Prospects of Privacy Advantage in Quantum Machine Learning

Jamie Heredge,[1, 2, *] Niraj Kumar,[1] Dylan Herman,[1] Shouvanik Chakrabarti,[1]
Romina Yalovetzky,[1] Shree Hari Sureshbabu,[1] Changhao Li,[1] and Marco Pistoia[1]

[1] *Global Technology Applied Research, JPMorgan Chase, New York, NY 10017*
[2] *School of Physics, The University of Melbourne, Parkville, VIC 3010, Australia*
(Dated: May 24, 2024)

I. OVERVIEW

Ensuring data privacy in machine learning models is critical, particularly in distributed settings where model gradients are typically shared among multiple parties to allow collaborative learning. Motivated by the increasing success of gradient inversion attacks recovering private and potentially confidential input data from the gradients of classical models, this study addresses a central question: *How hard is it to recover the input data from the gradients of quantum machine learning models?* Focusing on variational quantum circuits (VQC) as learning models, we uncover the crucial role played by the dynamical Lie algebra (DLA) of the VQC ansatz in determining privacy vulnerabilities. While the DLA has previously been linked to the classical simulatability and trainability of VQC models, this work, for the first time, establishes its connection to the privacy of VQC models. In particular, we show that properties conducive to the trainability of VQCs, such as a polynomial-sized DLA, also facilitate the extraction of detailed *snapshots* of the input. We term this a weak privacy breach, as the snapshots enable training VQC models for distinct learning tasks without direct access to the original input. Further, we investigate the conditions for a strong privacy breach where the original input data can be recovered from these snapshots by classical or quantum-assisted polynomial time methods. We establish conditions on the encoding map such as classical simulatability, overlap with DLA basis, and its Fourier frequency characteristics that enable such a privacy breach of VQC models. Our findings thus play a crucial role in detailing the prospects of quantum privacy advantage by guiding the requirements for designing quantum machine learning models that balance trainability with robust privacy protection.

ArXiv Pre-print of this work [1] for additional details is available at : `https://arxiv.org/abs/2405.08801`

II. SUMMARY OF KEY RESULTS

In the contemporary technological landscape, data privacy concerns command increasing attention, particularly within the domain of machine learning (ML) models that are trained on sensitive datasets, such as financial records, healthcare information or geo-location data [2–7]. The multi-national adoption of stringent legal frameworks [8] has further amplified the urgency to improve data privacy. In distributed learning settings, where many different nodes can train models collaboratively, the user does not need to send their confidential data, but rather their training gradients. This was originally proposed as a solution to enable collaborative learning while preventing data leakage [9]. However, subsequent works have shown that neural networks are particularly susceptible to gradient inversion-based attacks to recover the original input data [10–14].

To address these concerns, we introduce a framework designed to assess the possibility of retrieving classical inputs from the gradients observed in VQC models. We consider VQCs that satisfy the Lie algebra supported ansatz (LASA) property [15], which has been key in establishing connections with the trainability and classical simulatability of VQCs [15–17]. Our study systematically differentiates the separate prerequisites for input reconstruction across both the variational ansatz and encoding map architectures of these VQC models.

We consider input data $\mathbf{x}$ which is encoded into a quantum state $\rho(\mathbf{x})$ via some encoding circuit $V(\mathbf{x})$. This is then evolved by some variational circuit $W(\theta)$ before measurement of an observable O. The loss function is then defined as $y_\theta = \mathrm{Tr}(W(\theta)\rho(\mathbf{x})W(\theta)^\dagger O)$. We consider the case where we know the circuit architecture, the variational parameters θ and the gradients of the model $\frac{\partial y_\theta}{\partial \theta_j}, \forall j \in \dim(\theta)$. The goal of the attacker is to use this information to retrieve $\mathbf{x}$.

Our first result establishes a connection between the dynamical Lie algebra (DLA) $\mathfrak{g}$ of circuit generators and the privacy of the model. We achieve this by proving the following theorem.

