

Adversarial Effects on Expressibility and Trainability in Distributed Variational Quantum Algorithms

Abhishek Sadhu^{1,*} and Sharu Theresa Jose¹

¹*School of Computer Science, University of Birmingham,
Edgbaston, Birmingham B15 2TT, United Kingdom*

(Dated: June 22, 2026)

Distributed quantum algorithms offer a promising pathway to scale variational quantum algorithms beyond the constraints of noisy intermediate-scale quantum hardware. However, existing approaches implicitly assume a trusted entanglement-sharing layer across quantum processors. We show that this assumption introduces a fundamental vulnerability: adversarial perturbations of shared entanglement induce structured gate-level noise that directly impacts quantum learning. We develop a framework that maps entanglement-level perturbations to gate-level noise via an explicit Kraus representation. To quantify their impact, we introduce Kraus expressibility, a metric that generalizes unitary expressibility to noisy quantum channels. We then establish a trade-off between Kraus expressibility and trainability of noisy quantum circuits through gradient variance analysis. Our analysis reveals that an adversary can manipulate Kraus expressibility to maintain sufficiently large cost gradients (avoiding barren plateaus) while systematically biasing optimization toward incorrect solutions. We validate these findings through numerical simulations, demonstrating adversarial degradation of expressibility and trainability.

I. INTRODUCTION

The realization of a desirably functional quantum machine learning (QML) framework [1, 2] beyond the limitations of near-term hardware is a central challenge for the field. In this era of noisy quantum devices, variational quantum algorithms (VQAs) [3, 4] have emerged as a promising framework for achieving practical quantum utility [5]. However, their scalability remains fundamentally constrained by the limitations of currently available noisy-intermediate scale quantum (NISQ) [6] hardware. *Distributed quantum computing* [7, 8] offers a pathway to overcome these hardware limitations by partitioning a quantum circuit across multiple, small-scale NISQ quantum processing units (QPUs), thereby enabling the execution of circuits beyond the capacity of single devices.

Distributed VQAs however requires sharing quantum correlations across QPUs to implement non-local multi-qubit quantum gates across them. This architecture introduces new vulnerabilities. Adversarial perturbations to the shared entanglement can corrupt non-local gate execution, leading to cascading errors across distributed quantum circuits. From a machine learning perspective, this vulnerability is particularly critical. VQAs rely on precise circuit evaluations to estimate gradients and optimize model parameters. The gate noise introduced via the entanglement-level perturbations directly affect circuit expressibility, and optimization dynamics. *Despite this, the effect of entanglement-level attacks on distributed quantum learning remains largely unexplored.*

We address this gap by introducing a novel adversarial framework for distributed VQAs, where the attacker perturbs the shared entangled qubit pair underlying inter-device communication. We focus on the *cat-entangler/cat-disentangler* protocol [9], which implements distributed, non-local CNOT gates using a single entangled qubit pair and two bits of classical communication. As a minimal and widely studied primitive for distributed quantum computation, it provides a natural setting for adversarial analysis. Our main contributions [10] are as follows:

(1) We show that perturbations to shared entanglement between QPUs systematically corrupt non-local multi-qubit gates, transforming ideal unitary gates into *noisy quantum channels*. We derive an explicit Kraus representation of the resulting channel, parameterized by the adversary, providing a principled mapping from entanglement-level attacks to gate-level noise.

(2) We present *Kraus expressibility*, a novel expressibility measure that quantifies how well ‘noisy’ quantum circuits can uniformly explore the space of ideal unitaries. This generalizes *unitary expressibility* for noise-less PQCs [11] to account for both adversarial and intrinsic hardware noise. We show that adversarial perturbations degrade Kraus expressibility through two mechanisms: (i) state purity degradation and (ii) noise-induced cross-correlations. Our analysis introduces a framework to evaluate vulnerability of PQCs to stealthy noise manipulation.

(3) We present a theoretical link between Kraus expressibility and *trainability* of noisy PQCs by cost gradient variance analysis. We show that while noise typically induces barren plateaus, an adversary can balance noise to

* a.sadhu@bham.ac.uk

evade barren plateaus while systematically biasing the optimization landscape.

