FINITE-SHOT QUANTUM FEDERATED LEARNING: TRAJECTORY-DEPENDENT UPDATE DISCREPANCIES AND NONCONVEX CONVERGENCE UNDER PROXIMAL DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study quantum federated learning with variational quantum classifiers under finite-shot measurements, heterogeneous clients, multiple local updates, and partial participation. Finite-shot parameter-shift gradients are evaluated along recursively evolving proximal client trajectories, so their effect cannot be represented by simply adding an $O(1/S)$ variance term to a standard proximal federated bound. We therefore analyze the trajectory-dependent discrepancy between the actual federated update and an ideal global-gradient step. At the circuit level, we establish unbiasedness of the parameter-shift estimator, coordinatewise variance $O(1/S)$, and gradient mean-squared error $O(d_q/S)$. At the optimization level, we establish first-order stationarity guarantees for smooth nonconvex objectives with time-varying learning rates and proximal coefficients, without a Polyak–Łojasiewicz condition. The analysis identifies two finite-shot effects: a direct stochastic-gradient residual and an indirect trajectory-mediated residual. It also shows that heterogeneity does not create an independent common-iterate $O(\zeta^2)$ error; instead, it acts through subsequent client drift. For uniform sampling without replacement, we derive an additional correction proportional to $(K-m)/(m(K-1))$ and terminal drift energy, which vanishes under full participation. Experiments confirm near-inverse-shot estimator scaling and illustrate the resulting drift and stationarity behavior.

1 INTRODUCTION

Finite-shot uncertainty in quantum federated learning (QFL) is not merely an additional stochastic-gradient variance. A measurement error at one local step changes the client model and therefore every subsequent gradient evaluation along its trajectory. With heterogeneous objectives, multiple local updates, proximal regularization, and partial client participation, finite-shot uncertainty becomes a trajectory-dependent optimization error rather than an isolated additive term. This issue is intrinsic to variational quantum classifiers (VQCs), whose parameters are optimized through measurement-based circuit evaluations Beer et al. (2020); Cerezo et al. (2021). The parameter-shift rule constructs expectation-value derivatives from shifted circuit evaluations, each based on finitely many measurements. Although an individual circuit-derivative estimator can be unbiased, its finite-shot fluctuation perturbs the local model and changes the points at which later gradients are evaluated.

QFL enables clients to collaboratively train quantum or hybrid quantum–classical models without sharing private data Chen & Yoo (2021); Xia & Li (2021); Innan et al. (2024). Existing work has studied QFL model design, centralized finite-shot optimization, and distributed quantum learning under device heterogeneity or partial participation Jose & Simeone (2022); Pokharel et al. (2026); Rahman et al. (2026). Our focus is the interaction among these effects: *how finite-shot parameter-shift uncertainty propagates through heterogeneous, recursively evolving proximal client trajectories in nonconvex federated optimization*. In classical federated learning, multiple local updates create client drift because clients optimize different objectives McMahan et al. (2017). Proximal methods such as FedProx regularize local training toward the broadcast model Li et al. (2020a). In finite-shot QFL, however, the proximal term also changes the memory through which mini-batch and measurement errors propagate across local steps. The relevant question is therefore not whether proximal federated

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.

learning can be combined with quantum gradients, but how finite-shot uncertainty, heterogeneity, proximal memory, and client subsampling jointly alter the global update.

Let $\boldsymbol{\Delta}^{(t)} := \boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}$ denote the aggregated update in round t. We define the central quantity of the analysis as the discrepancy between this update and an idealized $\mathcal{T}$-step global-gradient update: $\mathcal{E}_t := \boldsymbol{\Delta}^{(t)} + \eta_t \mathcal{T} \nabla F(\boldsymbol{\theta}^{(t)})$. After the first local step, clients evaluate stochastic gradients at different random models determined by previous mini-batch samples, finite-shot outcomes, heterogeneous objectives, and proximal corrections. Thus, the server update is not simply a noisy global gradient evaluated at a common point. In particular, the local displacement recursion contains $q_t = 1 - \eta_t \mu_t$, so earlier stochastic gradients are weighted by powers of q_t. Proximal regularization therefore *changes both the memory of local perturbations and the sensitivity of the global update to client displacement, producing a drift–stability trade-off.*

At the circuit level, finite-shot parameter-shift estimation has coordinatewise variance $O(1/S)$ and a d_q-coordinate gradient mean-squared error of $O(d_q/S)$. At the optimization level, we impose conditional unbiasedness for the complete local stochastic-gradient oracle rather than infer it automatically from circuit-level unbiasedness. The analysis decomposes $\mathcal{E}_t$ into a centered stochastic-oracle component and a trajectory/proximal component controlled by the client-drift energy $D_{t,\tau}$. Consequently, *finite-shot uncertainty enters through two channels: directly through stochastic-gradient fluctuations and indirectly by perturbing later client iterates and gradients.*

The same decomposition clarifies heterogeneity. In the broadcast model, weighted client-gradient deviations cancel: $\sum_{k=1}^{K} p_k \left[\nabla F_k(\boldsymbol{\theta}^{(t)}) - \nabla F(\boldsymbol{\theta}^{(t)})\right] = \mathbf{0}$. Hence, heterogeneity does not create an independent common-iterate $O(\zeta^2)$ residual. It re-enters after clients leave the broadcast model and follow different local trajectories. We first retain this drift explicitly to obtain a first-order stationarity guarantee for smooth, lower-bounded, generally nonconvex objectives without a Polyak–Łojasiewicz (PL) condition, and then close the drift recursion in terms of statistical heterogeneity.

Partial participation introduces an additional layer of randomness because the server samples client endpoints that are already random under mini-batch and finite-shot updates. Conditioning first on the realized endpoints and applying the exact without-replacement finite-population identity yields a correction proportional to $\frac{K-m}{m(K-1)}$ times the terminal client-drift energy. This term vanishes under full participation. Experiments verify near-inverse-shot variance scaling ($\widehat{\rho} = -0.9817$), examine the analyzed proximal-SGD dynamics under constant and annealed schedules, and provide complementary Adam/AdamW evaluations on practical IID and non-IID federated VQC tasks.

Our contributions are as follows.

- **Trajectory-dependent nonconvex convergence.** We identify the global-update discrepancy $\mathcal{E}_t$ and derive its exact decomposition into centered stochastic-oracle and trajectory/proximal components. Controlling these new trajectory-dependent terms before the standard descent and telescoping steps yields first-order stationarity for smooth nonconvex proximal QFL with multiple local updates and time-varying schedules, without a PL condition.
- **Two-channel finite-shot effects.** We connect the measurement budget to coordinatewise $O(1/S)$ variance and gradient MSE $O(d_q/S)$, and show that finite-shot uncertainty contributes both directly through the stochastic oracle and indirectly through perturbed client trajectories.
- **Recursive heterogeneity and proximal drift.** We show that heterogeneity cancels at the broadcast iterate and re-enters through subsequent client displacement. We close this drift recursion in terms of statistical heterogeneity and stochastic uncertainty, exposing a schedule-dependent drift–stability trade-off rather than a monotonic benefit from increasing the proximal coefficient.
- **Partial participation through endpoint dispersion.** By conditioning first on the realized stochastic client endpoints, we derive the exact without-replacement correction $(K-m)/(m(K-1))$ coupled to terminal client-drift energy. The resulting participation penalty depends jointly on subset size and endpoint dispersion and vanishes under full participation.
- **Theory-aligned empirical validation.** We verify near-inverse-shot scaling of the quantum estimator and separately evaluate the same proximal-SGD dynamics used in the analysis, alongside practical IID and non-IID federated VQC experiments.

Table 1 positions our analysis relative to representative theoretical work on finite-shot quantum optimization and distributed quantum learning.

Table 1: Analytical comparison with representative work on finite-shot quantum optimization and distributed quantum learning. "PL" denotes the Polyak–Łojasiewicz condition.

Work	Optimization setting	Quantum uncertainty	Heterogeneity and local updates	Participation and propagated object
Jose & Simeone (2022)	Centralized stochastic quantum optimization	Finite measurements, gate noise, and error mitigation	No federated client trajectories	Not applicable, centralized stochastic-gradient effects
Rahman et al. (2026)	QFL convergence under PL-type condition	Device noise and measurement uncertainty	Non-IID data with personalized regularization	Sporadic participation, PL-based convergence
Pokharel et al. (2026)	Distributed learning under PL geometry	Finite-shot measurement dependence	Non-IID data with multi-step local SGD	Fidelity-aware partial selection
QProxFL (ours)	**Smooth nonconvex first-order stationarity without PL**	**Parameter-shift variance $O(1/S)$ and MSE $O(d_q/S)$**	**Heterogeneous proximal client trajectories**	**Drift envelope, exact discrepancy decomposition, direct and trajectory-mediated residuals, and finite-population factor $(K-m)/(m(K-1))$ times terminal client-drift energy**

2 PRELIMINARIES

We consider clients $\mathcal{K} = \{1, \dots, K\}$, where client k holds private data $\mathcal{D}_k = \{(\mathbf{x}_i^{(k)}, y_i^{(k)})\}_{i=1}^{N_k}$ drawn from a client-specific distribution $\mathcal{P}_k$. Let $\boldsymbol{\theta} \in \mathbb{R}^d$ denote the shared parameters, $f_{\boldsymbol{\theta}}$ the prediction model, and $\mathcal{L}$ a sample-wise loss. The local and global objectives are $F_k(\boldsymbol{\theta}) := N_k^{-1} \sum_{i=1}^{N_k} \mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}_i^{(k)}), y_i^{(k)})$ and $F(\boldsymbol{\theta}) := \sum_{k=1}^{K} p_k F_k(\boldsymbol{\theta})$, where $p_k := N_k / \sum_j N_j$. FedAvg McMahan et al. (2017) approximately minimizes F through local stochastic updates and server aggregation. Under non-IID data and multiple local steps, clients follow different trajectories, producing client drift. FedProx Li et al. (2020a) regularizes these deviations toward the broadcast model. At round t, the server broadcasts $\boldsymbol{\theta}^{(t)}$. Each participating client initializes $\boldsymbol{\theta}_k^{(t,0)} = \boldsymbol{\theta}^{(t)}$ and performs $\mathcal{T}$ proximal-SGD updates, $\boldsymbol{\theta}_k^{(t,\tau+1)} = \boldsymbol{\theta}_k^{(t,\tau)} - \eta_t \big[\mathbf{g}_k(\boldsymbol{\theta}_k^{(t,\tau)}, \boldsymbol{\xi}_k^{(t,\tau)}) + \mu_t(\boldsymbol{\theta}_k^{(t,\tau)} - \boldsymbol{\theta}^{(t)})\big]$, $\tau = 0, \dots, \mathcal{T}-1$, where η_t and μ_t are the learning rate and proximal coefficient, $\boldsymbol{\xi}_k^{(t,\tau)}$ denotes mini-batch randomness, and $\mathbf{g}_k$ estimates ∇F_k. Its conditional unbiasedness and variance, including finite-shot effects, appear in Section 3. After $\mathcal{T}$ local updates, participating clients transmit endpoints for server aggregation. The VQC model and finite-shot parameter-shift estimator are described next.

3 QUANTUM FEDERATED LEARNING

Quantum federated learning (QFL) uses a variational quantum circuit (VQC) as the trainable local model. Each client optimizes its circuit using private data and periodically sends its parameters or updates to the server. We first specify the VQC and then characterize the finite-shot gradient uncertainty used in the federated analysis. Additional circuit details are provided in Appendix D.

3.1 LOCAL VARIATIONAL QUANTUM CLASSIFIER

After common classical preprocessing, client k maps feature $\mathbf{z}_i^{(k)} \in \mathbb{R}^n$ to an n-qubit state using encoding, $|\phi(\mathbf{z}_i^{(k)})\rangle = \Phi(\mathbf{z}_i^{(k)}) |0\rangle^{\otimes n}$, $\Phi(\mathbf{z}_i^{(k)}) = \bigotimes_{j=1}^{n} R_y(z_{i,j}^{(k)})$. The encoded state is transformed by a trainable hardware-efficient ansatz $U(\boldsymbol{\theta}_k)$ consisting of parameterized single-qubit rotations and fixed entangling layers: $|\psi(\mathbf{z}, \boldsymbol{\theta}_k)\rangle = U(\boldsymbol{\theta}_k) |\phi(\mathbf{z})\rangle$. Let d_q denote the number of trainable quantum parameters. For a bounded Hermitian observable M_c, the c-th circuit output is $f_{\boldsymbol{\theta}_k, c}(\mathbf{z}) = \langle \psi(\mathbf{z}, \boldsymbol{\theta}_k)| M_c |\psi(\mathbf{z}, \boldsymbol{\theta}_k)\rangle$. The readout vector $\mathbf{f}_{\boldsymbol{\theta}_k}(\mathbf{z})$ is converted to logits or probabilities according to the classifier, including binary Pauli-Z readout and multiclass readout schemes. For a differentiable loss, the complete local gradient is obtained by combining the Jacobian of the quantum readout with the loss derivative through the chain rule. The finite-shot result below concerns the readout derivatives; unbiasedness and variance of the complete mini-batch loss gradient are imposed separately through the stochastic-gradient oracle in Section 3.2.

Parameter-shift (PS) gradients. Consider a parameterized gate $U_j(\theta_j) = e^{(-i\theta_j G_j)}$ whose generator has eigenvalues $\pm\lambda_j$, with $\lambda_j > 0$. The PS rule gives $\frac{\partial f_{\boldsymbol{\theta},c}(\mathbf{z})}{\partial \theta_j} = \lambda_j \big[f_{\boldsymbol{\theta}+s_j\mathbf{e}_j,c}(\mathbf{z}) - f_{\boldsymbol{\theta}-s_j\mathbf{e}_j,c}(\mathbf{z}) \big], s_j = \frac{\pi}{4\lambda_j}$. For Pauli rotations, $\lambda_j = 1/2$ and $s_j = \pi/2$, so each derivative requires two shifted expectation-value evaluations. Unlike a finite difference, this identity is exact and does not require access to the complete quantum state.

Finite-shot PS estimation. On hardware or a shot-based simulator, an expectation value is estimated from independent measurement outcomes: $\widehat{f}^{(S)}_{\boldsymbol{\theta},c}(\mathbf{z}) = \frac{1}{S}\sum_{s=1}^{S} X_s, \mathbb{E}[X_s] = f_{\boldsymbol{\theta},c}(\mathbf{z})$. Thus, even for a fixed circuit and input, finite measurements introduce stochasticity.

Lemma 3.1 (Finite-shot PS estimation)**.** *Suppose $\|M_c\| \le \Lambda_M$ and each shifted expectation is estimated using S independent shots. Then $\mathbb{E}\big[\widehat{f}^{(S)}_{\boldsymbol{\theta},c}(\mathbf{z})\big] = f_{\boldsymbol{\theta},c}(\mathbf{z})$, $\mathrm{Var}\big[\widehat{f}^{(S)}_{\boldsymbol{\theta},c}(\mathbf{z})\big] \le \frac{\Lambda_M^2}{S}$. If the positive and negative shifts are measured independently, then $\widehat{\partial_{\theta_j} f_{\boldsymbol{\theta},c}} := \lambda_j \big[\widehat{f}^{(S)}_{\boldsymbol{\theta}+s_j\mathbf{e}_j,c} - \widehat{f}^{(S)}_{\boldsymbol{\theta}-s_j\mathbf{e}_j,c} \big]$ is unbiased and satisfies $\mathrm{Var}\big[\widehat{\partial_{\theta_j} f_{\boldsymbol{\theta},c}}\big] \le \frac{2\lambda_j^2\Lambda_M^2}{S} = O(S^{-1})$. For Pauli-rotation generators and measurements, $\lambda_j = 1/2$ and $\Lambda_M = 1$, giving $\mathrm{Var}\big[\widehat{\partial_{\theta_j} f_{\boldsymbol{\theta},c}}\big] \le \frac{1}{2S}$.*

For a common shot budget, summing over the d_q shifted coordinates gives $\mathbb{E}\big[\big\|\widehat{\nabla}_q f_{\boldsymbol{\theta},c} - \nabla_q f_{\boldsymbol{\theta},c}\big\|_2^2\big] \le \frac{2\Lambda_M^2}{S}\sum_{j=1}^{d_q}\lambda_j^2$, which reduces to $d_q/(2S)$ for Pauli rotations and Pauli measurements. Thus, finite-shot uncertainty decreases as S^{-1} and grows with the number of shifted coordinates. This lemma establishes unbiasedness for a circuit-expectation derivative, not automatically for the gradient of a nonlinear empirical loss. The complete loss-gradient estimator also depends on the chain-rule composition and mini-batch sampling. We therefore impose its conditional unbiasedness and variance explicitly through the stochastic-gradient oracle in Section 3.2.

3.2 Convergence Analysis

The convergence analysis tracks how finite-shot uncertainty propagates through the federated update rather than treating measurement noise as an isolated variance term. The analysis has four stages: *(i) a recursive bound on proximal client trajectories, (ii) an exact decomposition of the global update discrepancy, (iii) a one-round descent inequality that separates direct stochastic-oracle error from trajectory-mediated error, and (iv) nonconvex stationarity bounds under full and partial participation.*

The central quantity is the discrepancy $\mathcal{E}_t := \boldsymbol{\Delta}^{(t)} + \eta_t \mathcal{T} \nabla F(\boldsymbol{\theta}^{(t)})$ between the actual aggregated update and an idealized $\mathcal{T}$-step global-gradient update. The analysis first controls the client-drift energy generating this discrepancy and only then applies smoothness-based descent and telescoping arguments. This order is essential because the stochastic gradients are evaluated along distinct, recursively evolving proximal trajectories rather than at a common iterate.

Assumption 3.2 (Lower-bounded objective)**.** *The global objective $F : \mathbb{R}^d \to \mathbb{R}$ is lower bounded, i.e., there exists $F^\star > -\infty$ such that $F(\boldsymbol{\theta}) \ge F^\star, \forall\, \boldsymbol{\theta} \in \mathbb{R}^d$.*

Assumption 3.3 (β-smoothness)**.** *For every client $k \in \mathcal{K}$, the local objective F_k is β-smooth. Consequently, the global objective $F(\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k F_k(\boldsymbol{\theta}), p_k \ge 0, \sum_{k=1}^{K} p_k = 1$ is also β-smooth.*

Assumption 3.4 (Stochastic-gradient oracle)**.** *Let $\mathcal{F}_{t,\tau}$ denote the optimization history immediately before the stochastic-oracle evaluation at local step τ of round t. For each client k, assume*

$$\mathbb{E}\big[\mathbf{g}_k^{(t,\tau)} \,|\, \mathcal{F}_{t,\tau}\big] = \nabla F_k(\boldsymbol{\theta}_k^{(t,\tau)}), \quad \mathbb{E}\big[\big\|\mathbf{g}_k^{(t,\tau)} - \nabla F_k(\boldsymbol{\theta}_k^{(t,\tau)})\big\|_2^2 \,|\, \mathcal{F}_{t,\tau}\big] \le \sigma_{k,t}^2, \tag{1}$$

where $\mathbf{g}_k^{(t,\tau)} := \mathbf{g}_k(\boldsymbol{\theta}_k^{(t,\tau)}, \boldsymbol{\xi}_k^{(t,\tau)})$, $\sigma_{k,t}^2 := \frac{\sigma_{\mathrm{mb},k}^2}{B} + \frac{c_{\mathrm{ps},k} d_q}{S_{k,t}}$. Here, B is the mini-batch size, $S_{k,t}$ is the shot budget per shifted circuit evaluation, d_q is the number of quantum parameters estimated through PS measurements, and $c_{\mathrm{ps},k}$ collects bounded estimator-dependent constants. The second term represents the circuit-level finite-shot contribution at the complete-gradient-oracle level; it does not assert unbiasedness of the gradient of an arbitrary nonlinear loss solely from PS unbiasedness.

Assumption 3.5 (Bounded second moment)**.** *There exists a finite constant $G \ge 0$ such that, for every client, communication round, and local step, $\mathbb{E}\big[\|\mathbf{g}_k^{(t,\tau)}\|_2^2 \,|\, \mathcal{F}_{t,\tau}\big] \le G^2$ almost surely.*

Assumption 3.6 (Heterogeneity)**.** *There exists $\zeta \ge 0$ such that $\sum_k p_k \big\|\nabla F_k(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta})\big\|_2^2 \le \zeta^2, \forall \boldsymbol{\theta}$.*

Assumption 3.7 (Partial participation)**.** *For the partial-participation analysis, let $p_k = 1/K$ with $K \geq 2$. At round t, the server samples $\mathcal{S}^{(t)} \subseteq \mathcal{K}$ uniformly without replacement with $|\mathcal{S}^{(t)}| = m$. Conditional on $\boldsymbol{\theta}^{(t)}$, the subset draw is independent of the potential local-update randomness.*

The assumptions serve distinct roles. Lower boundedness and smoothness support descent and telescoping, while Assumption 3.4 links finite-shot measurements and mini-batch sampling to a common stochastic-oracle model. Assumption 3.5 controls the recursive trajectories generated by multiple local updates. Assumption 3.6 is needed only to bound explicit drift by ζ^2, not for the full-participation theorem. Assumption 3.7 applies only to the client-subsampling extension.

Client drift and proximal memory. Define the displacement of client k from the broadcast model at local step τ by $\mathbf{d}_k^{(t,\tau)} := \boldsymbol{\theta}_k^{(t,\tau)} - \boldsymbol{\theta}^{(t)}$. Since $\boldsymbol{\theta}_k^{(t,0)} = \boldsymbol{\theta}^{(t)}$, thus $\mathbf{d}_k^{(t,0)} = \mathbf{0}$, and the local update yields $\mathbf{d}_k^{(t,\tau+1)} = (1-\eta_t\mu_t)\mathbf{d}_k^{(t,\tau)} - \eta_t \mathbf{g}_k^{(t,\tau)}$. Define $q_t := 1-\eta_t\mu_t$, $A_{t,\tau} := \sum_{r=0}^{\tau-1} q_t^r$, $A_{t,0} := 0$, then $\mathbf{d}_k^{(t,\tau)} = -\eta_t \sum_{s=0}^{\tau-1} q_t^{\tau-1-s} \mathbf{g}_k^{(t,s)}$. Under the stepsize conditions $\eta_t \leq \frac{1}{\beta+\mu_t}$, and $\beta\eta_t\mathcal{T} \leq \frac{1}{6}$, $q_t \in [0,1]$. If $\mu_t > 0$, then $A_{t,\tau} = \frac{1-q_t^\tau}{\eta_t\mu_t} \leq \min\{\tau, \frac{1}{\eta_t\mu_t}\}$, whereas $A_{t,\tau} = \tau$ when $\mu_t = 0$.

The quantity $A_{t,\tau}$ is the effective memory of stochastic local gradients in the client trajectory. Without proximal regularization, this memory grows linearly with the number of local steps. When $\mu_t > 0$, earlier gradients are geometrically discounted, yielding a proximal drift envelope that can saturate at $(\eta_t\mu_t)^{-1}$. This is the contraction side of proximal regularization. As shown below, μ_t also enters the trajectory-sensitivity factor $(\beta + \mu_t)^2$; consequently, increasing the proximal coefficient does not imply a uniformly tighter convergence bound.

Lemma 3.8 (Client-drift bound)**.** *Under Assumption 3.5, for every client k, communication round t, and local step $\tau \in \{0, \ldots, \mathcal{T}\}$, it holds $\mathbb{E}\big[\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2 \,|\, \boldsymbol{\theta}^{(t)}\big] \leq \eta_t^2 G^2 A_{t,\tau}^2$. Consequently, when $\mu_t > 0$, $\mathbb{E}\big[\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2 \,|\, \boldsymbol{\theta}^{(t)}\big] \leq \eta_t^2 G^2 \min\{\tau^2, \frac{1}{(\eta_t\mu_t)^2}\}$, while for $\mu_t = 0$, $\mathbb{E}\big[\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2 \,|\, \boldsymbol{\theta}^{(t)}\big] \leq \eta_t^2 G^2 \tau^2$.*

For the server-level analysis, define the weighted client-drift energy $D_{t,\tau} := \sum_k p_k \mathbb{E}\big[\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2 \,|\, \boldsymbol{\theta}^{(t)}\big]$. Lemma 3.8 then immediately yields $D_{t,\tau} \leq \eta_t^2 G^2 A_{t,\tau}^2$. This bound uses only the stochastic-gradient second-moment condition; the heterogeneity assumption is not needed at this stage. Its significance is that proximal regularization changes how local-update errors accumulate with the local-step index. Without proximal regularization, $A_{t,\tau} = \tau$, and the drift envelope can grow quadratically as $\eta_t^2 G^2 \tau^2$. When $\mu_t > 0$, earlier local gradients are geometrically discounted through $q_t = 1 - \eta_t\mu_t$, yielding $A_{t,\tau} \leq \min\left\{\tau, (\eta_t\mu_t)^{-1}\right\}$. This is the contraction side of proximal regularization. The server-level analysis below shows that μ_t simultaneously enters the trajectory-sensitivity factor $(\beta + \mu_t)^2$, so a larger proximal coefficient does not imply a uniformly tighter convergence bound.

The resulting drift energy connects the remaining results: $\sum_{\tau=0}^{\mathcal{T}-1} D_{t,\tau}$ determines the trajectory-mediated contribution under full participation, whereas the terminal energy $D_{t,\mathcal{T}}$ determines the additional client-subsampling fluctuation under partial participation.

Exact global-update discrepancy. Under full participation, the server increment is $\boldsymbol{\Delta}^{(t)} := \boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)} = \sum_k^K p_k \mathbf{d}_k^{(t,\mathcal{T})}$. For multiple local steps, local gradients are evaluated along client-specific trajectories rather than at the common broadcast model. The next step is thus to identify exactly how these trajectory deviations enter the server update. Define the centered stochastic-gradient error $\mathbf{e}_k^{(t,\tau)} := \mathbf{g}_k^{(t,\tau)} - \nabla F_k(\boldsymbol{\theta}_k^{(t,\tau)})$, the aggregated stochastic-oracle error $\mathbf{N}^{(t,\tau)} := \sum_k^K p_k \mathbf{e}_k^{(t,\tau)}$, the trajectory-induced gradient deviation $\mathbf{R}^{(t,\tau)} := \sum_k^K p_k [\nabla F_k(\boldsymbol{\theta}_k^{(t,\tau)}) - \nabla F_k(\boldsymbol{\theta}^{(t)})]$, and the weighted mean client displacement as $\overline{\mathbf{d}}^{(t,\tau)} := \sum_k^K p_k \mathbf{d}_k^{(t,\tau)}$. We combine the trajectory and proximal deviations $\mathbf{W}^{(t,\tau)} := \mathbf{R}^{(t,\tau)} + \mu_t \overline{\mathbf{d}}^{(t,\tau)}$. Here, $\mathbf{N}^{(t,\tau)}$ is the aggregated centered stochastic-oracle error, while $\mathbf{W}^{(t,\tau)}$ contains the trajectory-induced gradient deviation and the proximal restoring term.

Lemma 3.9 (Exact full-participation decomposition)**.** *Under full participation, the server increment satisfies $\boldsymbol{\Delta}^{(t)} = -\eta_t \mathcal{T} \nabla F(\boldsymbol{\theta}^{(t)}) - \eta_t \sum_{\tau=0}^{\mathcal{T}-1} \big[\mathbf{N}^{(t,\tau)} + \mathbf{W}^{(t,\tau)}\big]$.*

Lemma 3.9 identifies the central quantity that must be controlled before the nonconvex descent analysis can be applied: $\mathcal{E}_t = -\eta_t \sum_{\tau=0}^{\mathcal{T}-1} [\mathbf{N}^{(t,\tau)} + \mathbf{W}^{(t,\tau)}]$. Here, $\mathbf{N}^{(t,\tau)}$ is *a centered stochastic-oracle fluctuation containing mini-batch and finite-shot uncertainty*, whereas $\mathbf{W}^{(t,\tau)}$ contains *the*

trajectory-induced gradient deviation and the proximal restoring term. Thus, finite-shot uncertainty enters the global update directly through the centered oracle and indirectly by perturbing future client iterates and gradients. *This decomposition explains why the resulting convergence behavior cannot be captured by simply appending an $O(1/S)$ variance term to a standard FedProx bound.*

At the broadcast iterate, the weighted client-gradient deviations cancel exactly. Thus, statistical heterogeneity does not create an independent common-iterate bias term in the aggregated update. It enters the analysis through the gradients evaluated after clients depart from the broadcast model and follow their distinct local trajectories. The subsequent bounds therefore control the trajectory energy $D_{t,\tau}$ rather than introduce a separate common-iterate $O(\zeta^2)$ residual. The term ζ^2 appears only when the trajectory recursion is closed using the heterogeneity assumption.

Lemma 3.10 (One-round global descent)**.** *Suppose Assumptions 3.3–3.5 hold under full participation,*

$$\mathbb{E}\big[F(\boldsymbol{\theta}^{(t+1)})\,|\,\boldsymbol{\theta}^{(t)}\big] \le F(\boldsymbol{\theta}^{(t)}) - \frac{\eta_t\mathcal{T}}{2}\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2 + \frac{5}{4}\eta_t(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1} D_{t,\tau} + \frac{3}{2}\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2. \quad (2)$$

Lemma 3.10 is the point at which the trajectory-level analysis interfaces with the smoothness-based descent argument. Its three terms have distinct meanings: the term $-\frac{\eta_t\mathcal{T}}{2}\|\nabla F(\boldsymbol{\theta}^{(t)})\|_2^2$ is the desired global descent, the term $\frac{5}{4}\eta_t(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1} D_{t,\tau}$ is the trajectory-mediated penalty, and the term $\frac{3}{2}\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2$ is the direct stochastic-oracle penalty.

The heterogeneity assumption is not required at this stage because the client-drift energy $D_{t,\tau}$ is retained explicitly. This distinction is important for finite-shot QFL: the finite-shot component of $\overline{\sigma}_t^2$, which contains terms proportional to $d_q/S_{k,t}$, enters directly through the stochastic oracle, while the same measurement uncertainty can perturb future local iterates and therefore re-enter indirectly through $D_{t,\tau}$. Thus, *finite-shot uncertainty has both a direct and a trajectory-mediated effect.* The next theorem first establishes a general stationarity guarantee with the drift term explicit; a subsequent drift recursion then closes this term using the heterogeneity assumption.

Nonconvex stationarity under full participation. Define the learning-rate-weighted stationarity measure $\mathcal{G}_T := \frac{1}{\sum_{t=0}^{T-1}\eta_t\mathcal{T}}\sum_{t=0}^{T-1}\eta_t\mathcal{T}\,\mathbb{E}[\|\nabla F(\boldsymbol{\theta}^{(t)})\|_2^2]$. Since $D_{t,\tau}$ is conditional on the round-t iterate, define its unconditional counterpart $\overline{D}_{t,\tau} := \mathbb{E}\big[D_{t,\tau}\big] = \sum_k p_k\mathbb{E}[\|\mathbf{d}_k^{(t,\tau)}\|_2^2]$.

Theorem 3.11 (Nonconvex stationarity under full participation)**.** *Suppose full client participation is used and Assumptions 3.2, 3.3, 3.4, and 3.5 hold for every communication round with learning rate $\eta_t \le \frac{1}{\beta+\mu_t}, \beta\eta_t\mathcal{T} \le \frac{1}{6}$. Then, for any $T \ge 1$, the iterates generated by QFL-VQC satisfy*

$$\mathcal{G}_T \le \frac{2\big[F(\boldsymbol{\theta}^{(0)}) - F^\star\big]}{\sum_{t=0}^{T-1}\eta_t\mathcal{T}} + \frac{\sum_{t=0}^{T-1}\big[\frac{5}{2}\eta_t(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1}\overline{D}_{t,\tau} + 3\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2\big]}{\sum_{t=0}^{T-1}\eta_t\mathcal{T}}. \quad (3)$$

Theorem 3.11 provides a weighted finite-horizon first-order stationarity bound for a smooth, lower-bounded, generally nonconvex objective without requiring convexity, strong convexity, or a Polyak–Łojasiewicz condition. The heterogeneity assumption is not needed because the client-drift contribution is retained explicitly through $\overline{D}_{t,\tau}$; it is introduced only to obtain the heterogeneity-dependent corollary. The finite-shot component of $\overline{\sigma}_t^2$ contributes both directly through the stochastic-oracle term and indirectly by perturbing subsequent client trajectories, as made explicit in Remark 3.14.

Explicit heterogeneity dependence. Theorem 3.11 retains the client-drift energy explicitly. We now close this term in terms of the statistical heterogeneity level ζ^2. The key step is a heterogeneity-sensitive drift recursion: heterogeneity cancels in the weighted client-gradient average at the common broadcast iterate and affects the global update only through the client-specific trajectories generated afterward. For any $\gamma_t > 0$, define $\rho_t := (1+\gamma_t)(1-\eta_t\mu_t+\eta_t\beta)^2$ and $\Phi_t := \sum_{\tau=0}^{\mathcal{T}-1}\sum_{r=0}^{\tau-1}\rho_t^r$. Here, γ_t is the auxiliary parameter introduced in the heterogeneity-sensitive drift recursion. The quantity ρ_t is the one-step amplification factor for client drift, while Φ_t accumulates this amplification over all local steps in a communication round. The resulting recursion, proved in the appendix, yields

$$\sum_{\tau=0}^{\mathcal{T}-1} D_{t,\tau} \le \eta_t^2\Phi_t\Big[\big(1+\frac{1}{\gamma_t}\big)\big(\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2 + \zeta^2\big) + \overline{\sigma}_t^2\Big]. \quad (4)$$

Corollary 3.12 (Heterogeneity-explicit nonconvex stationarity)**.** *Suppose the conditions of Theorem 3.11 hold and Assumption 3.6 is satisfied. Suppose further that, for each communication round t,*

there exists $\gamma_t > 0$ *such that* $5\eta_t^2(\beta+\mu_t)^2\Phi_t(1+\frac{1}{\gamma_t}) \le \mathcal{T}$. *Then* $\mathcal{G}_T \le \frac{4}{\sum_{t=0}^{T-1}\eta_t\mathcal{T}}(\Delta_F + \sum_{t=0}^{T-1}\mathcal{R}_t)$, *where* $\Delta_F = F(\boldsymbol{\theta}^{(0)}) - F^\star$, $\mathcal{R}_t := \frac{3}{2}\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2 + \frac{5}{4}\eta_t^3(\beta+\mu_t)^2\Phi_t\big[(1+\frac{1}{\gamma_t})\zeta^2 + \overline{\sigma}_t^2\big]$.

Remark 3.13 (How heterogeneity enters the convergence bound)**.** *Corollary 3.12 makes explicit what remains implicit in Theorem 3.11. Statistical heterogeneity does not produce a separate first-order residual at the synchronized model. Instead,* ζ^2 *enters only after the heterogeneity-sensitive client-drift recursion is substituted into the server-level descent bound, producing a contribution proportional to* $\eta_t^3(\beta+\mu_t)^2\Phi_t\zeta^2$. *Thus, the effect of heterogeneity is jointly controlled by the learning rate, the local-step horizon through* Φ_t, *and the proximal coefficient through both the trajectory amplification factor and the sensitivity factor* $(\beta+\mu_t)^2$.

Remark 3.14 (Finite-shot dependence)**.** *Let the finite-shot component of the effective oracle variance be* $\overline{\sigma}^2_{\mathrm{shot},t} := \sum_k^K p_k \frac{c_{\mathrm{ps},k} d_q}{S_{k,t}}$. *Substitution into the stationarity bound reveals two finite-shot channels:*

$$\underbrace{O\big(\beta\eta_t^2\mathcal{T}\,\overline{\sigma}^2_{\mathrm{shot},t}\big)}_{\textit{direct stochastic-oracle contribution}} \qquad \textit{and} \qquad \underbrace{O\big(\eta_t^3(\beta+\mu_t)^2\Phi_t\,\overline{\sigma}^2_{\mathrm{shot},t}\big)}_{\textit{trajectory-mediated contribution}}. \tag{5}$$

Increasing the shot budget reduces both channels with inverse-shot dependence for fixed d_q, *but their effects on stationarity are shaped differently by the stochastic-oracle and local-trajectory dynamics. Thus, finite-shot uncertainty is not represented by a single additive* $O(1/S)$ *term: it also perturbs the client trajectories through which heterogeneity and local computation affect the global update.*

For constant schedules, the bound scales $\mathcal{G}_T = O\big(\frac{\Delta_F}{T\eta\mathcal{T}}\big) + O\big(\beta\eta\overline{\sigma}^2\big) + O\big(\frac{\eta^2(\beta+\mu)^2\Phi}{\mathcal{T}}[\zeta^2+\overline{\sigma}^2]\big)$, with $\overline{\sigma}^2 = \sum_k p_k[\frac{\sigma^2_{\mathrm{mb},k}}{B} + \frac{c_{\mathrm{ps},k}d_q}{S_k}]$. Thus, under a fixed stepsize, the attainable stationarity neighborhood depends jointly on the learning rate, local-step dynamics, proximal coefficient, heterogeneity, and stochastic variance. In particular, finite-shot noise does not appear as a standalone optimization-independent $O(1/S)$ term, and heterogeneity does not appear as an independent common-iterate $O(\zeta^2)$ bias. Having characterized the full-participation dynamics, we next examine how sampling only a subset of these trajectory-dependent client endpoints changes the convergence guarantee.

Partial client participation. It introduces a second source of randomness on top of the stochastic local trajectories. The endpoint $\mathbf{d}_k^{(t,\mathcal{T})}$ is random because of mini-batch and finite-shot updates, while the server additionally samples a random subset of these endpoints. To separate these two effects, we first condition on the complete collection of potential client endpoints and only then average over the subset selection.

Under equal client weights, let $\overline{\boldsymbol{\Delta}}^{(t)} := \frac{1}{K}\sum_{k=1}^K \mathbf{d}_k^{(t,\mathcal{T})}$ and $\boldsymbol{\Delta}_{\mathrm{pp}}^{(t)} := \frac{1}{m}\sum_{k\in\mathcal{S}^{(t)}} \mathbf{d}_k^{(t,\mathcal{T})}$. Conditional on the realized endpoints, uniform sampling without replacement gives a zero-mean sampling fluctuation and the exact finite-population factor $c_m := \frac{K-m}{m(K-1)}$. Averaging subsequently over the local-update randomness yields $\mathbb{E}\big[\big\|\boldsymbol{\Delta}_{\mathrm{pp}}^{(t)} - \overline{\boldsymbol{\Delta}}^{(t)}\big\|_2^2 \,\big|\, \boldsymbol{\theta}^{(t)}\big] \le c_m D_{t,\mathcal{T}}$. Thus client sampling is controlled not only by m, but also by the dispersion of the terminal local trajectories.

