How Noisy Training Reduces Model Discrepancy in Quantum Machine Unlearning

Anonymous authors
Paper under double-blind review

Abstract

Machine unlearning aims to remove the influence of selected training data without retraining from scratch. While calibrated training noise can support unlearning guarantees in classical learning, native noise in quantum neural networks (QNNs) is not designed for this purpose. We study how native quantum noise affects the *pre-to-post discrepancy* between the original (pre-unlearn) model distribution and its retrained gold-standard reference under the same noise regime, before any unlearning update is applied. We develop a trajectory-level 2-Wasserstein analysis that decomposes this discrepancy into deletion-induced update differences, their subsequent propagation, and finite-shot gradient fluctuations. For exact-expectation readout noise, this analysis yields explicit channel-dependent bounds and identifies conditions under which their geometric envelopes become smaller than the noiseless counterpart. We further construct a fixed-QNN example showing that arbitrarily small Wasserstein discrepancy can coexist with maximal deletion sensitivity. Experiments on a 4-qubit QNN show discrepancy reductions in selected readout and interleaved channel regimes, while finite-shot measurement provides no comparable reduction in the tested settings. A 10-qubit MNIST study additionally shows that selected noise regimes reduce unlearning optimization costs. Together, these findings suggest that native quantum noise can facilitate unlearning in selected regimes, while distinguishing geometric proximity and optimization savings from certified removal of data influence.

1 Introduction

Growing interest in the 'right to be forgotten' for personal data in AI models has heightened the need to remove the influence of certain data from learned models, with applications such as mitigating copyright infringement and eliminating harmful content. However, once data are incorporated into training, they may leave sample-specific information in model parameters, such that the naive deletion of the original data does not necessarily eliminate their influence on the learned model. Retraining the model after deleting the target data provides a straightforward solution, widely accepted as a gold standard of data removal (Guo et al., 2020), but it can require substantial computational resources. Machine unlearning (Cao & Yang, 2015) has therefore emerged as an effective approach to remove the influence of target training data without full retraining. Although machine unlearning has been extensively studied in classical learning systems, its behavior in quantum learning models remains considerably less understood.

A prominent aspect of machine unlearning in quantum neural networks (QNNs) stems from inevitable quantum noise during training. As a representative architecture, a parameterized quantum circuit encodes inputs into quantum states, processes them through quantum operations, and yields predictions via quantum measurements (Cerezo et al., 2021). The practical implementation of QNNs is subject to quantum hardware-induced channel noise—such as depolarization, amplitude damping, phase damping, and bit flip (Wang et al., 2021)—as well as finite-shot measurement fluctuations arising from the probabilistic nature of quantum systems.

On the classical machine learning side, noise intentionally introduced during training can regulate model sensitivity to individual training samples (Mou et al., 2018), thereby mitigating memorization (Carlini et al., 2019), and appropriately calibrated noisy training can provide sample-level privacy guarantees (Abadi et al., 2016). In particular, Sekhari et al. (2021) showed that differentially private

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.

training with privacy parameters calibrated to the deletion budget satisfies approximate unlearning guarantees without any post-training model updates. Its privacy guarantee bounds the statistical discrepancy between the model trained on the full dataset and its counterpart trained on the retained data with the same private algorithm.

These observations motivate us to ask how native quantum noise changes the gap to the retrained model distribution. Native quantum noise, however, is not intentionally calibrated to satisfy such privacy requirements. We therefore examine how it shapes this discrepancy, without presupposing that it yields certified unlearning. Our key question is: *How does native quantum noise in QNNs affect the discrepancy between the original and retrained model distributions?*

To answer this question, we provide both a theoretical bound on the discrepancy between original and retrained QNN distributions under noisy training, and empirical investigations:

- **Geometric analysis.** We derive a trajectory-level 2-Wasserstein bound that separates deletion-induced discrepancy, its propagation through retraining updates, and finite-shot gradient errors.
- **Channel-specific Bounds.** We specialize the pre-to-post discrepancy analysis to Pauli-Z readout channels, obtaining closed-form bounds on their deletion-induced update perturbations.
- **Empirical investigation.** We empirically examine noise effects on pre-to-post discrepancy and unlearning cost, identifying channel regimes that reduce both, with cost gains depending on the unlearning method. Finite-shot measurement shows no comparable discrepancy reduction.

2 Related Works & Preliminaries

Machine Unlearning. Machine unlearning aims to remove the influence of selected training data without full retraining (Cao & Yang, 2015; Ginart et al., 2019; Bourtoule et al., 2021). Retraining on the retained data serves as a standard target of unlearning: *exact unlearning* produces a model distribution identical to this retrained reference, whereas certified *approximate unlearning* bounds their statistical distinguishability (Guo et al., 2020; Neel et al., 2021; Koloskova et al., 2025). Calibrated training noise can provide differential privacy (DP) (Abadi et al., 2016); when its privacy parameters are adapted to the deletion budget, it can simultaneously guarantee approximate unlearning without post-processing (Sekhari et al., 2021). Noisy stochastic gradient descent (SGD) also supports exact unlearning through total-variation stability (Ullah et al., 2021), while Langevin unlearning links stronger initial privacy to fewer unlearning updates (Chien et al., 2024).

Quantum Neural Networks & Quantum Noise. QNNs typically encode inputs into quantum states, process them through parameterized quantum circuits (PQCs), and obtain predictions from measurements (Cerezo et al., 2021). In practical QNN implementations, noise enters during both circuit evolution and measurement: hardware noise perturbs quantum states through quantum channels, whereas finite-shot stochasticity arises from estimating expectation values with a finite number of measurements (Wang et al., 2021; Sweke et al., 2020). Noise channels considered in this work include *depolarization*, which drives states toward the maximally mixed state; *amplitude damping*, which models relaxation from $|1\rangle$ to $|0\rangle$; *phase damping*, which suppresses coherence while preserving basis state populations; and *bit flip*, which randomly exchanges $|0\rangle$ and $|1\rangle$. Beyond channel type, channel noise can also be characterized by its placement within the circuit. As shown in Figure 1, we consider two channel placements: readout, where the channel is applied immediately before the measurement, and interleaved, where the channels are repeatedly inserted throughout the circuit evaluation. Due to space constraints, we defer the background on QNNs and the mathematical formulations of their noise channels to Appendix A. Quantum noise can suppress gradients (Wang et al., 2021) and improve generalization (Kuzmin et al., 2025), but these effects do not necessarily imply weaker memorization. Related stability analyses show that depolarizing noise can improve noise-dependent generalization bounds under specific assumptions (Yang et al., 2025). Separately, Gil-Fuster et al. (2024) show that QNNs can fit randomly labeled quantum data, demonstrating their capacity for memorization. However, how native quantum noise affects the discrepancy between original and retrained QNNs remains insufficiently characterized.

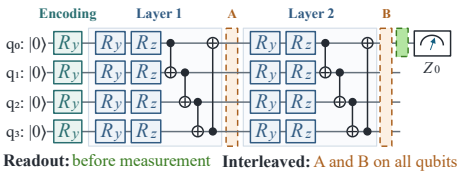

Figure 1: Example of 4-qubit PQC and our noise placement setting.

Quantum Machine Unlearning. Research on machine unlearning in QNNs has only recently emerged. QNNs have been reported to be more amenable to approximate unlearning than classical models under corrupted training data (Chen & Zhang, 2026). Concurrently, other studies have evaluated quantum unlearning in simulations and on hardware (Su et al., 2026), adapted classical methods to hybrid models (Crivoi & Ionescu, 2026), and investigated targeted noise injection (Masud et al., 2026) or poisoned-sample removal (Gadani et al., 2026). These studies highlight the need to expand our understanding of noisy QNN unlearning from deliberately introduced noise to native quantum noise. We investigate the latter through the geometric discrepancy between trained and retrained model distributions under identical noise regimes, before any unlearning update is applied.

3 Unlearning Discrepancy under Quantum Noise

We study how native quantum noise affects the discrepancy between the original trained-model distribution and its retrained reference under the same noise regime, before any unlearning update is applied. Section 3.1 defines the retraining reference and statistical deletion sensitivity. Sections 3.2-3.3 develop a complementary geometric analysis through 2-Wasserstein stability bounds, separating channel and finite-shot effects and deriving channel-specific bounds. Section 3.4 shows why geometric proximity alone does not imply certified unlearning.

3.1 Quantum Noisy Retraining and Approximate Unlearning

Fix a finite training dataset D, a deletion request set $F \subsetneq D$, and a retained set $R = D \setminus F$. A quantum noise regime $\nu = (\mathcal{N}, M)$ specifies noise channel configuration and shot budget, where $\mathcal{N}$ represents the quantum channel applied during circuit execution, including its channel type, placement and noise strength, while M denotes the number of measurement shots per circuit evaluation. The ideal noiseless regime is $\nu_0 = (\mathcal{I}, \infty)$, where $\mathcal{I}$ indicates the absence of channel noise and $M = \infty$ corresponds to the infinite number of measurements converging to exact expectation values. Let A_ν denote a learning mechanism under ν, with training randomness ξ. For any measurable set B in the model space $\mathcal{H}$, let us define distributions of the learned models as

$$P_D^\nu(B) \coloneqq \Pr[\mathsf{A}_\nu(D;\xi) \in B], \qquad P_R^\nu(B) \coloneqq \Pr[\mathsf{A}_\nu(R;\xi) \in B]. \tag{1}$$

P_D^ν is the noisy-trained original model distribution where unlearning applies, whereas P_R^ν is its noisy-retrained reference. For an unlearning mechanism U_ν, define an unlearned model distribution:

$$P_{D,F}^\nu(B) \coloneqq \Pr\left[\mathsf{U}_\nu\left(\mathsf{A}_\nu(D;\xi), F\right) \in B\right], \tag{2}$$

where the probability is taken over the training randomness and any internal randomness of the unlearning mechanism. The gold standard for $P_{D,F}^\nu(B)$ is $P_R^\nu(B)$: retraining holds the learning procedure fixed (Guo et al., 2020). Notably, we argue that the gold-standard for QNN unlearning must account for quantum noise. If a model were trained under ideal, noiseless conditions, the corresponding retrained standard becomes $P_R^{\nu_0}(B)$. Exact unlearning requires $P_{D,F}^\nu = P_R^\nu$, whereas approximate unlearning is quantified by the statistical distinguishability between $P_{D,F}^\nu$ and P_R^ν using hockey-stick divergence $H_\alpha(P\|Q) \coloneqq \sup_B\{P(B) - \alpha Q(B)\}$, where P, Q are model distributions, and the supremum is over measurable sets $B \subseteq \mathcal{H}$ (Balle et al., 2019; Koloskova et al., 2025).

Definition 1 (Quantum noisy approximate unlearning)**.** *Fix (D, F), a noise regime ν, and $\epsilon \geq 0$. For an unlearning mechanism U_ν, define its* deletion sensitivity *as*

$$\mathcal{S}_\epsilon^\nu(\mathsf{U}_\nu) \coloneqq \max\{\mathsf{H}_{e^\epsilon}\left(P_{D,F}^\nu \middle\| P_R^\nu\right), \mathsf{H}_{e^\epsilon}\left(P_R^\nu \middle\| P_{D,F}^\nu\right)\}. \tag{3}$$

For $0 \leq \delta \leq 1$, we say that U_ν satisfies (ϵ, δ)-approximate unlearning for (D, F) under ν if $\mathcal{S}_\epsilon^\nu(\mathsf{U}_\nu) \leq \delta$. Here, e^ϵ and δ specify the multiplicative factor and additive tolerance, respectively, in two-sided comparison.

To isolate the discrepancy present before any unlearning update, consider the identity mechanism $\mathsf{U}_{\mathrm{id}}(\theta, F) = \theta$, which leaves the trained model unchanged and thus gives $P_{D,F}^\nu = P_D^\nu$. Then, $S_\epsilon^\nu(\mathsf{U}_{\mathrm{id}})$ measures the statistical distinguishability already present between the trained original model (i.e., the pre-unlearn model) and the retrained distributions with the deletion of forget data (i.e., the post-unlearn target). At fixed ϵ, it is the smallest additive tolerance δ for which leaving the trained model unchanged satisfies approximate unlearning. We call the difference between P_D^ν and P_R^ν the *pre-to-post discrepancy*: "pre" denotes the model distribution obtained by training on the original dataset D, before applying unlearning, whereas "post" denotes the target distribution that unlearning aims to achieve, defined by retraining from scratch on the retained dataset R.

3.2 Geometric Model Discrepancy under Quantum Noise

To understand how quantum noise shapes the pre-to-post discrepancy, we couple full-data training and retraining from a shared initialization and track their root-mean-square parameter separation. This trajectory-level view is particularly useful because it allows the effects of channel noise and finite-shot measurement to be isolated through their contributions to the mean updates and stochastic gradient errors. This quantity naturally controls the 2-Wasserstein distance W_2 between the resulting model distributions, providing a geometric characterization of how training noise changes their separation. Importantly, such geometric control is distinct from the event-level distinguishability used for approximate unlearning: without additional distributional assumptions, a small Wasserstein distance need not imply a small hockey-stick divergence. We return to this distinction in Section 3.4.

We parameterize a model by $\theta \in \mathbb{R}^p$ and consider K full-batch gradient updates with shared deterministic step sizes $\eta_k \geq 0$, $k = 0, \ldots, K-1$. We use θ for a generic parameter value and $\Theta^{\nu}_{Z,k}$ for the random iterate produced by training on $Z \in \{D, R\}$ under regime ν, so that $\Theta^{\nu}_{Z,K} \sim P^{\nu}_{Z}$. For $Z \in \{D, R\}$, let $\mathcal{L}^{\mathcal{N}}_{Z}$ denote the sample-averaged objective based on exact channel-conditioned expectations, with a common differentiable regularizer. We define

$$T^{\mathcal{N}}_{Z,k}(\theta) := \theta - \eta_k \nabla \mathcal{L}^{\mathcal{N}}_{Z}(\theta), \qquad \Theta^{\nu}_{Z,k+1} = T^{\mathcal{N}}_{Z,k}(\Theta^{\nu}_{Z,k}) - \eta_k \Xi^{M}_{Z,k}. \tag{4}$$

Here, $T^{\mathcal{N}}_{Z,k}$ is the channel-conditioned mean update, whereas $\Xi^{M}_{Z,k}$ is the finite-shot loss-gradient estimation error. We assume that the finite-shot gradient estimator is unbiased for the channel-conditioned gradient at the current iterate, conditional on the coupled training history. Thus, the shot error is centered around the mean update, with $\Xi^{\infty}_{Z,k} = 0$. We focus on full-batch GD in the main analysis to isolate channel and finite-shot effects without introducing additional stochasticity from minibatch sampling. Appendix B.3 extends the argument to unbiased minibatch SGD.

To track how mean-update differences accumulate, consider a parameter region $\mathcal{P} \subseteq \mathbb{R}^p$. At each step k, assume finite nonnegative constants $\beta^{\mathcal{N}}_k$ and $c^{\mathcal{N}}_k$ satisfying

$$\beta^{\mathcal{N}}_k := \sup_{\theta \in \mathcal{P}} \left\| T^{\mathcal{N}}_{D,k}(\theta) - T^{\mathcal{N}}_{R,k}(\theta) \right\|_2, \qquad \left\| T^{\mathcal{N}}_{R,k}(\theta) - T^{\mathcal{N}}_{R,k}(\theta') \right\|_2 \leq c^{\mathcal{N}}_k \|\theta - \theta'\|_2, \tag{5}$$

for all $\theta, \theta' \in \mathcal{P}$. For any $\theta_D, \theta_R \in \mathcal{P}$, these quantities separate the one-step mean-update discrepancy into deletion-induced and propagated components:

$$\begin{aligned} \left\| T^{\mathcal{N}}_{D,k}(\theta_D) - T^{\mathcal{N}}_{R,k}(\theta_R) \right\|_2 &\leq \left\| T^{\mathcal{N}}_{D,k}(\theta_D) - T^{\mathcal{N}}_{R,k}(\theta_D) \right\|_2 + \left\| T^{\mathcal{N}}_{R,k}(\theta_D) - T^{\mathcal{N}}_{R,k}(\theta_R) \right\|_2 \\ &\leq \beta^{\mathcal{N}}_k + c^{\mathcal{N}}_k \|\theta_D - \theta_R\|_2. \end{aligned} \tag{6}$$

Thus, $\beta^{\mathcal{N}}_k$ controls the discrepancy newly introduced by deleting F at step k, whereas $c^{\mathcal{N}}_k$ controls how previously accumulated discrepancy is propagated by the retraining update. Our argument builds on coupled-iterate stability analyses of gradient methods (Hardt et al., 2016), but here we track the distributional discrepancy between full-data training and retraining under quantum noise. When $c^{\mathcal{N}}_k < 1$, existing discrepancies shrink under this update; $c^{\mathcal{N}}_k = 1$ is non-expansive, while $c^{\mathcal{N}}_k > 1$ permits amplification.

While parameter-space discrepancy captures geometric differences between the learned models, it does not directly characterize how these differences manifest in their predictions. To understand the discrepancy at the prediction level, we therefore transfer the comparison to the prediction space. To this end, fix an L_0-Lipschitz clean prediction map $\Psi_0 : \mathcal{P} \to \mathbb{R}^m$ that stacks exact, channel-free predictions on a fixed set of m inputs. We use the same map across noise regimes to provide a common functional comparison and denote the pushforward of a distribution P under Ψ_0 by $(\Psi_0)_P$.

