TWO-LEVEL NORMALIZATION FOR FEDERATED OPTIMIZATION UNDER GENERALIZED SMOOTHNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study federated optimization with two-level normalization: clients perform normalized stochastic local updates, and the server applies a normalized update along the aggregated client displacement. This structure is natural under generalized (L_0, L_1)-smoothness, where gradient magnitudes may be large and classical smoothness-based FedAvg analyses do not directly apply. The main difficulty is that exact normalization introduces a bias: the server direction is no longer an unbiased gradient estimate, and the bias depends on the dispersion of stochastic gradient norms within a communication round. We make this effect explicit through a norm-dispersion parameter δ. Under standard stochastic-noise and heterogeneity assumptions, we prove two-phase optimization rates up to an explicit error floor with client-and-step averaging of the stochastic term. In the non-convex setting, the rate contains both the standard sublinear component and an improved generalized-smoothness component. In the convex setting, we derive both sublinear and linear optimization behavior, up to the same bias/noise floor. We further extend the guarantees to mini-batching, partial participation, and heavy-tailed local stochastic noise, showing how batching and participation reduce stochastic and sampling-heterogeneity terms while the normalization-bias term remains explicit.

1 INTRODUCTION

Federated learning (FL) enables many clients to train a shared model while keeping their data local (McMahan et al., 2017; Kairouz et al., 2021). This paradigm is central in applications where data are naturally decentralized, privacy-sensitive, or expensive to move, such as mobile-device training and cross-silo collaboration (Bonawitz et al., 2019). Its practical appeal comes from replacing frequent centralized communication by local computation: in each communication round, selected clients perform several local optimization steps and send only model updates to a server. This same design also creates the main analytical difficulty. The participating clients may have different data distributions, different stochastic gradients, and different availability patterns, so the server must aggregate local trajectories that can differ both in direction and in magnitude.

A large body of work has developed convergence theory for federated averaging and related local methods under data heterogeneity and partial participation (Stich, 2019; Li et al., 2020b; Wang et al., 2020; Karimireddy et al., 2020; Yang et al., 2021; Wang & Ji, 2022). These results explain several important sources of error, including client drift, objective inconsistency, and the statistical effect of sampling only a subset of clients. Most of this theory, however, is built around standard smoothness assumptions and unnormalized local stochastic-gradient steps. This leaves open a complementary question that is motivated by modern deep learning: *what can be proved for federated methods when gradients may be large, local smoothness may grow with the gradient norm, and both clients and server use normalized directions?*

The smoothness issue is not merely technical. Gradient clipping and normalization are standard stabilization mechanisms for training deep networks, especially in regimes with exploding gradients or highly non-uniform curvature. Theoretical work on clipping introduced (L_0, L_1)-smoothness as a relaxed smoothness model in which the local Lipschitz constant of the gradient can scale with the gradient norm (Zhang et al., 2020b;a). This condition captures functions that are not globally smooth in the classical sense and has become a useful abstraction for understanding clipping and normalized methods. Recent work has begun to study distributed or federated optimization under related generalized-smooth regimes (Liu et al., 2022; Crawshaw et al., 2023; Demidovich et al., 2025).

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.

These analyses are based on clipping, control variates, or local-step/random-reshuffling mechanisms. In contrast, we study a different normalization pattern: every local stochastic step is normalized, and the server also takes a normalized step along the aggregated displacement.

This two-level normalization is algorithmically natural but analytically delicate. A normalized local step preserves direction while discarding magnitude; after several such steps, server normalizes average of already normalized local trajectories. The resulting direction is neither an unbiased gradient estimator nor a standard clipped gradient. Thus, the usual descent proof cannot be applied directly. The key obstacle is to compare the aggregate normalized displacement with an ideal descent direction while tracking the separate effects of local stochastic noise, worker heterogeneity, local drift, partial participation, and dispersion of stochastic-gradient norms within a communication round.

We address this problem by comparing the normalized server direction with an ideal full-gradient direction under (L_0, L_1)-smoothness. A round-dependent normalization scale separates local stochastic noise, worker heterogeneity, local drift, and an explicit normalization-bias term δ, which measures the mismatch created by replacing heterogeneous stochastic-gradient magnitudes with a common scale. Since this bias is not implied by standard variance or heterogeneity assumptions, keeping it separate makes the price of two-level normalization explicit.

We make the following contributions.

- **Convergence theory for two-level normalized federated optimization.** We study a federated method in which clients perform normalized stochastic local steps and the server performs a normalized update along the aggregated displacement. To the best of our knowledge, this gives the first convergence analysis of this two-level normalized FL dynamics under generalized (L_0, L_1)-smoothness. This setting is not covered by classical FedAvg or local-SGD analyses, and it is distinct from clipping-based generalized-smooth FL and from smoothed-normalization methods analyzed under standard smoothness.
- **Two-phase rates with explicit bias and noise terms.** Our bounds show convergence to an explicit error floor. In the non-convex case, for a target stationarity level ε, the optimization complexity contains two sublinear components: the standard L_0/ε^2 term and the improved generalized-smoothness L_1/ε term. Under convexity, for a target function-value accuracy ε, the complexity combines a sublinear component and a linear/geometric component; see Theorems 3.1 and 3.3, respectively. These phases match the qualitative behavior of recent generalized-smooth stochastic optimization results, up to the unavoidable stochastic and normalization-bias floor. In the base full-participation result, local stochastic noise is averaged as $\sigma_{loc}/\sqrt{mK}$, while δ captures norm dispersion induced by normalization. Worker heterogeneity affects local drift and appears explicitly under partial participation.
- **A proof framework for normalized federated descent.** The main technical step is to compare the server-normalized aggregate displacement with an ideal full-gradient direction inside the one-step descent bound. Since the local updates contain random inverse norms of stochastic gradients, we introduce a round-dependent common scale and isolate the resulting mismatch as the normalization-bias term δ. This comparison separates the effects of local noise, heterogeneity, local drift, and exact normalization.
- **Extensions beyond the base full-participation setting.** We extend both guarantees to local mini-batches and uniform partial worker participation. The corresponding rates contain the stochastic-noise and worker-sampling terms $\sigma_{loc}/\sqrt{nKB}$ and $\sigma_{glob}/\sqrt{n}$. We also cover heavy-tailed local stochastic gradients with $p \in (1, 2]$, where the mini-batch local-noise contribution becomes $\sigma_{loc}/(nKB)^{(p-1)/p}$.
- **Empirical validation.** We evaluate on FLAIR and FEMNIST using their original user and writer partitions. Hyperparameters are fixed before a fresh 40-seed paired comparison. Two-level normalization gives a higher final validation metric than local normalization or per-mini-batch clipping, without artificial perturbations.

2 Problem Setup and Two-Level Normalized Algorithm

We follow the common federated optimization setup used in analyses of FedAvg, local SGD, and partial participation (McMahan et al., 2017; Stich, 2019; Li et al., 2020b; Yang et al., 2021; Wang & Ji, 2022). There are m workers, each associated with a local objective F_i. The central optimization

problem is to minimize the average of the local objectives:

$$\min_{x\in\mathbb{R}^d} f(x), \qquad f(x) := \frac{1}{m}\sum_{i=1}^{m} F_i(x), \qquad F_i(x) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left[F_i(x,\xi_i)\right]. \tag{1}$$

Problem equation 1 is the standard finite-sum FL objective. We write $[m] := \{1,\dots,m\}$, and $\mathcal{D}_i$ denotes the local data distribution on worker i. At communication round t, the server may sample a subset $S_t \subseteq [m]$ of size n. The base convergence results are stated for full participation, $S_t = [m]$, and the extensions in Section 4 cover partial participation.

2.1 Assumptions on the Objective

We first state the structural assumptions on the objective. We start with a lower bound on the objective value, which lets us define the initial optimality gap and telescope the descent estimates.

Assumption 2.1 (Lower boundedness)**.** There exists $f^* > -\infty$ such that $f(x) \geq f^*$ for all $x \in \mathbb{R}^d$.

Assumption 2.1 is standard in non-convex convergence analyses and is used, for example, in recent work on generalized-smooth federated optimization and smoothed normalization (Demidovich et al., 2025; Shulgin et al., 2025).

We next impose generalized smoothness. This condition allows the local curvature to grow with the gradient norm, a regime in which clipping and normalization are particularly natural.

Assumption 2.2 ((L_0, L_1)-smoothness)**.** The global objective f and each local objective F_i, $i \in [m]$, are (L_0, L_1)-smooth: for all $x, y \in \mathbb{R}^d$ with $\|y - x\| \leq 1/L_1$,

$$\|\nabla f(y) - \nabla f(x)\| \leq (L_0 + L_1 \|\nabla f(x)\|)\, \|y - x\|,$$

and

$$\|\nabla F_i(y) - \nabla F_i(x)\| \leq (L_0 + L_1 \|\nabla F_i(x)\|)\, \|y - x\|.$$

Assumption 2.2 reduces to classical smoothness when $L_1 = 0$. For $L_1 > 0$, it allows the local Lipschitz constant of the gradient to scale with the gradient norm. This condition was introduced to explain clipping and normalization phenomena in neural-network training (Zhang et al., 2020b;a) and has since become a standard model in generalized-smooth optimization (Chen et al., 2023; Koloskova et al., 2023; Lobanov et al., 2024; Hübler et al., 2024; Gorbunov et al., 2025; Lobanov & Gasnikov, 2025; Vankov et al., 2025). Federated analyses under related unbounded or generalized smoothness assumptions include Crawshaw et al. (2023) and Demidovich et al. (2025).

2.2 Assumptions on Stochastic and Heterogeneous Noise

We now specify the two noise sources that appear in the bounds. The first one is the stochastic noise of the local gradient oracle.

Assumption 2.3 (Local stochastic noise)**.** For every worker $i \in [m]$, point $x \in \mathbb{R}^d$, and local sample $\xi_i \sim \mathcal{D}_i$, the stochastic gradient satisfies

$$\mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left[\nabla F_i(x,\xi_i)\right] = \nabla F_i(x), \qquad \mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left[\left\|\nabla F_i(x,\xi_i) - \nabla F_i(x)\right\|^2\right] \leq \sigma_{loc}^2.$$

Assumption 2.3 is the standard unbiased stochastic-gradient model with bounded second moment around the local gradient, as used in local SGD and FedAvg-type analyses (Stich, 2019; Li et al., 2020b; Wang et al., 2020).

The second source is worker heterogeneity, which captures the discrepancy between local gradients and the global gradient.

Assumption 2.4 (Worker heterogeneity)**.** For every worker $i \in [m]$ and point $x \in \mathbb{R}^d$,

$$\left\|\nabla F_i(x) - \nabla f(x)\right\|^2 \leq \sigma_{glob}^2.$$

Assumption 2.4 quantifies non-i.i.d. data across workers. Assumptions of this type, often called bounded gradient dissimilarity or bounded client drift, are common in heterogeneous FL analyses (Li et al., 2020b; Karimireddy et al., 2020; Yang et al., 2021; Wang & Ji, 2022).

2.3 TWO-LEVEL NORMALIZED ALGORITHM

Algorithm 1 gives the method studied in this paper. Each communication round has two stages. First, the selected workers start from the current server model x_t and perform K local stochastic steps.

Algorithm 1 Two-Level Normalized Federated Optimization

Input: initial point $x_0 \in \mathbb{R}^d$, local steps K, stepsizes η_{loc}, η_{glob}
for $t = 0$ **to** $T-1$ **do**
  The server samples a subset S_t of workers with $|S_t| = n$
  for each worker $i \in S_t$ in parallel **do**
    $x^i_{t,0} = x_t$
    for $k = 0$ **to** $K-1$ **do**
      $g^i_{t,k} = \nabla F_i(x^i_{t,k}, \xi^i_{t,k})$
      $x^i_{t,k+1} = x^i_{t,k} - \eta_{loc} \cdot \frac{g^i_{t,k}}{\|g^i_{t,k}\|}$
    end for
    $\Delta^i_t = x^i_{t,K} - x_t = -\eta_{loc} \sum_{k=0}^{K-1} \frac{g^i_{t,k}}{\|g^i_{t,k}\|}$
  end for
  At Server:
    Receive Δ^i_t for $i \in S_t$
    $\Delta_t = \frac{1}{|S_t|} \sum_{i \in S_t} \Delta^i_t$
    $x_{t+1} = x_t + \eta_{glob} \cdot \frac{\Delta_t}{\|\Delta_t\|}$
end for
Output: x_T

We denote the stochastic gradient computed by worker i at local step k by $g^i_{t,k}$. The unusual feature of the local update is that the step uses only the direction of $g^i_{t,k}$, namely $x^i_{t,k+1} = x^i_{t,k} - \eta_{loc} g^i_{t,k} / \left\| g^i_{t,k} \right\|$. Thus, η_{loc} controls the local step length directly, while the stochastic gradient determines only the direction. After K local steps, worker i sends the model displacement $\Delta^i_t := x^i_{t,K} - x_t = -\eta_{loc} \sum_{k=0}^{K-1} g^i_{t,k} / \left\| g^i_{t,k} \right\|$ to the server.

Second, the server averages the received displacements, $\Delta_t := |S_t|^{-1} \sum_{i \in S_t} \Delta^i_t$, and normalizes this aggregate direction before updating the model, $x_{t+1} = x_t + \eta_{glob} \Delta_t / \|\Delta_t\|$. This is the second level of normalization. It makes the global step length equal to η_{glob} and uses the averaged client displacement only as a direction. As is common for normalized methods, the practical implementation can use the convention $v / \|v\| = 0$ when $v = 0$; the analysis below is stated for the case where the displayed normalization factors are well defined almost surely.

With these notations in place, Algorithm 1 differs from FedAvg (McMahan et al., 2017) and local SGD (Stich, 2019) because each local stochastic step is normalized. It also differs from FedNova (Wang et al., 2020), where normalization corrects for the amount of local work rather than normalizing update directions. Compared with clipping-based generalized-smooth FL methods (Liu et al., 2022; Demidovich et al., 2025), our method uses exact normalization at both levels. Compared with smoothed-normalization methods, e.g., Shulgin et al. (2025), we study multi-step local stochastic trajectories under (L_0, L_1)-smoothness and explicitly track the bias induced by exact normalization.

2.4 NORMALIZATION BIAS

Exact normalization changes not only the length of a stochastic gradient, but also the relative weights with which different local gradients enter the averaged direction. Hence, even when the stochastic gradients are unbiased before normalization, the server-normalized direction in Algorithm 1 is generally biased. The key quantity is the variability of gradient norms within a round: if they are close to a common scale, replacing their inverse norms by this scale creates only a small error.

For the base full-participation analysis, we introduce the round-dependent scale

$$\alpha_\xi = \frac{1}{mK} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \frac{1}{\left\| g^i_{t,k} \right\|}.$$

Thus, $1/\alpha_\xi$ is the harmonic-mean norm scale for round t. It is computed from the same stochastic gradients that define the local updates and is used only in the analysis, not as an algorithmic parameter. Appendix E gives the common-scale calculation motivating the following assumption.

Assumption 2.5 (Norm dispersion)**.** For each communication round t,

$$\mathbb{E}\left[\frac{1}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\left|\left\|g_{t,k}^{i}\right\|-\frac{1}{\alpha_\xi}\right|\right]\le\delta_t,\qquad\delta:=\sup_{t=0,\dots,T-1}\delta_t.$$

Assumption 2.5 has a direct interpretation. It allows the stochastic gradients to have different directions, but requires their lengths within a round not to be too dispersed around the common normalization scale. Thus, δ measures the scalar bias introduced by exact normalization. This bias is different from local stochastic noise and worker heterogeneity, which are controlled by Assumptions 2.3 and 2.4; therefore, it appears as a separate error-floor term in the final rates.

2.5 Notation

We use $[m] := \{1, \dots, m\}$ for the worker set, $S_t \subseteq [m]$ for the workers selected at round t, and $n := |S_t|$ for the participation level. The number of communication rounds is T, the number of local steps is K, and B denotes the local mini-batch size in the extensions. The server iterate is x_t, the local iterate of worker i after k local steps is $x_{t,k}^i$, and $g_{t,k}^i$ is the corresponding single-sample stochastic gradient. In the mini-batch extensions, $\mathbf{g}_{t,k}^i$ denotes the averaged mini-batch stochastic gradient. The local and global stepsizes are η_{loc} and η_{glob}, while Δ_t^i and Δ_t denote the local model displacement and the averaged server displacement. We denote the generalized-smoothness constants by L_0, L_1, the local stochastic-noise and worker-heterogeneity levels by $\sigma_{loc}, \sigma_{glob}$, and the normalization-bias quantities by $\alpha_\xi, \delta_t, \delta$. Finally, λ is a hyperparameter, ε is the target accuracy, and $F_t := f(x_t) - f^*$ in the convex analysis. There, R_T denotes a deterministic finite-horizon bound on $\max_{0\le t\le T-1}\|x_t - x^*\|$, and p is the moment parameter in the heavy-tailed extension. All norms are Euclidean, and expectations are taken over the stochastic gradients and worker sampling unless stated otherwise.

3 Main Convergence Results

We now state the main convergence guarantees for the single-sample, full-participation setting. The common message is that Algorithm 1 admits two-phase optimization rates up to an explicit error floor. The optimization terms decrease with the same qualitative rates as in stochastic optimization (Lobanov & Gasnikov, 2025). The non-vanishing part of the full-participation bound is controlled by averaged local stochastic noise and the normalization-bias term δ; worker heterogeneity affects the local-drift step-size condition and appears explicitly under partial participation. The extensions to mini-batching, partial participation, and heavy-tailed local noise are stated in the next section.

3.1 Non-Convex Case

We first consider the non-convex setting, where the standard goal is to find an approximately stationary point. We state the result for the base full-participation, single-sample regime; the extensions to mini-batching, partial participation, and heavy-tailed local noise are given in Section 4.

Theorem 3.1 (Non-convex)**.** *Suppose Assumptions 2.1–2.5 hold. Consider the full-participation case $S_t = [m]$ with fresh independent single-sample local stochastic gradients. Let $\lambda > 0$ and choose*

$$\eta_{loc}\le\frac{\lambda}{8K\left(L_0+L_1\sigma_{glob}+L_1\lambda\right)},\qquad\eta_{glob}\le\frac{\lambda}{2\left(L_0+L_1\lambda\right)}.$$

Then the iterates of Algorithm 1 satisfy

$$\min_{t\in[0,T-1]}\mathbb{E}\left[\|\nabla f(x_t)\|\right]\le\frac{2F_0}{\eta_{glob}T}+4\left(\frac{\sigma_{loc}}{\sqrt{mK}}+\delta\right)+\lambda,$$

where $F_0 = f(x_0) - f^*$.

The proof is given in Appendix B. Theorem 3.1 separates decreasing optimization terms from the residual $\sigma_{loc}/\sqrt{mK} + \delta$, caused by averaged local stochastic noise and exact normalization bias.