(4) Through numerical simulations, we validate our theory and show that weakly cross-correlated, adversarial noise accelerates the decay of both the Kraus expressibility norm and gradient variance. In particular, using the hardware-efficient ansatz as a testbed, we demonstrate that increasing noise induces noise-induced barren plateaus even in otherwise trainable shallow circuits, and completely overpowers mitigation strategies such as restricted parameter initialization and local cost functions, ultimately rendering the ansatz untrainable at moderate system sizes.

II. KRAUS EXPRESSIBILITY FOR NOISY QUANTUM CIRCUITS

The success of Variational Quantum Algorithms depends critically on the choice of the sequence of transformations applied to the initial state. In the presence of adversarially-induced noise, the effective transformations are quantum channels. To capture this, we generalize the standard unitary-ansatz to a parameterized quantum channel (*PQCh*)-ansatz

$$\mathcal{E}_\theta^P = \mathcal{K}_L(\mathbf{p}_L, \vec{\theta}_L) \circ \mathcal{K}_{L-1}(\mathbf{p}_{L-1}, \vec{\theta}_{L-1}) \circ \dots \circ \mathcal{K}_1(\mathbf{p}_1, \vec{\theta}_1), \quad (1)$$

where $\mathbf{p}_j$ and $\vec{\theta}_j$ denote the noise and trainable parameters of the j -th channel $\mathcal{K}_j$, respectively. For fixed noise parameter $\mathbf{p} = (\mathbf{p}_1, \dots, \mathbf{p}_L)$, varying $\vec{\theta}$ induces an ensemble $\mathcal{T}_\mathbf{p} = \{\mathcal{E}_\theta^P : \vec{\theta} \in \Theta\}$ of fixed-noise trainable channels. Further variation over noise parameters $\mathbf{p} \in \Pi$ yields the full ensemble $\mathcal{T} = \{\mathcal{T}_\mathbf{p} : \mathbf{p} \in \Pi\}$ of noisy trainable transformations.

In the adversarial setting, the adversarially-induced noise can distort the exploration of the ansatz and drive the PQCh-ansatz away from the optimal transformation. This motivates us to define *Kraus expressibility*, which quantifies how well noisy quantum circuits explore the space of ideal unitary transformations.

Definition 1. *The Kraus expressibility of the PQCh-ansatz is defined via the following super-operator,*

$$\mathcal{A}_\mathcal{T}^t(\cdot) := \int_{\mathcal{U}(d)} d\mu(V) V^{\otimes t}(\cdot) (V^\dagger)^{\otimes t} - \int_{\mathcal{T}} d\nu(E) \sum_{k_1 \dots k_t} (\otimes_{j=1}^t E_k^{(j)}) (\cdot) (\otimes_{j=1}^t E_k^{(j)\dagger}), \quad (2)$$

where $d\mu(V)$ is the volume element of the Haar measure over the unitary group $\mathcal{U}(d)$ with $d = 2^n$, $d\nu(E)$ is the volume element corresponding to uniform distribution over the PQCh-ensemble $\mathcal{T}$, and $E_k^{(j)}$ denote the k -th Kraus operator of the PQCh-ansatz for the j -th copy.

Intuitively, Kraus expressibility quantifies how well the PQCh-ensemble $\mathcal{T}$ uniformly explores the unitary group $\mathcal{U}(d)$. In particular, if $\mathcal{A}_\mathcal{T}^t(X) = 0$ for all operators X , then averaging over $\mathcal{T}$ matches Haar averaging over $\mathcal{U}(d)$ upto the t -th moment, i.e., $\mathcal{T}$ forms a t -design. In this work, we focus on $t = 2$. For VQA optimization, we evaluate the Kraus expressibility of $\mathcal{T}$ with respect to the input state ρ via the norm

$$\Delta_\mathcal{T}^\rho := \|\mathcal{A}_\mathcal{T}^2(\rho^{\otimes 2})\|_2. \quad (3)$$

Smaller values of $\Delta_\mathcal{T}^\rho$ indicate higher Kraus expressibility. This extends unitary expressibility [11] to noisy and adversarial settings.