Theorem 1 (Geometric stability and prediction transfer)**.** *Under the preceding setup, consider a coupling of the full-data training and retraining trajectories with a shared initialization. For $Z \in \{D, R\}$, assume $\Theta^{\nu}_{Z,k} \in \mathcal{P}$ almost surely, $\mathbb{E}\|\Theta^{\nu}_{Z,k}\|_2^2 < \infty$ for $k = 0, \ldots, K$, and $\mathbb{E}\|\Xi^{M}_{Z,k}\|_2^2 < \infty$ for $k = 0, \ldots, K-1$. For this coupling, define the root-mean-square trajectory separation and the shot-induced separation by*

$$u_k := \left(\mathbb{E} \left\| \Theta^{\nu}_{D,k} - \Theta^{\nu}_{R,k} \right\|_2^2 \right)^{1/2}, \qquad v^{\mathcal{N},M}_k := \eta_k^2 \mathbb{E} \left\| \Xi^{M}_{D,k} - \Xi^{M}_{R,k} \right\|_2^2. \tag{7}$$

Then $u_0 = 0$ and, for $k = 0, \ldots, K-1$,

$$u_{k+1}^2 \leq \left(c^{\mathcal{N}}_k u_k + \beta^{\mathcal{N}}_k \right)^2 + v^{\mathcal{N},M}_k. \tag{8}$$

Rather than characterizing the coupled trajectory separation u_k exactly, we control its evolution using the geometric discrepancy envelope *a_k^ν, defined recursively by*

$$a_0^\nu = 0, \qquad a_{k+1}^\nu = \sqrt{\left(c_k^{\mathcal{N}} a_k^\nu + \beta_k^{\mathcal{N}}\right)^2 + v_k^{\mathcal{N},M}}, \tag{9}$$

for $k = 0, \ldots, K-1$. This envelope satisfies $u_k \le a_k^\nu$ for every $k = 0, \ldots, K$. In particular,

$$W_2(P_D^\nu, P_R^\nu) \le a_K^\nu, \qquad W_2\big((\Psi_0)_{P_D^\nu}, (\Psi_0)_{P_R^\nu}\big) \le L_0 a_K^\nu. \tag{10}$$

For $M = \infty$, the envelope has the explicit endpoint

$$a_K^\nu = \sum_{s=0}^{K-1} \beta_s^{\mathcal{N}} \prod_{j=s+1}^{K-1} c_j^{\mathcal{N}}. \tag{11}$$

Formal conditions and proof are in Appendix B.2. Theorem 1 separates the channel-dependent mean-update terms, $\beta_k^{\mathcal{N}}$ and $c_k^{\mathcal{N}}$, from the nonnegative finite-shot term $v_k^{\mathcal{N},M}$, which can only loosen the bound when the mean-update terms are fixed. For the channel-free independent parameter-shift estimator in Appendix B.4, $v_k^{\mathcal{N},M} = O(M^{-1})$; under common mean-update bounds across shot budgets, the resulting envelope is non-increasing in M and approaches the exact-expectation bound with an $O(M^{-1/2})$ excess over a fixed training horizon.

3.3 Quantum Noise Channel-Specific Bounds on Discrepancy

We now derive explicit channel-specific bounds from Theorem 1. Assume that the propagation factors $c_k^{\mathcal{N}}$ in Eq. 5 satisfy $c_k^{\mathcal{N}} \le c$ for a common $0 \le c < 1$, at every step and across all compared noise settings. By Eq. 5, this means that each retain-data update map is c-Lipschitz on $\mathcal{P}$ and therefore contracts existing parameter differences to at most a factor c. This is a contraction assumption of the parameter-space update map, not of the quantum channel itself. Under this common propagation guarantee, channel-dependent changes in the bound can be compared through $\beta_k^{\mathcal{N}}$.

We consider exact-expectation training ($M = \infty$) with a parameter-independent local channel applied to the measured qubit immediately before the final Pauli-Z readout. This readout provides a bounded scalar prediction directly from computational-basis measurements, as the difference between the probabilities of outcomes 0 and 1. For input x, denote this prediction by $f_\theta^{\mathcal{N}}(x)$. As a general supervision form, we use squared loss $\ell(f, y) = (f - y)^2/2$ with $|y| \le 1$ and the common regularizer. At this placement, the channel's effect on predictions and gradients is independent of the preceding ansatz and qubit count. For interleaved channels, subsequent gates act on noise-perturbed states, making the effect circuit-dependent. Repeated depolarizing channels can be reduced to a single effective channel in special cases (Du et al., 2021), but this simplification does not generally apply to our local interleaved placement. We therefore focus our analysis below to the readout.

Corollary 1 (Channel-specific bounds under common contraction)**.** *Under the preceding readout setting and the assumptions of Theorem 1, take constants $\eta > 0$ and $0 \le c < 1$ such that $\eta_k = \eta$ and $c_k^{\mathcal{N}} \le c$ for every step. Assume that the channel-free predictor $f_\theta^{\mathcal{I}}(x)$ has a uniformly bounded parameter-gradient norm: $\|\nabla_\theta f_\theta^{\mathcal{I}}(x)\|_2 \le L_f$ for a finite $L_f \ge 0$, all $\theta \in \mathcal{P}$, and all $(x, y) \in D$. Then a_K^ν, and hence $W_2(P_D^\nu, P_R^\nu)$, is bounded above by the corresponding expression in Table 1.*

Because $c < 1$, each retain-data update attenuates parameter differences accumulated from earlier steps. Accordingly, the common term, i.e., $(1 - c^K)/(1 - c) \le 1/(1 - c)$, prevents these propagated differences from accumulating in an unbounded manner over training, provided the same assumptions and constants remain valid. This does not imply that the discrepancy vanishes as K increases, since each step may introduce a deletion-induced perturbation bounded by $\beta_k^{\mathcal{N}}$. Appendix C provides the proof, sufficient conditions for common contraction, and a finite-shot extension.

Table 1 shows how the channel parameters modify the discrepancy bound under this common contraction condition. **Depolarization**, parameterized by the noise strength $0 \le q \le 1$, monotonically tightens the bound through $(1 - q)(2 - q)$, while **amplitude damping** yields a linear reduction through $1 - \gamma$, whose strength is governed by $0 \le \gamma \le 1$. **Phase damping** leaves the bound unchanged from the noiseless case because a phase-damping channel immediately before a Pauli-Z readout preserves the corresponding expectation and gradient. For **bit flip**, whose flip probability is

Table 1: Channel-specific upper bounds under the common contraction assumption of Corollary 1.

Channel	Parameter	Upper bound on a_K^ν	
Noiseless / Phase damping	$-$ / γ	$4\eta\frac{\lvert F\rvert}{\lvert D\rvert}L_f\frac{1-c^K}{1-c}$	(12)
Depolarization	q	$2\eta\frac{\lvert F\rvert}{\lvert D\rvert}(1-q)(2-q)L_f\frac{1-c^K}{1-c}$	(13)
Amplitude damping	γ	$4\eta\frac{\lvert F\rvert}{\lvert D\rvert}(1-\gamma)L_f\frac{1-c^K}{1-c}$	(14)
Bit flip	p	$2\eta\frac{\lvert F\rvert}{\lvert D\rvert}\lvert 1-2p\rvert\big(\lvert 1-2p\rvert+1\big)L_f\frac{1-c^K}{1-c}$	(15)

$0 \le p \le 1$, the bound term $|1-2p|(|1-2p|+1)$ decreases up to $p = 0.5$ and increases thereafter; for $p \le 0.5$, the bound coincides with depolarization under $q = 2p$.

These readout-level relations need not persist under interleaved placement, where subsequent circuit layers act on noise-perturbed states. Section 4 therefore examines the actual pre-to-post discrepancy under both placements and reports trajectory-level contraction diagnostics.

3.4 From Geometric Proximity to Unlearning Certification

The preceding analysis shows how quantum noise can tighten a geometric bound between the original and retrained model distributions. However, geometric proximity alone need not imply small deletion sensitivity. The following proposition gives an explicit counterexample.

Proposition 1 (Vanishing geometric discrepancy with maximal deletion sensitivity)**.** *There exist a fixed single-qubit QNN, dataset pair (D, F), squared loss, common quadratic regularizer, fixed step size, and a smooth nondegenerate initialization distribution such that, for every $0 \le q < 1$, one training step under $\nu = (\mathcal{D}_q, \infty)$ satisfies the assumptions of Theorem 1 and*

$$W_2(P_D^\nu, P_R^\nu) \le a_1^\nu = \frac{1-q}{8}, \tag{16}$$

$$\mathcal{S}_\epsilon^\nu(\mathsf{U}_{\mathrm{id}}) = 1 \quad \textit{for every finite } \epsilon \ge 0. \tag{17}$$

The initialization distribution is fixed independently of q. Hence, $W_2(P_D^\nu, P_R^\nu) \to 0$ as $q \to 1$, while the deletion sensitivity remains maximal for every $q < 1$.

The significance of the construction is not merely that training on D and R can produce distinguishable models. Rather, a fixed QNN, dataset pair, optimization rule, and smooth initialization distribution yield endpoint distributions that become arbitrarily close in Wasserstein distance as $q \to 1$, while remaining perfectly distinguishable for every $q < 1$. Thus, under the assumptions considered here, there is no distribution-free implication from vanishing geometric discrepancy to vanishing deletion sensitivity. The single-qubit construction shows that this separation is intrinsic to the two notions of discrepancy rather than a consequence of high-dimensional model geometry. Appendix B.5 gives the construction and verifies the assumptions.

4 Experiment 1: Measurements of Pre-to-Post Discrepancy

We examine whether noisy training reduces the empirical pre-to-post discrepancy relative to noiseless training while preserving predictive utility. We measure this discrepancy in parameter and prediction spaces under both readout and interleaved placements, and separately diagnose retain-update contraction in selected exact-expectation readout settings.

4.1 Experimental Settings

Data and model. We use the two-class versicolor–virginica subset of Iris (Fisher, 1936) and construct five repeated stratified splits of 60/20/20 train/validation/test examples. For each partition, the deletion set contains $|F| = 6$ class-balanced examples, corresponding to 10% of D. We use a

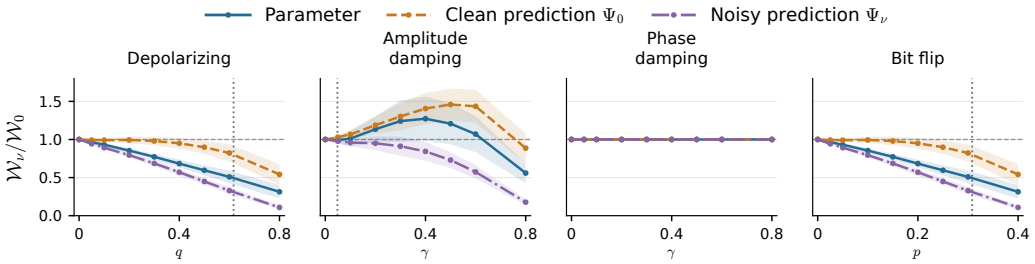

Figure 2: Readout placement and empirical pre-to-post discrepancy. Lines and bands show the mean and one standard deviation of five within partition ratios. Dotted vertical lines mark the 80% crossings of the lower full/retrain mean accuracy.

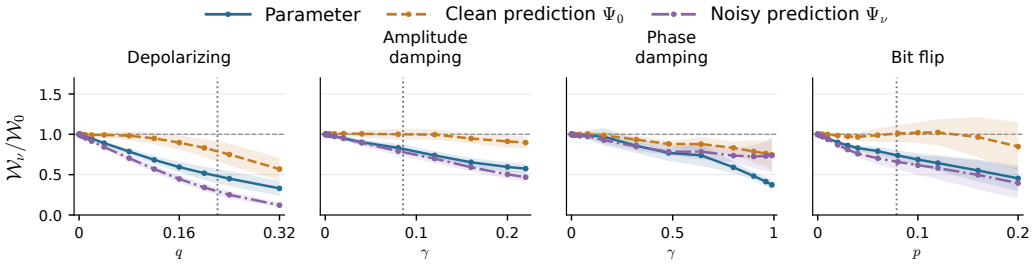

Figure 3: Interleaved placement and empirical pre-to-post discrepancy.

4-qubit classifier with two variational layers, 16 trainable parameters, and a Z_0 readout, trained with regularized squared loss. Full-data training and retraining use identical optimization settings.

Noise settings. We consider depolarizing, amplitude damping, phase damping, and bit flip channels under readout and interleaved placements. Readout noise acts on the measured qubit immediately before the Z_0 measurement, matching Section 3.3, whereas interleaved noise acts on all qubits after each ansatz layer. Full-data training and retraining use the same channel type, strength, and placement. Channel sweeps use exact expectations ($M = \infty$). Separately, we isolate finite-shot effects without channel noise by varying M with conditionally unbiased gradients and independent measurement draws across the two training trajectories.

Model distributions and evaluation. For each noise setting, we train 1,024 paired initialization full-data/retrained model pairs per partition, totaling 10,240 models across five partitions. We measure empirical W_2 distances in parameter space and under two prediction audits, using $\mathcal{W}_\nu$ to denote the empirical distance in each respective space. For prediction audits, each model is represented by its predictions on a common set of inputs. The *clean audit* uses the exact, channel-free map Ψ_0 to isolate training-induced changes, whereas the *noise-matched audit* uses Ψ_ν and also captures noise effects at evaluation. We report $\mathcal{W}_\nu/\mathcal{W}_0$, where $\mathcal{W}_0$ is the corresponding distance under noiseless training, $\nu_0 = (\mathcal{I}, \infty)$. The parameter-space ratio uses the noiseless parameter distance; both prediction-space ratios use the noiseless Ψ_0 distance. Ratios below one indicate a smaller empirical discrepancy in that space. Appendix D details the training, noise, and evaluation protocols; Appendix E.1 examines how quantum noise affects retraining discrepancy when deleting mislabeled (poisoned) data (Chen & Zhang, 2026).

4.2 Results

Readout placement and contraction diagnosis. Figure 2 evaluates the analytically tractable readout setting of Section 3.3. Under the common contraction assumption, Eq. 13 gives a discrepancy upper bound whose channel-dependent factor decreases monotonically with q. Empirically, at $q = 0.2$, the parameter-space discrepancy decreases by 14.58%. Clean-prediction reductions are smaller, whereas noise-matched prediction distances decrease more strongly because they additionally include direct readout attenuation.

Table 3 further examines retain-update propagation along the sampled trajectories. Across the noiseless and tested depolarizing settings, 82.3–90.7% of pair-step observations are contractive. This indicates that contraction is common along the sampled trajectories, but does not establish the uniform condition $c_k^{\mathcal{N}} < 1$ required by Corollary 1. Amplitude damping exhibits less frequent pairwise

contraction as γ increases. Under normal deletion, the pairwise contraction rate falls to 64.1% at $\gamma = 0.4$. This helps explain why the decreasing amplitude damping bound in Eq. 14, which assumes a common $c < 1$, is not reflected by the empirical parameter discrepancy: as indicated by Theorem 1, weaker propagation control can preserve or amplify previously accumulated differences.

Interleaved placement. Figure 3 examines interleaved noise, where the readout-level relations of Section 3.3 need not hold. Depolarization shows the most consistent behavior, reducing both parameter- and prediction-space discrepancies. Amplitude damping at $\gamma = 0.04$ reduces the parameter discrepancy by 9.87%, although the corresponding clean-prediction change is small. Phase damping, which suppresses coherence while preserving computational basis populations, now affects the learned models because later circuit operations process the dephased states; its reductions persist under clean evaluation. Bit flip likewise no longer follows the exact depolarizing correspondence of the readout case: intermediate basis flips are propagated through subsequent layers, producing clear parameter-space reductions but a less uniform clean-prediction response.

Taken together, selected channel noise regimes reduce the empirical pre-to-post discrepancy while preserving predictive utility, with the effect depending on placement. These empirical reductions do not establish the ordering or tightness of the upper bounds in Table 1, but are consistent with the mechanism of Theorem 1: channel-conditioned training can reduce discrepancy when smaller deletion-induced update differences are not offset by stronger propagation.

Finite-shot measurement. We further conducted an experiment to examine the empirical pre-to-post discrepancy induced by finite-shot measurement. Figure 4 isolates finite-shot training without channel noise. We use $\mathcal{W}_M/\mathcal{W}_\infty$ for the finite-shot sweep, where $\mathcal{W}_M$ with $\nu = (\mathcal{I}, M)$, and $\mathcal{W}_\infty$ with $\nu = (\mathcal{I}, \infty)$. At $M = 32$, the parameter-space discrepancy exceeds the exact-expectation reference by 16.06%, and approaches the reference as M increases. This trend is consistent with Theorem 1, whose finite-shot envelope decreases toward the exact-expectation bound as the shot budget increases. The larger discrepancy under noise-matched prediction also includes finite-shot sampling at evaluation, as shown by the control analysis in Appendix D.3. Overall, finite-shot measurement shows no reduction comparable to the selected channel-noise regimes in the settings examined.

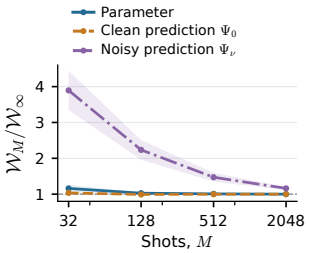

Figure 4: Finite-shot measurement and empirical pre-to-post discrepancy.

5 Experiment 2: Unlearning Optimization Cost

We extend the study to a 10-qubit QNN on MNIST to examine whether noise-conditioned training changes the effort required by unlearning. We compare depolarization and amplitude damping because they exhibit contrasting parameter-space discrepancy trends in Section 4, and ask whether these different starting models lead to different unlearning costs under fixed unlearning procedures.

5.1 Setting

We study digit-0 unlearning on MNIST using a 10-qubit QNN with 16 ansatz layers and 10 Pauli-Z readouts. Using amplitude-encoded images and five stratified 2,000/1,000/2,000 train/validation/test splits, full training and retraining run for 2,000 minibatch-SGD updates (batch size 32) with squared loss on centered one-hot targets. We compare noiseless training with depolarizing and amplitude damping readout noise applied to all qubits immediately before measurement. Each condition uses eight distinct initializations, paired across full training, retraining, and noise settings. We measure empirical W_2 in parameter and prediction spaces.