* jamie.heredge@student.unimelb.edu.au

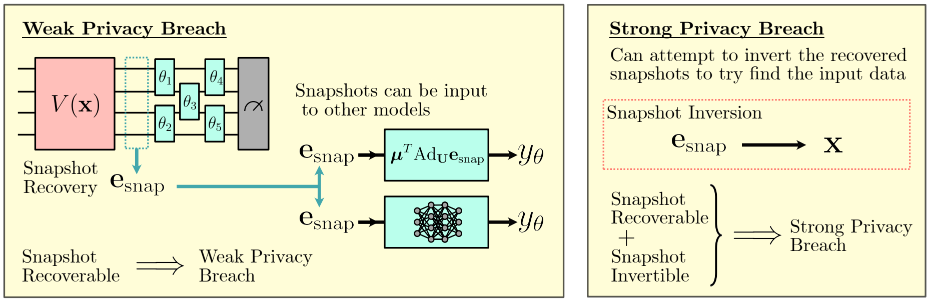

FIG. 1: Overview of the general framework and definitions. Weak privacy breach corresponds to attacks where snapshots of the data are retrieved. These can be used as inputs to other models, without explicitly needing the exact data, allowing one to potentially learn characteristics of the data. If these snapshots can then be further inverted to retrieve the input data $\mathbf{x}$ explicitly, we say the attack has succeeded in a strong privacy breach.

Theorem 1 (Snapshot Recovery)**.** *Given the requirements that the VQC satisfies the LASA condition and the DLA basis can be found in polynomial time, along with the assumption that the number of variational parameters $D \geq dim(\mathfrak{g})$, where $dim(\mathfrak{g})$ is the dimension of the DLA $\mathfrak{g}$, the VQC model admits snapshot $\mathbf{e}_{snap} = Tr(\mathbf{B_k}\rho(\mathbf{x})), \mathbf{B_k} \in \mathfrak{g}$ recovery with complexity scaling as $\mathcal{O}(poly(\dim(\mathfrak{g})))$.*

We prove the theorem by working in the Lie algebra framework by effectively implementing the reverse operation of Lie-algebraic simulation [17] to construct a system of simultaneous equations that can be solved in time $\mathcal{O}(\text{poly}(\dim(\mathfrak{g})))$. The implication of this is that when DLA scales polynomially with the number of qubits, it is possible to efficiently extract meaningful *snapshots* of the input, defined as $\mathbf{e}_{\text{snap}} = \text{Tr}(\mathbf{B_k}\rho(\mathbf{x})), \mathbf{B_k} \in \mathfrak{g}$, enabling training and evaluation of VQC models for other learning tasks without having direct access to the original input. For example, if an image had been encoded in a quantum state, and the snapshot is retrieved, it could be possible to pass this snapshot through different classification algorithms to learn properties of the image (e.g. if the picture is of a dog or a cat), even if it is not possible to fully recreate and view the original image. We term this a *weak privacy* breach of the model.

We can therefore conclude that a possible condition for protection against classical input recovery using gradients in a VQC model is to choose an ansatz that exhibits an exponentially large dynamical Lie algebra dimension, as this would render snapshot recovery difficult. Reviewing previous work within our new framework we see that previous privacy-demonstrating VQC models [18] effectively relied upon this exponentially large DLA to achieve privacy. However, we note that this may cause trainability issues for the models as recent works suggest that exponentially-sized DLA, may lead to the presence of barren plateaus drastically deteriorating the trainability of such models [15, 19]. The privacy of the model is linked to the DLA dimension and furthermore that there is a direct tradeoff between privacy and trainability of the model. As exponentially sized DLA models are expected to be untrainable in the LASA case, this means that for realistic applications, it does not seem feasible to rely on quantum privacy derived from an exponential DLA precluding snapshot recoverability in the model. This suggests that any privacy enhancement from quantum VQCs should not derive itself from the variational part of the circuit and focus should be directly towards the encoding portion.