Worker heterogeneity enters through the admissible local stepsize via local drift. In particular, with the largest admissible global stepsize and $\lambda = \varepsilon/2$, where $\varepsilon > 0$ is the target optimization accuracy,

$$\min_{t\in[0,T-1]} \mathbb{E}\left[\|\nabla f(x_t)\|\right] \le \underbrace{\frac{8L_0F_0}{\varepsilon T} + \frac{4L_1F_0}{T}}_{\text{optimization terms}} + 4\left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right) + \frac{\varepsilon}{2}.$$

Thus, before the residual floor is reached, Theorem 3.1 exhibits an analogous two-phase optimization behavior to clipped SGD under generalized smoothness (Koloskova et al., 2023): a classical $\mathcal{O}(\varepsilon^{-2})$ term and an $\mathcal{O}(\varepsilon^{-1})$ term induced by the generalized-smooth regime. Compared with FedAvg-type analyses under standard smoothness (Yang et al., 2021), Theorem 3.1 applies to two-level normalized local training under (L_0, L_1)-smoothness. In the homogeneous single-step regime $K = 1$, $n = m$, and $F_i = f$ for all workers, the heterogeneity term vanishes, $\sigma_{glob} = 0$, so the base residual is $4(\sigma_{loc}/\sqrt{m} + \delta)$. The normalization-bias term δ vanishes only under the stronger condition that the stochastic-gradient norms coincide within each round; in particular, $\delta = 0$ automatically in the single-worker single-step case $m = n = 1$ and $K = 1$, where the method reduces to ordinary NSGD even with stochastic gradients. Compared with clipped generalized-smooth FL methods (Demidovich et al., 2025), Theorem 3.1 analyzes exact normalization rather than clipping. Finally, unlike the smoothed-normalization method of Shulgin et al. (2025), which is analyzed under standard L-smoothness and yields the classical $\mathcal{O}(LF_0/\varepsilon^2)$ optimization term, Theorem 3.1 is stated under (L_0, L_1)-smoothness and gives the refined terms $\mathcal{O}(L_0F_0/\varepsilon^2)$ and $\mathcal{O}(L_1F_0/\varepsilon)$. This can be sharper when L_0 and L_1 are much smaller than a global smoothness constant L (Zhang et al., 2020b).

3.2 Convex Case

We next turn to convex objectives. Convex optimization under generalized smoothness has recently been studied for clipping, normalization, and adaptive gradient methods (Lobanov et al., 2024; Gorbunov et al., 2025; Lobanov & Gasnikov, 2025). We use the following convexity assumption.

Assumption 3.2 (Convexity)**.** The objective $f : \mathbb{R}^d \to \mathbb{R}$ is convex and attains its minimum at some point $x^* \in \arg\min_x f(x)$. We denote $f^* := f(x^*)$. Equivalently, for all $x, y \in \mathbb{R}^d$,

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle. \tag{2}$$

The next theorem states an a posteriori finite-horizon convex guarantee, up to the same normalization-bias and noise floor as in the non-convex case.

Theorem 3.3 (Convex)**.** *Suppose Assumptions 2.1–2.5 and Assumption 3.2 hold. Consider the full-participation case $S_t = [m]$ with fresh independent single-sample local stochastic gradients. Fix T and a finite deterministic R_T such that $\max_{0\le t\le T-1}\|x_t - x^*\| \le R_T$. Let $\lambda > 0$ and choose*

$$\eta_{loc} \le \frac{\lambda}{8K\left(L_0 + L_1\sigma_{glob} + L_1\lambda\right)}, \qquad \eta_{glob} \le \frac{\lambda}{2\left(L_0 + L_1\lambda\right)}.$$

Then the iterates of Algorithm 1 satisfy

$$\mathbb{E}\left[F_T\right] \le \left(1 - \frac{\eta_{glob}}{2R_T}\right)^T F_0 + 4R_T\left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right) + R_T\lambda.$$

Here $F_t := f(x_t) - f^*$. The proof in Appendix C gives a geometrically decaying optimization term and the residual scaled by the finite-horizon radius. If such an upper bound R_T is available, choosing $\lambda = \varepsilon/(2R_T)$ and the largest admissible $\eta_{glob} = \left[4R_TL_0/\varepsilon + 2L_1\right]^{-1}$ gives

$$\mathbb{E}\left[F_T\right] \le \exp\left(-\frac{T}{2R_T\left(\frac{4R_TL_0}{\varepsilon} + 2L_1\right)}\right) F_0 + 4R_T\left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right) + \frac{\varepsilon}{2}.$$

Thus, before the residual floor is reached, Theorem 3.3 exhibits two-phase convex behavior analogous to normalized stochastic methods under generalized smoothness (Lobanov & Gasnikov, 2025): a sublinear term $\mathcal{O}\left(\frac{L_0R_T^2}{\varepsilon}\log\frac{F_0}{\varepsilon}\right)$ and a linear term $\mathcal{O}\left(L_1R_T\log\frac{F_0}{\varepsilon}\right)$. In the single-worker single-step regime $m = n = 1$ and $K = 1$, the heterogeneity and normalization-bias terms vanish, $\sigma_{glob} = \delta = 0$, so Algorithm 1 reduces to NSGD even with stochastic gradients. In this regime, Theorem 3.3 matches the behavior of Lobanov & Gasnikov (2025), with explicit σ_{loc} and δ terms and without the extra M-dependent variance amplification in their NSGD bound.

4 EXTENSIONS

This section collects extensions of Theorems 3.1 and 3.3 beyond the base single-sample, full-participation setting, showing which parts of the error floor can be reduced by variance-reduction mechanisms. In the base theorem, local stochastic noise is already averaged over m workers and K local steps. With mini-batching and partial participation, this term becomes $\sigma_{loc}/\sqrt{nKB}$, while worker sampling contributes $\sigma_{glob}/\sqrt{n}$. By contrast, the normalization-bias term is unchanged, $\delta \mapsto \delta$, and remains explicit in all extensions. We first discuss mini-batching and partial participation, and then replace the second-moment local-noise assumption by a heavy-tailed moment condition.

4.1 MINI-BATCHING AND PARTIAL PARTICIPATION

Let $B \geq 1$ be the local mini-batch size and let $n \leq m$ be the number of workers sampled at each round. In Algorithm 1, mini-batching only replaces the local gradient by

$$\mathbf{g}_{t,k}^i = \frac{1}{B}\sum_{b=1}^{B} \nabla F_i(x_{t,k}^i, \xi_{t,k,b}^i),$$

where the bold symbol distinguishes it from the single-sample gradient $g_{t,k}^i$. Partial participation uses a uniformly sampled subset $S_t \subseteq [m]$ with $|S_t| = n$ and the server average $\Delta_t = n^{-1}\sum_{i\in S_t}\Delta_t^i$. The normalization-bias bound is applied to the same averaged quantity: replace $g_{t,k}^i$ by $\mathbf{g}_{t,k}^i$ when $B > 1$ and average over $i \in S_t$ when $n < m$.

The following result shows how these two modifications reduce the noise terms in the main guarantees.

Theorem 4.1 (Mini-batching and partial participation)**.** *Suppose the assumptions of Theorems 3.1 and 3.3 hold, with independent mini-batch samples and uniform worker sampling without replacement. Then Algorithm 1 satisfies, in the non-convex case,*

$$\min_{t\in[0,T-1]} \mathbb{E}\left[\|\nabla f(x_t)\|\right] \leq \frac{2F_0}{\eta_{glob}T} + 4\left(\frac{\sigma_{loc}}{\sqrt{nKB}} + \frac{\sigma_{glob}}{\sqrt{n}} + \delta\right) + \lambda.$$

Moreover, under Assumption 3.2, it satisfies

$$\mathbb{E}\left[F_T\right] \leq \left(1 - \frac{\eta_{glob}}{2R_T}\right)^T F_0 + 4R_T\left(\frac{\sigma_{loc}}{\sqrt{nKB}} + \frac{\sigma_{glob}}{\sqrt{n}} + \delta\right) + R_T\lambda.$$

The proof combines Appendices D.1 and D.2. The optimization terms are unchanged; the local stochastic-noise term becomes $\sigma_{loc}/\sqrt{nKB}$, the sampling heterogeneity term is bounded by $\sigma_{glob}/\sqrt{n}$, and δ is unchanged. The σ_{glob} term appears only because the full worker average of local gradients is replaced by the sampled average over S_t, vanishing when $n = m$.

4.2 HEAVY-TAILED LOCAL NOISE

Heavy-tailed stochastic gradients are common in modern stochastic optimization and motivate robust methods based on clipping or normalization (Gorbunov et al., 2020; Cutkosky & Mehta, 2021; Hübler et al., 2025; Liu & Zhou, 2025). We capture this regime by replacing the second-moment local-noise assumption with a finite p-th moment condition.

Assumption 4.2 (Heavy-tailed local noise)**.** For some $p \in (1, 2]$ and $\sigma_{loc} \geq 0$, the stochastic gradients are unbiased and satisfy, for all workers $i \in [m]$ and all $x \in \mathbb{R}^d$,

$$\mathbb{E}\left[\nabla F_i(x,\xi)\right] = \nabla F_i(x), \qquad \mathbb{E}\left[\|\nabla F_i(x,\xi) - \nabla F_i(x)\|^p\right] \leq \sigma_{loc}^p.$$

Assumption 4.2 coincides with the bounded-variance model Assumption 2.3 when $p = 2$, while for $p < 2$ it allows infinite variance. To quantify the effect of mini-batching in this regime, we use a martingale moment inequality for sums of independent centered stochastic gradients (von Bahr & Esseen, 1965; Pisier, 1975; Cox, 1982; Kornilov et al., 2023; Hübler et al., 2025). In particular,

$$\mathbb{E}\left[\left\|\mathbf{g}_{t,k}^i - \nabla F_i(x_{t,k}^i)\right\|^p\right] \leq \frac{2\sigma_{loc}^p}{B^{p-1}}.$$

The following result states the corresponding guarantees under heavy-tailed local noise.

Theorem 4.3 (Heavy-tailed local noise)**.** *Suppose Assumptions 2.1, 2.2, 2.4, 2.5, and 4.2 hold. Consider the mini-batch and partial-participation variant of Algorithm 1 with local batch size B and n workers sampled uniformly without replacement. Then, up to universal constants, the non-convex guarantee becomes*

$$\min_{t\in[0,T-1]} \mathbb{E}\left[\|\nabla f(x_t)\|\right] \lesssim \frac{F_0}{\eta_{glob} T} + \frac{\sigma_{loc}}{(nKB)^{(p-1)/p}} + \frac{\sigma_{glob}}{\sqrt{n}} + \delta + \lambda.$$

Moreover, under Assumption 3.2, the convex guarantee becomes

$$\mathbb{E}\left[F_T\right] \lesssim \left(1 - \frac{\eta_{glob}}{2R_T}\right)^T F_0 + R_T \left(\frac{\sigma_{loc}}{(nKB)^{(p-1)/p}} + \frac{\sigma_{glob}}{\sqrt{n}} + \delta + \lambda\right).$$

The proof is given in Appendix D.3 and uses the martingale moment inequality of Kornilov et al. (2023); Hübler et al. (2025). Theorem 4.3 preserves the same structure as the preceding guarantees. The optimization terms are the same as in Theorems 3.1 and 3.3; the partial-participation term remains $\sigma_{glob}/\sqrt{n}$ as in Theorem 4.1; and the normalization-bias term remains the same explicit δ. Only the local stochastic-noise term changes, from $\sigma_{loc}/\sqrt{nKB}$ to $\sigma_{loc}/(nKB)^{(p-1)/p}$. Thus, the two-phase behavior and explicit error floor are preserved under the weaker p-th moment assumption.

5 Comparison with Prior Works

We organize the comparison around the guarantees in Sections 3 and 4: federated local training, generalized smoothness, exact normalization, partial participation, and convexity.

Federated local training and participation. FedAvg and local SGD reduce communication by allowing several local steps between server aggregations (McMahan et al., 2017; Stich, 2019; Li et al., 2020b). Their convergence theory has been extended to heterogeneous and partially participating clients (Yang et al., 2021; Wang & Ji, 2022; Cho et al., 2023). Methods such as SCAFFOLD and FedNova address client drift or objective inconsistency through control variates or normalization by the amount of local work (Karimireddy et al., 2020; Wang et al., 2020). These results form the standard FL baseline, but their local steps are not norm-normalized. In contrast, Algorithm 1 normalizes both local stochastic steps and the server aggregate. Theorem 4.1 shows that partial participation contributes $\sigma_{glob}/\sqrt{n}$, while the normalization-bias term δ remains separate.

Generalized smoothness, clipping, and non-convex rates. The (L_0, L_1)-smoothness model was introduced to explain why clipping and normalization can be effective when the local curvature grows with the gradient norm (Zhang et al., 2020b;a). Non-convex clipping guarantees under this model were developed by Koloskova et al. (2023), and federated clipping methods by Demidovich et al. (2025). In contrast, Theorem 3.1 analyzes exact two-level normalization: the full-participation floor contains averaged local stochastic noise and the separate normalization-bias term δ, while worker heterogeneity enters through local drift and, under partial participation, through $\sigma_{glob}/\sqrt{n}$. The optimization terms keep the generalized-smooth two-phase behavior.

Convex generalized-smooth optimization. Recent convex analyses under generalized smoothness show that clipping, normalization, and adaptive gradient methods can improve over worst-case L-smooth bounds (Lobanov et al., 2024; Gorbunov et al., 2025; Lobanov & Gasnikov, 2025; Vankov et al., 2025). These works are primarily centralized. The closest comparison for our convex guarantee is Lobanov & Gasnikov (2025), who analyze stochastic normalized methods under (L_0, L_1)-smoothness. Theorem 3.3 preserves the same qualitative convex behavior, with a sublinear L_0-dependent term and a linear/geometric L_1-dependent term, but adds the federated effects: local trajectories, server normalization, averaged stochastic noise, worker sampling under partial participation, and the explicit normalization-bias floor.

Normalization in distributed optimization. Several distributed methods include normalization, but usually for a different purpose. FedNova normalizes client updates by local work to remove objective inconsistency (Wang et al., 2020), while adaptive federated optimizers use server-side adaptivity rather than normalized local directions (Reddi et al., 2021). The closest normalization-based distributed method is Shulgin et al. (2025), which combines smoothed normalization, error compensation, server normalization, and a differentially private extension under standard smoothness. Our setting keeps exact normalization at both levels and studies multi-step local stochastic trajectories under (L_0, L_1)-smoothness. Exact normalization requires tracking how far the local stochastic-gradient norms are from a common round scale; we make this effect explicit through the normalization-bias term δ in Assumption 2.5. Moreover, unlike the standard-smooth analysis of Shulgin et al. (2025), the (L_0, L_1) model yields two-phase optimization behavior before the residual error floor is reached.

6 EXPERIMENTS

We ask whether server normalization adds value beyond normalized local updates on datasets with *natural* client identities and no injected gradient, direction, or label perturbations. We compare two-level normalization against local normalization and per-mini-batch clipping.

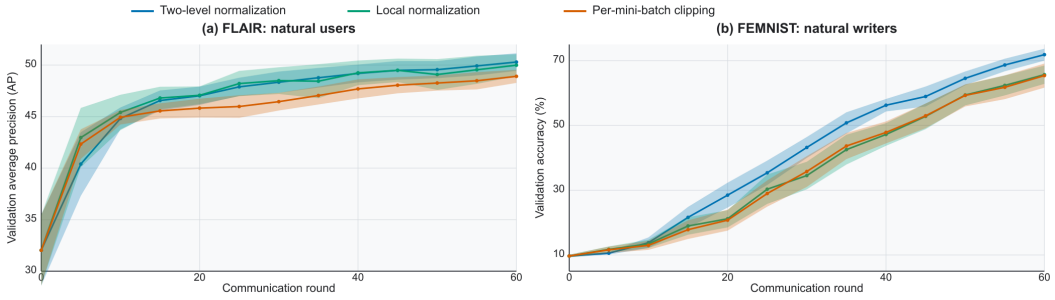


Figure 1: Five-round validation checkpoints over 40 paired seeds. Bands match Appendix A.1: ± 1 SD (FLAIR) and mean ± 1.96 SEM (FEMNIST).

FLAIR: natural users and fixed multi-step work. We use the official FLAIR user partition and imbalanced binary material label (Song et al., 2022). Images are represented by fixed 512-dimensional ImageNet-pretrained ResNet-18 features, followed by a trainable 512–128–2 MLP; average precision (AP) is the primary metric. Runs use 60 rounds, $K = 5$, $n = 100$, and $B = 32$. The base search evaluates the same 9×9 numerical grid for every method and batch-size regime; the selected regime is refined with the same union of 27 parameter pairs. The frozen settings are $(\eta_{loc}, \eta_{glob}) = (0.8, 0.375)$ for two-level, $(\eta_{loc}, \eta_{server}) = (0.8, 2.5)$ for local normalization, and $(\eta, C) = (1.2, 0.8)$ for clipping. On fresh seeds 41–80, final validation AP is 50.293 ± 0.850, 49.989 ± 1.066, and 48.907 ± 0.634, respectively. The paired gains are $+0.305$ AP points over local normalization (95% CI $[0.094, 0.519]$; $27/40$ positive; $p = 0.0385$) and $+1.386$ over clipping ($[1.085, 1.680]$; $37/40$; $p = 1.95 \times 10^{-8}$). With the settings unchanged, the pooled 100-seed effects remain positive: $+0.216$ ($[0.046, 0.383]$) and $+1.208$ ($[1.009, 1.406]$).

FEMNIST: natural writers and an almost single-step regime. We use FEMNIST digits with original writers (Caldas et al., 2019): 2707 training writers, 338 validation writers, and 338 unopened reserved-test writers. A two-convolution CNN is trained for 60 rounds with $n = 10$, $B = 64$, and $E = 0.5$ local epochs; 99.62% of client-round trajectories contain one local update. The screen gives every method the same 12 pairs in each of nine (n, B) regimes, followed by the same 27-pair refinement union. The frozen settings are $(0.4, 0.6)$ for two-level, $(0.5, 1.5)$ for local normalization, and $(\eta, C) = (6.0, 0.1)$ for clipping. On fresh seeds 161–200, final validation accuracy is $71.85 \pm 5.94\%$, $65.67 \pm 9.10\%$, and $65.40 \pm 11.99\%$, respectively. The paired gains are $+6.18$ points over local normalization (95% CI $[3.92, 8.56]$; $33/40$ positive; $p = 4.23 \times 10^{-5}$) and $+6.45$ over clipping ($[3.80, 9.74]$; $35/40$; $p = 1.38 \times 10^{-6}$). The pooled 100-seed effects are $+4.03$ ($[1.72, 6.29]$) and $+5.16$ ($[2.84, 7.58]$). In a separately tuned $n = 20$ boundary, the effect versus local normalization is unresolved at -0.35 ($[-4.19, 3.50]$), while the effect versus clipping is $+6.53$ ($[1.52, 11.97]$).

