavy regularization ($\epsilon \gg q^{2\alpha}$), the regularizer dominates, $(\tilde{g}^{(\epsilon)})^{-1} \to \frac{1}{\epsilon} I$, collapsing the update back into the Euclidean barren plateau $\mathcal{O}(\kappa_{\text{err}}^{L})$.

□

H Full DFS-QNG Results

This appendix provides the complete convergence tables and trajectories supporting the empirical claims. The purpose is not only to report additional benchmark outcomes, but also to show the same qualitative pattern across problem families: as the number of qubits and strata increases, standard QNG becomes increasingly unstable, while DFS-QNG remains more stable because it avoids global metric inversion and restricts each update to a commuting local geometry.

Across these additional results, the same pattern appears as in the main text: standard QNG becomes increasingly sensitive as the circuit dimension and number of strata grow, while DFS-QNG

Table 2: Normalized Area Above Optimal (↓) on QAOA MaxCut ring graph. Budget equals total epochs (30). Best per cell in **bold**.

n	Optimizer	2 Strata	4 Strata	6 Strata	8 Strata	10 Strata
2	DFS-QNG	**0.10**±0.04	**0.05**±0.02	**0.04**±0.03	**0.05**±0.05	**0.03**±0.01
	QNG	0.17±0.16	0.64±0.61	1.43±0.99	2.00±1.00	2.09±0.91
	Adam	0.30±0.20	0.17±0.12	0.13±0.08	0.17±0.11	0.20±0.11
4	DFS-QNG	**2.12**±0.05	**0.26**±0.09	**0.09**±0.02	**0.14**±0.09	**0.07**±0.02
	QNG	2.14±0.08	0.38±0.22	0.95±0.43	2.02±0.82	2.61±0.92
	Adam	2.23±0.14	0.94±0.38	0.60±0.22	0.45±0.12	0.41±0.15
6	DFS-QNG	**2.36**±0.20	**2.28**±0.12	**0.57**±0.26	**0.35**±0.14	**0.27**±0.06
	QNG	2.52±0.40	2.34±0.14	1.17±0.58	1.99±0.76	2.58±0.60
	Adam	2.92±0.30	2.66±0.15	1.07±0.39	1.18±0.26	0.77±0.20
8	DFS-QNG	**3.41**±0.21	**2.64**±0.46	**2.37**±0.10	**1.17**±0.66	**0.63**±0.32
	QNG	3.50±0.27	2.82±0.44	2.69±0.28	3.35±0.83	4.08±0.81
	Adam	4.10±0.38	3.41±0.33	2.97±0.18	2.45±0.42	1.97±0.82
10	DFS-QNG	**4.59**±0.23	**3.27**±0.32	**2.52**±0.15	**2.51**±0.13	**1.49**±0.80
	QNG	4.63±0.24	3.42±0.16	3.15±0.38	4.29±0.56	5.43±0.76
	Adam	5.20±0.25	4.23±0.41	3.49±0.44	3.16±0.24	2.95±0.24

Table 3: Normalized Area Above Optimal (↓) on VQE Ising Model. Budget equals 30 epochs. Best per cell in **bold**.