We then investigate conditions for a *strong privacy* breach, i.e., recoverability of the original input $\mathbf{x}$ from these snapshots $\mathbf{e}_{\text{snap}}$, which is a process we term snapshot inversion. Fully reconstructing the input data from these snapshots to perform a strong privacy breach presents a further challenge which we show is dependent on properties of the encoding map, such as the hardness of classically simulating the encoding, overlap of the DLA basis with encoding circuit generators, and its Fourier frequency characteristics. The two types of privacy breach we introduce are summarised in Figure 1.

We consider various general architectures and show if they are susceptible to strong privacy breaches through snapshot inversion. The first encoding is the commonly used Pauli product encoding defined as $\bigotimes_{j_1}^{n} \rho_j(x_j) = \bigotimes_{j_1}^{n} R_{\mathsf{X}}(x_j)|0\rangle\langle 0|R_{\mathsf{X}}(-x_j)$. We show that in this case the input can be found from the snapshots in a closed form solution for each x_j input value given by

$$x_j = \cos^{-1}\left[\frac{2}{\text{sign}(\alpha)\sqrt{\alpha^2+\beta^2}} \sum_{k=1}^{\dim(\mathfrak{g})} \gamma_k [\mathbf{e}_{\text{snap}}]_k\right] - \tan^{-1}(\beta/\alpha), \tag{1}$$

where $\mathbf{B}_k \in \mathfrak{g}$ are the DLA basis elements and we define $\sum_{k=1}^{\dim(\mathfrak{g})} \gamma_i \mathbf{B}_k = \alpha i \mathsf{Z}_j + \beta i \mathsf{Y}_j \in \mathfrak{g}$. Hence models of this type will be vulnerable to strong privacy breaches. This is a significant result as this encoding map is commonly used, having been implemented in previous quantum federated learning models [18]. The only way in which models can remain private with this encoding, is if they are immune to snapshot recovery via an exponential DLA as in [18], although as mentioned this could significantly hinder the trainability.

We then turn our attention to more general Pauli encodings where $\rho(\mathbf{x}) = \bigotimes_{J\in\mathcal{P}} \rho_J(\mathbf{x})$ can be written as a separable state and prove the theorem

Theorem 2. *Suppose that the feature encoding state $\rho(\mathbf{x})$ is a multipartite state, specifically there exists a partition $\mathcal{P}$ of qubits $[n]$ such that $\rho(\mathbf{x}) = \bigotimes_{J\in\mathcal{P}} \rho_J(\mathbf{x})$, where we define $\mathrm{x}_J \subseteq \mathbf{x}$ to be components of $\mathbf{x}$ on which ρ_J depends. In addition, we have as input an $\mathcal{O}(\mathrm{poly}(n))$-dimensional snapshot vector $\mathbf{e}_{snap}$ with respect to a known basis $\mathbf{B}_k$ for the DLA of the VQC. Suppose that for $\rho_J(\mathbf{x})$ the following conditions are satisfied: **1):** $\dim(\mathbf{x}_J) = \mathcal{O}(1)$. **2):** Each x_k is encoded at most $R = \mathcal{O}(\mathrm{poly}(n))$ times in, potentially multi-qubit, Pauli rotations. **3):** The set $\mathcal{S}_J = \{k : Tr(\mathbf{B}_k\rho_J(\mathbf{x})) \neq 0 \ \&\ Tr(\mathbf{B}_k\rho_{J^c}(\mathbf{x})) = 0\}$ has cardinality at least $\dim(\mathrm{x}_J)$, where $J^c := [n] - J$. Then the model admits quantum-assisted snapshot inversion for recovering x_J. Furthermore, a classical snapshot inversion can be performed if $\forall k, \mathrm{Tr}(\mathbf{B}_k\rho_J(\mathbf{x}))$ can be evaluated classically for all $\mathbf{x}$. Overall, ignoring error in estimating $\mathrm{Tr}(\mathbf{B}_k\rho_J(\mathbf{x}))$, with the chosen parameters, this leads to a $\mathcal{O}(\mathrm{poly}(n, \log(1/\epsilon)))$ algorithm.*

This is proven by considering the Fourier decomposition of the quantum encoded state which can lead to the construction of a system of trigonometric simultaneous equations. Converting these to Chebyshev polynomials and considering bounds from computational geometry we show that Buchberger's algorithm [20] can be guaranteed to solve the equations under these circumstances in polynomial time. This demonstrate a wider class of encoding circuits that are guaranteed to be vulnerable to strong privacy breaches.