Scope and additional evidence. The FLAIR and FEMNIST results show that server normalization can improve over both comparators with natural client identities, while the separately tuned $n = 20$ FEMNIST setting shows that the improvement over local normalization is not universal. Appendix A.1 reports full-resolution learning curves, paired-seed effects, tuning procedures, and per-seed checks. Appendix A.2 summarizes favorable and unfavorable regimes, and Appendix A.3 examines server-scale variation and variable local work. Controlled CIFAR-10 gradient-scale outliers, AG News, and synthetic diagnostics are reported in Appendices A.4–A.6.

7 CONCLUSION

We analyze exact two-level normalization under (L_0, L_1)-smoothness, with two-phase guarantees that separate local noise, partial participation, and normalization bias δ. On FLAIR and FEMNIST with natural client identities, two-level normalization yields statistically resolved paired improvements over local normalization and clipping under the evaluated protocols. Additional experiments identify settings in which the improvement is absent or statistically unresolved.

AI USE STATEMENT

Generative AI tools were used for language editing and to improve the clarity of the manuscript. We reviewed all resulting changes and take full responsibility for the final content.

REFERENCES

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé Kiddon, Jakub Konečný, Stefano Mazzocchi, H. Brendan McMahan, et al. Towards federated learning at scale: System design. In *Proceedings of Machine Learning and Systems*, volume 1, pp. 374–388, 2019.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konecný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2019.

Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5396–5427, 2023.

Yeshwanth Cherapanamjeri, Nilesh Tripuraneni, Peter Bartlett, and Michael Jordan. Optimal mean estimation without a variance. In *Conference on Learning Theory*, pp. 356–357. PMLR, 2022.

Yae Jee Cho, Pranay Sharma, Gauri Joshi, Zheng Xu, Satyen Kale, and Tong Zhang. On the convergence of federated averaging with cyclic client participation. In *International Conference on Machine Learning*, pp. 5677–5721. PMLR, 2023.

David C Cox. Note on a martingale inequality of pisier. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 92, pp. 163–165. Cambridge University Press, 1982.

Michael Crawshaw, Yajie Bao, and Mingrui Liu. Federated learning with client subsampling, data heterogeneity, and unbounded smoothness: A new algorithm and lower bounds. In *Advances in Neural Information Processing Systems*, volume 36, pp. 22459–22507, 2023.

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4883–4895, 2021.

Yury Demidovich, Petr Ostroukhov, Grigory Malinovsky, Samuel Horváth, Martin Takáč, Peter Richtárik, and Eduard Gorbunov. Methods with local steps and random reshuffling for generally smooth non-convex federated optimization. In *International Conference on Learning Representations*, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15042–15053, 2020.

Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex (l_0, l_1)-smooth optimization: Clipping, acceleration, and adaptivity. In *International Conference on Learning Representations*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Florian Hübler, Junchi Yang, Xiang Li, and Niao He. Parameter-agnostic optimization under relaxed smoothness. In *International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4861–4869, 2024.

Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed sgd. In *International Conference on Artificial Intelligence and Statistics*, pp. 2413–2421. PMLR, 2025.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14 (1–2):1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U. Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pp. 17343–17363. PMLR, 2023.

Nikita Kornilov, Ohad Shamir, Aleksandr Lobanov, Darina Dvinskikh, Alexander Gasnikov, Innokentiy Shibaev, Eduard Gorbunov, and Samuel Horváth. Accelerated zeroth-order method for non-smooth stochastic convex optimization problem with infinite variance. *Advances in Neural Information Processing Systems*, 36:64083–64102, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. In *Technical report*, 2009.

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.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-IID data. In *International Conference on Learning Representations*, 2020b.

Mingrui Liu, Zhenxun Zhuang, Yunwen Lei, and Chunyang Liao. A communication-efficient distributed gradient clipping algorithm for training deep neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26204–26217, 2022.

Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *International Conference on Learning Representations*, 2025.

Aleksandr Lobanov and Alexander Gasnikov. Power of generalized smoothness in stochastic convex optimization: First-and zero-order algorithms. *arXiv preprint arXiv:2501.18198*, 2025.

Aleksandr Lobanov, Alexander Gasnikov, Eduard Gorbunov, and Martin Takáč. Linear convergence rate in convex setup is possible! gradient descent method variants under (l_0, l_1)-smoothness. *arXiv preprint arXiv:2412.17050*, 2024.

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, 2017.

Gilles Pisier. Martingales with values in uniformly convex spaces. *Israel Journal of Mathematics*, 20 (3):326–350, 1975.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

Egor Shulgin, Sarit Khirirat, and Peter Richtárik. Private federated learning with provable convergence via smoothed normalization. In *ICML 2025 Workshop on Collaborative and Federated Agentic Workflows*, 2025.

Congzheng Song, Filip Granqvist, and Kunal Talwar. FLAIR: Federated learning annotated image repository. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, volume 35, 2022.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

Daniil Vankov, Anton Rodomanov, Angelia Nedić, Lalitha Sankar, and Sebastian U. Stich. Optimizing (l_0, l_1)-smooth functions by gradient methods. In *International Conference on Learning Representations*, 2025.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, pp. 299–303, 1965.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7611–7623, 2020.

Shiqiang Wang and Mingyue Ji. A unified analysis of federated learning with arbitrary client participation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19124–19137, 2022.

Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-IID federated learning. In *International Conference on Learning Representations*, 2021.

Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15511–15521, 2020a.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020b.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

APPENDIX

CONTENTS

1 Introduction **1**

2 Problem Setup and Two-Level Normalized Algorithm **2**

2.1 Assumptions on the Objective 3

2.2 Assumptions on Stochastic and Heterogeneous Noise 3

2.3 Two-Level Normalized Algorithm 4

2.4 Normalization Bias 4

2.5 Notation 5

3 Main Convergence Results **5**

3.1 Non-Convex Case 5

3.2 Convex Case 6

4 Extensions **7**

4.1 Mini-Batching and Partial Participation 7

4.2 Heavy-Tailed Local Noise 7

5 Comparison with Prior Works **8**

6 Experiments **9**

7 Conclusion **9**

A Additional Experimental Details and Results **14**

A.1 Natural-Client Experiments 14

A.2 Where the Second Level Helps—and Where It Does Not 16

A.3 Controlled Server-Normalization Diagnostics 17

A.4 CIFAR-10 Gradient-Scale Robustness Stress Test 18

A.5 Clean Real-Data Experiment: AG News 19

A.6 Synthetic Mechanism Diagnostics 20

A.7 Data, Evaluation, and Compute Details 21

B Proof of Theorem 3.1 **21**

C Proof of Theorem 3.3 **28**

D Extension Proofs **29**

D.1 Extension to mini-batch stochastic gradients 29

D.2 Extension to partial worker participation 30

D.3 Extension to heavy-tailed stochastic gradients 31

E Motivation for the Normalization-Bias Assumption **32**

F Pre-Hitting Fejér Stability under Uniform Directional Control **34**

F.1 Pathwise directional-error decomposition 35

F.2 Pre-hitting Fejér stability 36

F.3 A conservative high-probability realization 39

A Additional Experimental Details and Results

This appendix provides the experimental details and additional diagnostics supporting Section 6. We first report the complete FLAIR and FEMNIST analyses with natural client identities. We then compare favorable and unfavorable regimes and examine server-scale variation, variable local work, gradient-scale outliers, AG News, and synthetic settings. Appendix A.7 provides data, evaluation, and compute details.

A.1 Natural-Client Experiments

Both primary experiments use natural client identities and introduce no synthetic partition, gradient rescaling, direction corruption, or label corruption. Each method receives the same number of tuned numerical pairs. After method-specific selection, all settings are frozen and evaluated on 40 fresh paired seeds; paired seeds share initialization and client sampling. The reported endpoints are validation metrics.

FLAIR natural users. We use the official FLAIR train/validation user partition and binary material label (Song et al., 2022). Images are represented by 512-dimensional features from an ImageNet-pretrained ResNet-18, followed by a trainable 512–128–2 MLP. Each run has 60 rounds, fixed $K = 5$, $n = 100$, and $B = 32$. The selected settings are $(\eta_{loc}, \eta_{glob}) = (0.8, 0.375)$ for two-level normalization, $(\eta_{loc}, \eta_{server}) = (0.8, 2.5)$ for local normalization, and $(\eta, C) = (1.2, 0.8)$ for clipping. On 40 fresh seeds, final validation AP is 50.293 ± 0.850 for two-level, 49.989 ± 1.066 for local normalization, and 48.907 ± 0.634 for clipping. The paired effects are $+0.305$ AP points (95% CI $[0.094, 0.519]$) and $+1.386$ ($[1.085, 1.680]$), respectively.

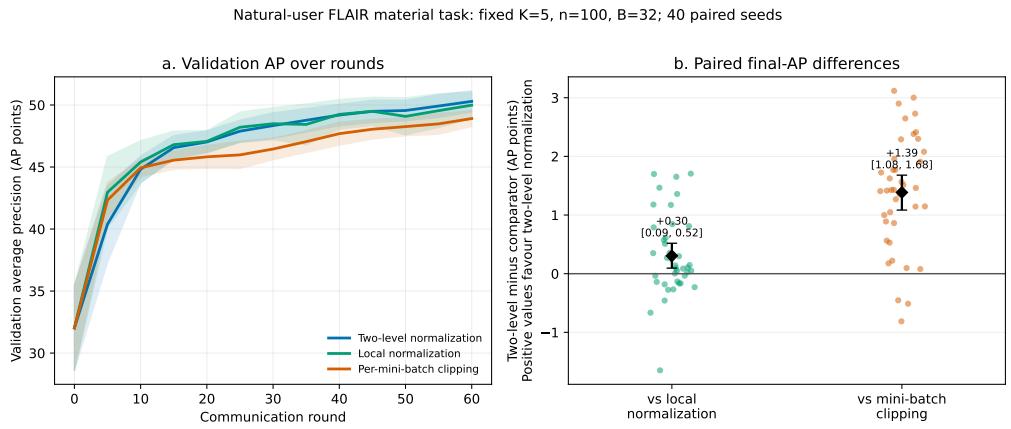


Figure 2: FLAIR with natural user identities. Left: validation-AP trajectories over 40 fresh paired seeds. Right: paired final-AP effects of two-level normalization; positive values favor the proposed method.

Extending the same frozen configurations to 100 paired seeds gives effects of $+0.216$ AP points versus local normalization (95% CI $[0.046, 0.383]$) and $+1.208$ versus clipping ($[1.009, 1.406]$). We treat this pooled analysis as a consistency check, not as the primary test, because the additional seeds were accumulated after the initial block.

FEMNIST natural writers. We use the LEAF/TensorFlow Federated FEMNIST digit subset (Caldas et al., 2019), with a client equal to an original writer. A fixed split contains 2707 training, 338 validation, and 338 reserved-test writers; the reserved test remains unopened. A two-convolution CNN is trained for 60 rounds with $n = 10$, $B = 64$, and $E = 0.5$ local epochs. The selected settings are $(\eta_{loc}, \eta_{glob}) = (0.4, 0.6)$, $(\eta_{loc}, \eta_{server}) = (0.5, 1.5)$, and $(\eta, C) = (6.0, 0.1)$ for two-level, local-only, and clipping, respectively. In 99.62% of method-round trajectories, a selected client takes one local step. On 40 fresh seeds, final validation accuracy is $71.85 \pm 5.94\%$, $65.67 \pm 9.10\%$, and $65.40 \pm 11.99\%$. The paired effects are $+6.18$ percentage points over local normalization (95% CI $[3.92, 8.56]$) and $+6.45$ over clipping ($[3.80, 9.74]$).

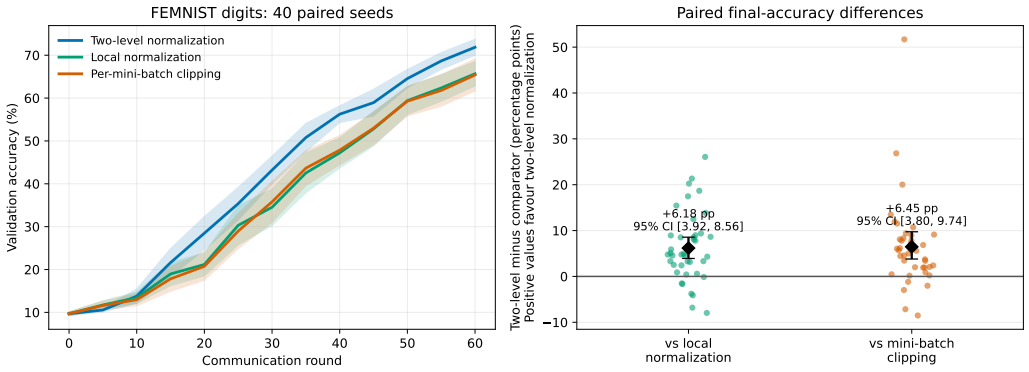


Figure 3: FEMNIST with natural writer identities at $n = 10$. Left: validation-accuracy trajectories over 40 fresh paired seeds. Right: paired endpoint effects; positive values favor two-level normalization.

Pooling the frozen configurations across 100 paired seeds gives $+4.03$ points versus local normalization (95% CI $[1.72, 6.29]$) and $+5.16$ versus clipping ($[2.84, 7.58]$). A separately tuned $n = 20$ protocol supplies a nearby boundary: the two-level–local effect is -0.35 points ($[-4.19, 3.50]$), while the two-level–clipping effect remains positive at $+6.53$ ($[1.52, 11.97]$). Thus the incremental gain over local normalization is present at $n = 10$ but is not resolved at $n = 20$; this is an empirical boundary, not a claim that participation alone causes the difference.

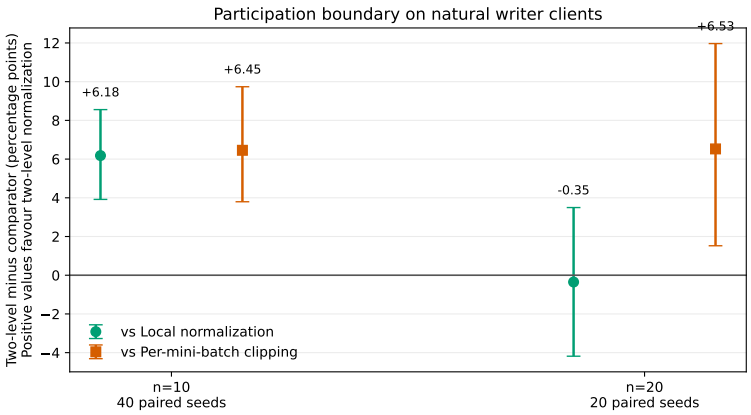


Figure 4: FEMNIST participation boundary under separately tuned protocols. The gain over local normalization is resolved at $n = 10$ and unresolved at $n = 20$.

A.2 Where the Second Level Helps—and Where It Does Not

Table 1 summarizes all paired endpoint comparisons reported in this appendix. Each entry is signed as two-level normalization minus the named comparator; intervals crossing zero indicate statistically unresolved differences. The two natural-client experiments provide evidence of improvement, whereas matched clean CIFAR-10, directional corruption, and clean AG News identify settings in which two-level normalization has no resolved advantage.

Table 1: Paired endpoint effects across natural and controlled regimes. Values are mean differences with paired-bootstrap 95% CIs. Units are AP points for FLAIR and percentage points otherwise.

Regime	vs. local norm	vs. clipping
FLAIR material, natural users	$+0.305$ $[0.094, 0.519]$	$+1.386$ $[1.085, 1.680]$
FEMNIST digits, $n = 10$	$+6.18$ $[3.92, 8.56]$	$+6.45$ $[3.80, 9.74]$
FEMNIST digits, $n = 20$	-0.35 $[-4.19, 3.50]$	$+6.53$ $[1.52, 11.97]$
Matched clean CIFAR-10	-2.90 $[-3.42, -2.38]$	-3.19 $[-3.63, -2.92]$
Clean CIFAR-10, high heterogeneity	$+1.05$ $[0.56, 1.63]$	-0.97 $[-2.89, 0.29]$
Mild scale outliers	$+1.31$ $[0.35, 2.27]$	$+1.41$ $[0.50, 2.50]$
Severe scale outliers	$+1.05$ $[-0.75, 2.80]$	$+0.95$ $[-0.36, 2.09]$
Directional corruption	-1.37 $[-2.49, 0.16]$	-2.56 $[-4.13, -0.02]$
AG News, clean ($K = 5$)	$+1.30$ $[0.11, 2.65]$	-1.71 $[-2.48, -0.94]$

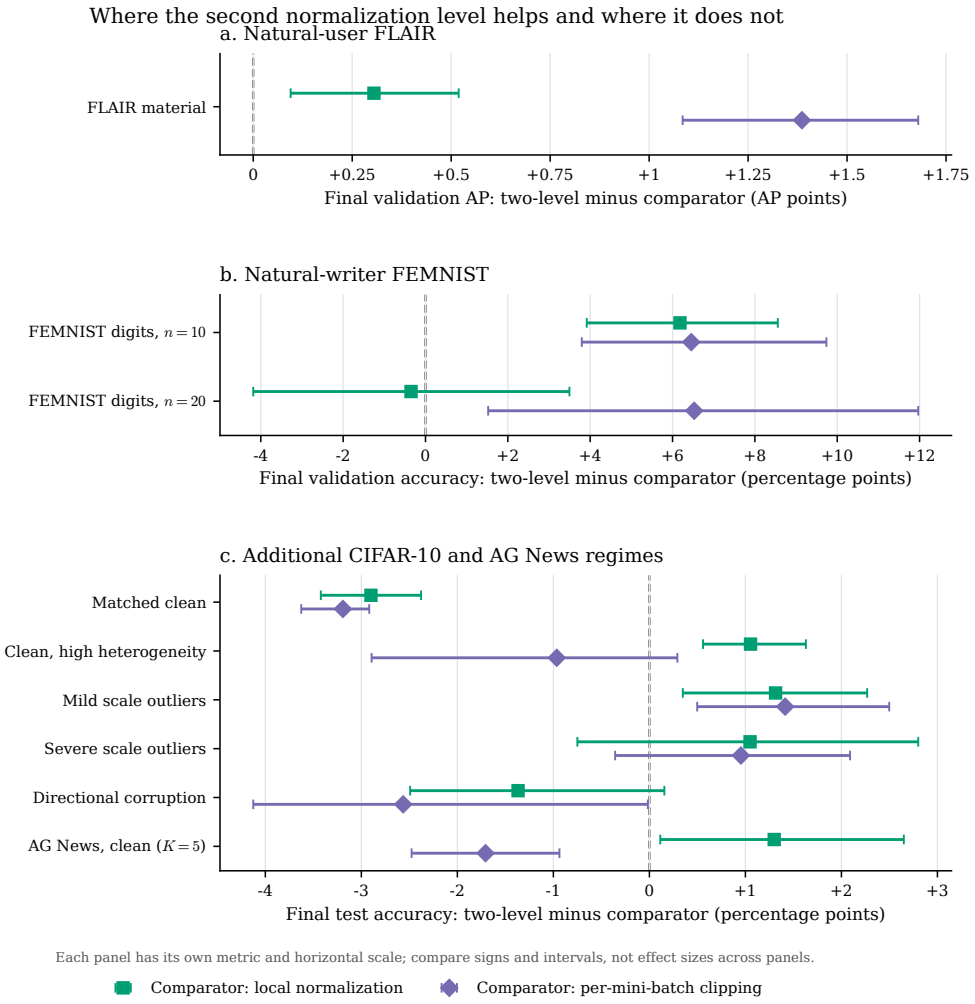


Figure 5: Visual summary of Table 1. The sign and uncertainty of the paired effects expose both positive and unfavorable regimes rather than selecting only favorable comparisons.

