

# Private Training in Quantum Machine Learning

Extended Abstract

Tigran Sedrakyan<sup>1</sup>   Frédéric Grosshans<sup>1</sup>   Elham Kashefi<sup>1,2</sup>

<sup>1</sup>Sorbonne Université, CNRS, LIP6, Paris, France

<sup>2</sup>School of Informatics, University of Edinburgh, United Kingdom

## Abstract

As machine learning models are increasingly trained on sensitive data, a central question in AI safety is how to make learning private with respect to the training set. Differential privacy (DP) is the standard tool for this, but its interplay with quantum machine learning (QML) is poorly understood. We study private training of hybrid variational QML models with classical inputs and outputs using a classical DP-SGD optimizer. We (i) explain why native quantum noise is not a satisfactory replacement for the calibrated noise of DP-SGD, (ii) prove a deterministic bound on the DP-SGD clipping bias for a broad class of quantum models, arising from unitary evolution and bounded observables, and (iii) formulate a matched-model comparison protocol at fixed clipping threshold and privacy budget. On synthetic and image-classification tasks, quantum models retain higher accuracy than equivalent classical baselines in private-training regimes where the formal guarantee is supplied by classical DP-SGD.

## 1 Background

**Differential privacy and DP-SGD.** Differential privacy [1] bounds how much a single record can change a randomized algorithm’s output distribution. A mechanism  $\mathcal{M}$  is  $(\epsilon, \delta)$ -DP if, for all datasets  $D \sim D'$  differing in one record and all measurable  $S$ ,

$$\Pr[\mathcal{M}(D) \in S] \leq e^\epsilon \Pr[\mathcal{M}(D') \in S] + \delta. \quad (1)$$

$(\epsilon, \delta)$  pair defines the privacy *budget*. The dominant way to train deep models privately is DP-SGD [2], which computes per-example gradients  $g_i(\theta) = \partial_\theta \ell(f(x_i, \theta), y_i)$ , clips each to an  $\ell_2$  norm threshold  $C$ ,

$$\Pi_C(g_i) = g_i \min\{1, C/\|g_i\|_2\}, \quad (2)$$

and adds Gaussian noise to the clipped batch average, with privacy composed across training iterations via Rényi accounting [3]. Over a minibatch  $B$  this yields the privatized gradient

$$\widehat{g}_B(\theta) = \frac{1}{|B|} \sum_{i \in B} \Pi_C(g_i(\theta)) + Z, \quad Z \sim \mathcal{N}\left(0, \frac{\sigma^2 C^2}{|B|^2} \mathbf{1}_p\right), \quad (3)$$

on which the parameter update is performed. Clipping bounds sensitivity but also distorts the optimization: it biases the expected update by  $b_C(\theta) = \mathbb{E}[\Pi_C(g_i)] - g(\theta)$ , a non-vanishing source of error that does not shrink with batch size. The quality of this private update is captured by its mean-squared error relative to the true population gradient  $g$ ,

$$\begin{aligned} \text{MSE}(\theta) &= \mathbb{E} [\|\widehat{g}_B(\theta) - g(\theta)\|_2^2] \\ &= \underbrace{\frac{1}{|B|} \text{Tr Cov}[\bar{g}_i]}_{\text{sampling}} + \underbrace{p(\sigma C/|B|)^2}_{\text{DP noise}} + \underbrace{\|b_C(\theta)\|_2^2}_{\text{clipping bias}}, \end{aligned} \quad (4)$$

so smaller  $C$  is desirable (less noise) only if the bias term can be kept under control. Figure 1 shows where privatization can enter a hybrid quantum/classical training pipeline.

**Hybrid quantum models.** We use the standard hybrid variational classifier, in which classical inputs  $x$  are encoded in a state  $\rho(x)$ , a trainable unitary  $U(\theta) = \prod_{j=1}^p e^{-i\theta_j H_j}$  is applied, and class logits are read out as expectation values of bounded observables, optionally post-processed classically. The one structural fact our results exploit is that the trainable part evolves unitarily and the readout is bounded — unlike a classical network, whose per-example gradients are unbounded a priori and so offer no native control over the clipping sensitivity above.