Theorem 3.15 (Nonconvex stationarity under partial participation)**.** *Suppose the conditions of Theorem 3.11 hold, specialize to equal client weights, and suppose Assumption 3.7 holds. Then*

$$\mathcal{G}_T \le \frac{2\big[F(\boldsymbol{\theta}^{(0)}) - F^\star\big]}{\sum_{t=0}^{T-1}\eta_t\mathcal{T}} + \frac{\sum_{t=0}^{T-1}\Big[\frac{5}{2}\eta_t(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1}\overline{D}_{t,\tau} + 3\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2\Big]}{\sum_{t=0}^{T-1}\eta_t\mathcal{T}} + \frac{\sum_{t=0}^{T-1}\beta c_m \overline{D}_{t,\mathcal{T}}}{\sum_{t=0}^{T-1}\eta_t\mathcal{T}}. \tag{6}$$

Compared to Theorem 3.11, partial participation contributes exactly one additional term, $\frac{\sum_{t=0}^{T-1}\beta c_m\overline{D}_{t,\mathcal{T}}}{\sum_{t=0}^{T-1}\eta_t\mathcal{T}}$. This term has a direct interpretation: participation becomes costly when client endpoints are widely dispersed, but it has little effect when local trajectories stay close. At the one-round level, the corresponding objective penalty is $\frac{\beta}{2}c_m D_{t,\mathcal{T}}$; the factor becomes $\beta c_m\overline{D}_{t,\mathcal{T}}$ in Theorem 3.15 after the factor-of-two rearrangement used to obtain the stationarity bound. The factor c_m also provides two consistency checks. It decreases as more clients participate and satisfies $c_K = 0$. Hence the partial-participation theorem recovers Theorem 3.11 exactly when all clients participate.

Remark 3.16. *Using Lemma 3.8,* $\overline{D}_{t,\mathcal{T}} \le \eta_t^2 G^2 A_{t,\mathcal{T}}^2$ *and therefore* $\beta c_m\overline{D}_{t,\mathcal{T}} \le \beta c_m\eta_t^2 G^2 A_{t,\mathcal{T}}^2$. *For* $\mu_t > 0$, $\beta c_m\overline{D}_{t,\mathcal{T}} \le \beta c_m G^2 \min\big\{\eta_t^2\mathcal{T}^2, \frac{1}{\mu_t^2}\big\}$ *whereas* $\mu_t = 0$ *yields the Q-FedAvg envelope*

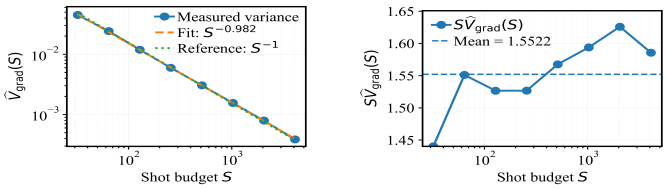
$\beta c_m \eta_t^2 G^2 \mathcal{T}^2$. Hence, proximal memory couples client participation to endpoint dispersion and can reduce subsampling sensitivity, but larger μ_t need not improve the complete bound monotonically because it also appears through $(\beta + \mu_t)^2$.

4 EXPERIMENTS

We evaluate three complementary aspects of the framework: *(i)* inverse-shot estimator variance predicted by Lemma 3.1; *(ii)* proximal client dynamics under the theory-aligned stochastic Prox-SGD recursion of Section 3; and *(iii)* practical federated classification under IID and non-IID partitions. The first two directly probe the analytical mechanisms, whereas the adaptive-optimizer classification experiments provide complementary practical evidence rather than direct verification of the SGD convergence theorem. Analytic circuit evaluations serve as references, while finite-shot experiments use genuine measurement sampling. For the proximal diagnostic, we compare Q-FedAvg ($\mu_t = 0$), Q-FedProx with constant $\mu_t = 1$, and Q-FedProx with μ_t annealed from 1 to 0, using $K = 5$ clients, a four-qubit one-layer VQC, $S = 128$ shots, and eight paired seeds. Practical classification uses MNIST and a six-class medical-image benchmark with $K \in \{3, 5, 10\}$ clients, together with a 15-client non-IID stress test. Detailed circuit configurations, shot-scaling settings, training schedules, and reproducibility details are provided in Appendix D.

4.1 DIRECT FINITE-SHOT PARAMETER-SHIFT DIAGNOSTIC

Lemma 3.1 predicts inverse-shot variance for each finite-shot parameter-shift derivative. Under the diagnostic setup described above, we compute the centered sample variance $\widehat{V}_{\mathrm{grad}}(S)$ and compare the finite-shot estimates with the corresponding analytic parameter-shift derivatives. Full MSE, bias, and shot-by-shot results are reported in Appendix D.



(a) Gradient variance. (b) Scaled variance.

Figure 1: Finite-shot parameter-shift scaling. The gradient variance follows a near-S^{-1} law ($\widehat{\rho} = -0.9817$), while $S\widehat{V}_{\mathrm{grad}}(S)$ remains approximately constant.

A log–log fit $\widehat{V}_{\mathrm{grad}}(S) = CS^{\widehat{\rho}}$ in Fig. 1(a) gives $\widehat{\rho} = -0.9817$, close to the predicted exponent -1. Figure 1(b) provides a complementary check: $S\widehat{V}_{\mathrm{grad}}(S)$ remains approximately constant, with coefficient of variation 3.64%. Thus, doubling the shot budget approximately halves the parameter-shift variance, while the empirical bias remains small relative to the measured variance.

A complementary downstream study in Appendix D shows that classification accuracy, Macro-F1, AUC, and agreement with analytic evaluation improve with the shot budget. These nonlinear quantities are not expected to follow an S^{-1} law and are interpreted only as practical consequences of estimator-level variance reduction. The experiment therefore validates measurement-level S^{-1} scaling, not an unconditional $O(1/S)$ optimization-error floor: finite-shot noise contributes directly through the learning-rate-weighted stochastic term and indirectly through subsequent client trajectories.

4.2 THEORY-ALIGNED PROXIMAL-SGD DIAGNOSTIC

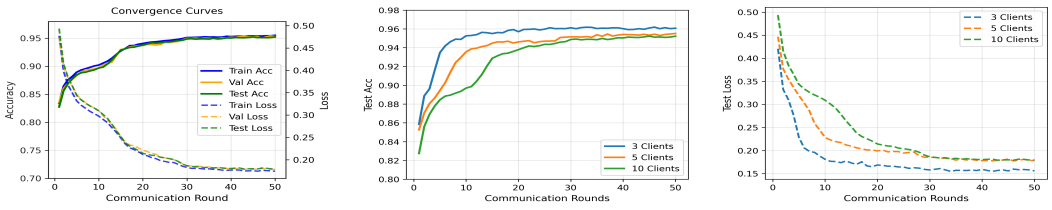
This experiment follows the analyzed stochastic Prox-SGD recursion and probes trajectory stabilization and stationarity. The diagnostic isolates trajectory effects from classification performance. For round t, define the terminal drift energy $D_t^{\mathrm{end}} := \sum_k p_k \|\boldsymbol{\theta}_k^{(t,\mathcal{T})} - \boldsymbol{\theta}^{(t)}\|_2^2$, with cumulative and peak summaries $D_{\mathrm{cum}}^{\mathrm{end}} := \sum_t D_t^{\mathrm{end}}$ and $D_{\mathrm{peak}}^{\mathrm{end}} := \max_t D_t^{\mathrm{end}}$. These endpoint measures differ from the theorem's within-round $\sum_{\tau=0}^{\mathcal{T}-1} D_{t,\tau}$, and all comparisons use the same eight paired seeds.

Table 2: Paired effects relative to Q-FedAvg over eight matched seeds. Negative values denote reductions; brackets show paired-bootstrap 95% confidence intervals.

Metric	Constant μ	Annealed μ_t
Cumulative endpoint drift	$+1.40\%$ $[-3.97, 8.59]$	-1.72% $[-5.48, 2.56]$
Peak endpoint drift	-1.19% $[-6.67, 5.75]$	-5.25% $[-12.07, 1.32]$
Final stationarity surrogate	$\mathbf{-26.28\%}$ $[\mathbf{-52.96}, \mathbf{-5.30}]$	-14.70% $[-44.75, 14.23]$

Table 2 shows directional drift reductions of 1.72% and 5.25% for annealing, although both confidence intervals include zero. Constant Q-FedProx gives the clearest stationarity improvement (-26.28%, bootstrap 95% CI $[-52.96\%, -5.30\%]$). Neither schedule universally dominates: annealing controls drift more consistently, whereas constant regularization better stabilizes difficult trajectories, supporting the theoretical drift–stability trade-off predicted by the convergence analysis.

4.3 Practical Federated Classification



(a) Convergence. (b) Accuracy. (c) Loss.

Figure 2: Federated classification under IID client partitions.

IID federation. With full participation and IID client partitions, we vary the number of clients. The 10-client model stabilizes near 95% test accuracy with closely aligned training, validation, and test performance (Fig. 2a). Across 3, 5, and 10 clients, final accuracy remains between 95.22% and 96.06%, while the rounds required to reach 95% increase from 10 to 39 (Fig. 2b–2c), indicating slower convergence rather than substantial final-performance loss. Larger federations mainly slow optimization, not degrade accuracy. With 50% participation, the 10-client model achieves 95.08% accuracy and loss 0.1810, close to full participation. Complete IID results are reported in Appendix D.

Statistical heterogeneity. Under Dirichlet client partitions, decreasing α produces more skewed label distributions. Test accuracy changes from 95.3% under IID data to 95.2%, 94.8%, and 92.0% for $\alpha = 1.0$, 0.5, and 0.3, respectively, while AUC decreases from 0.9957 to 0.9910. This trend is qualitatively consistent with the trajectory-based analysis: heterogeneous objectives generate increasingly different client trajectories after departure from the broadcast model. Because α is not the theoretical gradient-dissimilarity parameter ζ^2, these results are not a quantitative test of the ζ^2-dependent bound. Complete results are provided in Appendix D.

Larger non-IID federation. As a stress test, we evaluate a six-class medical-image task with 15 non-IID clients, 18,000 training examples, and 11,791 test examples. Client sizes range from 591 to 1,932 examples and exhibit class imbalance. After 25 rounds, the model achieves 87.38% test accuracy, macro-F1 $= 0.874$, macro-AUC $= 0.980$, and ECE $= 0.034$, extending the evaluation to a larger multiclass heterogeneous federation. Appendix D, Section D.7, compares Q-FedProx-Anneal with Q-FedAvg, centralized VQC, and local-only VQC baselines on the 15-client non-IID task. Q-FedProx-Anneal reaches 96.00% accuracy versus 95.67%, 93.45%, and 78.39%, respectively; these structural comparisons do not claim quantum advantage or intrinsic federated superiority.

5 Conclusion

We analyzed finite-shot quantum federated optimization with heterogeneous clients, proximal local updates, and partial participation. Finite-shot uncertainty contributes directly through stochastic gradients and indirectly through recursively evolving client trajectories. We established nonconvex first-order stationarity without a PL condition, with heterogeneity entering through post-broadcast drift and partial participation governed by the finite-population factor $(K-m)/(m(K-1))$. Experiments confirm near-inverse-shot scaling and the predicted proximal drift–stability trade-off.

REFERENCES

Azim Akhtarshenas, Mohammad Ali Vahedifar, Navid Ayoobi, Behrouz Maham, Tohid Alizadeh, Sina Ebrahimi, and David López-Pérez. Federated learning: A cutting-edge survey of the latest advancements and applications. *Computer Communications*, 228:107964, Dec. 2024.

Kerstin Beer, Dmytro Bondarenko, Terry Farrelly, Tobias J. Osborne, Robert Salzmann, Daniel Scheiermann, and Ramona Wolf. Training deep quantum neural networks. *Nature Communications*, 11(1):808, Feb. 2020.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191. Association for Computing Machinery, 2017.

Nader Bouacida, Jiahui Hou, Hui Zang, and Xin Liu. Adaptive federated dropout: Improving communication efficiency and generalization for federated learning. In *IEEE INFOCOM 2021 – IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 1–6, 2021.

M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3(9): 625–644, Sep. 2021.

Samuel Yen-Chi Chen and Shinjae Yoo. Federated quantum machine learning, 2021.

Iris Cong, Soonwon Choi, and Mikhail D. Lukin. Quantum convolutional neural networks. *Nature Physics*, 15(12):1273–1278, Aug. 2019.

Nouhaila Innan, Muhammad Al-Zafar Khan, Alberto Marchisio, Muhammad Shafique, and Mohamed Bennai. Fedqnn: Federated learning using quantum neural networks. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9, 2024.

Nouhaila Innan, Alberto Marchisio, Mohamed Bennai, and Muhammad Shafique. Qfnn-ffd: Quantum federated neural network for financial fraud detection. In *2025 IEEE International Conference on Quantum Software (QSW)*, pp. 41–47, 2025.

Sofiene Jerbi, Lukas J. Fiderer, Hendrik Poulsen Nautrup, Jonas M. Kübler, Hans J. Briegel, and Vedran Dunjko. Quantum machine learning beyond kernel methods. *Nature Communications*, 14 (1):517, Jan. 2023.

Sharu Theresa Jose and Osvaldo Simeone. Error-mitigation-aided optimization of parameterized quantum circuits: Convergence analysis. *IEEE Transactions on Quantum Engineering*, 3:1–19, 2022.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020a.

Zengpeng Li, Vishal Sharma, and Saraju P. Mohanty. Preserving data privacy via federated learning: Challenges and solutions. *IEEE Consumer Electronics Magazine*, 9(3):8–16, May 2020b.

Fabio Valerio Massoli, Lucia Vadicamo, Giuseppe Amato, and Fabrizio Falchi. A leap among quantum computing and quantum neural networks: A survey. *ACM Computing Surveys*, 55(5):98, Dec. 2022.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017.

Atit Pokharel, Shaba Shaon, Thomas Morris, and Dinh C. Nguyen. Distributed quantum learning over near-term devices: Convergence analysis and security design. *IEEE Journal on Selected Areas in Communications*, 44:4733–4747, 2026.

Cheng Qiao, Mianjie Li, Yuan Liu, and Zhihong Tian. Transitioning from federated learning to quantum federated learning in internet of things: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 27(1):509–545, 2025.

Ratun Rahman, Shaba Shaon, and Dinh C. Nguyen. Tackling heterogeneity in quantum federated learning: An integrated sporadic-personalized approach. *IEEE Transactions on Computers*, 75(4): 1418–1432, 2026.

Rod Rofougaran, Shinjae Yoo, Huan-Hsin Tseng, and Samuel Yen-Chi Chen. Federated quantum machine learning with differential privacy, Oct. 2023.

Jinjing Shi, Wenxuan Wang, Xiaoping Lou, Shichao Zhang, and Xuelong Li. Parameterized hamiltonian learning with quantum circuit. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6086–6095, May 2023.

Leo Sünkel, Michael Kölle, Tobias Rohe, and Thomas Gabor. Towards federated learning on the quantum internet. In *Computational Science – ICCS 2024: 24th International Conference, Malaga, Spain, July 2–4, 2024, Proceedings, Part VI*, pp. 330–344. Springer-Verlag, 2024.

Chuhan Wu, Fan Wu, Lingjuan Lyu, Xing Xie, Bo Li, and Qiang Zhang. Communication-efficient federated learning via knowledge distillation. *Nature Communications*, 13(1):2032, Apr. 2022.

Qi Xia and Qun Li. Quantumfed: A federated learning framework for collaborative quantum training. In *2021 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, 2021.

Qi Xia, Zeyi Tao, and Qun Li. Defending against byzantine attacks in quantum federated learning. In *2021 17th International Conference on Mobility, Sensing and Networking (MSN)*, pp. 145–152, 2021.

Contents

1 Introduction **1**

2 PRELIMINARIES **3**

3 QUANTUM FEDERATED LEARNING **3**
3.1 Local Variational Quantum Classifier 3
3.2 Convergence Analysis 4

4 EXPERIMENTS **8**
4.1 Direct Finite-Shot Parameter-Shift Diagnostic 8
4.2 Theory-Aligned Proximal-SGD Diagnostic 8
4.3 Practical Federated Classification 9

5 CONCLUSION **9**

Appendix **12**

A RELATED WORK **13**

B ADDITIONAL DETAILS ON VARIATIONAL QUANTUM CLASSIFIERS **15**
B.1 Data Encoding and Variational Circuit 15
B.2 Measurement Model and Predictive Readout 18
B.3 Bounds on Expectation Derivatives 19
B.4 Parameter-Shift Differentiation and Finite-Shot Estimation 21

C ADDITIONAL DETAILS ON THE CONVERGENCE ANALYSIS **22**
C.1 Analytical Setup and Local Proximal Descent 23
C.1.1 Additional stochastic-oracle and variance model 23
C.1.2 Local proximal dynamics 24
C.2 Proximal Memory and Client-Drift 26
C.2.1 Exact proximal-memory representation 26
C.2.2 Geometric memory factors 27
C.2.3 General client-drift envelope 28
C.2.4 Endpoint and cumulative client drift 30
C.2.5 Geometric filtering of stochastic perturbations 32
C.2.6 Heterogeneity-sensitive drift recursion 33
C.3 Full-Participation Update and One-round Descent 35
C.3.1 Exact server-update decomposition 35
C.3.2 One-round global descent 37
C.4 Nonconvex Stationarity under Full Participation 39
C.4.1 Drift-explicit stationarity guarantee 39
C.4.2 Explicit finite-shot and heterogeneity dependence 40
C.4.3 Constant-schedule specialization 41
C.5 Partial Client Participation 41
C.5.1 Sampling Model and Potential-Trajectory Coupling 41
C.5.2 Finite-Population Identity and Endpoint Deviation 42
C.5.3 One-Round Descent under Partial Participation 45
C.5.4 Nonconvex Stationarity under Partial Participation 46
C.6 Bias-Aware and Time-Varying Extensions 47
C.6.1 Gradient bias under noisy hardware 47
C.6.2 Time-Varying Learning Rate and Proximal Weight 47
C.6.3 Barren plateaus and shot complexity 48

D ADDITIONAL EXPERIMENTAL RESULTS **49**
D.1 Experimental Configurations 49
D.2 Finite-Shot Parameter-Shift Diagnostics 50
D.2.1 Controlled diagnostic protocol 50
D.2.2 Measurement cost 51

D.3 Sensitivity to the Shot Budget 52
D.4 Theory-Aligned Proximal-SGD Diagnostics 52
D.4.1 Endpoint Client Drift 53
D.4.2 Paired Drift Analysis 53
D.4.3 Stationarity and predictive diagnostics 54
D.5 Binary MNIST IID Check 55
D.6 MedNIST Medical-Image Experiments 57
D.7 Comparison with Centralized and Local VQC Baselines 62
D.8 Experimental Scope and Limitations 63

A RELATED WORK

Federated optimization under statistical heterogeneity. FL enables distributed model training across decentralized clients without exchanging raw data McMahan et al. (2017). A central difficulty is statistical heterogeneity: when local data distributions are non-IID, repeated local updates may cause client models to drift away from the broadcast global model and from one another. FedProx addresses this issue by augmenting each local objective with a proximal term that penalizes deviation from the current global iterate Li et al. (2020a). A broad classical FL literature has further investigated communication-efficient optimization Bouacida et al. (2021); Wu et al. (2022); Akhtarshenas et al. (2024), model compression and quantization Li et al. (2020b), secure aggregation Bonawitz et al. (2017), differential privacy Rofougaran et al. (2023), and personalized FL.

Our work adopts the smooth nonconvex federated-optimization perspective but focuses on a different source of stochasticity introduced by finite-shot quantum measurements. The main theoretical question is *not merely* whether client objectives are heterogeneous, but *how statistical gradient dissimilarity propagates through multiple finite-shot local updates and proximal client trajectories before aggregation*.

Variational quantum learning and gradient estimation. Variational quantum algorithms use parameterized quantum circuits whose parameters are optimized by a classical procedure from quantum measurement outcomes Cerezo et al. (2021). Representative quantum-learning architectures include variational quantum classifiers (VQCs), quantum neural networks, and quantum convolutional models Cong et al. (2019); Beer et al. (2020); Massoli et al. (2022); Jerbi et al. (2023); Shi et al. (2023). For many commonly used parameterized gates, circuit-expectation derivatives can be evaluated exactly through parameter-shift identities. On physical quantum devices or shot-based simulators, however, each shifted expectation value must be estimated using a finite number of measurements, making the resulting derivative estimate stochastic.

Finite-measurement and hardware-noise effects in variational quantum optimization have been studied outside the federated setting. Jose & Simeone (2022), for example, analyze stochastic optimization of parameterized quantum circuits in the presence of quantum gate noise and quantum error mitigation, characterizing the resulting convergence and bias–variance tradeoffs. Other work has similarly examined the role of sampling complexity, measurement allocation, and noisy gradient estimation in variational quantum optimization. Accordingly, the inverse-shot behavior of parameter-shift estimation is not itself the novelty of this work. Our analysis instead retains its explicit dependence on the number of parameter-shift coordinates and studies how finite-shot uncertainty is propagated by federated local-update dynamics.

Quantum federated learning. Quantum federated learning (QFL) extends federated optimization to quantum or hybrid quantum–classical models. Early studies established basic protocols for collaborative quantum-model training Chen & Yoo (2021); Xia & Li (2021), while subsequent work developed frameworks such as FedQNN and related schemes for federated training of shallow quantum models Innan et al. (2024); Qiao et al. (2025). Other QFL studies have investigated privacy and differential privacy Rofougaran et al. (2023), Byzantine robustness Xia et al. (2021), communication or systems aspects, and application-oriented deployments such as biomedical learning and fraud detection Sünkel et al. (2024); Innan et al. (2025). These studies establish that QFL is a broader research area than the specific optimization problem considered here. Our contribution is therefore not the introduction of federated VQC training, finite-shot measurements, or heterogeneous

client data. We instead focus on a convergence analysis that tracks how these ingredients interact through client trajectories under multiple local updates and proximal regularization.

Heterogeneity and personalization in QFL. Heterogeneity has also been studied explicitly in recent QFL work. Rahman et al. (2026) introduce an integrated sporadic-personalized QFL framework designed to address both quantum-device noise heterogeneity and non-IID data distributions. Their sporadic component accounts for heterogeneous quantum-device quality, while personalized local regularization is used to mitigate data heterogeneity, and the resulting algorithm is accompanied by convergence analysis.

Our setting is complementary. We do not model device-quality heterogeneity inside the main unbiased-oracle theory, nor do we claim that proximal or personalized regularization is new to QFL. Instead, we isolate statistical heterogeneity through the gradient-dissimilarity condition and characterize how this dispersion enters the convergence analysis through the subsequent client trajectories. Systematic hardware-induced bias is treated separately in our bias-aware theoretical extension rather than being identified with the statistical heterogeneity parameter ζ.

Pokharel et al. (2026) study distributed quantum learning under non-IID data, finite measurement budgets, and partial device participation. Their work derives a convergence guarantee with explicit dependence on measurement shots and participating subset size and employs fidelity-based partial device selection. Their analysis assumes smoothness together with a Polyak–Łojasiewicz (PL) condition on the global objective and a bounded gradient-diversity condition, while the local optimization follows an SGD/FedAvg-style update.

Although the two settings share several problem ingredients, distributed variational quantum optimization, finite-shot stochasticity, heterogeneous data, and partial participation, the analytical objectives and proof structures are different. Pokharel et al. (2026) analyze objective convergence under PL geometry, whereas we establish first-order stationarity for a smooth, lower-bounded nonconvex objective without a PL assumption. Our analysis further develops an explicit trajectory-based treatment of proximal local dynamics: statistical heterogeneity enters through client-drift evolution after the broadcast iterate, finite-shot stochasticity contributes both directly and through these trajectories, and uniform client subsampling induces a finite-population correction coupled to terminal drift.

First, we establish a first-order stationarity guarantee for a smooth, lower-bounded nonconvex global objective without requiring a Polyak–Łojasiewicz (PL) condition. Accordingly, the convergence target is a learning-rate-weighted squared gradient norm, rather than an objective-gap guarantee derived through PL geometry.

Second, we explicitly characterize proximal local dynamics. Defining the client displacement from the broadcast model as $\mathbf{d}_k^{(t,\tau)} = \boldsymbol{\theta}_k^{(t,\tau)} - \boldsymbol{\theta}^{(t)}$, the proximal recursion introduces the local-memory retention factor $1-\eta_t\mu_t$. Unrolling this recursion yields a trajectory-level drift analysis in which proximal regularization controls how earlier local stochastic updates propagate to subsequent local iterates. Thus, the proximal coefficient acts through the local trajectory dynamics rather than as a standalone regularization-error term.

Third, our treatment of statistical heterogeneity is trajectory based. At the common broadcast iterate, the weighted first-order client-gradient deviations cancel exactly at the broadcast model, although their squared dispersion can remain nonzero. Once clients perform local updates, heterogeneous objectives generate different trajectories; consequently, ζ^2 enters the convergence analysis by controlling the evolution of client-drift energy along these trajectories, rather than through an additional common-iterate gradient bias.

Fourth, finite-shot stochasticity enters the global convergence bound through two distinct channels. The direct stochastic-oracle variance has the dimension-aware form $\sigma_{k,t}^2 = \frac{\sigma_{\mathrm{mb},k}^2}{B} + \frac{c_{\mathrm{ps},k} d_q}{S_{k,t}}$ while oracle noise also perturbs the local iterates and therefore propagates indirectly through client drift. Thus, for fixed d_q, finite-shot effects retain inverse-shot dependence but do not reduce to an unconditional standalone $O(1/S)$ optimization floor.

Finally, under uniform size-m sampling without replacement, the partial-participation analysis yields the finite-population factor $c_m = \frac{K-m}{m(K-1)}$. The resulting participation residual is proportional to the terminal client-drift energy, $c_m \overline{D}_{t,\mathcal{T}}$, thereby separating client-subsampling fluctuations from the

full-participation optimization terms. Since $c_K = 0$, the partial-participation guarantee recovers the full-participation result exactly when $m = K$.

Positioning of the present work. Prior studies establish important results on finite-shot quantum optimization, heterogeneous QFL, and convergence of distributed quantum learning. Our contribution addresses a complementary analytical regime: *first-order nonconvex stationarity with dimension-aware finite-shot uncertainty, proximal client-trajectory dynamics, statistical gradient heterogeneity, and uniform partial participation*. The resulting analysis tracks how these effects propagate jointly through client drift and also motivates the theory-aligned empirical diagnostics used in our experiments.

B ADDITIONAL DETAILS ON VARIATIONAL QUANTUM CLASSIFIERS

This appendix presents the quantum-model properties used by the convergence analysis. We distinguish carefully between deterministic properties of the circuit expectation, finite-shot estimation of circuit expectations and their parameter-shift derivatives, and assumptions imposed separately on the stochastic gradient of the complete learning objective.

B.1 Data Encoding and Variational Circuit

Definition B.1 (Data embedding)**.** *Let $\mathcal{H} = (\mathbb{C}^2)^{\otimes n}$ denote the Hilbert space of an n-qubit quantum register. Given an input feature vector $\mathbf{x} \in \mathbb{R}^{d_x}$, a data-encoding unitary $\Phi(\mathbf{x})$ maps the computational basis state $|0\rangle^{\otimes n}$ to the quantum state $|\phi(\mathbf{x})\rangle := \Phi(\mathbf{x})\,|0\rangle^{\otimes n} \in \mathcal{H}$.*

The data-encoding map $\Phi(\cdot)$ is shared across all clients, ensuring a common feature-embedding scheme throughout the federated system. Thus, while the input data are client-specific, the embedding map remains common to all clients.

Assumption B.2 (Bounded trainable generators)**.** *Each trainable gate in the ansatz is of the form $U_j(\theta_j) = \exp(-i\theta_j G_j)$, where G_j is a Hermitian generator. Since G_j is Hermitian, its eigenvalues $\lambda_r(G_j)$ are real, and its spectral norm satisfies $\|G_j\|_2 = \max_r |\lambda_r(G_j)|$. There exists $G_{\max} < \infty$, independent of the gate index j, such that $\|G_j\|_2 \le G_{\max}$ for every trainable gates j.*

Assumption B.2 imposes a uniform bound on the generators of all trainable gates, thereby controlling the sensitivity of the quantum circuit with respect to variations in its parameters. In particular, this assumption provides a uniform bound on parameter derivatives and facilitates the derivation of bounded gradient estimators under finite-shot measurements.

Lemma B.3 (Parameterized-gate derivative)**.** *Let $G_j \in \mathbb{C}^{m\times m}$, $G_j = G_j^\dagger$, acting on a finite-dimensional Hilbert space, and define the parameterized unitary $U_j(\theta_j) = \exp(-i\theta_j G_j), \theta_j \in \mathbb{R}$. Then $U_j(\theta_j)$ is unitary for all trainable θ_j, and its derivatives satisfy:*

$$\frac{\partial U_j(\theta_j)}{\partial \theta_j} = -iG_j U_j(\theta_j), \quad \frac{\partial U_j(\theta_j)^\dagger}{\partial \theta_j} = iU_j(\theta_j)^\dagger G_j. \tag{7}$$

Proof of Lemma B.3. Since $G_j = G_j^\dagger$, we have $U_j(\theta_j)^\dagger = \exp(i\theta_j G_j)$, and therefore

$$U_j(\theta_j)^\dagger U_j(\theta_j) = e^{i\theta_j G_j} e^{-i\theta_j G_j} = I, \tag{8}$$

which shows that $U_j(\theta_j)$ is unitary. To compute the derivative, we use the power-series expansion of the matrix exponential, $U_j(\theta_j) = \sum_{m=0}^{\infty} \frac{(-i\theta_j G_j)^m}{m!}$. By differentiating $U_j(\theta_j)$ with respect to θ_j, we can obtain the derivative as follows:

$$\frac{\partial U_j(\theta_j)}{\partial \theta_j} = \sum_{m=1}^{\infty} \frac{(-i)^m m \theta_j^{m-1} G_j^m}{m!} = -iG_j \sum_{m=1}^{\infty} \frac{(-i\theta_j G_j)^{m-1}}{(m-1)!} \tag{9a}$$

$$= -iG_j \sum_{m=0}^{\infty} \frac{(-i\theta_j G_j)^m}{m!} = -iG_j U_j(\theta_j). \tag{9b}$$

Similarly, $\frac{\partial U_j(\theta_j)^\dagger}{\partial \theta_j} = iU_j(\theta_j)^\dagger G_j$. This completes the proof. □

Lemma B.4 (Generator norm and Lipschitz continuity). *Let $U_j(\theta_j) = \exp(-i\theta_j G_j)$ where G_j is Hermitian. Then the derivative of $U_j(\theta_j)$ satisfies*

$$\left\| \frac{\partial U_j(\theta_j)}{\partial \theta_j} \right\|_2 = \|G_j\|_2. \tag{10}$$

Moreover, $U_j(\theta_j)$ is Lipschitz continuous with respect to θ_j under the spectral norm, satisfying

$$\|U_j(\theta_j) - U_j(\theta_j')\|_2 \leq \|G_j\|_2 |\theta_j - \theta_j'|, \quad \forall \theta_j, \theta_j' \in \mathbb{R}. \tag{11}$$

Under Assumption B.2, this further implies the uniform bound

$$\|U_j(\theta_j) - U_j(\theta_j')\|_2 \leq G_{\max} |\theta_j - \theta_j'|, \quad \forall \theta_j, \theta_j' \in \mathbb{R}. \tag{12}$$

Proof of Lemma B.4. From Lemma B.3, $\frac{\partial U_j(\theta_j)}{\partial \theta_j} = -iG_j U_j(\theta_j)$. Taking the spectral norm yields:

$$\left\| \frac{\partial U_j(\theta_j)}{\partial \theta_j} \right\|_2 = \left\| G_j U_j(\theta_j) \right\|_2. \tag{13}$$

Since $U_j(\theta_j)$ is unitary and the spectral norm is unitarily invariant, which proves (10).

For the Lipschitz bound, express the difference between the operators as an integral of their derivative:

$$U_j(\theta_j) - U_j(\theta_j') = \int_{\theta_j'}^{\theta_j} \frac{\partial U_j(s)}{\partial s} ds. \tag{14}$$

Taking the spectral norm on both sides and applying the Minkowski inequality for integrals:

$$\left\| U_j(\theta_j) - U_j(\theta_j') \right\|_2 \leq \int_{\theta_j'}^{\theta_j} \left\| \frac{\partial U_j(s)}{\partial s} \right\|_2 ds = \int_{\theta_j'}^{\theta_j} \|G_j\|_2 \, ds = \|G_j\|_2 \left| \theta_j - \theta_j' \right|. \tag{15}$$

By Assumption B.2, $\|G_j\|_2 \leq G_{\max}$, completing the proof. □

Lemma B.4 establishes that the spectral norm of its generator directly controls the sensitivity of each parameterized gate to variations in its trainable parameter. Combined with Assumption B.2, the response of every trainable gate to parameter perturbations is therefore uniformly bounded by $G_{\max}$. This result provides a key regularity condition for the variational circuit, as it prevents arbitrarily large changes in the gate operation under small parameter perturbations. Moreover, it serves as a building block for deriving Lipschitz and smoothness bounds for the overall quantum model and its gradients, which are subsequently used in the convergence analysis.

Lemma B.5 (Pauli-generator specialization). *Let $P = \bigotimes_{q=1}^{n} \sigma_{\ell_q}$ be an n-qubit Pauli string with $\ell_q \in \{0, x, y, z\}$, where $\sigma_0 = I$. Then, $P^2 = I$ and $\operatorname{spec}(P) \subseteq \{-1, +1\}$, so $\|P\|_2 = 1$. More generally, for any $\alpha \in \mathbb{R}$, if $G_j = \alpha P$, then $\|G_j\|_2 = |\alpha|$. In particular, single-qubit rotation $U_j(\theta_j) = \exp\big(-i\frac{\theta_j}{2}\sigma_\alpha\big)$ has $\|G_j\|_2 = 1/2$ for $\ell \in \{x, y, z\}$.*

Proof of Lemma B.5. Each Pauli matrix σ_ℓ, $\ell \in \{x, y, z\}$, is Hermitian and satisfies $\sigma_\ell^2 = I$. Tensor products preserve Hermiticity and multiplication, therefore, for $P = \bigotimes_{q=1}^{n} \sigma_{\ell_q}$,

$$P^\dagger = \bigotimes_{q=1}^{n} \sigma_{\ell_q}^\dagger = \bigotimes_{q=1}^{n} \sigma_{\ell_q} = P, \quad P^2 = \bigotimes_{q=1}^{n} \sigma_{\ell_q}^2 = \bigotimes_{j=1}^{n} I = I. \tag{16}$$

Thus every eigenvalue λ of P satisfies $\lambda^2 = 1$, so $\operatorname{spec}(P) \subseteq \{-1, +1\}$. Since P is Hermitian,

$$\|P\|_2 = \max_{\lambda \in \operatorname{spec}(P)} |\lambda| = 1. \tag{17}$$

If $G_j = \alpha P$, then the eigenvalues of G_j are $\{\alpha\lambda : \lambda \in \operatorname{spec}(P)\}$, and therefore

$$\|G_j\|_2 = \max_{\lambda \in \operatorname{spec}(P)} |\alpha\lambda| = |\alpha|. \tag{18}$$

The special case $G_j = \sigma_\ell/2$ gives $\|G_j\|_2 = 1/2$, which completes the proof. □

Lemma B.5 verifies Assumption B.2 for Pauli-string generators, which are commonly employed in hardware-efficient VQCs. Since each G_j is Hermitian, the corresponding operator $U_j(\theta_j) = \exp(-i\theta_j G_j)$ is unitary for every $\theta_j \in \mathbb{R}$. Beyond ensuring boundedness of the gate generators, Assumption B.2 provides several useful properties for the analysis and optimization of quantum circuits: (i) *First, it imposes a uniform upper bound on the sensitivity of trainable gates to their respective parameters, preventing individual gates from exhibiting arbitrarily large parameter derivatives.* (ii) *Second, for normalized Pauli-string generators, it provides a consistent parameterization across trainable gates, such that comparable changes in the circuit parameters induce uniformly bounded changes in the corresponding quantum operations.* (iii) *Third, it establishes a uniform Lipschitz bound for the parameterized gates, thereby limiting their sensitivity to parameter perturbations and providing a useful regularity condition for subsequent gradient and convergence analyses.*

To enhance the expressive capacity of the local VQC models, each layer of parameterized single-qubit rotations is followed by an entangling unitary $U_{\text{ent}}^{(\ell)}$, typically constructed using two-qubit gates such as controlled-NOT (CNOT) or controlled-Z (CZ) gates. The entangling operation generates nonseparable correlations among qubits, allowing the circuit to capture feature interactions beyond those accessible through independent single-qubit operations. For instance, a ring-entanglement architecture at layer ℓ can be implemented using

$$U_{\text{ent}}^{(\ell)} = \text{CNOT}_{n\to 1} \prod_{q=1}^{n-1} \text{CNOT}_{q\to q+1} \tag{19}$$

where the periodic indexing connects the final qubit back to the first, thereby forming a closed entanglement ring. Let the trainable unitary associated with layer ℓ at client k be defined as

$$U_{\text{train}}^{(\ell)}(\boldsymbol{\theta}_k^{(\ell)}) := \prod_{j=1}^{p_\ell} U_j^{(\ell)}(\theta_{k,j}^{(\ell)}), \tag{20}$$

where p_ℓ denotes the number of trainable gates in layer ℓ and $\boldsymbol{\theta}_k^{(\ell)} = [\theta_{k,1}^{(\ell)}, \ldots, \theta_{k,p_\ell}^{(\ell)}]^\top$ represents the corresponding trainable parameters. By stacking L such variational layers, the local VQC can successively transform the encoded quantum state through alternating trainable and entangling operations. Accordingly, the complete variational circuit for client k is defined as

$$\begin{aligned} U(\boldsymbol{\theta}_k) &= \prod_{\ell=1}^{L} \left[U_{\text{ent}}^{(\ell)} U_{\text{train}}^{(\ell)}(\boldsymbol{\theta}_k^{(\ell)}) \right] \\ &= \underbrace{U_{\text{ent}}^{(L)} U_{\text{train}}^{(L)}(\boldsymbol{\theta}_k^{(L)})}_{\text{Layer } L} \cdots \underbrace{U_{\text{ent}}^{(2)} U_{\text{train}}^{(2)}(\boldsymbol{\theta}_k^{(2)})}_{\text{Layer 2}} \underbrace{U_{\text{ent}}^{(1)} U_{\text{train}}^{(1)}(\boldsymbol{\theta}_k^{(1)})}_{\text{Layer 1}}, \end{aligned} \tag{21}$$

where $\boldsymbol{\theta}_k = [(\boldsymbol{\theta}_k^{(1)})^\top, \ldots, (\boldsymbol{\theta}_k^{(L)})^\top]^\top$ denotes the complete trainable parameter vector of client k. Products are interpreted in circuit-application order, with the rightmost operator acting first.