For unlearning, we apply GA (Thudi et al., 2022), FT, GA+FT, SalUn (Fan et al., 2024) and DELETE (Zhou et al., 2025) adaptations using Adam with learning rate 0.007. We consider unlearning performed on the same quantum device used for training. Success requires forget training and validation accuracies both at most 3%. Table 2 reports mean first-hit steps and corresponding test forget/retain accuracies over successful runs only. Appendix F provides the details of training, noise, and unlearning process, and Appendix G reports baseline performances.

Table 2: MNIST digit-0 unlearning across five partitions and eight paired initializations per partition. Forget Accuracy (FA), Retain Accuracy (RA), and Steps average only successful runs reaching both forget training and validation accuracy $\leq 3\%$. Green/red indicate lower/higher mean steps than Noiseless.

	Noiseless			Depol $q = 0.05$			Depol $q = 0.20$			Amp $\gamma = 0.02$			Amp $\gamma = 0.08$		
Method	FA	RA	Steps	FA	RA	Steps	FA	RA	Steps	FA	RA	Steps	FA	RA	Steps
Original	94.95%	71.34%	–	94.76%	71.03%	–	94.27%	70.24%	–	94.95%	71.68%	–	94.16%	69.73%	–
Retrain	0.00%	75.17%	–	0.06%	74.83%	–	0.03%	74.07%	–	0.00%	74.78%	–	0.04%	72.77%	–
GA	1.44%	69.81%	8.3	1.29%	69.33%	8.5	1.55%	68.23%	8.4	1.31%	69.87%	8.5	1.34%	68.32%	8.3
FT	3.01%	73.91%	159.5	2.85%	73.64%	176.7	2.63%	72.98%	156.1	3.35%	73.87%	200.1	2.68%	72.41%	204.2
GA+FT	1.85%	70.45%	13.4	1.95%	70.33%	13.4	1.96%	69.44%	13.4	1.91%	70.95%	13.6	1.80%	69.27%	13.4
SalUn	3.00%	72.55%	306.7	2.56%	71.92%	218.7	2.95%	71.64%	241.6	3.45%	72.78%	321.2	2.58%	70.75%	274.2
DELETE	1.55%	69.79%	8.4	1.59%	69.53%	8.4	1.78%	68.34%	8.4	1.46%	69.72%	8.6	1.36%	68.24%	8.3
$\mathcal{W}_\nu/\mathcal{W}_0$	1.0000			0.9900			0.9608			1.0096			1.0227		
Mean cost reduction	0.00%			+3.18%			+4.50%			−7.08%			−3.23%		

Model discrepancy. Figure 5 examines whether the channel-dependent discrepancy trends persist in the larger QNN. Under depolarization, the parameter-space distance decreases with q, reaching a 3.92% reduction at $q = 0.20$ relative to noiseless training. The clean-prediction distance instead increases by 3.50%, indicating that the parameter-space reduction does not necessarily transfer to prediction space. Amplitude damping instead increases the mean parameter-space distance by 0.96% at $\gamma = 0.02$ and 2.27% at $\gamma = 0.08$. This increase is qualitatively consistent with the readout results in Figure 2 and contrasts with the reductions under depolarization.

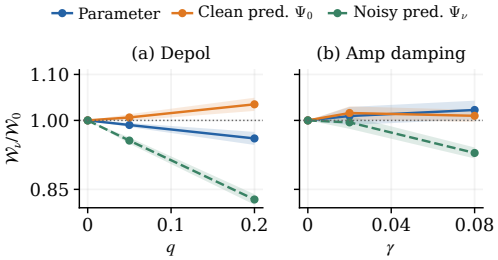

Figure 5: Channel noise effects on the pre-to-post discrepancy for digit-0 deletion in the 10-qubit QNN MNIST experiment.

Unlearning cost. Table 2 shows contrasting cost trends under the two channels. We compute each method's relative reduction in success-conditioned mean steps from its noiseless baseline, then average equally across the five methods. Depolarization lowers cost by 3.18% at $q = 0.05$ and 4.50% at $q = 0.20$, whereas amplitude damping increases it by 7.08% at $\gamma = 0.02$ and 3.23% at $\gamma = 0.08$. These directions align with the parameter-space discrepancy: depolarization brings the full-training and retraining distributions closer, while amplitude damping increases their mean separation. Benefits nevertheless vary by method, as their objectives do not directly minimize distance to the retrained distribution. SalUn shows the largest depolarizing gains, requiring 218.7 and 241.6 steps versus 306.7 without noise. FT improves slightly at $q = 0.20$ but slows at $q = 0.05$. Amplitude damping increases FT's cost at both strengths, although SalUn benefits at $\gamma = 0.08$. GA, GA+FT, and DELETE remain near their noiseless costs. Overall, these results show that favorable quantum-noise regimes can reduce both pre-to-post discrepancy and practical unlearning costs.

6 Conclusion

We studied how native quantum noise shapes the discrepancy between original and retrained QNN distributions. Our analysis separates discrepancy newly introduced by deletion from the propagation of previously accumulated model differences through retraining updates, while also accounting for finite-shot errors, yielding a Wasserstein stability bound. Under a common contraction assumption, we further derive closed-form channel-specific bounds that characterize how quantum noise changes a geometric upper bound under common contraction. Experiments identify readout and interleaved channel regimes that reduce empirical discrepancy. On MNIST with a 10-qubit QNN, selected readout-noise regimes also reduce unlearning costs, although the gains depend on the channel and unlearning method.

Limitations. Our explicit channel-specific bounds require common contraction and final-readout noise; extending them to interleaved channels and more general optimization dynamics remains open. Parameter-space Wasserstein proximity need not imply functional proximity or certified unlearning. Our cost results use success-conditioned first-hit updates to a class-forgetting criterion, and thus do not establish unconditional speedups or convergence to the retrained distribution.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing, assist in the writing of proofs, assist with translation. We have not used generative AI tools for Generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. We have reviewed all AI-assisted work. All sentences and grammar refined through AI were rechecked 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.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or newly collected sensitive data. Our experiments use standard benchmark datasets (Iris and MNIST) and simulated quantum neural networks under controlled noise models. Although machine unlearning is motivated in part by privacy and data-removal concerns, our geometric discrepancy analysis does not constitute a privacy guarantee, and we explicitly distinguish these notions throughout the paper. We are not aware of additional ethical concerns specific to the methodology presented in this work.

REPRODUCIBILITY STATEMENT

We provide detailed experimental settings in Appendices D and F, including data preprocessing, model architectures, training procedures, noise configurations, unlearning methods, and evaluation protocols. Upon acceptance, we will publicly release the complete experimental code, configurations, and random seeds used to produce the reported results.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pp. 308–318, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341394. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

Borja Balle, Gilles Barthe, Marco Gaboardi, and Joseph Geumlek. Privacy amplification by mixing and diffusion mechanisms. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/c4c42505a03f2e969b4c0a97ee9b34e7-Paper.pdf.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (S&P)*, pp. 141–159, 2021. doi: 10.1109/SP40001.2021.00019.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy (S&P)*, pp. 463–480, 2015. doi: 10.1109/SP.2015.35.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, pp. 267–284, USA, 2019. USENIX Association. ISBN 9781939133069.

M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, 2021. ISSN 2522-5820. doi: 10.1038/s42254-021-00348-9. URL https://doi.org/10.1038/s42254-021-00348-9.

Yu-Qin Chen and Shi-Xin Zhang. Superior resilience to poisoning and amenability to unlearning in quantum machine learning. *Nature Communications*, 17(1):3716, 2026. ISSN 2041-1723. doi: 10.1038/s41467-026-70420-4. URL https://doi.org/10.1038/s41467-026-70420-4.

Eli Chien, Haoyu Wang, Ziang Chen, and Pan Li. Langevin unlearning: A new perspective of noisy gradient descent for machine unlearning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 79666–79703. Curran Associates, Inc., 2024. doi: 10.52202/079017-2530. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9161a3e810cc0542718abddb791e8aed-Paper-Conference.pdf.

Carla Crivoi and Radu Tudor Ionescu. Machine unlearning in the era of quantum machine learning: An empirical study. In *Pattern Recognition: 28th International Conference, ICPR 2026, Lyon, France, August 17–22, 2026, Proceedings, Part VI*, pp. 31–47, Berlin, Heidelberg, 2026. Springer-Verlag. ISBN 978-3-032-31672-1. doi: 10.1007/978-3-032-31673-8_3. URL https://doi.org/10.1007/978-3-032-31673-8_3.

Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, Shan You, and Dacheng Tao. Learnability of quantum neural networks. *PRX Quantum*, 2:040337, Nov 2021. doi: 10.1103/PRXQuantum.2.040337. URL https://link.aps.org/doi/10.1103/PRXQuantum.2.040337.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations (ICLR)*, pp. 53643–53673, 2024.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179–188, 1936. doi: https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Oum Gadani, Kandarp Gajjar, Himani Trivedi, and Hiren Patel. Self-revealing poisons: loss-guided forensic detection and quantum unlearning of corrupted training data. *Physica Scripta*, 101(33): 336001, aug 2026. doi: 10.1088/1402-4896/ae9810. URL https://doi.org/10.1088/1402-4896/ae9810.

Elies Gil-Fuster, Jens Eisert, and Carlos Bravo-Prieto. Understanding quantum machine learning also requires rethinking generalization. *Nature Communications*, 15(1):2277, 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-45882-z. URL https://doi.org/10.1038/s41467-024-45882-z.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making AI forget you: Data deletion in machine learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc., 2019.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3832–3842. PMLR, July 2020.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1225–1234, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/hardt16.html.

Anastasia Koloskova, Youssef Allouah, Animesh Jha, Rachid Guerraoui, and Sanmi Koyejo. Certified unlearning for neural networks. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31275–31298. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/koloskova25a.html.

Viacheslav Kuzmin, Wilfrid Somogyi, Ekaterina Pankovets, and Alexey Melnikov. Method for noise-induced regularization in quantum neural networks. *Advanced Quantum Technologies*, 8 (12):e00603, 2025. doi: https://doi.org/10.1002/qute.202400603. URL https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/qute.202400603.

Md Akmol Masud, Zannat Hossain Tamim, Md Musa Kalimullah Ratul, and Md Biplob Hosen. Squ: Sensitivity-guided quantum machine unlearning for privacy-preserving ml. In *2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence Networking (QPAIN)*, pp. 1–6, 2026. doi: 10.1109/QPAIN69676.2026.11546660.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 605–638. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/mou18a.html.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-Delete: Gradient-based methods for machine unlearning. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato (eds.), *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 931–962. PMLR, March 2021.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 18075–18086. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/9627c45df543c816a3ddf2d8ea686a99-Paper.pdf.

Junjian Su, Runze He, Guanghui Li, Sujuan Qin, Zhimin He, Haozhen Situ, and Fei Gao. From membership-privacy leakage to quantum machine unlearning. *Phys. Rev. Appl.*, 25:044056, Apr 2026. doi: 10.1103/mdfc-pgsl. URL https://link.aps.org/doi/10.1103/mdfc-pgsl.

Ryan Sweke, Frederik Wilde, Johannes Meyer, Maria Schuld, Paul K. Faehrmann, Barthélémy Meynard-Piganeau, and Jens Eisert. Stochastic gradient descent for hybrid quantum-classical optimization. *Quantum*, 4:314, August 2020. ISSN 2521-327X. doi: 10.22331/q-2020-08-31-314. URL https://doi.org/10.22331/q-2020-08-31-314.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pp. 303–319, 2022. doi: 10.1109/EuroSP53844.2022.00027.

Enayat Ullah, Tung Mai, Anup Rao, Ryan A. Rossi, and Raman Arora. Machine unlearning via algorithmic stability. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4126–4142. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/ullah21a.html.

Samson Wang, Enrico Fontana, M. 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. ISSN 2041-1723. doi: 10.1038/s41467-021-27045-6. URL https://doi.org/10.1038/s41467-021-27045-6.

Jiaqi Yang, Wei Xie, and Xiaohua Xu. Stability and generalization of quantum neural networks. *arXiv preprint arXiv:2501.12737*, 2025.

Yu Zhou, Dian Zheng, Qijie Mo, Renjie Lu, Kun-Yu Lin, and Wei-Shi Zheng. Decoupled distillation to erase: A general unlearning method for any class-centric tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20350–20359, June 2025. doi: 10.1109/CVPR52734.2025.01895.

A PRELIMINARIES

A.1 PARAMETERIZED QUANTUM CIRCUIT

In classical learning, a classical computer processes an input through a parameterized model, produces a prediction, evaluates a loss, and updates parameters through an optimization procedure. In contrast, the QNN framework considered here uses a quantum circuit to implement the trainable transformation, while loss evaluation and parameter updates remain classical. In this hybrid workflow, classical input x is first encoded in a quantum state $\rho_{\mathrm{enc}}(x)$ and then processed by a unitary trainable transformation U_θ called a parameterized quantum circuit. Finally, measurement/readout produces the classical prediction used to evaluate the loss and update the circuit parameters with a classical optimizer.

To understand this workflow in concrete terms, we consider the four-qubit parameterized quantum circuit. Let $x = (x_0, \ldots, x_3)$ denote the preprocessed input angles supplied to the circuit. Starting from $|0\rangle^{\otimes 4}$, the four input angles are encoded into the quantum state through single-qubit R_y rotations,

$$|\psi_{\mathrm{enc}}(x)\rangle = \bigotimes_{j=0}^{3} R_y(x_j)|0\rangle, \tag{18}$$

with $\rho_{\mathrm{enc}}(x) = |\psi_{\mathrm{enc}}(x)\rangle\langle\psi_{\mathrm{enc}}(x)|$. The encoded state is then processed by a two-layer parameterized quantum circuit. In each layer, every qubit undergoes a trainable R_y rotation followed by a trainable R_z rotation, after which a ring of CNOT gates entangles neighboring qubits in the order $0 \to 1$, $1 \to 2$, $2 \to 3$, and $3 \to 0$.

More formally, let $\theta_\ell := \left(\theta^y_{\ell,0}, \theta^z_{\ell,0}, \ldots, \theta^y_{\ell,3}, \theta^z_{\ell,3}\right) \in \mathbb{R}^8$ for each layer $\ell \in \{1, 2\}$ denote its trainable parameters. Then,

$$U_\ell(\theta_\ell) = U_{\mathrm{ent}} \left[\bigotimes_{j=0}^{3} R_z(\theta^z_{\ell,j}) R_y(\theta^y_{\ell,j}) \right], \tag{19}$$

where the entangling operation is $U_{\mathrm{ent}} = \mathrm{CNOT}_{3\to0}\mathrm{CNOT}_{2\to3}\mathrm{CNOT}_{1\to2}\mathrm{CNOT}_{0\to1}$. The complete transformation is $U_\theta = U_2(\theta_2)U_1(\theta_1)$. Since each of the two layers contains an R_y and an R_z rotation on each of the four qubits, the circuit contains $2 \times 4 \times 2 = 16$ trainable parameters.

After transformation, the output quantum state is $\rho_\theta(x) := U_\theta \rho_{\mathrm{enc}}(x) U_\theta^\dagger$. The classifier then converts this state into a classical prediction by measuring the Pauli-Z observable on qubit 0. The corresponding expectation value of the ideal (channel-free) prediction is

$$f_\theta^{\mathcal{I}}(x) = \mathrm{Tr}\left[Z_0 \, U_\theta \rho_{\mathrm{enc}}(x) U_\theta^\dagger \right], \tag{20}$$

where $Z_0 = Z \otimes I^{\otimes 3}$. When a quantum channel $\mathcal{N}$ is present, we denote the resulting channel-conditioned prediction by $f_\theta^{\mathcal{N}}(x)$. For the classifier considered in Experiment 1, given a training set Z, we optimize the circuit parameters using the regularized squared loss

$$L_Z^{\mathcal{N}}(\theta) = \frac{1}{2|Z|} \sum_{(x,y)\in Z} \left(f_\theta^{\mathcal{N}}(x) - y \right)^2 + \frac{\lambda}{2} \|\theta\|_2^2. \tag{21}$$

A.2 QUANTUM NOISE CHANNEL

Practical quantum circuits are inevitably affected by hardware-induced noise, which perturbs the quantum state during circuit execution, and consequently alters the predictions and gradients used for training. Such physical noise is represented as a quantum channel, described by a completely positive trace-preserving (CPTP) map $\mathcal{N} : \rho \mapsto \mathcal{N}(\rho)$. Any finite-dimensional CPTP map admits a Kraus representation

$$\mathcal{N}(\rho) = \sum_i K_i \rho K_i^\dagger, \qquad \sum_i K_i^\dagger K_i = I, \tag{22}$$

where $\{K_i\}$ are called Kraus operators, which are not uniquely determined because different sets of Kraus operators can represent the same quantum channel.

Depolarization. Depolarization uniformly reduces the information contained in a quantum state, driving the state toward the maximally mixed state as the noise strength q increases. For a single qubit, this effect is described by $\mathcal{D}_q(\rho) = (1-q)\rho + q\frac{I}{2},\ 0 \le q \le 1$ and one Kraus representation consistent with this convention is

$$K_0 = \sqrt{1-\frac{3q}{4}}\, I, \qquad K_1 = \sqrt{\frac{q}{4}}\, X, \\ K_2 = \sqrt{\frac{q}{4}}\, Y, \qquad K_3 = \sqrt{\frac{q}{4}}\, Z. \tag{23}$$

Geometrically, depolarization contracts the Bloch vector isotropically by a factor $1-q$ and a traceless Pauli observable satisfies $f_\theta^{\mathcal{D}_q}(x) = (1-q) f_\theta^{\mathcal{I}}(x)$, when it is applied immediately before readout.

