ADAPTIVE FEDERATED OPTIMIZATION ACROSS SMOOTH AND NONSMOOTH REGIMES

Anonymous authors
Paper under double-blind review

ABSTRACT

AdaGrad-style methods can simultaneously adapt to both the order and constants of smoothness in centralized optimization. However, existing extensions to federated optimization with heterogeneous client objectives and multiple local steps require the smoothness order to be specified in advance. We address this limitation with AdaFed, a federated method that adapts across several regimes under a single choice of parameters. In the deterministic setting, AdaFed achieves an $O(1/T)$ convergence rate for both smooth nonconvex and smooth convex objectives, and an $\widetilde{O}(1/\sqrt{T})$ rate for Lipschitz nonsmooth convex objectives. Under a light-tailed stochastic-gradient noise assumption, these guarantees extend to the stochastic setting with additional logarithmic factors and noise-dependent terms. Experiments across the three problem classes further support our theoretical findings.

1 INTRODUCTION

Federated learning (FL) aims to train a shared model from data distributed across many clients without requiring the clients to transmit their raw data. The resulting optimization problem is often formulated as

$$\min_{x\in\mathbb{R}^d} F(x) := \frac{1}{m}\sum_{i=1}^{m} F_i(x). \tag{1}$$

Here $F_i : \mathbb{R}^d \to \mathbb{R}$ is the local loss associated with the i-th client, which in stochastic settings can be expressed as:

$$F_i(x) = \mathbb{E}_{\xi\sim\mathcal{D}_i}[\ell_i(x;\xi)],$$

and the data distributions $\mathcal{D}_i$ can differ across clients. To reduce communication cost when m is large, classic algorithms such as FedAvg (McMahan et al., 2017; Li et al., 2020) allow each participating client to perform multiple local stochastic-gradient steps before its update is aggregated at the server. When the difference among the data distributions $\{\mathcal{D}_i\}_{i=1}^m$ is significant, methods such as SCAFFOLD (Karimireddy et al., 2020) and FedVARP (Jhunjhunwala et al., 2022) address the local drift arising from multiple local steps through control variates or server-side variance reduction. However, both theoretical guarantees and practical behavior of these methods depend on step-size schedules whose effective choices vary with the problem class and its related parameters (e.g., smoothness, gradient noise, heterogeneity, etc.).

In the centralized setting, adaptive gradient methods offer a natural way to overcome this reliance on prior problem knowledge. The pioneering method AdaGrad (Duchi et al., 2011) adjusts its step sizes adaptively using accumulated gradient information. Beyond empirical success, AdaGrad-type adaptive methods (Levy, 2017; Levy et al., 2018; Kavis et al., 2019; Ene et al., 2021) also exhibit an important theoretical advantage: a single update rule can achieve *universal* (in the sense of Nesterov (2015)) convergence by attaining the rates associated with smooth and nonsmooth objectives without knowing in advance the problem class or smoothness parameter.

However, transferring this universal convergence property to FL settings is highly nontrivial, where algorithms couple client-side local updates with server-side aggregation. Server-level adaptation alone is insufficient to guarantee universal convergence because the main optimization steps occur locally on clients, and their resulting updates directly shape server aggregation. Client-level adaptation can also fail because individual client gradients need not vanish even at a stationary point of F. Hence naively accumulating the client gradients during local updates may cause the step size to continuously

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.

decay. Several adaptive FL methods have been proposed in the literature. Specifically, server-level methods include FedOpt (Reddi et al., 2021) and FedAMS/FedCAMS (Wang et al., 2022), client-level methods include Δ-SGD (Kim et al., 2024) and FedSPS (Mukherjee et al., 2023), and methods operating at both levels include Mime (Karimireddy et al., 2021), FAFED (Wu et al., 2023), and FedAda^2 (Lee et al., 2026). Despite these efforts, the convergence results established in these works are tailored to specific problem classes, and some guarantees still require knowledge of smoothness parameters. Consequently, none of these works addresses the question central to this work:

Can a single adaptive FL method, without prior knowledge of the problem class or related parameters, automatically attain the convergence rates associated with smooth nonconvex, smooth convex, and Lipschitz convex objectives?

1.1 Our contributions

We answer this question affirmatively with a new AdaGrad-type FL method. Our contributions are summarized as follows.