Lemma B.6 (Unitarity preservation)**.** *Let $U(\boldsymbol{\theta}_k)$ be the VQC ansatz, composed of parameterized gates $U_j(\theta_{k,j}) = \exp(-i\theta_{k,j}G_j)$ with Hermitian G_j, and fixed entangling gates such as CNOT or CZ. Then $U(\boldsymbol{\theta}_k)$ is unitary and $U(\boldsymbol{\theta}_k)^\dagger U(\boldsymbol{\theta}_k) = I$. Consequently, the client-k variational state $|\psi(\mathbf{x}, \boldsymbol{\theta}_k)\rangle = U(\boldsymbol{\theta}_k)\Phi(\mathbf{x})\,|0\rangle^{\otimes n}$ satisfies $\big\| \, |\psi(\mathbf{x}, \boldsymbol{\theta}_k)\rangle \, \big\|_2 = 1$.*

Proof of Lemma B.6. Each trainable gate in the variational circuit is defined as $U_j(\theta_{k,j}) = e^{-i\theta_{k,j}G_j}$, where G_j is a Hermitian generator satisfying $G_j^\dagger = G_j$. Since the adjoint of the exponential of a Hermitian satisfies $\left(e^{-i\theta_{k,j}G_j}\right)^\dagger = e^{i\theta_{k,j}G_j}$, it follows that:

$$U_j^\dagger(\theta_{k,j})\, U_j(\theta_{k,j}) = e^{i\theta_{k,j}G_j} e^{-i\theta_{k,j}G_j} = I. \tag{22}$$

Hence, every parameterized rotation gate $U_j(\theta_{k,j})$ is unitary. Additionally, the entangling operation $U_{\text{ent}}^{(\ell)}$ is constructed as a finite product of two-qubit gates chosen from $\{\text{CNOT}, \text{CZ}\}$. Both CNOT and CZ are standard gates represented by unitary matrices, i.e., $\text{CNOT}^\dagger\text{CNOT} = I$ and $\text{CZ}^\dagger\text{CZ} = I$, thus each entangling operation $U_{\text{ent}}^{(\ell)}$ is unitary, satisfying $U_{\text{ent}}^{(\ell)\dagger} U_{\text{ent}}^{(\ell)} = I$. Because the product of

unitary operators is itself unitary, the entire circuit, comprising both trainable and entangling blocks, remains unitary. Specifically, for the full variational circuit in (21), taking the adjoint of $U(\boldsymbol{\theta}_k))$ yields the reversed product of the adjoints of all gates, as follows:

$$U^\dagger(\boldsymbol{\theta}_k) = \prod_{\ell=L}^{1} \left[\left(\prod_{j=1}^{p_\ell} U_j^{(\ell)}(\theta_{k,j}) \right)^\dagger U_{\text{ent}}^{(\ell)\dagger} \right]. \tag{23}$$

Multiplying the adjoint by the original circuit gives:

$$U^\dagger(\boldsymbol{\theta}_k)\, U(\boldsymbol{\theta}_k) = \underbrace{\prod_{\ell=L}^{1} \left[\left(\prod_{j=1}^{p_\ell} U_j^{(\ell)}(\theta_{k,j}) \right)^\dagger U_{\text{ent}}^{(\ell)\dagger} \right]}_{U^\dagger} \underbrace{\prod_{\ell=1}^{L} \left[U_{\text{ent}}^{(\ell)} \prod_{j=1}^{p_\ell} U_j^{(\ell)}(\theta_{k,j}) \right]}_{U} = I \tag{24}$$

Therefore, the full variational circuit $U(\boldsymbol{\theta}_k)$ is unitary. Moreover, by Definition B.1, the data-encoding unitary $\Phi^{(k)}(\mathbf{x})$ also satisfies $\Phi^{(k)}(\mathbf{x})^\dagger \Phi^{(k)}(\mathbf{x}) = I$. The composition $U(\boldsymbol{\theta}_k)\, \Phi^{(k)}(\mathbf{x})$ is a unitary transformation. Consequently, for any quantum state $|\psi\rangle \in \mathcal{H}$, the transformed state is expressed as:

$$\left\| U(\boldsymbol{\theta}_k) \Phi^{(k)}(\mathbf{x}) \, |\psi\rangle \right\|_2 = \left\| \, |\psi\rangle \, \right\|_2, \tag{25}$$

which indicates that the evolution of the quantum state within each local VQC model preserves both the norm and the total probability, which completes the proof. □

Lemma B.6 ensures that the VQC evolution is norm-preserving, so measurement probabilities remain well-defined at every client and every communication round.

B.2 Measurement Model and Predictive Readout

Definition B.7 (Measurement)**.** *Let M be a Hermitian observable acting on the n-qubit Hilbert space $\mathcal{H}$. Given an input $\mathbf{x}_i^{(k)}$, client k prepares the variational state $|\psi(\mathbf{x}_i^{(k)}, \boldsymbol{\theta}_k)\rangle$. The VQC output is given by the measurement $\langle M \rangle$:*

$$f_{\boldsymbol{\theta}_k}(\mathbf{x}_i^{(k)}) := \langle \psi(\mathbf{x}_i^{(k)}, \boldsymbol{\theta}_k)| \, M \, |\psi(\mathbf{x}_i^{(k)}, \boldsymbol{\theta}_k)\rangle \, . \tag{26}$$

Substituting the variational state, the local VQC model output for each client k is expressed as:

$$f_{\boldsymbol{\theta}_k}(\mathbf{x}_i^{(k)}) = \langle 0|^{\otimes n} \, \Phi^{(k)}(\mathbf{x}_i^{(k)})^\dagger \, U^\dagger(\boldsymbol{\theta}_k) M \, U(\boldsymbol{\theta}_k) \, \Phi^{(k)}(\mathbf{x}_i^{(k)}) \, |0\rangle^{\otimes n} \, . \tag{27}$$

The expectation value (27) can equivalently be expressed in trace form by introducing the density operator of the encoded input state, defined by:

$$\rho_{\text{in}}^{(k)}(\mathbf{x}_i^{(k)}) := |\phi^{(k)}(\mathbf{x}_i^{(k)})\rangle \, \langle \phi^{(k)}(\mathbf{x}_i^{(k)})| \, . \tag{28}$$

Hence, the measurement result, can be then expressed in the Heisenberg-representation as:

$$f_{\boldsymbol{\theta}_k}(\mathbf{x}_i^{(k)}) = \text{Tr} \left[\rho_{\text{in}}^{(k)}(\mathbf{x}_i^{(k)}) \, M_{\boldsymbol{\theta}_k} \right], \tag{29}$$

where $M_{\boldsymbol{\theta}_k} := U^\dagger(\boldsymbol{\theta}_k) M U(\boldsymbol{\theta}_k)$ corresponds to the Heisenberg -representation operator under the variational circuit $U(\boldsymbol{\theta}_k)$. The trace form provides a general representation of the quantum measurement, which is suitable for both pure and mixed states, and enables a convenient expression for gradient derivation in hybrid optimization frameworks. The expression (27) shows how the data-encoding unitary $\Phi^{(k)}(\mathbf{x})$ embeds classical features into the Hilbert space, while the parameterized unitary $U(\boldsymbol{\theta}_k)$ learns an adaptive transformation that captures complex nonlinear relationships in the data. The measurement of M then projects the resulting quantum state into a scalar value, serving as the model's prediction for the input $\mathbf{x}_i^{(k)}$.

Since any Hermitian operator has a real and finite spectrum, the model output always is bounded between the smallest and largest eigenvalues of the observable M, as follows:

$$\lambda_{\min}(M) \leq f_{\boldsymbol{\theta}_k}(\mathbf{x}_i^{(k)}) \leq \lambda_{\max}(M). \tag{30}$$

Binary classification. In this work, we choose the measurement observable M as the Pauli-Z operator acting on a designated readout qubit ($M = Z_r$). Since the eigenvalues of Pauli-Z lie in $\{-1, 1\}$, the prediction in (26) satisfies $f_{\boldsymbol{\theta}_k}(\mathbf{x}) \in [-1, 1]$. The corresponding probability map is

$$p_{\boldsymbol{\theta}_k}(y = 1 \mid \mathbf{x}) = \frac{1 + f_{\boldsymbol{\theta}_k}(\mathbf{x})}{2}, \quad p_{\boldsymbol{\theta}_k}(y = 0 \mid \mathbf{x}) = \frac{1 - f_{\boldsymbol{\theta}_k}(\mathbf{x})}{2}. \tag{31}$$

Accordingly, for a binary label $y \in \{0, 1\}$, the binary cross-entropy (BCE) loss is defined as

$$\mathcal{L}_{\mathrm{BCE}}\left(f_{\boldsymbol{\theta}_k}(\mathbf{x}), y\right) = -y \log p_{\boldsymbol{\theta}_k}(y = 1 \mid \mathbf{x}) - (1 - y) \log p_{\boldsymbol{\theta}_k}(y = 0 \mid \mathbf{x}). \tag{32}$$

Since the derivatives of the logarithmic terms become unbounded as the predicted probabilities approach 0 or 1, a probability margin is introduced to ensure bounded loss derivatives. Specifically, for some $\varepsilon \in (0, 1/2)$, the predicted probability is restricted to

$$p_{\boldsymbol{\theta}_k}(y = 1 \mid \mathbf{x}) \in [\varepsilon, 1 - \varepsilon]. \tag{33}$$

Alternatively, to enforce this condition without explicit clipping, we may use the smooth margin-stabilized probability map

$$p_{\boldsymbol{\theta}_k, \varepsilon}(y = 1 \mid \mathbf{x}) := \varepsilon + (1 - 2\varepsilon) \frac{1 + f_{\boldsymbol{\theta}_k}(\mathbf{x})}{2}, \tag{34}$$

which guarantees $p_{\boldsymbol{\theta}_k, \varepsilon}(y \mid \mathbf{x}) \in [\varepsilon, 1 - \varepsilon], y \in \{0, 1\}$.

Multiclass classification. For a general C-class classification problem, the scalar readout used in the binary setting is extended to a vector of quantum expectation values. Specifically, let

$$\mathbf{z}_{\boldsymbol{\theta}_k}(\mathbf{x}) := \begin{bmatrix} f_{\boldsymbol{\theta}_k, 1}(\mathbf{x}) \\ \vdots \\ f_{\boldsymbol{\theta}_k, R}(\mathbf{x}) \end{bmatrix} \in \mathbb{R}^R, \tag{35}$$

where each component is obtained from a readout observable M_r as

$$f_{\boldsymbol{\theta}_k, r}(\mathbf{x}) := \langle \psi_{\boldsymbol{\theta}_k}(\mathbf{x}) | \, M_r \, | \psi_{\boldsymbol{\theta}_k}(\mathbf{x}) \rangle, \quad r = 1, \ldots, R. \tag{36}$$

The resulting quantum readout is passed through a classical affine head to produce the class logits

$$\mathbf{a}_k(\mathbf{x}) = W_k \mathbf{z}_{\boldsymbol{\theta}_k}(\mathbf{x}) + \mathbf{b}_k \in \mathbb{R}^C, \tag{37}$$

where $W_k \in \mathbb{R}^{C \times R}$ and $\mathbf{b}_k \in \mathbb{R}^C$ denote the weight matrix and bias vector, respectively. The corresponding class probabilities are obtained through the softmax map,

$$p_k(y = c \mid \mathbf{x}) = \frac{\exp(a_{k,c}(\mathbf{x}))}{\sum_{s=1}^{C} \exp(a_{k,s}(\mathbf{x}))}, \quad c = 1, \ldots, C. \tag{38}$$

With class label $y \in \{1, \ldots, C\}$, the corresponding sample-wise loss is the multiclass cross-entropy

$$\mathcal{L}_{\mathrm{CE}}\left(\mathbf{a}_k(\mathbf{x}), y\right) = -\log p_k(y \mid \mathbf{x}). \tag{39}$$

The subsequent bounds derived for a scalar quantum expectation $f_{\boldsymbol{\theta}_k}(\mathbf{x})$ can be applied componentwise to each $f_{\boldsymbol{\theta}_k, r}(\mathbf{x})$ in $\mathbf{z}_{\boldsymbol{\theta}_k}(\mathbf{x})$. Consequently, bounds on the individual quantum readouts can be propagated through the affine head and softmax cross-entropy loss to establish corresponding regularity properties for the multiclass model.

B.3 Bounds on Expectation Derivatives

Lemma B.8 (Bounded Expectation Derivatives)**.** *For each client $k \in \mathcal{K}$ with input data $\mathbf{x} \in \mathcal{D}_k$, let M be a Hermitian observable, then, for every trainable parameter $\theta_{k,j}^{(\ell)}$, $f_{\boldsymbol{\theta}_k}(\mathbf{x})$ is differentiable in $\theta_{k,j}^{(\ell)}$ and its satisfies an uniform bound $\left|\frac{\partial}{\partial \theta_{k,j}^{(\ell)}} f_{\boldsymbol{\theta}_k}(\mathbf{x})\right| \leq 2 \|M\|_2 \|G_j^{(\ell)}\|_2$.*

Proof of Lemma B.8. For simplicity, we write $\mathbf{x} = \mathbf{x}_i^{(k)}$ and by definition of the density matrix $\rho_{\text{in}}^{(k)}(\mathbf{x}_i^{(k)})$ in (28), the expectation value can be simply expressed as follows:

$$f_{\boldsymbol{\theta}_k}(\mathbf{x}) = \langle \phi^{(k)}(\mathbf{x}) | U(\boldsymbol{\theta}_k)^\dagger M U(\boldsymbol{\theta}_k) | \phi^{(k)}(\mathbf{x}) \rangle = \operatorname{Tr}\left(\rho_{\text{in}}^{(k)}(\mathbf{x})\, U(\boldsymbol{\theta}_k)^\dagger M U(\boldsymbol{\theta}_k) \right). \tag{40}$$

Hence, we can then write $f_{\boldsymbol{\theta}_k}(\mathbf{x}) = \operatorname{Tr}\big(\rho_{\text{in}}^{(k)}(\mathbf{x}) M_{\boldsymbol{\theta}_k}\big)$. By differentiating $f_{\boldsymbol{\theta}_k}(\mathbf{x})$ with respect to $\theta_{k,j}^{(\ell)}$, we can obtain its derivative as follows:

$$\frac{\partial f_{\boldsymbol{\theta}_k}(\mathbf{x})}{\partial \theta_{k,j}^{(\ell)}} = \operatorname{Tr}\left(\rho_{\text{in}}^{(k)}(\mathbf{x}) \frac{\partial}{\partial \theta_{k,j}^{(\ell)}} \Big(U(\boldsymbol{\theta}_k)^\dagger M U(\boldsymbol{\theta}_k) \Big) \right). \tag{41}$$

Applying the product rule to $U(\boldsymbol{\theta}_k)^\dagger M U(\boldsymbol{\theta}_k)$ in (41), we can derive:

$$\frac{\partial}{\partial \theta_{k,j}^{(\ell)}} \Big(U(\boldsymbol{\theta}_k)^\dagger M U(\boldsymbol{\theta}_k) \Big) = \left(\frac{\partial U(\boldsymbol{\theta}_k)^\dagger}{\partial \theta_{k,j}^{(\ell)}} \right) M U(\boldsymbol{\theta}_k) + U(\boldsymbol{\theta}_k)^\dagger M \left(\frac{\partial U(\boldsymbol{\theta}_k)}{\partial \theta_{k,j}^{(\ell)}} \right). \tag{42}$$

Therefore, (41) can be written as follows:

$$\frac{\partial f_{\boldsymbol{\theta}_k}(\mathbf{x})}{\partial \theta_{k,j}^{(\ell)}} = \operatorname{Tr}\left[\rho_{\text{in}}^{(k)}(\mathbf{x}) \Big(\frac{\partial U(\boldsymbol{\theta}_k)^\dagger}{\partial \theta_{k,j}^{(\ell)}} \Big) M U(\boldsymbol{\theta}_k) \right] + \operatorname{Tr}\left[\rho_{\text{in}}^{(k)}(\mathbf{x})\, U(\boldsymbol{\theta}_k)^\dagger M \Big(\frac{\partial U(\boldsymbol{\theta}_k)}{\partial \theta_{k,j}^{(\ell)}} \Big) \right]. \tag{43}$$

By assumption, each circuit parameter $\theta_{k,j}^{(\ell)}$ is associated with the single gate $U_j^{(\ell)}(\theta_{k,j}^{(\ell)})$ within the structure of the full ansatz $U(\boldsymbol{\theta}_k)$. Let's define $U_{\text{pos}}^{(\ell,j)}(\boldsymbol{\theta}_k)$ as the product of all circuit blocks that occur after $U_j^{(\ell)}(\theta_{k,j}^{(\ell)})$ in the circuit order, and $U_{\text{pre}}^{(\ell,j)}(\boldsymbol{\theta}_k)$ as the product of all circuit blocks that occur before $U_j^{(\ell)}(\theta_{k,j}^{(\ell)})$. Fix a specific trainable gate $U_j^{(\ell)}(\theta_{k,j}^{(\ell)})$ in layer ℓ and position j inside $\prod_{j=1}^{p_\ell} U_j^{(\ell)}(\cdot)$. Then the full unitary can be written as follows:

$$U(\boldsymbol{\theta}_k) = U_{\text{pos}}^{(\ell,j)}(\boldsymbol{\theta}_k)\, U_j^{(\ell)}(\theta_{k,j}^{(\ell)})\, U_{\text{pre}}^{(\ell,j)}(\boldsymbol{\theta}_k). \tag{44}$$

By differentiating $U(\boldsymbol{\theta}_k)$ with respect to $\theta_{k,j}^{(\ell)}$, we can derive its derivative as follows:

$$\frac{\partial U(\boldsymbol{\theta}_k)}{\partial \theta_{k,j}^{(\ell)}} = U_{\text{pos}}^{(\ell,j)}(\boldsymbol{\theta}_k) \frac{\partial U_j^{(\ell)}(\theta_{k,j}^{(\ell)})}{\partial \theta_{k,j}^{(\ell)}} U_{\text{pre}}^{(\ell,j)}(\boldsymbol{\theta}_k). \tag{45}$$

From Lemma B.3, the derivative satisfies $\partial_\theta U_j^{(\ell)} = -i G_j^{(\ell)} U_j^{(\ell)}$. Using the circuit factorization in (45), the derivative of the full ansatz with respect to the parameter $\theta_{k,j}^{(\ell)}$ can therefore be written as:

$$\frac{\partial U(\boldsymbol{\theta}_k)}{\partial \theta_{k,j}^{(\ell)}} = -i U_{\text{pos}}^{(\ell,j)}(\boldsymbol{\theta}_k) G_j^{(\ell)} U_j^{(\ell)}(\theta_{k,j}^{(\ell)}) U_{\text{pre}}^{(\ell,j)}(\boldsymbol{\theta}_k). \tag{46}$$

Let's define a conjugated operator as follows:

$$\widetilde{G}_{k,j}^{(\ell)} := U_{\text{pos}}^{(\ell,j)}(\boldsymbol{\theta}_k)\, G_j^{(\ell)}\, \big(U_{\text{pos}}^{(\ell,j)}(\boldsymbol{\theta}_k) \big)^\dagger. \tag{47}$$

Multiplying the full unitary $U(\boldsymbol{\theta}_k)$ on the right of the conjugated operator in (47), we obtain:

$$\widetilde{G}_{k,j}^{(\ell)}\, U(\boldsymbol{\theta}_k) = \Big(U_{\text{pos}}^{(\ell,j)}(\boldsymbol{\theta}_k)\, G_j^{(\ell)}\, (U_{\text{pos}}^{(\ell,j)}(\boldsymbol{\theta}_k))^\dagger \Big) \Big(U_{\text{pos}}^{(\ell,j)}(\boldsymbol{\theta}_k)\, U_j^{(\ell)}(\theta_{k,j}^{(\ell)})\, U_{\text{pre}}^{(\ell,j)}(\boldsymbol{\theta}_k) \Big) \tag{48}$$

$$= U_{\text{pos}}^{(\ell,j)}(\boldsymbol{\theta}_k)\, G_j^{(\ell)}\, U_j^{(\ell)}(\theta_{k,j}^{(\ell)})\, U_{\text{pre}}^{(\ell,j)}(\boldsymbol{\theta}_k). \tag{49}$$

Combining (46) and (48), the derivative of the full ansatz can be represented $\frac{\partial U(\boldsymbol{\theta}_k)}{\partial \theta_{k,j}^{(\ell)}} = -i \widetilde{G}_{k,j}^{(\ell)} U(\boldsymbol{\theta}_k)$. Taking the adjoint, $\frac{\partial U(\boldsymbol{\theta}_k)^\dagger}{\partial \theta_{k,j}^{(\ell)}} = i U(\boldsymbol{\theta}_k)^\dagger \widetilde{G}_{k,j}^{(\ell)}$, and then substituting the expressions for $\partial U(\boldsymbol{\theta}_k)/\partial \theta_{k,j}^{(\ell)}$ and $\partial U(\boldsymbol{\theta}_k)^\dagger/\partial \theta_{k,j}^{(\ell)}$ into (43), we obtain:

$$\frac{\partial f_{\boldsymbol{\theta}_k}(\mathbf{x})}{\partial \theta_{k,j}^{(\ell)}} = i \operatorname{Tr}\left(\rho_{\text{in}}^{(k)}(\mathbf{x})\, U(\boldsymbol{\theta}_k)^\dagger \big[\widetilde{G}_{k,j}^{(\ell)}, M \big] U(\boldsymbol{\theta}_k) \right). \tag{50}$$

Taking the absolute value and then applying the Hölder's inequality to the trace

$$\left|\frac{\partial f_{\boldsymbol{\theta}_k}(\mathbf{x})}{\partial \theta_{k,j}^{(\ell)}}\right| \leq \left\|\rho_{\text{in}}^{(k)}(\mathbf{x})\right\|_1 \left\|U(\boldsymbol{\theta}_k)^\dagger \big[\widetilde{G}_{k,j}^{(\ell)}, M\big] U(\boldsymbol{\theta}_k)\right\|_2. \tag{51}$$

Since $\rho_{\text{in}}^{(k)}(\mathbf{x})$ is a density matrix, $\|\rho_{\text{in}}^{(k)}(\mathbf{x})\|_1 = \text{Tr}(\rho_{\text{in}}^{(k)}(\mathbf{x})) = 1$. Hence, the derivative of $f_{\boldsymbol{\theta}_k}(\mathbf{x})$ with respect to $\theta_{k,j}^{(\ell)}$ can be bounded:

$$\left|\frac{\partial f_{\boldsymbol{\theta}_k}(\mathbf{x})}{\partial \theta_{k,j}^{(\ell)}}\right| \leq \left\|U(\boldsymbol{\theta}_k)^\dagger \big[\widetilde{G}_{k,j}^{(\ell)}, M\big] U(\boldsymbol{\theta}_k)\right\|_2 = \left\|\big[\widetilde{G}_{k,j}^{(\ell)}, M\big]\right\|_2. \tag{52}$$

Finally, using the triangle inequality and the sub-multiplicative property of the spectral norm:

$$\left|\frac{\partial f_{\boldsymbol{\theta}_k}(\mathbf{x})}{\partial \theta_{k,j}^{(\ell)}}\right| \leq \left\|\widetilde{G}_{k,j}^{(\ell)} M\right\|_2 + \left\|M \widetilde{G}_{k,j}^{(\ell)}\right\|_2 \leq 2\left\|\widetilde{G}_{k,j}^{(\ell)}\right\|_2 \|M\|_2. \tag{53}$$

Since $\widetilde{G}_{k,j}^{(\ell)}$ is a unitary conjugation of $G_j^{(\ell)}$, its spectral norm is invariant under such a transformation. This completes the proof. □

Proposition B.9. *Under the assumptions of Lemma B.8, the following special-case bounds hold:*

1. *If the readout observable M is a Pauli observable, then $\|M\|_2 = 1$ and $\left|\frac{\partial f}{\partial \theta_{k,j}^{(\ell)}}\right| \leq 2\left\|G_j^{(\ell)}\right\|_2$.*
2. *If the operator is $G_j^{(\ell)} = \sigma_\alpha/2$ for some Pauli σ_α, then $\|G_j^{(\ell)}\|_2 = 1/2$ and $\left|\frac{\partial f}{\partial \theta_{k,j}^{(\ell)}}\right| \leq 1$.*

Eq. (50) is useful analytically but need not correspond to a single directly measurable observable in the implemented circuit. Depending on the operators involved, its evaluation may require additional observable decompositions or circuit constructions. The parameter-shift rule used below instead evaluates the derivative directly from shifted expectation values.

Remark B.10. *The coordinatewise derivative bound above controls circuit sensitivity but does not by itself establish a dimension-free smoothness constant for the complete empirical objective. In particular, the smoothness constant of a BCE or multiclass cross-entropy objective also depends on the loss curvature, the number of trainable parameters, the readout map, and the parameter-domain assumptions. Accordingly, the β-smoothness condition used in the convergence analysis is retained as an explicit optimization assumption rather than inferred solely from Assumption B.2.*

B.4 Parameter-Shift Differentiation and Finite-Shot Estimation

Proof of Lemma 3.1. Let Y denote the single-shot outcome of a bounded observable with support $[a, b]$, and suppose S independent shots are used to estimate an expectation value $\widehat{f} = \frac{1}{S}\sum_{s=1}^{S} Y_s$. The unbiasedness of $\widehat{f}$ follows from linearity of expectation as

$$\mathbb{E}[\widehat{f}] = \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}[Y_s] = \mathbb{E}[Y] = f. \tag{54}$$

Since the Y_s are i.i.d., from its independence, we can obtain:

$$\text{Var}(\widehat{f}) = \text{Var}\left(\frac{1}{S}\sum_{s=1}^{S} Y_s\right) = \frac{1}{S^2}\sum_{s=1}^{S} \text{Var}(Y_s) = \frac{1}{S}\text{Var}(Y). \tag{55}$$

Additionally, Popoviciu's variance inequality for a random variable supported on $[a, b]$ yields $\text{Var}(Y) \leq (b-a)^2/4$. Substituting it into $\text{Var}(\widehat{f})$ in (55), we can derive the variance bound.

For the shifted estimator, applying the definition of the PSR and linearity of expectation, we obtain

$$\mathbb{E}[\widehat{\partial_\theta f}] = \lambda_j(\mathbb{E}[\widehat{f}_+] - \mathbb{E}[\widehat{f}_-]) = \lambda_j(f_+ - f_-) = \partial_\theta f. \tag{56}$$

Suppose that the S shots used for $\widehat{f}_+$ and $\widehat{f}_-$ are independent we have:

$$\mathrm{Var}(\widehat{\partial_\theta f}) = \lambda_j^2\big(\mathrm{Var}(\widehat{f}_+) + \mathrm{Var}(\widehat{f}_-)\big) \le \lambda_j^2\Big(\frac{(b-a)^2}{4S} + \frac{(b-a)^2}{4S}\Big) = \frac{\lambda_j^2(b-a)^2}{2S}, \tag{57}$$

For a Pauli measurement, $b-a=2$, while a Pauli-rotation generator has $\lambda_j = 1/2$, yielding $1/(2S)$. This completes the proof. □

From Lemma 3.1, it is obvious that increasing the number of shots improves gradient accuracy, but incurs additional computational resources.

Corollary B.11 (Full-gradient parameter dependence)**.** *Let $\boldsymbol{\theta} \in \mathbb{R}^{d_\theta}$ contain d_θ trainable Pauli-rotation parameters, and estimate every coordinate using S shots per shifted expectation. Then*

$$\mathbb{E}\Big[\Big\|\widehat{\nabla_{\boldsymbol{\theta}} f} - \nabla_{\boldsymbol{\theta}} f\Big\|_2^2\Big] = \sum_{j=1}^{d_\theta} \mathbb{E}\Big[\Big(\widehat{\partial_{\theta_j} f} - \partial_{\theta_j} f\Big)^2\Big] \le \frac{d_\theta}{2S}. \tag{58}$$

Thus, the commonly stated $O(S^{-1})$ dependence is an inverse-shot law for fixed parameter dimension; coordinatewise parameter-shift estimation also contains an explicit linear dependence on d_θ.

For one scalar observable and one sample, a complete coordinatewise parameter-shift gradient uses $2d_\theta$ shifted circuit expectations. Consequently, the direct measurement budget scales as $O(d_\theta S)$ shots per sample, before accounting for mini-batching or multiple readout observables.

Remark B.12. *Lemma 3.1 establishes unbiasedness of the circuit expectation and its parameter-shift derivative. It does* not *by itself imply that a plug-in estimator of the gradient of a nonlinear loss is unbiased. For example, $\nabla_{\boldsymbol{\theta}}\mathcal{L} = \frac{\partial \mathcal{L}}{\partial f}\nabla_{\boldsymbol{\theta}} f$. If both $\partial\mathcal{L}/\partial f$ and $\nabla_{\boldsymbol{\theta}} f$ are evaluated using the same finite-shot estimate $\widehat{f}$, then in general $\mathbb{E}\big[\widehat{\nabla_{\boldsymbol{\theta}}\mathcal{L}}\big] \neq \nabla_{\boldsymbol{\theta}}\mathcal{L}$, for nonlinear losses such as BCE or softmax cross-entropy. Accordingly, the unbiased stochastic-gradient condition used in the optimization analysis is stated separately as an oracle assumption. A bias-aware extension may instead write $\mathbb{E}\left[\mathbf{g}_k(\boldsymbol{\theta},\xi) \mid \boldsymbol{\theta}\right] = \nabla F_k(\boldsymbol{\theta}) + \mathbf{b}_k(\boldsymbol{\theta})$, where $\mathbf{b}_k$ captures systematic bias arising from nonlinear finite-shot plug-in effects or hardware noise.*

Remark B.13. *The preceding finite-shot result assumes that the circuit expectation at each shift is sampled from the circuit model appearing in the objective. On noisy hardware, finite-shot sampling may remain unbiased for the expectation of the implemented noisy circuit, but this expectation can differ systematically from the ideal noiseless objective. Hence, arbitrary gate noise should be modeled through an additional gradient-bias term rather than absorbed into the shot-variance contribution.*

The direct gradient-variance diagnostic reported in the experiments tests the estimator-level prediction in Corollary B.11 by fixing the circuit checkpoint, client, mini-batch, and circuit coordinates while varying only the number of measurement shots.

C ADDITIONAL DETAILS ON THE CONVERGENCE ANALYSIS

We analyze QFL-VQC as a stochastic proximal federated optimization method under smoothness and lower-boundedness conditions. The analysis characterizes four principal mechanisms governing the optimization dynamics: *(i) finite-shot and stochastic mini-batch gradient noise, (ii) heterogeneity-induced local client drift, (iii) time-varying proximal contraction, and (iv) client-sampling fluctuations under partial participation.*

An important distinction is that client heterogeneity has different first- and second-order effects when evaluated at a common broadcast iterate. Since $F(\boldsymbol{\theta}) = \sum_k p_k F_k(\boldsymbol{\theta}), \sum_k p_k = 1$, the weighted first-order deviations cancel exactly, $\sum_{k=1}^K p_k\big[\nabla F_k(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta})\big] = \mathbf{0}$. Their second-order dispersion, however, is generally nonzero $\sum_{k=1}^K p_k\big\|\nabla F_k(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta})\big\|_2^2 \le \zeta^2$. Consequently, heterogeneity does not contribute to the full-participation analysis through a nonzero averaged gradient mismatch at the broadcast iterate. Rather, different client objectives drive different local trajectories, which then generate nonzero client drift. We therefore track the weighted client-drift energy explicitly.

Appendix B establishes the inverse-shot variance scaling of finite-shot parameter-shift estimators at the circuit-expectation level. Unbiasedness of the complete stochastic gradient of the learning

objective is imposed separately below as a stochastic-oracle assumption, because unbiasedness of an expectation-value derivative does not automatically imply unbiasedness of a nonlinear finite-shot loss gradient. We later give a bias-aware extension for hardware noise.

C.1 Analytical Setup and Local Proximal Descent

C.1.1 Additional stochastic-oracle and variance model

Recall at communication round t, the server broadcasts $\boldsymbol{\theta}^{(t)}$. Each participating client initializes $\boldsymbol{\theta}_k^{(t,0)} = \boldsymbol{\theta}^{(t)}$ and performs $\mathcal{T}$ local proximal-SGD updates:

$$\boldsymbol{\theta}_k^{(t,\tau+1)} = \boldsymbol{\theta}_k^{(t,\tau)} - \eta_t \Big[\mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,\tau)}, \boldsymbol{\xi}_k^{(t,\tau)}\big) + \mu_t \big(\boldsymbol{\theta}_k^{(t,\tau)} - \boldsymbol{\theta}^{(t)}\big) \Big], \tag{59}$$

where $\mathbf{g}_k^{(t,\tau)} := \mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,\tau)}, \boldsymbol{\xi}_k^{(t,\tau)}\big)$ is the stochastic gradient estimator and $\boldsymbol{\xi}_k^{(t,\tau)}$ collects mini-batch and finite-shot randomness. The corresponding round-t proximal objective is

$$J_{k,t}(\boldsymbol{\theta}) := F_k(\boldsymbol{\theta}) + \frac{\mu_t}{2} \big\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}\big\|_2^2. \tag{60}$$

For communication round t and local step τ,

$$\mathbf{e}_k^{(t,\tau)} := \mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,\tau)}, \boldsymbol{\xi}_k^{(t,\tau)}\big) - \nabla F_k\big(\boldsymbol{\theta}_k^{(t,\tau)}\big). \tag{61}$$

Assumption C.1 (Stochastic-gradient oracle)**.** *Let $\mathcal{F}_{t,\tau}$ denote the sigma-field generated by the complete optimization history immediately before the oracle evaluations at local step τ of communication round t. We assume $\mathbb{E}\big[\mathbf{g}_k^{(t,\tau)} \mid \mathcal{F}_{t,\tau}\big] = \nabla F_k\big(\boldsymbol{\theta}_k^{(t,\tau)}\big)$, and $\mathbb{E}\big[\big\|\mathbf{e}_k^{(t,\tau)}\big\|_2^2 \mid \mathcal{F}_{t,\tau}\big] \le \sigma_{k,t}^2$.*

The sequence $\{\mathbf{e}_k^{(t,\tau)}\}_\tau$ is therefore a martingale-difference sequence. Full independence between different local steps is not required. For finite-shot parameter-shift gradients, we expose the measurement-budget and parameter-dimension dependence through $\sigma_{k,t}^2 := \frac{\sigma_{\mathrm{mb},k}^2}{B} + \frac{c_{\mathrm{ps},k} d_q}{S_{k,t}}$ where B is the mini-batch size, $S_{k,t}$ is the shot budget, d_q is the number of quantum parameters estimated through coordinatewise parameter-shift measurements, and $c_{\mathrm{ps},k}$ collects bounded loss, observable, and measurement-allocation constants. Define the client-weighted variance $\overline{\sigma}_t^2 := \sum_{k=1}^K p_k \sigma_{k,t}^2$. It makes explicit that the familiar $O(S^{-1})$ law is an inverse-shot dependence for fixed model dimension. A complete coordinate-wise parameter-shift gradient additionally incurs linear dependence on d_q.

Assumption 3.4 characterizes the stochasticity of the local gradient oracle arising from two distinct sources: mini-batch sampling and finite-shot quantum measurement. Specifically, for client k, we write the conditional gradient-variance bound as

$$\mathbb{E}\Big[\big\|\mathbf{g}_k(\boldsymbol{\theta}, \boldsymbol{\xi}) - \nabla F_k(\boldsymbol{\theta})\big\|_2^2 \;\Big|\; \boldsymbol{\theta} \Big] \le \frac{\sigma_{\mathrm{mb},k}^2}{B} + \frac{c_{\mathrm{ps},k} d_q}{S_k}, \tag{62}$$

where $\sigma_{\mathrm{mb},k}^2/B$ represents the variance induced by mini-batch sampling, while $c_{\mathrm{ps},k} d_q / S_k$ represents the finite-shot contribution associated with coordinatewise parameter-shift gradient estimation. Here, B is the mini-batch size, S_k is the number of measurement shots allocated to each shifted circuit evaluation, d_q is the number of trainable quantum parameters estimated through the parameter-shift rule, and $c_{\mathrm{ps},k}$ collects bounded constants associated with the loss, measurement observables, and shot-allocation scheme.

The two terms describe different stochastic mechanisms. Increasing the mini-batch size reduces the classical data-sampling contribution, whereas increasing the shot budget reduces the quantum measurement contribution. In particular, for fixed d_q, $\frac{c_{\mathrm{ps},k} d_q}{S_k} = O(S_k^{-1})$ which is the inverse-shot scaling established at the circuit-estimator level. Increasing S_k does not remove mini-batch variability, and increasing B does not remove finite-shot variability. Under the unbiased stochastic-oracle model, using the full local batch together with $S_k \to \infty$ eliminates these two stochastic contributions and recovers the gradient of the modeled client objective. This statement does not apply to systematic hardware-induced bias, which is treated separately in the bias-aware analysis. In the convergence bound, reducing either source of variance decreases the stochastic stationarity residual rather than changing the deterministic client-drift mechanism itself.

Assumption 3.6 imposes a bounded gradient-dissimilarity condition across client objectives. The constant ζ^2 therefore measures the weighted dispersion of local objective gradients around the global gradient and primarily captures statistical heterogeneity arising from differences among the clients' local data distributions. All client objectives are assumed to be defined over a common parameter space, consistent with federated parameter aggregation. Importantly, the corresponding weighted first-order deviations cancel exactly at a common iterate whereas their squared dispersion is generally nonzero. Consequently, heterogeneity does not contribute to the full-participation analysis through a nonzero averaged gradient mismatch at the broadcast iterate. Instead, it drives different clients along different local trajectories and thereby increases the weighted client-drift energy

$$D_{t,\tau} := \sum_{k=1}^{K} p_k \mathbb{E}\Big[\big\|\boldsymbol{\theta}_k^{(t,\tau)} - \boldsymbol{\theta}^{(t)}\big\|_2^2 \mid \boldsymbol{\theta}^{(t)}\Big]. \tag{63}$$

When all clients have identical objective gradients, $\nabla F_k(\boldsymbol{\theta}) = \nabla F(\boldsymbol{\theta})$ for every k and $\boldsymbol{\theta}$, we have $\zeta = 0$. Larger values of ζ enlarge the theoretical upper bound on heterogeneity-induced client drift and, consequently, the corresponding stationarity residual. Systematic differences between an implemented noisy quantum circuit and the intended ideal objective are not absorbed into ζ^2. Such effects are modeled separately through a gradient-bias term $\mathbf{b}_k(\boldsymbol{\theta})$. Likewise, arbitrary differences in client circuit architectures or parameter dimensions fall outside the common-parameter-space aggregation model considered in this analysis.