Amplitude Damping. Amplitude damping models a one-way transition from $|1\rangle$ to $|0\rangle$, with larger γ producing a stronger bias toward $|0\rangle$. Unlike depolarization, which uniformly suppresses state information, amplitude damping captures a directional change toward a particular state. Its Kraus operators are

$$K_0 = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-\gamma} \end{pmatrix}, \qquad K_1 = \begin{pmatrix} 0 & \sqrt{\gamma} \\ 0 & 0 \end{pmatrix}, \qquad 0 \le \gamma \le 1. \tag{24}$$

Because a Pauli-Z measurement assigns $+1$ to $|0\rangle$ and -1 to $|1\rangle$, this one-way transition attenuates the original Z expectation while shifting it toward $+1$: $\mathcal{A}_\gamma^*(Z) = (1-\gamma)Z + \gamma I$. Consequently, when amplitude damping is applied immediately before a Z readout, $f_\theta^{\mathcal{A}_\gamma}(x) = (1-\gamma) f_\theta^{\mathcal{I}}(x) + \gamma$.

Phase Damping. Phase damping models the relative-phase information loss between $|0\rangle$ and $|1\rangle$ without changing the probabilities of these two basis states. With damping strength γ, let $\kappa = \sqrt{1-\gamma},\ 0 \le \gamma \le 1$, where κ determines how much of the phase information is retained. A Kraus representation is

$$K_0 = \sqrt{\kappa}\, I, \qquad K_1 = \sqrt{1-\kappa}\, |0\rangle\langle 0|, \qquad K_2 = \sqrt{1-\kappa}\, |1\rangle\langle 1|. \tag{25}$$

For a single-qubit state

$$\rho = \begin{pmatrix} \rho_{00} & \rho_{01} \\ \rho_{10} & \rho_{11} \end{pmatrix}, \tag{26}$$

the channel acts as

$$\mathcal{P}_\gamma(\rho) = \begin{pmatrix} \rho_{00} & \kappa\rho_{01} \\ \kappa\rho_{10} & \rho_{11} \end{pmatrix}. \tag{27}$$

Thus, phase damping suppresses only the off-diagonal terms while leaving the diagonal terms unchanged. Because a Pauli-Z readout depends only on the diagonal terms, phase damping applied immediately before the readout leaves its expectation value unchanged, i.e., $f_\theta^{\mathcal{P}_\gamma}(x) = f_\theta^{\mathcal{I}}(x)$.

Bit Flip Noise. Bit flip noise models errors that interchange the computational basis states $|0\rangle$ and $|1\rangle$. With flip probability p, the channel is defined as $\mathcal{B}_p(\rho) = (1-p)\rho + pX\rho X,\ 0 \le p \le 1$. Its Kraus operators are

$$K_0 = \sqrt{1-p}\, I, \qquad K_1 = \sqrt{p}\, X. \tag{28}$$

Because a Pauli-X operation reverses the sign of a Z measurement, $XZX = -Z$, applying bit flip noise immediately before a Z readout gives $f_\theta^{\mathcal{B}_p}(x) = (1-2p) f_\theta^{\mathcal{I}}(x)$. Thus, increasing p progressively suppresses the Z expectation up to $p = 1/2$, after which its sign is reversed.

B Proofs and Extensions of the Main Results

Following Section 3, we first interpret deletion sensitivity, prove Theorem 1, and extend its recursion to minibatch SGD. We then bound the finite-shot contribution and prove the separation and audit results from Section 3.4. Channel-specific calculations are collected in Appendix C.

B.1 Interpretation of Deletion Sensitivity

Definition 1 fixes the dataset pair (D, F), noise regime ν, and learning procedure. In particular, retraining uses the same initialization law and training schedule; the ideal regime removes quantum noise but need not remove other training randomness. For fixed $\epsilon \geq 0$, the event inequalities

$$\begin{aligned} P^{\nu}_{D,F}(B) &\leq e^{\epsilon} P^{\nu}_{R}(B) + \delta, \\ P^{\nu}_{R}(B) &\leq e^{\epsilon} P^{\nu}_{D,F}(B) + \delta \end{aligned} \tag{29}$$

hold for every measurable B exactly when $\mathcal{S}^{\nu}_{\epsilon}(\mathsf{U}_{\nu}) \leq \delta$. This follows by taking the event supremum in each direction. Thus $\mathcal{S}^{\nu}_{\epsilon} \in [0, 1]$ is the smallest admissible additive tolerance. For U_{id}, its first argument is the original trained-model law P^{ν}_{D}.

At $\epsilon = 0$, the complement identity $P(B) - Q(B) = Q(B^c) - P(B^c)$ gives $\mathsf{H}_1(P\|Q) = \mathsf{H}_1(Q\|P) = \mathrm{TV}(P, Q)$, where $\mathrm{TV}(P, Q) = \sup_B |P(B) - Q(B)|$. Hence

$$\mathcal{S}^{\nu}_{0}(\mathsf{U}_{\nu}) = \mathrm{TV}(P^{\nu}_{D,F}, P^{\nu}_{R}). \tag{30}$$

In particular, $(0, 0)$-unlearning means equality of the two laws; zero additive tolerance at $\epsilon > 0$ need not imply equality.

B.2 Full-Batch Stability: Assumptions and Proof

Objective and coupling. Theorem 1 concerns the released parameter vector in a fixed Euclidean representation $\mathcal{P} \subseteq \mathbb{R}^p$. Its distance does not identify functionally equivalent parameters or cover additional released optimizer states. For $Z \in \{D, R\}$, let

$$\mathcal{L}^{\mathcal{N}}_{Z}(\theta) = \frac{1}{|Z|} \sum_{(x,y)\in Z} \ell(f^{\mathcal{N}}_{\theta}(x), y) + \mathcal{R}(\theta), \tag{31}$$

with a common differentiable regularizer. The predictions here are exact channel-conditioned expectations. Both runs use the same deterministic step sizes and horizon K.

Let $\mathcal{F}_k$ contain both histories before the current shot draws. An admissible coupling preserves each run's prescribed conditional sampling law at its current iterate and satisfies

$$\mathbb{E}[\Xi^{M}_{Z,k} \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}\|\Xi^{M}_{Z,k}\|_2^2 < \infty, \qquad \Xi^{\infty}_{Z,k} = 0. \tag{32}$$

The shared initialization and all iterates have finite second moments. Conditional independence between runs is sufficient but not necessary. Conditioning on the joint history is essential for the cross-term cancellation below; unbiasedness only with respect to each run's own history is insufficient under an arbitrary coupling. The assumption concerns the complete loss gradient, not merely the measured output; Appendix B.4 gives an estimator that satisfies it.

Propagation and trajectory containment. An $H^{\mathcal{N}}$-Lipschitz gradient of $\mathcal{L}^{\mathcal{N}}_{R}$ gives the valid factor $c^{\mathcal{N}}_{k} = 1 + \eta_k H^{\mathcal{N}}$ by the triangle inequality. Convexity of the objective is not required. For the stronger, contractive case, suppose on a convex region $\mathcal{P}$ that

$$0 < \mu^{\mathcal{N}} I \preceq \nabla^2 \mathcal{L}^{\mathcal{N}}_{R}(\theta) \preceq H^{\mathcal{N}} I. \tag{33}$$

Integrating the continuous Hessian along a line segment gives

$$c^{\mathcal{N}}_{k} = \max\{|1 - \eta_k \mu^{\mathcal{N}}|, |1 - \eta_k H^{\mathcal{N}}|\} < 1 \quad \text{if } 0 < \eta_k < 2/H^{\mathcal{N}}. \tag{34}$$

These are properties of the optimizer update, not consequences of contraction of quantum states under a channel. In either case, the theorem separately requires both trajectories to stay in $\mathcal{P}$ almost surely through step K. Local curvature alone does not ensure this, and conditioning afterward on staying in a chosen basin changes the endpoint laws.

For the rotation circuit, squared loss, and quadratic regularizer of Experiment 1 (Appendix D), one can instead take $\mathcal{P} = \mathbb{R}^p$. The Pauli expectation and its first two derivatives are uniformly bounded, by smooth periodicity of the rotation angles. Writing $L^{\mathcal{N}}_{f}$ and $B^{\mathcal{N}}_{f}$ for bounds on the gradient norm and Hessian operator norm, respectively, and using $|f|, |y| \leq 1$, gives

$$H^{\mathcal{N}} = (L^{\mathcal{N}}_{f})^2 + 2B^{\mathcal{N}}_{f} + \lambda, \qquad \beta^{\mathcal{N}}_{k} \leq 4\eta_k \frac{|F|}{|D|} L^{\mathcal{N}}_{f}.$$

The second bound follows from cancellation of the common regularizer and Eq. 63. The data-gradient estimates are bounded and the regularizer gradient is linear, so finite second moments of the Gaussian initialization persist at every finite step. The clean audit is globally Lipschitz by the same derivative argument. These observations ensure applicability without asserting strict contraction or tightness of the global bound.

Proof of Theorem 1. Under the stipulated coupling, set $\Delta_k = \Theta^{\nu}_{D,k} - \Theta^{\nu}_{R,k}$, $\Upsilon_k = \Xi^{M}_{D,k} - \Xi^{M}_{R,k}$, and $\Gamma_k = T^{\mathcal{N}}_{D,k}(\Theta^{\nu}_{D,k}) - T^{\mathcal{N}}_{R,k}(\Theta^{\nu}_{R,k})$. Adding and subtracting $T^{\mathcal{N}}_{R,k}(\Theta^{\nu}_{D,k})$ yields

$$\|\Gamma_k\|_2 \le c^{\mathcal{N}}_k \|\Delta_k\|_2 + \beta^{\mathcal{N}}_k. \tag{35}$$

Since $\Delta_{k+1} = \Gamma_k - \eta_k \Upsilon_k$, Γ_k is $\mathcal{F}_k$-measurable, and $\mathbb{E}[\Upsilon_k \mid \mathcal{F}_k] = 0$,

$$\mathbb{E}[\|\Delta_{k+1}\|_2^2 \mid \mathcal{F}_k] = \|\Gamma_k\|_2^2 + \eta_k^2 \mathbb{E}[\|\Upsilon_k\|_2^2 \mid \mathcal{F}_k]. \tag{36}$$

Let $u_k = (\mathbb{E}\|\Delta_k\|_2^2)^{1/2}$. Taking expectations and using $\mathbb{E}\|\Delta_k\|_2 \le u_k$ gives

$$u_{k+1}^2 \le (c^{\mathcal{N}}_k u_k + \beta^{\mathcal{N}}_k)^2 + v^{\mathcal{N},M}_k. \tag{37}$$

Shared initialization gives $u_0 = a^{\nu}_0 = 0$. Monotonicity of the scalar recursion in its nonnegative argument implies $u_k \le a^{\nu}_k$. Our coupled endpoints have the prescribed marginals, so the definition of W_2 as the infimum of root-mean-square coupling costs gives

$$W_2(P^{\nu}_D, P^{\nu}_R) \le u_K \le a^{\nu}_K. \tag{38}$$

Applying the fixed L_0-Lipschitz map Ψ_0 to these endpoints gives

$$W_2\big((\Psi_0)_{P^{\nu}_D}, (\Psi_0)_{P^{\nu}_R}\big) \le L_0 u_K \le L_0 a^{\nu}_K. \tag{39}$$

For $M = \infty$, the error term vanishes and the recursion becomes $a^{\nu}_{k+1} = c^{\mathcal{N}}_k a^{\nu}_k + \beta^{\mathcal{N}}_k$. Unrolling it proves Eq. 11, with empty products equal to one. □

B.3 Extension to Minibatch SGD

The same argument extends to unbiased minibatch gradients by separating data-sampling error from shot-estimation error. In this subsection, P^{ν}_Z denotes the endpoint law of the specified SGD procedure; its batch sizes and sampling rule are fixed when comparing noise regimes. Write $n_Z = |Z|$ and

$$g^{\mathcal{N}}_z(\theta) = \nabla_\theta \ell(f^{\mathcal{N}}_\theta(x), y), \qquad \overline{g}^{\mathcal{N}}_Z(\theta) = \frac{1}{n_Z} \sum_{z \in Z} g^{\mathcal{N}}_z(\theta), \quad z = (x, y).$$

Let $\mathcal{F}_k$ now denote the joint history before both batch and shot sampling. At step k, sample a batch $B_{Z,k}$ of deterministic size $b_{Z,k}$ and form the exact minibatch gradient

$$G_{Z,k} = \frac{1}{b_{Z,k}} \sum_{z \in B_{Z,k}} g^{\mathcal{N}}_z(\Theta^{\nu}_{Z,k}) + \nabla \mathcal{R}(\Theta^{\nu}_{Z,k}), \qquad E_{Z,k} = G_{Z,k} - \nabla \mathcal{L}^{\mathcal{N}}_Z(\Theta^{\nu}_{Z,k}). \tag{40}$$

The sum counts repeated draws if sampling is with replacement. Assume $\mathbb{E}[E_{Z,k} \mid \mathcal{F}_k] = 0$ and finite second moments. Fresh uniform draws with replacement, or a fresh uniform subset without replacement at each step, satisfy this assumption. Let $\mathcal{H}_k$ additionally contain both sampled batches, and let the measured gradient be $G_{Z,k} + \Xi^{M}_{Z,k}$, where $\mathbb{E}[\Xi^{M}_{Z,k} \mid \mathcal{H}_k] = 0$ and the errors have finite second moments. All couplings must preserve these conditional laws.

Proposition 2 (Minibatch extension of geometric stability)**.** *Under the sampling conditions above, assume the shared initialization, finite moments, trajectory containment, and mean-update bounds of Theorem 1. Define*

$$v^{\mathrm{batch}}_k = \eta_k^2 \mathbb{E}\|E_{D,k} - E_{R,k}\|_2^2, \qquad v^{\mathrm{shot}}_k = \eta_k^2 \mathbb{E}\|\Xi^{M}_{D,k} - \Xi^{M}_{R,k}\|_2^2. \tag{41}$$

For $a_0 = 0$ and

$$a_{k+1} = \sqrt{(c^{\mathcal{N}}_k a_k + \beta^{\mathcal{N}}_k)^2 + v^{\mathrm{batch}}_k + v^{\mathrm{shot}}_k}, \tag{42}$$

we have $W_2(P^{\nu}_D, P^{\nu}_R) \le a_K$ and $W_2((\Psi_0)_{P^{\nu}_D}, (\Psi_0)_{P^{\nu}_R}) \le L_0 a_K$. The factors $c^{\mathcal{N}}_k, \beta^{\mathcal{N}}_k$ concern the dataset-averaged mean maps $T^{\mathcal{N}}_{Z,k}$, as in the original theorem.

Proof. The update is $\Theta^{\nu}_{Z,k+1} = T^{\mathcal{N}}_{Z,k}(\Theta^{\nu}_{Z,k}) - \eta_k(E_{Z,k} + \Xi^{M}_{Z,k})$. Set $E_k = E_{D,k} - E_{R,k}$ and $\Upsilon_k = \Xi^{M}_{D,k} - \Xi^{M}_{R,k}$. The tower property gives $\mathbb{E}[\Upsilon_k \mid \mathcal{F}_k] = 0$. Moreover, E_k is $\mathcal{H}_k$-measurable, so

$$\mathbb{E}[\langle E_k, \Upsilon_k\rangle \mid \mathcal{F}_k] = \mathbb{E}[\langle E_k, \mathbb{E}[\Upsilon_k \mid \mathcal{H}_k]\rangle \mid \mathcal{F}_k] = 0.$$

Thus both errors are centered relative to the mean update and their cross term vanishes. With Γ_k, Δ_k as in the preceding proof,

$$\mathbb{E}\|\Delta_{k+1}\|_2^2 = \mathbb{E}\|\Gamma_k\|_2^2 + v_k^{\mathrm{batch}} + v_k^{\mathrm{shot}}.$$

Eq. 35, the same scalar induction, and the endpoint coupling bounds complete the proof. □

Explicit batch-size dependence. For conditionally independent batch sampling between runs, define

$$\sigma^2_{Z,k} = \frac{1}{n_Z}\sum_{z\in Z}\|g^{\mathcal{N}}_z(\Theta^{\nu}_{Z,k}) - \overline{g}^{\mathcal{N}}_Z(\Theta^{\nu}_{Z,k})\|_2^2.$$

Then

$$v_k^{\mathrm{batch}} = \eta_k^2 \sum_{Z\in\{D,R\}} \kappa_{Z,k}\mathbb{E}\sigma^2_{Z,k}, \qquad \kappa_{Z,k} = \begin{cases} 1/b_{Z,k}, & \text{with replacement,} \\ \dfrac{n_Z - b_{Z,k}}{b_{Z,k}(n_Z-1)}, & \text{without replacement.} \end{cases} \tag{43}$$

For the second case, $1 \le b_{Z,k} \le n_Z$; set $\kappa_{Z,k} = 0$ when $n_Z = 1$. To verify this formula, center each example gradient. Independent draws give variance $\sigma^2_{Z,k}/b_{Z,k}$; for two distinct uniform draws without replacement, the expected inner product of their centered gradients is $-\sigma^2_{Z,k}/(n_Z - 1)$. Expanding the squared norm of the average gives the stated correction. The common regularizer cancels from $E_{Z,k}$. If $\|g^{\mathcal{N}}_z(\theta)\|_2 \le G^{\mathcal{N}}$ uniformly, then $\sigma^2_{Z,k} \le (G^{\mathcal{N}})^2$, giving a deterministic upper envelope by substitution.