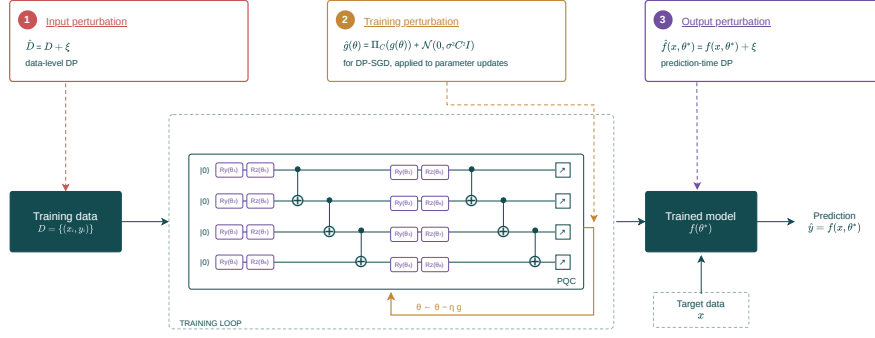


**Figure 1:** Where differential privacy can be injected in a (quantum) machine-learning pipeline: at the input (1), during training (2), or at the output (3). We study *training-time* privacy of classical data via DP-SGD applied to a parameterized quantum circuit (PQC).

## 2 Main results

We work in a  $K$ -class setting with logits  $f_y(x, \theta) = \text{Tr}(O_y U(\theta) \rho(x) U(\theta)^\dagger)$ , temperature- $\tau$  cross-entropy loss  $\ell = -\log p_y$ , with partial derivatives given by  $\partial_{\theta_j} \ell = \frac{1}{\tau} (\sum_k p_k \partial_{\theta_j} f_k - \partial_{\theta_j} f_y)$ , which transfers gradient control from the logits to the loss gradients used in DP-SGD.

**(1) Quantum noise is not a DP substitute.** A natural hope is that intrinsic quantum randomness (finite-shot and hardware noise) could supply the privacy noise for free.

**Theorem 1** (Noise propagation, informal). *For traceless-Pauli logits under global depolarizing noise of strength  $\lambda$  and  $S$  shots, the induced perturbation of each loss-gradient coordinate is approximately additive Gaussian,  $\widehat{\partial_\theta \ell} \approx \partial_\theta \ell + \mathcal{N}(m, v)$ , with  $m, v$  computable from the logits, their gradients,  $\lambda$ , and  $S$ .*

While this resembles the DP-SGD perturbation (3), it is not a valid substitute. Finite-shot readout of a Pauli expectation  $f$  has variance  $(1 - f^2)/S$ , which collapses as  $f \rightarrow \pm 1$ , and depolarizing noise of strength  $\lambda$  rescales the gradient by  $(1 - \lambda)$  and adds a data-dependent bias  $-\lambda \partial_\theta \ell$ . Both the mean  $m$  and variance  $v$  therefore depend on the example, whereas DP demands a perturbation of fixed, example-independent sensitivity, and the worst-case bound here is too loose to give a competitive privacy budget. Shot noise alone gives  $m = 0$  and symmetrizes the pre-clipping gradient distribution and can marginally improve utility [4], but the formal guarantee must still come from classical calibrated noise in DP-SGD.

**(2) A deterministic clipping-bias bound.** Since quantum models admit a data-independent gradient bound, the harmful clipping bias in (4) can be controlled deterministically in DP-SGD.

**Theorem 2** (Deterministic clipping-bias bound). *For the hybrid classifier with cross-entropy loss at temperature  $\tau$ , set*

$$G := 2 \max_{1 \leq k \leq K} \|O_k\| \left( \sum_{j=1}^p \|H_j\|^2 \right)^{1/2}.$$

*Then every per-example loss gradient obeys  $\|g_i(\theta)\|_2 \leq 2G/\tau$ , and for every clipping threshold  $C > 0$  the clipping bias satisfies*

$$\|b_C(\theta)\|_2 \leq (2G/\tau - C)_+. \quad (5)$$

The bound is deterministic, computable from the circuit description, and uniform over the encoding strategy. In particular, choosing  $C \geq 2G/\tau$  makes the bias vanish; more usefully, a small  $C$  can be used without incurring large bias, which lowers the DP noise scale  $\sigma C$  in (4) and improves the privacy-utility tradeoff. For local, low-weight observables and generators the constant  $G$  stays tight; this regime also coincides with the trainable (barren-plateau-free) and classically simulable regime. This is akin to the classical Lipschitz networks [5], except that here gradient control is inherited quantum phenomena.