Remark C.2. *The condition $\mathbb{E}\big[\big\|\mathbf{g}_k(\boldsymbol{\theta}, \boldsymbol{\xi})\big\|_2^2 \mid \boldsymbol{\theta}\big] \le G^2$ is an upper second-moment bound. It does not lower bound the true gradient norm and therefore does not exclude barren-plateau regimes. The convergence result below is a stationarity result and should not be interpreted as a large-scale VQC trainability guarantee.*

C.1.2 Local proximal dynamics

Before analyzing client drift and global aggregation, we present several standard smoothness properties used throughout the subsequent proofs. These results isolate the deterministic effect of the proximal regularizer from the stochasticity of the local gradient oracle.

Lemma C.3 (Smoothness descent)**.** *Under Assumption 3.3, each local objective function F_k is β_k-smooth, implying that F_k can be locally upper bounded by a quadratic function. That is, for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p$, the following hold:*

$$F_k(\boldsymbol{\theta}') \le F_k(\boldsymbol{\theta}) + \big\langle \nabla F_k(\boldsymbol{\theta}),\, \boldsymbol{\theta}' - \boldsymbol{\theta} \big\rangle + \frac{\beta}{2}\big\|\boldsymbol{\theta}' - \boldsymbol{\theta}\big\|_2^2. \tag{64}$$

Proof of Lemma C.3. Let's define $\boldsymbol{\theta}(\nu) = \boldsymbol{\theta} + \nu(\boldsymbol{\theta}' - \boldsymbol{\theta})$ for $\nu \in \Big[0, 1\Big]$, then we can compute the gap of the local loss F_k of client k when the model parameter moves from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ as follows:

$$F_k\big(\boldsymbol{\theta}'\big) - F_k\big(\boldsymbol{\theta}\big) = \int_0^1 \big\langle \nabla F_k(\boldsymbol{\theta}(\nu)),\, \boldsymbol{\theta}' - \boldsymbol{\theta} \big\rangle \, d\nu. \tag{65}$$

By adding and subtracting $\nabla F_k(\boldsymbol{\theta})$ inside the integral we can express the gap as follows:

$$F_k\big(\boldsymbol{\theta}'\big) - F_k\big(\boldsymbol{\theta}\big) = \big\langle \nabla F_k(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta} \big\rangle + \int_0^1 \big\langle \nabla F_k\big(\boldsymbol{\theta}(\nu)\big) - \nabla F_k(\boldsymbol{\theta}), (\boldsymbol{\theta}' - \boldsymbol{\theta}) \big\rangle \, d\nu. \tag{66}$$

By applying the Cauchy–Schwarz inequality, we can obtain the bound for the gap as follows:

$$F_k\big(\boldsymbol{\theta}'\big) - F_k\big(\boldsymbol{\theta}\big) \le \big\langle \nabla F_k(\boldsymbol{\theta}),\, \boldsymbol{\theta}' - \boldsymbol{\theta} \big\rangle + \int_0^1 \big\|\nabla F_k\big(\boldsymbol{\theta}(\nu)\big) - \nabla F_k(\boldsymbol{\theta})\big\|_2 \big\|\boldsymbol{\theta}' - \boldsymbol{\theta}\big\|_2 \, d\nu. \tag{67}$$

From the β_k-Lipschitz continuity of the gradient (Assumption 3.3) and the definition of $\boldsymbol{\theta}(\nu)$, we can obtain the rate of the change of the gradients bounded as:

$$\big\|\nabla F_k\big(\boldsymbol{\theta}(\nu)\big) - \nabla F_k(\boldsymbol{\theta})\big\|_2 \;\le\; \beta_k\, \nu \, \big\|\boldsymbol{\theta}' - \boldsymbol{\theta}\big\|_2. \tag{68}$$

Substituting (68) into (67) and selecting the Lipschitz constant $\beta = \max_{k \in \mathcal{K}} \beta_k$ over all clients, complete the proof. □

For the global objective, $\nabla F(\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k \nabla F_k(\boldsymbol{\theta})$, it follows that,

$$\left\|\nabla F(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta}')\right\|_2 = \left\| \sum_{k=1}^{K} p_k \left[\nabla F_k(\boldsymbol{\theta}) - \nabla F_k(\boldsymbol{\theta}')\right] \right\|_2 \leq \sum_{k=1}^{K} p_k \left\|\nabla F_k(\boldsymbol{\theta}) - \nabla F_k(\boldsymbol{\theta}')\right\|_2 \quad (69)$$

$$\leq \beta \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_2 \sum_{k=1}^{K} p_k = \beta \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_2. \quad (70)$$

Thus F is also β-smooth, and the same descent lemma applies.

Remark C.4. *Assumption 3.3 is imposed directly on the complete client objectives F_k. The bounded-generator properties of the VQC derived in Appendix B control circuit sensitivity but are not, by themselves, sufficient to establish a dimension-free smoothness constant for an arbitrary nonlinear loss. We therefore retain β-smoothness as an explicit optimization assumption.*

Lemma C.5 (Smoothness of theproximal local objective)**.** *Under Assumption 3.3, the proximal objective $J_{k,t}$ is $(\beta + \mu_t)$-smooth.*

Remark C.6. *The proximal term introduces no stochasticity, but it changes the deterministic smoothness scale of the local objective from β to $\beta + \mu_t$. Consequently, increasing μ_t strengthens the pull toward the broadcast model while simultaneously imposing a more restrictive stable local stepsize. This is one source of the drift–adaptation trade-off appearing in the later convergence analysis.*

For compactness, we express the stochastic gradient of the local proximal objective by

$$\widetilde{\mathbf{g}}_{k,t}^{(t,\tau)} := \mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,\tau)}, \boldsymbol{\xi}_k^{(t,\tau)}\big) + \mu_t\big(\boldsymbol{\theta}_k^{(t,\tau)} - \boldsymbol{\theta}^{(t)}\big). \quad (71)$$

The local update becomes

$$\boldsymbol{\theta}_k^{(t,\tau+1)} = \boldsymbol{\theta}_k^{(t,\tau)} - \eta_t \widetilde{\mathbf{g}}_{k,t}^{(t,\tau)}. \quad (72)$$

The deterministic proximal correction is measurable with respect to $\mathcal{F}_{t,\tau}$. Thus, by Assumption C.1,

$$\mathbb{E}\big[\widetilde{\mathbf{g}}_{k,t}^{(t,\tau)} \mid \mathcal{F}_{t,\tau}\big] = \nabla F_k\big(\boldsymbol{\theta}_k^{(t,\tau)}\big) + \mu_t\big(\boldsymbol{\theta}_k^{(t,\tau)} - \boldsymbol{\theta}^{(t)}\big) = \nabla J_{k,t}\big(\boldsymbol{\theta}_k^{(t,\tau)}\big). \quad (73)$$

Moreover,

$$\widetilde{\mathbf{g}}_{k,t}^{(t,\tau)} - \nabla J_{k,t}\big(\boldsymbol{\theta}_k^{(t,\tau)}\big) = \mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,\tau)}, \boldsymbol{\xi}_k^{(t,\tau)}\big) - \nabla F_k\big(\boldsymbol{\theta}_k^{(t,\tau)}\big) = \mathbf{e}_k^{(t,\tau)}. \quad (74)$$

The proximal changes the mean local direction but introduces no new stochastic-gradient variance.

Lemma C.7 (One-step proximal-SGD descent)**.** *Under Assumptions 3.3 and C.1,*

$$\mathbb{E}\Big[J_{k,t}\big(\boldsymbol{\theta}_k^{(t,\tau+1)}\big)\Big|\mathcal{F}_{t,\tau}\Big] \leq J_{k,t}\big(\boldsymbol{\theta}_k^{(t,\tau)}\big) - \eta_t\Big[1 - \frac{(\beta+\mu_t)\eta_t}{2}\Big]\Big\|\nabla J_{k,t}\big(\boldsymbol{\theta}_k^{(t,\tau)}\big)\Big\|_2^2 + \frac{(\beta+\mu_t)\eta_t^2}{2}\sigma_{k,t}^2. \quad (75)$$

Consequently, if $\eta_t \leq \frac{1}{\beta+\mu_t}$, then

$$\mathbb{E}\left[J_{k,t}\big(\boldsymbol{\theta}_k^{(t,\tau+1)}\big)\Big|\mathcal{F}_{t,\tau}\right] \leq J_{k,t}\big(\boldsymbol{\theta}_k^{(t,\tau)}\big) - \frac{\eta_t}{2}\Big\|\nabla J_{k,t}\big(\boldsymbol{\theta}_k^{(t,\tau)}\big)\Big\|_2^2 + \frac{(\beta+\mu_t)\eta_t^2}{2}\sigma_{k,t}^2. \quad (76)$$

Proof of Lemma C.7. For compactness, let $\boldsymbol{\vartheta} := \boldsymbol{\theta}_k^{(t,\tau)}, L_t := \beta + \mu_t$ and write $\widetilde{\mathbf{g}} := \widetilde{\mathbf{g}}_{k,t}^{(t,\tau)}$. By Lemma C.5 and (72),

$$J_{k,t}\left(\boldsymbol{\vartheta} - \eta_t\widetilde{\mathbf{g}}\right) \leq J_{k,t}(\boldsymbol{\vartheta}) - \eta_t\left\langle\nabla J_{k,t}(\boldsymbol{\vartheta}), \widetilde{\mathbf{g}}\right\rangle + \frac{L_t\eta_t^2}{2}\left\|\widetilde{\mathbf{g}}\right\|_2^2. \quad (77)$$

Taking conditional expectation given $\mathcal{F}_{t,\tau}$ and using (73),

$$\mathbb{E}\left[\left\langle\nabla J_{k,t}(\boldsymbol{\vartheta}), \widetilde{\mathbf{g}}\right\rangle | \mathcal{F}_{t,\tau}\right] = \left\|\nabla J_{k,t}(\boldsymbol{\vartheta})\right\|_2^2. \quad (78)$$

Next, from $\widetilde{\mathbf{g}} = \nabla J_{k,t}(\boldsymbol{\vartheta}) + \mathbf{e}_k^{(t,\tau)}$, we have

$$\left\|\widetilde{\mathbf{g}}\right\|_2^2 = \left\|\nabla J_{k,t}(\boldsymbol{\vartheta})\right\|_2^2 + 2\left\langle\nabla J_{k,t}(\boldsymbol{\vartheta}), \mathbf{e}_k^{(t,\tau)}\right\rangle + \left\|\mathbf{e}_k^{(t,\tau)}\right\|_2^2. \quad (79)$$

Conditioning on $\mathcal{F}_{t,\tau}$ eliminates the cross term since $\mathbb{E}\big[\mathbf{e}_k^{(t,\tau)} \mid \mathcal{F}_{t,\tau}\big] = \mathbf{0}$. Therefore,

$$\mathbb{E}\big[\,\|\widetilde{\mathbf{g}}\|_2^2 \mid \mathcal{F}_{t,\tau}\big] = \|\nabla J_{k,t}(\boldsymbol{\vartheta})\|_2^2 + \mathbb{E}\big[\big\|\mathbf{e}_k^{(t,\tau)}\big\|_2^2 \mid \mathcal{F}_{t,\tau}\big] \tag{80}$$

$$\leq \|\nabla J_{k,t}(\boldsymbol{\vartheta})\|_2^2 + \sigma_{k,t}^2. \tag{81}$$

Substituting (78) and (80) into (77) gives

$$\mathbb{E}\Big[J_{k,t}\big(\boldsymbol{\theta}_k^{(t,\tau+1)}\big) \mid \mathcal{F}_{t,\tau}\Big] \leq J_{k,t}\big(\boldsymbol{\theta}_k^{(t,\tau)}\big) - \eta_t\Big[1 - \frac{L_t\eta_t}{2}\Big]\Big\|\nabla J_{k,t}\big(\boldsymbol{\theta}_k^{(t,\tau)}\big)\Big\|_2^2 + \frac{L_t\eta_t^2}{2}\sigma_{k,t}^2, \tag{82}$$

which proves (75). Finally, if $\eta_t \leq \frac{1}{L_t}$ then $1 - \frac{L_t\eta_t}{2} \geq \frac{1}{2}$ yielding (76). □

Remark C.8. *For the finite-shot quantum oracle, $\sigma_{k,t}^2 = \frac{\sigma_{\mathrm{mb},k}^2}{B} + \frac{c_{\mathrm{ps},k}d_q}{S_{k,t}}$. Therefore, the one-step stochastic residual is $\frac{(\beta+\mu_t)\eta_t^2}{2}\big[\frac{\sigma_{\mathrm{mb},k}^2}{B} + \frac{c_{\mathrm{ps},k}d_q}{S_{k,t}}\big]$. Thus, increasing the shot budget reduces the quantum component of the one-step stochastic error, whereas the deterministic proximal term affects the descent condition through $\beta + \mu_t$.*

Lemma C.7 establishes descent of the round-dependent proximal objective $J_{k,t}$ along an individual client's local trajectory. It does not by itself imply descent of the global federated objective F. When $\mathcal{T} > 1$, local gradients are evaluated at client-specific iterates and the resulting client models can drift away from the broadcast model. The following subsections therefore quantify this local drift before deriving the global server-level stationarity bound.

C.2 Proximal Memory and Client-Drift

We next characterize how the proximal term controls the departure of each local model from the global model broadcast at the beginning of communication. This is important because local gradients are evaluated along the client-specific trajectories $\{\boldsymbol{\theta}_k^{(t,\tau)}\}_{\tau=0}^{\mathcal{T}}$ rather than at the common iterate $\boldsymbol{\theta}^{(t)}$. Consequently, even though the weighted gradient heterogeneity cancels exactly at a common iterate, different local objectives can generate different local trajectories after one or more local updates.

Recall that for client k, the displacement from the round-t broadcast model is $\mathbf{d}_k^{(t,\tau)} := \boldsymbol{\theta}_k^{(t,\tau)} - \boldsymbol{\theta}^{(t)}$. Since every client initializes its local optimization from the broadcast model, $\mathbf{d}_k^{(t,0)} = \mathbf{0}$, using the proximal local update, this drift satisfies

$$\mathbf{d}_k^{(t,\tau+1)} = (1 - \eta_t\mu_t)\,\mathbf{d}_k^{(t,\tau)} - \eta_t\mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,\tau)}, \xi_k^{(t,\tau)}\big), \quad \mathbf{d}_k^{(t,0)} = \mathbf{0}. \tag{83}$$

For convenience, define $q_t := 1 - \eta_t\mu_t$. We assume $0 \leq \eta_t\mu_t \leq 1$, so that $q_t \in [0,1]$. The quantity q_t can be interpreted as a *local-memory retention factor*. When $\mu_t = 0$, we have $q_t = 1$ and recover the ordinary local-SGD recursion used by Q-FedAvg: previous displacements are carried forward without proximal attenuation. When $\mu_t > 0$, $q_t < 1$, and displacement created by earlier local gradients is geometrically attenuated as subsequent proximal steps pull the client model toward $\boldsymbol{\theta}^{(t)}$.

C.2.1 Exact proximal-memory representation

The drift recursion in (83) admits an exact unrolled form that makes explicit how gradients generated at previous local steps propagate into the current client displacement. Recall that η_t and μ_t are round-dependent but are held fixed across the $\mathcal{T}$ local steps within communication round t, so that $q_t = 1 - \eta_t\mu_t$ is constant within the round.

Lemma C.9 (Exact proximal-memory representation)**.** *For every client k, communication round t, and local step $\tau \geq 1$, the local drift satisfies*

$$\mathbf{d}_k^{(t,\tau)} = -\eta_t \sum_{s=0}^{\tau-1} q_t^{\tau-1-s}\mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,s)}, \boldsymbol{\xi}_k^{(t,s)}\big). \tag{84}$$

Proof of Lemma C.9. We proceed by induction on the local-step index τ.

For $\tau = 1$, using $\mathbf{d}_k^{(t,0)} = \mathbf{0}$ in the recursion (83) yields $\mathbf{d}_k^{(t,1)} = -\eta_t\mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,0)}, \boldsymbol{\xi}_k^{(t,0)}\big)$ which coincides with (84) for $\tau = 1$.

Now suppose that (84) holds for some $\tau \geq 1$. Substituting the induction hypothesis into (83) yields

$$\mathbf{d}_k^{(t,\tau+1)} = q_t \mathbf{d}_k^{(t,\tau)} - \eta_t \mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,\tau)}, \boldsymbol{\xi}_k^{(t,\tau)}\big) \tag{85}$$

$$= -\eta_t \sum_{s=0}^{\tau-1} q_t^{\tau-s} \mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,s)}, \boldsymbol{\xi}_k^{(t,s)}\big) - \eta_t \mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,\tau)}, \boldsymbol{\xi}_k^{(t,\tau)}\big) \tag{86}$$

$$= -\eta_t \sum_{s=0}^{\tau} q_t^{\tau-s} \mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,s)}, \boldsymbol{\xi}_k^{(t,s)}\big). \tag{87}$$

This is precisely 84 with τ replaced by $\tau + 1$. Hence the representation holds for every $\tau \geq 1$. □

Lemma C.9 shows that the local displacement is a geometrically weighted accumulation of the stochastic gradients generated earlier in the same communication round. Specifically, the gradient computed at local step s contributes to the displacement at step τ through the kernel weight

$$q_t^{\tau-1-s} = (1 - \eta_t \mu_t)^{\tau-1-s}. \tag{88}$$

Hence the effect of a gradient depends on its temporal distance from the current local step.

When $\mu_t = 0$, we have $q_t = 1$, and (84) reduces to $\mathbf{d}_k^{(t,\tau)} = -\eta_t \sum_{s=0}^{\tau-1} \mathbf{g}_k^{(t,s)}$ which is the usual accumulation produced by local SGD in Q-FedAvg. In contrast, whenever $\mu_t > 0$ and $0 < \eta_t \mu_t \leq 1$, we have $0 \leq q_t < 1$, and the contribution of a gradient generated r local steps in the past is multiplied by q_t^r. Thus the proximal term introduces a finite-memory effect: recent gradients receive larger temporal weights, while older gradients are progressively attenuated.

For example, after three local steps,

$$\mathbf{d}_k^{(t,3)} = -\eta_t \Big[q_t^2 \mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,0)}, \boldsymbol{\xi}_k^{(t,0)}\big) + q_t \mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,1)}, \boldsymbol{\xi}_k^{(t,1)}\big) + \mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,2)}, \boldsymbol{\xi}_k^{(t,2)}\big) \Big]. \tag{89}$$

For $\mu_t = 0$, all three temporal weights equal one. For $\mu_t > 0$, the kernel weights satisfy $q_t^2 \leq q_t \leq 1$ so earlier gradient contributions are attenuated more strongly. Importantly, Lemma C.9 characterizes the propagation of a given stochastic-gradient sequence; by itself, it does not yet compare the realized FedAvg and FedProx trajectories, since the gradients $\mathbf{g}_k^{(t,s)}$ are evaluated at algorithm-dependent local iterates. The subsequent drift bound converts this geometric-memory representation into a rigorous second-moment envelope for $\|\mathbf{d}_k^{(t,\tau)}\|_2^2$.

C.2.2 Geometric memory factors

The exact representation in Lemma C.9 shows that proximal local optimization induces the geometric temporal kernel $\big(1, q_t, q_t^2, \ldots, q_t^{\tau-1}\big), q_t = 1 - \eta_t \mu_t$. Two different norms of this kernel naturally arise in the subsequent analysis. For $\tau \geq 1$, define

$$A_{t,\tau} := \sum_{r=0}^{\tau-1} q_t^r, \quad B_{t,\tau} := \sum_{r=0}^{\tau-1} q_t^{2r}, \tag{90}$$

with $A_{t,0} = B_{t,0} = 0$. The two quantities encode different notions of temporal memory. Specifically,

$$A_{t,\tau} = \big\| \big(1, q_t, \ldots, q_t^{\tau-1}\big) \big\|_1 \tag{91}$$

is the ℓ_1 mass of the proximal-memory kernel. Its square, $A_{t,\tau}^2$, appears when bounding the accumulated effect of a general sequence of local gradient vectors without assuming cancellation among different local steps. In contrast,

$$B_{t,\tau} = \big\| \big(1, q_t, \ldots, q_t^{\tau-1}\big) \big\|_2^2 \tag{92}$$

is the squared ℓ_2 norm of the same kernel. It arises when the accumulated perturbations form a martingale-difference sequence, because the temporal cross terms vanish after taking expectation. Accordingly, $A_{t,\tau}^2$ governs the conservative client-drift envelope, whereas $B_{t,\tau}$ governs the sharper accumulation of zero-mean mini-batch and finite-shot stochastic noise.

Drift-memory factor. Suppose $\mu_t > 0$. Since $1 - q_t = \eta_t\mu_t$ the geometric-series identity yields

$$A_{t,\tau} = \frac{1-q_t^\tau}{1-q_t} = \frac{1-(1-\eta_t\mu_t)^\tau}{\eta_t\mu_t}. \tag{93}$$

Since $q_t \in [0,1]$, each term in $A_{t,\tau}$ is at most one. Therefore, we obtain $A_{t,\tau} \le \tau$. Moreover, $1-(1-\eta_t\mu_t)^\tau \le 1$, which implies $A_{t,\tau} \le \frac{1}{\eta_t\mu_t}$. Combining the two inequalities yields

$$A_{t,\tau} = \frac{1-q_t^\tau}{1-q_t} \le \min\Big\{\tau, \frac{1}{\eta_t\mu_t}\Big\}. \tag{94}$$

Thus $A_{t,\tau}$ initially grows approximately linearly with the number of local steps but cannot exceed the proximal-memory scale $(\eta_t\mu_t)^{-1}$.

Stochastic-memory factor. Similarly, $B_{t,\tau} = \frac{1-q_t^{2\tau}}{1-q_t^2}$, using $1-q_t^2 = \eta_t\mu_t\,(2-\eta_t\mu_t)$, we obtain

$$B_{t,\tau} = \frac{1-(1-\eta_t\mu_t)^{2\tau}}{\eta_t\mu_t(2-\eta_t\mu_t)}. \tag{95}$$

Again, because $q_t^{2r} \le 1$, it follows that $B_{t,\tau} \le \tau$. The numerator of (95) is also at most one, so $B_{t,\tau} \le \frac{1}{\eta_t\mu_t(2-\eta_t\mu_t)}$. Consequently,

$$B_{t,\tau} = \frac{1-q_t^{2\tau}}{1-q_t^2} \le \min\Big\{\tau, \frac{1}{\eta_t\mu_t(2-\eta_t\mu_t)}\Big\}. \tag{96}$$

Since $0 \le q_t \le 1$, we have $q_t^{2r} \le q_t^r$ and hence $B_{t,\tau} \le A_{t,\tau}$. For $\tau \ge 1$, $A_{t,\tau} \ge 1$, so we also have $B_{t,\tau} \le A_{t,\tau} \le A_{t,\tau}^2$. This explains why exploiting the martingale structure of stochastic gradient noise yields a sharper temporal accumulation factor $B_{t,\tau}$ than the conservative factor $A_{t,\tau}^2$ used for the general drift sequence.

Q-FedAvg limit. When $\mu_t = 0$, we have $q_t = 1$, and by the sum definitions $A_{t,\tau} = B_{t,\tau} = \tau$. Moreover, $\lim_{\mu_t\to 0^+} A_{t,\tau} = \tau, \lim_{\mu_t\to 0^+} B_{t,\tau} = \tau$. Thus the memory factors admit a continuous extension to the Q-FedAvg case and remain well-defined when an annealed proximal schedule reaches $\mu_t = 0$.

Remark C.10. *For positive proximal regularization, the two memory factors saturate differently. In the long-local-horizon regime, $\tau \gg \frac{1}{\eta_t\mu_t}$ the conservative drift-memory factor satisfies $A_{t,\tau}^2 \lesssim \frac{1}{(\eta_t\mu_t)^2}$ whereas the zero-mean stochastic-memory factor satisfies $B_{t,\tau} \lesssim \frac{1}{\eta_t\mu_t(2-\eta_t\mu_t)}$. When $\eta_t\mu_t \ll 1$, the latter behaves approximately as $B_{t,\tau} \lesssim \frac{1}{2\eta_t\mu_t}$. Hence the accumulated second moment of martingale stochastic perturbations has a substantially milder dependence on the proximal-memory scale than the worst-case squared drift envelope. This distinction is exploited in Section C.2.5.*

C.2.3 General client-drift envelope

We now use the geometric-memory factor $A_{t,\tau}$ to control the second moment of the total displacement generated by the local stochastic gradient sequence. Unlike the later noise-specific bound based on $B_{t,\tau}$, the result below does not require cancellation among gradients generated at different local steps and therefore applies to the complete stochastic-gradient sequence.

Lemma C.11 (Proximal client-drift bound)**.** *Suppose Assumption 3.5 holds, i.e.,*

$$\mathbb{E}\Big[\big\|g_k\big(\boldsymbol{\theta}_k^{(t,s)}, \boldsymbol{\xi}_k^{(t,s)}\big)\big\|_2^2 \,\Big|\, \mathcal{F}_{t,s}\Big] \le G^2 \tag{97}$$

almost surely for every client k, communication round t, and local step s. Then, for every $\tau \ge 0$,

$$\mathbb{E}\Big[\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \le \eta_t^2 G^2 A_{t,\tau}^2. \tag{98}$$

Consequently, when $\mu_t > 0$,

$$\mathbb{E}\Big[\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \le \eta_t^2 G^2 \min\Big\{\tau^2, \frac{1}{(\eta_t\mu_t)^2}\Big\}. \tag{99}$$

When $\mu_t = 0$, the bound reduces to

$$\mathbb{E}\Big[\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \le \eta_t^2 G^2 \tau^2. \tag{100}$$

Proof of Lemma C.11. We first consider the case $\tau = 0$. Since no local update has been performed, we have $\mathbf{d}_k^{(t,0)} = \mathbf{0}$, and the claim holds trivially.

Now we consider $\tau \geq 1$. From Lemma C.9, we have

$$\mathbf{d}_k^{(t,\tau)} = -\eta_t \sum_{s=0}^{\tau-1} q_t^{\tau-1-s} \mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,s)}, \boldsymbol{\xi}_k^{(t,s)}\big). \tag{101}$$

For arbitrary vectors $\{\mathbf{v}_s\}$ and nonnegative coefficients $\{a_s\}$, weighted Cauchy–Schwarz gives

$$\Big\| \sum_s a_s \mathbf{v}_s \Big\|_2^2 \leq \Big(\sum_s a_s \Big) \Big(\sum_s a_s \|\mathbf{v}_s\|_2^2 \Big). \tag{102}$$

Taking the Euclidean norm on both sides of (101) and applying (102) with $a_s := q_t^{\tau-1-s} \geq 0$ yields the pathwise inequality

$$\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2 \leq \eta_t^2 A_{t,\tau} \sum_{s=0}^{\tau-1} q_t^{\tau-1-s} \big\|\mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,s)}, \boldsymbol{\xi}_k^{(t,s)}\big)\big\|_2^2. \tag{103}$$

Taking conditional expectation given the broadcast model $\boldsymbol{\theta}^{(t)}$, we can obatin

$$\mathbb{E}\Big[\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \leq \eta_t^2 \mathbb{E}\Big[\big(A_{t,\tau} \sum_{s=0}^{\tau-1} q_t^{\tau-1-s}\big) \big\|\mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,s)}, \boldsymbol{\xi}_k^{(t,s)}\big)\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big]. \tag{104}$$

$$= \eta_t^2 A_{t,\tau} \sum_{s=0}^{\tau-1} q_t^{\tau-1-s} \mathbb{E}\Big[\big\|\mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,s)}, \boldsymbol{\xi}_k^{(t,s)}\big)\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \tag{105}$$

Since $\boldsymbol{\theta}^{(t)}$ is measurable with respect to the history preceding the local oracle call, the tower property and Assumption 3.5 imply

$$\mathbb{E}\Big[\big\|\mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,s)}, \boldsymbol{\xi}_k^{(t,s)}\big)\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] = \mathbb{E}\Big[\mathbb{E}\Big[\big\|\mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,s)}, \boldsymbol{\xi}_k^{(t,s)}\big)\big\|_2^2 \,\Big|\, \mathcal{F}_{t,s}\Big] \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \leq G^2. \tag{106}$$

Substituting (106) into (104) yields

$$\mathbb{E}\Big[\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \leq \eta_t^2 G^2 A_{t,\tau} \sum_{s=0}^{\tau-1} q_t^{\tau-1-s} = \eta_t^2 G^2 A_{t,\tau}^2, \tag{107}$$

which proves (98).

It remains to bound the geometric sum. For $\mu_t > 0$, using $A_{t,\tau} \leq \min\big\{\tau, \frac{1}{\eta_t \mu_t}\big\}$ we then have

$$A_{t,\tau}^2 \leq \min\Big\{\tau^2, \frac{1}{(\eta_t \mu_t)^2}\Big\}. \tag{108}$$

Combining these two bounds 107 and 108 yields the following bound:

$$\mathbb{E}\Big[\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \leq \eta_t^2 G^2 \min\Big\{\tau^2, \frac{1}{(\eta_t \mu_t)^2}\Big\}. \tag{109}$$

When $\mu_t = 0$, $A_{t,\tau} = \tau$, giving 100. This completes the proof. □

Remark C.12. *The proof uses only the per-step conditional second-moment bound in Assumption 3.5. In particular, it does not require a bound of the form*

$$\mathbb{E}\Big[\max_{0 \leq s < \tau} \big\|\mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,s)}, \boldsymbol{\xi}_k^{(t,s)}\big)\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \leq G^2, \tag{110}$$

which would be strictly stronger and does not follow from a separate second-moment bound on each stochastic gradient. The weighted Cauchy–Schwarz inequality therefore provides the drift envelope directly under the stated oracle assumptions.

Corollary C.13 (Weighted client-drift energy)**.** *For the weighted client-drift energy at local step τ of round t, $D_{t,\tau} := \sum_{k=1}^{K} p_k \mathbb{E}\big[\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2 \big| \boldsymbol{\theta}^{(t)}\big]$, under the conditions of Lemma C.11,*

$$D_{t,\tau} \le \eta_t^2 G^2 A_{t,\tau}^2. \tag{111}$$

Consequently, if $\mu_t > 0$, $D_{t,\tau} \le \eta_t^2 G^2 \min\big\{\tau^2, \frac{1}{(\eta_t \mu_t)^2}\big\}$. When $\mu_t = 0$, the corresponding Q-FedAvg bound is $D_{t,\tau} \le \eta_t^2 G^2 \tau^2$.

Proof of Corollary C.13. Lemma C.11 establishes, separately for every client k, that

$$\mathbb{E}\Big[\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \le \eta_t^2 G^2 A_{t,\tau}^2. \tag{112}$$

Since the client weights satisfy $p_k \ge 0, \sum_{k=1}^{K} p_k = 1$ multiplying (112) by p_k preserves the inequality:

$$p_k \mathbb{E}\Big[\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \le p_k \eta_t^2 G^2 A_{t,\tau}^2. \tag{113}$$

Summing (113) over $k = 1, \dots, K$ we obtain

$$\sum_{k=1}^{K} p_k \mathbb{E}\Big[\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \le \eta_t^2 G^2 A_{t,\tau}^2 \sum_{k=1}^{K} p_k. \tag{114}$$

By the definition of $D_{t,\tau}$ and the normalization $\sum_{k=1}^{K} p_k = 1$, we obtain the following bound

$$D_{t,\tau} \le \eta_t^2 G^2 A_{t,\tau}^2. \tag{115}$$

Now suppose $\mu_t > 0$. From (94), $A_{t,\tau}^2 \le \min\big\{\tau^2, \frac{1}{(\eta_t \mu_t)^2}\big\}$. Substituting it into (111) yields the bound for $D_{t,\tau}$.Finally, when $\mu_t = 0$, we have $q_t = 1$ and $A_{t,\tau} = \tau$. Therefore, $D_{t,\tau} \le \eta_t^2 G^2 \tau^2$, which completes the proof. □

Remark C.14. *Corollary C.13 requires neither equal client weights nor independence among client trajectories. The result follows from the uniform per-client second-moment bound and the convex weighting $\sum_k p_k = 1$. The bound is deliberately heterogeneity-agnostic: it provides a robust upper envelope for $D_{t,\tau}$ based only on G^2. It does not yet explain how statistical heterogeneity affects client drift. That dependence is derived separately through the heterogeneity-sensitive recursion, where ζ^2 enters the evolution of $D_{t,\tau}$ through the dispersion of the local objective gradients.*

C.2.4 Endpoint and cumulative client drift

Lemma C.11 controls the client displacement at an arbitrary local step. Two consequences of this result are particularly important for the subsequent global analysis. The first is the *endpoint drift*, $D_{t,\mathcal{T}}$, which controls the second moment of the local models relative to the broadcast model at the end of local optimization. This quantity will later be used to control the additional aggregation error introduced by partial client participation. The second is the *cumulative drift*, $\sum_{\tau=0}^{\mathcal{T}-1} D_{t,\tau}$, which controls the accumulated effect of evaluating local gradients along client-specific trajectories rather than at the common broadcast iterate. This quantity enters directly into the global-update deviation and the resulting stationarity analysis.

Endpoint drift. At the end of the $\mathcal{T}$ local updates, Lemma C.11 gives, for every client k,

$$\mathbb{E}\Big[\big\|\mathbf{d}_k^{(t,\mathcal{T})}\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \le \eta_t^2 G^2 A_{t,\mathcal{T}}^2. \tag{116}$$

Averaging over clients with the weights $\{p_k\}_{k=1}^{K}$ therefore yields $D_{t,\mathcal{T}} \le \eta_t^2 G^2 A_{t,\mathcal{T}}^2$. If $\mu_t > 0$, then $A_{t,\mathcal{T}} \le \min\big\{\mathcal{T}, \frac{1}{\eta_t \mu_t}\big\}$ and therefore

$$D_{t,\mathcal{T}} \le \eta_t^2 G^2 \min\Big\{\mathcal{T}^2, \frac{1}{(\eta_t \mu_t)^2}\Big\}. \tag{117}$$

When $\mu_t = 0$, the Q-FedAvg limit satisfies $D_{t,\mathcal{T}} \le \eta_t^2 G^2 \mathcal{T}^2$. The quantity $D_{t,\mathcal{T}}$ is a second moment relative to the common broadcast model rather than the variance of the client endpoints around their

population mean. Nevertheless, it upper-bounds that dispersion. To see this, consider, for example, equal client weights and define $\overline{\mathbf{d}}^{(t,\mathcal{T})} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{d}_k^{(t,\mathcal{T})}$. The standard variance decomposition gives

$$\frac{1}{K}\sum_{k=1}^{K}\left\|\mathbf{d}_k^{(t,\mathcal{T})} - \overline{\mathbf{d}}^{(t,\mathcal{T})}\right\|_2^2 = \frac{1}{K}\sum_{k=1}^{K}\left\|\mathbf{d}_k^{(t,\mathcal{T})}\right\|_2^2 - \left\|\overline{\mathbf{d}}^{(t,\mathcal{T})}\right\|_2^2 \le \frac{1}{K}\sum_{k=1}^{K}\left\|\mathbf{d}_k^{(t,\mathcal{T})}\right\|_2^2. \tag{118}$$

After taking conditional expectation, the endpoint dispersion is therefore upper-bounded by $D_{t,\mathcal{T}}$. This observation explains why endpoint drift becomes the relevant trajectory quantity in the later partial-participation analysis: sampling only a subset of clients is more consequential when the potential client endpoints are more dispersed.

We defer the exact finite-population sampling factor to Section C.5, where the additional assumptions on client weights and subset sampling are introduced explicitly.

Cumulative drift. The global-update analysis depends not only on the final client endpoint but also on the entire local trajectory. Indeed, by smoothness,

$$\left\|\nabla F_k\big(\boldsymbol{\theta}_k^{(t,\tau)}\big) - \nabla F_k\big(\boldsymbol{\theta}^{(t)}\big)\right\|_2 \le \beta\left\|\mathbf{d}_k^{(t,\tau)}\right\|_2. \tag{119}$$

Thus the accumulation of local gradient mismatch across $\tau = 0, \dots, \mathcal{T}-1$ is naturally controlled by the cumulative client-drift energy $\sum_{\tau=0}^{\mathcal{T}-1} D_{t,\tau}$. To quantify this term compactly, define the cumulative geometric-memory factor $\Xi_{t,\mathcal{T}} := \sum_{\tau=0}^{\mathcal{T}-1} A_{t,\tau}^2$. Then (111) yields immediately $\sum_{\tau=0}^{\mathcal{T}-1} D_{t,\tau} \le \eta_t^2 G^2 \Xi_{t,\mathcal{T}}$. Similarly, for every individual client,

$$\sum_{\tau=0}^{\mathcal{T}-1}\mathbb{E}\left[\left\|\mathbf{d}_k^{(t,\tau)}\right\|_2^2 \,\middle|\, \boldsymbol{\theta}^{(t)}\right] \le \eta_t^2 G^2 \Xi_{t,\mathcal{T}}. \tag{120}$$

When $\mu_t > 0$, the bound of $A_{t,\tau}^2$ implies

$$\Xi_{t,\mathcal{T}} \le \sum_{\tau=0}^{\mathcal{T}-1}\min\left\{\tau^2, \frac{1}{(\eta_t\mu_t)^2}\right\} \le \min\left\{\sum_{\tau=0}^{\mathcal{T}-1}\tau^2, \frac{\mathcal{T}-1}{(\eta_t\mu_t)^2}\right\} \tag{121}$$

$$= \min\left\{\frac{(\mathcal{T}-1)\mathcal{T}(2\mathcal{T}-1)}{6}, \frac{\mathcal{T}-1}{(\eta_t\mu_t)^2}\right\}. \tag{122}$$

The factor $\mathcal{T}-1$ in the second branch follows from $A_{t,0} = 0$, so only the $\mathcal{T}-1$ nonzero local-step terms need to be bounded by $(\eta_t\mu_t)^{-2}$. Consequently,

$$\sum_{\tau=0}^{\mathcal{T}-1} D_{t,\tau} \le \eta_t^2 G^2 \min\left\{\frac{(\mathcal{T}-1)\mathcal{T}(2\mathcal{T}-1)}{6}, \frac{\mathcal{T}-1}{(\eta_t\mu_t)^2}\right\}. \tag{123}$$

For Q-FedAvg, $\mu_t = 0$ and $A_{t,\tau} = \tau$, so $\Xi_{t,\mathcal{T}} = \sum_{\tau=0}^{\mathcal{T}-1}\tau^2 = \frac{(\mathcal{T}-1)\mathcal{T}(2\mathcal{T}-1)}{6}$. Therefore,

$$\sum_{\tau=0}^{\mathcal{T}-1} D_{t,\tau} \le \eta_t^2 G^2 \frac{(\mathcal{T}-1)\mathcal{T}(2\mathcal{T}-1)}{6} = O\left(\eta_t^2 G^2 \mathcal{T}^3\right). \tag{124}$$

For $\mu_t > 0$, two local-computation regimes arise. Before the geometric memory has saturated, the first branch of (123) yields $\sum_{\tau=0}^{\mathcal{T}-1} D_{t,\tau} = O\big(\eta_t^2 G^2 \mathcal{T}^3\big)$. Once the local horizon is sufficiently long relative to the proximal-memory scale, the second branch gives $\sum_{\tau=0}^{\mathcal{T}-1} D_{t,\tau} = O\big(\frac{G^2\mathcal{T}}{\mu_t^2}\big)$. Thus the available drift envelope changes from cubic dependence on the local horizon to linear dependence after the geometric memory has saturated. This statement concerns the local drift envelope only. It should not be interpreted as an $O(\mathcal{T})$ global convergence rate. The subsequent stationarity bound additionally contains the learning-rate normalization, the factor $(\beta+\mu_t)^2$, stochastic-gradient variance, and, under partial participation, the client-sampling term.