Under the independent parameter-shift assumptions of Proposition 3, applied separately to every sampled occurrence and independently between runs,

$$v_k^{\mathrm{shot}} \le \eta_k^2 p\left(\frac{1}{b_{D,k}} + \frac{1}{b_{R,k}}\right)\left(\frac{3}{M} + \frac{1}{2M^2}\right) \qquad (\mathcal{N} = \mathcal{I}). \tag{44}$$

This follows by conditioning on $\mathcal{H}_k$ and averaging the per-example variance bound over each batch. Repeated occurrences must use fresh shot draws for this $1/b_{Z,k}$ reduction.

When $M = \infty$, only v_k^{shot} disappears; the batch term generally remains. Taking every example without replacement recovers the full-batch theorem. The extension therefore preserves the channel-dependent mean-update analysis but adds a separate sampling contribution. It does not assert an ordering of actual endpoint distances. Random reshuffling within an epoch is not automatically conditionally unbiased relative to the full objective, and adaptive optimizers such as Adam are outside this SGD statement.

B.4 Finite-Shot Moments and Shot-Budget Dependence

The role of the coupling. Under Eq. 32, expansion gives

$$v_k^{\mathcal{N},M} = \eta_k^2\big(\mathbb{E}\|\Xi^{M}_{D,k}\|_2^2 + \mathbb{E}\|\Xi^{M}_{R,k}\|_2^2 - 2\mathbb{E}\langle \Xi^{M}_{D,k}, \Xi^{M}_{R,k}\rangle\big).$$

The cross term is zero for conditionally independent draws. Shared random numbers need not give identical errors at different circuit states. Accordingly, this term depends on the admissible coupling: even identical endpoint laws can give a positive contribution under independent draws. It is part of an upper bound, not an intrinsic distance between the laws.

Using the unbiased squared-loss gradient estimator of Sweke et al. (2020), we bound its finite-shot contribution to the coupled training–retraining recursion.

Proposition 3 (Shot-budget bound for independent parameter-shift estimates)**.** *Consider channel-free full-batch training with $\ell(f,y) = (f-y)^2/2$, $|y| \le 1$, a normalized Pauli readout, and p parameter coordinates satisfying*

$$\partial_j f^{\mathcal{I}}_\theta(x) = \frac{f^{\mathcal{I}}_{\theta+(\pi/2)e_j}(x) - f^{\mathcal{I}}_{\theta-(\pi/2)e_j}(x)}{2}. \tag{45}$$

At the current iterate, use the estimator

$$\widehat{g}_{Z,k,j} = \frac{1}{|Z|} \sum_{(x,y)\in Z} (\widehat{f}(x) - y) \frac{\widehat{f}_{j,+}(x) - \widehat{f}_{j,-}(x)}{2} + \partial_j \mathcal{R}(\Theta^{\nu}_{Z,k}). \tag{46}$$

Given the joint pre-update history, each circuit estimate averages $M \geq 1$ independent Pauli measurements. For each example and coordinate, the forward and two shifted estimates use independent batches; draws are also independent across examples and between runs. The regularizer gradient is exact. Independence across coordinates is unnecessary, and one forward estimate may be shared across them. Writing $h(M) = 3/M + 1/(2M^2)$, the errors $\Xi^M_{Z,k} = \widehat{g}_{Z,k} - \nabla \mathcal{L}^{\mathcal{I}}_Z(\Theta^{\nu}_{Z,k})$ are conditionally centered and satisfy

$$\mathbb{E}[\|\Xi^M_{Z,k}\|_2^2 \mid \mathcal{F}_k] \leq \frac{p}{|Z|} h(M), \tag{47}$$

$$v_k^{\mathcal{I},M} \leq \eta_k^2 p \left(\frac{1}{|D|} + \frac{1}{|R|} \right) h(M) =: \overline{v}_k(M). \tag{48}$$

Proof. Condition on $\mathcal{F}_k$ and fix one example and coordinate. Write $d_j = (f_{j,+} - f_{j,-})/2$ and $\widehat{d}_j = (\widehat{f}_{j,+} - \widehat{f}_{j,-})/2$. The Pauli sample means are unbiased, with

$$\sigma_f^2 = \mathrm{Var}(\widehat{f} \mid \mathcal{F}_k) = \frac{1 - f^2}{M} \leq \frac{1}{M}, \qquad \sigma_{d,j}^2 = \mathrm{Var}(\widehat{d}_j \mid \mathcal{F}_k) = \frac{2 - f_{j,+}^2 - f_{j,-}^2}{4M} \leq \frac{1}{2M}.$$

Independence of $\widehat{f}$ and $\widehat{d}_j$ implies $\mathbb{E}[(\widehat{f} - y)\widehat{d}_j \mid \mathcal{F}_k] = (f - y)d_j$. Using $|f - y| \leq 2$ and $|d_j| \leq 1$, their product variance is

$$\mathrm{Var}((\widehat{f} - y)\widehat{d}_j \mid \mathcal{F}_k) = (f - y)^2 \sigma_{d,j}^2 + d_j^2 \sigma_f^2 + \sigma_f^2 \sigma_{d,j}^2 \leq \frac{3}{M} + \frac{1}{2M^2}. \tag{49}$$

Averaging independent examples divides this bound by $|Z|$. Summing coordinate variances gives Eq. 47; coordinate independence is not needed for a squared Euclidean norm. Finally, independence between runs makes the difference-error variances add, proving Eq. 48. □

This construction applies to the estimator in Appendix D.2. Reusing correlated forward and shifted estimates can bias their product. Likewise, an unbiased circuit estimate does not in general give an unbiased nonlinear plug-in loss gradient; such estimators require separate bias control.

Propagation over a fixed horizon. For comparison across shot budgets, keep datasets, initialization, step sizes, and K fixed, and require common valid constants $c_k^{\mathcal{I}}, \beta_k^{\mathcal{I}}$ on a region containing all relevant trajectories almost surely. Define

$$\overline{a}_0(M) = 0, \qquad \overline{a}_{k+1}(M) = \sqrt{(c_k^{\mathcal{I}} \overline{a}_k(M) + \beta_k^{\mathcal{I}})^2 + \overline{v}_k(M)}. \tag{50}$$

Theorem 1 bounds the parameter and clean-prediction distances by $\overline{a}_K(M)$ and $L_0 \overline{a}_K(M)$, respectively. Set $\overline{v}_k(\infty) = 0$, so $\overline{a}_K(\infty) = a_K^{(\mathcal{I},\infty)}$. Since $h(M)$ decreases, monotonicity of the recursion shows that $\overline{a}_K(M)$ is non-increasing in M. Moreover, $\sqrt{x^2 + v} \leq x + \sqrt{v}$ for $x, v \geq 0$ gives, by iteration,

$$0 \leq \overline{a}_K(M) - a_K^{(\mathcal{I},\infty)} \leq \sum_{s=0}^{K-1} \sqrt{\overline{v}_s(M)} \prod_{j=s+1}^{K-1} c_j^{\mathcal{I}} = O(M^{-1/2}). \tag{51}$$

Thus the explicit squared-error envelope is $O(M^{-1})$, and its propagated excess tends to zero at any fixed finite horizon. The actual error moments also depend on the shot-dependent iterate laws, so this argument orders neither those moments nor the actual endpoint distances across budgets. With minibatch sampling fixed, the same reasoning approaches the exact-expectation *SGD* envelope, retaining its batch term, whenever common deterministic envelopes for that term and the mean-update constants are available.

B.5 Geometric Proximity Does Not Imply Certification

Proof of Proposition 1. Prepare $HR_y(\theta)|0\rangle$, where H is a Hadamard gate, apply depolarization immediately before a Pauli-Z measurement, and use the resulting expectation as the prediction. This circuit ignores its input and has $f_\theta^{\mathcal{I}}(x) = \sin\theta$ and $f_\theta^{\mathcal{D}_q}(x) = (1-q)\sin\theta$. Fix $R = \{(0,0)\}$, $F = \{(1,1)\}$, $D = R \cup F$, loss $(f-y)^2/2$, and regularizer $\mathcal{R}(\theta) = 2\theta^2$. With $K = 1$ and $\eta_0 = 1/4$, differentiating Eq. 31 gives

$$T_{R,0}^{\mathcal{D}_q}(\theta) = -\frac{(1-q)^2}{4}\sin\theta\cos\theta, \tag{52}$$

$$T_{D,0}^{\mathcal{D}_q}(\theta) = T_{R,0}^{\mathcal{D}_q}(\theta) + \frac{1-q}{8}\cos\theta. \tag{53}$$

All these choices are independent of q.

On $\mathcal{P} = [-1/4, 1/4]$, the derivative of the retain map is $-(1-q)^2\cos(2\theta)/4$, giving $c_0^{\mathcal{D}_q} = (1-q)^2/4 < 1$. Also $|T_{R,0}^{\mathcal{D}_q}(\theta)| \le 1/16$ and $|T_{D,0}^{\mathcal{D}_q}(\theta)| \le 3/16 < 1/4$, so both maps preserve $\mathcal{P}$. Their difference gives exactly

$$\beta_0^{\mathcal{D}_q} = \sup_{\theta\in\mathcal{P}} \frac{1-q}{8}|\cos\theta| = \frac{1-q}{8}. \tag{54}$$

Use a shared initialization with any smooth probability density compactly supported in $(1/10, 1/5)$, fixed independently of q. The trajectories have finite moments, and $M = \infty$ makes the shot errors zero. The clean audit $\Psi_0(\theta) = \sin\theta$ is 1-Lipschitz. All assumptions of Theorem 1 therefore hold and $a_1^\nu = \beta_0^{\mathcal{D}_q}$.

For θ in the initialization support and $q < 1$, the retain update is negative, whereas

$$T_{D,0}^{\mathcal{D}_q}(\theta) = \frac{1-q}{8}\cos\theta\,[1 - 2(1-q)\sin\theta] > 0, \tag{55}$$

since $\cos\theta > 0$ and $2(1-q)\sin\theta < 2/5$. The positive and negative half-lines consequently separate the two endpoint laws. Each directional hockey-stick divergence equals one for every finite $\epsilon \ge 0$, proving Eq. 17. Meanwhile, the shared coupling gives

$$W_2^2(P_D^\nu, P_R^\nu) \le \frac{(1-q)^2}{64}\mathbb{E}[\cos^2\Theta_0] \le (a_1^\nu)^2, \tag{56}$$

which proves Eq. 16.

This separation also holds for smooth endpoint densities: on the initialization interval the retain derivative is strictly negative, as is $\partial_\theta T_{D,0}^{\mathcal{D}_q} = -(1-q)^2\cos(2\theta)/4 - (1-q)\sin\theta/8$ for every $q < 1$. The change-of-variables formula therefore preserves smoothness. Moreover, $\sin\theta$ preserves signs on $\mathcal{P}$, so the clean prediction laws remain perfectly distinguishable as well. At $q = 1$ both updates equal zero; the sensitivity is maximal only for $q < 1$, even as the geometric bound tends to zero. □

B.6 Fixed Clean Audits and Evaluation-Only Attenuation

Data processing. For any fixed measurable audit Ψ_0 and $\alpha \ge 1$, taking preimages of prediction events yields

$$\mathsf{H}_\alpha((\Psi_0)_P\|(\Psi_0)_Q) \le \mathsf{H}_\alpha(P\|Q). \tag{57}$$

Indeed, each event difference on the left is $P(\Psi_0^{-1}(B)) - \alpha Q(\Psi_0^{-1}(B))$ and is bounded by the supremum over all model events. Applying this in both directions to $P_{D,F}^\nu, P_R^\nu$ transfers the model-level guarantee to the audit. The converse fails for non-injective audits: a constant map merges even distinct point masses whose model-level sensitivity is one.

Exact evaluation-only attenuation. Fix the trained-model laws P, Q and a clean audit with finite second moments under both laws. For exact effective depolarizing readout of traceless observables, $\Psi_q = (1-q)\Psi_0$. Scaling both vectors in any prediction coupling, and using inverse scaling for $q < 1$, gives

$$W_2((\Psi_q)_P, (\Psi_q)_Q) = (1-q)W_2((\Psi_0)_P, (\Psi_0)_Q). \tag{58}$$

For $q < 1$, scaling is a measurable bijection with a measurable inverse, so data processing in both directions also gives

$$\mathsf{H}_\alpha((\Psi_q)_P\|(\Psi_q)_Q) = \mathsf{H}_\alpha((\Psi_0)_P\|(\Psi_0)_Q), \qquad \alpha \geq 1. \tag{59}$$

At $q = 1$, both noisy prediction laws are point masses at zero, so their Wasserstein and hockey-stick discrepancies vanish. These statements concern exact predictions, not finite-shot outcomes. Thus evaluation alone can shrink a prediction-space distance without changing the model laws or their prediction-level distinguishability for $q < 1$. A common clean audit removes this scaling effect when comparing training regimes. It does not turn geometric proximity into a certificate or a guarantee of lower unlearning cost; the retraining reference remains the noise-matched law P_R^ν.

C Channel Noise Analyses

This appendix derives Table 1 and gives supporting conditions and channel identities. We derive the deletion perturbation and the conditional contractive bound, give sufficient conditions for contraction, and then detail the channel identities and amplitude-damping example. All channels in this appendix act on the final state before a fixed readout and are independent of the parameters and dataset.

C.1 Readout Gradients and Deletion Perturbations

Affine readout and data-gradient bound. Write $f_\theta^{\mathcal{I}}(x) = \mathrm{Tr}[O\rho_\theta(x)]$. If the adjoint channel satisfies

$$\mathcal{N}^*(O) = s_{\mathcal{N}} O + b_{\mathcal{N}} I, \qquad f_\theta^{\mathcal{N}} = s_{\mathcal{N}} f_\theta^{\mathcal{I}} + b_{\mathcal{N}}, \tag{60}$$

then $\nabla f_\theta^{\mathcal{N}} = s_{\mathcal{N}} \nabla f_\theta^{\mathcal{I}}$. Here $s_{\mathcal{N}}, b_{\mathcal{N}}$ are scalar readout coefficients, not optimizer propagation factors or regularization parameters. For squared loss, suppose uniformly on the parameter region and training inputs that

$$|f_\theta^{\mathcal{I}}(x)| \leq \bar{f}, \quad |y| \leq \bar{y}, \quad \|\nabla f_\theta^{\mathcal{I}}(x)\|_2 \leq L_f.$$

The chain rule gives the following valid, possibly loose, bound:

$$\|\nabla_\theta \ell(f_\theta^{\mathcal{N}}(x), y)\|_2 \leq G^{\mathcal{N}} := |s_{\mathcal{N}}|\big(|s_{\mathcal{N}}|\bar{f} + |b_{\mathcal{N}}| + \bar{y}\big)L_f. \tag{61}$$

The main-text table uses the natural normalized-Pauli bounds $\bar{f} = \bar{y} = 1$. The common regularizer is excluded from $G^{\mathcal{N}}$.

Deletion perturbation. Let $\alpha_F = |F|/|D|$ and let $\overline{g}_Z^{\mathcal{N}}$ denote the average data gradient on Z. For $F \neq \varnothing$, the common regularizer in Eq. 31 cancels, giving

$$\nabla\mathcal{L}_D^{\mathcal{N}} - \nabla\mathcal{L}_R^{\mathcal{N}} = \alpha_F\big(\overline{g}_F^{\mathcal{N}} - \overline{g}_R^{\mathcal{N}}\big). \tag{62}$$

Taking norms, multiplying by the step size, and taking the supremum over the parameter region yields

$$\beta_k^{\mathcal{N}} \leq 2\eta_k\alpha_F G^{\mathcal{N}}. \tag{63}$$

For $F = \varnothing$, both sides vanish. In particular, depolarization with a traceless readout and the ideal channel give

$$\beta_k^{\mathcal{D}_q} \leq 2\eta_k\alpha_F(1-q)\big((1-q)\bar{f} + \bar{y}\big)L_f, \tag{64}$$

$$\beta_k^{\mathcal{I}} \leq 2\eta_k\alpha_F(\bar{f} + \bar{y})L_f. \tag{65}$$

These inequalities compare uniform bounds; they do not order the actual gradient differences at individual parameters.

C.2 Contractive Finite-Horizon Bounds

Proof of Corollary 1. Theorem 1 itself requires no contraction. Under the additional assumption $c_k^{\mathcal{N}} \leq c < 1$, Eq. 11 and the perturbation bound in Eq. 63 give

$$\begin{aligned} a_K^{(\mathcal{N},\infty)} &\leq 2\eta\alpha_F G^{\mathcal{N}} \sum_{t=0}^{K-1} c^t \\ &= 2\eta\alpha_F G^{\mathcal{N}} \frac{1-c^K}{1-c} \leq \frac{2\eta\alpha_F G^{\mathcal{N}}}{1-c}. \end{aligned} \tag{66}$$

For the Pauli-Z readout, the adjoint identities in Appendix C.4 give

$\mathcal{N}$	$\mathcal{I}$	$\mathcal{D}_q$	$\mathcal{A}_\gamma$	$\mathcal{P}_{\gamma_\phi}$	$\mathcal{B}_p$
$s_{\mathcal{N}}$	1	$1-q$	$1-\gamma$	1	$1-2p$
$b_{\mathcal{N}}$	0	0	γ	0	0

Substituting these coefficients and $\bar{f} = \bar{y} = 1$ into Eq. 61 yields

$$\begin{aligned} G^{\mathcal{I}} = G^{\mathcal{P}_{\gamma_\phi}} &= 2L_f, & G^{\mathcal{D}_q} &= (1-q)(2-q)L_f, \\ G^{\mathcal{A}_\gamma} &= 2(1-\gamma)L_f, & G^{\mathcal{B}_p} &= |1-2p|\big(|1-2p|+1\big)L_f. \end{aligned} \tag{67}$$

Eq. 66 now gives every row of Table 1. With the same c, L_f, η, K and dataset pair, only these nonnegative channel factors vary: the depolarizing and amplitude-damping envelopes are nonincreasing, phase damping leaves the envelope unchanged, and the bit flip envelope decreases up to $p = 1/2$ and increases thereafter. □

Channel-dependent contraction and nonexpansion. A common c is a shared upper bound on propagation, not necessarily the smallest factor for any channel. If instead each regime has its own $c_k^{\mathcal{N}} \le c^{\mathcal{N}} < 1$, then

$$a_K^{(\mathcal{N},\infty)} \le 2\eta\alpha_F G^{\mathcal{N}} \frac{1-(c^{\mathcal{N}})^K}{1-c^{\mathcal{N}}}. \tag{68}$$

A smaller gradient bound alone need not reduce this expression when the propagation factor also changes. For $c = 1$, the geometric sum equals K and the corresponding nonexpansive bound is $2\eta\alpha_F G^{\mathcal{N}} K$. Neither comparison orders the actual Wasserstein distances.