- **Algorithm.** We develop a new adaptive FL method for solving (1) that accommodates heterogeneous client objectives, stochastic local gradients, and multiple local updates per communication round. Its key feature is a scaled AdaGrad-type step size that is shared across clients and fixed throughout each round, and is updated after server aggregation based on the server displacement.
- **Universal convergence.** The same algorithm automatically attains appropriate convergence rates for smooth nonconvex, smooth convex, and Lipschitz convex objectives, without prior knowledge of the problem class or associated parameters. These rates match those of classical AdaGrad in centralized settings up to logarithmic factors, and benefit jointly from the number of clients and multiple local updates (see Table 1 for a summary). To the best of our knowledge, this is the first universal convergence guarantee in FL that covers all three problem classes with a single algorithm and without problem-dependent tuning.
- **Further extensions.** We extend the algorithm to independent, nonuniform partial participation and establish universal convergence with exact local gradients across the same three problem classes; see Appendix E. We also prove that, with exact local gradients, the full-participation algorithm automatically adapts to Hölder smoothness of any order $\alpha \in [0, 1]$ for convex objectives, attaining the corresponding $\widetilde{O}(T^{-(1+\alpha)/2})$ rate; see Appendix F.

Table 1: Summary of convergence results for Algorithm 1. $\widetilde{O}(\cdot)$ hides logarithmic factors.

Problem class			**Gradient type**	**Assumptions**	**Convergence**
Smooth	**Lipschitz**	**Convex**			
✓	×	×	Exact	Assm. 4.1	$O\big(\frac{1}{T}\big)$
✓	×	×	Stochastic	Assms. 4.1, 4.2	$\widetilde{O}\big(\frac{1}{T} + \frac{\sigma}{\sqrt{mKT}}\big)$
✓	×	✓	Exact	Assm. 4.1	$O\big(\frac{1}{T}\big)$
✓	×	✓	Stochastic	Assms. 4.1, 4.2	$\widetilde{O}\big(\frac{1}{T} + \frac{\sigma}{\sqrt{mKT}}\big)$
×	✓	✓	Exact	Assm. 4.1	$\widetilde{O}\big(\frac{1}{\sqrt{T}}\big)$
×	✓	✓	Stochastic	Assms. 4.1, 4.2	$\widetilde{O}\big(\frac{1}{\sqrt{T}} + \frac{\sigma}{\sqrt{mKT}}\big)$

Compared to centralized analyses of AdaGrad (Duchi et al., 2011; Ward et al., 2020; Wang et al., 2023), FL creates two additional difficulties: local drift from multiple client updates and delayed accumulation due to unavailable aggregated gradient information. To handle these challenges, we link local drift to the server displacement, thereby reducing the analysis to server updates. We then bound this displacement almost surely by the maximal noise, whose moments admit logarithmic bounds, and use the result with a shifted summation lemma to control delayed accumulation. Notably, our analysis removes bounded-domain or bounded-iterate assumptions, which are commonly used in AdaGrad-type analyses (Levy, 2017; Levy et al., 2018; Kavis et al., 2019; Ene et al., 2021).

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 presents the algorithm. Section 4 gives the main convergence results and a proof sketch. Section 5 reports

the numerical experiments and Section 6 concludes the paper. All technical proofs are deferred to Appendix A–D. Extensions to partial participation and Hölder smoothness are established in Appendix E and F, respectively.

2 Related Work

Universal adaptive optimization. AdaGrad (Duchi et al., 2011) introduced adaptive step sizes based on accumulated gradients, with popular variants including Adam (Kingma & Ba, 2015) and AMSGrad (Reddi et al., 2018). Scalar AdaGrad steps have since been analyzed for stochastic and nonconvex optimization under progressively weaker assumptions. Kavis et al. (2022); Ward et al. (2020) provide high-probability guarantees under uniformly bounded gradients, which is later extended in Attia & Koren (2023); Faw et al. (2022); Wang et al. (2023) with relaxed assumptions. For convex optimization, AcceleGrad (Levy et al., 2018), universal accelerated methods (Ene et al., 2021; Kavis et al., 2019; Levy et al., 2018) improve over AdaGrad by achieving near-optimal rates simultaneously for convex smooth and nonsmooth objectives. Nesterov's universal method (Nesterov, 2015) instead uses line search. Further extensions treat inexact stochastic oracles (Rodomanov et al., 2024) and Hölder smoothness (Orabona, 2023). More closely related, Wang & Yurtsever (2026) replace gradient accumulation with gradient-mapping accumulation and obtain a method for smooth nonconvex, smooth convex, and nonsmooth convex composite problems.