Exact cumulative-memory expression. For completeness, when $\mu_t > 0$ and therefore $q_t < 1$, the cumulative geometric-memory factor admits the exact closed form

$$\Xi_{t,\mathcal{T}} = \sum_{\tau=0}^{\mathcal{T}-1}\left(\frac{1-q_t^\tau}{1-q_t}\right)^2 = \frac{1}{(1-q_t)^2}\sum_{\tau=0}^{\mathcal{T}-1}\left(1 - 2q_t^\tau + q_t^{2\tau}\right)$$

$$= \frac{1}{(1-q_t)^2}\Big[\mathcal{T} - \frac{2(1-q_t^{\mathcal{T}})}{1-q_t} + \frac{1-q_t^{2\mathcal{T}}}{1-q_t^2}\Big]. \tag{125}$$

Since $1 - q_t = \eta_t\mu_t$ the leading term for a long local horizon is $\Xi_{t,\mathcal{T}} \sim \frac{\mathcal{T}}{(\eta_t\mu_t)^2}, \mathcal{T} \to \infty$ with t fixed and $\mu_t > 0$. Hence (121) captures the correct long-horizon scaling while providing a simpler expression for the subsequent convergence analysis.

Remark C.15. *Let $a_t := \eta_t\mu_t$ When $a_t\tau \ll 1$ the local horizon is short relative to the proximal-memory scale. Expanding $A_{t,\tau} = \frac{1-(1-a_t)^\tau}{a_t}$ around $a_t = 0$ gives $A_{t,\tau} = \tau - \frac{a_t}{2}\tau(\tau-1) + O\left(a_t^2\tau^3\right)$. Thus, in this* unsaturated-memory regime*, the geometric kernel remains close to the Q-FedAvg kernel $A_{t,\tau} = \tau$. In contrast, when $\tau \gg \frac{1}{\eta_t\mu_t}$ the local horizon exceeds the induced proximal-memory scale and $A_{t,\tau} \le \frac{1}{\eta_t\mu_t}$. We refer to $\tau_{\mathrm{mem},t} := \frac{1}{\eta_t\mu_t}, \mu_t > 0$ as a convenient* proximal-memory scale*. It is not an additional algorithmic hyperparameter; it is induced jointly by the local learning rate and the proximal coefficient. More precisely, whenever $0 < q_t < 1$, the geometric kernel q_t^r has the e-folding decay time $\tau_{\mathrm{decay},t} = \frac{1}{-\log(1-\eta_t\mu_t)}$. Using $-\log(1-a_t) = a_t + \frac{a_t^2}{2} + O(a_t^3)$ we obtain, for $a_t \ll 1$, $\tau_{\mathrm{decay},t} = \frac{1}{a_t}\left[1 - \frac{a_t}{2} + O(a_t^2)\right]$. Therefore, $\tau_{\mathrm{mem},t} = 1/(\eta_t\mu_t)$ is the natural first-order approximation to the exact exponential decay timescale in the small $\eta_t\mu_t$ regime.*

The endpoint and cumulative quantities play complementary roles:

- $D_{t,\mathcal{T}}$: controls endpoint-sensitive client-sampling fluctuations;
- $\sum_{\tau=0}^{\mathcal{T}-1} D_{t,\tau}$: controls trajectory-induced error in the global update.

The first role will be made precise only after introducing the sampling model for partial participation, while the second enters immediately in the full-participation update decomposition developed below.

C.2.5 Geometric filtering of stochastic perturbations

The previous drift envelope treats the complete stochastic-gradient sequence conservatively and therefore depends on the ℓ_1 memory factor $A_{t,\tau}^2$. We now exploit the martingale-difference structure of the stochastic-gradient error to obtain a sharper characterization of how mini-batch and finite-shot perturbations propagate through the proximal local recursion.

Define the direct martingale component of the unrolled local displacement by

$$\mathbf{u}_k^{(t,\tau)} := -\eta_t \sum_{s=0}^{\tau-1} q_t^{\tau-1-s}\mathbf{e}_k^{(t,s)}. \tag{126}$$

Although $\mathbf{u}_k^{(t,\tau)}$ represents the explicit contribution of the stochastic-gradient errors in the unrolled recursion, previous errors also affect future local iterates and hence future deterministic gradients $\nabla F_k(\boldsymbol{\theta}_k^{(t,s)})$. Thus $\mathbf{u}_k^{(t,\tau)}$ should not be interpreted as the complete difference between a stochastic local trajectory and a hypothetical noise-free trajectory.

Lemma C.16 (Geometric filtering of stochastic perturbations)**.** *Under Assumption C.1, for every client k, communication round t, and local step $\tau \ge 1$,*

$$\mathbb{E}\left[\left\|\mathbf{u}_k^{(t,\tau)}\right\|_2^2 \,\middle|\, \boldsymbol{\theta}^{(t)}\right] \le \eta_t^2\sigma_{k,t}^2 B_{t,\tau}, \tag{127}$$

where $B_{t,\tau} = \sum_{r=0}^{\tau-1} q_t^{2r}$.

Proof of Lemma C.16. Expanding the squared norm in (126) we can obtain

$$\left\|\mathbf{u}_k^{(t,\tau)}\right\|_2^2 = \eta_t^2 \sum_{s=0}^{\tau-1} q_t^{2(\tau-1-s)}\left\|\mathbf{e}_k^{(t,s)}\right\|_2^2 + 2\eta_t^2 \sum_{0\le r<s\le\tau-1} q_t^{\tau-1-r}q_t^{\tau-1-s}\left\langle \mathbf{e}_k^{(t,r)}, \mathbf{e}_k^{(t,s)}\right\rangle. \tag{128}$$

We first show that every temporal cross term vanishes after expectation. For $r < s$, the earlier error $\mathbf{e}_k^{(t,r)}$ is measurable with respect to $\mathcal{F}_{t,s}$. Therefore, using the tower property,

$$\mathbb{E}\left[\left\langle \mathbf{e}_k^{(t,r)}, \mathbf{e}_k^{(t,s)}\right\rangle \,\middle|\, \boldsymbol{\theta}^{(t)}\right] = \mathbb{E}\left[\mathbb{E}\left[\left\langle \mathbf{e}_k^{(t,r)}, \mathbf{e}_k^{(t,s)}\right\rangle \,\middle|\, \mathcal{F}_{t,s}\right] \,\middle|\, \boldsymbol{\theta}^{(t)}\right] \tag{129}$$

$$= \mathbb{E}\left[\left\langle \mathbf{e}_k^{(t,r)}, \mathbb{E}\left[\mathbf{e}_k^{(t,s)} \,\middle|\, \mathcal{F}_{t,s}\right]\right\rangle \,\middle|\, \boldsymbol{\theta}^{(t)}\right] = 0. \tag{130}$$

Taking conditional expectation in (128) therefore yields

$$\mathbb{E}\Big[\big\|\mathbf{u}_k^{(t,\tau)}\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] = \eta_t^2 \sum_{s=0}^{\tau-1} q_t^{2(\tau-1-s)} \mathbb{E}\Big[\big\|\mathbf{e}_k^{(t,s)}\big\|_2^2 \,\big|\, \boldsymbol{\theta}^{(t)}\Big]. \tag{131}$$

By the tower property and Assumption C.1,

$$\mathbb{E}\Big[\big\|\mathbf{e}_k^{(t,s)}\big\|_2^2 \,\big|\, \boldsymbol{\theta}^{(t)}\Big] = \mathbb{E}\Big[\mathbb{E}\Big[\big\|\mathbf{e}_k^{(t,s)}\big\|_2^2 \,\big|\, \mathcal{F}_{t,s}\Big] \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \le \sigma_{k,t}^2. \tag{132}$$

Consequently,

$$\mathbb{E}\Big[\big\|\mathbf{u}_k^{(t,\tau)}\big\|_2^2 \,\big|\, \boldsymbol{\theta}^{(t)}\Big] \le \eta_t^2 \sigma_{k,t}^2 \sum_{s=0}^{\tau-1} q_t^{2(\tau-1-s)} = \eta_t^2 \sigma_{k,t}^2 B_{t,\tau}, \tag{133}$$

which proves the claim. □

For the finite-shot VQC, $\sigma_{k,t}^2 = \frac{\sigma_{\mathrm{mb},k}^2}{B} + \frac{c_{\mathrm{ps},k} d_q}{S_{k,t}}$ and therefore Lemma C.16 gives

$$\mathbb{E}\Big[\big\|\mathbf{u}_k^{(t,\tau)}\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \le \eta_t^2 \Big(\frac{\sigma_{\mathrm{mb},k}^2}{B} + \frac{c_{\mathrm{ps},k} d_q}{S_{k,t}}\Big) B_{t,\tau}. \tag{134}$$

When $\mu_t = 0$, $B_{t,\tau} = \tau$, so the direct stochastic perturbation accumulates linearly in second moment:

$$\mathbb{E}\Big[\big\|\mathbf{u}_k^{(t,\tau)}\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \le \eta_t^2 \sigma_{k,t}^2 \tau. \tag{135}$$

For $\mu_t > 0$, $B_{t,\tau} \le \min\big\{\tau, \frac{1}{\eta_t \mu_t (2 - \eta_t \mu_t)}\big\}$ and hence

$$\mathbb{E}\Big[\,\Big\|\mathbf{u}_k^{(t,\tau)}\Big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \le \eta_t^2 \sigma_{k,t}^2 \min\left\{\tau, \frac{1}{\eta_t \mu_t (2 - \eta_t \mu_t)}\right\}. \tag{136}$$

Thus positive proximal regularization limits the temporal accumulation of the direct martingale component. Importantly, it does not alter the single-oracle variance $\sigma_{k,t}^2$ itself. Increasing the number of quantum measurement shots reduces the finite-shot source variance through $S_{k,t}^{-1}$, whereas the proximal coefficient controls how perturbations generated at different local steps propagate through the local trajectory.

Remark C.17 (Why $B_{t,\tau}$ is sharper than $A_{t,\tau}^2$)**.** *The generic client-drift envelope depends on $A_{t,\tau}^2$ because it allows temporally coherent accumulation of arbitrary local gradients. For the martingale component, temporal cross terms vanish and the corresponding factor is only $B_{t,\tau} = \sum_{r=0}^{\tau-1} q_t^{2r}$ Since $B_{t,\tau} \le A_{t,\tau}^2$ the noise-specific analysis is strictly sharper whenever the geometric kernel contains more than one nonzero term.*

C.2.6 Heterogeneity-sensitive drift recursion

The preceding G^2-based analysis provides a robust upper envelope for client drift, but it does not expose how statistical heterogeneity affects the local trajectories. We now derive a complementary recursion that makes the dependence on the gradient-dissimilarity ζ^2 explicit. The latter drives the different clients away from the common broadcast model and therefore enters the evolution of $D_{t,\tau}$.

Lemma C.18 (Heterogeneity-sensitive drift recursion)**.** *Fix any nonnegative $\gamma_t > 0$ and define the drift-propagation factor $\rho_t := (1 + \gamma_t)\,(1 - \eta_t \mu_t + \eta_t \beta)^2$ and the round-dependent forcing term*

$$C_t := \eta_t^2 \Big[\big(1 + \frac{1}{\gamma_t}\big)\big(\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2 + \zeta^2\big) + \overline{\sigma}_t^2\Big]. \tag{137}$$

Then, for every local step τ, $D_{t,\tau+1} \le \rho_t D_{t,\tau} + C_t$. Since $D_{t,0} = 0$, it follows that

$$D_{t,\tau} \le C_t \sum_{r=0}^{\tau-1} \rho_t^r. \tag{138}$$

Proof of Lemma C.18. For compactness, we define

$$\mathbf{g}_t := \nabla F(\boldsymbol{\theta}^{(t)}), \quad \mathbf{h}_k^{(t)} := \nabla F_k(\boldsymbol{\theta}^{(t)}) - \mathbf{g}_t, \quad \mathbf{r}_k^{(t,\tau)} := \nabla F_k\big(\boldsymbol{\theta}_k^{(t,\tau)}\big) - \nabla F_k\big(\boldsymbol{\theta}^{(t)}\big). \tag{139}$$

By construction, we have

$$\nabla F_k\big(\boldsymbol{\theta}_k^{(t,\tau)}\big) = \mathbf{g}_t + \mathbf{h}_k^{(t)} + \mathbf{r}_k^{(t,\tau)}. \tag{140}$$

Using the oracle decomposition in the local drift recursion gives

$$\mathbf{d}_k^{(t,\tau+1)} = q_t \mathbf{d}_k^{(t,\tau)} - \eta_t \mathbf{r}_k^{(t,\tau)} - \eta_t\big(\mathbf{g}_t + \mathbf{h}_k^{(t)}\big) - \eta_t \mathbf{e}_k^{(t,\tau)}. \tag{141}$$

We define $\mathbf{a}_k^{(t,\tau)} := q_t \mathbf{d}_k^{(t,\tau)} - \eta_t \mathbf{r}_k^{(t,\tau)}$, and $\mathbf{b}_k^{(t)} := -\eta_t\big(\mathbf{g}_t + \mathbf{h}_k^{(t)}\big)$. Both are measurable with respect to $\mathcal{F}_{t,\tau}$. Hence, using $\mathbb{E}\big[\mathbf{e}_k^{(t,\tau)} \,\big|\, \mathcal{F}_{t,\tau}\big] = \mathbf{0}$ we obtain

$$\mathbb{E}\Big[\big\|\mathbf{d}_k^{(t,\tau+1)}\big\|_2^2 \,\big|\, \mathcal{F}_{t,\tau}\Big] = \big\|\mathbf{a}_k^{(t,\tau)} + \mathbf{b}_k^{(t)}\big\|_2^2 + \eta_t^2 \mathbb{E}\Big[\big\|\mathbf{e}_k^{(t,\tau)}\big\|_2^2 \,\big|\, \mathcal{F}_{t,\tau}\Big] \tag{142}$$

$$\leq \big\|\mathbf{a}_k^{(t,\tau)} + \mathbf{b}_k^{(t)}\big\|_2^2 + \eta_t^2 \sigma_{k,t}^2. \tag{143}$$

For any $\gamma_t > 0$, Young's inequality gives

$$\big\|\mathbf{a}_k^{(t,\tau)} + \mathbf{b}_k^{(t)}\big\|_2^2 \leq (1+\gamma_t)\big\|\mathbf{a}_k^{(t,\tau)}\big\|_2^2 + \big(1 + \frac{1}{\gamma_t}\big)\big\|\mathbf{b}_k^{(t)}\big\|_2^2. \tag{144}$$

By β-smoothness, we obtain $\|\mathbf{r}_k^{(t,\tau)}\|_2 \leq \beta \|\mathbf{d}_k^{(t,\tau)}\|_2$. Since $q_t \geq 0$, we have

$$\big\|\mathbf{a}_k^{(t,\tau)}\big\|_2 \leq q_t \big\|\mathbf{d}_k^{(t,\tau)}\big\|_2 + \eta_t \big\|\mathbf{r}_k^{(t,\tau)}\big\|_2 \leq (q_t + \eta_t\beta)\,\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2 \tag{145}$$

$$= (1 - \eta_t\mu_t + \eta_t\beta)\,\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2, \tag{146}$$

it follows that,

$$\big\|\mathbf{a}_k^{(t,\tau)}\big\|_2^2 \leq (1 - \eta_t\mu_t + \eta_t\beta)^2\,\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2. \tag{147}$$

Moreover, we have $\big\|\mathbf{b}_k^{(t)}\big\|_2^2 = \eta_t^2\big\|\mathbf{g}_t + \mathbf{h}_k^{(t)}\big\|_2^2$. Combining it with (142), and (147), taking conditional expectation given $\boldsymbol{\theta}^{(t)}$, multiplying by p_k, and summing over clients we then obtain

$$D_{t,\tau+1} \leq \rho_t D_{t,\tau} + \eta_t^2\Big(1 + \frac{1}{\gamma_t}\Big)\sum_{k=1}^K p_k \big\|\mathbf{g}_t + \mathbf{h}_k^{(t)}\big\|_2^2 + \eta_t^2 \overline{\sigma}_t^2. \tag{148}$$

It remains to simplify the client-gradient term. Since

$$\sum_{k=1}^K p_k \mathbf{h}_k^{(t)} = \sum_{k=1}^K p_k \nabla F_k(\boldsymbol{\theta}^{(t)}) - \mathbf{g}_t = \mathbf{0}, \tag{149}$$

we have the exact weighted variance decomposition

$$\sum_{k=1}^K p_k \big\|\mathbf{g}_t + \mathbf{h}_k^{(t)}\big\|_2^2 = \sum_{k=1}^K p_k \Big[\|\mathbf{g}_t\|_2^2 + 2\big\langle \mathbf{g}_t, \mathbf{h}_k^{(t)}\big\rangle + \|\mathbf{h}_k^{(t)}\|_2^2\Big] \tag{150}$$

$$= \|\mathbf{g}_t\|_2^2 + 2\Big\langle \mathbf{g}_t, \sum_{k=1}^K p_k \mathbf{h}_k^{(t)}\Big\rangle + \sum_{k=1}^K p_k \|\mathbf{h}_k^{(t)}\|_2^2 \tag{151}$$

$$= \|\mathbf{g}_t\|_2^2 + \sum_{k=1}^K p_k \|\mathbf{h}_k^{(t)}\|_2^2 \leq \|\mathbf{g}_t\|_2^2 + \zeta^2. \tag{152}$$

Substituting this bound into (148) proves the bound for $D_{t,\tau+1}$.

Finally, since $D_{t,0} = 0$, repeated application of the recursion (148) gives

$$D_{t,1} \leq C_t, D_{t,2} \leq \rho_t C_t + C_t, D_{t,3} \leq \rho_t^2 C_t + \rho_t C_t + C_t. \tag{153}$$

Therefore, for any $\tau \geq 1$, $D_{t,\tau} \leq C_t \sum_{r=0}^{\tau-1} \rho_t^r$. This completes the proof. □

Remark C.19. *The weighted first-order heterogeneity deviations cancel at the common broadcast iterate, $\sum_{k=1}^{K} p_k \mathbf{h}_k^{(t)} = \mathbf{0}$ but their squared dispersion does not, $\sum_{k=1}^{K} p_k \|\mathbf{h}_k^{(t)}\|_2^2 \le \zeta^2$. Hence ζ^2 affects convergence by enlarging the forcing term in the client-drift recursion. It does not appear as a nonzero averaged common-iterate gradient vector.*

Remark C.20. *The factor ρ_t should be interpreted as a* drift-propagation factor. *The finite-horizon bound (138) is valid for every $\rho_t \ge 0$ and does not require contraction. If $\rho_t < 1$, however, the recursion is contractive and $D_{t,\tau} \le C_t \frac{1-\rho_t^\tau}{1-\rho_t} \le \frac{C_t}{1-\rho_t}$. Under the present smoothness-only analysis, $\rho_t < 1$ requires $1 + \gamma_t)^{1/2} (1 - \eta_t\mu_t + \eta_t\beta) < 1$. Thus contractivity of this particular bound is a stronger property than the basic proximal-memory condition $q_t < 1$. We therefore do not assume $\rho_t < 1$ in the main finite-horizon convergence result.*

C.3 Full-Participation Update and One-round Descent

C.3.1 Exact server-update decomposition

We next connect the local trajectory analysis to the global server update. Under full participation, all K clients return their terminal local models, and the server uses the same weights $\{p_k\}_{k=1}^K$ as in the global objective. Define the full-participation server increment by $\mathbf{\Delta}^{(t)} := \boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)} = \sum_{k=1}^{K} p_k \mathbf{d}_k^{(t,\mathcal{T})}$. For every local step, define the aggregated martingale noise $\mathbf{N}^{(t,\tau)} := \sum_{k=1}^{K} p_k \mathbf{e}_k^{(t,\tau)}$, the trajectory-induced gradient deviation $\mathbf{R}^{(t,\tau)} := \sum_{k=1}^{K} p_k \left[\nabla F_k\big(\boldsymbol{\theta}_k^{(t,\tau)}\big) - \nabla F_k\big(\boldsymbol{\theta}^{(t)}\big)\right]$ the weighted mean local displacement $\overline{\mathbf{d}}^{(t,\tau)} := \sum_{k=1}^{K} p_k \mathbf{d}_k^{(t,\tau)}$ and the combined trajectory/proximal deviation $\mathbf{W}^{(t,\tau)} := \mathbf{R}^{(t,\tau)} + \mu_t \overline{\mathbf{d}}^{(t,\tau)}$. The following identity is the central bridge between the local dynamics and the server-level convergence analysis.

Lemma C.21 (Exact full-participation update decomposition)**.** *Under full client participation,*

$$\mathbf{\Delta}^{(t)} + \eta_t \mathcal{T} \nabla F(\boldsymbol{\theta}^{(t)}) = -\eta_t \sum_{\tau=0}^{\mathcal{T}-1} \Big[\mathbf{N}^{(t,\tau)} + \mathbf{W}^{(t,\tau)}\Big]. \tag{154}$$

Proof of Lemma C.21. From the local update, we have

$$\mathbf{d}_k^{(t,\tau+1)} - \mathbf{d}_k^{(t,\tau)} = -\eta_t \Big[\mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,\tau)}, \boldsymbol{\xi}_k^{(t,\tau)}\big) + \mu_t \mathbf{d}_k^{(t,\tau)}\Big]. \tag{155}$$

Summing from $\tau = 0$ to $\mathcal{T} - 1$ and using $\mathbf{d}_k^{(t,0)} = \mathbf{0}$, we can obtain

$$\mathbf{d}_k^{(t,\mathcal{T})} = -\eta_t \sum_{\tau=0}^{\mathcal{T}-1} \Big[\mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,\tau)}, \boldsymbol{\xi}_k^{(t,\tau)}\big) + \mu_t \mathbf{d}_k^{(t,\tau)}\Big]. \tag{156}$$

Multiplying by p_k and summing over clients yields

$$\mathbf{\Delta}^{(t)} = -\eta_t \sum_{\tau=0}^{\mathcal{T}-1} \Big[\sum_{k=1}^{K} p_k \mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,\tau)}, \boldsymbol{\xi}_k^{(t,\tau)}\big) + \mu_t \overline{\mathbf{d}}^{(t,\tau)}\Big]. \tag{157}$$

Using the stochastic-oracle decomposition,

$$\sum_{k=1}^{K} p_k \mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,\tau)}, \boldsymbol{\xi}_k^{(t,\tau)}\big) = \sum_{k=1}^{K} p_k \nabla F_k\big(\boldsymbol{\theta}_k^{(t,\tau)}\big) + \mathbf{N}^{(t,\tau)}. \tag{158}$$

Adding and subtracting $\nabla F_k(\boldsymbol{\theta}^{(t)})$ inside the weighted sum we have

$$\sum_{k=1}^{K} p_k \nabla F_k\big(\boldsymbol{\theta}_k^{(t,\tau)}\big) = \sum_{k=1}^{K} p_k \nabla F_k\big(\boldsymbol{\theta}^{(t)}\big) + \mathbf{R}^{(t,\tau)} = \nabla F\big(\boldsymbol{\theta}^{(t)}\big) + \mathbf{R}^{(t,\tau)}. \tag{159}$$

Therefore, we can obtain

$$\sum_{k=1}^{K} p_k \mathbf{g}_k\big(\boldsymbol{\theta}_k^{(t,\tau)}, \boldsymbol{\xi}_k^{(t,\tau)}\big) + \mu_t \overline{\mathbf{d}}^{(t,\tau)} = \nabla F(\boldsymbol{\theta}^{(t)}) + \mathbf{N}^{(t,\tau)} + \mathbf{W}^{(t,\tau)}. \tag{160}$$

Substituting this into (157) and summing over $\mathcal{T}$ local steps completes the proof. □

Eq. (154) separates the difference between the actual federated update and an idealized global-gradient step into two terms: *stochastic oracle error* $\mathbf{N}^{(t,\tau)}$ and *trajectory/proximal deviation* $\mathbf{W}^{(t,\tau)}$. Notably, there is no additional nonzero common-iterate heterogeneity vector: that component cancels exactly through $\sum_k p_k \nabla F_k = \nabla F$.

Lemma C.22 (Aggregated stochastic-noise bound)**.** *Under Assumption C.1, it holds*

$$\mathbb{E}\Big[\big\|\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{N}^{(t,\tau)}\big\|_2^2\,\Big|\,\boldsymbol{\theta}^{(t)}\Big] \le \mathcal{T}\overline{\sigma}_t^2. \tag{161}$$

Proof of Lemma C.22. Since $\mathbf{N}^{(t,\tau)} = \sum_{k=1}^K p_k \mathbf{e}_k^{(t,\tau)}$ the stochastic-oracle assumption implies

$$\mathbb{E}\big[\mathbf{N}^{(t,\tau)}\,\big|\,\mathcal{F}_{t,\tau}\big] = \sum_{k=1}^K p_k \mathbb{E}\big[\mathbf{e}_k^{(t,\tau)}\,\big|\,\mathcal{F}_{t,\tau}\big] = \mathbf{0}. \tag{162}$$

Thus $\{\mathbf{N}^{(t,\tau)}\}_\tau$ is a martingale-difference sequence. For $r<s$, $\mathbb{E}\big[\langle \mathbf{N}^{(t,r)}, \mathbf{N}^{(t,s)}\rangle \big| \boldsymbol{\theta}^{(t)}\big] = 0$. Hence,

$$\mathbb{E}\Big[\big\|\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{N}^{(t,\tau)}\big\|_2^2\,\big|\,\boldsymbol{\theta}^{(t)}\Big] = \sum_{\tau=0}^{\mathcal{T}-1}\mathbb{E}\Big[\big\|\mathbf{N}^{(t,\tau)}\big\|_2^2\,\big|\,\boldsymbol{\theta}^{(t)}\Big]. \tag{163}$$

For each τ, convexity of the squared norm yields the pathwise inequality $\big\|\sum_{k=1}^K p_k \mathbf{e}_k^{(t,\tau)}\big\|_2^2 \le \sum_{k=1}^K p_k \big\|\mathbf{e}_k^{(t,\tau)}\big\|_2^2$. Consequently, $\mathbb{E}\big[\big\|\mathbf{N}^{(t,\tau)}\big\|_2^2 \big| \mathcal{F}_{t,\tau}\big] \le \sum_{k=1}^K p_k \sigma_{k,t}^2 = \overline{\sigma}_t^2$. Applying the tower property in (163) and summing over τ proves (161). □

Remark C.23. *The bound in Lemma C.22 does not assume independence among client oracle errors at the same local step. It follows from convexity and is therefore valid under arbitrary within-step dependence consistent with the marginal variance bounds. If the client oracle errors are additionally conditionally uncorrelated given $\mathcal{F}_{t,\tau}$, then the sharper bound $\mathbb{E}\big[\big\|\mathbf{N}^{(t,\tau)}\big\|_2^2 \big| \mathcal{F}_{t,\tau}\big] \le \sum_{k=1}^K p_k^2 \sigma_{k,t}^2$ is available. We retain the more conservative $\overline{\sigma}_t^2 = \sum_k p_k \sigma_{k,t}^2$ in the main theorem to avoid requiring cross-client independence.*

Lemma C.24 (Trajectory/proximal deviation bound)**.** *For every local step τ, it holds*

$$\mathbb{E}\big[\big\|\mathbf{W}^{(t,\tau)}\big\|_2^2 \big| \boldsymbol{\theta}^{(t)}\big] \le (\beta+\mu_t)^2 D_{t,\tau}. \tag{164}$$

Consequently,

$$\mathbb{E}\Big[\big\|\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{W}^{(t,\tau)}\big\|_2^2\,\big|\,\boldsymbol{\theta}^{(t)}\Big] \le \mathcal{T}(\beta+\mu_t)^2 \sum_{\tau=0}^{\mathcal{T}-1} D_{t,\tau}. \tag{165}$$

Proof of Lemma C.24. By the definition of $\mathbf{R}^{(t,\tau)}$ and the triangle inequality,

$$\big\|\mathbf{R}^{(t,\tau)}\big\|_2 \le \sum_{k=1}^K p_k \big\|\nabla F_k\big(\boldsymbol{\theta}_k^{(t,\tau)}\big) - \nabla F_k\big(\boldsymbol{\theta}^{(t)}\big)\big\|_2 \le \beta \sum_{k=1}^K p_k \big\|\mathbf{d}_k^{(t,\tau)}\big\|_2. \tag{166}$$

Since $\sum_k p_k = 1$, weighted Cauchy–Schwarz implies

$$\sum_{k=1}^K p_k \big\|\mathbf{d}_k^{(t,\tau)}\big\|_2 \le \Big[\sum_{k=1}^K p_k \big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2\Big]^{1/2}. \tag{167}$$

Therefore, we can obtain

$$\big\|\mathbf{R}^{(t,\tau)}\big\|_2 \le \beta\Big[\sum_{k=1}^K p_k \big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2\Big]^{1/2}. \tag{168}$$

Similarly, Jensen's inequality gives

$$\big\|\overline{\mathbf{d}}^{(t,\tau)}\big\|_2 = \Big\|\sum_{k=1}^K p_k \mathbf{d}_k^{(t,\tau)}\Big\|_2 \le \Big[\sum_{k=1}^K p_k \big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2\Big]^{1/2}. \tag{169}$$

Using $\mathbf{W}^{(t,\tau)} = \mathbf{R}^{(t,\tau)} + \mu_t \overline{\mathbf{d}}^{(t,\tau)}$ the triangle inequality together with (168) and (169) yields

$$\big\|\mathbf{W}^{(t,\tau)}\big\|_2 \le (\beta+\mu_t)\Big[\sum_{k=1}^{K} p_k \big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2\Big]^{1/2}. \tag{170}$$

Squaring and taking conditional expectation given $\boldsymbol{\theta}^{(t)}$ gives

$$\mathbb{E}\Big[\big\|\mathbf{W}^{(t,\tau)}\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \le (\beta+\mu_t)^2 \sum_{k=1}^{K} p_k \mathbb{E}\Big[\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2 \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] = (\beta+\mu_t)^2 D_{t,\tau}, \tag{171}$$

which proves (164). Finally, Cauchy–Schwarz across the $\mathcal{T}$ local steps gives

$$\Big\|\sum_{\tau=0}^{\mathcal{T}-1} \mathbf{W}^{(t,\tau)}\Big\|_2^2 \le \mathcal{T} \sum_{\tau=0}^{\mathcal{T}-1} \big\|\mathbf{W}^{(t,\tau)}\big\|_2^2. \tag{172}$$

Taking conditional expectation and applying (164) proves (165). □

Remark C.25. *The analysis now provides two complementary controls of $D_{t,\tau}$, that is $D_{t,\tau} \le \eta_t^2 G^2 A_{t,\tau}^2$, and $D_{t,\tau} \le C_t \sum_{r=0}^{\tau-1} \rho_t^r$. The first is a uniform second-moment envelope requiring no explicit heterogeneity calculation. The second exposes how statistical heterogeneity, stochastic-gradient variance, smoothness, and proximal regularization jointly drive the local trajectory. The former is especially convenient for endpoint and partial-participation bounds, whereas the latter is used to expose the ζ^2 dependence of the global stationarity result.*

C.3.2 One-round global descent

We now convert the full-participation update decomposition into a descent bound for the global objective. The result retains the martingale structure of the aggregated stochastic-gradient noise instead of treating the entire server-update deviation as an arbitrary deterministic perturbation. This distinction is important because it yields a stochastic residual of order $\eta_t^2 \overline{\sigma}_t^2$ at the one-round level.

Recall from Lemma C.21 that

$$\boldsymbol{\Delta}^{(t)} = -\eta_t \mathcal{T} \nabla F(\boldsymbol{\theta}^{(t)}) - \eta_t \sum_{\tau=0}^{\mathcal{T}-1} \Big[\mathbf{N}^{(t,\tau)} + \mathbf{W}^{(t,\tau)}\Big]. \tag{173}$$

The following lemma quantifies the resulting one-round descent.

Lemma C.26 (One-round global descent)**.** *Suppose that the local step size satisfies $\beta\eta_t\mathcal{T} \le \frac{1}{6}$. Then the expected global objective after one communication round satisfies*

$$\mathbb{E}\Big[F(\boldsymbol{\theta}^{(t+1)}) \,\Big|\, \boldsymbol{\theta}^{(t)}\Big] \le F(\boldsymbol{\theta}^{(t)}) - \frac{\eta_t\mathcal{T}}{2}\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2 + \frac{5}{4}\eta_t(\beta+\mu_t)^2 \sum_{\tau=0}^{\mathcal{T}-1} D_{t,\tau} + \frac{3}{2}\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2. \tag{174}$$

Proof of Lemma C.26. By Assumption 3.3, the global objective F is β-smooth. Hence, by Lemma C.3 for $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \boldsymbol{\Delta}^{(t)}$, the smoothness inequality gives:

$$F\big(\boldsymbol{\theta}^{(t+1)}\big) \le F\big(\boldsymbol{\theta}^{(t)}\big) + \Big\langle \nabla F\big(\boldsymbol{\theta}^{(t)}\big), \boldsymbol{\Delta}^{(t)} \Big\rangle + \frac{\beta}{2}\Big\|\boldsymbol{\Delta}^{(t)}\Big\|_2^2. \tag{175}$$

We control the first- and second-order terms separately.

First-order term. Substituting (173) in to (175), we have

$$\begin{aligned}\big\langle \nabla F(\boldsymbol{\theta}^{(t)}), \boldsymbol{\Delta}^{(t)} \big\rangle &= -\eta_t\mathcal{T}\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2 \\ &\quad - \eta_t\Big\langle \nabla F(\boldsymbol{\theta}^{(t)}), \sum\nolimits_{\tau=0}^{\mathcal{T}-1} \mathbf{W}^{(t,\tau)} \Big\rangle - \eta_t\Big\langle \nabla F(\boldsymbol{\theta}^{(t)}), \sum_{\tau=0}^{\mathcal{T}-1} \mathbf{N}^{(t,\tau)} \Big\rangle.\end{aligned} \tag{176}$$

By Lemma C.22, $\{\mathbf{N}^{(t,\tau)}\}_\tau$ is a martingale-difference sequence. Thus, $\mathbb{E}\big[\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{N}^{(t,\tau)}\big|\boldsymbol{\theta}^{(t)}\big]=\mathbf{0}$. Since $\nabla F(\boldsymbol{\theta}^{(t)})$ is measurable with respect to $\boldsymbol{\theta}^{(t)}$,

$$\mathbb{E}\Big[\Big\langle\nabla F(\boldsymbol{\theta}^{(t)}),\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{N}^{(t,\tau)}\Big\rangle\,\Big|\,\boldsymbol{\theta}^{(t)}\Big]=0. \tag{177}$$

Therefore, we can obtain

$$\mathbb{E}\Big[\big\langle\nabla F(\boldsymbol{\theta}^{(t)}),\boldsymbol{\Delta}^{(t)}\big\rangle\,\Big|\,\boldsymbol{\theta}^{(t)}\Big]=-\eta_t\mathcal{T}\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2-\eta_t\mathbb{E}\Big[\Big\langle\nabla F(\boldsymbol{\theta}^{(t)}),\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{W}^{(t,\tau)}\Big\rangle\,\Big|\,\boldsymbol{\theta}^{(t)}\Big]. \tag{178}$$

To control the remaining inner product, by applying Young's inequality $-\langle\mathbf{a},\mathbf{b}\rangle\le\frac{\lambda}{2}\|\mathbf{a}\|_2^2+\frac{1}{2\lambda}\|\mathbf{b}\|_2^2$ with $\lambda=\mathcal{T}/2$ we have

$$-\eta_t\Big\langle\nabla F(\boldsymbol{\theta}^{(t)}),\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{W}^{(t,\tau)}\Big\rangle\le\frac{\eta_t\mathcal{T}}{4}\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2+\frac{\eta_t}{\mathcal{T}}\big\|\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{W}^{(t,\tau)}\big\|_2^2. \tag{179}$$

Hence, we obtain

$$\mathbb{E}\Big[\big\langle\nabla F(\boldsymbol{\theta}^{(t)}),\boldsymbol{\Delta}^{(t)}\big\rangle\,\Big|\,\boldsymbol{\theta}^{(t)}\Big]\le-\frac{3}{4}\eta_t\mathcal{T}\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2+\frac{\eta_t}{\mathcal{T}}\mathbb{E}\Big[\big\|\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{W}^{(t,\tau)}\big\|_2^2\,\Big|\,\boldsymbol{\theta}^{(t)}\Big]. \tag{180}$$

Second-order smoothness term. From (173), we have

$$\big\|\boldsymbol{\Delta}^{(t)}\big\|_2^2=\eta_t^2\Big\|\mathcal{T}\nabla F(\boldsymbol{\theta}^{(t)})+\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{W}^{(t,\tau)}+\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{N}^{(t,\tau)}\Big\|_2^2. \tag{181}$$

Applying the Cauchy-Schwarz inequality

$$\big\|\mathbf{a}+\mathbf{b}+\mathbf{c}\big\|_2^2\le3\left(\|\mathbf{a}\|_2^2+\|\mathbf{b}\|_2^2+\|\mathbf{c}\|_2^2\right), \tag{182}$$

we then obtain

$$\begin{aligned}\frac{\beta}{2}\mathbb{E}\Big[\big\|\boldsymbol{\Delta}^{(t)}\big\|_2^2\,\Big|\,\boldsymbol{\theta}^{(t)}\Big]\le\frac{3}{2}\beta\eta_t^2\Big[\mathcal{T}^2\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2\\+\mathbb{E}\Big[\big\|\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{W}^{(t,\tau)}\big\|_2^2\,\Big|\,\boldsymbol{\theta}^{(t)}\Big]+\mathbb{E}\Big[\big\|\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{N}^{(t,\tau)}\big\|_2^2\,\Big|\,\boldsymbol{\theta}^{(t)}\Big]\Big].\end{aligned} \tag{183}$$

We use the deterministic three-term inequality here rather than discarding a cross term between $\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{W}^{(t,\tau)}$ and $\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{N}^{(t,\tau)}$. Although $\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{N}^{(t,\tau)}$ is a martingale sum, $\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{W}^{(t,\tau)}$ is constructed from the stochastic local trajectories and can depend on previous oracle errors. Thus, in general, $\mathbb{E}\big[\big\langle\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{W}^{(t,\tau)},\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{N}^{(t,\tau)}\big\rangle\big|\boldsymbol{\theta}^{(t)}\big]$ need not vanish under the stated assumptions.