Finite-shot extension. Retain all coupling, moment, conditional-unbiasedness, and trajectory assumptions of Theorem 1 at the finite shot budget. If $c_k^{\mathcal{N}} \le c < 1$, $\beta_k^{\mathcal{N}} \le \bar{\beta}^{\mathcal{N}}$, and $v_k^{\mathcal{N},M} \le \bar{v}^{\mathcal{N},M}$ for every step, then

$$a_K^{\nu} \le \bar{\beta}^{\mathcal{N}} \frac{1-c^K}{1-c} + \sqrt{\bar{v}^{\mathcal{N},M} \frac{1-c^{2K}}{1-c^2}}. \tag{69}$$

To verify this, let $u_0 = z_0 = 0$, $u_{k+1} = cu_k + \bar{\beta}^{\mathcal{N}}$, and $z_{k+1} = \sqrt{c^2 z_k^2 + \bar{v}^{\mathcal{N},M}}$. If $a_k^{\nu} \le u_k + z_k$, the theorem's recursion and the triangle inequality in $\mathbb{R}^2$ give

$$a_{k+1}^{\nu} \le \left\| (cu_k + \bar{\beta}^{\mathcal{N}}, 0) + (cz_k, \sqrt{\bar{v}^{\mathcal{N},M}}) \right\|_2 \le u_{k+1} + z_{k+1}.$$

Induction and the two geometric sums prove Eq. 69. One may use $\bar{\beta}^{\mathcal{N}} = 2\eta\alpha_F G^{\mathcal{N}}$. For the channel-free estimator of Proposition 3, $\bar{v}^{\mathcal{I},M} = \eta^2 p(1/|D| + 1/|R|)(3/M + 1/(2M^2))$ gives an explicit shot contribution. The comparison across budgets requires common valid contraction and perturbation bounds; it does not order the actual distances.

C.3 Sufficient Conditions for Common Contraction

Uniform positive curvature. Suppose the retain objectives are twice continuously differentiable on a common convex region and, for every compared channel setting,

$$\mu I \preceq \nabla^2 \mathcal{L}_R^{\mathcal{N}}(\theta) \preceq LI, \qquad 0 < \mu \le L < \infty. \tag{70}$$

For $0 < \eta \le 1/L$, the update Jacobian satisfies

$$\|I - \eta\nabla^2 \mathcal{L}_R^{\mathcal{N}}(\theta)\|_2 \le 1 - \eta\mu < 1.$$

Integrating the Jacobian along the segment between two parameters shows that the common contraction condition holds with $c = 1 - \eta\mu < 1$. Consequently,

$$a_K^{(\mathcal{N},\infty)} \le \frac{2\alpha_F G^{\mathcal{N}}}{\mu}\big[1-(1-\eta\mu)^K\big] \le \frac{2\alpha_F G^{\mathcal{N}}}{\mu}. \tag{71}$$

These curvature bounds are sufficient conditions for the corollary, not requirements of Theorem 1.

Quadratic regularization and readout coefficients. For a concrete sufficient condition, take $\mathcal{R}(\theta) = \lambda\|\theta\|_2^2/2$, $\lambda \geq 0$, and additionally assume $\|\nabla^2 f_\theta^{\mathcal{I}}(x)\|_2 \leq B_f$ uniformly on the region. Writing $f = f_\theta^{\mathcal{I}}(x)$, the data-loss Hessian is exactly

$$\nabla_\theta^2 \ell(s_{\mathcal{N}} f + b_{\mathcal{N}}, y) = s_{\mathcal{N}}^2 \nabla f \nabla f^\top + s_{\mathcal{N}}(s_{\mathcal{N}} f + b_{\mathcal{N}} - y)\nabla^2 f. \tag{72}$$

The first term is positive semidefinite. Define the auxiliary residual factor $r_{\mathcal{N}} = |s_{\mathcal{N}}|(|s_{\mathcal{N}}|\bar{f} + |b_{\mathcal{N}}| + \bar{y})$. Averaging and adding the regularizer gives

$$(\lambda - r_{\mathcal{N}} B_f) I \preceq \nabla^2 \mathcal{L}_R^{\mathcal{N}}(\theta) \preceq H^{\mathcal{N}} I, \qquad H^{\mathcal{N}} := \lambda + s_{\mathcal{N}}^2 L_f^2 + r_{\mathcal{N}} B_f. \tag{73}$$

If $\lambda - r_{\mathcal{N}} B_f \geq \mu > 0$ and $H^{\mathcal{N}} \leq L$ hold for all compared settings, Eq. 70 follows. For $\bar{f} = \bar{y} = 1$, the respective lower bounds for depolarization and amplitude damping are

$$\nabla^2 \mathcal{L}_R^{\mathcal{D}_q} \succeq [\lambda - (1-q)(2-q)B_f] I, \qquad \nabla^2 \mathcal{L}_R^{\mathcal{A}_\gamma} \succeq [\lambda - 2(1-\gamma)B_f] I. \tag{74}$$

A sufficient common choice for all four channels and the noiseless setting is $\lambda > 2B_f$, $\mu = \lambda - 2B_f$, and $L = \lambda + L_f^2 + 2B_f$. This can require strong regularization and is not asserted for the experimental hyperparameters.

Region and experimental scope. The hypotheses require both trajectories to remain in the stated region almost surely, not only the retained trajectory or its endpoint. Local positive curvature by itself does not establish this containment. In particular, no bounded region contains a nondegenerate Gaussian initialization almost surely. Moreover, contraction observed along finitely many sampled trajectories does not establish the uniform Lipschitz condition required by Corollary 1. Contraction of a physical quantum channel also does not by itself imply contraction of the parameter-space training update. The experiments therefore test actual endpoint behavior rather than certify the contraction assumption of Corollary 1.

C.4 Channel Identities and Other Readouts

Depolarization. For $\mathcal{D}_q(\rho) = (1-q)\rho + qI/d$, the adjoint satisfies

$$\mathcal{D}_q^*(O) = (1-q)O + q\,\mathrm{Tr}[O]I/d. \tag{75}$$

A traceless observable therefore has $s_{\mathcal{D}_q} = 1 - q$, $b_{\mathcal{D}_q} = 0$. The same identity holds for a Pauli observable on a locally depolarized measured qubit, with identity on the other qubits. At $q = 1$, this readout and its data gradients vanish.

Amplitude damping. We use the relaxation probability $\gamma \in [0, 1]$ throughout, with Kraus operators

$$K_0 = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-\gamma} \end{pmatrix}, \qquad K_1 = \begin{pmatrix} 0 & \sqrt{\gamma} \\ 0 & 0 \end{pmatrix}.$$

Computing $\mathcal{A}_\gamma^*(O) = \sum_i K_i^\dagger O K_i$ gives

$$\mathcal{A}_\gamma^*(X) = \sqrt{1-\gamma}\,X, \quad \mathcal{A}_\gamma^*(Y) = \sqrt{1-\gamma}\,Y, \quad \mathcal{A}_\gamma^*(Z) = (1-\gamma)Z + \gamma I. \tag{76}$$

The Z readout has $s_{\mathcal{A}_\gamma} = 1 - \gamma$, $b_{\mathcal{A}_\gamma} = \gamma$, whereas an X or Y readout has $s_{\mathcal{A}_\gamma} = \sqrt{1-\gamma}$ and zero offset. Their data-gradient bounds follow from Eq. 61; under a valid contraction assumption, Eq. 66 gives the endpoint bound for the chosen observable. At $\gamma = 1$, the Z prediction equals $+1$ and its data gradient vanishes.

Phase damping. Use the same local-channel convention as the experiments: the damping parameter $\gamma_\phi \in [0, 1]$ multiplies off-diagonal density-matrix entries by $\sqrt{1-\gamma_\phi}$ and preserves the diagonal entries. One Kraus representation is

$$K_0 = \mathrm{diag}(1, \sqrt{1-\gamma_\phi}), \qquad K_1 = \mathrm{diag}(0, \sqrt{\gamma_\phi}).$$

Consequently,

$$\mathcal{P}_{\gamma_\phi}^*(Z) = Z, \qquad \mathcal{P}_{\gamma_\phi}^*(X) = \sqrt{1-\gamma_\phi}\,X, \qquad \mathcal{P}_{\gamma_\phi}^*(Y) = \sqrt{1-\gamma_\phi}\,Y. \tag{77}$$

The Z readout, its gradients, and exact-expectation training trajectories coincide with the ideal ones under shared initialization. For a transverse observable $O = aX + bY$, the affine relation instead has $s_{\mathcal{P}_{\gamma_\phi}} = \sqrt{1-\gamma_\phi}$ and zero offset. For mixed transverse and longitudinal components, only the former are attenuated, so a single scalar attenuation need not describe the whole readout. If phase damping is alternatively parameterized by the coherence multiplier $1 - q_\phi$, the conversion is

$$q_\phi = 1 - \sqrt{1-\gamma_\phi}, \qquad \gamma_\phi = 2q_\phi - q_\phi^2. \tag{78}$$

Table 1 and the phase-damping axes in the experimental figures abbreviate γ_ϕ as γ; this is distinct from the amplitude-damping relaxation probability.

Bit flip. For $\mathcal{B}_p(\rho) = (1-p)\rho + pX\rho X$, the adjoint is

$$\mathcal{B}_p^*(O) = (1-p)O + pXOX. \tag{79}$$

For $O = Z$ or Y, $XOX = -O$, giving $s_{\mathcal{B}_p} = 1 - 2p$, $b_{\mathcal{B}_p} = 0$. The absolute value in the gradient bound is necessary for $p > 1/2$. For $p \le 1/2$, setting $q = 2p$ gives exactly the same predictions and squared-loss gradients as depolarization for this readout. At $p = 1/2$ the affected readout is constant; at $p = 1$ it is sign-reversed without attenuation. An X readout is unchanged for all p.

Scope. These identities describe effective readout channels and do not establish the same scalar relations for channels interleaved with trainable layers. At noise-strength endpoints where the data gradients vanish identically, exact-expectation full-data training and retraining follow the same regularizer-only updates from a shared initialization. Evaluation-only attenuation for fixed model distributions is treated separately in Appendix B.6.

D Experiment 1: Settings and Evaluation Protocol

D.1 Data, Model, and Training

We assign labels -1 and $+1$ to versicolor and virginica, respectively, and use five stratified 60/20/20 train/validation/test partitions. Within each partition, three training examples per class are selected uniformly without replacement for deletion, leaving 54 retained examples. Normal deletion uses the original labels; corrupted-label deletion flips only the deletion-set labels before full-data training, following the mislabeled-data setting of Chen & Zhang (2026). Both scenarios use the same retained set and the original test labels. The partitions may share examples and are repeated partitions of one dataset. Each input feature is mapped to an angle by $(\pi/2)\tanh((x_j - \mu_j)/\sigma_j)$, with μ_j and σ_j fitted on the full training set. This transformation is fixed during retraining, so deletion is scoped to QNN training conditional on the preprocessing.

The circuit starts in $|0\rangle^{\otimes 4}$ and encodes the four input angles with R_y gates. Each of its two trainable layers applies R_y and then R_z on every qubit, followed by CNOTs in the order $0 \to 1$, $1 \to 2$, $2 \to 3$, and $3 \to 0$. The output is the expectation of Z_0. For $Z \in \{D, R\}$, training minimizes

$$\mathcal{L}_Z^{\mathcal{N}}(\theta) = \frac{1}{2|Z|} \sum_{(x,y)\in Z} \left(f_\theta^{\mathcal{N}}(x) - y\right)^2 + \frac{\lambda}{2}\|\theta\|_2^2, \qquad \lambda = 0.01. \tag{80}$$

All 16 parameters are initialized independently from $\mathcal{N}(0, 0.3^2)$ and trained for 120 full-batch gradient-descent steps with learning rate 0.15. Each ensemble contains 1,024 distinct initial parameter vectors. For a given model index, the same vector is used for full training and retraining across noise settings and deletion scenarios. Thus, each scenario and noise setting involves 10,240 full-data and retrained models across five partitions; retained-data ensembles are reused between scenarios.

D.2 Noise Settings and Finite-Shot Training

Channels and placements. Depolarization is parameterized as $\mathcal{D}_q(\rho) = (1-q)\rho + qI/2$. Amplitude damping uses relaxation probability γ. Phase damping multiplies the off-diagonal density-matrix entries by $\sqrt{1-\gamma}$, with the parameter conversion in Eq. 78; bit flip applies X with probability p. Readout noise acts on qubit 0 after the final layer. Interleaved noise acts locally on all four

qubits after each complete layer, including the last, with the same strength at each application. The placement and strength are identical during full training and retraining. Equal local parameters do not imply equal total noise across placements. Channel sweeps use exact expectations throughout training.

Noise grids. The readout grid for q and γ is $\{0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8\}$; the bit flip grid is half of these values. The interleaved grids share

$$G = \{0, 0.0001, 0.0005, 0.001, 0.0025, 0.005, 0.01, 0.02\}$$

and add the following channel-specific values:

$$\begin{aligned} \text{Depolarization:} &\quad \{0.04, 0.08, 0.12, 0.16, 0.20, 0.24, 0.32\}, \\ \text{Amplitude damping:} &\quad \{0.04, 0.08, 0.12, 0.16, 0.20, 0.22\}, \\ \text{Phase damping:} &\quad \{0.04, 0.08, 0.16, 0.32, 0.48, 0.64, 0.80, 0.90, 0.96, 0.99\}, \\ \text{Bit flip:} &\quad \{0.03, 0.04, 0.06, 0.08, 0.10, 0.12, 0.16, 0.20\}. \end{aligned}$$

The interleaved ranges were extended after initial validation pilots with separate initializations to include visible accuracy degradation.

Finite-shot gradient estimator. The channel-free shot sweep uses $M \in \{32, 128, 512, 2048\}$ measurements per circuit and otherwise keeps the training protocol fixed. For each training input and update, independent batches of Pauli measurements estimate the forward output $\widehat{f}$ and the shifted outputs $\widehat{f}_{j,+}$ and $\widehat{f}_{j,-}$ at $\theta \pm (\pi/2)e_j$. The loss-gradient estimate is

$$\widehat{g}_j = \frac{1}{|Z|} \sum_{(x,y)\in Z} (\widehat{f}(x) - y) \frac{\widehat{f}_{j,+}(x) - \widehat{f}_{j,-}(x)}{2} + \lambda\theta_j. \tag{81}$$

Independence of the forward and shifted estimates makes this estimator conditionally unbiased; the regularizer gradient is exact. All 16 coordinates are sampled, and full training and retraining receive independent measurement draws even when initialized identically. The exact-expectation baseline uses the same initializations without measurement sampling. Evaluation uses either exact predictions or independent finite-shot predictions, as specified below.

D.3 Evaluation Protocol

Empirical distances and prediction audits. For each partition and noise setting, the parameter-space distance is

$$\mathcal{W}_\nu = \left[\min_{\pi\in\mathfrak{S}_n} \frac{1}{n} \sum_{i=1}^{n} \|\theta^\nu_{D,i} - \theta^\nu_{R,\pi(i)}\|_2^2 \right]^{1/2}, \qquad n = 1{,}024, \tag{82}$$

where $\mathfrak{S}_n$ is the set of permutations. We solve the full assignment problem with squared Euclidean costs and take the square root of the minimum average cost. Shared initialization does not fix the matching. For prediction-space distances, each model is represented by its prediction vector on the partition's 20 test inputs, divided by $\sqrt{20}$. The clean audit Ψ_0 uses exact, channel-free predictions; the noise-matched audit Ψ_ν uses the training channel and placement, or the same shot budget for channel-free finite-shot training. The optimal-transport objective is defined separately in each space. For the affine readout maps, noise-matched distances can equivalently be obtained by exact scaling of the clean prediction distance; interleaved predictions require a separate assignment. As a robustness check, we recomputed the empirical parameter-space W_2 distances after excluding parameters that do not affect the circuit output and observed negligible changes in the tested settings.

We use $\mathcal{W}_\nu$ for the empirical distance in each respective space and divide it by the corresponding noiseless-training reference $\mathcal{W}_0$. Both prediction audits use the clean prediction distance after noiseless training as their denominator. For the shot sweep, the analogous notation is $\mathcal{W}_M/\mathcal{W}_\infty$. Ratios are computed within each partition and then averaged over the five partitions; shaded bands show their sample standard deviation. Each deletion scenario uses its own reference, so normalized curves compare changes relative to that scenario's baseline.

Finite-shot inference. For an exact output f, inference samples $K \sim \mathrm{Binomial}(M, (1+f)/2)$ and returns $\widehat{f} = 2K/M - 1$. Draws are independent across models, inputs, and full/retrained ensembles, and independent of training. For each fixed ensemble pair, we repeat inference 32 times, recompute prediction-space optimal matching for each repetition, and average the resulting empirical distances before computing partition ratios. Each repetition uses one 20-dimensional vector per model; the repetitions are not pooled into a larger ensemble. The plotted bands describe variation across five partition means, not Monte Carlo error. As a control for separation caused by measurement sampling alone, we also evaluate the same retained-model ensemble on both sides with independent inference draws.