Adaptive federated optimization. Existing adaptive FL methods differ mainly in where the adaptive step size is maintained. At the server, FedOpt (Reddi et al., 2021) applies AdaGrad, Adam, or Yogi to aggregated client displacements, while FedAMS/FedCAMS (Wang et al., 2022) add AMSGrad-style stabilization. Other server-side variants include FedUR (Zhang et al., 2023), FedDuA (Takakura et al., 2025), FedAgg Yuan & Wang (2025), and the asynchronous method FADAS (Wang et al., 2024). At the client level, Local AdaAlter (Xie et al., 2020) synchronizes AdaGrad denominators, whereas Δ-SGD (Kim et al., 2024) follows a different type of adaptivity (Malitsky & Mishchenko, 2020) to adapt each client's step size to its local smoothness. Other local adaptive methods include FedSPS (Mukherjee et al., 2023) and FedLADA (Sun et al., 2023). Methods acting at both levels include FAFED (Wu et al., 2023) and FedAda^2 (Lee et al., 2026), which coordinate server statistics, preconditioners, or global corrections with local adaptation. However, their analyses largely focus on one specific problem class (mostly smooth nonconvex or smooth convex), and some results require smoothness-dependent step-size conditions (e.g., FedOpt (Reddi et al., 2021) and FAFED (Wu et al., 2023)). In contrast, our method uses one unchanged rule to adapt across three problem classes, without knowledge of problem-related parameters. A more detailed summary of existing adaptive FL methods is provided in Appendix H.

3 Algorithm Design

Notation. We use $\|\cdot\|$ for the Euclidean norm and denote $[n] := \{1, 2, \dots, n\}$. For a stochastic algorithm, we denote by $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$ the expectation conditioned on $\mathcal{F}_t$, which contains all the information available at step t. Depending on the differentiability of F, we use ∇F to denote the gradient or a subgradient of F.

Algorithm overview. We present the adaptive federated learning method in Algorithm 1. It initializes the server model at x_1 and the shared step size at $\eta_1 = \eta/\gamma$. At each round t, the server broadcasts (x_t, η_t, τ_t) to all m clients. Each client i starts from $x_{i,0}^t = x_t$, performs K local (stochastic) gradient /subgradient steps with step size $\tau_t\eta_t$, and returns its local displacement $\Delta_i^t = x_{i,K}^t - x_t$ to server. The server then computes the new model and uses it to update the step size.

Note that the scales $\{\tau_t\}$ are introduced to mitigate local drift due to multiple local updates. At the server, the factor $1/\tau_t$ in the aggregation cancels this scaling, so that τ_t does not directly affect the server-level update. Indeed, let $q_t := 1/\eta_t$, and define

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

Algorithm 1 Adaptive Federated Learning Method (AdaFed)

Require: initial model $x_1 \in \mathbb{R}^d$; local-step count $K \geq 1$; server iterations $T \geq 1$; prescribed local scales $\tau_t > 0$; parameters $\gamma, \eta > 0$.
Initialize the step size $\eta_1 := \eta/\gamma$.
for $t = 1, 2, \ldots, T$ **do**
  Server broadcasts (x_t, η_t, τ_t).
  for each **client** $i \in [m]$ **in parallel do**
    Initialize $x_{i,0}^t := x_t$.
    for $k = 0, \ldots, K-1$ **do**
      Update $x_{i,k+1}^t := x_{i,k}^t - \tau_t \eta_t g_{i,k}^t$ with $g_{i,k}^t \approx \nabla F_i(x_{i,k}^t)$.
    end for
    Set $\Delta_i^t := x_{i,K}^t - x_t$ and send Δ_i^t to the server.
  end for
  Server updates the model and the step size:

$$x_{t+1} := x_t + \frac{1}{mK\tau_t} \sum_{i=1}^{m} \Delta_i^t, \qquad \frac{1}{\eta_{t+1}^2} := \frac{1}{\eta_t^2} \left(1 + \frac{\|x_{t+1} - x_t\|^2}{\eta^2}\right).$$

end for
Output: $\bar{x}_T$ chosen uniformly at random from $\{x_t\}_{t=1}^T$.

Then, since $\Delta_i^t = -(K\tau_t/q_t)\bar{g}_t^{(i)}$, the server update in Algorithm 1 can be written as

$$x_{t+1} = x_t - \frac{\bar{g}_t}{q_t}, \qquad q_{t+1}^2 = q_t^2 + \frac{\|\bar{g}_t\|^2}{\eta^2}. \tag{Core-FP}$$

Motivation for the adaptive step size. Individual client gradients generally do not vanish at a stationary point of the global objective. As a result, accumulating their squared norms can force the adaptive step size to keep decaying even as the global gradient vanishes, preventing the method from exploiting global smoothness. We therefore construct the adaptive accumulator from the aggregated server direction $\bar{g}_t$. In the special case $K = 1$ with exact local gradients, $\bar{g}_t = \nabla F(x_t)$, and the recursion in (Core-FP) reduces directly to a centralized AdaGrad-type update on the global objective. With multiple local updates, however, $\bar{g}_t$ consists of gradients evaluated along different client trajectories and is no longer the global gradient at x_t. Additionally, the adaptive accumulator is delayed since the server displacement $x_{t+1} - x_t$ is unavailable during clients' local updates in round t. Establishing that this adaptive mechanism retains universal convergence despite these discrepancies is one of the main technical challenges of our analysis.