A.3 Controlled Server-Normalization Diagnostics

To isolate the server-level operation after normalized local training, a controlled CIFAR-10/ResNet-18 phase map applies an intervention only after the aggregate client displacement is formed. The study uses a Dirichlet-0.1 partition, $m = 100$, $n = 10$, $K = 5$, $B = 16$, 240 rounds, and twelve paired seeds. Aggregate scale is stable, moderately variable, or highly variable; directional fidelity to the unperturbed aggregate is independently set to cosine 1.00, 0.94, or 0.71. The measured outcome is normalized test-accuracy AUC (AccAUC), not endpoint accuracy.

Two-level minus local-only AccAUC ranges from $+0.47$ to $+0.86$ points under stable scale, from $+1.34$ to $+1.45$ under moderate variation, and from $+6.25$ to $+8.85$ under high variation. Under high variation, two-level normalization also exceeds frozen clean-selected server clipping by $+2.75$ to $+3.60$ points across the three direction cells. This result is a controlled mechanism diagnostic: it is consistent with exact server normalization removing common positive scale variation, but it does not by itself identify the cause of the natural-data gains or exclude clipping retuned for every intervention.

The clean variable-work boundary gives the complementary result. With exact mean local work $K = 5$, two-level minus local-only AccAUC is $+0.86$ ($[0.51, 1.26]$) for fixed work and $+0.50$ ($[0.23, 0.73]$) when work varies across clients within every round, but -1.56 ($[-2.07, -1.03]$) when all clients share round-dependent work. Aggregate magnitude can therefore be nuisance variation or useful information; server normalization should not be read as uniformly beneficial.

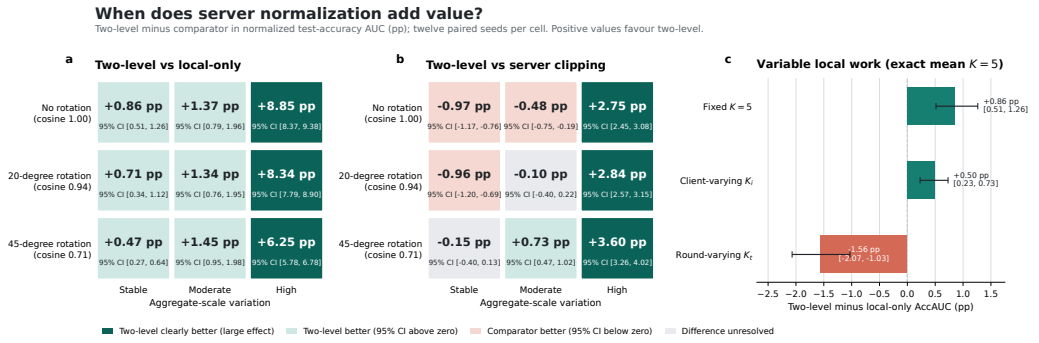


Figure 6: Controlled decision map for the second normalization level. Panels vary post-aggregate scale, direction, and endogenous local work. The figure distinguishes favorable nuisance-scale variation from a boundary where aggregate magnitude carries information.

A.4 CIFAR-10 Gradient-Scale Robustness Stress Test

The experiment in Figure 7 is designed to stress-test the central motivation for exact normalization: stochastic-gradient magnitudes may be unstable even when their directions remain informative. It instantiates the non-convex federated setup in Sections 3 and 4: we train a ResNet-18 image classifier (He et al., 2016) on CIFAR-10 (Krizhevsky et al., 2009) with cross-entropy loss, stochastic mini-batches, local client steps, and partial participation, in a standard supervised FL setting. The training data are split over $m = 100$ clients by a Dirichlet label partition with concentration 0.5, creating a standard non-IID FL benchmark. Each communication round samples $n = 10$ clients, and each selected client performs $K = 5$ local steps with mini-batch size $B = 16$. We inject multiplicative scale outliers by multiplying each local mini-batch gradient by 20 independently with probability 0.1. This perturbation leaves the direction of the affected mini-batch gradient unchanged but makes its norm unreliable, making the setting a direct test of whether local and server normalization can remove harmful magnitude information while preserving useful direction information.

We compare six methods: FedAvg (McMahan et al., 2017), FedAvgM (Reddi et al., 2021), FedProx (Li et al., 2020a), local normalization, server normalization, and two-level normalization (Algorithm 1). As in the main text, local normalization normalizes the local stochastic steps and uses the unnormalized server aggregate with $\eta_{glob} = 1$, whereas server normalization keeps unnormalized local SGD steps and normalizes only the server aggregate. Hyperparameters are selected separately for each method on a validation split using 40 communication rounds and then fixed for the final 240-round test runs. Table 2 reports the selected hyperparameters. Final results are averaged over three independent seeds, and the complete validation grid is stored in `tuning_grid.csv`.

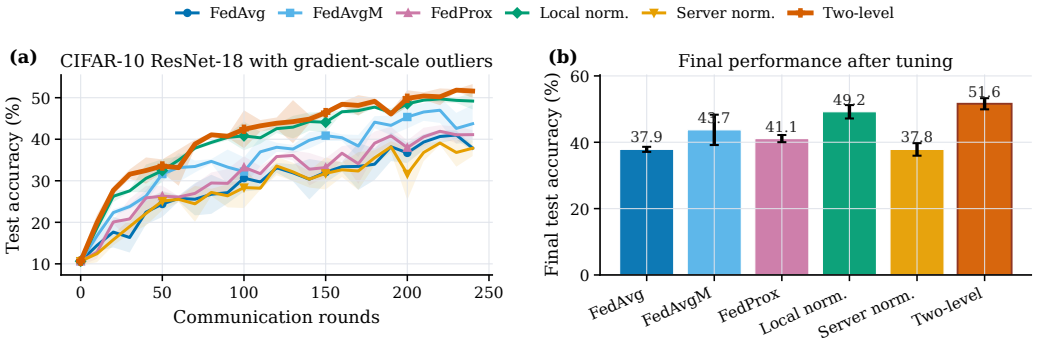


Figure 7: Robustness to gradient-scale outliers on CIFAR-10 with ResNet-18. Curves and bars show mean $\pm$ one standard deviation over three seeds after validation-based hyperparameter selection.

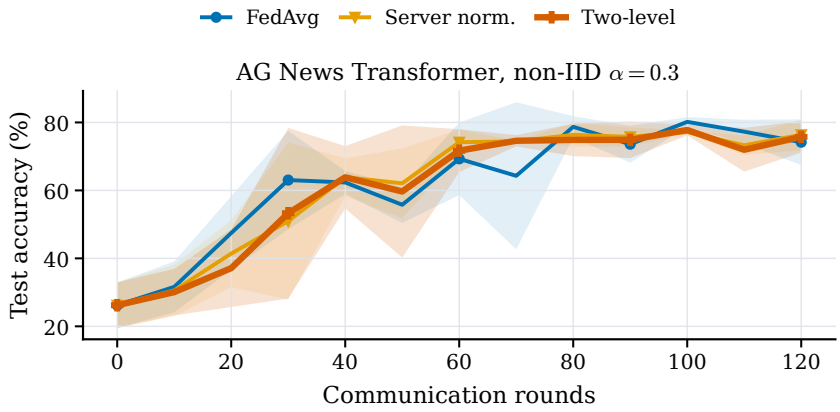


Figure 8: Clean AG News experiment after validation-based hyperparameter selection. Curves and shaded regions show mean $\pm$ one standard deviation over three seeds.

Table 2: Validation-selected hyperparameters and final test accuracy for the CIFAR-10 outlier stress test. Accuracies are reported in percent.

Method	Selected hyperparameters	Validation acc.	Test acc.
FedAvg	$\eta = 0.01$	23.76	37.86 ± 0.77
FedAvgM	$\eta = 0.001,\ \beta = 0.9$	27.12	43.74 ± 4.57
FedProx	$\eta = 0.0025,\ \mu = 0.01$	27.22	41.11 ± 1.08
Local norm	$\eta_{loc} = 0.12$	33.62	49.19 ± 2.02
Server norm	$\eta = 0.005,\ \eta_{glob} = 0.08$	24.50	37.84 ± 1.89
Two-level	$\eta_{loc} = 0.12,\ \eta_{glob} = 0.16$	33.00	51.61 ± 1.68

Figure 7 and Table 2 show a clear separation between the two normalization mechanisms. Local normalization substantially improves over the FedAvg-type baselines, which indicates that removing unreliable stochastic-gradient magnitudes before the local trajectory is formed is crucial in this outlier regime. Server normalization alone does not provide the same benefit: although it controls the length of the aggregated displacement, it cannot undo distortions accumulated during the local client updates. Two-level normalization combines both effects and obtains the best final accuracy, $51.61 \pm 1.68\%$, as well as the lowest final test loss. This behavior is consistent with our theory: the descent bound separates local stochastic noise, worker heterogeneity, local drift, and normalization bias, and the experiment shows that controlling normalization at both levels is beneficial when the dominant perturbation is the scale of the stochastic gradients.

A.5 Clean Real-Data Experiment: AG News

We next consider a clean real-data text-classification task without injected gradient-scale outliers. The goal is not to claim state-of-the-art accuracy, but to verify that the same federated normalization implementation remains competitive on a standard supervised learning task. We use AG News (Zhang et al., 2015), a four-class news classification dataset, and fine-tune a compact Transformer classifier `google/bert_uncased_L-2_H-128_A-2` (Devlin et al., 2019) by minimizing cross-entropy loss.

The training split uses 6000 examples per class, with 1000 validation examples per class and 1500 test examples per class. We distribute the training data across $m = 40$ clients using a Dirichlet label partition with concentration 0.3, so each client observes a different class mixture. At each round, the server samples $n = 8$ clients; each selected client performs $K = 5$ stochastic local step with mini-batch size $B = 16$. Hyperparameters are tuned for 60 communication rounds on the validation split and then fixed for 120-round test runs over three seeds. Since this experiment is intended as a clean real-data normalization diagnostic rather than an outlier stress test, we compare FedAvg (McMahan et al., 2017) with the two normalization variants that performed best in the robustness study: server normalization, which normalizes only the server aggregate, and two-level normalization (Algorithm 1). Table 3 reports the selected hyperparameters.

Table 3: Validation-selected hyperparameters and final test accuracy for the AG News clean real-data experiment. Accuracies are reported in percent.

Method	Selected hyperparameters	Validation acc.	Test acc.
FedAvg	$\eta = 0.02$	67.97	74.18 ± 6.74
Server norm	$\eta = 0.005,\ \eta_{glob} = 0.024$	68.88	76.36 ± 4.52
Two-level	$\eta_{loc} = 0.006,\ \eta_{glob} = 0.024$	69.12	75.77 ± 4.62

Figure 8 and Table 3 show that the normalization variants remain competitive on clean real data. Two-level normalization gives the best validation accuracy, suggesting that local normalization can stabilize early stochastic client updates even without injected scale outliers. On the final test runs, server normalization has the highest mean accuracy, while two-level normalization is close but slightly lower. A likely reason is that this experiment uses only one local step, $K = 5$: without a multi-step local trajectory, scale distortions cannot accumulate before aggregation, and server normalization already controls the aggregate displacement. Thus, AG News serves as a clean-data sanity check, whereas the constructed gradient-scale stress test in Appendix A.4 favors two-level normalization.

A.6 Synthetic Mechanism Diagnostics

The synthetic experiment is designed to isolate the three error-floor mechanisms that are difficult to disentangle on real data. It directly probes the extensions in Section 4: the normalization-bias term δ, the mini-batch reduction of local stochastic noise, and the partial-participation reduction of worker-sampling error. We use a controlled quadratic federated objective with

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x), \qquad f_i(x) = \frac{s_i}{2}\left\|x - a_i\right\|^2,$$

where $a_i \in \mathbb{R}^d$ is the client center and $s_i > 0$ is a client-specific curvature scale. The stochastic local gradient is

$$g_{t,k}^i = s_i(x_{t,k}^i - a_i) + \zeta_{t,k}^i, \qquad \zeta_{t,k}^i \sim \mathcal{N}\left(0, \tau_i^2 I_d/B\right),$$

so B directly controls the local stochastic-noise level. We use $m = 128$ clients, dimension $d = 256$, 600 communication rounds, and five seeds. The first sweep increases the dispersion of stochastic-gradient norms by varying the curvature scales s_i and noise scales τ_i. The second sweep varies the mini-batch size B in a high local-noise regime. The third sweep varies the number of participating clients n in a high-heterogeneity regime. All other parameters are kept fixed within each sweep so that the changed mechanism and its corresponding bound term are isolated in the plotted trend. At each communication round we compute the empirical normalization-bias proxy $\widehat{\delta}_t = \frac{1}{nK}\sum_{i\in S_t}\sum_{k=0}^{K-1}\left|\left\|g_{t,k}^i\right\| - 1/\widehat{\alpha}_t\right|$, where $\widehat{\alpha}_t = \frac{1}{nK}\sum_{i\in S_t}\sum_{k=0}^{K-1} 1/\left\|g_{t,k}^i\right\|$. The quantity reported in Figure 9 is the average of $\widehat{\delta}_t$ over the final 20% of logged rounds.

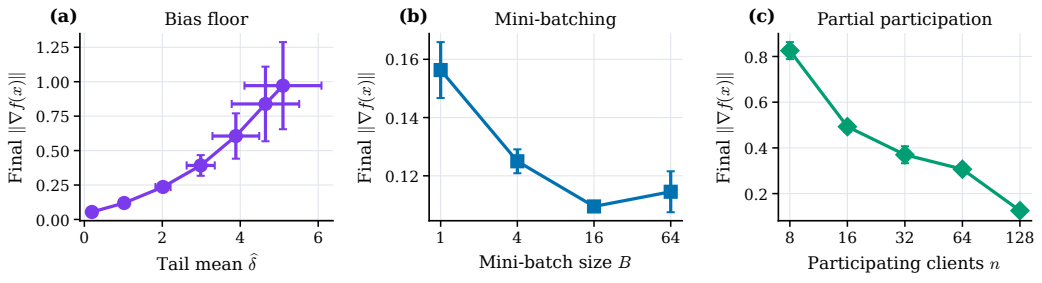


Figure 9: Synthetic diagnostics illustrating the mechanisms in the theory: normalization bias sets the limiting floor, while mini-batching and participation reduce noise-related components.

Figure 9 matches the qualitative structure of the theoretical bounds. Panel (a) shows that increasing the empirical normalization-bias proxy raises the final stationarity floor, supporting the role of δ as a non-vanishing bias term. Panel (b) shows that increasing B reduces the local stochastic-noise contribution, in line with the $\sigma_{loc}/\sqrt{nKB}$ dependence when n and K are fixed. Panel (c) shows that increasing n reduces the worker-sampling contribution, consistent with the $\sigma_{glob}/\sqrt{n}$ term.

Importantly, the bias proxy does not systematically decrease with B or n, which reflects the theory: mini-batching and participation reduce noise-related components, whereas δ remains a separate normalization effect.

A.7 Data, Evaluation, and Compute Details

All datasets are public. FLAIR uses the official reduced-image release; its image license requires users to obtain the images directly from the data provider. FEMNIST is loaded from the public TensorFlow Federated/LEAF release, with the supplied writer identities retained. CIFAR-10 is loaded through `torchvision.datasets.CIFAR10`. AG News is loaded through the HuggingFace datasets library, and its classifier uses the compact pretrained `google/bert_uncased_L-2_H-128_A-2` checkpoint. The synthetic diagnostics generate their data procedurally. No private dataset is used.

All methods are tuned on validation data before the reported evaluation runs, and competing methods receive the same number of candidate numerical parameter pairs within each study. Paired runs share the model initialization and client-sampling randomness. The natural-client experiments report validation endpoints because the FLAIR test data and FEMNIST reserved-test writers were not used. The reported confidence intervals are paired-bootstrap 95% intervals; exact two-sided sign tests and positive-pair counts are additionally reported for the primary 40-seed comparisons in Section 6. The relevant subsections and tables specify the model, number of communication rounds, client participation, local work, batch size, seed count, hyperparameter search, and selected configuration for each experiment.

The experiments were implemented in Python and PyTorch and run on NVIDIA A100 80GB PCIe GPUs. The complete validation grids dominate the computational cost. The CIFAR-10 robustness study took 7148.3 seconds on one GPU, including validation tuning and final evaluation; the AG News grid took 999.4 seconds, and the synthetic diagnostics took 95.5 seconds. The server-normalization phase map and variable-work study comprise 432 final method–seed runs. They took approximately 2.3 hours of wall-clock time on four GPUs with two concurrent shards per GPU, corresponding to approximately 9 allocated GPU-hours. Deterministic seeds were used, and nondeterministic cuDNN autotuning was disabled where applicable; minor numerical variation can nevertheless arise across software and hardware versions.