Predictive accuracy. All accuracies use the original test labels, predicting $+1$ for a nonnegative output and -1 otherwise. We average accuracy over models within each partition and then report the mean and sample standard deviation across partitions. Channel-noise accuracy uses exact expectations under either Ψ_0 or Ψ_ν. For finite-shot inference, expected accuracy is computed directly from the binomial probabilities, including the $+1$ tie rule, rather than from a single sampled evaluation. The dotted vertical lines in the channel-distance figures interpolate crossings of 80% by the smaller of the full-model and retrained-model mean noise-matched accuracies. They provide a utility reference and do not exclude any plotted points.

E Experiment 1: Deletion Scenarios and Accuracy

This appendix compares normal and corrupted-label deletion under the same training and evaluation protocols, and reports predictive accuracy for both scenarios.

E.1 Comparison with Corrupted-Label Deletion

Corrupted-label deletion removes examples whose labels were flipped before full-data training. Because the retained set, initializations, and noise settings are shared with normal deletion, the comparison isolates the effect of changing the labels of the deleted examples. Figures 6–8 show both scenarios, using the parameter distance and the clean and noise-matched prediction audits defined in Appendix D.3. Each curve is normalized by its own scenario-specific noiseless reference; a lower ratio therefore indicates a larger relative reduction, not necessarily a smaller absolute distance than in the other scenario.

Readout noise. Figure 6 shows similar depolarizing reductions in parameter space: at $q = 0.20$, the distance decreases by 14.58% for normal deletion and 13.85% for corrupted-label deletion. The clean prediction reductions differ more, at 0.65% and 5.23%, respectively. Amplitude damping is more sensitive to the deletion scenario: at $\gamma = 0.20$, the parameter distance increases by 13.48% under normal deletion but stays within 1% of its baseline under corrupted-label deletion. In both cases, the noise-matched prediction distance is lower than the clean prediction distance because the affine readout attenuates prediction differences. Phase damping leaves the Z_0 output unchanged, and bit flip reproduces the depolarizing curves under $q = 2p$ in both scenarios.

Interleaved noise. Figure 7 applies local channels to all four qubits after each of the two layers. Depolarization reduces parameter and prediction distances in both scenarios. For amplitude damping at $\gamma = 0.04$, the parameter distance decreases by 9.87% under normal deletion and 3.90% under corrupted-label deletion, while both clean prediction ratios remain close to one. Phase damping also reduces discrepancy, unlike in the readout setting, but the reductions are weaker and more variable under corrupted-label deletion. Thus, changing the deleted labels can alter the size of a noise-induced reduction, and parameter-space improvements need not transfer proportionally to clean prediction space.

Finite-shot measurement. Figure 8 shows no comparable parameter-distance reduction from finite measurement budgets in either scenario. At $M = 32$, the parameter distance is 16.06% above its exact-expectation reference for normal deletion and 5.53% above it for corrupted-label deletion; these deviations shrink as M increases. Clean prediction ratios remain much closer to one, whereas noise-matched prediction ratios reach 3.89 and 2.27, respectively, at $M = 32$. The latter include

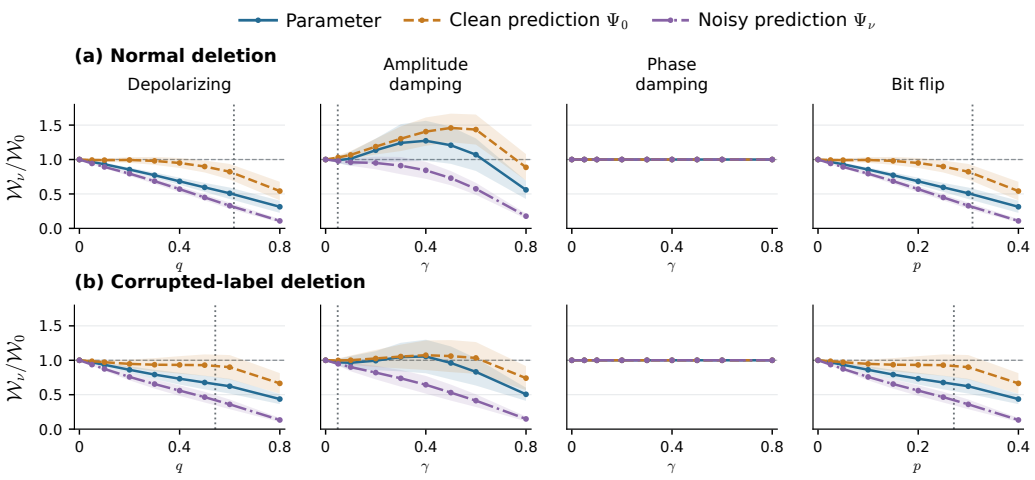

Figure 6: Readout-noise discrepancy for (a) normal and (b) corrupted-label deletion. Curves show parameter-space, clean prediction (Ψ_0), and noise-matched prediction (Ψ_ν) distances, each normalized by its corresponding noiseless reference. Lines and bands show the mean and one sample standard deviation across five partitions. Dotted vertical lines mark 80% crossings of the lower full/retrain mean noise-matched accuracy.

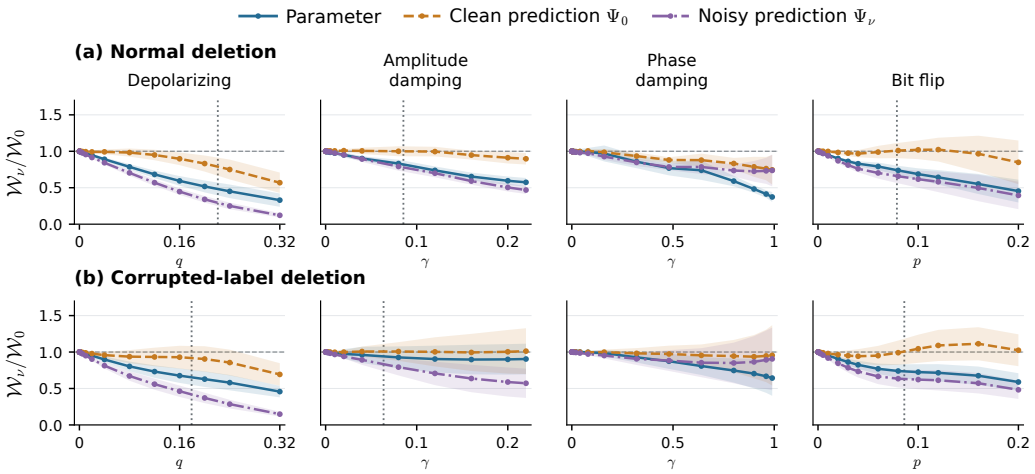

Figure 7: Interleaved-noise discrepancy for (a) normal and (b) corrupted-label deletion. Channels act on all four qubits after each complete layer. Curves, normalization, bands, and accuracy reference lines follow Figure 6; each column uses its channel-specific strength range.

evaluation sampling: the independent-inference control using the same retained ensemble on both sides accounts for approximately 98% and 91% of the corresponding empirical distances at this budget. Averaging inference repetitions reduces Monte Carlo variability but does not remove finite-sample transport discrepancy. These curves therefore describe empirical finite-shot outputs, rather than the exact-prediction distributions in the theorem's fixed-audit transfer result.

E.2 Predictive Accuracy under Clean and Noise-Matched Evaluation

Channel noise. Figures 9–12 report accuracy for both deletion scenarios under clean and noise-matched evaluation. Removing channels at evaluation tests the learned predictor without their direct output distortion, whereas noise-matched evaluation retains the training channels. For readout depolarization and bit flip over the displayed ranges, the output is multiplied by a positive factor, preserving its sign; clean and noisy classification accuracies therefore agree even when prediction distances differ. Readout phase damping leaves predictions unchanged. Amplitude damping adds a

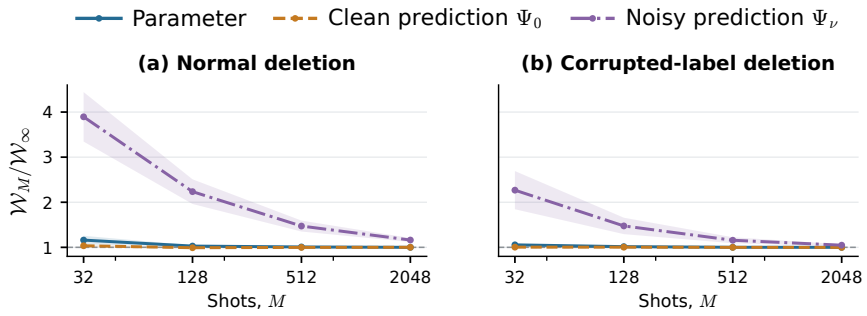

Figure 8: Finite-shot discrepancy for (a) normal and (b) corrupted-label deletion. Moving right increases the training and noise-matched inference budget M. The three curves use their corresponding exact-training references, with exact clean predictions as the denominator for both prediction audits. Noise-matched prediction distances average 32 independent inference repetitions per partition; bands show one sample standard deviation across five partition means.

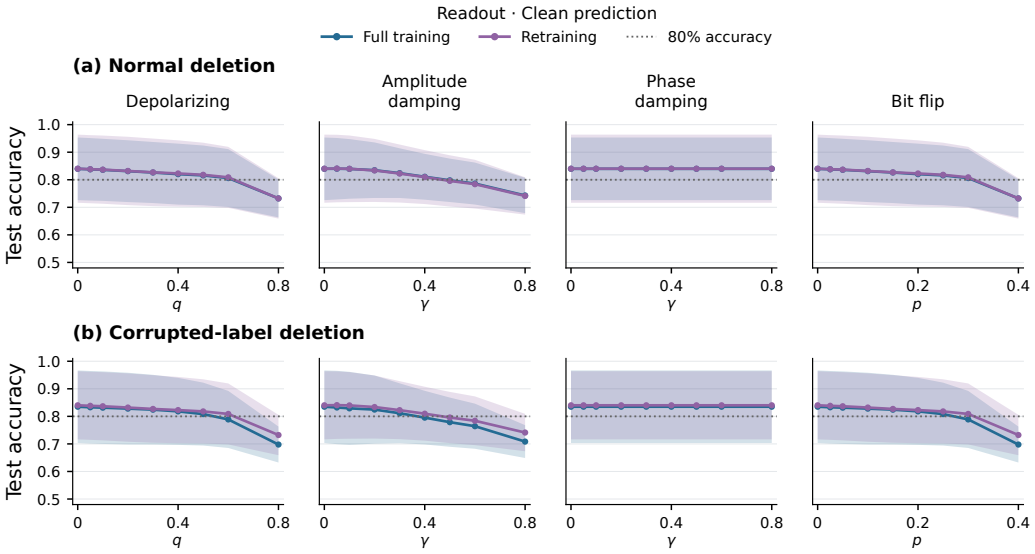

Figure 9: Clean accuracy after readout-noise training. Channels are removed at evaluation. Blue and purple curves show full training and retraining, respectively; rows distinguish normal and corrupted-label deletion. Bands show one sample standard deviation over five partitions and dotted horizontal lines mark 80% accuracy.

positive offset and can change predicted labels, explaining the larger accuracy decline under noise-matched evaluation. For interleaved noise, earlier channels also change the state processed by the next layer, so the two evaluations need not preserve the same decisions. The accuracy curves show where a discrepancy reduction is accompanied by loss of predictive utility.

Finite-shot measurement. Figure 13 combines clean and finite-shot accuracy for both deletion scenarios. With exact inference, full-model accuracy stays near 84.0% and 83.5% across training shot budgets. With M shots also used at inference, expected full-model accuracy is 79.31% and 78.61% at $M = 32$, increasing to 83.64% and 83.21% at $M = 2048$. Stable clean accuracy therefore does not imply insensitivity to a limited inference measurement budget. Finite-shot accuracy is averaged analytically over measurement outcomes, as described in Appendix D.3; bands still represent variation across the five data partitions.

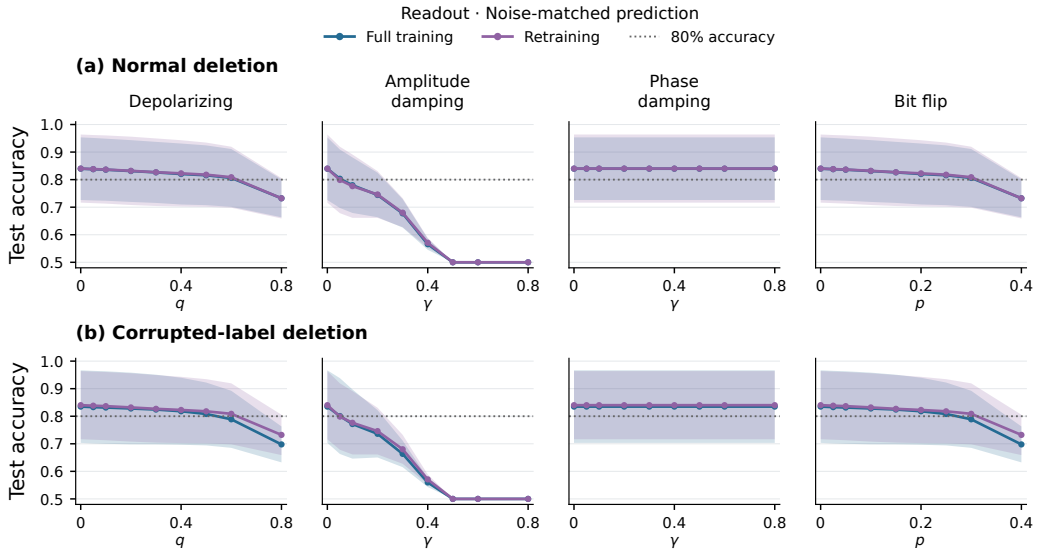

Figure 10: Noise-matched accuracy after readout-noise training, corresponding to Figure 6. The models in Figure 9 are evaluated with their training channel and strength. Colors, rows, bands, and the 80% reference follow the clean-accuracy figure.

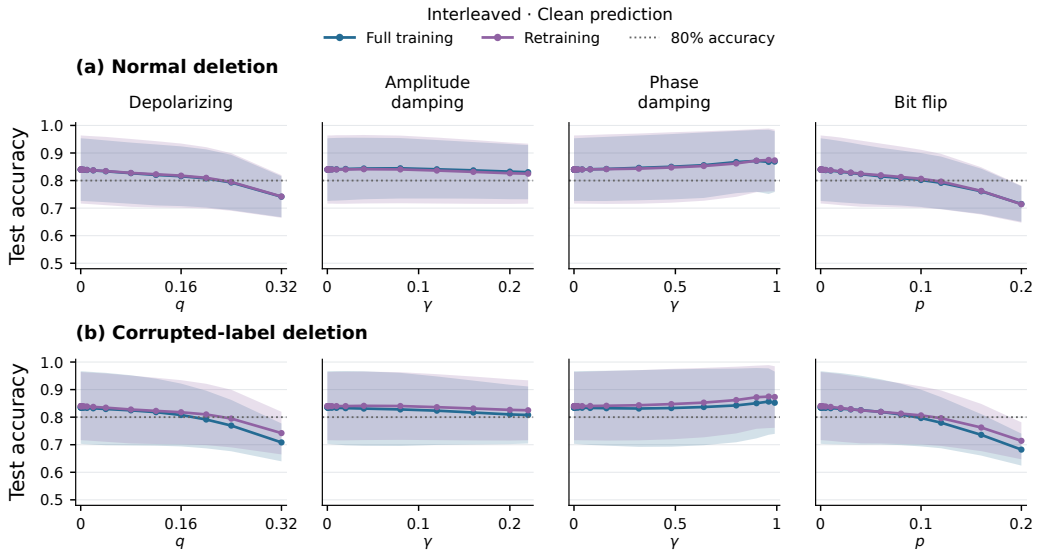

Figure 11: Clean accuracy after interleaved-noise training. Channels act after both layers during training and are removed at evaluation. Plot conventions follow Figure 9; each column uses its channel-specific strength range.

E.3 Trajectory-Level Retain-Update Contraction Diagnostics

Corollary 1 studies a regime in which the retain-data update map is uniformly contractive on a parameter region. The experiments do not establish this uniform condition. Instead, we examine how often the same retain-data update contracts the parameter differences encountered along the sampled training trajectories. This provides a trajectory-level diagnostic of the propagation behavior represented by $c_k^{\mathcal{N}}$ in Eq. 5.

For each full-training/retraining pair i sharing the same initialization, let $\theta_{D,i,k}$ and $\theta_{R,i,k}$ denote their parameters at step k. We apply the same retain-data update map $T_{R,k}^{\mathcal{N}}$ to both parameters and

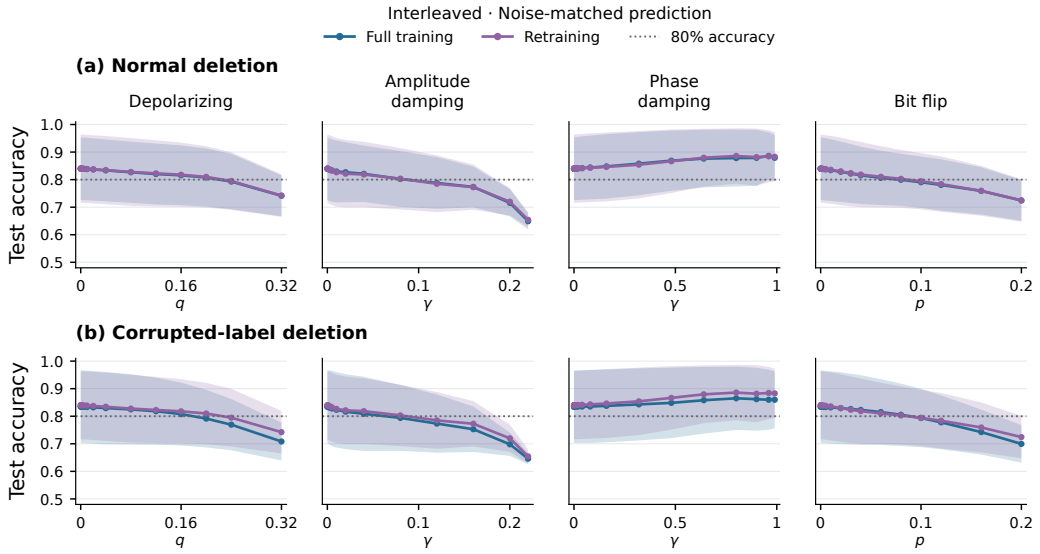

Figure 12: Noise-matched accuracy after interleaved-noise training, corresponding to Figure 7. The models in Figure 11 are evaluated with local channels on all four qubits after each complete layer. Colors, rows, bands, and the 80% reference follow the clean-accuracy figure.