4 Convergence Analysis

In this section, we present the convergence results of Algorithm 1. We start by making the following standard assumption on problem (1).

Assumption 4.1. F is lower bounded, i.e, $F^\star := \inf_{x \in \mathbb{R}^d} F(x) > -\infty$. For convex F, we further assume that the minimum is attained at some point $x^\star \in \mathbb{R}^d$.

Let $\mathcal{F}_t$ contain the history before round t, including (x_t, q_t). For client i, let $\mathcal{F}_{t,k}^{(i)}$ contain $\mathcal{F}_t$ and its first k oracle calls in round t, so $x_{i,k}^t$ is $\mathcal{F}_{t,k}^{(i)}$-measurable. Define the local stochastic noise

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

For the stochastic setting, we make the following assumption.

Assumption 4.2. The noise variables $\{\xi_{i,k}^t\}$ are conditionally independent and satisfy

$$\mathbb{E}[\xi_{i,k}^t \mid \mathcal{F}_{t,k}^{(i)}] = 0, \qquad \mathbb{E}\left[\exp\left(\frac{\|\xi_{i,k}^t\|^2}{\sigma^2}\right) \middle| \mathcal{F}_{t,k}^{(i)}\right] \leq \exp(1)$$

for all $t \geq 1$, $i \in [m]$, and $k \in \{0, \ldots, K-1\}$.

Assumption 4.2 is stated for $\sigma > 0$. When $\sigma = 0$, it means $\xi_{i,k}^t = 0$ almost surely. This assumption is also used in (Li & Orabona, 2019; Nemirovski et al., 2009). It allows unbounded noise but is stronger than bounded conditional variance. Indeed, under Assumption 4.2, the convexity of $\exp(\cdot)$ implies

$$\mathbb{E}[\|\xi_{i,k}^t\|^2 \mid \mathcal{F}_{t,k}^{(i)}] \le \sigma^2. \tag{2}$$

However, the bounded variance assumption alone is insufficient to guarantee convergence of Algorithm 1, as indicated by the following proposition. The proof is deferred to Appendix A.2.

Proposition 4.3 (Bounded variance is insufficient)**.** *Fix any $m, K \ge 1$. Let $\{\tau_t\}$ be chosen arbitrarily and set $\gamma = \eta = 1$. Consider the following convex, 1-smooth, and 1-Lipschitz client objectives:*

$$F_i(x) = F(x) := \sqrt{1 + x^2} - 1, \quad i = 1, \dots, m.$$

If $x_1 = 0$, then for any T divisible by 4, there exist independent noises satisfying $\mathbb{E}[\xi_{i,k}^t \mid \mathcal{F}_{t,k}^{(i)}] = 0$ and $\mathbb{E}[|\xi_{i,k}^t|^2 \mid \mathcal{F}_{t,k}^{(i)}] \le 1$ such that

$$\mathbb{E}[|F'(\bar{x}_T)|^2] \ge \frac{1}{16384mK^2}, \qquad \mathbb{E}[F(\bar{x}_T) - F^\star] \ge \frac{\sqrt{T}}{2048mK^2}.$$

The counterexample exploits rare but large noise realizations permitted under bounded variance. A single such realization can move the server iterate far from the minimizer, and also produce a large server displacement before it is incorporated into the adaptive accumulator. The resulting increase in the accumulator then makes subsequent steps too small for the iterates to recover within the remaining rounds.

4.1 Main results

We are now ready to state our main convergence results, which establish universal convergence for three distinct problem classes: smooth nonconvex, smooth convex, and Lipschitz convex. For ease of presentation, we define the following quantities:

$$\mathcal{U}_T := 1 + \sigma\sqrt{\frac{T}{mK}} + \sigma^2\big(1 + \log(mKT)\big)\log\big(1 + \sigma^2(1 + \log(mKT))\big),$$
$$\mathcal{C}_T := \big(1 + \sigma^2(1 + \log(mKT))\big)\big(1 + \log \mathcal{U}_T\big), \tag{3}$$
$$\mathcal{V}_T := \big(1 + \sigma^2(1 + \log(mKT))\big)\left(1 + \log\left(\sqrt{T} + \sigma\sqrt{T/(mK)}\right)\right) + (1 + \sigma)\log T. \tag{4}$$

We use $\lesssim$ to hide constants depending only on fixed quantities (e.g., L, G, γ, η) hereafter, i.e., $A \lesssim B$ if there exists a constant $c > 0$ such that $A \le cB$. We also use $\widetilde{O}(\cdot)$ to hide polylogarithmic factors.