B Proof of Theorem 3.1

Proof. We consider the full-participation case, i.e., $S_t = [m]$ and $\Delta_t = \frac{1}{m}\sum_{i=1}^m \Delta_t^i$. Let $\mathcal{F}_t$ denote the history before round t. All one-round expectations below can be read conditionally on $\mathcal{F}_t$; to keep the notation light, we omit this conditioning and take total expectation after the one-step estimates are derived. For a fixed round t, recall that

$$\alpha_\xi = \frac{1}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\frac{1}{\left\|g_{t,k}^i\right\|}.$$

When $L_1 > 0$, the stepsize conditions of Theorem 3.1 imply

$$\eta_{glob}L_1 \le \frac{1}{2}, \qquad K\eta_{loc}L_1 \le \frac{1}{8}.$$

Hence, $\|x_{t+1} - x_t\| \le 1/L_1$ and $\left\|x_{t,k}^i - x_t\right\| \le K\eta_{loc} \le 1/L_1$. Thus, all applications of Assumption 2.2 below are within its admissible neighborhood. When $L_1 = 0$, the distance restriction is vacuous. Then we have:

$$\begin{aligned}
K\eta_{loc}\cdot\mathbb{E}\left[f(x_{t+1}) - f(x_t)\right] &\overset{①}{\le} \mathbb{E}\left[K\eta_{loc}\cdot\langle\nabla f(x_t), x_{t+1} - x_t\rangle\right] \\
&\quad + K\eta_{loc}\cdot\frac{L_0 + L_1\left\|\nabla f(x_t)\right\|}{2}\mathbb{E}\left[\left\|x_{t+1} - x_t\right\|^2\right] \\
&\overset{②}{=} \mathbb{E}\left[\frac{\eta_{glob}}{\alpha_\xi\left\|\Delta_t\right\|}\langle\alpha_\xi K\eta_{loc}\cdot\nabla f(x_t), \Delta_t\rangle\right] + K\eta_{loc}\cdot\eta_{glob}^2\cdot\frac{L_0 + L_1\left\|\nabla f(x_t)\right\|}{2} \\
&= \mathbb{E}\left[\frac{\eta_{glob}}{\alpha_\xi\left\|\Delta_t\right\|}\langle\alpha_\xi K\eta_{loc}\cdot\nabla f(x_t) \pm \Delta_t, \Delta_t\rangle\right]
\end{aligned}$$

$$
\begin{aligned}
&\quad + K\eta_{loc}\cdot\eta_{glob}^2\cdot\frac{L_0+L_1\left\|\nabla f(x_t)\right\|}{2}\\
&= -\mathbb{E}\left[\frac{\eta_{glob}}{\alpha_\xi}\left\|\Delta_t\right\|\right]+\mathbb{E}\left[\frac{\eta_{glob}}{\alpha_\xi\left\|\Delta_t\right\|}\left\langle\alpha_\xi K\eta_{loc}\cdot\nabla f(x_t)+\Delta_t,\Delta_t\right\rangle\right]\\
&\quad + K\eta_{loc}\cdot\eta_{glob}^2\cdot\frac{L_0+L_1\left\|\nabla f(x_t)\right\|}{2}\\
&\overset{③}{\leq} -\mathbb{E}\left[\frac{\eta_{glob}}{\alpha_\xi}\left\|\Delta_t\right\|\right]+\mathbb{E}\left[\frac{\eta_{glob}}{\alpha_\xi}\left\|\alpha_\xi K\eta_{loc}\cdot\nabla f(x_t)+\Delta_t\right\|\right]\\
&\quad + K\eta_{loc}\cdot\eta_{glob}^2\cdot\frac{L_0+L_1\left\|\nabla f(x_t)\right\|}{2}\\
&\overset{④}{\leq} -K\eta_{loc}\cdot\eta_{glob}\left\|\nabla f(x_t)\right\|+2\cdot\mathbb{E}\left[\frac{\eta_{glob}}{\alpha_\xi}\left\|\alpha_\xi K\eta_{loc}\cdot\nabla f(x_t)+\Delta_t\right\|\right]\\
&\quad + K\eta_{loc}\cdot\eta_{glob}^2\cdot\frac{L_0+L_1\left\|\nabla f(x_t)\right\|}{2}\\
&\overset{⑤}{=} -K\eta_{loc}\cdot\eta_{glob}\left\|\nabla f(x_t)\right\|\\
&\quad + 2\cdot\mathbb{E}\left[\frac{\eta_{glob}}{\alpha_\xi}\left\|\alpha_\xi K\eta_{loc}\cdot\nabla f(x_t)+\frac{1}{m}\sum_{i=1}^{m}\Delta_t^i\right\|\right]\\
&\quad + K\eta_{loc}\cdot\eta_{glob}^2\cdot\frac{L_0+L_1\left\|\nabla f(x_t)\right\|}{2}\\
&\overset{⑥}{=} -K\eta_{loc}\cdot\eta_{glob}\left\|\nabla f(x_t)\right\|\\
&\quad + 2\cdot\mathbb{E}\left[\frac{\eta_{glob}}{\alpha_\xi}\left\|\alpha_\xi K\eta_{loc}\cdot\nabla f(x_t)-\frac{\eta_{loc}}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\frac{g_{t,k}^i}{\left\|g_{t,k}^i\right\|}\right\|\right]\\
&\quad + K\eta_{loc}\cdot\eta_{glob}^2\cdot\frac{L_0+L_1\left\|\nabla f(x_t)\right\|}{2}\\
&= -K\eta_{loc}\cdot\eta_{glob}\left\|\nabla f(x_t)\right\|\\
&\quad + 2\cdot\mathbb{E}\left[\eta_{loc}\cdot\eta_{glob}\left\|K\cdot\nabla f(x_t)-\frac{1}{\alpha_\xi m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\frac{g_{t,k}^i}{\left\|g_{t,k}^i\right\|}\pm\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}g_{t,k}^i\right\|\right]\\
&\quad + K\eta_{loc}\cdot\eta_{glob}^2\cdot\frac{L_0+L_1\left\|\nabla f(x_t)\right\|}{2}\\
&\overset{⑦}{\leq} -K\eta_{loc}\cdot\eta_{glob}\left\|\nabla f(x_t)\right\|\\
&\quad + 2\cdot\mathbb{E}\left[\eta_{loc}\cdot\eta_{glob}\left\|K\cdot\nabla f(x_t)-\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}g_{t,k}^i\right\|\right]\\
&\quad + 2\cdot\mathbb{E}\left[\eta_{loc}\cdot\eta_{glob}\left\|\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}g_{t,k}^i-\frac{1}{\alpha_\xi m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\frac{g_{t,k}^i}{\left\|g_{t,k}^i\right\|}\right\|\right]\\
&\quad + K\eta_{loc}\cdot\eta_{glob}^2\cdot\frac{L_0+L_1\left\|\nabla f(x_t)\right\|}{2}\\
&\leq -K\eta_{loc}\cdot\eta_{glob}\left\|\nabla f(x_t)\right\|\\
&\quad + 2\cdot\mathbb{E}\left[\eta_{loc}\cdot\eta_{glob}\left\|K\cdot\nabla f(x_t)-\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}g_{t,k}^i\right\|\right]\\
&\quad + 2\cdot\frac{\eta_{loc}\cdot\eta_{glob}}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}\left[\left|\left\|g_{t,k}^i\right\|-\frac{1}{\alpha_\xi}\right|\right]
\end{aligned}
$$

$$
\begin{aligned}
&\quad + K\eta_{loc}\cdot\eta_{glob}^2\cdot\frac{L_0+L_1\left\|\nabla f(x_t)\right\|}{2}\\
&= -K\eta_{loc}\cdot\eta_{glob}\left\|\nabla f(x_t)\right\|\\
&\quad + 2\eta_{loc}\cdot\eta_{glob}\cdot\mathbb{E}\left[\left\|K\cdot\nabla f(x_t)-\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\left(g_{t,k}^i\pm\nabla F_i(x_{t,k}^i)\right)\right\|\right]\\
&\quad + 2\cdot\frac{\eta_{loc}\cdot\eta_{glob}}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}\left[\left|\left\|g_{t,k}^i\right\|-\frac{1}{\alpha_\xi}\right|\right]\\
&\quad + K\eta_{loc}\cdot\eta_{glob}^2\cdot\frac{L_0+L_1\left\|\nabla f(x_t)\right\|}{2}\\
&\overset{⑦}{\leq}\underbrace{-K\eta_{loc}\cdot\eta_{glob}\left\|\nabla f(x_t)\right\|}_{T_1}\\
&\quad\underbrace{+2\eta_{loc}\cdot\eta_{glob}\cdot\mathbb{E}\left[\left\|K\cdot\nabla f(x_t)-\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^i)\right\|\right]}_{T_2}\\
&\quad\underbrace{+2\eta_{loc}\cdot\eta_{glob}\cdot\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\left(\nabla F_i(x_{t,k}^i)-g_{t,k}^i\right)\right\|\right]}_{T_3}\\
&\quad\underbrace{+2\cdot\frac{\eta_{loc}\cdot\eta_{glob}}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}\left[\left|\left\|g_{t,k}^i\right\|-\frac{1}{\alpha_\xi}\right|\right]}_{T_4}\\
&\quad\underbrace{+K\eta_{loc}\cdot\eta_{glob}^2\cdot\frac{L_0+L_1\left\|\nabla f(x_t)\right\|}{2}}_{T_5},
\end{aligned}
\tag{3}
$$

where in ① we used descent lemma of (L_0, L_1)-smoothness: $\forall y, x \in \mathbb{R}^d$ with $\|y-x\| \leq 1/L_1$: $f(y) \leq f(x) + \langle \nabla f(x), y-x\rangle + \frac{L_0+L_1\|\nabla f(x)\|}{2}\|y-x\|^2$, in ② we used $x_{t+1} = x_t + \eta_{glob}\frac{\Delta_t}{\|\Delta_t\|}$, in ③ we used Cauchy–Schwarz inequality $|\langle a, b\rangle| \leq \|a\|\,\|b\|$, in ④ we used triangle inequality $-\|b\| \leq -\|a\| + \|a+b\|$, in ⑤ we used $\Delta_t = \frac{1}{m}\sum_{i=1}^m \Delta_t^i$ (m is the total worker number), in ⑥ we used $\Delta_t^i = -\eta_{loc}\sum_{k=0}^{K-1}\frac{g_{t,k}^i}{\|g_{t,k}^i\|}$, in ⑦ we used triangle inequality $\|a+b\| \leq \|a\| + \|b\|$.

The full-participation cancellation used in the estimate of T_2 is

$$
K\nabla f(x_t) - \frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_t) = K\left(\nabla f(x_t) - \frac{1}{m}\sum_{i=1}^{m}\nabla F_i(x_t)\right) = 0. \tag{4}
$$

Now let us start with the term T_2 from equation 3:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|K\cdot\nabla f(x_t)-\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^i)\right\|\right]\\
&= \mathbb{E}\left[\left\|\underbrace{K\cdot\nabla f(x_t)-\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_t)}_{①}+\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\left(\nabla F_i(x_t)-\nabla F_i(x_{t,k}^i)\right)\right\|\right]\\
&\overset{②}{=}\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\left(\nabla F_i(x_t)-\nabla F_i(x_{t,k}^i)\right)\right\|\right]
\end{aligned}
$$

$$
\begin{aligned}
&\overset{③}{\leq} \frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla F_i(x_t)-\nabla F_i(x_{t,k}^i)\right\|\right]\\
&\overset{④}{\leq} \frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}\left[\left(L_0+L_1\left\|\nabla F_i(x_t)\right\|\right)\left\|x_{t,k}^i-x_t\right\|\right]\\
&\overset{⑤}{\leq} \frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}\left[\left(L_0+L_1\left\|\nabla F_i(x_t)\right\|\right)\eta_{loc}k\right]\\
&\leq \frac{K\cdot\eta_{loc}}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}\left[\left(L_0+L_1\left\|\nabla F_i(x_t)\right\|\right)\right]\\
&= \frac{K\cdot\eta_{loc}}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}\left[\left(L_0+L_1\left\|\nabla F_i(x_t)\pm\nabla f(x_t)\right\|\right)\right]\\
&\overset{⑥}{\leq} \frac{K\cdot\eta_{loc}}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}\left[\left(L_0+L_1\left\|\nabla f(x_t)\right\|+L_1\sigma_{glob}\right)\right]\\
&\leq K^2\cdot\eta_{loc}\left(L_0+L_1\left\|\nabla f(x_t)\right\|+L_1\sigma_{glob}\right). \qquad (5)
\end{aligned}
$$

where in ① we isolated the full-participation heterogeneity component, in ② we used $\frac{1}{m}\sum_{i=1}^{m}\nabla F_i(x_t)=\nabla f(x_t)$, equivalently equation 4, which makes the bracketed term exactly zero, in ③ we used triangle inequality and linearity of expectation, in ④ we used (L_0,L_1)-smoothness: $\|\nabla F_i(y)-\nabla F_i(x)\|\leq\left(L_0+L_1\|\nabla F_i(x)\|\right)\|y-x\|$, in ⑤ we used $x_{t,k}^i-x_t=-\eta_{loc}\sum_{\ell=0}^{k-1}\frac{g_{t,\ell}^i}{\|g_{t,\ell}^i\|}$ and hence $\left\|x_{t,k}^i-x_t\right\|\leq\eta_{loc}k\leq K\eta_{loc}$, and in ⑥ we used $\|\nabla F_i(x_t)\|\leq\|\nabla f(x_t)\|+\|\nabla F_i(x_t)-\nabla f(x_t)\|\leq\|\nabla f(x_t)\|+\sigma_{glob}$.

Let us move on to term T_3 from equation 3. Denote $Z_{t,k}^i:=\nabla F_i(x_{t,k}^i)-g_{t,k}^i$. Then

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\left(\nabla F_i(x_{t,k}^i)-g_{t,k}^i\right)\right\|\right]\\
&\overset{①}{\leq}\left(\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}Z_{t,k}^i\right\|^2\right]\right)^{1/2}\\
&\overset{②}{=}\left(\frac{1}{m^2}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|Z_{t,k}^i\right\|^2\right]+\frac{1}{m^2}\sum_{(i,k)\neq(j,\ell)}\mathbb{E}\left[\left\langle Z_{t,k}^i,Z_{t,\ell}^j\right\rangle\right]\right)^{1/2}\\
&\overset{③}{=}\left(\frac{1}{m^2}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|Z_{t,k}^i\right\|^2\right]\right)^{1/2}\\
&\overset{④}{\leq}\left(\frac{mK\sigma_{loc}^2}{m^2}\right)^{1/2}=\sigma_{loc}\sqrt{\frac{K}{m}}. \qquad (6)
\end{aligned}
$$

where in ① we applied Jensen's inequality, in ② we expanded the squared norm, in ③ we used that, conditioned on the local histories, the variables $Z_{t,k}^i$ form martingale differences across local steps and are independent across workers, so all cross terms vanish by the tower property, and in ④ we used $\mathbb{E}\left[\left\|\nabla F_i(x_{t,k}^i)-g_{t,k}^i\right\|^2\right]\leq\sigma_{loc}^2$.

Now consider term T_4 from equation 3:

$$
\begin{aligned}
2\cdot\frac{\eta_{loc}\cdot\eta_{glob}}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}\left[\left|\left\|g_{t,k}^i\right\|-\frac{1}{\alpha_\xi}\right|\right] &\overset{①}{\leq} 2K\cdot\eta_{loc}\cdot\eta_{glob}\cdot\delta_t\\
&\overset{②}{\leq} 2K\cdot\eta_{loc}\cdot\eta_{glob}\cdot\delta
\end{aligned}
$$

$$= 2K \cdot \eta_{loc} \cdot \eta_{glob} \cdot \delta, \tag{7}$$

where in ① we used the assumption on δ_t with $\alpha_\xi = \frac{1}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\frac{1}{\left\|g_{t,k}^i\right\|}$, in ② we used $\delta := \sup_{t=0,\ldots,T-1}\delta_t$.

Substituting equation 5-equation 7 into equation 3 and dividing both sides by $K\eta_{loc}$, we obtain the following:

$$\begin{aligned}
\mathbb{E}\left[f(x_{t+1}) - f(x_t)\right] &\overset{equation\ 3}{\leq} -\eta_{glob}\left\|\nabla f(x_t)\right\| \\
&\quad + \frac{2\cdot\eta_{glob}}{K}\cdot\mathbb{E}\left[\left\|K\cdot\nabla f(x_t) - \frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^i)\right\|\right] \\
&\quad + \frac{2\cdot\eta_{glob}}{K}\cdot\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\left(\nabla F_i(x_{t,k}^i) - g_{t,k}^i\right)\right\|\right] \\
&\quad + 2\cdot\frac{\eta_{glob}}{m\cdot K}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}\left[\left|\left\|g_{t,k}^i\right\| - \frac{1}{\alpha_\xi}\right|\right] \\
&\quad + \eta_{glob}^2\cdot\frac{L_0 + L_1\left\|\nabla f(x_t)\right\|}{2} \\
&\overset{equation\ 5-equation\ 7}{\leq} -\eta_{glob}\left\|\nabla f(x_t)\right\| \\
&\quad + 2K\cdot\eta_{glob}\cdot\eta_{loc}\left(L_0 + L_1\left\|\nabla f(x_t)\right\| + L_1\sigma_{glob}\right) \\
&\quad + 2\cdot\eta_{glob}\cdot\frac{\sigma_{loc}}{\sqrt{mK}} \\
&\quad + 2\cdot\eta_{glob}\cdot\delta \\
&\quad + \eta_{glob}^2\cdot\frac{L_0 + L_1\left\|\nabla f(x_t)\right\|}{2} \\
&= -\eta_{glob}\left\|\nabla f(x_t)\right\| + 2\cdot\eta_{glob}\left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right) \\
&\quad + 2K\cdot\eta_{glob}\cdot\eta_{loc}\left(L_0 + L_1\sigma_{glob} + L_1\left\|\nabla f(x_t)\right\|\right) \\
&\quad + \eta_{glob}^2\cdot\frac{L_0 + L_1\left\|\nabla f(x_t)\right\|}{2}.
\end{aligned} \tag{8}$$

Let $\lambda > 0$ be a hyperparameter. Then, we consider two cases: $\|\nabla f(x_t)\| \geq \lambda$ and $\|\nabla f(x_t)\| < \lambda$.

Case 1: $\|\nabla f(x_t)\| \geq \lambda$ Let us introduce notations $F_t := f(x_t) - f^*$. Then from equation 8 we have:

$$\begin{aligned}
\mathbb{E}\left[F_{t+1}\right] - F_t &\overset{equation\ 8}{\leq} -\eta_{glob}\left\|\nabla f(x_t)\right\| + 2\cdot\eta_{glob}\left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right) \\
&\quad + 2K\cdot\eta_{glob}\cdot\eta_{loc}\left(L_0 + L_1\sigma_{glob} + L_1\left\|\nabla f(x_t)\right\|\right) \\
&\quad + \eta_{glob}^2\cdot\frac{L_0 + L_1\left\|\nabla f(x_t)\right\|}{2} \\
&= -\eta_{glob}\left\|\nabla f(x_t)\right\| + 2\cdot\eta_{glob}\left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right) \\
&\quad + \frac{2K\cdot\eta_{glob}\cdot\eta_{loc}\left(L_0 + L_1\sigma_{glob} + L_1\left\|\nabla f(x_t)\right\|\right)}{\left\|\nabla f(x_t)\right\|}\left\|\nabla f(x_t)\right\| \\
&\quad + \frac{\eta_{glob}^2\cdot\left(L_0 + L_1\left\|\nabla f(x_t)\right\|\right)}{2\left\|\nabla f(x_t)\right\|}\cdot\left\|\nabla f(x_t)\right\| \\
&= -\eta_{glob}\left\|\nabla f(x_t)\right\| + 2\cdot\eta_{glob}\left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right) \\
&\quad + 2K\cdot\eta_{glob}\cdot\eta_{loc}\left(\frac{L_0 + L_1\sigma_{glob}}{\left\|\nabla f(x_t)\right\|} + L_1\right)\cdot\left\|\nabla f(x_t)\right\|
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{\eta_{glob}^2 \cdot \left(\frac{L_0}{\|\nabla f(x_t)\|} + L_1\right)}{2} \cdot \|\nabla f(x_t)\| \\
&\overset{①}{\leq} -\eta_{glob} \|\nabla f(x_t)\| + 2 \cdot \eta_{glob} \left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right) \\
&\quad + 2K \cdot \eta_{glob} \cdot \eta_{loc} \left(\frac{L_0 + L_1\sigma_{glob}}{\lambda} + L_1\right) \cdot \|\nabla f(x_t)\| \\
&\quad + \frac{\eta_{glob}^2 \cdot \left(\frac{L_0}{\lambda} + L_1\right)}{2} \cdot \|\nabla f(x_t)\| \\
&= -\eta_{glob} \|\nabla f(x_t)\| + 2 \cdot \eta_{glob} \left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right) \\
&\quad + 2K \cdot \eta_{glob} \cdot \eta_{loc} \left(\frac{L_0 + L_1\sigma_{glob} + L_1\lambda}{\lambda}\right) \cdot \|\nabla f(x_t)\| \\
&\quad + \frac{\eta_{glob}^2 \cdot (L_0 + L_1\lambda)}{2\lambda} \cdot \|\nabla f(x_t)\| \\
&\overset{②}{\leq} -\eta_{glob} \|\nabla f(x_t)\| + 2 \cdot \eta_{glob} \left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right) \\
&\quad + \frac{\eta_{glob}}{4} \cdot \|\nabla f(x_t)\| \\
&\quad + \frac{\eta_{glob}^2 \cdot (L_0 + L_1\lambda)}{2\lambda} \cdot \|\nabla f(x_t)\| \\
&\overset{③}{\leq} -\eta_{glob} \|\nabla f(x_t)\| + 2 \cdot \eta_{glob} \left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right) \\
&\quad + \frac{\eta_{glob}}{4} \cdot \|\nabla f(x_t)\| \\
&\quad + \frac{\eta_{glob}}{4} \cdot \|\nabla f(x_t)\| \\
&= -\frac{\eta_{glob}}{2} \|\nabla f(x_t)\| + 2 \cdot \eta_{glob} \left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right), \qquad (9)
\end{aligned}
$$

where in ① we used $\|\nabla f(x_t)\| \geq \lambda$, in ② we chose step size $\eta_{loc} \leq \frac{\lambda}{8K \cdot (L_0 + L_1\sigma_{glob} + L_1\lambda)}$, in ③ we chose step size $\eta_{glob} \leq \frac{\lambda}{2(L_0 + L_1\lambda)}$.