## 3 Numerical experiments

To compare fairly, quantum and classical models are *equivalent*: matched parameter count  $p$  and matched accuracy under non-private training, so any gap under DP-SGD reflects privacy robustness rather than capacity. We benchmark against standard tanh networks [6] and Lipschitz networks [5].

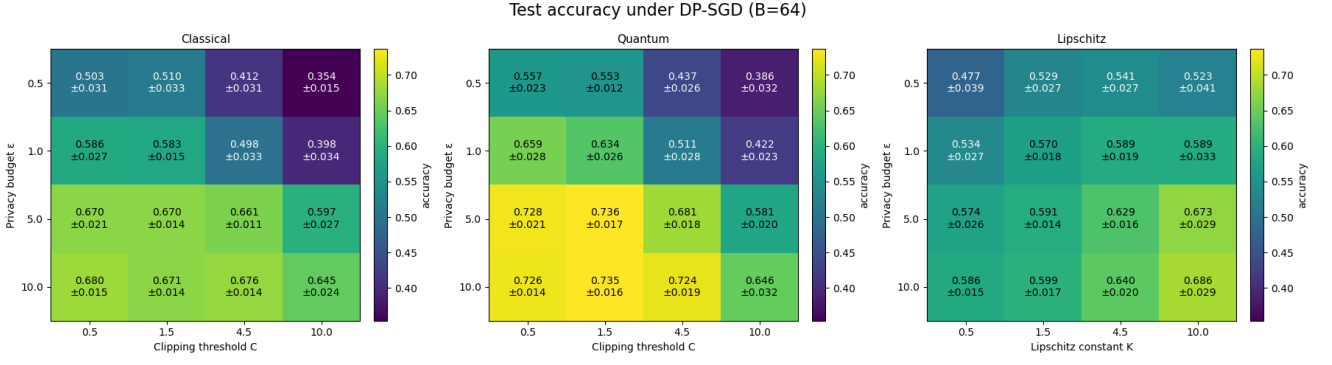


**Figure 2:** Final test accuracy vs. clipping threshold  $C$  and privacy budget  $\epsilon$  ( $|B| = 64$ ,  $\delta = 10^{-5}$ ) for matched classical, quantum, and Lipschitz models. The quantum model achieves the best accuracy at the small threshold  $C = 1.5$ , where the calibrated DP noise  $\sigma C$  is minimized while its bounded gradients keep the clipping bias small.

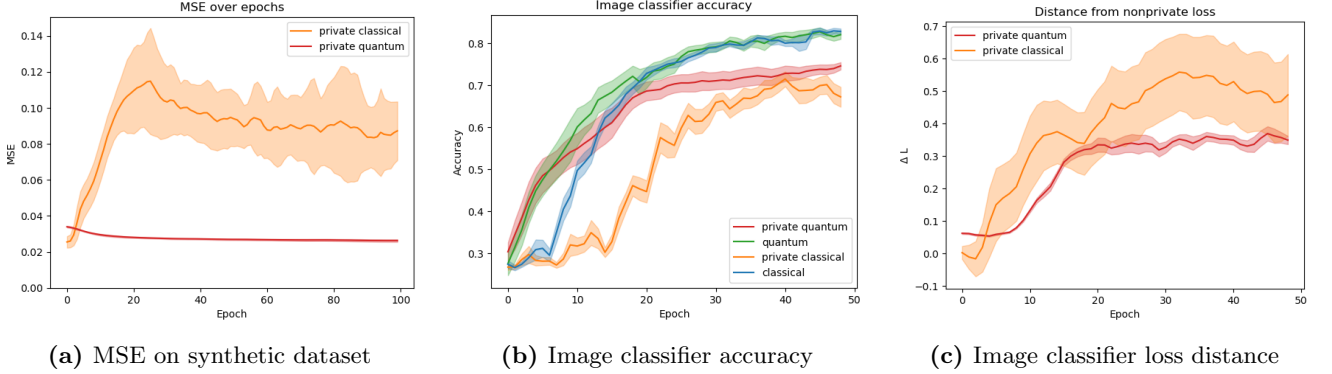


**Figure 3:** Private training under DP-SGD ( $\epsilon = 10$ ,  $\delta = 10^{-5}$ ,  $C = 1.5$ ; mean  $\pm$  std, 10 seeds). The quantum model keeps a lower gradient MSE (4) on a synthetic task (a) and on the image classification task reaches higher accuracy (b) and stays closer to the non-private loss (c) than the matched CNN.