Theorem 4.4 (Universal convergence of Algorithm 1)**.** *Let Assumptions 4.1 and 4.2 hold. Let $\{x_t\}_{t\ge1}$ be the iterates generated by Algorithm 1 with arbitrary $\gamma, \eta > 0$ and τ_t chosen as*

$$\tau_t = 1 \text{ if } K = 1; \qquad \tau_t \le \frac{1}{(K-1)t} \text{ if } K \ge 2. \tag{5}$$

Then, for any $T \ge 1$, we have the following results:

(i) Smooth nonconvex. *If each F_i is L-smooth, then*

$$\mathbb{E}[\|\nabla F(\bar{x}_T)\|^2] \lesssim \frac{\mathcal{C}_T^2}{T} + \frac{\sigma\mathcal{C}_T}{\sqrt{mKT}}.$$

(ii) Smooth convex. *If each F_i is L-smooth and convex, then*

$$\mathbb{E}[F(\bar{x}_T) - F^\star] \lesssim \frac{\mathcal{C}_T^2}{T} + \frac{\sigma\mathcal{C}_T}{\sqrt{mKT}}.$$

(iii) Lipschitz convex. *If each F_i is convex and G-Lipschitz, then*

$$\mathbb{E}[F(\bar{x}_T) - F^\star] \lesssim \left(\frac{1}{\sqrt{T}} + \frac{\sigma}{\sqrt{mKT}}\right)\mathcal{V}_T.$$

Case (i) is proved as Theorem B.4 in Appendix B, case (ii) as Theorem C.3 in Appendix C, and case (iii) as Theorem D.4 in Appendix D. In fact, for the smooth convex case and $K \geq 2$, we can use the more relaxed choice $\tau_t \leq \frac{\tau}{(K-1)t^p}$ for any $p > 1/2$ and $\tau > 0$ without degrading the convergence rate, see Theorem C.3 for details.

To provide more explicit convergence rates in the stochastic setting, note that $\mathcal{C}_T \lesssim \log^2(mKT)$ and $\mathcal{V}_T \lesssim \log^2(mKT)$. Consequently, the three convergence rates in Theorem 4.4 can be simplified to

$$\widetilde{O}\Big(\frac{1}{T} + \frac{\sigma}{\sqrt{mKT}}\Big), \quad \widetilde{O}\Big(\frac{1}{T} + \frac{\sigma}{\sqrt{mKT}}\Big), \quad \widetilde{O}\Big(\frac{1}{\sqrt{T}} + \frac{\sigma}{\sqrt{mKT}}\Big), \tag{6}$$

respectively. In every setting, the rate matches that of classical AdaGrad in the non-federated case (up to logarithmic factors captured by $\mathcal{C}_T$ and $\mathcal{V}_T$), while the $\sigma/\sqrt{mKT}$ term shows that the convergence benefits jointly from the number of clients m and local updates K.

4.2 Proof sketch

Here we provide a proof sketch for the smooth nonconvex case, which also serves as the basis of the analysis for convex cases. Besides stochasticity, the main challenges in the analysis arise from *local drift* and *delayed accumulator*, which are absent in the classical AdaGrad analysis.

Note that $\bar{g}_t$ averages gradients along moving local paths, whereas stationarity is measured by $g_t := \nabla F(x_t)$. To estimate their difference, we decompose $\bar{g}_t = g_t + \hat{b}_t + \bar{\xi}_t$, where $\bar{\xi}_t$ denotes average of the stochastic errors and $\hat{b}_t$ is the local gradient drift. Then it follows that

$$\mathbb{E}_t[\|\bar{g}_t\|^2] \leq 2\|g_t\|^2 + 2\mathbb{E}_t[\|\hat{b}_t\|^2] + \frac{2\sigma^2}{mK}. \tag{7}$$

To further bound $\mathbb{E}_t[\|\hat{b}_t\|^2]$, we unroll the local updates and obtain the local distance $\mathbb{E}_t[\|x_{i,k}^t - x_t\|^2] \lesssim (k\tau_t/q_t)^2(\|\nabla F_i(x_t)\| + \sigma)^2$, which together with L-smoothness then implies

$$\mathbb{E}_t[\|\hat{b}_t\|^2] \leq \frac{L^2}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}_t[\|x_{i,k}^t - x_t\|^2] \lesssim \frac{\tau_t^2}{mq_t^2}\sum_{i=1}^{m}\big(\|\nabla F_i(x_t)\| + \sigma\big)^2.$$