Thus, we have obtained the inequality for the first case:

$$
\|\nabla f(x_t)\| \overset{equation\ 9}{\leq} \frac{2\left(F_t - \mathbb{E}\left[F_{t+1}\right]\right)}{\eta_{glob}} + 4\left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right). \qquad (10)
$$

Case 2: $\|\nabla f(x_t)\| < \lambda$ Let us introduce notations $F_t := f(x_t) - f^*$. Then from equation 8 we have:

$$
\begin{aligned}
\mathbb{E}\left[F_{t+1}\right] - F_t &\overset{equation\ 8}{\leq} -\eta_{glob} \|\nabla f(x_t)\| + 2 \cdot \eta_{glob} \left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right) \\
&\quad + 2K \cdot \eta_{glob} \cdot \eta_{loc} \left(L_0 + L_1\sigma_{glob} + L_1 \|\nabla f(x_t)\|\right) \\
&\quad + \eta_{glob}^2 \cdot \frac{L_0 + L_1 \|\nabla f(x_t)\|}{2} \\
&= -\eta_{glob} \|\nabla f(x_t)\| + 2 \cdot \eta_{glob} \left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right) \\
&\quad + \frac{2K \cdot \eta_{glob} \cdot \eta_{loc} \left(L_0 + L_1\sigma_{glob} + L_1 \|\nabla f(x_t)\|\right)}{\lambda} \cdot \lambda \\
&\quad + \frac{\eta_{glob}^2 \cdot \left(L_0 + L_1 \|\nabla f(x_t)\|\right)}{2\lambda} \cdot \lambda
\end{aligned}
$$

$$
\begin{aligned}
&\overset{①}{\leq} -\eta_{glob}\,\|\nabla f(x_t)\| + 2\cdot\eta_{glob}\left(\frac{\sigma_{loc}}{\sqrt{mK}}+\delta\right)\\
&\qquad + \frac{2K\cdot\eta_{glob}\cdot\eta_{loc}\,(L_0+L_1\sigma_{glob}+L_1\lambda)}{\lambda}\cdot\lambda\\
&\qquad + \frac{\eta_{glob}^2\cdot(L_0+L_1\lambda)}{2\lambda}\cdot\lambda\\
&\overset{②}{\leq} -\eta_{glob}\,\|\nabla f(x_t)\| + 2\cdot\eta_{glob}\left(\frac{\sigma_{loc}}{\sqrt{mK}}+\delta\right)\\
&\qquad + \frac{\eta_{glob}}{4}\cdot\lambda\\
&\qquad + \frac{\eta_{glob}^2\cdot(L_0+L_1\lambda)}{2\lambda}\cdot\lambda\\
&\overset{③}{\leq} -\eta_{glob}\,\|\nabla f(x_t)\| + 2\cdot\eta_{glob}\left(\frac{\sigma_{loc}}{\sqrt{mK}}+\delta\right)\\
&\qquad + \frac{\eta_{glob}}{4}\cdot\lambda\\
&\qquad + \frac{\eta_{glob}}{4}\cdot\lambda\\
&= -\eta_{glob}\,\|\nabla f(x_t)\| + 2\cdot\eta_{glob}\left(\frac{\sigma_{loc}}{\sqrt{mK}}+\delta\right) + \frac{\eta_{glob}}{2}\cdot\lambda\\
&\leq -\frac{\eta_{glob}}{2}\,\|\nabla f(x_t)\| + 2\cdot\eta_{glob}\left(\frac{\sigma_{loc}}{\sqrt{mK}}+\delta\right) + \frac{\eta_{glob}}{2}\cdot\lambda, \qquad (11)
\end{aligned}
$$

where in ① we used $\|\nabla f(x_t)\| < \lambda$, in ② we chose step size $\eta_{loc} \leq \frac{\lambda}{8K\cdot(L_0+L_1\sigma_{glob}+L_1\lambda)}$, in ③ we chose step size $\eta_{glob} \leq \frac{\lambda}{2(L_0+L_1\lambda)}$.

Thus, we have obtained the inequality for the second case:

$$
\|\nabla f(x_t)\| \overset{equation\ 11}{\leq} \frac{2\left(F_t - \mathbb{E}\left[F_{t+1}\right]\right)}{\eta_{glob}} + 4\left(\frac{\sigma_{loc}}{\sqrt{mK}}+\delta\right) + \lambda. \qquad (12)
$$

Final convergence rate. It is easy to see that the second-case bound equation 12 covers the first-case bound equation 10. Therefore, summing up over all indices $0 \leq t \leq T-1$, we get:

$$
\min_{t\in[0,T-1]} \mathbb{E}\left[\|\nabla f(x_t)\|\right] \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(x_t)\|\right] \leq \underbrace{\frac{2F_0}{\eta_{glob}T} + 4\left(\frac{\sigma_{loc}}{\sqrt{mK}}+\delta\right)}_{①} + \underbrace{\lambda}_{②}.
$$

To guarantee convergence to the desired accuracy $\min_{t\in[0,T-1]} \mathbb{E}\left[\|\nabla f(x_t)\|\right] \leq \varepsilon$, the two terms ① and ② should each be bounded above by $\varepsilon/2$. This is possible only if the residual term in ① is small enough, e.g.,

$$
4\left(\frac{\sigma_{loc}}{\sqrt{mK}}+\delta\right) \leq \frac{\varepsilon}{2},
$$

Using this, we obtain a constraint on λ:

$$
② : \quad \lambda \leq \frac{\varepsilon}{2}.
$$

Taking the maximum value $\lambda = \frac{\varepsilon}{2}$ and the maximum step sizes $\eta_{glob} = \frac{\lambda}{2(L_0+L_1\lambda)} = \left[\frac{4L_0}{\varepsilon} + 2L_1\right]^{-1}$, as well as $\eta_{loc} = \frac{\lambda}{8K\cdot(L_0+L_1\sigma_{glob}+L_1\lambda)} = \left[8K\left(\frac{2L_0+2L_1\sigma_{glob}}{\varepsilon}+L_1\right)\right]^{-1}$ we obtain the following inequality:

$$
① : \quad \frac{2F_0}{\eta_{glob}T} + 4\left(\frac{\sigma_{loc}}{\sqrt{mK}}+\delta\right) = \frac{8L_0F_0}{\varepsilon T} + \frac{4L_1F_0}{T} + 4\left(\frac{\sigma_{loc}}{\sqrt{mK}}+\delta\right).
$$

Thus, T should be chosen so that the last display is at most $\varepsilon/2$. □

C Proof of Theorem 3.3

Proof. The proof converts the stationarity estimate obtained in the non-convex analysis into a recursion for the function gap. Let $F_t := f(x_t) - f^*$ and fix a finite deterministic radius R_T such that $\|x_t - x^*\| \le R_T$ for all $t = 0, \dots, T-1$. We first record the consequence of convexity used below. We shall use the Cauchy–Schwarz inequality

$$|\langle a, b\rangle| \le \|a\| \, \|b\| \,. \tag{13}$$

Therefore, by Assumption 3.2, for every communication round $t \in \{0, 1, \dots, T-1\}$,

$$\begin{aligned} F_t &= f(x_t) - f^* \\ &\overset{equation\ 2}{\le} \langle \nabla f(x_t), x_t - x^* \rangle \\ &\overset{equation\ 13}{\le} \|\nabla f(x_t)\| \, \|x_t - x^*\| \\ &\le \|\nabla f(x_t)\| \, R_T. \end{aligned} \tag{14}$$

As in the nonconvex case, the second-case bound equation 12 dominates the first-case bound equation 10. Therefore, taking expectation with respect to all randomness in equation 12 yields

$$\mathbb{E}\left[\|\nabla f(x_t)\|\right] \le \frac{2\left(\mathbb{E}\left[F_t\right] - \mathbb{E}\left[F_{t+1}\right]\right)}{\eta_{glob}} + 4\left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right) + \lambda. \tag{15}$$

Combining equation 15 and equation 14, we obtain

$$\frac{\mathbb{E}\left[F_t\right]}{R_T} \le \frac{2\left(\mathbb{E}\left[F_t\right] - \mathbb{E}\left[F_{t+1}\right]\right)}{\eta_{glob}} + 4\left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right) + \lambda, \tag{16}$$

and hence

$$\mathbb{E}\left[F_{t+1}\right] \le \left(1 - \frac{\eta_{glob}}{2R_T}\right)\mathbb{E}\left[F_t\right] + \frac{\eta_{glob}}{2}\left(4\left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right) + \lambda\right). \tag{17}$$

Applying equation 17 recursively for $t = 0, \dots, T-1$ yields

Convex convergence rate.

$$\mathbb{E}\left[F_T\right] \le \underbrace{\left(1 - \frac{\eta_{glob}}{2R_T}\right)^T F_0 + 4R_T\left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right)}_{①} + \underbrace{R_T\lambda}_{②}.$$

To guarantee the target accuracy $\mathbb{E}\left[F_T\right] \le \varepsilon$, it suffices to bound each of the two terms ① and ② by $\varepsilon/2$. This requires the residual term in ① to be sufficiently small; for instance, it is sufficient that

$$4R_T\left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right) \le \frac{\varepsilon}{2}.$$

The corresponding constraint on λ is

$$② : \quad R_T\lambda \le \frac{\varepsilon}{2}, \qquad \lambda \le \frac{\varepsilon}{2R_T}.$$

Choosing the largest admissible value $\lambda = \frac{\varepsilon}{2R_T}$ together with the largest admissible global and local step sizes

$$\eta_{glob} = \frac{\lambda}{2\left(L_0 + L_1\lambda\right)} = \left[\frac{4R_TL_0}{\varepsilon} + 2L_1\right]^{-1},$$

and

$$\eta_{loc} = \frac{\lambda}{8K\left(L_0 + L_1\sigma_{glob} + L_1\lambda\right)} = \left[8K\left(\frac{2R_T\left(L_0 + L_1\sigma_{glob}\right)}{\varepsilon} + L_1\right)\right]^{-1},$$

respectively, gives the following bound:

$$① : \quad \left(1 - \frac{\eta_{glob}}{2R_T}\right)^T F_0 + 4R_T\left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right) \le \exp\left(-\frac{T}{2R_T\left(\frac{4R_TL_0}{\varepsilon} + 2L_1\right)}\right) F_0 + 4R_T\left(\frac{\sigma_{loc}}{\sqrt{mK}} + \delta\right).$$

□

D EXTENSION PROOFS

This section proves the extension results stated in Section 4. The mini-batch argument in Appendix D.1 shows how the local stochastic-noise term in Theorems 3.1 and 3.3 changes under local batching. Appendix D.2 adds uniform partial worker participation and completes the proof of Theorem 4.1. Finally, Appendix D.3 proves Theorem 4.3 by replacing the second-moment local-noise estimate with a p-th moment mini-batch estimate. Throughout, we reuse the descent inequalities from Appendices B and C and only track the terms affected by the extension under consideration.

D.1 EXTENSION TO MINI-BATCH STOCHASTIC GRADIENTS

Proof of the mini-batch part of Theorem 4.1. We now extend the nonconvex and convex bounds above to the mini-batch setting. Let $B \geq 1$ be the local batch size. In the mini-batch variant, each worker replaces the single-sample stochastic gradient in Algorithm 1 by

$$\mathbf{g}_{t,k}^i = \frac{1}{B}\sum_{b=1}^{B} \nabla F_i(x_{t,k}^i, \xi_{t,k,b}^i).$$

For fixed $x_{t,k}^i$, Assumption 2.3 and linearity of expectation give

$$\mathbb{E}\left[\mathbf{g}_{t,k}^i\right] = \frac{1}{B}\sum_{b=1}^{B} \mathbb{E}\left[\nabla F_i(x_{t,k}^i, \xi_{t,k,b}^i)\right] = \nabla F_i(x_{t,k}^i).$$

Assume that the individual stochastic gradients satisfy

$$\mathbb{E}\left[\left\|\nabla F_i(x,\xi) - \nabla F_i(x)\right\|^2\right] \leq \sigma_{loc}^2 \qquad \forall i \in [m],\ x \in \mathbb{R}^d,$$

and that samples inside each mini-batch are independent. For brevity, denote $Z_{t,k,b}^i := \nabla F_i(x_{t,k}^i, \xi_{t,k,b}^i) - \nabla F_i(x_{t,k}^i)$. Then

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{g}_{t,k}^i - \nabla F_i(x_{t,k}^i)\right\|^2\right] &\overset{①}{=} \mathbb{E}\left[\left\|\frac{1}{B}\sum_{b=1}^{B} Z_{t,k,b}^i\right\|^2\right] \\
&\overset{②}{=} \frac{1}{B^2}\sum_{b=1}^{B} \mathbb{E}\left[\left\|Z_{t,k,b}^i\right\|^2\right] \\
&\quad + \frac{1}{B^2}\sum_{b\neq b'} \mathbb{E}\left[\left\langle Z_{t,k,b}^i, Z_{t,k,b'}^i\right\rangle\right] \\
&\overset{③}{=} \frac{1}{B^2}\sum_{b=1}^{B} \mathbb{E}\left[\left\|Z_{t,k,b}^i\right\|^2\right] \\
&\overset{④}{\leq} \frac{\sigma_{loc}^2}{B},
\end{aligned}$$

where in ① we used the definition of $\mathbf{g}_{t,k}^i$, in ② we expanded the squared norm of the sum, in ③ we used independence of the mini-batch samples and $\mathbb{E}\left[Z_{t,k,b}^i\right] = 0$, and in ④ we used the bounded second-moment assumption and the fact that there are B diagonal terms. Therefore, the only term in the preceding proof that changes is the martingale estimate for T_3. Instead of equation 6, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\left(\nabla F_i(x_{t,k}^i) - \mathbf{g}_{t,k}^i\right)\right\|\right] \\
&\overset{①}{\leq} \left(\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\left(\nabla F_i(x_{t,k}^i) - \mathbf{g}_{t,k}^i\right)\right\|^2\right]\right)^{1/2} \\
&\overset{②}{\leq} \left(\frac{mK}{m^2}\cdot\frac{\sigma_{loc}^2}{B}\right)^{1/2} = \sigma_{loc}\sqrt{\frac{K}{mB}},
\end{aligned} \tag{18}$$

where in ① we used Jensen's inequality as in the estimate of T_3 in equation 6, and in ② we used the mini-batch second-moment bound above together with the same martingale-difference cancellation of cross terms across workers and local steps. Thus, both the nonconvex bound equation 12 and the convex recursion equation 17 remain valid after replacing $\sigma_{loc}/\sqrt{mK}$ by $\sigma_{loc}/\sqrt{mKB}$. Consequently, the mini-batch nonconvex rate is

$$\min_{t\in[0,T-1]} \mathbb{E}\left[\|\nabla f(x_t)\|\right] \le \frac{2F_0}{\eta_{glob}T} + 4\left(\frac{\sigma_{loc}}{\sqrt{mKB}} + \delta\right) + \lambda,$$

and, under Assumption 3.2, the mini-batch convex rate is

$$\mathbb{E}\left[F_T\right] \le \left(1 - \frac{\eta_{glob}}{2R_T}\right)^T F_0 + 4R_T\left(\frac{\sigma_{loc}}{\sqrt{mKB}} + \delta\right) + R_T\lambda.$$

The case $B = 1$ recovers the single-sample rates derived above. □

D.2 Extension to partial worker participation

Proof of the partial-participation part of Theorem 4.1. We now describe how the preceding bounds change when the server samples only n workers per communication round, as in Algorithm 1. In this case, stochastic noise is averaged over the n participating workers and K local steps, while worker heterogeneity appears through the sampling error of the selected client set. More precisely, partial participation changes exactly the component that vanishes in equation 4.