**Clipping effect.** On a synthetic 16-feature, 4-class task (no DP noise, so only clipping acts), all models match at large  $C$  but classical models degrade sharply for  $C \leq 1.5$ . Tracking the clipping probability  $\Pr[\|g_i\|_2 > C]$  and the directional alignment  $A_C = \langle \bar{g}_B, g \rangle / (\|\bar{g}_B\|_2 \|g\|_2)$ , we see classical gradients routinely exceed  $C$ , so  $A_C$  falls below 1 and the population gradient plateaus short of a stationary point, whereas the quantum model respects the ceiling  $2G/\tau$ , keeps  $A_C \approx 1$ , and is unaffected — a direct empirical signature of Theorem 2.

**Private training.** Under actual DP-SGD (Fig. 2), the same structure pays off: the quantum model attains the best accuracy at a small threshold  $C = 1.5$ , outperforming the best classical equivalent in accuracy by  $\sim 5\%$  at fixed budget, because small  $C$  minimizes the calibrated noise  $\sigma C$  while the bias stays bounded. At  $C = 1.5$ ,  $\epsilon = 10$  the quantum update has lower mean-squared error (4) (Fig. 3a) — both its clipping-bias and  $\sigma C$  noise terms are smaller — which results in a higher accuracy. The advantage persists on a real image classification task (Fig. 3b,c): a quantum image-loading-and-classification pipeline [7] on the Honda Scenes dataset [8], with  $\sim 500$  parameters matched to a CNN.

## 4 Conclusion

The picture is twofold. Formal record-level privacy should still come from classical DP-SGD — native quantum noise results in a privacy budget too loose to serve as a mechanism on its own — but the unitary-plus-bounded-readout structure of quantum variational models deterministically bounds the clipping bias, letting them tolerate smaller thresholds and hence less injected noise at fixed budget. This yields a measurable privacy-utility advantage over matched classical and Lipschitz baselines, and suggests that norm-preserving, bounded-output layers could carry part of the advantage to classical architectures. Full manuscript is available [9].

## References

- [1] Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9**(3–4), 211–407 (2014) <https://doi.org/10.1561/04000000042>
- [2] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318 (2016). <https://doi.org/10.1145/2976749.2978318>
- [3] Mironov, I.: Rényi differential privacy. 2017 IEEE 30th Computer Security Foundations Symposium (CSF), 263–275 (2017) <https://doi.org/10.48550/arXiv.1702.07476> [arXiv:1702.07476](https://arxiv.org/abs/1702.07476)
- [4] Chen, X., Wu, Z.S., Hong, M.: Understanding gradient clipping in private sgd: a geometric perspective. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20*. Curran Associates Inc., Red Hook, NY, USA (2020)
- [5] Béthune, L., Massena, T., Boissin, T., Bellet, A., Mamalet, F., Prudent, Y., Friedrich, C., Serrurier, M., Vigouroux, D.: DP-SGD without clipping: The lipschitz neural network way. In: *The Twelfth International Conference on Learning Representations* (2024). <https://openreview.net/forum?id=BEyEziZ4R6>
- [6] Papernot, N., Thakurta, A., Song, S., Chien, S., Erlingsson, U.: Tempered sigmoid activations for deep learning with differential privacy. *Proceedings of the AAAI Conference on Artificial Intelligence* **35**(10), 9312–9321 (2021) <https://doi.org/10.1609/aaai.v35i10.17123>
- [7] Gharibyan, H., Karapetyan, H., Sedrakyan, T., Subasic, P., Su, V.P., Tanin, R.H., Tepanyan, H.: Quantum image loading and classification: experiments on utility-scale quantum computers. *Quantum Machine Intelligence* **8**(1), 57 (2026) <https://doi.org/10.1007/s42484-026-00388-3>
- [8] Narayanan, A., Dwivedi, I., Dariush, B.: Dynamic Traffic Scene Classification with Space-Time Coherence (2019). <https://arxiv.org/abs/1905.12708>
- [9] Sedrakyan, T., Grosshans, F., Kashefi, E.: Private training in quantum machine learning (2026). <https://arxiv.org/abs/2606.29293>