n	Optimizer	2 Strata	4 Strata	6 Strata	8 Strata	10 Strata
2	DFS-QNG	0.88±0.11	0.65±0.05	**0.14**±0.09	**0.10**±0.06	**0.08**±0.03
	QNG	0.88±0.11	0.67±0.06	1.08±0.41	1.49±0.69	2.14±0.92
	Adam	1.46±0.68	1.23±0.42	0.60±0.24	0.43±0.21	0.58±0.32
	EGT	**0.54**±0.30	**0.54**±0.30	0.54±0.30	0.54±0.30	0.54±0.30
4	DFS-QNG	1.67±0.32	1.33±0.23	0.94±0.29	0.53±0.20	**0.38**±0.17
	QNG	1.67±0.32	1.39±0.27	1.85±0.66	1.97±0.71	2.17±0.62
	Adam	2.76±0.52	2.49±0.84	2.26±0.55	1.49±0.33	0.95±0.17
	EGT	**0.48**±0.09	**0.48**±0.09	**0.48**±0.09	**0.48**±0.09	0.48±0.09
6	DFS-QNG	2.76±1.37	3.16±1.48	2.24±0.56	1.73±0.39	1.48±0.28
	QNG	2.76±1.37	3.25±1.49	3.15±0.69	3.34±0.71	3.76±0.83
	Adam	4.32±0.67	4.23±1.54	3.75±1.05	3.59±0.81	3.43±0.95
	EGT	**0.70**±0.12	**0.70**±0.12	**0.70**±0.12	**0.70**±0.12	**0.70**±0.12
8	DFS-QNG	3.61±1.00	3.96±0.65	3.67±1.02	3.33±0.93	3.56±1.08
	QNG	3.61±1.00	4.37±0.94	4.34±0.97	4.24±0.65	5.83±1.31
	Adam	5.12±0.67	5.15±0.99	5.26±0.70	5.50±1.11	5.27±0.74
	EGT	**1.91**±1.04	**1.91**±1.04	**1.91**±1.04	**1.91**±1.04	**1.91**±1.04
10	DFS-QNG	3.90±0.75	5.05±1.60	4.45±0.70	5.49±1.37	6.02±1.22
	QNG	3.90±0.75	5.19±1.60	5.25±0.64	6.43±1.04	7.47±1.16
	Adam	5.87±1.17	6.63±1.07	6.85±1.03	7.40±1.18	7.27±1.08
	EGT	**1.33**±0.21	**1.33**±0.21	**1.33**±0.21	**1.33**±0.21	**1.33**±0.21

maintains smoother descent trajectories. These results support the central diagnosis of the paper: the issue is not that the quantum statistical manifold lacks useful geometry, but that the full QNG update uses a global metric inverse whose conditioning and cross-parameter coupling become unstable at scale. DFS-QNG improves robustness by localizing each update to a commuting stratum. We organize the result as follows:

The NAAO metric results are reported in Tables: 1, 2, and 3.

The stable energy and epoch when reached it metric is reported in Tables 4, 5 and 6.

The full training trajectories on the Maxcut 3-regular graph problem is shown in Figure 4.

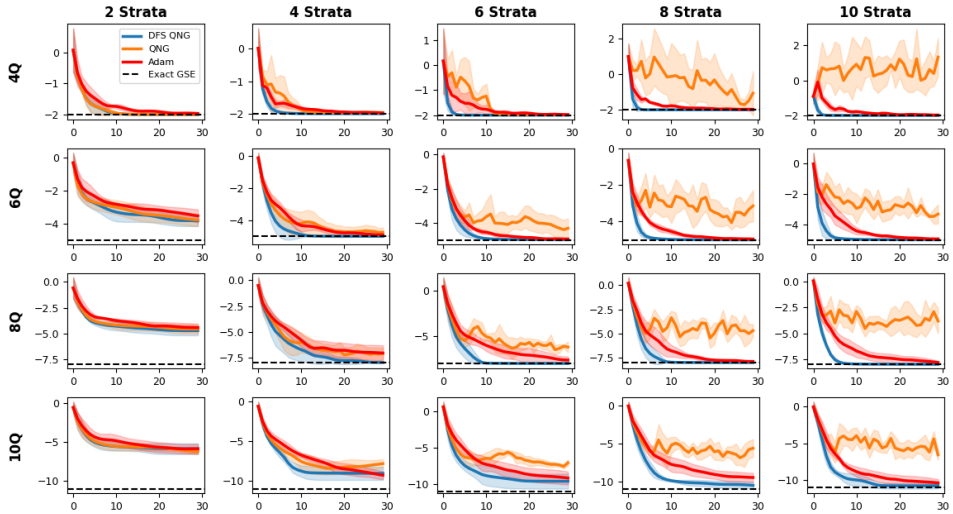


Figure 4: DFS-QNG vs QNG on QAOA Maxcut3 problem with $n = \{4, 6, 8, 10\}$ qubits and $\{2, 4, 6, 8, 10\}$ stratums. The X-axis represents the number of Epochs.