Theorem 1. *The Kraus expressibility norm of the PQCh-ansatz with respect to the state ρ is given by*

$$(\Delta_\mathcal{T}^\rho)^2 = (\alpha^2 + \beta^2) d^2 + 2\alpha\beta d - 2[\alpha + \beta\bar{\nu}] + \mathcal{N}_{noise}, \quad (4)$$

where $\alpha = \frac{d - \text{Tr} \rho^2}{d(d^2 - 1)}$, $\beta = \frac{d \text{Tr} \rho^2 - 1}{d(d^2 - 1)}$, $\bar{\nu}$ is the ensemble averaged purity¹ of the input state ρ after evolving through the PQCh-ansatz, and $\mathcal{N}_{noise}$ is the noise fluctuation term defined as $\mathcal{N}_{noise} = \int_{\mathcal{T}} \int_{\mathcal{T}} d\tau(E) d\nu(F) \left[\sum_{j,k} \text{Tr}(E_k \rho E_k^\dagger F_j \rho F_j^\dagger) \right]^2$ with $\{E_k\}$ and $\{F_k\}$ representing Kraus operators corresponding to independent channel realizations sampled from $\mathcal{T}$.

Proof is provided in [10]. Intuitively, a highly Kraus expressive PQCh-ensemble requires both high average purity (to preserve the quantum structure) and sufficiently diverse channel realizations (to explore the transformation space). Adversarial noise can disrupt either (or both), making the resulting ansatz less capable of approximating target solutions. Notably, even with negligible purity loss, strong noise-induced correlations can significantly degrade expressibility. In the absence of adversarial noise, these effects vanish, and Kraus expressibility reduces to standard unitary expressibility.

¹ A quantum state ρ is pure if $\text{Tr}(\rho^2) = 1$ and is mixed if $\text{Tr}(\rho^2) < 1$.

Corollary 1. When the PQCh-ensemble reduces to a composite channel $\mathcal{E}_\theta^p$ with fixed $(\mathbf{p}, \vec{\theta})$, the Kraus expressibility norm with respect to the input state ρ is given by

$$(\Delta_{\mathcal{E}_\theta^p}^\rho)^2 = (\alpha^2 + \beta^2)d^2 + 2\alpha\beta d - 2[\alpha + \beta\nu] + \nu^2, \quad (5)$$

where ν is the purity of the state after evolving through $\mathcal{E}_\theta^p$, and (α, β) are as defined in Theorem 1.

In the post-training regime of VQAs, the expressibility norm $(\Delta_{\mathcal{E}_\theta^p}^\rho)^2$ depends solely on the output state purity, which is governed by the noise parameters $\mathbf{p}$. In this regime, the ensemble-level noise correlation term $\mathcal{N}_{\text{noise}}$ in Theorem 1 collapses to a function of ν . Moreover, since $(\Delta_{\mathcal{E}_\theta^p}^\rho)^2$ is monotonically increasing in ν , increasing noise (which reduces purity) decreases $(\Delta_{\mathcal{E}_\theta^p}^\rho)^2$.

III. TRAINABILITY-EXPRESSIBILITY TRADE-OFFS FOR NOISY QUANTUM CIRCUITS

While Kraus expressibility determines whether a PQCh-ansatz can represent the target solution, it does not guarantee that the solution can be *efficiently learned*. Successful optimization additionally requires the ansatz to be *trainable*, i.e., the cost landscape must exhibit sufficiently large gradients for the solution to be found.

For a PQCh-ansatz as in Eq. (1) with input state ρ_0 , we define the cost function as $C_{\rho_0, \mathcal{H}} = \text{Tr}[\mathcal{H} \mathcal{E}_\theta^p]$. We next bound the deviation of the expected gradient variance via the Kraus expressibility norm.

Theorem 2. Consider the generic cost function $C_{\rho_0, \mathcal{H}}$ for a PQCh-ansatz as in Eq. (1). Then, the deviation of the expected gradient variance from the reference variance $\text{var}_{\mathcal{R}}(\partial_k C_{\rho_0, \mathcal{H}})$ is upper bounded as

$$|\mathbb{E}_{\mathcal{E}_\mathcal{L} \sim \mathcal{T}_\mathcal{L}, \mathcal{E}_\mathcal{R} \sim \mathcal{T}_\mathcal{R}}[\text{var}_{\vec{\theta}_k}(\partial_k C_{\rho_0, \mathcal{H}})] - \text{var}_{\mathcal{R}}(\partial_k C_{\rho_0, \mathcal{H}})| \leq 4\|\mathcal{A}_{\mathcal{T}_\mathcal{R}}^2(\rho_0^{\otimes 2})\|_2 \int_{\mathcal{T}_\mathcal{L}} d\nu(\mathcal{E}_\mathcal{L}) \|\mathcal{E}_\mathcal{L}^\dagger(\mathcal{H})\|_2^2. \quad (6)$$