Assume that S_t is sampled uniformly without replacement from $[m]$, with $|S_t| = n$, and that

$$\frac{1}{m}\sum_{i=1}^{m}\|\nabla F_i(x) - \nabla f(x)\|^2 \le \sigma_{glob}^2 \qquad \forall x \in \mathbb{R}^d.$$

Let

$$\gamma_{m,n} := \sqrt{\frac{m-n}{n(m-1)}} \le \frac{1}{\sqrt{n}}.$$

The analogue of equation 4 under partial participation is no longer zero:

$$\begin{aligned} &K\nabla f(x_t) - \frac{1}{n}\sum_{i\in S_t}\sum_{k=0}^{K-1}\nabla F_i(x_t) \\ &\overset{①}{=} K\left(\nabla f(x_t) - \frac{1}{n}\sum_{i\in S_t}\nabla F_i(x_t)\right). \end{aligned} \tag{19}$$

where in ① we used that $\nabla F_i(x_t)$ does not depend on the local index k. Thus, the only new heterogeneity contribution is the deviation between the full worker average $\nabla f(x_t) = m^{-1}\sum_{i=1}^{m}\nabla F_i(x_t)$ and the sampled worker average $n^{-1}\sum_{i\in S_t}\nabla F_i(x_t)$. For any fixed x_t, the finite-population variance identity gives

$$\begin{aligned} \mathbb{E}_{S_t}\left[\left\|\frac{1}{n}\sum_{i\in S_t}\nabla F_i(x_t) - \nabla f(x_t)\right\|^2\right] &\overset{①}{=} \frac{m-n}{n(m-1)}\frac{1}{m}\sum_{i=1}^{m}\|\nabla F_i(x_t) - \nabla f(x_t)\|^2 \\ &\overset{②}{\le} \gamma_{m,n}^2\sigma_{glob}^2, \end{aligned}$$

where in ① we used the finite-population variance identity for sampling without replacement, and in ② we used the worker-heterogeneity assumption together with the definition of $\gamma_{m,n}$. Hence,

$$\mathbb{E}_{S_t}\left[\left\|\frac{1}{n}\sum_{i\in S_t}\nabla F_i(x_t) - \nabla f(x_t)\right\|\right] \overset{①}{\le} \gamma_{m,n}\sigma_{glob},$$

where in ① we applied Jensen's inequality to the previous second-moment bound. Therefore, in the estimate of T_2, the sampling component equation 19 is bounded as

$$\mathbb{E}_{S_t}\left[\left\|K\nabla f(x_t) - \frac{1}{n}\sum_{i\in S_t}\sum_{k=0}^{K-1}\nabla F_i(x_t)\right\|\right] \overset{equation\ 19}{=} K\mathbb{E}_{S_t}\left[\left\|\nabla f(x_t) - \frac{1}{n}\sum_{i\in S_t}\nabla F_i(x_t)\right\|\right]$$

$$\overset{②}{\leq} K\gamma_{m,n}\sigma_{glob}, \tag{20}$$

where in ② we used the Jensen bound above. The local-drift part of T_2 is treated exactly as in equation 5; in particular, the step-size restrictions still depend on σ_{glob} through the local smoothness estimate. The stochastic term T_3 is estimated as in equation 18, with m replaced by n because only the selected workers contribute to the server average:

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i\in S_t}\sum_{k=0}^{K-1}\left(\nabla F_i(x_{t,k}^i)-\mathbf{g}_{t,k}^i\right)\right\|\right] \leq \sigma_{loc}\sqrt{\frac{K}{nB}}.$$

Here we used the same Jensen step and martingale cross-term cancellation as in equation 18, together with the fact that the selected average has normalization factor $1/n$ instead of $1/m$. Thus, after division by K in the descent estimate, the local stochastic-noise contribution becomes $\sigma_{loc}/\sqrt{nKB}$, while the sampling component of T_2 contributes $\gamma_{m,n}\sigma_{glob}$. Combining the two estimates, and using $\gamma_{m,n} \leq 1/\sqrt{n}$, we obtain the following rate for the nonconvex case:

$$\min_{t\in[0,T-1]}\mathbb{E}\left[\|\nabla f(x_t)\|\right] \leq \frac{2F_0}{\eta_{glob}T} + 4\left(\frac{\sigma_{loc}}{\sqrt{nKB}}+\frac{\sigma_{glob}}{\sqrt{n}}+\delta\right)+\lambda,$$

and, under Assumption 3.2, the following rate for the convex case:

$$\mathbb{E}\left[F_T\right] \leq \left(1-\frac{\eta_{glob}}{2R_T}\right)^T F_0 + 4R_T\left(\frac{\sigma_{loc}}{\sqrt{nKB}}+\frac{\sigma_{glob}}{\sqrt{n}}+\delta\right)+R_T\lambda.$$

□

D.3 Extension to heavy-tailed stochastic gradients

We now combine the mini-batch and partial-participation extensions under the heavy-tailed model in Assumption 4.2. Compared with Appendix D.1, only the local-noise mini-batch estimate changes; it is controlled by the following martingale moment inequality.

Lemma D.1 (Kornilov et al. (2023); Hübler et al. (2025)). *Let $p \in [1,2]$, and let $X_1,\ldots,X_q \in \mathbb{R}^d$ be a martingale difference sequence, i.e., $\mathbb{E}\left[X_j \mid X_{j-1},\ldots,X_1\right]=0$ a.s. for all $j=1,\ldots,q$, with $\mathbb{E}\left[\|X_j\|^p\right]<\infty$. Define $S_q := \sum_{j=1}^q X_j$. Then*

$$\mathbb{E}\left[\|S_q\|^p\right] \leq 2\sum_{j=1}^q \mathbb{E}\left[\|X_j\|^p\right].$$

Historical background. Lemma D.1 goes back to von Bahr & Esseen (1965) for $d=1$ and was extended to Banach spaces by (Pisier, 1975, Proposition 2.4); optimal constants were derived by Cox (1982). Later proofs and higher-dimensional extensions appear in Cherapanamjeri et al. (2022); Kornilov et al. (2023), with corrections by Hübler et al. (2025).

Proof of Theorem 4.3. Let the mini-batch stochastic gradient be defined as in Section D.1. Applying Lemma D.1 for fixed $x_{t,k}^i$ with $X_b = \nabla F_i(x_{t,k}^i,\xi_{t,k,b}^i)-\nabla F_i(x_{t,k}^i)$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{g}_{t,k}^i-\nabla F_i(x_{t,k}^i)\right\|^p\right] &\overset{①}{=} \frac{1}{B^p}\mathbb{E}\left[\left\|\sum_{b=1}^B X_b\right\|^p\right] \\
&\overset{②}{\leq} \frac{2}{B^p}\sum_{b=1}^B \mathbb{E}\left[\left\|\nabla F_i(x_{t,k}^i,\xi_{t,k,b}^i)-\nabla F_i(x_{t,k}^i)\right\|^p\right] \\
&\overset{③}{\leq} \frac{2\sigma_{loc}^p}{B^{p-1}},
\end{aligned}$$

where in ① we used the definition of $\mathbf{g}_{t,k}^i$, in ② we applied Lemma D.1, and in ③ we used Assumption 4.2.

For the descent proof, we apply the same martingale moment inequality to the full noise sum over the selected workers, local steps, and mini-batch samples. For fixed S_t, define $X^i_{t,k,b} := \nabla F_i(x^i_{t,k}, \xi^i_{t,k,b}) - \nabla F_i(x^i_{t,k})$. Then

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i\in S_t}\sum_{k=0}^{K-1}\left(\nabla F_i(x^i_{t,k}) - \mathbf{g}^i_{t,k}\right)\right\|\right] \\
&\overset{①}{=} \mathbb{E}\left[\left\|\frac{1}{nB}\sum_{i\in S_t}\sum_{k=0}^{K-1}\sum_{b=1}^{B} X^i_{t,k,b}\right\|\right] \\
&\overset{②}{\le} \left(\frac{1}{(nB)^p}\mathbb{E}\left[\left\|\sum_{i\in S_t}\sum_{k=0}^{K-1}\sum_{b=1}^{B} X^i_{t,k,b}\right\|^p\right]\right)^{1/p} \\
&\overset{③}{\le} \left(\frac{2}{(nB)^p}\sum_{i\in S_t}\sum_{k=0}^{K-1}\sum_{b=1}^{B}\mathbb{E}\left[\left\|X^i_{t,k,b}\right\|^p\right]\right)^{1/p} \\
&\overset{④}{\le} \frac{2^{1/p}K^{1/p}\sigma_{loc}}{(nB)^{(p-1)/p}}.
\end{aligned}$$

where in ① we used the definition of $\mathbf{g}^i_{t,k}$, in ② we applied Jensen's inequality, in ③ we applied Lemma D.1 to the martingale difference sequence indexed by (i,k,b), and in ④ we used Assumption 4.2. After division by K in the descent estimate, this gives the local stochastic-noise contribution $\sigma_{loc}/(nKB)^{(p-1)/p}$ up to a universal constant. Combining this modification with the partial worker participation estimate from Section D.2, we obtain the following combined rate for the nonconvex case:

$$\min_{t\in[0,T-1]}\mathbb{E}\left[\|\nabla f(x_t)\|\right] \lesssim \frac{F_0}{\eta_{glob}T} + \frac{\sigma_{loc}}{(nKB)^{(p-1)/p}} + \frac{\sigma_{glob}}{\sqrt{n}} + \delta + \lambda,$$

and, under Assumption 3.2, the following combined rate for the convex case:

$$\mathbb{E}\left[F_T\right] \lesssim \left(1 - \frac{\eta_{glob}}{2R_T}\right)^T F_0 + R_T\left(\frac{\sigma_{loc}}{(nKB)^{(p-1)/p}} + \frac{\sigma_{glob}}{\sqrt{n}} + \delta + \lambda\right).$$

Thus, in the heavy-tailed setting, the local stochastic-noise contribution $\sigma_{loc}/\sqrt{nKB}$ from the light-tailed mini-batch analysis is replaced by $\sigma_{loc}/(nKB)^{(p-1)/p}$, while the partial-participation contribution remains $\sigma_{glob}/\sqrt{n}$. □

E Motivation for the Normalization-Bias Assumption

Exact normalization affects the descent direction in two distinct ways. It fixes the length of each local step, which is the intended stabilization effect, but it also changes the relative weights with which stochastic gradients of different magnitudes enter the averaged direction. The local-noise and worker-heterogeneity assumptions control vector errors before normalization; they do not control this scalar reweighting by inverse norms. Assumption 2.5 isolates this additional normalization effect. This section explains why the quantity is the right one to control, when it is small, why it is not implied by the other assumptions, and how it can be measured.

Common-scale approximation. Fix a communication round t in the full-participation setting and define the harmonic-mean norm scale

$$H_t := \frac{1}{\alpha_\xi} = \left(\frac{1}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\frac{1}{\left\|g^i_{t,k}\right\|}\right)^{-1}.$$

For each worker i and local step k, we have

$$g^i_{t,k} - H_t\frac{g^i_{t,k}}{\left\|g^i_{t,k}\right\|} = \left(\left\|g^i_{t,k}\right\| - H_t\right)\frac{g^i_{t,k}}{\left\|g^i_{t,k}\right\|}.$$

Averaging this identity and applying the triangle inequality gives

$$\left\| \frac{1}{mK} \sum_{i=1}^{m} \sum_{k=0}^{K-1} g_{t,k}^i - H_t \frac{1}{mK} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \frac{g_{t,k}^i}{\left\| g_{t,k}^i \right\|} \right\| \leq \frac{1}{mK} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \left| \left\| g_{t,k}^i \right\| - H_t \right|.$$

The right-hand side is exactly the empirical quantity whose expectation is bounded in Assumption 2.5. Hence, δ_t controls the deterministic error of approximating the average stochastic gradient by a rescaled average normalized direction.

Sufficient conditions and examples. The assumption is mild when gradient magnitudes within a communication round are close to a common scale. If all stochastic-gradient norms in a round coincide, then H_t equals this common norm and the common-scale error is zero. This can happen even when the directions are very different, so Assumption 2.5 does not impose angular alignment. More generally, if for some $s_t > 0$ and $\rho_t \in [0, 1)$,

$$\left\| g_{t,k}^i \right\| \in [s_t(1-\rho_t), s_t(1+\rho_t)] \qquad \text{for all } i \in [m],\ k = 0, \ldots, K-1,$$

then H_t belongs to the same interval and

$$\frac{1}{mK} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \left| \left\| g_{t,k}^i \right\| - H_t \right| \leq 2\rho_t s_t.$$

Thus, a small relative dispersion of gradient magnitudes is sufficient for a small normalization-bias term.

The same interpretation relates the bias to the local stepsize in a deterministic local-gradient regime. Suppose $g_{t,k}^i = \nabla F_i(x_{t,k}^i)$, $\|\nabla F_i(x_t)\| \in [s_t(1-\rho_t), s_t(1+\rho_t)]$ for all workers, and $K\eta_{loc}L_1 \leq 1$. Since normalized local steps satisfy $\left\| x_{t,k}^i - x_t \right\| \leq K\eta_{loc} \leq 1/L_1$, Assumption 2.2 gives

$$\left| \left\| \nabla F_i(x_{t,k}^i) \right\| - \|\nabla F_i(x_t)\| \right| \leq K\eta_{loc} \left(L_0 + L_1 s_t (1+\rho_t) \right).$$

Consequently, whenever the lower endpoint remains positive, the within-round norm dispersion is bounded by

$$2\rho_t s_t + 2K\eta_{loc} \left(L_0 + L_1 s_t (1+\rho_t) \right).$$

This simple bound is not used in the main theorem, which keeps δ explicit for stochastic gradients, but it illustrates the expected dependence on the local drift scale $K\eta_{loc}$ and the generalized-smoothness constants.

Why a separate assumption is needed. The term δ_t is not implied by the local stochastic-noise or worker-heterogeneity assumptions. Consider the deterministic single-worker case $m = 1$ with $K = 2$ and $f(x) = \frac{1}{2}x^2$ on the positive real line. If $x_t > \eta_{loc} > 0$, then the two normalized local steps give

$$g_{t,0} = x_t, \qquad x_{t,1} = x_t - \eta_{loc}, \qquad g_{t,1} = x_t - \eta_{loc}.$$

There is no stochastic noise and no worker heterogeneity, so $\sigma_{loc} = \sigma_{glob} = 0$. Nevertheless, the two gradient norms in the same communication round are different. Since the harmonic mean H_t lies between $x_t - \eta_{loc}$ and x_t, we have

$$\frac{1}{2} \left(|x_t - H_t| + |x_t - \eta_{loc} - H_t| \right) = \frac{\eta_{loc}}{2}.$$

Thus, the normalization-bias term captures a genuine effect of exact normalization that can appear even without stochastic noise or worker heterogeneity. The example also shows that the effect is tied to the within-round variation of gradient magnitudes along the local trajectory.

Empirical diagnostic. The normalization-bias proxy can be measured directly from the gradients computed during training. In the full-participation single-sample setting, define

$$\widehat{\delta}_t = \frac{1}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\left|\left\|g_{t,k}^i\right\| - \widehat{H}_t\right|, \qquad \widehat{H}_t = \left(\frac{1}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\frac{1}{\left\|g_{t,k}^i\right\|}\right)^{-1}.$$

For mini-batching and partial participation, the diagnostic becomes

$$\widehat{\delta}_t^{B,n} = \frac{1}{nK}\sum_{i\in S_t}\sum_{k=0}^{K-1}\left|\left\|\mathbf{g}_{t,k}^i\right\| - \widehat{H}_t^{B,n}\right|, \qquad \widehat{H}_t^{B,n} = \left(\frac{1}{nK}\sum_{i\in S_t}\sum_{k=0}^{K-1}\frac{1}{\left\|\mathbf{g}_{t,k}^i\right\|}\right)^{-1}.$$

This diagnostic requires no additional gradient-oracle calls; it only logs the norms of the gradients already computed by the clients. In Appendix A.6, we report the average of this quantity over the final 20% of logged communication rounds to verify that the normalization-bias proxy tracks the observed stationarity floor.

Table 4 reports this proxy on the real-data experiments and on a representative synthetic diagnostic. The values are averages over the last 20% of logged communication rounds. Since gradient norms depend on the model and loss scale, the absolute values should be compared within a task rather than across different datasets.

Table 4: Empirical normalization-bias proxy averaged over the last 20% of logged communication rounds. Values are mean $\pm$ one standard deviation over seeds.

Experiment	Method/setting	Seeds	Tail mean $\widehat{\delta}_t$
CIFAR-10 outlier stress test	Local normalization	3	23.25 ± 1.20
CIFAR-10 outlier stress test	Two-level normalization	3	25.06 ± 1.27
AG News clean real-data task	Server normalization, diagnostic on raw local gradients	3	2.64 ± 0.79
AG News clean real-data task	Two-level normalization	3	2.96 ± 1.05
Synthetic diagnostics	Delta sweep, scale std. 1.0	5	3.89 ± 0.60

The table should be interpreted as a diagnostic rather than as an additional tuning criterion. The CIFAR-10 outlier stress test has a substantially larger normalization-bias proxy than the clean AG News task, which is consistent with its construction: multiplicative scale outliers directly increase the dispersion of stochastic-gradient magnitudes within a round. In this regime, local normalization and two-level normalization are designed to remove unreliable magnitude information while preserving directions. By contrast, the clean AG News task has a smaller proxy and uses only one local step, so the additional benefit of local normalization is less pronounced. This matches the empirical observation that two-level normalization is most beneficial in the outlier stress test and remains competitive, but not uniformly superior, in the clean text experiment.

Overall, δ should be read as a measurable normalization-bias term: it quantifies the additional scalar error introduced by exact normalization, beyond the local stochastic noise and worker heterogeneity controlled by Assumptions 2.3 and 2.4. The role of Assumption 2.5 is not to claim that exact normalization is always unbiased, but to make its additional price explicit. When $\widehat{\delta}_t$ is small, the normalized direction behaves like a rescaled average stochastic gradient up to a small scalar mismatch. When it is large, the theory predicts a larger limiting neighborhood, and the method should be interpreted as trading magnitude information for directional robustness. This is why we report $\widehat{\delta}_t$ in the experimental appendix.

F Pre-Hitting Fejér Stability under Uniform Directional Control

The convex guarantee in Theorem 3.3 uses a deterministic finite-horizon bound R_T on the distance to a minimizer. This section gives a complementary pathwise result showing that a trajectory-dependent radius is not intrinsic to the convex stability argument once sufficiently accurate uniform directional control is available. Under explicit conditions on the local drift, stochastic error, and normalization dispersion, the iterates are Fejér monotone until they first reach a prescribed target sublevel set. The analysis can then use any deterministic bound $R \geq \|x_0 - x^*\|$ instead of R_T.

The result is deliberately stated as a hitting-time guarantee. A fixed nonzero normalized step can leave the target sublevel set after first entering it, so the argument does not provide a last-iterate guarantee. It also gives an analytical sufficient condition rather than an online certificate: one of the directional-error terms below depends on exact local gradients that are not observed by the algorithm.