Table 4: Stable Energy Reached and Epoch of Convergence on QAOA MaxCut3. A run is considered stable if it remains within 10% of its minimum energy for 5 consecutive epochs. Best stable energy per cell in bold.

n	Optimizer	1 Stratum	2 Strata	3 Strata	4 Strata	5 Strata
4	DFS-QNG	**-1.84** @ 7	**-1.83** @ 3	**-1.87** @ 2	**-1.87** @ 2	**-1.96** @ 2
	QNG	-1.82 @ 7	-1.80 @ 9	-1.79 @ 11	N/A	N/A
	Adam	-1.80 @ 12	-1.80 @ 9	-1.82 @ 11	-1.79 @ 11	-1.80 @ 9
6	DFS-QNG	**-3.45** @ 14	**-4.55** @ 6	**-4.62** @ 6	**-4.58** @ 3	**-4.72** @ 4
	QNG	-3.44 @ 17	-4.33 @ 14	-4.02 @ 12	N/A	N/A
	Adam	-3.19 @ 20	-4.47 @ 14	-4.53 @ 9	-4.47 @ 9	-4.49 @ 11
8	DFS-QNG	**-4.26** @ 9	**-7.31** @ 13	**-7.31** @ 7	**-7.37** @ 6	**-7.55** @ 5
	QNG	-4.15 @ 9	-6.67 @ 13	N/A	N/A	N/A
	Adam	-4.02 @ 15	-6.49 @ 12	-6.93 @ 17	-7.23 @ 13	-7.18 @ 13
10	DFS-QNG	-5.37 @ 7	-8.30 @ 8	**-8.68** @ 9	**-9.61** @ 8	**-9.83** @ 8
	QNG	**-5.59** @ 12	-7.78 @ 11	-6.84 @ 20	N/A	N/A
	Adam	-5.32 @ 15	**-8.41** @ 19	-8.29 @ 17	-8.63 @ 16	-9.39 @ 14

Table 5: Stable Energy Reached and Epoch of Convergence on QAOA MaxCut ring graph. A run is considered stable if it remains within 10% of its minimum energy for 5 consecutive epochs. Best stable energy per cell in **bold**.

n	Optimizer	1 Stratum	2 Strata	3 Strata	4 Strata	5 Strata
2	DFS-QNG	-1.95 @ 3	**-2.00** @ 2	**-2.00** @ 2	**-2.00** @ 2	**-2.00** @ 1
	QNG	**-1.99** @ 6	N/A	N/A	N/A	N/A
	Adam	-1.81 @ 14	-1.82 @ 8	-1.79 @ 4	-1.88 @ 10	-1.85 @ 9
4	DFS-QNG	-1.93 @ 4	**-3.90** @ 4	**-3.92** @ 2	**-3.97** @ 3	**-3.75** @ 1
	QNG	**-1.93** @ 5	-3.73 @ 5	-3.32 @ 23	N/A	N/A
	Adam	-1.83 @ 12	-3.62 @ 16	-3.60 @ 10	-3.57 @ 7	-3.57 @ 8
6	DFS-QNG	**-3.63** @ 7	**-3.63** @ 4	**-5.52** @ 8	**-5.61** @ 4	**-5.61** @ 3
	QNG	-3.58 @ 9	-3.62 @ 5	-5.36 @ 12	N/A	-3.71 @ 15
	Adam	-3.33 @ 15	-3.60 @ 14	-5.34 @ 15	-5.37 @ 14	-5.41 @ 10
8	DFS-QNG	-4.52 @ 6	**-5.40** @ 6	**-5.46** @ 4	**-6.94** @ 10	**-7.29** @ 7
	QNG	**-4.52** @ 7	-5.29 @ 8	-5.30 @ 5	N/A	N/A
	Adam	-4.44 @ 18	-5.07 @ 16	-5.33 @ 13	-6.23 @ 20	-6.47 @ 12
10	DFS-QNG	-5.47 @ 6	**-6.64** @ 5	**-7.34** @ 6	-7.20 @ 5	**-8.34** @ 8
	QNG	**-5.58** @ 7	-6.49 @ 5	-6.82 @ 7	-6.67 @ 24	N/A
	Adam	-5.35 @ 15	-6.33 @ 16	-6.77 @ 11	**-7.24** @ 13	-7.46 @ 12