To proceed, we use smoothness to write $\|\nabla F_i(x_t)\| \leq \|\nabla F_i(x_1)\| + L\|x_t - x_1\|$ and then bound sum of $\tau_t^2\|x_t - x_1\|^2/q_t$. Set $r_t := x_{t+1} - x_t$ and $\mathcal{R}_T := \sum_{t=1}^T \|r_t\|^2$. Since $x_t - x_1 = \sum_{j<t} r_j$, Hardy's inequality gives $\sum_{t=1}^T \tau_t^2\|x_t - x_1\|^2/q_t^2 \lesssim \mathcal{R}_T$. Consequently, we obtain

$$\sum_{t=1}^{T}\mathbb{E}_t[\|\hat{b}_t\|^2] \lesssim 1 + \mathcal{R}_T. \tag{8}$$

Note that when $K = 1$, $\hat{b}_t = 0$ exactly.

Since (8) converts the local gradient drift into server displacement, we may now focus on the server-level analysis, where the natural next step is to analyze $\mathcal{R}_T$. However, delayed accumulator directly introduces a significant challenge: q_t does not contain $\bar{g}_t$, so $\|r_t\| = \|\bar{g}_t\|/q_t$ is not automatically bounded, which prevents us to apply an AdaGrad-type summation property (cf. Lemma A.4). To address this issue, we use the decomposition $\bar{g}_t = \bar{g}_{t-1} + (g_t - g_{t-1}) + (b_t - b_{t-1})$ to write

$$\|r_t\| = \frac{\|\bar{g}_t\|}{q_t} \leq \frac{\|\bar{g}_{t-1}\|}{q_t} + \frac{\|g_t - g_{t-1}\|}{q_t} + \frac{\|b_t\| + \|b_{t-1}\|}{q_t},$$

where $b_t = \hat{b}_t + \bar{\xi}_t$. Then the first term is bounded trivially due to $q_t \geq \|\bar{g}_{t-1}\|/\eta$. Moreover, smoothness and (Core-FP) also gives $\|g_t - g_{t-1}\|/q_t \leq L\|x_t - x_{t-1}\|/q_t = L\|\bar{g}_{t-1}\|/(q_{t-1}q_t) \leq L\eta/q_1$. To bound the last term, let $M_T := \max_{t\leq T,i,k}\|\xi_{i,k}^t\|$ be the largest noise, then by bounding $\hat{b}_t$ similarly as before and noting that $\bar{\xi}_t \leq M_T$, we can show $\|b_t\| \lesssim 1 + M_T$. Combining these bounds together gives

$$\|r_t\| \lesssim 1 + M_T.$$

This means $\|r_t\|$ is controlled *almost surely* for all $t \le T$, which allows us to apply the summation property in Lemma A.4 to obtain

$$\mathcal{R}_T \lesssim \big(1 + M_T^2\big) \log \frac{q_{T+1}}{q_1}.$$

Assumption 4.2 implies $(\mathbb{E}[M_T^{2p}])^{1/p} = \widetilde{O}(\sigma^2)$ for $p \ge 1$, which further gives $\mathbb{E}[\mathcal{R}_T] = \widetilde{O}\left(\log \mathbb{E}[q_{T+1}]\right) = \widetilde{O}(1)$. This property matches the classic AdaGrad up to logarithmic factors and reveals the reason why Algorithm 1 preserves the rate of AdaGrad. In contrast, bounded variance assumption (2) may instead lead to a weaker linear reliance $\mathbb{E}[M_T^2] \lesssim \sigma^2 T$.

Let $\mathcal{G}_T := \sum_{t=1}^T \|g_t\|^2/q_t$ and $\zeta_t := \bar{g}_t - \mathbb{E}_t[\bar{g}_t]$. Smoothness and decomposition $\bar{g}_t = g_t + \mathbb{E}_t[\hat{b}_t] + \zeta_t$ yield

$$F(x_{t+1}) \le F(x_t) - \frac{\|g_t\|^2}{2q_t} + \frac{\mathbb{E}_t[\|\hat{b}_t\|^2]}{2q_t} - \frac{\langle g_t, \zeta_t\rangle}{q_t} + \frac{L}{2}\|r_t\|^2.$$