Proof is provided in [10]. Theorem 2 provides a direct, analytical link between the trainability and expressibility of the PQCh-ansatz. A strong adversary perturbs shared entanglement to induce noise in non-local gates, reducing $\bar{\nu}$ by δ while remaining ϵ -indistinguishable from ideal CNOT implementation. For a PQCh-ansatz with L layers, this results in $\bar{\nu} \approx (1 - \mathcal{O}(\delta^2))^L$. While reduced purity affects Kraus expressibility, Theorem 2 reveals a layered asymmetry in trainability. For parameters near the end of the PQCh-ansatz, trainability is bottlenecked by preceding layers approaching a 2-design, driving the expressibility norm $\|\mathcal{A}_{\mathcal{T}_\mathcal{R}}^{(2)}(\rho_0^{\otimes 2})\|_2$ towards zero. In contrast, for parameters near the beginning of the ansatz, the trainability is limited by the observable attenuation factor, $\|\mathcal{E}_\mathcal{L}^\dagger(\mathcal{H})\|_2^2$. A budget-constrained adversary can exploit this symmetry: attacking early layers flatten the cost landscape via signal attenuation, while attacking later layers suppress gradients through increased expressibility. By calibrating these effects, the adversary can systematically bias the optimization objective, degrading the final solution quality.

IV. DISCUSSION

Distributed implementations of VQAs rely critically on shared entanglement across QPUs, which is typically assumed to be secure. We show that this assumption introduces a fundamental vulnerability. We develop an adversarial framework that maps entanglement-level perturbations to gate-level noise via an explicit Kraus representation. To quantify the impact of such perturbations, we introduce Kraus expressibility, a metric that extends unitary expressibility to noisy and adversarial settings. Our analysis establishes a principled connection between Kraus expressibility and trainability by bounding the variance of cost-function gradients. This reveals a key insight: by balancing purity loss and noise-induced correlations, a stealthy adversary can avoid barren plateaus, by maintaining non-vanishing gradients, while systematically biasing the optimization towards incorrect solutions. Numerical results corroborate this behavior and show that standard mitigation strategies, such as restricted parameter initialization and local cost functions, can fail under adversarial noise.

[1] Dan Ventura and Tony Martinez. A quantum associative memory based on grover's algorithm. In *Artificial Neural Nets and Genetic Algorithms: Proceedings of the International Conference in Portorož, Slovenia, 1999*, pages 22–27. Springer, 1999.

- [2] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017.
- [3] Jarrod R McClean, Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik. The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics*, 18(2):023023, 2016.
- [4] 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.
- [5] Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermann Heimonen, Jakob S Kottmann, Tim Menke, et al. Noisy intermediate-scale quantum algorithms. *Reviews of Modern Physics*, 94(1):015004, 2022.
- [6] John Preskill. Quantum computing in the nisq era and beyond. *Quantum*, 2:79, 2018.
- [7] J Ignacio Cirac, AK Ekert, Susana F Huelga, and Chiara Macchiavello. Distributed quantum computation over noisy channels. *Physical Review A*, 59(6):4249, 1999.
- [8] Stephen DiAdamo, Marco Ghibaudi, and James Cruise. Distributed quantum computing and network control for accelerated vqe. *IEEE Transactions on Quantum Engineering*, 2:1–21, 2021.
- [9] Anocha Yimsiriwattana and Samuel J Lomonaco Jr. Generalized ghz states and distributed quantum computing. *arXiv preprint quant-ph/0402148*, 2004.
- [10] Abhishek Sadhu and Sharu Theresa Jose. Adversarial effects on expressibility and trainability in distributed variational quantum algorithms. *arXiv preprint arXiv:2605.03629*, 2026.
- [11] Zoë Holmes, Kunal Sharma, Marco Cerezo, and Patrick J Coles. Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *PRX quantum*, 3(1):010313, 2022.