Table 6: Stable Energy Reached and Epoch of Convergence on VQE Ising Model. A run is considered stable if it remains within 10% of its minimum energy for 5 consecutive epochs. Budget equals 30 epochs. Best stable energy per cell in **bold**.

n	Optimizer	2 Strata	4 Strata	6 Strata	8 Strata	10 Strata
2	DFS-QNG	-1.82 @ 3	-2.08 @ 2	**-2.57** @ 4	**-2.62** @ 3	**-2.70** @ 2
	QNG	-1.82 @ 3	-2.05 @ 3	N/A	N/A	N/A
	Adam	-1.79 @ 19	-1.95 @ 17	-2.39 @ 13	-2.53 @ 14	-2.51 @ 14
	EGT	**-2.36** @ 13	**-2.36** @ 13	-2.36 @ 13	-2.36 @ 13	-2.36 @ 13
4	DFS-QNG	-3.81 @ 8	-4.01 @ 6	-4.28 @ 6	**-4.71** @ 5	**-4.72** @ 4
	QNG	-3.81 @ 8	-4.01 @ 7	-3.61 @ 12	N/A	-3.66 @ 15
	Adam	-3.52 @ 19	-3.59 @ 23	-4.03 @ 19	-4.09 @ 15	-4.52 @ 12
	EGT	**-4.52** @ 6	**-4.52** @ 6	**-4.52** @ 6	-4.52 @ 6	-4.52 @ 6
6	DFS-QNG	-5.50 @ 24	-5.00 @ 10	-6.19 @ 13	-6.19 @ 9	-6.53 @ 8
	QNG	-5.50 @ 24	-4.84 @ 10	-5.18 @ 11	N/A	N/A
	Adam	-5.30 @ 21	-4.65 @ 24	-4.92 @ 19	-5.36 @ 23	-5.25 @ 21
	EGT	**-6.83** @ 8	**-6.83** @ 8	**-6.83** @ 8	**-6.83** @ 8	**-6.83** @ 8
8	DFS-QNG	-7.23 @ 14	-7.67 @ 13	-7.64 @ 15	-7.53 @ 10	-7.84 @ 12
	QNG	-7.23 @ 14	-7.57 @ 18	-7.53 @ 20	-6.78 @ 16	N/A
	Adam	-7.13 @ 21	-6.83 @ 21	-6.53 @ 18	-6.23 @ 22	-7.37 @ 23
	EGT	**-8.87** @ 16	**-8.87** @ 16	**-8.87** @ 16	**-8.87** @ 16	**-8.87** @ 16
10	DFS-QNG	-9.02 @ 9	-8.63 @ 12	-9.35 @ 11	-8.81 @ 16	-8.05 @ 13
	QNG	-9.02 @ 9	-8.74 @ 13	-8.65 @ 11	-8.62 @ 23	N/A
	Adam	-9.01 @ 20	-8.59 @ 22	-7.86 @ 20	-7.60 @ 22	-7.57 @ 21
	EGT	**-11.39** @ 11	**-11.39** @ 11	**-11.39** @ 11	**-11.39** @ 11	**-11.39** @ 11