Throughout this section, we consider the full-participation, single-sample setting of Theorem 3.3. We work on sample paths on which the local normalization factors in Algorithm 1 are well defined. Write

$$F_t := f(x_t) - f^*, \qquad g_t := \nabla f(x_t).$$

F.1 Pathwise directional-error decomposition

For a communication round t, define the harmonic-mean gradient-norm scale

$$H_t := \left(\frac{1}{mK} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \frac{1}{\left\| g_{t,k}^i \right\|} \right)^{-1}$$

and the scaled average normalized direction

$$v_t := H_t \frac{1}{mK} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \frac{g_{t,k}^i}{\left\| g_{t,k}^i \right\|}.$$

Whenever $v_t \neq 0$, the server update in Algorithm 1 can equivalently be written as

$$x_{t+1} = x_t - \eta_{glob} \frac{v_t}{\|v_t\|}. \tag{21}$$

Thus, v_t is an inexact gradient whose normalized direction is exactly the direction used by the server. Define the empirical within-round normalization dispersion

$$\widehat{\delta}_t := \frac{1}{mK} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \left| \left\| g_{t,k}^i \right\| - H_t \right|$$

and the averaged stochastic-gradient error

$$N_t := \left\| \frac{1}{mK} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \left(g_{t,k}^i - \nabla F_i(x_{t,k}^i) \right) \right\|.$$

Lemma F.1 (Directional-error decomposition)**.** *Suppose Assumptions 2.2 and 2.4 hold and $K\eta_{loc}L_1 \leq 1$. Then, for every communication round t,*

$$\|v_t - g_t\| \leq K\eta_{loc} \left(L_0 + L_1 \sigma_{glob} + L_1 \|g_t\| \right) + N_t + \widehat{\delta}_t.$$

Proof. Let

$$\bar{g}_t := \frac{1}{mK} \sum_{i=1}^{m} \sum_{k=0}^{K-1} g_{t,k}^i.$$

The common-scale calculation in Appendix E gives, pathwise,

$$\|\bar{g}_t - v_t\| \leq \widehat{\delta}_t.$$

Moreover, using $m^{-1} \sum_{i=1}^{m} \nabla F_i(x_t) = \nabla f(x_t)$,

$$\|\bar{g}_t - g_t\| \leq \left\| \frac{1}{mK} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \left(\nabla F_i(x_{t,k}^i) - \nabla F_i(x_t) \right) \right\| + N_t.$$

Every local step has length η_{loc}, and hence $\left\| x_{t,k}^i - x_t \right\| \leq k\eta_{loc} \leq K\eta_{loc} \leq \frac{1}{L_1}$. Therefore,

$$\begin{aligned} \left\| \nabla F_i(x_{t,k}^i) - \nabla F_i(x_t) \right\| &\leq K\eta_{loc} \left(L_0 + L_1 \|\nabla F_i(x_t)\| \right) \\ &\leq K\eta_{loc} \left(L_0 + L_1 \sigma_{glob} + L_1 \|g_t\| \right), \end{aligned}$$

where the first inequality uses Assumption 2.2 and the second uses Assumption 2.4. Averaging over i and k and applying the triangle inequality proves the claim. □

For a deterministic horizon $T \geq 1$ and thresholds $\nu, \delta_{\mathrm{unif}} \geq 0$, define the uniform directional-control event

$$\mathcal{E}_T(\nu, \delta_{\mathrm{unif}}) := \left\{ \max_{0 \leq t < T} N_t \leq \nu, \quad \max_{0 \leq t < T} \widehat{\delta}_t \leq \delta_{\mathrm{unif}} \right\}. \tag{22}$$

The next theorem is deterministic on this event. A conservative probability bound for the event under the assumptions of the main paper is given later in this section.

F.2 Pre-hitting Fejér stability

We first record two elementary geometric facts used in the stability proof.

Lemma F.2 (Alignment of an inexact normalized direction)**.** *Let f be convex and differentiable, let $x^* \in \arg\min_x f(x)$, and set*

$$F := f(x) - f^* > 0, \qquad g := \nabla f(x).$$

Let $v \neq 0$, $e := \|v - g\|$, and suppose that, for some $\theta \in [0, 1)$,

$$e \|x - x^*\| \leq \theta F.$$

Then

$$\left\langle \frac{v}{\|v\|}, x - x^* \right\rangle \geq \frac{1-\theta}{1+\theta} \frac{F}{\|g\|}, \qquad \left\langle g, \frac{v}{\|v\|} \right\rangle \geq \frac{1-\theta}{1+\theta} \|g\|.$$

Proof. Convexity gives $F \leq \langle g, x - x^* \rangle$, and hence

$$\langle v, x - x^* \rangle \geq F - e \|x - x^*\| \geq (1 - \theta) F.$$

It also gives $F \leq \|g\| \|x - x^*\|$. Since $F > 0$, we have $\|x - x^*\| > 0$ and $\|g\| > 0$, and the assumed error bound implies $e \leq \theta \|g\|$. Consequently,

$$\|v\| \leq \|g\| + e \leq (1 + \theta) \|g\|,$$

which proves the first inequality. Similarly,

$$\langle g, v \rangle \geq \|g\|^2 - \|g\| e \geq (1 - \theta) \|g\|^2.$$

Dividing by $\|v\| \leq (1 + \theta) \|g\|$ proves the second inequality. □

Lemma F.3 (Generalized-smooth self-bounding inequality)**.** *Under Assumption 2.2, for every x,*

$$f(x) - f^* \geq \frac{\|\nabla f(x)\|^2}{2 (L_0 + L_1 \|\nabla f(x)\|)},$$

whenever the displayed denominator is positive.

Proof. Set $g := \nabla f(x)$ and

$$y := x - \frac{g}{L_0 + L_1 \|g\|}.$$

If $L_1 > 0$, then

$$\|y - x\| = \frac{\|g\|}{L_0 + L_1 \|g\|} \leq \frac{1}{L_1};$$

when $L_1 = 0$, the distance restriction is vacuous. Hence the descent inequality is applicable to the pair (x, y). The descent inequality associated with Assumption 2.2 gives

$$\begin{aligned} f(y) &\leq f(x) - \frac{\|g\|^2}{L_0 + L_1 \|g\|} + \frac{L_0 + L_1 \|g\|}{2} \frac{\|g\|^2}{(L_0 + L_1 \|g\|)^2} \\ &= f(x) - \frac{\|g\|^2}{2 (L_0 + L_1 \|g\|)}. \end{aligned}$$

Since $f^* \leq f(y)$, the result follows. □

For a target accuracy $\varepsilon > 0$, define the first hitting time

$$\tau_\varepsilon := \inf\{t \geq 0 : F_t \leq \varepsilon\}.$$

Theorem F.4 (Pre-hitting Fejér stability on a uniform-control event)**.** *Suppose Assumptions 2.1–2.4 and Assumption 3.2 hold. Fix $x^* \in \arg\min_x f(x)$ and a deterministic $R \geq \|x_0 - x^*\| > 0$. Let $0 < \varepsilon < F_0$ and define*

$$q_\varepsilon := \frac{\varepsilon}{R}.$$

For fixed thresholds $\nu, \delta_{\mathrm{unif}} \geq 0$, set

$$A := K\eta_{loc}(L_0 + L_1\sigma_{glob}) + \nu + \delta_{\mathrm{unif}}, \qquad B := K\eta_{loc}L_1.$$

Suppose

$$K\eta_{loc}L_1 \leq 1, \qquad A \leq \frac{\varepsilon}{6R}, \qquad B \leq \frac{q_\varepsilon}{12R(L_0 + L_1 q_\varepsilon)}. \tag{23}$$

and choose

$$0 < \eta_{glob} \leq \frac{q_\varepsilon}{2(L_0 + L_1 q_\varepsilon)}. \tag{24}$$

Define the deterministic analytical horizon

$$T_\varepsilon := \left\lceil \frac{4R}{\eta_{glob}} \log \frac{F_0}{\varepsilon} \right\rceil. \tag{25}$$

Then, on the event $\mathcal{E}_{T_\varepsilon}(\nu, \delta_{\mathrm{unif}})$, for every $0 \leq t < \min\{\tau_\varepsilon, T_\varepsilon\}$,

$$\|x_{t+1} - x^*\| \leq \|x_t - x^*\| \leq R, \tag{26}$$

and

$$F_{t+1} \leq \left(1 - \frac{\eta_{glob}}{4R}\right) F_t. \tag{27}$$

Consequently,

$$\tau_\varepsilon \leq T_\varepsilon, \qquad \textit{equivalently} \qquad \min_{0 \leq t \leq T_\varepsilon} F_t \leq \varepsilon.$$

Proof. We work on $\mathcal{E}_{T_\varepsilon}(\nu, \delta_{\mathrm{unif}})$ and argue by induction. We first verify that the contraction factor in equation 27 lies strictly between zero and one.

If $L_1 > 0$ and $R \geq 1/L_1$, then

$$\eta_{glob} \leq \frac{q_\varepsilon}{2(L_0 + L_1 q_\varepsilon)} \leq \frac{1}{2L_1} \leq \frac{R}{2}.$$

Otherwise, either $L_1 = 0$ or $R < 1/L_1$. In this case, $\|x_0 - x^*\| \leq R$ places x_0 and x^* within the admissible neighborhood of Assumption 2.2. Since $\nabla f(x^*) = 0$, the generalized descent inequality applied at x^* gives

$$F_0 \leq \frac{L_0}{2}\|x_0 - x^*\|^2 \leq \frac{L_0 R^2}{2}.$$

Because $0 < \varepsilon < F_0$, we have $L_0 > 0$, and hence

$$\eta_{glob} \leq \frac{q_\varepsilon}{2(L_0 + L_1 q_\varepsilon)} \leq \frac{\varepsilon}{2RL_0} < \frac{F_0}{2RL_0} \leq \frac{R}{4}.$$

Thus, in all cases, $0 < \eta_{glob} < R$, and the contraction factor in equation 27 lies strictly between zero and one.

Suppose that $t < \min\{\tau_\varepsilon, T_\varepsilon\}$ and $\|x_t - x^*\| \leq R$. Since $F_t > \varepsilon$, convexity gives

$$F_t \leq \|g_t\|\,\|x_t - x^*\| \leq R\,\|g_t\|,$$

so

$$\|g_t\| > q_\varepsilon. \tag{28}$$

By Lemma F.1 and the uniform-control event,

$$\|v_t - g_t\| \leq A + B\,\|g_t\|.$$

The first condition in equation 23 gives $AR \le \varepsilon/6 < F_t/6$. For the second term, Lemma F.3 and equation 28 give

$$\frac{BR\,\|g_t\|}{F_t} \le 2BR\left(\frac{L_0}{\|g_t\|} + L_1\right) \le 2BR\frac{L_0 + L_1 q_\varepsilon}{q_\varepsilon} \le \frac{1}{6}.$$

Consequently,

$$\|v_t - g_t\|\,R \le \frac{F_t}{3}, \qquad \|v_t - g_t\|\,\|x_t - x^*\| \le \frac{F_t}{3}. \tag{29}$$

Moreover, $F_t \le R\,\|g_t\|$ and the first inequality in equation 29 imply $\|v_t - g_t\| \le \|g_t\|/3$. Hence $\|v_t\| \ge 2\,\|g_t\|/3 > 0$, so the server direction in equation 21 is well defined before the hitting time.

Applying Lemma F.2 with $\theta = 1/3$ yields

$$\left\langle \frac{v_t}{\|v_t\|}, x_t - x^* \right\rangle \ge \frac{1}{2}\frac{F_t}{\|g_t\|}, \qquad \left\langle g_t, \frac{v_t}{\|v_t\|} \right\rangle \ge \frac{1}{2}\,\|g_t\|. \tag{30}$$

Using the first bound in equation 30,

$$\|x_{t+1} - x^*\|^2 = \|x_t - x^*\|^2 - 2\eta_{glob}\left\langle \frac{v_t}{\|v_t\|}, x_t - x^* \right\rangle + \eta_{glob}^2 \le \|x_t - x^*\|^2 - \eta_{glob}\frac{F_t}{\|g_t\|} + \eta_{glob}^2.$$

Lemma F.3, equation 28, and monotonicity of $z \mapsto z/(L_0 + L_1 z)$ give

$$\frac{F_t}{\|g_t\|} \ge \frac{\|g_t\|}{2(L_0 + L_1\,\|g_t\|)} \ge \frac{q_\varepsilon}{2(L_0 + L_1 q_\varepsilon)} \ge \eta_{glob}.$$

Thus $\|x_{t+1} - x^*\| \le \|x_t - x^*\| \le R$, which closes the induction and proves equation 26. Moreover, when $L_1 > 0$, condition equation 24 gives

$$\|x_{t+1} - x_t\| = \eta_{glob} \le \frac{1}{2L_1} \le \frac{1}{L_1}.$$

Thus, the generalized descent inequality is applicable to (x_t, x_{t+1}); for $L_1 = 0$, the distance restriction is vacuous. For the function gap, the generalized descent inequality and the second bound in equation 30 imply

$$F_{t+1} \le F_t - \eta_{glob}\left\langle g_t, \frac{v_t}{\|v_t\|} \right\rangle + \frac{\eta_{glob}^2}{2}(L_0 + L_1\,\|g_t\|) \le F_t - \frac{\eta_{glob}}{2}\,\|g_t\| + \frac{\eta_{glob}^2}{2}(L_0 + L_1\,\|g_t\|).$$

The same monotonicity argument and equation 24 give

$$\eta_{glob} \le \frac{\|g_t\|}{2(L_0 + L_1\,\|g_t\|)}.$$

Therefore,

$$F_{t+1} \le F_t - \frac{\eta_{glob}}{4}\,\|g_t\| \le \left(1 - \frac{\eta_{glob}}{4R}\right)F_t,$$

where the last inequality again uses $F_t \le R\,\|g_t\|$. This proves equation 27.

If $\tau_\varepsilon > T_\varepsilon$, iterating equation 27 for $t = 0, \ldots, T_\varepsilon - 1$ would give

$$F_{T_\varepsilon} \le \exp\left(-\frac{\eta_{glob} T_\varepsilon}{4R}\right)F_0 \le \varepsilon,$$

contradicting the definition of τ_ε. Hence $\tau_\varepsilon \le T_\varepsilon$. □

Corollary F.5 (Two-phase communication-round complexity)**.** *Under the conditions of Theorem F.4, choose*

$$\eta_{glob} = \frac{q_\varepsilon}{2(L_0 + L_1 q_\varepsilon)}.$$

Then, on $\mathcal{E}_{T_\varepsilon}(\nu, \delta_{\mathrm{unif}})$, the method visits the target sublevel set within

$$T_\varepsilon = \left\lceil 8 \left(\frac{L_0 R^2}{\varepsilon} + L_1 R \right) \log \frac{F_0}{\varepsilon} \right\rceil$$

communication rounds. This is a communication-round hitting-time bound, subject to the explicit directional-error feasibility conditions in equation 23; it is not an end-to-end stochastic-gradient complexity bound.

A convenient sufficient choice for the local stepsize is obtained by requiring

$$\nu + \delta_{\mathrm{unif}} \le \frac{\varepsilon}{12R}$$

and

$$\eta_{loc} \le \min \left\{ \frac{\varepsilon}{12KR(L_0 + L_1 \sigma_{glob})}, \frac{q_\varepsilon}{12KL_1 R(L_0 + L_1 q_\varepsilon)}, \frac{1}{KL_1} \right\}, \tag{31}$$

where the second and third restrictions are omitted when $L_1 = 0$.

F.3 A CONSERVATIVE HIGH-PROBABILITY REALIZATION

The event-wise theorem isolates the deterministic optimization argument from the probability model. For completeness, the assumptions of the main paper imply the following conservative realization. Its dependence on the horizon reflects the use of only second-moment and expectation bounds.

Lemma F.6 (Uniform control from the main-paper assumptions)**.** *Suppose Assumptions 2.3 and 2.5 hold and the stochastic gradients are fresh and conditionally independent as in Theorem 3.3. Fix a deterministic horizon $T \ge 1$ and $\rho \in (0, 1)$. Define*

$$\nu_{T,\rho} := \sigma_{loc} \sqrt{\frac{2T}{mK\rho}}, \qquad \delta_{T,\rho}^{\mathrm{unif}} := \frac{2T\delta}{\rho},$$

where $\delta = \sup_{0 \le t < T} \delta_t$. Then

$$\mathbb{P}\left(\mathcal{E}_T(\nu_{T,\rho}, \delta_{T,\rho}^{\mathrm{unif}}) \right) \ge 1 - \rho.$$

Proof. Fix a round t and enumerate the mK worker-step pairs in an order that respects the local-step order within every worker. Let $\mathcal{G}_{t,r-1}$ contain the history before the rth gradient in this enumeration is sampled, and write

$$Z_{t,r} := g_{t,k}^i - \nabla F_i(x_{t,k}^i)$$

for its centered stochastic error. The local iterate $x_{t,k}^i$ is $\mathcal{G}_{t,r-1}$-measurable, and freshness of the sample gives

$$\mathbb{E}[Z_{t,r} \mid \mathcal{G}_{t,r-1}] = 0, \qquad \mathbb{E}[\|Z_{t,r}\|^2 \mid \mathcal{G}_{t,r-1}] \le \sigma_{loc}^2.$$

The cross terms therefore vanish by the tower property, and

$$\mathbb{E}[N_t^2] \le \frac{\sigma_{loc}^2}{mK}.$$

Markov's inequality and a union bound over rounds yield

$$\mathbb{P}\left(\max_{0 \le t < T} N_t > \nu_{T,\rho} \right) \le \sum_{t=0}^{T-1} \frac{\mathbb{E}[N_t^2]}{\nu_{T,\rho}^2} \le \frac{\rho}{2}.$$

Similarly, Assumption 2.5 and the nonnegativity of $\widehat{\delta}_t$ give

$$\mathbb{P}\left(\max_{0 \le t < T} \widehat{\delta}_t > \delta_{T,\rho}^{\mathrm{unif}} \right) \le \sum_{t=0}^{T-1} \frac{\mathbb{E}[\widehat{\delta}_t]}{\delta_{T,\rho}^{\mathrm{unif}}} \le \frac{\rho}{2}.$$

When $\sigma_{loc} = 0$ or $\delta = 0$, the corresponding bound holds almost surely and the associated Markov step is omitted. A final union bound proves the result. □

Corollary F.7 (Conservative high-probability hitting-time guarantee)**.** *Let the conditions of Theorem F.4 hold, set $T = T_\varepsilon$, and use the thresholds from Lemma F.6. If these thresholds satisfy equation 23, then*

$$\mathbb{P}(\tau_\varepsilon \le T_\varepsilon) \ge 1 - \rho.$$

A fully explicit sufficient realization is

$$mK \ge \frac{1152\,\sigma_{loc}^2 R^2 T_\varepsilon}{\rho\varepsilon^2}, \qquad \delta \le \frac{\rho\varepsilon}{48RT_\varepsilon},$$

together with the local-stepsize condition equation 31 evaluated at $\nu = \nu_{T_\varepsilon,\rho}$ and $\delta_{\mathrm{unif}} = \delta^{\mathrm{unif}}_{T_\varepsilon,\rho}$. The displayed scaling is intentionally conservative: it makes explicit the price of obtaining uniform control from only the second-moment and expectation assumptions of the main paper.

Remark F.8 (Stronger tail assumptions)**.** *Sharper, logarithmic-in-T/ρ realizations of the stochastic-error threshold are possible under an appropriate conditional vector-valued sub-Gaussian or bounded-increment assumption. Such an assumption on the additive stochastic error does not by itself control $\widehat{\delta}_t$, because this quantity contains the harmonic mean of stochastic-gradient norms and is sensitive to behavior near zero. A sharper realization therefore also requires a corresponding tail condition for the within-round norm-dispersion term. We do not impose these stronger assumptions in the main results.*

Remark F.9 (Scope of the guarantee)**.** *Theorem F.4 proves that the method visits the target sublevel set and that the iterates are Fejér monotone only before this first visit. It does not guarantee $F_{T_\varepsilon} \le \varepsilon$ for an algorithm that continues taking fixed normalized steps, nor does it provide an online stopping certificate. In particular, $\widehat{\delta}_t$ is observable from the gradients already computed during a round, but N_t depends on the unknown exact local gradients $\nabla F_i(x^i_{t,k})$. A last-iterate or implementable stopping guarantee would require additional structure, a stopping rule, or a vanishing server stepsize.*