Combining the two terms. Combining (180) and (183), the coefficient of $\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2$ is $-\frac{3}{4}\eta_t\mathcal{T}+\frac{3}{2}\beta\eta_t^2\mathcal{T}^2$. Under the local step size condition , we have

$$-\frac{3}{4}\eta_t\mathcal{T}+\frac{3}{2}\beta\eta_t^2\mathcal{T}^2\le-\frac{1}{2}\eta_t\mathcal{T}. \tag{184}$$

The coefficient multiplying $\mathbb{E}\big[\big\|\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{W}^{(t,\tau)}\big\|_2^2\,\big|\,\boldsymbol{\theta}^{(t)}\big]$ is $\frac{\eta_t}{\mathcal{T}}+\frac{3}{2}\beta\eta_t^2$. By using $\beta\eta_t\mathcal{T}\le1/6$, $\frac{\eta_t}{\mathcal{T}}+\frac{3}{2}\beta\eta_t^2\le\frac{5\eta_t}{4\mathcal{T}}$. By Lemma C.24, we have

$$\mathbb{E}\Big[\big\|\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{W}^{(t,\tau)}\big\|_2^2\,\Big|\,\boldsymbol{\theta}^{(t)}\Big]\le\mathcal{T}(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1}D_{t,\tau}. \tag{185}$$

Therefore, we can obtain

$$\Big(\frac{\eta_t}{\mathcal{T}}+\frac{3}{2}\beta\eta_t^2\Big)\mathbb{E}\Big[\big\|\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{W}^{(t,\tau)}\big\|_2^2\,\Big|\,\boldsymbol{\theta}^{(t)}\Big]\le\frac{5}{4}\eta_t(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1}D_{t,\tau}. \tag{186}$$

Finally, Lemma C.22 yields $\mathbb{E}\big[\big\|\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{N}^{(t,\tau)}\big\|_2^2\big|\boldsymbol{\theta}^{(t)}\big]\le\mathcal{T}\overline{\sigma}_t^2$ and consequently,

$$\frac{3}{2}\beta\eta_t^2\mathbb{E}\Big[\big\|\sum_{\tau=0}^{\mathcal{T}-1}\mathbf{N}^{(t,\tau)}\big\|_2^2\,\Big|\,\boldsymbol{\theta}^{(t)}\Big]\le\frac{3}{2}\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2. \tag{187}$$

Substituting (184), (186), and (187) into the smoothness inequality proves (174). □

Remark C.27. *Lemma C.26 separates the one-round optimization error into three components:*

$$\underbrace{-\frac{1}{2}\eta_t\mathcal{T}\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2}_{\textit{global descent}}+\underbrace{\frac{5}{4}\eta_t(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1}D_{t,\tau}}_{\textit{local trajectory/proximal error}}+\underbrace{\frac{3}{2}\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2}_{\textit{stochastic oracle error}}. \tag{188}$$

The heterogeneity parameter ζ^2 does not appear as an independent common-iterate error term. Instead, it affects the first residual indirectly through the evolution of $D_{t,\tau}$.

Remark C.28. *Using $\overline{\sigma}_t^2=\sum_{k=1}^K p_k\big[\frac{\sigma_{\mathrm{mb},k}^2}{B}+\frac{c_{\mathrm{ps},k}d_q}{S_{k,t}}\big]$ the direct stochastic residual in Lemma C.26 becomes $\frac{3}{2}\beta\eta_t^2\mathcal{T}\sum_{k=1}^K p_k\frac{\sigma_{\mathrm{mb},k}^2}{B}+\frac{3}{2}\beta\eta_t^2\mathcal{T}\sum_{k=1}^K p_k\frac{c_{\mathrm{ps},k}d_q}{S_{k,t}}$. Thus, for fixed model dimension, the direct quantum-measurement contribution decreases as $S_{k,t}^{-1}$ and is multiplied by η_t^2 at the one-round level. It should therefore be interpreted as a stepsize-dependent stochastic residual rather than an unconditional $O(S^{-1})$ optimization floor.*

C.4 Nonconvex Stationarity under Full Participation

C.4.1 Drift-explicit stationarity guarantee

We first derive a global stationarity guarantee while leaving the client-drift energy explicit. This separates the server-level descent argument from the particular assumptions used to control the local trajectories. Define the weighted stationarity measure by

$$\mathcal{G}_T:=\frac{1}{\sum_{t=0}^{T-1}\eta_t\mathcal{T}}\sum_{t=0}^{T-1}\eta_t\mathcal{T}\,\mathbb{E}\Big[\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2\Big]. \tag{189}$$

We assume throughout that the stepsizes $\eta_t>0$ are deterministic, so that the denominator in (189) is positive and nonrandom. Since $D_{t,\tau}$ denotes the drift energy conditional on the broadcast iterate, it is convenient at the global level to define its unconditional counterpart

$$\overline{D}_{t,\tau}:=\mathbb{E}\big[D_{t,\tau}\big]=\sum_{k=1}^K p_k\,\mathbb{E}\Big[\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2\Big]. \tag{190}$$

Theorem C.29 (Full-participation nonconvex stationarity)**.** *Suppose the local objectives are β-smooth, the global objective is lower bounded by $F^\star$, and Assumption C.1 holds. Suppose further that $\beta\eta_t\mathcal{T}\le\frac{1}{6}$ for every communication round t. Then*

$$\mathcal{G}_T\le\frac{2\big[F(\boldsymbol{\theta}^{(0)})-F^\star\big]}{\sum_{t=0}^{T-1}\eta_t\mathcal{T}}+\frac{\sum_{t=0}^{T-1}\big[\frac{5}{2}\eta_t(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1}\overline{D}_{t,\tau}+3\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2\big]}{\sum_{t=0}^{T-1}\eta_t\mathcal{T}}. \tag{191}$$

Proof of Theorem C.29. Lemma C.26 gives, conditional on the broadcast iterate,

$$\mathbb{E}\big[F(\boldsymbol{\theta}^{(t+1)})\,\big|\,\boldsymbol{\theta}^{(t)}\big]\le F(\boldsymbol{\theta}^{(t)})-\frac{\eta_t\mathcal{T}}{2}\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2+\frac{5}{4}\eta_t(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1}D_{t,\tau}+\frac{3}{2}\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2. \tag{192}$$

Taking total expectation and using the tower property yields

$$\mathbb{E}\big[F(\boldsymbol{\theta}^{(t+1)})\big]\le\mathbb{E}\big[F(\boldsymbol{\theta}^{(t)})\big]-\frac{\eta_t\mathcal{T}}{2}\mathbb{E}\Big[\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2\Big]+\frac{5}{4}\eta_t(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1}\mathbb{E}\big[D_{t,\tau}\big]+\frac{3}{2}\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2 \tag{193}$$

$$=\mathbb{E}\big[F(\boldsymbol{\theta}^{(t)})\big]-\frac{\eta_t\mathcal{T}}{2}\mathbb{E}\Big[\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2\Big]+\frac{5}{4}\eta_t(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1}\overline{D}_{t,\tau}+\frac{3}{2}\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2. \tag{194}$$

Rearranging and summing over $t = 0, \dots, T-1$ we can obtain

$$\frac{1}{2}\sum_{t=0}^{T-1}\eta_t\mathcal{T}\mathbb{E}\Big[\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2\Big] \le F(\boldsymbol{\theta}^{(0)}) - \mathbb{E}\big[F(\boldsymbol{\theta}^{(T)})\big] + \sum_{t=0}^{T-1}\Big[\frac{5}{4}\eta_t(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1}\overline{D}_{t,\tau} + \frac{3}{2}\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2\Big]. \quad (195)$$

Since $F(\boldsymbol{\theta}^{(T)}) \ge F^\star$ almost surely, it follows that $\mathbb{E}\big[F(\boldsymbol{\theta}^{(T)})\big] \ge F^\star$. Therefore, we can obtain

$$\frac{1}{2}\sum_{t=0}^{T-1}\eta_t\mathcal{T}\mathbb{E}\Big[\Big\|\nabla F(\boldsymbol{\theta}^{(t)})\Big\|_2^2\Big] \le F(\boldsymbol{\theta}^{(0)}) - F^\star + \sum_{t=0}^{T-1}\Big[\frac{5}{4}\eta_t(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1}\overline{D}_{t,\tau} + \frac{3}{2}\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2\Big]. \quad (196)$$

Multiplying by two and dividing by $\sum_{t=0}^{T-1}\eta_t\mathcal{T}$ proves (191). □

Theorem C.29 establishes first-order stationarity for a smooth, lower-bounded, potentially nonconvex global objective. In particular, the theorem does not require convexity, strong convexity, or a Polyak–Łojasiewicz condition. It deliberately leaves $\overline{D}_{t,\tau}$ explicit and therefore does not require the gradient-dissimilarity assumption involving ζ^2. The heterogeneity assumption is introduced only when a more explicit upper bound on the client-drift energy is required. This separates the global descent mechanism from the particular model used to control local trajectory divergence.

C.4.2 Explicit finite-shot and heterogeneity dependence

Theorem C.29 retains the local drift energy explicitly in the stationarity bound. We now apply the heterogeneity-sensitive drift recursion established in Lemma C.18 to make explicit its dependence on statistical heterogeneity and stochastic-gradient variance. For each round t, fix $\gamma_t > 0$ and define $\Phi_t := \sum_{\tau=0}^{\mathcal{T}-1}\sum_{r=0}^{\tau-1}\rho_t^r$ where an empty inner sum is interpreted as zero. Equivalently, $\Phi_t = \sum_{r=0}^{\mathcal{T}-2}(\mathcal{T}-1-r)\rho_t^r$. In particular, $\Phi_t = \frac{\mathcal{T}(\mathcal{T}-1)}{2}$ when $\rho_t = 1$. From (138), $D_{t,\tau} \le C_t\sum_{r=0}^{\tau-1}\rho_t^r$ and therefore $\sum_{\tau=0}^{\mathcal{T}-1}D_{t,\tau} \le C_t\Phi_t$ Since C_t contains the current global-gradient norm, part of this trajectory term must be absorbed into the descent term. The following condition ensures that the remaining descent coefficient stays strictly negative.

Corollary C.30 (Heterogeneity-explicit nonconvex stationarity)**.** *Suppose the conditions of Theorem C.29 hold and Assumption 3.6 is satisfied. Suppose further that for each round t there exists $\gamma_t > 0$ such that $5\eta_t^2(\beta+\mu_t)^2\Phi_t\big(1+\frac{1}{\gamma_t}\big) \le \mathcal{T}$. Then,*

$$\mathcal{G}_T \le \frac{4\big[F(\boldsymbol{\theta}^{(0)}) - F^\star\big]}{\sum_{t=0}^{T-1}\eta_t\mathcal{T}} + \frac{4\sum_{t=0}^{T-1}\mathcal{R}_t}{\sum_{t=0}^{T-1}\eta_t\mathcal{T}}, \quad (197)$$

where $\mathcal{R}_t := \frac{3}{2}\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2 + \frac{5}{4}\eta_t^3(\beta+\mu_t)^2\Phi_t\big[\big(1+\frac{1}{\gamma_t}\big)\zeta^2 + \overline{\sigma}_t^2\big]$.

Proof of Corollary C.30. Lemma C.18 gives, conditional on $\boldsymbol{\theta}^{(t)}$,

$$\sum_{\tau=0}^{\mathcal{T}-1}D_{t,\tau} \le \eta_t^2\Phi_t\Big[\big(1+\frac{1}{\gamma_t}\big)\big(\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2 + \zeta^2\big) + \overline{\sigma}_t^2\Big]. \quad (198)$$

Substituting (198) directly into the conditional one-round descent inequality 174 yields

$$\begin{aligned}\mathbb{E}\big[F(\boldsymbol{\theta}^{(t+1)}) \,\big|\, \boldsymbol{\theta}^{(t)}\big] &\le F(\boldsymbol{\theta}^{(t)}) - \frac{\eta_t\mathcal{T}}{2}\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2 \\ &\quad + \frac{5}{4}\eta_t^3(\beta+\mu_t)^2\Phi_t\big(1+\frac{1}{\gamma_t}\big)\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2 + \mathcal{R}_t.\end{aligned} \quad (199)$$

The absorption condition implies $\frac{5}{4}\eta_t^3(\beta+\mu_t)^2\Phi_t\big(1+\frac{1}{\gamma_t}\big) \le \frac{1}{4}\eta_t\mathcal{T}$. Consequently, we can obtain

$$\mathbb{E}\big[F(\boldsymbol{\theta}^{(t+1)}) \,\big|\, \boldsymbol{\theta}^{(t)}\big] \le F(\boldsymbol{\theta}^{(t)}) - \frac{\eta_t\mathcal{T}}{4}\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2 + \mathcal{R}_t. \quad (200)$$

Taking total expectation then gives

$$\mathbb{E}\big[F(\boldsymbol{\theta}^{(t+1)})\big] \le \mathbb{E}\big[F(\boldsymbol{\theta}^{(t)})\big] - \frac{\eta_t\mathcal{T}}{4}\mathbb{E}\Big[\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2\Big] + \mathcal{R}_t. \quad (201)$$

Summing over $t = 0, \dots, T-1$ and using $F(\boldsymbol{\theta}^{(T)}) \geq F^\star$ gives

$$\frac{1}{4}\sum_{t=0}^{T-1} \eta_t \mathcal{T} \mathbb{E}\Big[\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2\Big] \leq F(\boldsymbol{\theta}^{(0)}) - F^\star + \sum_{t=0}^{T-1} \mathcal{R}_t. \tag{202}$$

Multiplying by four and normalizing by $\sum_{t=0}^{T-1} \eta_t \mathcal{T}$ proves (197). □

The residual in Corollary C.30 separates the direct stochastic oracle contribution from the stochastic and heterogeneous perturbations propagated through the local trajectories. In particular, $\overline{\sigma}_t^2 = \sum_{k=1}^K p_k \big[\frac{\sigma_{\mathrm{mb},k}^2}{B} + \frac{c_{\mathrm{ps},k} d_q}{S_{k,t}}\big]$. Hence finite-shot variance appears both through the direct stochastic term $\frac{3}{2}\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2$ and through the drift-propagation term $\frac{5}{4}\eta_t^3(\beta+\mu_t)^2\Phi_t\overline{\sigma}_t^2$. Statistical heterogeneity enters only through $\frac{5}{4}\eta_t^3(\beta+\mu_t)^2\Phi_t\big(1+\frac{1}{\gamma_t}\big)\zeta^2$. Thus neither finite-shot noise nor client heterogeneity appears as an unconditional additive floor independent of the optimization dynamics. Their contributions depend explicitly on the learning rate, local horizon, smoothness, and proximal schedule.

C.4.3 Constant-schedule specialization

To expose the principal scaling more transparently, suppose $\eta_t = \eta, \mu_t = \mu, \gamma_t = \gamma, S_{k,t} = S_k$, for all communication rounds. Then $\rho_t = \rho$ and $\Phi_t = \Phi$ are constant across rounds, where $\rho = (1+\gamma)\,(1-\eta\mu+\eta\beta)^2$. Under the conditions of Corollary C.30,

$$\mathcal{G}_T = O\Big(\frac{F(\boldsymbol{\theta}^{(0)}) - F^\star}{T\eta\mathcal{T}}\Big) + O\left(\beta\eta\overline{\sigma}^2\right) + O\Big(\eta^2\frac{(\beta+\mu)^2\Phi}{\mathcal{T}}\Big[\big(1+\frac{1}{\gamma}\big)\zeta^2 + \overline{\sigma}^2\Big]\Big), \tag{203}$$

where $\overline{\sigma}^2 = \sum_{k=1}^K p_k\big[\frac{\sigma_{\mathrm{mb},k}^2}{B} + \frac{c_{\mathrm{ps},k} d_q}{S_k}\big]$. If all clients use a common shot budget S, the direct finite-shot contribution has the scaling $O\big(\beta\eta\frac{d_q}{S}\sum_{k=1}^K p_k c_{\mathrm{ps},k}\big) = O\big(\eta\frac{d_q}{S}\big)$ when β and the weighted parameter-shift variance constant are absorbed into the problem-dependent constant. In addition, finite-shot variance enters through the trajectory term with scaling $O\big(\eta^2\frac{(\beta+\mu)^2\Phi}{\mathcal{T}}\frac{d_q}{S}\big)$. Therefore, the analysis does not predict an unconditional $O(S^{-1})$ optimization floor. For a fixed stepsize, finite-shot measurements contribute to a fixed-stepsize stationarity neighborhood whose shot-dependent part decreases inversely with S. Under a diminishing stepsize schedule satisfying the theorem conditions, these stochastic residuals can decrease with the stepsize as well.

C.5 Partial Client Participation

C.5.1 Sampling Model and Potential-Trajectory Coupling

We extend the convergence analysis to the setting of partial client participation, where only a subset of clients contributes to the server update at each communication round. To obtain an exact finite-population sampling identity, this subsection specializes to equal client weights, $p_k = \frac{1}{K}, k = 1, \dots, K$. Accordingly, the server aggregates the selected client endpoints by their uniform sample mean. At communication round t, let

$$\mathcal{S}^{(t)} \subseteq \mathcal{K} = \{1, \dots, K\}, \quad \big|\mathcal{S}^{(t)}\big| = m, \quad 1 \leq m \leq K, \tag{204}$$

where $\mathcal{S}^{(t)}$ is sampled uniformly without replacement.

A technical issue under partial participation is that the realized algorithm executes local optimization only on the selected clients, whereas the finite-population sampling identity is naturally formulated for a fixed collection of K candidate endpoints. To bridge these two views, we use an analytical *potential-trajectory coupling*.

Conditional on the broadcast iterate $\boldsymbol{\theta}^{(t)}$, associate each client k with the local stochastic trajectory that it would generate if it were selected at round t. Let $\mathbf{d}_k^{(t,\mathcal{T})} := \boldsymbol{\theta}_k^{(t,\mathcal{T})} - \boldsymbol{\theta}^{(t)}$ denote the corresponding potential endpoint client drift (displacement). Let $\mathcal{U}_t$ collect the stochastic-oracle randomness used to generate the K potential local trajectories. We assume that the uniform subset draw is conditionally independent of this potential local-update randomness:

$$\mathcal{S}^{(t)} \;\perp\!\!\!\perp\; \mathcal{U}_t \mid \boldsymbol{\theta}^{(t)}. \tag{205}$$

Equivalently, conditional on the broadcast iterate and the complete collection of potential endpoints, the subset $\mathcal{S}^{(t)}$ remains a uniformly sampled size-m subset of $\{1,\dots,K\}$. Importantly, this construction does not require the potential trajectories of different clients to be mutually independent. Their stochastic-oracle randomness may be dependent; the sampling argument below requires only that the client-selection draw be conditionally independent of the collection of potential trajectories.

The coupling is purely analytical and does not require unselected clients to perform any computation in the actual algorithm. Provided that a client's local stochastic-update law, conditional on $\boldsymbol{\theta}^{(t)}$, does not depend on the identities of the other selected clients, the trajectories $\left\{\mathbf{d}_k^{(t,\mathcal{T})} : k \in \mathcal{S}^{(t)}\right\}$ obtained by revealing the selected subset from the potential collection have the same joint distribution as the local trajectories generated by executing the partial-participation algorithm only on the selected clients. We can therefore define the potential full-population endpoint increment as $\overline{\boldsymbol{\Delta}}^{(t)} := \frac{1}{K}\sum_{k=1}^{K}\mathbf{d}_k^{(t,\mathcal{T})}$ and the realized partial-participation increment as $\boldsymbol{\Delta}_{\mathrm{pp}}^{(t)} := \frac{1}{m}\sum_{k\in\mathcal{S}^{(t)}}\mathbf{d}_k^{(t,\mathcal{T})}$ Under the equal-weight specialization $p_k = 1/K$, $\overline{\boldsymbol{\Delta}}^{(t)}$ is exactly the server increment that would be obtained under full participation for the same realization of the potential local trajectories.

C.5.2 Finite-Population Identity and Endpoint Deviation

The potential-trajectory construction reduces client subsampling to uniform sampling from a fixed finite collection of endpoint displacements. The following identity characterizes the resulting conditional mean and variance.

Lemma C.31 (Finite-population sampling identity)**.** *Let $\mathbf{v}_1,\dots,\mathbf{v}_K \in \mathbb{R}^d$ be deterministic vectors and define their population mean as $\overline{\mathbf{v}} := \frac{1}{K}\sum_{k=1}^{K}\mathbf{v}_k$. Let $\mathcal{S}$ be sampled uniformly without replacement from $\{1,\dots,K\}$ with $|\mathcal{S}| = m$. Then the sample mean is unbiased, i.e., $\mathbb{E}_{\mathcal{S}}\left[\frac{1}{m}\sum_{k\in\mathcal{S}}\mathbf{v}_k\right] = \overline{\mathbf{v}}$. Moreover, its mean-squared sampling error satisfies*

$$\mathbb{E}_{\mathcal{S}}\left[\left\|\frac{1}{m}\sum_{k\in\mathcal{S}}\mathbf{v}_k - \overline{\mathbf{v}}\right\|_2^2\right] = \frac{K-m}{m(K-1)}\frac{1}{K}\sum_{k=1}^{K}\left\|\mathbf{v}_k - \overline{\mathbf{v}}\right\|_2^2. \tag{206}$$

Proof of Lemma C.31. For each client index k, define the inclusion indicator as

$$I_k := \mathbf{1}\{k \in \mathcal{S}\}. \tag{207}$$

Since $\mathcal{S}$ is sampled uniformly without replacement, the probability that a fixed client k is selected is $\mathbb{P}(k \in \mathcal{S}) = \frac{m}{K}$ so the first-order inclusion probability satisfies

$$\mathbb{E}_{\mathcal{S}}[I_k] = \frac{m}{K} \tag{208}$$

Similarly, for distinct $i \neq j$, the probability that both clients are selected is:

$$\mathbb{P}(i,j \in \mathcal{S}^{(t)}) = \frac{m(m-1)}{K(K-1)}. \tag{209}$$

Therefore, we have

$$\mathbb{E}_{\mathcal{S}}[I_i\, I_j] = \frac{m(m-1)}{K(K-1)}, \quad i \neq j. \tag{210}$$

For deterministic vectors $\mathbf{v}_k \in \mathbb{R}^d$, it can be expressed using the indicator (207) as:

$$\frac{1}{m}\sum_{k\in\mathcal{S}}\mathbf{v}_k = \frac{1}{m}\sum_{k=1}^{K}I_k\mathbf{v}_k. \tag{211}$$

Taking expectations and using linearity, we can obtain:

$$\mathbb{E}_{\mathcal{S}}\left[\frac{1}{m}\sum_{k=1}^{K}I_k\mathbf{v}_k\right] = \frac{1}{m}\sum_{k=1}^{K}\mathbb{E}_{\mathcal{S}}[I_k]\mathbf{v}_k = \frac{1}{m}\sum_{k=1}^{K}\frac{m}{K}\mathbf{v}_k = \frac{1}{K}\sum_{k=1}^{K}\mathbf{v}_k = \overline{\mathbf{v}}. \tag{212}$$

For the second-moment identity, define the centered vectors $\widetilde{\mathbf{v}}_k := \mathbf{v}_k - \overline{\mathbf{v}}$. Then, it follows that:

$$\frac{1}{m}\sum_{k\in\mathcal{S}}\mathbf{v}_k - \overline{\mathbf{v}} = \frac{1}{m}\sum_{k\in\mathcal{S}}\widetilde{\mathbf{v}}_k = \frac{1}{m}\sum_{k=1}^{K} I_k\widetilde{\mathbf{v}}_k. \tag{213}$$

Therefore, taking the squared Euclidean norm and then the expectation gives

$$\mathbb{E}_{\mathcal{S}}\Big[\Big\|\frac{1}{m}\sum_{k\in\mathcal{S}}\mathbf{v}_k - \overline{\mathbf{v}}\Big\|_2^2\Big] = \frac{1}{m^2}\mathbb{E}_{\mathcal{S}}\Big[\Big\|\sum_{k=1}^{K} I_k\widetilde{\mathbf{v}}_k\Big\|_2^2\Big] \tag{214}$$

$$= \frac{1}{m^2}\Big[\sum_{k=1}^{K}\mathbb{E}_{\mathcal{S}}[I_k]\big\|\widetilde{\mathbf{v}}_k\big\|_2^2 + \sum_{\substack{i,j=1\\ i\neq j}}\mathbb{E}_{\mathcal{S}}[I_i\, I_j]\,\langle\widetilde{\mathbf{v}}_i, \widetilde{\mathbf{v}}_j\rangle\Big]. \tag{215}$$

By construction of centered vector, $\sum_{k=1}^{K}\widetilde{\mathbf{v}}_k = \mathbf{0}$, we have

$$0 = \Big\|\sum_{k=1}^{K}\widetilde{\mathbf{v}}_k\Big\|_2^2 = \sum_{k=1}^{K}\big\|\widetilde{\mathbf{v}}_k\big\|_2^2 + \sum_{\substack{i,j=1\\ i\neq j}}\langle\widetilde{\mathbf{v}}_i, \widetilde{\mathbf{v}}_j\rangle\,, \tag{216}$$

and consequently,

$$\sum_{i\neq j}\langle\widetilde{\mathbf{v}}_i, \widetilde{\mathbf{v}}_j\rangle = -\sum_{k=1}^{K}\big\|\widetilde{\mathbf{v}}_k\big\|_2^2. \tag{217}$$

Substituting (208), (210), and (217), we can obtain

$$\mathbb{E}_{\mathcal{S}}\Big[\Big\|\frac{1}{m}\sum_{k\in\mathcal{S}}\mathbf{v}_k - \overline{\mathbf{v}}\Big\|_2^2\Big] = \frac{1}{m^2}\left[\frac{m}{K} - \frac{m(m-1)}{K(K-1)}\right]\sum_{k=1}^{K}\|\widetilde{\mathbf{v}}_k\|_2^2 \tag{218}$$

$$= \frac{K-m}{mK(K-1)}\sum_{k=1}^{K}\big\|\widetilde{\mathbf{v}}_k\big\|_2^2 \tag{219}$$

$$= \frac{K-m}{m(K-1)}\frac{1}{K}\sum_{k=1}^{K}\big\|\mathbf{v}_k - \overline{\mathbf{v}}\big\|_2^2, \tag{220}$$

which proves (206). □

For compactness, define the finite-population sampling factor $c_m := \frac{K-m}{m(K-1)}$. Observe that $c_K = 0, c_1 = 1$ so the sampling penalty vanishes under full participation and is largest when only one client is sampled.

Lemma C.32 (Partial-participation endpoint deviation)**.** *Suppose $p_k = 1/K$ and $\mathcal{S}^{(t)}$ is sampled uniformly without replacement with $|\mathcal{S}^{(t)}| = m$. Under the potential-trajectory coupling, with the subset draw conditionally independent of the potential local-update randomness given $\boldsymbol{\theta}^{(t)}$, the deviation of the realized partial-participation increment from the corresponding potential full-population increment satisfies*

$$\mathbb{E}\Big[\big\|\boldsymbol{\Delta}_{\mathrm{pp}}^{(t)} - \overline{\boldsymbol{\Delta}}^{(t)}\big\|_2^2\,\Big|\,\boldsymbol{\theta}^{(t)}\Big] \le c_m D_{t,\mathcal{T}}, \tag{221}$$

where $c_m = \frac{K-m}{m(K-1)}$ is finite-population sampling factor and under the equal-weight specialization $D_{t,\mathcal{T}} := \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\big[\big\|\mathbf{d}_k^{(t,\mathcal{T})}\big\|_2^2\big|\boldsymbol{\theta}^{(t)}\big]$ is the weighted endpoint client-drift energy.

Proof of Lemma C.32. Let $\mathcal{H}_t := \sigma\big(\boldsymbol{\theta}^{(t)}, \{\mathbf{d}_k^{(t,\mathcal{T})}\}_{k=1}^{K}\big)$ denote the sigma-field generated by the broadcast iterate and the complete collection of potential endpoint client drifts. Conditional on $\mathcal{H}_t$, the vectors $\{\mathbf{d}_k^{(t,\mathcal{T})}\}_{k=1}^{K}$ are fixed. Moreover, by the potential-trajectory coupling and the conditional

independence between the subset draw and the potential local-update randomness, the conditional law of $\mathcal{S}^{(t)}$ given $\mathcal{H}_t$ remains uniform over all size-m subsets of $\{1,\dots,K\}$. Equivalently, for every $\mathcal{S}\subseteq\{1,\dots,K\}$ with $|\mathcal{S}|=m$,

$$\mathbb{P}\left(\mathcal{S}^{(t)}=\mathcal{S}\,\middle|\,\mathcal{H}_t\right)=\binom{K}{m}^{-1}. \tag{222}$$

Therefore, applying the unbiasedness part of Lemma C.31 conditionally on $\mathcal{H}_t$ gives

$$\mathbb{E}\left[\boldsymbol{\Delta}_{\mathrm{pp}}^{(t)}\,\middle|\,\mathcal{H}_t\right]=\frac{1}{K}\sum_{k=1}^{K}\mathbf{d}_k^{(t,\mathcal{T})}=\overline{\boldsymbol{\Delta}}^{(t)}. \tag{223}$$

Hence the partial-participation sampling fluctuation $\boldsymbol{\varepsilon}_{\mathrm{pp}}^{(t)}:=\boldsymbol{\Delta}_{\mathrm{pp}}^{(t)}-\overline{\boldsymbol{\Delta}}^{(t)}$ is conditionally mean zero

$$\mathbb{E}\left[\boldsymbol{\varepsilon}_{\mathrm{pp}}^{(t)}\,\middle|\,\mathcal{H}_t\right]=\mathbf{0}. \tag{224}$$

Applying the second-moment identity of Lemma C.31 conditionally on $\mathcal{H}_t$ yields

$$\mathbb{E}\Big[\big\|\boldsymbol{\Delta}_{\mathrm{pp}}^{(t)}-\overline{\boldsymbol{\Delta}}^{(t)}\big\|_2^2\,\Big|\,\mathcal{H}_t\Big]=c_m\frac{1}{K}\sum_{k=1}^{K}\big\|\mathbf{d}_k^{(t,\mathcal{T})}-\overline{\boldsymbol{\Delta}}^{(t)}\big\|_2^2. \tag{225}$$

Since $\overline{\boldsymbol{\Delta}}^{(t)}=\frac{1}{K}\sum_{k=1}^{K}\mathbf{d}_k^{(t,\mathcal{T})}$, the standard finite-population variance decomposition yields

$$\frac{1}{K}\sum_{k=1}^{K}\big\|\mathbf{d}_k^{(t,\mathcal{T})}-\overline{\boldsymbol{\Delta}}^{(t)}\big\|_2^2=\frac{1}{K}\sum_{k=1}^{K}\big\|\mathbf{d}_k^{(t,\mathcal{T})}\big\|_2^2-\big\|\overline{\boldsymbol{\Delta}}^{(t)}\big\|_2^2. \tag{226}$$

Since the second term on the right-hand side is nonnegative,

$$\frac{1}{K}\sum_{k=1}^{K}\big\|\mathbf{d}_k^{(t,\mathcal{T})}-\overline{\boldsymbol{\Delta}}^{(t)}\big\|_2^2\le\frac{1}{K}\sum_{k=1}^{K}\big\|\mathbf{d}_k^{(t,\mathcal{T})}\big\|_2^2. \tag{227}$$

Combining (225) and (227), and then taking conditional expectation given $\boldsymbol{\theta}^{(t)}$, gives

$$\mathbb{E}\Big[\big\|\boldsymbol{\Delta}_{\mathrm{pp}}^{(t)}-\overline{\boldsymbol{\Delta}}^{(t)}\big\|_2^2\,\Big|\,\boldsymbol{\theta}^{(t)}\Big]=\mathbb{E}\Big[\mathbb{E}\Big[\big\|\boldsymbol{\Delta}_{\mathrm{pp}}^{(t)}-\overline{\boldsymbol{\Delta}}^{(t)}\big\|_2^2\,\Big|\,\mathcal{H}_t\Big]\,\Big|\,\boldsymbol{\theta}^{(t)}\Big] \tag{228}$$

$$\le c_m\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\big\|\mathbf{d}_k^{(t,\mathcal{T})}\big\|_2^2\,\Big|\,\boldsymbol{\theta}^{(t)}\Big]. \tag{229}$$

Under the equal-weight specialization $p_k=1/K$, the right-hand side is exactly $c_mD_{t,\mathcal{T}}$. Therefore,

$$\mathbb{E}\Big[\big\|\boldsymbol{\Delta}_{\mathrm{pp}}^{(t)}-\overline{\boldsymbol{\Delta}}^{(t)}\big\|_2^2\,\Big|\,\boldsymbol{\theta}^{(t)}\Big]\le c_mD_{t,\mathcal{T}}, \tag{230}$$

which proves (221). □

The endpoint drift envelope now yields an explicit upper bound on the additional update fluctuation induced by partial client participation. By Corollary C.13, $D_{t,\mathcal{T}}\le\eta_t^2G^2A_{t,\mathcal{T}}^2$. Combining this result with Lemma C.32 gives

$$\mathbb{E}\left[\big\|\boldsymbol{\Delta}_{\mathrm{pp}}^{(t)}-\overline{\boldsymbol{\Delta}}^{(t)}\big\|_2^2\,\middle|\,\boldsymbol{\theta}^{(t)}\right]\le c_m\eta_t^2G^2A_{t,\mathcal{T}}^2. \tag{231}$$

When $\mu_t>0$ and $0\le\eta_t\mu_t\le1$, the geometric-memory bound $A_{t,\mathcal{T}}\le\min\left\{\mathcal{T},\frac{1}{\eta_t\mu_t}\right\}$ implies $A_{t,\mathcal{T}}^2\le\min\left\{\mathcal{T}^2,\frac{1}{(\eta_t\mu_t)^2}\right\}$. Therefore, we can obtain

$$\mathbb{E}\Big[\big\|\boldsymbol{\Delta}_{\mathrm{pp}}^{(t)}-\overline{\boldsymbol{\Delta}}^{(t)}\big\|_2^2\,\Big|\,\boldsymbol{\theta}^{(t)}\Big]\le c_m\eta_t^2G^2\min\left\{\mathcal{T}^2,\frac{1}{(\eta_t\mu_t)^2}\right\}=c_mG^2\min\left\{\eta_t^2\mathcal{T}^2,\frac{1}{\mu_t^2}\right\}. \tag{232}$$

The equivalent form in the last line makes the two local-horizon regimes explicit. Before the proximal memory saturates, the participation-induced endpoint fluctuation is bounded by $O\big(c_m\eta_t^2G^2\mathcal{T}^2\big)$ whereas in the saturated-memory branch the available envelope is $O\big(c_m\frac{G^2}{\mu_t^2}\big)$. These are bounds on the endpoint sampling fluctuation itself and should not be interpreted as standalone convergence rates. When $\mu_t=0$, we have $A_{t,\mathcal{T}}=\mathcal{T}$, and hence

$$\mathbb{E}\Big[\big\|\boldsymbol{\Delta}_{\mathrm{pp}}^{(t)}-\overline{\boldsymbol{\Delta}}^{(t)}\big\|_2^2\,\Big|\,\boldsymbol{\theta}^{(t)}\Big]\le c_m\eta_t^2G^2\mathcal{T}^2. \tag{233}$$

Remark C.33. *The quantity $c_m = \frac{K-m}{m(K-1)}$ is the exact finite-population sampling factor for a uniformly sampled size-m subset without replacement. Conditional on the complete collection of potential client endpoints, the exact sampling variance is*

$$\mathbb{E}\left[\left\|\boldsymbol{\Delta}_{\mathrm{pp}}^{(t)} - \overline{\boldsymbol{\Delta}}^{(t)}\right\|_2^2 \,\middle|\, \mathcal{H}_t\right] = c_m \frac{1}{K}\sum_{k=1}^{K}\left\|\mathbf{d}_k^{(t,\mathcal{T})} - \overline{\boldsymbol{\Delta}}^{(t)}\right\|_2^2. \tag{234}$$

Thus c_m isolates the effect of participation level from the dispersion of the potential client endpoints. In particular, $c_K = 0, c_1 = 1$. Hence the additional client-sampling fluctuation vanishes exactly under full participation. Moreover, when $m \ll K$, $c_m = \frac{K-m}{m(K-1)} \approx \frac{1}{m}$ which recovers the familiar inverse-participation scaling.

The exact finite-population identity above relies on equal client weights, $p_k = 1/K$, together with uniform sampling without replacement. It therefore applies directly to the partial-participation server increment $\boldsymbol{\Delta}_{\mathrm{pp}}^{(t)} = \frac{1}{m}\sum_{k\in\mathcal{S}^{(t)}}\mathbf{d}_k^{(t,\mathcal{T})}$. For nonuniform client weights, importance sampling, or aggregation using selected-client-normalized weights, the conditional sampling mean and variance generally change and require a separate analysis. We retain the equal-weight specialization here to isolate the participation-induced fluctuation in exact closed form.

The preceding results quantify partial participation in update space. The next step is to determine how this zero-mean sampling fluctuation affects the global objective. Using smoothness around the potential full-population endpoint and the conditional mean-zero property

$$\mathbb{E}\left[\boldsymbol{\Delta}_{\mathrm{pp}}^{(t)} - \overline{\boldsymbol{\Delta}}^{(t)} \,\middle|\, \mathcal{H}_t\right] = \mathbf{0}, \tag{235}$$

the corresponding one-round objective penalty is controlled by $\frac{\beta}{2}c_m D_{t,\mathcal{T}}$.