For the most general encoding circuit, with no specified architecture, we treat snapshot inversion as a black box problem. In this setting guess inputs $\mathbf{x}'$ can be passed to a quantum circuit or classical simulator to find values $\mathbf{e}'_{\mathrm{snap}}$. A solution is found when the calculated $\mathbf{e}'_{\mathrm{snap}}$ is sufficiently close to the true snapshot values $\mathbf{e}_{\mathrm{snap}}$, meaning that $\mathbf{x}'$ will also be sufficiently close to $\mathbf{x}$. In general, it is not possible to guarantee snapshot inversion in polynomial time, as a grid search would scale $\mathcal{O}(\exp(n))$. However, we show that using perturbed gradient descent, the scaling can instead be formed with respect to the highest frequency Fourier term of the expectation values, assuming the optimisation does not fail due to bad local minima. This places the emphasis on encoding circuits with high frequency expectation values. We provide numerics exploring the frequency landscape for example circuits.

In the case that is not possible to classically simulate the expectation values of the quantum encoded state with respect to the DLA basis elements of the variational circuits, then this can guarantee privacy against strong breaches if the attacking agent is restricted to purely classical means. Providing a potentially useful framework for circuit construction against classical attackers. However, if the attacking agent has access to a quantum device that can take measurements of a circuit to calculate snapshots values.

In the case where the attacker can simulate expectation values of the DLA basis or has access to a quantum device to obtain the expectation values, then numerical classical snapshot inversion or numerical quantum-assisted snapshot inversion can be attempted, respectively. We have shown that in this case an important factor in preventing these techniques is that the expectation values have exponentially scaling frequency terms resulting in the attacks requiring to solve a system of high-degree polynomial equations. The implication of this is that to achieve useful privacy advantage in VQC it may require that the encoding circuit is constructed in such a way that the expectation values of the DLA basis elements of the variational circuit contain frequency terms that scale exponentially. Notably, we find that having high frequency terms in the gradients, as suggested in the encoding circuit of [18], does not necessarily protect against numerical snapshot inversion attacks. This is because the gradients inherit the highest frequency term from all expectation values, but there may be a sufficient number of polynomial frequency expectation values to perform snapshot inversion, even if direct input inversion is not possible. This introduces stricter requirements to prevent gradient inversion attacks then previously known.

It remains as a question for future research the types of data which may be trained appropriately using the privacy-preserving high frequency feature maps proposed. More generally, the prospect for quantum privacy rests on feature maps that are *untrainable* with regard to adjusting $\mathbf{x}'$ to recover expectation values $\mathbf{e}_{\mathrm{snap}}$, while at the same time remaining useful feature maps with respect to the underlying dataset and overall model.

In conclusion, we offer a methodological approach for classifying and analyzing the privacy features of VQC models, presenting conditions for weak and strong privacy breaches for a broad spectrum of possible VQC architectures. Our findings not only enhance the understanding of quantum privacy mechanisms but also offer strategic guidelines for the design of quantum circuits that prioritize security while at the same time maintaining trainability. We provide strict guidance on a large class of models that should be avoided, which significantly narrows the search for privacy advantage in VQC models and is therefore of significant importance to the field. Looking ahead, this research paves the way for more robust quantum machine learning model designs, where privacy and functionality are balanced. This knowledge offers the potential to deliver effective machine learning models that simultaneously demonstrate a privacy advantage over conventional classical methods.

DISCLAIMER

This work was prepared for informational purposes by the Global Technology Applied Research center of JPMorgan Chase & Co. This paper is not a product of the Research Department of JPMorgan Chase & Co. or its affiliates. Neither JPMorgan Chase & Co. nor any of its affiliates makes any explicit or implied representation or warranty and none of them accept any liability in connection with this paper, including, without limitation, with respect to the completeness, accuracy, or reliability of the information contained herein and the potential legal, compliance, tax, or accounting effects thereof. This document is not intended as investment research or investment advice, or as a recommendation, offer, or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction.