Summing this inequality and plugging (8), we obtain $\mathbb{E}[\mathcal{G}_T] \lesssim 1 + \mathbb{E}[\mathcal{R}_T]$ since ζ_t is unbiased. To bound $\mathbb{E}[q_{T+1}]$, we use $q_{T+1}^2 = q_1^2 + \sum_{t=1}^T \|\bar{g}_t\|^2/\eta^2$ and apply (7) and (8) to obtain

$$\mathbb{E}[q_{T+1}] \lesssim 1 + \sigma\sqrt{\frac{T}{mK}} + \mathbb{E}[\mathcal{G}_T] + \mathbb{E}[\mathcal{R}_T],$$

where we also use $\mathbb{E}\big[(\sum_{t=1}^T \|g_t\|^2)^{1/2}\big] \le \mathbb{E}[\sqrt{q_{T+1}\mathcal{G}_T}] \le \sqrt{\mathbb{E}[q_{T+1}]}\sqrt{\mathbb{E}[\mathcal{G}_T]}$ to solve a self-bounding inequality for $\mathbb{E}[q_{T+1}]$. Then substituting the established bounds on $\mathbb{E}[\mathcal{G}_T]$ and $\mathbb{E}[\mathcal{R}_T]$, we derive $\mathbb{E}[q_{T+1}] \lesssim \mathcal{U}_T$, which further gives $\mathbb{E}[\mathcal{R}_T] \lesssim \mathcal{C}_T$ and $\mathbb{E}[\mathcal{G}_T] \lesssim \mathcal{C}_T$.

It remains to bound the second moments $\mathbb{E}[\mathcal{G}_T^2]$ and $\mathbb{E}[q_{T+1}^2]$. Using the 4-th momentum $\mathbb{E}[M_T^4]$ and similar arguments for estimating $\mathbb{E}[\mathcal{R}_T]$, we can show $\sqrt{\mathbb{E}[\mathcal{R}_T^2]} \lesssim \mathcal{C}_T$, which further implies $\sqrt{\mathbb{E}[\mathcal{G}_T^2]} \lesssim \mathcal{C}_T$. Then by similarly using $q_{T+1}^2 = q_1^2 + \sum_{t=1}^T \|\bar{g}_t\|^2/\eta^2$ and applying (7) and (8), we obtain

$$\mathbb{E}[q_{T+1}^2] \lesssim 1 + \frac{T\sigma^2}{mK} + \mathbb{E}[\mathcal{G}_T^2] + \mathbb{E}[\mathcal{R}_T] \lesssim \mathcal{C}_T^2 + \frac{T\sigma^2}{mK}.$$

Finally, $q_t \le q_{T+1}$ and Cauchy–Schwarz imply

$$\mathbb{E}[\|\nabla F(\bar{x}_T)\|^2] = \frac{1}{T}\sum_{t=1}^T \mathbb{E}[\|g_t\|^2] \le \frac{1}{T}\sqrt{\mathbb{E}[q_{T+1}^2]}\sqrt{\mathbb{E}[\mathcal{G}_T^2]} \lesssim \frac{\mathcal{C}_T^2}{T} + \frac{\sigma\mathcal{C}_T}{\sqrt{mKT}}.$$

5 Numerical Experiments

In this section, we evaluate the empirical performance of Algorithm 1 with the primary goal of examining its adaptation to the underlying problem class. In particular, our experiments test whether a common parameterization adapts across the smooth nonconvex, smooth convex, and nonsmooth convex regimes, and whether increasing the number of local steps improves performance under stochastic gradients, as predicted by Theorem 4.4.

For all experiments, each local objective is a logistic loss with a regime-specific penalty. Specifically, if client i holds n_i samples (a_{ij}, y_{ij}) with $y_{ij} \in \{-1, +1\}$, we set

$$F_i(x) = \frac{1}{n_i}\sum_{j=1}^{n_i} \log\big(1 + \exp(-y_{ij} a_{ij}^\top x)\big) + \lambda r(x). \tag{9}$$

We set $\lambda = 0.01$ throughout, with $r(x) = \|x\|_1$ for the nonsmooth convex case, $\frac{1}{2}\|x\|_2^2$ for the smooth convex case, and $\sum_j x_j^2/(1 + x_j^2)$ in the smooth nonconvex case.

We use 32,000 examples from the LIBSVM (Chang & Lin, 2011) IJCNN1 binary-classification dataset ($d = 22$), partitioned equally among $m = 32$ clients. To induce data heterogeneity, we sort examples by their margin under a centralized reference solution, form blocks of 50, and distribute examples within each block round-robin across clients. Additional datasets and experimental details are provided in Appendix G.