C.5.3 One-Round Descent under Partial Participation

We next quantify how the client-sampling fluctuation induced by partial participation affects the global objective. Under the potential-trajectory coupling, define the reference full-population endpoint as $\widetilde{\boldsymbol{\theta}}^{(t+1)} := \boldsymbol{\theta}^{(t)} + \overline{\boldsymbol{\Delta}}^{(t)}$ and the realized partial-participation endpoint as $\boldsymbol{\theta}^{(t+1)} := \boldsymbol{\theta}^{(t)} + \boldsymbol{\Delta}_{\mathrm{pp}}^{(t)}$. The discrepancy between these two endpoints is induced solely by client subsampling. Define the corresponding sampling error as $\boldsymbol{\varepsilon}_{\mathrm{pp}}^{(t)} := \boldsymbol{\Delta}_{\mathrm{pp}}^{(t)} - \overline{\boldsymbol{\Delta}}^{(t)}$. It follows directly that

$$\boldsymbol{\theta}^{(t+1)} = \widetilde{\boldsymbol{\theta}}^{(t+1)} + \boldsymbol{\varepsilon}_{\mathrm{pp}}^{(t)}. \tag{236}$$

Thus, the realized partial-participation iterate can be viewed as the coupled full-population reference iterate perturbed by the finite-population client-sampling error.

Lemma C.34 (One-round descent under partial participation)**.** *Suppose the conditions of Lemma C.26 hold, and suppose that m clients are sampled uniformly without replacement under the equal-weight specialization $p_k = 1/K$. Then the expected global objective after one communication round satisfies*

$$\begin{aligned}\mathbb{E}\left[F\left(\boldsymbol{\theta}^{(t+1)}\right) \,\middle|\, \boldsymbol{\theta}^{(t)}\right] \le F\left(\boldsymbol{\theta}^{(t)}\right) - \frac{\eta_t\mathcal{T}}{2}\left\|\nabla F\left(\boldsymbol{\theta}^{(t)}\right)\right\|_2^2 \\ + \frac{5}{4}\eta_t(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1}D_{t,\tau} + \frac{3}{2}\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2 + \frac{\beta}{2}c_m D_{t,\mathcal{T}}.\end{aligned} \tag{237}$$

Proof of Lemma C.34. Recall the sigma-field

$$\mathcal{H}_t := \sigma\left(\boldsymbol{\theta}^{(t)}, \left\{\mathbf{d}_k^{(t,\mathcal{T})}\right\}_{k=1}^{K}\right). \tag{238}$$

Conditional on $\mathcal{H}_t$, the potential client endpoints and hence the full-population reference endpoint $\widetilde{\boldsymbol{\theta}}^{(t+1)}$ are fixed. The remaining randomness in $\boldsymbol{\theta}^{(t+1)}$ is thus due only to the random subset selection.

By the β-smoothness of F and (236), we have

$$F(\boldsymbol{\theta}^{(t+1)}) = F\left(\widetilde{\boldsymbol{\theta}}^{(t+1)} + \boldsymbol{\varepsilon}_{\mathrm{pp}}^{(t)}\right) \le F(\widetilde{\boldsymbol{\theta}}^{(t+1)}) + \left\langle\nabla F(\widetilde{\boldsymbol{\theta}}^{(t+1)}), \boldsymbol{\varepsilon}_{\mathrm{pp}}^{(t)}\right\rangle + \frac{\beta}{2}\left\|\boldsymbol{\varepsilon}_{\mathrm{pp}}^{(t)}\right\|_2^2. \tag{239}$$

Conditional on $\mathcal{H}_t$, $\nabla F(\widetilde{\boldsymbol{\theta}}^{(t+1)})$ is fixed. Moreover, by (224), $\mathbb{E}\big[\boldsymbol{\varepsilon}_{\mathrm{pp}}^{(t)}\big|\mathcal{H}_t\big]=\mathbf{0}$. Consequently,

$$\mathbb{E}\big[\big\langle\nabla F(\widetilde{\boldsymbol{\theta}}^{(t+1)}),\boldsymbol{\varepsilon}_{\mathrm{pp}}^{(t)}\big\rangle\,\big|\,\mathcal{H}_t\big]=\big\langle\nabla F(\widetilde{\boldsymbol{\theta}}^{(t+1)}),\mathbb{E}\big[\boldsymbol{\varepsilon}_{\mathrm{pp}}^{(t)}\,\big|\,\mathcal{H}_t\big]\big\rangle=0. \tag{240}$$

Taking conditional expectation of (239) given $\mathcal{H}_t$ therefore yields

$$\mathbb{E}\big[F(\boldsymbol{\theta}^{(t+1)})\,\big|\,\mathcal{H}_t\big]\le F(\widetilde{\boldsymbol{\theta}}^{(t+1)})+\frac{\beta}{2}\mathbb{E}\big[\big\|\boldsymbol{\varepsilon}_{\mathrm{pp}}^{(t)}\big\|_2^2\,\big|\,\mathcal{H}_t\big]. \tag{241}$$

Applying the exact finite-population identity yields

$$\mathbb{E}\big[F(\boldsymbol{\theta}^{(t+1)})\,\big|\,\mathcal{H}_t\big]\le F(\widetilde{\boldsymbol{\theta}}^{(t+1)})+\frac{\beta}{2}c_m\frac{1}{K}\sum_{k=1}^{K}\big\|\mathbf{d}_k^{(t,\mathcal{T})}-\overline{\boldsymbol{\Delta}}^{(t)}\big\|_2^2. \tag{242}$$

Under equal-weight aggregation, $\overline{\boldsymbol{\Delta}}^{(t)}=\frac{1}{K}\sum_{k=1}^{K}\mathbf{d}_k^{(t,\mathcal{T})}$. Hence, by the standard variance decomposition,

$$\frac{1}{K}\sum_{k=1}^{K}\big\|\mathbf{d}_k^{(t,\mathcal{T})}-\overline{\boldsymbol{\Delta}}^{(t)}\big\|_2^2=\frac{1}{K}\sum_{k=1}^{K}\big\|\mathbf{d}_k^{(t,\mathcal{T})}\big\|_2^2-\big\|\overline{\boldsymbol{\Delta}}^{(t)}\big\|_2^2\le\frac{1}{K}\sum_{k=1}^{K}\big\|\mathbf{d}_k^{(t,\mathcal{T})}\big\|_2^2. \tag{243}$$

Taking the conditional expectation with respect to $\boldsymbol{\theta}^{(t)}$ and applying the tower property yields

$$\mathbb{E}\big[F(\boldsymbol{\theta}^{(t+1)})\,\big|\,\boldsymbol{\theta}^{(t)}\big]\le\mathbb{E}\big[F(\widetilde{\boldsymbol{\theta}}^{(t+1)})\,\big|\,\boldsymbol{\theta}^{(t)}\big]+\frac{\beta}{2}c_mD_{t,\mathcal{T}}. \tag{244}$$

The potential full-population endpoint $\widetilde{\boldsymbol{\theta}}^{(t+1)}$ is generated from the complete collection of potential client trajectories using the same equal-weight aggregation as in the full-participation analysis. Hence Lemma C.26 applies:

$$\mathbb{E}\big[F(\widetilde{\boldsymbol{\theta}}^{(t+1)})\,\big|\,\boldsymbol{\theta}^{(t)}\big]\le F(\boldsymbol{\theta}^{(t)})-\frac{\eta_t\mathcal{T}}{2}\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2+\frac{5}{4}\eta_t(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1}D_{t,\tau}+\frac{3}{2}\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2. \tag{245}$$

Combining this inequality with (244) proves (237). □

C.5.4 Nonconvex Stationarity under Partial Participation

Summing the preceding one-round inequality over communication rounds yields the following global stationarity guarantee.

Theorem C.35 (Nonconvex stationarity under partial participation)**.** *Suppose the conditions of Lemma C.34 hold for every round. Define* $\overline{D}_{t,\tau}:=\mathbb{E}[D_{t,\tau}]=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\big\|\mathbf{d}_k^{(t,\tau)}\big\|_2^2\right]$*. Then,*

$$\mathcal{G}_T\le\frac{2\big[F(\boldsymbol{\theta}^{(0)})-F^\star\big]}{\sum_{t=0}^{T-1}\eta_t\mathcal{T}}+\frac{\sum_{t=0}^{T-1}\big[\frac{5}{2}\eta_t(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1}\overline{D}_{t,\tau}+3\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2+\beta c_m\overline{D}_{t,\mathcal{T}}\big]}{\sum_{t=0}^{T-1}\eta_t\mathcal{T}}. \tag{246}$$

Proof of Theorem C.35. Taking total expectation in Lemma C.34 and using definition of $\overline{D}_{t,\tau}$ yields

$$\begin{aligned}\mathbb{E}\big[F(\boldsymbol{\theta}^{(t+1)})\big]\le\mathbb{E}\big[F(\boldsymbol{\theta}^{(t)})\big]&-\frac{\eta_t\mathcal{T}}{2}\mathbb{E}\big[\big\|\nabla F(\boldsymbol{\theta}^{(t)})\big\|_2^2\big]\\&+\frac{5}{4}\eta_t(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1}\overline{D}_{t,\tau}+\frac{3}{2}\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2+\frac{\beta}{2}c_m\overline{D}_{t,\mathcal{T}}.\end{aligned} \tag{247}$$

Rearranging and summing over $t=0,\dots,T-1$ yields

$$\begin{aligned}\frac{1}{2}\sum_{t=0}^{T-1}\eta_t\mathcal{T}\mathbb{E}\big[\big\|\nabla F\big(\boldsymbol{\theta}^{(t)}\big)\big\|_2^2\big]&\le F\big(\boldsymbol{\theta}^{(0)}\big)-\mathbb{E}\big[F\big(\boldsymbol{\theta}^{(T)}\big)\big]\\&+\sum_{t=0}^{T-1}\Big[\frac{5}{4}\eta_t(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1}\overline{D}_{t,\tau}+\frac{3}{2}\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2+\frac{\beta}{2}c_m\overline{D}_{t,\mathcal{T}}\Big].\end{aligned} \tag{248}$$

Since $F(\boldsymbol{\theta}^{(T)}) \geq F^\star$, we have $F\big(\boldsymbol{\theta}^{(0)}\big) - \mathbb{E}\big[F\big(\boldsymbol{\theta}^{(T)}\big)\big] \leq F\big(\boldsymbol{\theta}^{(0)}\big) - F^\star$. Therefore,

$$\frac{1}{2}\sum_{t=0}^{T-1}\eta_t\mathcal{T}\mathbb{E}\left[\left\|\nabla F\big(\boldsymbol{\theta}^{(t)}\big)\right\|_2^2\right] \leq F\big(\boldsymbol{\theta}^{(0)}\big) - F^\star + \sum_{t=0}^{T-1}\left[\frac{5}{4}\eta_t(\beta+\mu_t)^2\sum_{\tau=0}^{\mathcal{T}-1}\overline{D}_{t,\tau} + \frac{3}{2}\beta\eta_t^2\mathcal{T}\overline{\sigma}_t^2 + \frac{\beta}{2}c_m\overline{D}_{t,\mathcal{T}}\right]. \tag{249}$$

Multiplying by 2 and dividing by $\sum_{t=0}^{T-1}\eta_t\mathcal{T}$ gives (246). □

The additional penalty due to partial client participation at the one-round objective level is $\frac{\beta}{2}c_m D_{t,\mathcal{T}}$. After summing the one-round descent inequalities over all communication rounds and multiplying the resulting inequality by two, this contribution becomes $\beta c_m\overline{D}_{t,\mathcal{T}}$ in the numerator of the global stationarity bound. Hence, the partial-participation residual in Theorem C.35 is $\frac{\sum_{t=0}^{T-1}\beta c_m\overline{D}_{t,\mathcal{T}}}{\sum_{t=0}^{T-1}\eta_t\mathcal{T}}$. Using the endpoint drift envelope, $\overline{D}_{t,\mathcal{T}} \leq \eta_t^2 G^2 A_{t,\mathcal{T}}^2$ this residual is further bounded by $\frac{\sum_{t=0}^{T-1}\beta c_m\eta_t^2 G^2 A_{t,\mathcal{T}}^2}{\sum_{t=0}^{T-1}\eta_t\mathcal{T}}$.

C.6 Bias-Aware and Time-Varying Extensions

C.6.1 Gradient bias under noisy hardware

The preceding analysis assumes that the stochastic gradient oracle is conditionally unbiased for the modeled client objective. Finite-shot parameter-shift estimation can satisfy this condition for the corresponding modeled circuit-gradient quantity when the implemented gradient estimator is unbiased. This assumption should be distinguished from systematic implementation errors: realistic hardware imperfections may cause the implemented gradient to differ persistently from the gradient of a chosen ideal reference objective.

Scope of the current experiments. The experiments in this work do not explicitly simulate hardware noise. Our empirical study focuses on finite-shot measurement stochasticity under the modeled quantum circuits and therefore evaluates the variance component of the finite-shot analysis rather than hardware-induced systematic bias. The bias-aware development in this subsection is included as a theoretical extension that characterizes how persistent implementation errors would enter the convergence analysis. Experimental evaluation under explicit gate noise, readout noise, decoherence, calibration errors, or physical quantum hardware is left for future work.

To separate systematic bias from zero-mean stochastic variation, suppose that $\mathbb{E}\big[\mathbf{g}_k^{(t,\tau)}\big|\mathcal{F}_{t,\tau}\big] = \nabla F_k\big(\boldsymbol{\theta}_k^{(t,\tau)}\big) + \mathbf{b}_k^{(t,\tau)}$ where $\mathbf{b}_k^{(t,\tau)}$ denotes the conditional systematic gradient bias relative to the reference objective.

Unlike pure finite-shot variance, systematic implementation bias is not, in general, removed by increasing the measurement-shot budget. Increasing S suppresses sampling variance, whereas persistent implementation bias requires separate calibration, error mitigation, or an analysis relative to the implemented noisy objective. Consequently, the finite-shot experiments in this work should not be interpreted as an empirical validation that proposition instead characterizes the additional theoretical effect that would arise in the presence of a bounded systematic oracle bias.

C.6.2 Time-Varying Learning Rate and Proximal Weight

The preceding convergence results permit round-dependent sequences $\{\eta_t\}_{t=0}^{T-1}$ and $\{\mu_t\}_{t=0}^{T-1}$. Their role should be interpreted through the quantities that appear explicitly in the drift and stationarity bounds, rather than through a standalone term such as $O(\sum_t \mu_t^2)$. In particular, the proximal coefficient μ_t enters the analysis through both

$$\underbrace{(\beta+\mu_t)^2}_{\text{trajectory-sensitivity factor}} \quad \text{and} \quad \underbrace{q_t = 1-\eta_t\mu_t}_{\text{local-memory retention factor}}. \tag{250}$$

For fixed η_t, increasing μ_t decreases q_t and therefore strengthens the geometric suppression of accumulated local drift. At the same time, it increases the smoothness constant $\beta+\mu_t$ of the local

proximal objective and the corresponding trajectory-sensitivity factor $(\beta+\mu_t)^2$. The analysis hence exhibits a drift–adaptation trade-off rather than predicting monotonic improvement as μ_t increases. This interaction is also reflected in the geometric-memory quantities $A_{t,\tau} = \sum_{r=0}^{\tau-1} q_t^r, B_{t,\tau} = \sum_{r=0}^{\tau-1} q_t^{2r}$ which control, respectively, worst-case accumulation along the local trajectory and the accumulation of martingale-difference stochastic noise.

A round-indexed warmup–cosine learning-rate schedule can be written as

$$\eta_t = \begin{cases} \eta_{\text{start}} + (\eta_{\max} - \eta_{\text{start}}) \frac{t+1}{T_w}, & 0 \le t < T_w, \\ \eta_{\min} + \frac{\eta_{\max}-\eta_{\min}}{2} \left[1 + \cos\left(\pi \frac{t-T_w}{T-1-T_w}\right)\right], & T_w \le t \le T-1, \end{cases} \tag{251}$$

for $0 \le T_w < T-1$. A linear proximal-annealing schedule is $\mu_t = \mu_0 \max\left\{0, 1 - \frac{t}{T_\mu}\right\}$

$$\mu_t = \mu_0 \max\left\{0, 1 - \frac{t}{T_\mu}\right\}, \tag{252}$$

where $T_\mu > 0$ denotes the proximal-annealing horizon. The schedules must satisfy the conditions required by the corresponding local and global descent arguments. In particular, $\eta_t \le \frac{1}{\beta+\mu_t}, \beta\eta_t\mathcal{T} \le \frac{1}{6}$. The first condition also implies $0 \le \eta_t\mu_t \le 1$ whenever $\beta, \mu_t \ge 0$, and is therefore consistent with the geometric-memory bounds used above. When the annealed schedule reaches $\mu_t = 0$, we have $q_t = 1, A_{t,\tau} = \tau, B_{t,\tau} = \tau$ so the definitions reduce continuously to their Q-FedAvg counterparts and the convergence analysis remains well-defined. The theory does not, by itself, imply that either warmup–cosine learning rates or proximal annealing accelerate convergence relative to every constant schedule. Rather, it identifies the quantities through which such schedules modify the optimization, stochastic, and client-drift terms.

C.6.3 Barren plateaus and shot complexity

Assumption 3.5 is an upper second-moment condition. It controls the magnitude of the stochastic gradient estimator, but it does not impose a positive lower bound on $\|\nabla F(\boldsymbol{\theta})\|_2$ It therefore does not exclude a barren-plateau regime in which the gradient signal becomes exponentially small as the number of qubits or circuit size increases.

For the modeled finite-shot parameter-shift estimator analyzed in Appendix B, the gradient-estimation mean-squared error obeys a dimension-aware upper bound of the form

$$\mathbb{E}\left[\left\|\widehat{\nabla f} - \nabla f\right\|_2^2\right] = O\left(\frac{d_q}{S}\right). \tag{253}$$

When the full modeled gradient estimator inherits this scaling, the corresponding relative-error upper bound behaves schematically as

$$\frac{\mathbb{E}\left[\left\|\widehat{\nabla F} - \nabla F\right\|_2^2\right]}{\|\nabla F\|_2^2} = O\left(\frac{d_q}{S\|\nabla F\|_2^2}\right). \tag{254}$$

Suppose, for example, that the squared gradient norm exhibits the barren-plateau scaling

$$\|\nabla F\|_2^2 = \Theta\left(e^{-\alpha n}\right), \quad \alpha > 0, \tag{255}$$

where n denotes the relevant circuit-size parameter, such as the number of qubits. Maintaining the upper bound in (254) at constant order then requires $S = \Omega(d_q e^{\alpha n})$. Thus, bounded absolute stochastic moments do not imply favorable relative gradient accuracy when the underlying gradient signal becomes exponentially small. In particular, increasing the number of shots can suppress measurement variance but does not remove the trainability mechanism responsible for a barren plateau.

Accordingly, the convergence theorem established in this work should be interpreted as an absolute first-order stationarity guarantee under the stated smoothness and stochastic-moment assumptions. It does not establish trainability in a barren-plateau regime or rule out an exponential shot requirement when the underlying gradient signal itself vanishes exponentially.

D ADDITIONAL EXPERIMENTAL RESULTS

This appendix provides additional implementation details, estimator-level diagnostics, optimization diagnostics, and qualitative visualizations supporting the experiments in Section 4. We distinguish throughout between three empirical roles: *(i) direct validation of finite-shot parameter-shift statistics*, *(ii) mechanism-level validation using the proximal-SGD recursion analyzed in the theory*, and *(iii) practical IID and non-IID federated classification experiments using adaptive optimizers*. This distinction prevents downstream classification behavior from being interpreted as a direct test of the stochastic-SGD convergence.

D.1 EXPERIMENTAL CONFIGURATIONS

All quantum experiments are implemented in PennyLane with PyTorch-based classical processing. Classical features are reduced to match the circuit width and rescaled to $[-\pi, \pi]$ before angle encoding. The variational circuits contain parameterized single-qubit rotations, fixed two-qubit entangling layers, and Pauli-Z expectation-value measurements.

Since the experiments span several distinct purposes, no single hyperparameter configuration describes all runs. Table 3 summarizes the two controlled diagnostics for which the experimental configuration is fixed exactly.

Table 3: Exact configurations of the two controlled experiments used to connect the empirical results to the theoretical analysis.

Experiment	**Qubits**	**Layers**	**Clients**	**Batch**	**Shots**	**Rounds / reps.**	**Optimizer**
Finite-shot estimator	8	2	1 fixed	4	32–4096	100 repetitions/S	Fixed checkpoint
Theory-aligned proximal	4	1	5	4	128	8 rounds, 8 seeds	SGD / Prox-SGD

The larger practical classification experiments use circuit widths between 4 and 12 qubits and two to three variational layers, with experiment-dependent mini-batch sizes, communication horizons, and learning-rate schedules. These runs use Adam or AdamW and should therefore be regarded as practical performance evaluations rather than direct instances of Theorem 3.11.

We explicitly distinguish circuit-evaluation modes throughout the paper

$$\text{analytic evaluation:} \quad \texttt{shots=None}, \tag{256}$$

$$\text{finite-shot evaluation:} \quad \texttt{shots} = S < \infty. \tag{257}$$

The finite-shot experiments use stochastic measurement sampling. No synthetic Gaussian perturbation is added to analytic expectation values. The experiments isolate statistical uncertainty due to finite quantum measurements. Unless explicitly stated otherwise, they do not simulate gate errors, decoherence, coherent calibration errors, or systematic readout bias. Consequently, these experiments should not be interpreted as physical-hardware noise benchmarks.

For federated optimization, all clients must operate in the same feature coordinate system. Accordingly, the preprocessing transformation used by the reported federated model is treated as a common map $\Phi(\cdot)$ and is held fixed across clients. In particular, a PCA transformation used for federated training should be fitted using only the designated training data (or an appropriate shared/public training subset), then frozen and applied consistently to every client and to validation/test data. Validation and test examples are not used to fit the transformation.

This common preprocessing requirement is important because independently fitted client-specific PCA bases would generally satisfy $\mathbf{P}_k \neq \mathbf{P}_{k'}$ so the same VQC parameter coordinate could act on different feature directions at different clients. The retained feature dimension is matched to the circuit width. Depending on the experiment, 4–12 PCA components are used and subsequently rescaled to the interval $[-\pi, \pi]$ for angle encoding.

D.2 Finite-Shot Parameter-Shift Diagnostics

D.2.1 Controlled diagnostic protocol

The direct finite-shot experiment in Section 4.1 uses a fixed trained 8-qubit, 2-layer VQC with 48 trainable quantum parameters. Throughout the experiment, we hold fixed a mini-batch of $B = 4$ samples, a Pauli-Z readout on wire 0, and the same 12 circuit coordinates across all shot budgets. Thus, the shot budget is the only measurement-related quantity varied in the diagnostic.

For reproducibility, the flattened circuit-coordinate indices used in the experiment are

$$\mathcal{J} = \{0, 4, 8, 12, 17, 21, 25, 29, 34, 38, 42, 47\}. \tag{258}$$

The set $\mathcal{J}$ constitutes a fixed diagnostic subset of the 48 trainable parameters and is used only for the finite-shot analysis; it is not part of the QFL training algorithm.

For each $j \in \mathcal{J}$, the finite-shot parameter-shift estimator is defined as

$$\widehat{\partial_{\theta_j} f}^{(S)} = \frac{1}{2}\left[\widehat{f}_S\left(\boldsymbol{\theta} + \frac{\pi}{2}\mathbf{e}_j\right) - \widehat{f}_S\left(\boldsymbol{\theta} - \frac{\pi}{2}\mathbf{e}_j\right)\right], \tag{259}$$

where $\mathbf{e}_j$ denotes the coordinate vector associated with parameter j, and the positive and negative shifted expectation values are estimated using independent finite-shot measurements.

For each shot budget $S \in \{32, 64, 128, 256, 512, 1024, 2048, 4096\}$, we generate $N_{\text{rep}} = 100$ independent realizations of the 12-dimensional parameter-shift gradient vector, $\widehat{\mathbf{g}}_S^{(r)} \in \mathbb{R}^{12}$, $r = 1, \ldots, N_{\text{rep}}$. The empirical mean gradient is $\overline{\mathbf{g}}_S := \frac{1}{N_{\text{rep}}} \sum_{r=1}^{N_{\text{rep}}} \widehat{\mathbf{g}}_S^{(r)}$ and the corresponding centered sample variance is defined as

$$\widehat{V}_{\text{grad}}(S) := \frac{1}{N_{\text{rep}} - 1} \sum_{r=1}^{N_{\text{rep}}} \left\|\widehat{\mathbf{g}}_S^{(r)} - \overline{\mathbf{g}}_S\right\|_2^2. \tag{260}$$

To assess the finite-shot estimation error relative to the exact circuit-gradient reference, let $\mathbf{g}_{\text{an}} \in \mathbb{R}^{12}$ denote the corresponding analytic parameter-shift gradient vector. We additionally compute the empirical gradient mean-squared error (MSE) and squared bias as

$$\text{MSE}_{\text{grad}}(S) := \frac{1}{N_{\text{rep}}} \sum_{r=1}^{N_{\text{rep}}} \left\|\widehat{\mathbf{g}}_S^{(r)} - \mathbf{g}_{\text{an}}\right\|_2^2, \tag{261}$$

$$\text{Bias}^2(S) := \left\|\overline{\mathbf{g}}_S - \mathbf{g}_{\text{an}}\right\|_2^2. \tag{262}$$

These quantities separately characterize the dispersion and systematic deviation of the finite-shot gradient estimates from the analytic reference.

Table 4: Complete finite-shot parameter-shift diagnostic over $N_{\text{rep}} = 100$ independent repetitions. All circuit, data, and parameter-coordinate quantities are fixed while only S is varied.

Shots S	$\widehat{V}_{\text{grad}}(S)$	$\text{MSE}_{\text{grad}}(S)$	$\text{Bias}^2(S)$	$S\widehat{V}_{\text{grad}}(S)$
32	4.501287×10^{-2}	4.511990×10^{-2}	5.57×10^{-4}	1.440412
64	2.423587×10^{-2}	2.412744×10^{-2}	1.34×10^{-4}	1.551096
128	1.192685×10^{-2}	1.194635×10^{-2}	1.39×10^{-4}	1.526636
256	5.963561×10^{-3}	5.964504×10^{-3}	6.10×10^{-5}	1.526672
512	3.061654×10^{-3}	3.054189×10^{-3}	2.30×10^{-5}	1.567567
1024	1.556462×10^{-3}	1.572401×10^{-3}	3.20×10^{-5}	1.593817
2048	7.939017×10^{-4}	7.923841×10^{-4}	6.00×10^{-6}	1.625911
4096	3.871330×10^{-4}	3.869867×10^{-4}	4.00×10^{-6}	1.585697

Table 4 exhibits the expected reduction in finite-shot gradient variance as the measurement budget increases. As S increases from 32 to 4096, corresponding to a 128-fold increase in the number of shots, the empirical gradient variance decreases from 4.501287×10^{-2} to 3.871330×10^{-4}. The corresponding variance-reduction factor is

$$\frac{4.501287 \times 10^{-2}}{3.871330 \times 10^{-4}} \approx 116.27, \tag{263}$$

which corresponds to an approximately 99.14% reduction in the measured gradient variance.

To quantify the scaling behavior, we fit the power-law model $\widehat{V}_{\mathrm{grad}}(S) = CS^{\rho}$ in log–log coordinates. The fitted exponent and goodness of fit are $\widehat{\rho} = -0.9817$ with $95\%\mathrm{CI} = [-0.9995, -0.9624]$ and $R^2 = 0.9999$. The estimated exponent differs from the theoretical inverse-shot exponent -1 by 0.0183. Although the fitted scaling is very close to S^{-1}, the reported confidence interval does not contain -1. We therefore characterize the empirical behavior as *near-inverse-shot scaling* rather than asserting an exponent statistically indistinguishable from -1.

More importantly, the rescaled quantity $S\widehat{V}_{\mathrm{grad}}(S)$ remains approximately constant across the entire range of shot budgets, varying only from approximately 1.44 to 1.63. This behavior is consistent with the inverse-shot variance scaling $\widehat{V}_{\mathrm{grad}}(S) \propto \frac{1}{S}$ predicted by the finite-shot parameter-shift analysis. Thus, increasing the shot budget primarily suppresses the stochastic dispersion of the gradient estimator at the expected $O(S^{-1})$ rate.

Across the tested shot budgets, we obtain $\mathrm{mean}\left[S\widehat{V}_{\mathrm{grad}}(S)\right] = 1.5522$, $\mathrm{std}\left[S\widehat{V}_{\mathrm{grad}}(S)\right] = 0.0565$, and $\mathrm{CV}\left[S\widehat{V}_{\mathrm{grad}}(S)\right] = 3.64\%$. Thus, the rescaled quantity $S\widehat{V}_{\mathrm{grad}}(S)$ remains approximately stable despite a 128-fold increase in the shot budget, providing additional empirical evidence for near-S^{-1} variance scaling. This behavior is also apparent from the variance-reduction factors obtained under successive doublings of the shot budget that are $1.857, 2.032, 2.000, 1.948, 1.967, 1.961, 2.051$ whose mean is 1.974. Hence, doubling the number of shots reduces the measured gradient variance by approximately a factor of two over the tested range.

The gradient MSE remains close to the centered variance across all shot budgets, indicating that the observed finite-shot estimation error is dominated by stochastic dispersion rather than systematic bias. In particular, the empirical squared-bias contribution remains small relative to the measured variance throughout the experiment. Its component is only approximately 0.55%–2.06% of the measured variance across the tested budgets.

For several shot budgets, $\mathrm{MSE}_{\mathrm{grad}}(S)$ is slightly smaller than $\widehat{V}_{\mathrm{grad}}(S)$. This behavior is consistent with their different finite-sample normalizations. Specifically, e centered variance $\widehat{V}_{\mathrm{grad}}(S)$ uses the unbiased $(N_{\mathrm{rep}}-1)^{-1}$ normalization, whereas $\mathrm{MSE}_{\mathrm{grad}}(S)$ uses N_{rep}^{-1}. The exact empirical decomposition, in finite samples, is

$$\mathrm{MSE}_{\mathrm{grad}}(S) = \frac{N_{\mathrm{rep}} - 1}{N_{\mathrm{rep}}} \widehat{V}_{\mathrm{grad}}(S) + \mathrm{Bias}^2(S). \tag{264}$$

Therefore, $\mathrm{MSE}_{\mathrm{grad}}(S) < \widehat{V}_{\mathrm{grad}}(S)$ can occur when the empirical squared bias is sufficiently small. Overall, the controlled diagnostic provides empirical support for the estimator-level inverse-shot variance scaling established in Lemma 3.1. The result pertains specifically to the finite-shot parameter-shift gradient estimator and should not be interpreted as implying that end-to-end training error, convergence rate, or classification performance necessarily scales as S^{-1}.

D.2.2 Measurement cost

Each gradient realization in the finite-shot diagnostic evaluates 12 parameter coordinates over a fixed mini-batch of $B = 4$ samples. Since the parameter-shift rule requires two shifted circuit evaluations for each coordinate and sample, each gradient realization requires

$$N_{\mathrm{eval}} = 12 \times 4 \times 2 = 96 \tag{265}$$

shifted-circuit evaluations. Across 8 shot budgets and $N_{\mathrm{rep}} = 100$ independent repetitions, the total number of finite-shot shifted-circuit executions is therefore

$$N_{\mathrm{circ}} = 8 \times 100 \times 96 = 76{,}800. \tag{266}$$

The physical measurement cost additionally depends on the number of shots used for each circuit evaluation. For the shot budgets S the total number of individual measurement shots is

$$N_{\mathrm{shots}} = N_{\mathrm{rep}}\, 96 \sum_{S \in \{32,\ldots,4096\}} S = 78{,}336{,}000. \tag{267}$$

This measurement accounting highlights the resource–variance trade-off associated with the empirical scaling $\widehat{V}_{\mathrm{grad}}(S) \propto S^{-1}$: increasing the shot budget reduces finite-shot sampling variance approximately inversely with S, while the corresponding measurement cost grows linearly with S. More generally, consider a parameter-shift gradient with d_q trainable quantum parameters, mini-batch size B, $\mathcal{T}$ local optimization steps per communication round, and m participating clients. Assuming two shifted circuit evaluations per trainable parameter, the leading shifted-circuit evaluation count per communication round scales as $N_{\mathrm{circ}}^{(\mathrm{round})} = O(m\mathcal{T}Bd_q)$ while the corresponding measurement-shot complexity scales as $N_{\mathrm{shots}}^{(\mathrm{round})} = O(m\mathcal{T}Bd_qS)$. These expressions represent the leading parameter-shift measurement cost before accounting for multiple readout observables, commuting-observable grouping, or other measurement-reduction strategies.

D.3 Sensitivity to the Shot Budget

The preceding diagnostic directly evaluates the finite-shot gradient estimator appearing in the theoretical analysis. To examine the downstream effect of finite-shot measurement noise, we additionally hold the trained classifier and evaluation data fixed and repeat inference under several shot budgets using independent measurement realizations.

Table 5: Downstream finite-shot evaluation under repeated independent measurement realizations. Results are reported as mean $\pm$ std across repetitions. Accuracy, macro-F1, AUC, and analytic agreement are nonlinear performance metrics and are not expected to follow an exact S^{-1} scaling.

Shots S	**Acc**	**Macro-F1**	**AUC**	**Analytic agreement**
128	0.3434 ± 0.0051	0.3487 ± 0.0058	0.6775 ± 0.0031	0.3480 ± 0.0057
512	0.5322 ± 0.0060	0.5412 ± 0.0057	0.8185 ± 0.0034	0.5468 ± 0.0078
1024	0.6389 ± 0.0055	0.6437 ± 0.0058	0.8795 ± 0.0032	0.6602 ± 0.0067
2048	0.7182 ± 0.0055	0.7212 ± 0.0056	0.9167 ± 0.0021	0.7534 ± 0.0041
4096	0.7767 ± 0.0043	0.7785 ± 0.0045	0.9433 ± 0.0011	0.8220 ± 0.0052
Analytic	0.8690	0.8704	0.9791	1.0000

As shown in Table 5, all reported metrics improve monotonically over the tested range as the shot budget increases. In particular, accuracy increases from 34.34% at $S = 128$ to 77.67% at $S = 4096$, corresponding to an absolute improvement of 43.33 percentage points. Macro-F1 follows a similar trend, increasing from 34.87% to 77.85% over the same range.

The AUC increases from 0.6775 at $S = 128$ to 0.9433 at $S = 4096$, approaching the analytic reference value of 0.9791. Relative to the hard classification metrics, AUC exhibits greater robustness to finite-shot measurement noise over the tested budgets. This behavior is consistent with the ranking-based nature of AUC: finite-shot perturbations can alter the predicted class without necessarily changing the relative ordering of the underlying class scores to the same extent.

Analytic agreement, defined as the fraction of finite-shot predictions that coincide with the corresponding analytic predictions, increases from 34.80% at $S = 128$ to 82.20% at $S = 4096$. Despite this substantial improvement, the agreement remains below unity at the largest tested shot budget, indicating that finite-shot fluctuations continue to affect a non-negligible fraction of the predicted labels. The results in Table 5 should not be interpreted as exhibiting an S^{-1} scaling law. Unlike the variance of the finite-shot parameter-shift estimator, these downstream quantities are nonlinear functions of the measured circuit outputs and may additionally depend on decision boundaries, score margins, and class-ranking structure. The two diagnostics therefore serve complementary purposes: the direct gradient experiment evaluates the estimator-level inverse-shot variance scaling predicted by the theory, while the downstream experiment demonstrates how increasing the measurement budget can improve the stability and predictive performance of a fixed trained classifier.

D.4 Theory-Aligned Proximal-SGD Diagnostics

The theory-aligned experiment implements the same proximal stochastic-gradient update analyzed in the convergence theory. To isolate the behavior predicted by the theoretical model, no momentum, Adam moments, adaptive preconditioning, weight decay, or server-side adaptive optimization is employed. The experimental configuration consists of a 4-qubit, one-layer VQC distributed across

$K = 5$ heterogeneous clients. Each client performs $\mathcal{T} = 4$ local steps per communication round using mini-batches of size $B = 4$ and finite-shot parameter-shift gradients with shot budget $S = 128$. Training is conducted for $T = 8$ communication rounds. Each proximal schedule is evaluated using the same eight random seeds, thereby enabling paired comparisons across the three methods.

The round-dependent learning-rate schedule is $\{\eta_t\}_{t=0}^{7} = \{0.01, 0.02, 0.01879, 0.0155, 0.011, 0.0065, 0.003206\}$. For the annealed proximal variant, the proximal coefficient follows $\{\mu_t\}_{t=0}^{7} = \{1, 0.8333, 0.6667, 0.5, 0.3333, 0.1667, 0, 0\}$. We compare three proximal schedules: Q-FedAvg ($\mu_t = 0$), Q-FedProx-Const ($\mu_t = 1$), Q-FedProx-Anneal ($\{\mu_t\}_{t=0}^{7}$) Thus, Q-FedAvg removes proximal regularization entirely, Q-FedProx-Const maintains a fixed proximal strength throughout training, and Q-FedProx-Anneal progressively reduces the proximal influence before setting $\mu_t = 0$ in the final two communication rounds. Under the proximal-memory definition $q_t := 1 - \eta_t\mu_t$ the schedules satisfy $\max_t\{\eta_t\mu_t\} = 0.02, \min_t\{q_t\} = 0.98$. Hence, $q_t \in [0.98, 1]$ throughout the experiment, satisfying the nonnegative proximal-memory regime assumed in the corresponding drift analysis. For the annealed schedule, q_t approaches unity as μ_t decreases, and the final two rounds with $\mu_t = 0$ recover the unregularized local-update dynamics.

D.4.1 Endpoint Client Drift

At the end of communication round t, we measure the terminal client-drift energy as $D_t^{\text{end}} := \sum_{k=1}^{K} p_k \left\| \boldsymbol{\theta}_k^{(t,\mathcal{T})} - \boldsymbol{\theta}^{(t)} \right\|_2^2$. To summarize the drift trajectory across communication rounds, we report the cumulative $D_{\text{cum}}^{\text{end}} := \sum_{t=0}^{T-1} D_t^{\text{end}}$, round-averaged $D_{\text{mean}}^{\text{end}} := \frac{1}{T}\sum_{t=0}^{T-1} D_t^{\text{end}}$, and peak endpoint drift $D_{\text{peak}}^{\text{end}} := \max_{0\le t<T} D_t^{\text{end}}$. These empirical quantities characterize the magnitude of client drift at the end of each communication round. They are closely related to, but distinct from, the accumulated within-round drift energy $\sum_{\tau=0}^{\mathcal{T}-1} D_{t,\tau}$ appearing in the convergence analysis. In particular, D_t^{end} corresponds to the terminal local step $\tau = \mathcal{T}$, whereas the theoretical quantity aggregates drift over all local steps within a round.