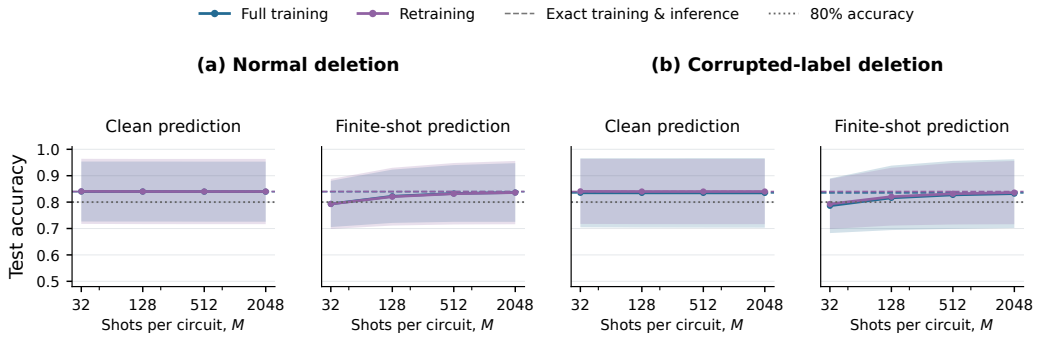

Figure 13: Accuracy after finite-shot training under clean and noise-matched inference. The four panels form one row: clean and finite-shot evaluation for normal deletion on the left, followed by the same evaluations for corrupted-label deletion on the right. Blue and purple curves show full training and retraining. Dashed lines mark the corresponding accuracies after exact training and exact inference, and dotted lines mark 80% accuracy. Bands show one sample standard deviation across five partitions.

define the pairwise propagation ratio

$$r_{i,k}^{\mathcal{N}} := \frac{\left\| T_{R,k}^{\mathcal{N}}(\theta_{D,i,k}) - T_{R,k}^{\mathcal{N}}(\theta_{R,i,k}) \right\|_2}{\|\theta_{D,i,k} - \theta_{R,i,k}\|_2}. \tag{83}$$

Thus, $r_{i,k}^{\mathcal{N}} < 1$ indicates contraction along the particular parameter pair encountered during training. Importantly, this differs from comparing the actual next full-training and retraining iterates: both arguments in Eq. 83 are propagated using the same retain-data map, thereby isolating the propagation effect associated with $c_k^{\mathcal{N}}$ from the deletion-induced update difference $\beta_k^{\mathcal{N}}$.

We additionally measure contraction at the ensemble RMS level. For the $n = 1{,}024$ paired trajectories within each data partition, define

$$\widehat{c}_k^{\mathcal{N}} := \sqrt{\frac{\sum_{i=1}^{n} \left\| T_{R,k}^{\mathcal{N}}(\theta_{D,i,k}) - T_{R,k}^{\mathcal{N}}(\theta_{R,i,k}) \right\|_2^2}{\sum_{i=1}^{n} \|\theta_{D,i,k} - \theta_{R,i,k}\|_2^2}}. \tag{84}$$

Table 3: Trajectory-level retain-update contraction in the readout setting. Pair denotes the percentage of model-pair–step observations satisfying $r_{i,k}^{\mathcal{N}} < 1 - 10^{-10}$, and RMS denotes the percentage of partition–step observations satisfying $\widehat{c}_k^{\mathcal{N}} < 1 - 10^{-10}$. Each channel includes five strengths, with zero represented by the shared noiseless row and the maximum matching Figure 2. Percentages are averaged over five partitions with 1,024 model pairs each, using $k = 1, \ldots, 119$.

		Normal deletion		Corrupted-label deletion	
Channel	Strength	Pair (%)	RMS (%)	Pair (%)	RMS (%)
Noiseless	0	87.0	63.7	90.7	100.0
Depolarizing	$q = 0.05$	86.9	65.2	90.5	96.3
	$q = 0.20$	86.4	71.1	89.8	97.5
	$q = 0.40$	85.5	81.2	88.1	100.0
	$q = 0.80$	83.0	100.0	82.3	100.0
Amplitude damping	$\gamma = 0.05$	84.9	66.6	89.9	99.2
	$\gamma = 0.20$	77.1	24.4	86.0	94.3
	$\gamma = 0.40$	64.1	0.0	78.3	72.8
	$\gamma = 0.80$	36.3	0.0	58.6	66.7

Unlike the fraction of individually contracting pairs, $\widehat{c}_k^{\mathcal{N}}$ weights each pair according to its squared pre-update separation. Indeed, writing $d_{i,k} = \|\theta_{D,i,k} - \theta_{R,i,k}\|_2$ gives

$$\left(\widehat{c}_k^{\mathcal{N}}\right)^2 = \sum_{i=1}^{n} \frac{d_{i,k}^2}{\sum_j d_{j,k}^2} \left(r_{i,k}^{\mathcal{N}}\right)^2. \tag{85}$$

Consequently, a majority of individual pairs may contract while a smaller number of sufficiently large or strongly expanding pairs make $\widehat{c}_k^{\mathcal{N}} \geq 1$.

The diagnostic uses the same sample-averaged squared loss, quadratic regularizer $\lambda\|\theta\|_2^2/2$ with $\lambda = 0.01$, learning rate $\eta = 0.15$, channel configuration, and exact-expectation full-batch gradients as the corresponding training condition. We analyze the readout placement for noiseless training, depolarization with $q \in \{0.05, 0.20, 0.40, 0.80\}$, and amplitude damping with $\gamma \in \{0.05, 0.20, 0.40, 0.80\}$, under both deletion scenarios. Finite-shot and interleaved settings are not included in this diagnostic. Steps $k = 1, \ldots, 119$ are analyzed; $k = 0$ is excluded because the shared initialization gives zero separation.

We classify a pair or RMS update as contractive when its ratio is below $1 - 10^{-10}$. For numerical stability, a ratio is set to one if its denominator—the pairwise Euclidean distance in Eq. 83, or the pre-update RMS distance in Eq. 84—is at most 10^{-14}. This fallback is never triggered at the analyzed steps. For each condition and deletion scenario, the pairwise statistic therefore aggregates $5 \times 1{,}024 \times 119 = 609{,}280$ pair-step observations, whereas the RMS statistic aggregates $5 \times 119 = 595$ partition-step observations.

Table 3 shows that contraction is common along the sampled trajectories, although it is neither uniform nor universal. In the noiseless setting, 87.0% and 90.7% of pair-step observations contract under normal and corrupted-label deletion, respectively. Depolarizing training preserves similarly high pairwise contraction frequencies across the tested strengths. At the RMS level, corrupted-label deletion is contractive at 96.3–100% of partition-steps under depolarization, while normal deletion increases from 63.7% in the noiseless setting to 81.2% at $q = 0.40$.

Amplitude damping exhibits a different propagation pattern. Under normal deletion, its pairwise contraction frequency decreases from 84.9% at $\gamma = 0.05$ to 64.1% at $\gamma = 0.40$. The corresponding RMS contraction frequency drops more sharply, from 66.6% to 0%. Thus, at $\gamma = 0.40$, a majority of pair-step observations still contract individually, but sufficiently strong expansion among the remaining pairs prevents the ensemble RMS separation from contracting at any analyzed partition-step. A weaker version of this effect is already visible at $\gamma = 0.20$, where the pairwise contraction rate remains 77.1% but the RMS rate falls to 24.4%.

These diagnostics help interpret the endpoint results in Section 4. The frequent contraction observed in the noiseless and depolarizing settings provides empirical motivation for studying the contractive regime of Corollary 1, while not verifying its uniform assumption. Conversely, the loss of RMS

contraction under stronger amplitude damping is consistent with its failure to produce the same parameter-discrepancy reductions as depolarization. In terms of Theorem 1, this illustrates why a smaller channel-conditioned deletion perturbation alone need not tighten the final discrepancy when propagation of previously accumulated differences becomes less contractive or expansive. The trajectory diagnostics therefore support the qualitative mechanism of the theory without constituting an empirical estimate of the uniform constants $c_k^{\mathcal{N}}$.

F Experiment 2: Details

F.1 Data, Circuit, and Training Protocol

Data preparation and deletion task. We use five fixed, class-balanced partitions of MNIST, each containing 200 training and 100 validation images per digit. These training and validation subsets are mutually disjoint across the five partitions. A single held-out set of 200 images per digit from the official test set is shared by all partitions and conditions. Each 28×28 image is flattened, padded with 240 zeros, and normalized to unit Euclidean norm to obtain 1,024 real amplitudes. No learned preprocessing or data augmentation is used during full training or retraining. The forget set contains all 200 digit-0 training images; the retained set contains the remaining 1,800 images. Digit-0 validation and test subsets contain 100 and 200 images, respectively. Retraining removes the forget examples but preserves the ten-output architecture.

Circuit and objective. Each of the 16 ansatz layers applies R_y and then R_z to every qubit, followed by CNOTs in the order $0 \to 1, 1 \to 2, \ldots, 8 \to 9, 9 \to 0$. This gives 320 trainable rotation angles and no trainable classical head. Writing $f_\theta(x) \in \mathbb{R}^{10}$ for the ten Pauli-Z expectations and e_y for the one-hot class vector, the training objective is

$$\mathcal{L}_Z^{\mathcal{N}}(\theta) = \frac{1}{2|Z|} \sum_{(x,y)\in Z} \left\| f_\theta^{\mathcal{N}}(x) - \left(e_y - \tfrac{1}{10}\mathbf{1}\right) \right\|_2^2 . \tag{86}$$

Thus, the target is 0.9 for the correct class and -0.1 for each other class. Predictions use the largest readout value. Training uses 2,000 minibatch-SGD updates with learning rate 0.5, batch size 32, no regularization, and no momentum. Each angle is initialized independently from $\mathcal{N}(0, 0.3^2)$.

Paired randomness and noise grids. The eight initializations have distinct random seeds; the same initial vector is reused for full training, retraining, and all noise settings at a given model index. The same eight seed indices are used in each partition, giving 40 full-data/retrained model pairs per noise condition across the five data partitions. At each SGD update, we draw a fresh permutation of the full training set: the full model uses its first 32 examples, while the retrained model uses its first 32 retained examples. Both batches are uniform without-replacement samples from their respective datasets. The permutation stream is reused across noise settings and differs across initialization indices. The five-partition comparisons use depolarization at $q \in \{0.05, 0.20\}$ and amplitude damping at $\gamma \in \{0.02, 0.08\}$, together with a shared noiseless reference. Channels act on all ten qubits after the final ansatz layer, and channel-conditioned training uses exact expectations and gradients. For these local readouts, depolarization gives $f^{\mathcal{D}_q} = (1-q)f^{\mathcal{I}}$ and amplitude damping gives $f^{\mathcal{A}_\gamma} = (1-\gamma)f^{\mathcal{I}} + \gamma\mathbf{1}$. Both full training and retraining use the same channel setting and the fixed 2,000-update budget.

F.2 Unlearning Implementations

Each method starts from the corresponding full-data checkpoint with a freshly initialized Adam optimizer. We use learning rate 0.007, $(\beta_1, \beta_2) = (0.9, 0.999)$, and $\epsilon = 10^{-8}$. The minibatch random seed is reset for each method and paired across noise settings at a given initialization index. Retrained checkpoints are used only as evaluation references. For the main experiment, the training readout channel is retained throughout unlearning, including loss gradients, SalUn mask construction, and DELETE teacher and student predictions. All these computations use exact expectations, without finite-shot sampling. Stopping decisions and reported accuracies instead use exact, channel-free predictions, so the success criterion is the same across noise conditions.

GA, FT, and GA+FT. GA minimizes $-\mathcal{L}_F^{\mathcal{N}}$ using 32 forget examples per update; FT minimizes $\mathcal{L}_R^{\mathcal{N}}$ using 32 retained examples. GA+FT minimizes $\mathcal{L}_R^{\mathcal{N}} - 0.25\mathcal{L}_F^{\mathcal{N}}$, with separate sample means over 16 retained and 16 forget examples in each update. Batches are sampled without replacement within each update.

SalUn. We compute the channel-conditioned mean forget-loss gradient over all 200 forget examples at the starting checkpoint and select the 160 of 320 coordinates with the largest absolute gradients. This mask remains fixed throughout unlearning, and unselected coordinates remain at their starting values. Each 64-update cycle consists of seven forget-set updates with independently sampled incorrect labels, uniform over digits 1–9, followed by 57 retained-set updates with original labels. Both phases use batches of 32 and the squared loss in Eq. 86.

DELETE. We adapt DELETE to the QNN's ten readouts using a fixed unit-temperature softmax, without adding trainable parameters. The starting model, evaluated under its training channel, supplies a frozen teacher distribution whose digit-0 probability is set to zero and whose remaining probabilities are renormalized. DELETE minimizes teacher-to-student KL divergence on batches of 32 forget examples; our implementation uses this masked-distillation objective. The student uses the same channel and temperature as the teacher. These are QNN adaptations of the methods, rather than reproductions of every component of their original architectures.

F.3 Evaluation and Aggregation

Distributional distances. Parameter distances use the 320 raw rotation angles, without periodic wrapping or dimension normalization. Prediction-space transport uses all ten exact readouts on the same 2,000 test images, flattened and divided by $\sqrt{2000 \times 10}$. Within each partition and space, we minimize the sum of squared Euclidean assignment costs between the eight full and eight retrained models, then take the square root of the mean matched cost. Initialization pairing does not constrain this assignment. Clean and noise-matched prediction distances are evaluated separately; for these readout channels, their exact scaling relation also provides a consistency check. Each distance ratio uses its corresponding noiseless reference within the same partition; both prediction audits use the noiseless clean-prediction reference. The reported ratios are means across the five partitions, with one sample standard deviation describing their variation. For parameter space, the ratios at $q = 0.05, 0.20$ are 0.9900 ± 0.0044 and 0.9608 ± 0.0142; at $\gamma = 0.02, 0.08$, they are 1.0096 ± 0.0197 and 1.0227 ± 0.0203. These distances use all eight model pairs in each partition, independently of subsequent unlearning success.

Stopping rule and reported cost. We check clean forget training and validation accuracy at initialization and after every unlearning update, stopping at the first simultaneous 3% hit. Retain accuracy is reported but is not a stopping constraint; test data are not used to choose the stopping point. Consequently, test FA can exceed 3% even for a successful run. Steps count optimizer updates and exclude saliency-mask and teacher preparation, as well as evaluation. Thus, steps measure update cost rather than total runtime or quantum-circuit evaluations. In Table 2, FA, RA, and Steps pool the same successful runs across the five partitions within each method and condition, whereas Original and Retrain average all 40 models. All methods except SalUn succeeded 40 times, while SalUn had some failures. This cost aggregation weights successful runs equally, rather than first averaging within each partition. For each method, relative cost reduction is $100(1 - \bar{s}_\nu/\bar{s}_0)$, where $\bar{s}_\nu$ and $\bar{s}_0$ are its unrounded success-conditioned mean steps under the corresponding noise condition and noiseless baseline. The table's mean cost reduction averages these percentages equally over the five methods. Successful subsets can differ across conditions, so a lower conditional mean does not by itself establish a paired speedup or account for non-attainment.

G Experiment 2: Predictive Accuracy

Figure 14 supplements the distributional comparisons in Figure 5 using the same trained model ensembles. All predictions are evaluated with channel-free exact expectations. Full-model accuracy is measured over all ten test classes, while retrained-model accuracy is measured over the nine retained classes. These curves therefore describe performance on the respective tasks, rather than a

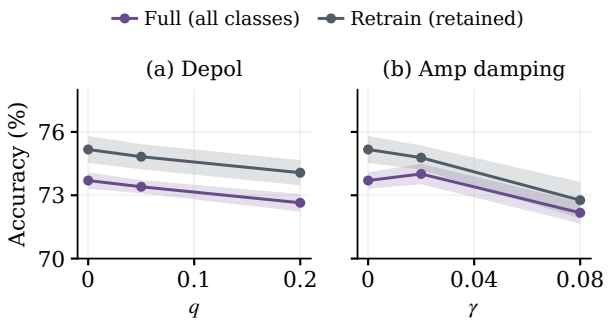

Figure 14: Clean test accuracy for the pure 10-qubit QNN in Figure 5. Accuracy is averaged over eight initializations within each partition; curves and bands show the mean and one sample standard deviation across five partition means. Full-model accuracy uses all ten classes, whereas retrained-model accuracy uses the nine retained classes.

direct comparison on an identical test subset. The figure includes all 40 models per condition and training procedure and is not conditioned on subsequent unlearning success.

Noiseless full-model accuracy is 73.70% and retrained-model retained-class accuracy is 75.17%. At $q = 0.20$, these values are 72.64% and 74.07%; at $\gamma = 0.08$, they are 72.17% and 72.77%. The distance changes thus occur alongside modest changes in clean predictive performance over the displayed range. These are pre-unlearning accuracies; the FA and RA values in Table 2 describe the subsequent unlearning endpoints.