[1] J. Heredge, N. Kumar, D. Herman, S. Chakrabarti, R. Yalovetzky, S. H. Sureshbabu, C. Li, and M. Pistoia, "Prospects of privacy advantage in quantum machine learning," 2024. https://doi.org/10.48550/arXiv.2405.08801.

[2] T. Liu, Z. Wang, H. He, W. Shi, L. Lin, W. Shi, R. An, and C. Li, "Efficient and secure federated learning for financial applications," 2023.

[3] T. Awosika, R. M. Shukla, and B. Pranggono, "Transparency and privacy: The role of explainable ai and federated learning in financial fraud detection," 2023.

[4] G. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, pp. 305 – 311, 2020.

[5] S. Ahamed, N. Nishant, A. Selvaraj, N. Gandhewar, S. A, and B. K, "Investigating privacy-preserving machine learning for healthcare data sharing through federated learning," *The Scientific Temper*, vol. 14, pp. 1308–1315, 12 2023.

[6] E. Aguiar, C. Jr, and A. Traina, "Security and privacy in machine learning for health systems: Strategies and challenges," *Yearbook of Medical Informatics*, vol. 32, pp. 269–281, 12 2023.

[7] C. Park, T. Choi, T. Kim, M. Cho, J. Hong, M. Choi, and J. Choo, "Fedgeo: Privacy-preserving user next location prediction with federated learning," 2023.

[8] J. P. Albrecht, "How the gdpr will change the world," *European Data Protection Law Review*, vol. 2, pp. 287–289, 2016.

[9] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," 2023.

[10] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

[11] Y. Huang, S. Gupta, Z. Song, K. Li, and S. Arora, "Evaluating gradient inversion attacks and defenses in federated learning," in *Advances in Neural Information Processing Systems* (A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), 2021.

[12] B. Zhao, K. R. Mopuri, and H. Bilen, "idlg: Improved deep leakage from gradients," *arXiv preprint arXiv:2001.02610*, 2020.

[13] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting Gradients - How easy is it to break privacy in federated learning?," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[14] H. Yin, A. Mallya, A. Vahdat, J. Alvarez, J. Kautz, and P. Molchanov, "See through Gradients: Image Batch Recovery via GradInversion," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16332–16341, 2021.

[15] E. Fontana, D. Herman, S. Chakrabarti, N. Kumar, R. Yalovetzky, J. Heredge, S. H. Sureshbabu, and M. Pistoia, "The adjoint is all you need: Characterizing barren plateaus in quantum ans\" atze," *arXiv preprint arXiv:2309.07902*, 2023.

[16] R. Somma, G. Ortiz, H. Barnum, E. Knill, and L. Viola, "Nature and measure of entanglement in quantum phase transitions," *Physical Review A*, vol. 70, no. 4, p. 042311, 2004.

[17] M. L. Goh, M. Larocca, L. Cincio, M. Cerezo, and F. Sauvage, "Lie-algebraic classical simulations for variational quantum computing," *arXiv preprint arXiv:2308.01432*, 2023.

[18] N. Kumar, J. Heredge, C. Li, S. Eloul, S. H. Sureshbabu, and M. Pistoia, "Expressive variational quantum circuits provide inherent privacy in federated learning," *arXiv preprint arXiv:2309.13002*, 2023.

[19] M. Ragone, B. N. Bakalov, F. Sauvage, A. F. Kemper, C. O. Marrero, M. Larocca, and M. Cerezo, "A unified theory of barren plateaus for deep parametrized quantum circuits," *arXiv preprint arXiv:2309.09342*, 2023.

[20] B. Buchberger, *Gröbner Bases: An Algorithmic Method in Polynomial Ideal Theory*, pp. 184–232. Reidel Publishing Company, 01 1985.