Table 6: Endpoint client-drift statistics for the theory-aligned proximal-SGD experiment over eight paired random seeds. Values are reported as mean ± std across seeds.

Method	**Cum. drift**	**Mean drift**	**Peak drift**
Q-FedAvg	0.1736 ± 0.0848	0.02170 ± 0.01060	0.0595 ± 0.0328
Q-FedProx-Const	0.1763 ± 0.0867	0.02204 ± 0.01084	0.0586 ± 0.0314
Q-FedProx-Anneal	0.1681 ± 0.0742	0.02101 ± 0.00928	0.0542 ± 0.0235

As shown in Table 6, Q-FedProx-Anneal attains the smallest mean value for all three endpoint-drift summaries. Its mean cumulative drift decreases from 0.1736 for Q-FedAvg to 0.1681, while the mean peak drift decreases from 0.0595 to 0.0542. In contrast, Q-FedProx-Const yields a cumulative and mean drift slightly larger than those of Q-FedAvg, although its mean peak drift is marginally smaller. These raw averages therefore do not indicate that proximal regularization uniformly reduces endpoint drift for every choice of proximal schedule. Rather, they suggest that the annealed schedule provides the most favorable endpoint-drift behavior among the three configurations considered here. Since all methods are evaluated using matched random seeds, however, comparisons based solely on unpaired ratios of the aggregate means can obscure seed-specific effects. The paired differences reported below provide a more appropriate assessment of the effect of the proximal schedules.

D.4.2 Paired Drift Analysis

Table 7: Paired changes in endpoint client drift relative to Q-FedAvg over eight matched seeds. Negative values indicate reductions. Confidence intervals are paired-bootstrap 95% intervals.

Method	**Cum. drift change**	**Seeds reduced**	**Peak drift change**	**Seeds reduced**
Constant μ	$+1.40\%$ $[-3.97, 8.59]$	$5/8$	-1.19% $[-6.67, 5.75]$	$6/8$
Annealed μ_t	-1.72% $[-5.48, 2.56]$	$6/8$	-5.25% $[-12.07, 1.32]$	$7/8$

Table 7 provides a seed-matched comparison of the endpoint client-drift statistics relative to Q-FedAvg. For the annealed schedule, cumulative endpoint drift decreases in 6 of the 8 paired seeds,

with a mean relative change of -1.72%. Peak endpoint drift decreases in 7 of the 8 seeds, with a larger mean relative change of -5.25%. However, the corresponding paired-bootstrap 95% confidence intervals, $[-5.48, 2.56]\%$ and $[-12.07, 1.32]\%$, respectively, both include zero. Thus, the observed reductions provide directional evidence but are not statistically resolved at the present number of seeds.

The stronger directional effect is observed for peak drift. The annealed schedule yields a mean peak-drift change of -5.25%, compared with -1.19% under the constant proximal schedule. This behavior is consistent with the proximal-drift mechanism identified in the theoretical analysis, under which proximal regularization limits the accumulation of local client updates. In the present experiment, applying stronger proximal regularization during the earlier communication rounds and subsequently annealing it is associated with smaller peak client excursions. The confidence interval for this reduction nevertheless crosses zero, so the experiment does not provide statistically conclusive evidence of a nonzero population-level reduction in peak drift. The constant proximal schedule exhibits a different pattern. Its mean cumulative-drift change is slightly positive at $+1.40\%$, despite reductions occurring in 5 of the 8 paired seeds, while its mean peak-drift change is only -1.19%. These results indicate that a fixed proximal coefficient does not uniformly improve the endpoint-drift statistics in this experiment. Its effect is therefore better assessed jointly with the stationarity diagnostic considered below, rather than interpreted solely through endpoint drift.

D.4.3 Stationarity and predictive diagnostics

Table 8: Paired final-stationarity changes relative to Q-FedAvg. Negative values denote reductions.

Method	Paired change	95% bootstrap CI
Constant μ	-26.28%	$[-52.96, -5.30]\%$
Annealed μ_t	-14.70%	$[-44.75, 14.23]\%$

As shown in Table 8, the constant proximal schedule yields a mean paired reduction of 26.28% in the terminal squared-gradient stationarity diagnostic relative to Q-FedAvg. Its paired-bootstrap 95% confidence interval, $[-52.96, -5.30]\%$, lies entirely below zero. In contrast, the annealed schedule yields a mean reduction of 14.70%, but its confidence interval, $[-44.75, 14.23]\%$, includes zero. Thus, among the two proximal schedules considered, only the constant schedule exhibits a paired reduction whose reported bootstrap interval excludes zero.

The magnitude of this result, however, is influenced by a difficult Q-FedAvg trajectory. For seed 6, the terminal stationarity diagnostic is approximately 117.98 for Q-FedAvg, 4.45 for Q-FedProx-Const, 5.96 for Q-FedProx-Anneal. Both proximal schedules therefore substantially reduce the terminal squared-gradient value for this seed, indicating that a considerable portion of the improvement in the mean paired statistic is associated with stabilization of this particularly difficult trajectory.

Table 9: Paired predictive-performance changes with respect to Q-FedAvg in the theory-aligned experiment. "pp" denotes percentage points.

Method	Acc change	95% CI	Macro-F1 change	95% CI
Constant μ	-0.71 pp	$[-2.63, 0.46]$ pp	-0.69 pp	$[-2.43, 0.36]$ pp
Annealed μ_t	-0.85 pp	$[-2.63, 0.31]$ pp	-0.84 pp	$[-2.60, 0.25]$ pp

A complementary robust summary is provided by the median terminal stationarity values with 1.161 for Q-FedAvg, 1.117 for Q-FedProx-Const, and 1.204 for Q-FedProx-Anneal. The substantially smaller separation among the medians than that suggested by the mean paired changes indicates that the stationarity benefit is not uniform across seeds. We therefore interpret the constant-proximal result primarily as evidence of improved robustness to difficult finite-shot optimization trajectories in this controlled experiment, rather than as a uniform 26.28% reduction in terminal stationarity error.

For both accuracy and macro-F1, the paired-bootstrap confidence intervals include zero, and the absolute mean changes remain below one percentage point, as shown in Table 9. The controlled configuration uses only four qubits, one variational layer, four local steps, and eight communication rounds. It is designed to permit repeated finite-shot parameter-shift training under an update rule

that closely matches the theoretical analysis, rather than to maximize classification performance. Taken together, the controlled diagnostics reveal different empirical effects of the two proximal schedules. The annealed schedule exhibits the more consistent directional reduction in peak endpoint drift, whereas the constant schedule provides the stronger observed stabilization of difficult terminal stationarity trajectories. These results therefore suggest a schedule-dependent trade-off between endpoint-drift control and optimization stability in the tested setting, rather than uniform dominance of either proximal strategy.

The controlled proximal experiment also clarifies why the theory does not predict monotonic improvement as μ_t increases. The local recursion contains the memory factor $q_t = 1 - \eta_t\mu_t$ so increasing μ_t strengthens contraction of earlier local-gradient contributions. However, the trajectory-sensitive error term also contains the factor $(\beta + \mu_t)^2$. Thus μ_t has competing effects:

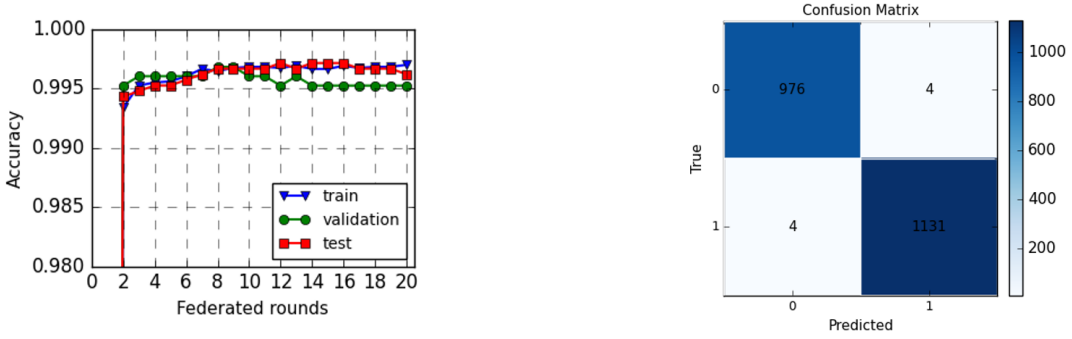
$$\text{larger } \mu_t \Longrightarrow \text{stronger local-memory damping,} \tag{268}$$

$$\text{larger } \mu_t \Longrightarrow \text{larger proximal-objective sensitivity factor.} \tag{269}$$

This analytical trade-off is reflected empirically. Annealing produces the most consistent directional reduction in peak endpoint drift, while the constant proximal coefficient provides the strongest stabilization of the difficult final-stationarity trajectory. The results therefore support the interpretation that proximal regularization modifies client dynamics rather than uniformly improving every optimization or predictive metric.

D.5 Binary MNIST IID Check

In this sub-section, we first evaluate the FL system under the assumption of IID client datasets and no secure aggregation, where the MNIST dataset is used for the binary classification task. The dataset is partitioned into three IID client subsets, ensuring that each client receives similar samples drawn from the same distribution. Local training is performed using the same VQC model architecture with 8 qubits and 3 entangling layers. Each sample of the data is a grayscale image of size 28×28, corresponding to 784 pixel intensities. These high-dimensional inputs are first compressed via principal component analysis (PCA) into an 8-dimensional feature vector, matching the number of qubits in the VQC model. This preprocessing not only reduces computational burden but also mitigates noise by maintaining the principal variance directions most relevant for classification. The transformed features are then embedded into an 8-qubit quantum state using angle encoding, followed by three layers of entangling unitaries and a linear readout.



(a) Global convergence.

(b) Test confusion matrix.

Figure 3: IID federated training performance of the proposed QFL-VQC framework for binary MNIST with three IID clients over 20 communication rounds.

Fig. 3(a) presents the learning curves of the global model's performance over $T = 20$ communication rounds. It is observed that in the baseline setting with IID client data and without secure aggregation, the QFL-VQCs model quickly converges within the first five rounds, achieving near-optimal accuracy of approximately 99.5% on all splits. In particular, within just two rounds, test accuracy increases from approximately 68% to more than 99%, with a corresponding loss decreases approximately from 0.61 to 0.08. This significant improvement results from the IID partitioning which minimizes heterogeneity among clients and ensures that local gradients are well aligned, enabling the federated

aggregation to approximate centralized training and quickly capture the global decision boundary. Furthermore, the quantum feature embedding maps the 8 PCA-reduced dimensions into an 8-qubit Hilbert space, and the variational ansatz provides an expressive representation that makes the binary classification task nearly separable within only a few communication rounds. Additionally, the use of a warmup–cosine learning rate schedule further stabilizes optimization by allowing large exploration updates in the early stages followed by gradual refinement, while the shallow circuit depth avoids *barren plateau* effects and supports efficient convergence.

Fig. 3(b) presents the confusion matrix of the global QFL-VQC model after $T = 20$ federated rounds under the IID data setting with three clients. The model achieves almost perfect classification on the binary MNIST task, correctly identifying 976 out of 980 digit-0 samples and 1131 out of 1135 digit-1 samples. Only eight misclassifications are observed in total, evenly distributed as four false positives and four false negatives. This balanced error pattern indicates that the learned decision boundary is highly symmetric, without bias toward either class. Combined with the learning curves, the confusion matrix confirms that the quantum-enhanced embedding and variational ansatz enable the model to converge to a highly expressive global solution, where class separability is nearly optimal. The extremely low error rate illustrates that under IID partitioning, federated aggregation is able to preserve local signal alignment and effectively exploit the quantum feature space for global generalization.

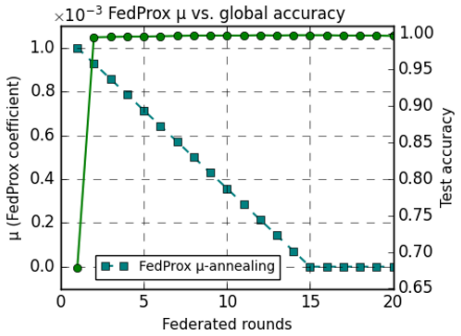


Figure 4: A visual representation of a simple quantum circuit.

Fig. 4 shows the evolution of the FedProx regularization parameter μ_t along with the global test accuracy over communication rounds in the IID three-client setting. The coefficient begins at $\mu = 10^{-3}$ and is gradually annealed toward zero, following a linear annealing schedule. This schedule design in FedProx ensures that stronger regularization is applied during the early rounds when client updates may diverge, preventing local models from drifting too far from the global trajectory. As communication rounds progress and the global model becomes more stable, μ diminishes, allowing clients to perform local updates with greater flexibility while still remaining close to the global model trajectory.

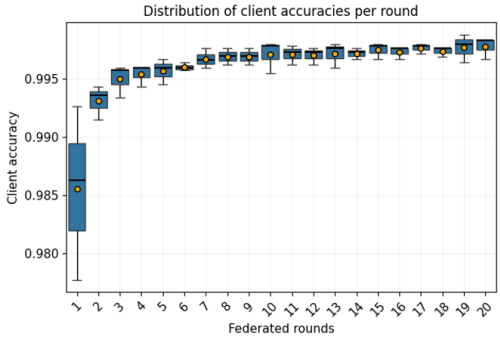


Figure 5: A visual representation of a simple quantum circuit.

In addition to the global QFL-VQC performance, it is critical to examine the distribution of accuracies achieved by individual clients across communication rounds, as displayed in Fig. 5. It is obvious

that in the first communication round, there is a relatively wide deviation in client performance, with accuracies ranging between approximately 98% and 99.2%. This variance arises primarily due to random initialization of model parameters and stochastic effects of mini-batch training, where some clients may perform poorly depending on their early gradient directions. However, as federated rounds progress, the distribution tightens rapidly concentrated around the global average after the first few aggregation rounds, with client accuracies exceeding 99.3% by round 3 and converging to above 99.7% by the last round. This outcome is expected, as the IID partitioning ensures that each client's training set is statistically similar to the others, leading to nearly identical local gradients. As a result, the aggregation process does not increase client-specific noise or bias, but instead preserves a stable global trajectory. Additionally, the expressive quantum feature map and variational ansatz allow the global model to quickly reach a separable decision boundary, minimizing divergence between local models. Furthermore, the stability of the accuracy distribution in later rounds also indicates fairness, as all clients benefit equally from the training process, with no lower-performing participants.

D.6 MEDNIST MEDICAL-IMAGE EXPERIMENTS

We further evaluate the performance of the proposed QFL-VQC framework on a medical imaging dataset under both IID and non-IID client partitioning scenarios. The dataset used in our experiments is the MedMNIST collection, which contains a total of $58,954$ radiological images across six medical imaging categories, *AbdomenCT, BreastMRI, CXR, ChestCT, Hand*, and *HeadCT*, each represented as a 64×64 grayscale image, as shown in Fig. 6. Each class exhibits mild imbalance, with counts ranging from 7000 to 10000 samples. To prevent class-frequency bias from influencing the Dirichlet partitioning in federated settings, each subset was first normalized to per-class proportions. This step ensures that observed non-IID effects arise primarily from label distribution skew, rather than pre-existing imbalance in the data.

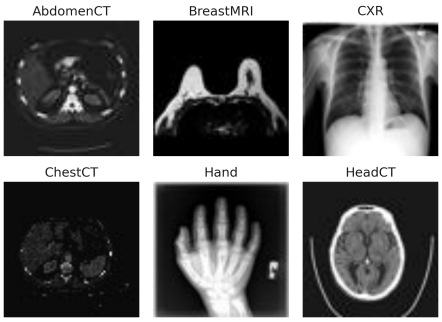
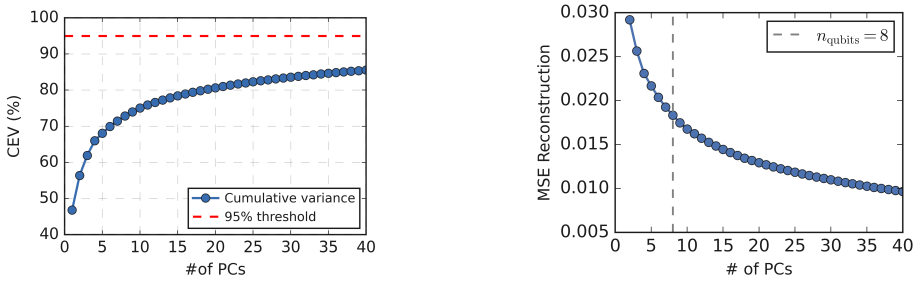


Figure 6: Representative examples from the six-class MedNIST medical-image dataset.

To enable quantum-compatible feature reduction for high-dimensional medical images, we employed principal component analysis (PCA) to project the original $4,096$-dimensional image vectors of the MedMNIST dataset into a lower-dimensional latent subspace whose dimensionality is aligned with the available qubit resources of the quantum device. PCA is applied to each client's local dataset using the training split only. The resulting cumulative explained variance (CEV) curve, shown in Fig. 7(a), illustrates the proportion of total variance captured by the first k principal components (PCs) after grayscale normalization and standardization. Each point on the curve represents the cumulative variance captured as additional components are included, reflecting how effectively the dataset can be represented in a lower-dimensional space. The red dashed line indicates the 95% variance threshold, which is a widely adopted benchmark in representation learning to preserve most of the relevant information while filtering out noise and redundancy to create an efficient feature subspace. In this work, the PCA analysis provides a principled basis for the quantum feature dimension selection. As shown in Fig. 7(a), the explained variance curve exhibits a steep increase over the first few components, followed by a gradual plateau, indicating that a relatively small subset of principal directions captures the majority of meaningful variance. In particular, approximately 80% of the total variance is captured within the first 15 PCs, whereas reaching the 95% threshold requires more than 50 components.

Additionally, Fig. 7(b) shows the relationship between the number of retained PCs and the resulting mean squared reconstruction error (MSE) for the MedMNIST dataset. As expected, the reconstruction



(a) Cumulative explained variance of MedNIST PCA. (b) Dimensionality–qubit trade-off: reconstruction error.

Figure 7: PCA analysis of the MedNIST dataset: (a) cumulative explained variance and (b) reconstruction error as a function of dimensionality.

error decreases monotonically with the number of components, confirming that most of the variance in MedNIST images is captured by the first few components. The steep decline within the first 6–10 components confirms that the majority of the dataset's statistical energy is concentrated in a small subspace of the overall feature space. Although the complete 95% variance requires more than 50 PCs, due to the physical constraints of near-term quantum hardware, particularly the limited number of available qubits, in our QFL-VQC configuration the encoding space is limited to 4–12 PCs, corresponding to the same number of qubits used in the quantum feature embedding. Despite this reduction, the selected subspace retained approximately 70–80% of the total data variance, which was empirically sufficient for accurate classification within our federated quantum variational classifier framework. This indicates that quantum feature maps and entangling circuits can recover discriminative correlations even from low-rank representations.

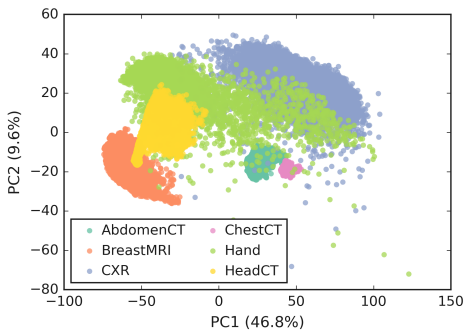


Figure 8: PCA-based feature-space visualization of MedMNIST.

Fig. 8 presents the two-dimensional PC projection of the data, where each point corresponds to an individual medical image and colors denote the six diagnostic categories. The first and second principal components account for approximately 46.8% and 9.6% of the total variance, respectively, providing an informative low-dimensional representation of the dataset's structural organization. As shown in the the scatter distribution in Fig. 8, the *BreastMRI* and *HeadCT* classes form highly compact clusters, indicating that images within each class share very similar visual characteristics and feature distributions. In contrast, the *CXR* and *Hand* categories exhibit broader spread, reflecting higher variability in texture and anatomical scale. Despite some overlap between neighboring classes, particularly *CXR* and *ChestCT*, the principal axes capture enough discriminative variance to suggest that the dataset possesses a well-structured, nearly linear separation among the classes.

It is obvious that the first two components explain approximately 56.4% of the total variance, showing that most of the meaningful variation in the dataset can be represented in a low-dimensional subspace. Moreover, Fig. 9 visualizes the spatial distribution of pixel contributions for the first $k = 4$ principal components. Each subpanel (PC1–PC4) represents a linear combination of original image features capturing the dominant variance directions in the dataset. Regions with warmer colors indicate positive correlations with the corresponding component, while cooler colors indicate negative correlations. These visual patterns highlight anatomical regions and texture gradients that primarily define the variance structure used for quantum feature encoding in the QFL–VQC framework. This

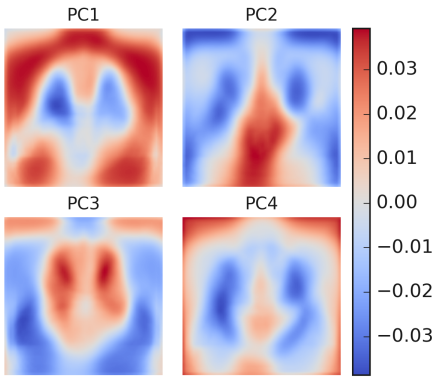


Figure 9: Spatial loadings of the first four principal components for MedNIST. Positive and negative values identify pixels contributing with opposite signs to each principal direction.

allows our design choice of using only 4–12 PCs for quantum feature encoding, consistent with the limited qubit resources available on current hardware. Mapping these PCA-derived components to qubits ensures that the quantum feature map operates on orthogonal, information-rich bases rather than raw pixel intensities. As a result, each qubit represents a distinct and interpretable global feature direction, allowing the VQC to focus on meaningful correlations instead of noise.

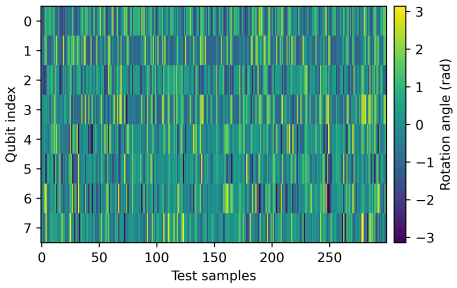


Figure 10: Example angle-encoded inputs for 300 MedNIST samples using an 8-qubit input register.

The next step involves quantum feature embedding, in which each classical vector in $\mathbb{R}^{n_q}$ is mapped into the 2^{n_q}-dimensional Hilbert space of an n_q-qubit system. In the proposed QFL-VQC model, we encode preprocessed MedNIST features into a quantum latent space using a hardware-efficient angle-embedding scheme. Each of the selected principal components is linearly mapped to a single-qubit rotation along the y-axis, forming the quantum state expression. This encoding approach was chosen after preliminary trials with amplitude-based encodings showed unstable gradients and significantly longer simulation times. Angle embedding provided consistently faster convergence and lower variance across training rounds while maintaining comparable classification accuracy. The dimensionality of the embedding was aligned with available quantum resources, in the experimental simulation we selected 6–12 principal components as the quantum input dimension, corresponding to the number of qubits per client model.

Fig. 10 presents an example visualization of the rotation-angle distribution across the 8-qubit register for a representative subset of 300 random MedMNIST samples. Each row corresponds to one qubit, and each column represents a test instance. The heatmap shows a balanced spread of rotation values within the range $[-\pi, \pi]$, indicating that the PCA-scaled features effectively utilize the qubit Hilbert space without saturating the encoding range. The spatial diversity observed along both axes suggests that each qubit receives distinct and non-redundant information, enabling the subsequent entangling layers to explore a rich subspace of qubit correlations.

Fig. 11 further examines the mean input rotation angles (in radians) per qubit, computed separately for each class in the dataset. Distinct angular patterns are observed among qubits, showing that different categories occupy different regions within the quantum rotation space. For instance, the *Hand* and *HeadCT* classes occupy higher-angle regions, while *BreastMRI* and *AbdomenCT* cluster in lower-angle zones. This variation indicates that the PCA-derived features already contain

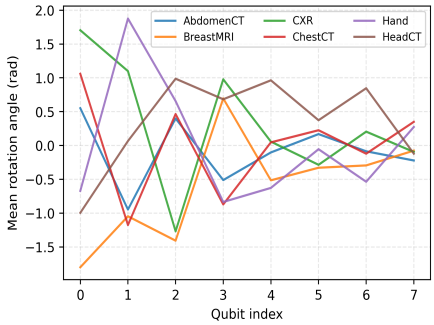


Figure 11: Class-conditional mean input rotation angles across the encoded features.

substantial diagnostic information, even before the application of entangling operations and variational parameter optimization. Within the proposed QFL–VQC circuit, these input-dependent rotations are coupled through entangling gates, generating non-local quantum correlations that further enhance the separation between diagnostic categories.

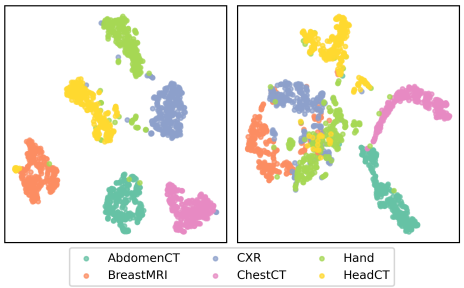


Figure 12: Two-dimensional t-SNE visualizations of the PCA-angle representation (left) and VQC-derived representation (right).

We visualize the comparison of the latent-space structures of MedNIST for the classical PCA-angle features and the VQC pre-readout features using t-SNE, as shown in Fig. 12. The left figure shows the distribution of samples in the PCA-reduced classical feature space (8 PCs), while the right figure depicts the same samples after processing through the variational circuit and measurement layer. Each point corresponds to an individual medical image, color-coded by its diagnostic category. It is observed that in the classical PCA space, the diagnostic classes form distinct but relatively overlapping clusters, suggesting that the discriminative structure is primarily governed by global variance directions. After the quantum transformation, the feature space exhibits a nonlinear structure where classes such as *BreastMRI* and *HeadCT* occupy highly compact regions with minimal overlap, while *CXR* and *Hand* data are spread across curved submanifolds that reflect the action of entangling layers, which mix feature information nonlocally across qubits. This behavior arises from the *StronglyEntanglingLayers* within the variational quantum circuit, which induce correlations among qubits that cannot be represented by classical linear operations. As a result, even though the same 8 PCs are used for encoding, the quantum transformation increases effective separability in the feature space without requiring additional dimensionality.

To investigate the effect of client number and participation level under identical data distributions, all clients were assigned IID partitions of the training dataset, as shown in Fig. 13. Each client participated in every communication round, ensuring unbiased and consistent global aggregation.

The full-participation results show a clear distinction between *convergence speed* and *final predictive performance*. Increasing the federation from three to ten clients changes final accuracy only from 96.06% to 95.22%, a difference of 0.84 percentage points. In contrast, the communication rounds required to cross the 95% threshold increase from 10 to 39. Thus, under the tested IID partition, increasing the number of clients is associated primarily with slower communication-round convergence rather than a large deterioration in the final classifier. The 10-client 50%-participation run reaches 95.08% accuracy, only 0.14 percentage points below its full-participation counterpart. It crosses 95% in fewer rounds in this particular run (30 versus 39). Because these practical results are not a

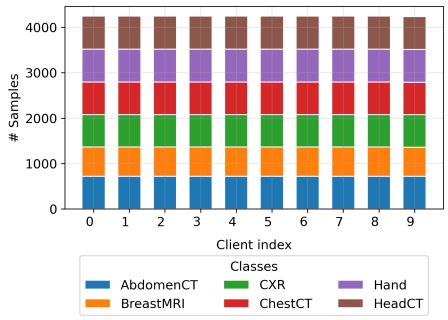


Figure 13: IID Data distribution for 10 clients.

Table 10: Practical classification performance under IID client data.

Clients	Particip.	Rounds to 95%	Final Acc.	Final Loss
3	100%	10	0.9606	0.1559
5	100%	28	0.9551	0.1783
10	100%	39	0.9522	0.1795
10	50%	30	0.9508	0.1810

paired multi-seed participation study, this difference should not be interpreted as evidence that lower participation improves convergence.

Statistical heterogeneity is generated through Dirichlet label partitioning. Smaller concentration parameter α produces more unequal client label proportions.

Table 11: Practical classification performance as Dirichlet heterogeneity increases.

Partition	α	Train	Val.	Test	AUC
IID	–	95.6	95.2	95.3	0.9957
Mild non-IID	1.0	95.6	95.4	95.2	0.9948
Moderate non-IID	0.5	94.9	94.8	94.8	0.9944
Severe non-IID	0.3	93.4	92.8	92.0	0.9910

At $\alpha = 1.0$, the test accuracy differs from the IID result by only 0.1 percentage points. At $\alpha = 0.5$, the difference grows to 0.5 percentage points, while at $\alpha = 0.3$ it reaches 3.3 percentage points. The degradation is therefore nonlinear over the tested range: mild and moderate skew have limited effect, whereas the strongest tested heterogeneity produces a clearer performance loss. AUC is considerably more stable. It decreases from 0.9957 under IID partitioning to 0.9910 at $\alpha = 0.3$, an absolute decrease of only 0.0047. Hence the strongest heterogeneous partition affects the hard classification accuracy more strongly than the ranking quality measured by AUC. The train–test gap also increases at $\alpha = 0.3$: training accuracy is 93.4% and test accuracy is 92.0%. This 1.4-percentage-point gap remains moderate, indicating that the dominant effect in this experiment is not severe overfitting but reduced global performance under the more skewed client distributions. These findings are qualitatively consistent with the trajectory-based role of statistical heterogeneity in the theory. At the synchronized global iterate, the heterogeneity does not appear as a nonzero first-order mean perturbation. Its effect arises as clients subsequently move along different local objectives.

Fig. 14 illustrates the evolution of the FedProx proximal coefficient μ_t and the learning rate during federated training under the non-IID setting with $\alpha = 0.5$ and 10 clients. The proximal coefficient decreases linearly from 5×10^{-3} to zero over approximately the first 35 communication rounds, while the learning rate remains nearly constant throughout training. This scheduling strategy stabilizes optimization during the early training stage, where heterogeneous client distributions induce larger divergence among local updates. The stronger proximal regularization constrains local models to remain closer to the global iterate, thereby suppressing excessive client drift and improving aggregation stability. As training progresses and the global model becomes increasingly consistent across clients, the proximal coefficient is gradually annealed toward zero, allowing greater flexibility for local adaptation while preserving convergence stability.

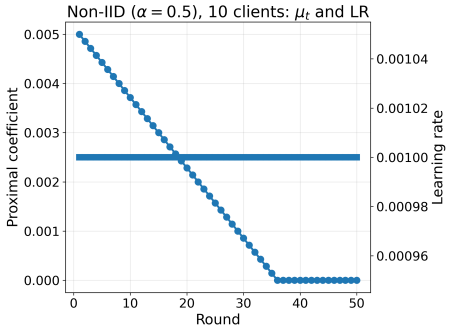


Figure 14: Evolution of the FedProx proximal coefficient μ_t and learning rate during federated training under the non-IID setting with $\alpha = 0.5$ and 10 clients.

To evaluate a larger and more heterogeneous federation, we additionally consider a six-class MedNIST experiment with 15 clients, 18,000 distributed training examples, and 11,791 test examples. Local client training-set sizes range from 591 to 1,932 examples, indicating substantial imbalance in the number of samples assigned to different clients as well as class-distribution skew.

Table 12: Summary of the 15-client non-IID MedNIST stress test.

Quantity	**Value**
Clients	15
Distributed training samples	18,000
Test samples	11,791
Client training-set range	591–1,932
Communication rounds	25
Test accuracy	87.38%
Macro-F1	0.874
Macro-AUC	0.980
Class-wise AUC range	0.960–1.000
Expected calibration error	≈ 0.034

After 25 communication rounds, the model reaches 87.38% test accuracy and macro-F1 0.874. Their close numerical values indicate that performance is not dominated by a single majority class despite the heterogeneous federation. The macro-AUC of approximately 0.980 is substantially higher than the hard accuracy, indicating that the model generally ranks the correct class well even for examples where the final argmax decision is incorrect. Class-wise AUC values range from approximately 0.960 to 1.000, which shows that the strong aggregate AUC is not solely generated by one particularly easy class. The expected calibration error is approximately 0.034, indicating a relatively small average discrepancy between confidence and empirical correctness under the calibration binning used in the experiment. Calibration should nevertheless be interpreted separately from classification accuracy: a well-ranked model can still be miscalibrated, and conversely. The training, validation, and test trajectories remain closely aligned over the reported horizon without pronounced late-round divergence. Together with the multiclass metrics, this experiment provides evidence that the practical implementation remains trainable beyond the original 3–10-client settings.

D.7 Comparison with Centralized and Local VQC Baselines

We next compare the proposed Q-FedProx method with three VQC-based structural baselines designed to isolate the roles of federation and proximal regularization. Specifically, we consider: (i) a *local-only VQC*, in which each client trains independently without server aggregation; (ii) a *centralized VQC*, trained on the pooled training data; and (iii) *Q-FedAvg*, which uses the same federated VQC architecture as the proposed method but removes the proximal term by setting $\mu_t = 0$. The proposed method, denoted *Q-FedProx-Anneal*, uses the same VQC model and federated protocol with the time-varying proximal schedule.

All four methods use the same 8-qubit, 2-layer VQC architecture and are evaluated on the same held-out test set. The federated methods are trained for 50 communication rounds with 5 local epochs

per round. Because the centralized and local-only baselines do not involve communication, we define one non-federated training round as one full epoch: the centralized model processes the pooled training set, whereas each local-only model processes only its own client data. Both non-federated baselines are trained for 50 such rounds.

Table 13: Comparison of VQC-based training on the 15-client non-IID medical-image task.

Method	Test Loss	Acc.	Macro-F1	Macro-AUC
Local-only VQC	0.6482	0.7839	0.7745	0.9542
Centralized VQC	0.2244	0.9345	0.9350	0.9933
Q-FedAvg	0.1779	0.9567	0.9568	0.9948
Q-FedProx-Anneal	**0.1663**	**0.9600**	**0.9600**	**0.9961**

Federation versus isolated local training. The local-only VQC exhibits the weakest global-test performance, reaching 78.39% accuracy, macro-F1 0.7745, and macro-AUC 0.9542. In contrast, the centralized VQC reaches 93.45% accuracy and macro-F1 0.9350. The approximately 15.1 percentage-point difference between these two non-federated references illustrates the difficulty of training independent client models under heterogeneous local distributions when no cross-client information is exchanged. Federated aggregation substantially improves the shared VQC relative to the isolated local-only baseline. Q-FedAvg reaches 95.67% test accuracy and macro-F1 0.9568, while Q-FedProx-Anneal reaches 96.00% accuracy and macro-F1 0.9600. Because the federated methods perform multiple local epochs within each communication round, we do not interpret their numerical performance relative to the 50-epoch centralized VQC as evidence that federation is intrinsically superior to pooled centralized training. Rather, the comparison shows that the federated protocol successfully combines heterogeneous local information and attains strong global predictive performance while preserving data locality.

Effect of proximal dynamics. The comparison between Q-FedAvg and Q-FedProx-Anneal directly isolates the practical effect of the proximal term, since both methods use the same VQC architecture, client partition, communication horizon, and local training budget. After 50 communication rounds, Q-FedProx-Anneal improves test accuracy from 95.67% to 96.00%, corresponding to an absolute difference of approximately 0.33 percentage points. The proximal model also reduces the test loss from 0.1779 to 0.1663, increases macro-F1 from 0.9568 to 0.9600, and increases macro-AUC from 0.9948 to 0.9961.

Although these differences are modest, they consistently favor the annealed proximal model at the end of the 50-round horizon. This result complements the theory-aligned proximal-SGD diagnostic, where proximal regularization primarily affects client trajectories and stationarity rather than guaranteeing a uniform improvement in classification accuracy. In particular, the convergence analysis does not predict that increasing μ_t must monotonically improve predictive performance. Instead, μ_t modifies the recursive accumulation of client displacement through the proximal memory term, producing a trade-off between trajectory contraction and the additional sensitivity introduced by the proximal correction. The long-horizon classification results are therefore consistent with the mechanism suggested by the theory: proximal regularization can alter the optimization trajectory in a way that modestly improves the final predictive solution, but the effect is schedule- and trajectory-dependent rather than universally beneficial.

Taken together, the VQC baselines separate two effects. First, isolated local training performs substantially worse than training mechanisms that can exploit information across client distributions, highlighting the importance of model aggregation under non-IID data. Second, within the federated VQC setting, the proposed annealed proximal dynamics yield a modest but consistent improvement over Q-FedAvg after 50 communication rounds. These findings support the role of proximal trajectory control in the practical QFL setting without making a claim of quantum predictive or computational advantage.

D.8 Experimental Scope and Limitations

The experimental results should be interpreted within several important limitations.

Finite-shot noise versus hardware bias. The finite-shot experiments directly simulate stochastic measurement sampling. Gate errors, decoherence, coherent control errors, and systematic readout bias may produce nonzero gradient bias that does not vanish as $S \to \infty$. Those effects require the bias-aware extension discussed in the theory and are not evaluated here.

No claim of quantum computational advantage. The reported classification experiments use PCA-reduced MNIST and MedNIST benchmarks. High predictive accuracy on these tasks does not establish a computational advantage over classical learning. The PCA, angle-encoding, loading-map, and t-SNE figures are diagnostic visualizations rather than comparisons against matched classical representation-learning baselines.

Barren plateaus. The bounded-second-moment assumption in the analysis is an upper bound and does not exclude exponentially small true gradients. The shallow circuits used in the experiments reduce practical trainability risk at the tested sizes, but the experiments do not establish asymptotic avoidance of barren plateaus.

Optimizer scope. The practical Adam/AdamW experiments and the proximal-SGD experiment serve different purposes. The proximal-SGD experiment directly matches the local update analyzed in the convergence theorem. The adaptive-optimizer experiments instead demonstrate practical predictive behavior and should not be treated as direct numerical verification of the SGD stationarity bound.

Partial participation. The analytical partial-participation result applies to uniform fixed-size sampling without replacement with equal client weights. Practical dropout, thresholded secure aggregation, or other client-availability mechanisms should not be identified with this theorem unless their sampling law matches the stated assumptions.