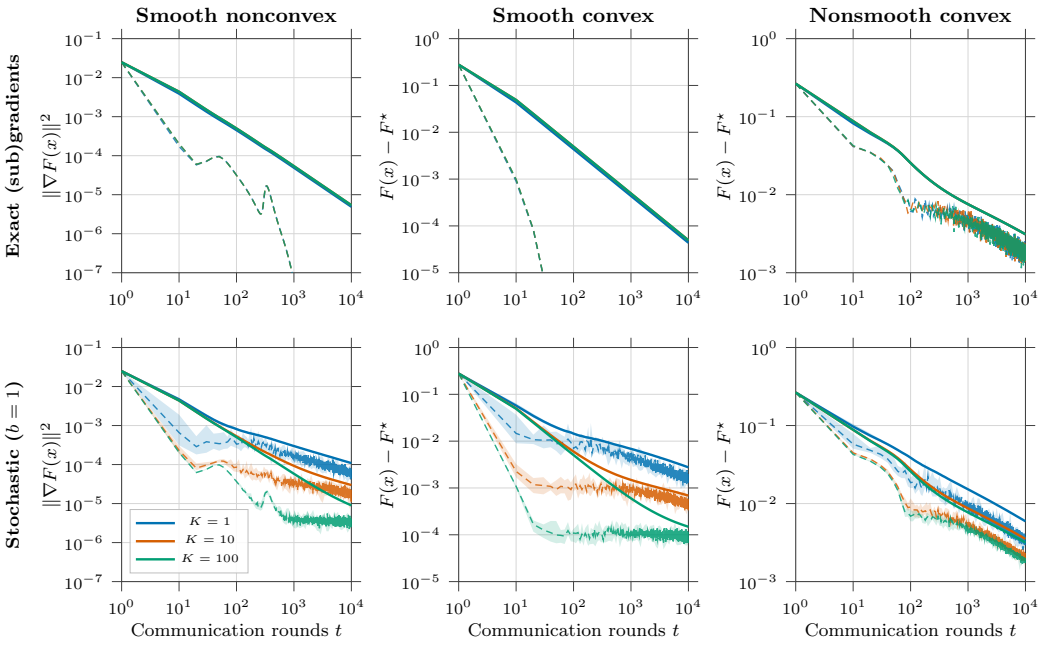

Figure 1: Performance of AdaFed on IJCNN1 across the three problem classes. Solid and dashed curves show the uniform-random and last-iterate metrics, respectively. Stochastic results show the mean over ten seeds, with shaded regions indicating the upward and downward RMS deviations.

For evaluation, we report $\|\nabla F(x)\|^2$ for the smooth nonconvex problem and the objective residual $F(x) - F^\star$ for the convex problems. The reference value $F^\star$ is computed to high accuracy using a centralized solver (damped Newton for the smooth convex problem and FISTA followed by active-set Newton for the nonsmooth convex problem). Following Theorem 4.4, the expected value for the uniform-random output metric is computed by averaging the corresponding metric along the trajectory. We additionally report the last-iterate metric for reference.

Adaptation across problem classes and local steps. We first examine whether Algorithm 1 adapts to the underlying problem class under a common parameterization and how its behavior changes with the number of local steps. Figure 1 shows the results for $K \in \{1, 10, 100\}$, with exact local gradients in the top row and single-sample stochastic gradients in the bottom row. Under exact gradients, the two smooth regimes exhibit the faster $1/T$ behavior, whereas the nonsmooth convex regime exhibits the slower $1/\sqrt{T}$ behavior, consistently across all three choices of K. Under stochastic gradients, increasing K delays the transition from the faster $1/T$ behavior to the noise-dominated $1/\sqrt{KT}$ regime in the two smooth settings. In the nonsmooth convex setting, increasing K instead improves the constant while preserving the $1/\sqrt{T}$ behavior. These observations are consistent with the dependence on the problem class, T, and K predicted by Theorem 4.4.

Comparison with adaptive FL methods. Figure 2 compares Algorithm 1 with representative methods using server-side, client-side, and both server- and client-side adaptation: FedAdagrad (Reddi et al., 2021), FedLADA (Sun et al., 2023), and FAFED (Wu et al., 2023), respectively. We use $K = 10$ local steps with full client participation and consider both exact local gradients and stochastic local gradients (minibatch size $b = 32$). Each method's main parameters are selected from short tuning runs, then fixed for the 10,000-round comparison (see Appendix G.6 for the tuning details).

Across all three problem classes, Algorithm 1 is competitive with the baselines while consistently following the rate behavior associated with the corresponding regime. In contrast, the baselines stagnate in some settings. This distinction is particularly clear under stochastic gradients and in the nonsmooth convex regime, where the random-output metrics of the baselines reach apparent accuracy floors, while Algorithm 1 continues to decrease. These results further illustrate the cross-regime adaptation predicted by Theorem 4.4.

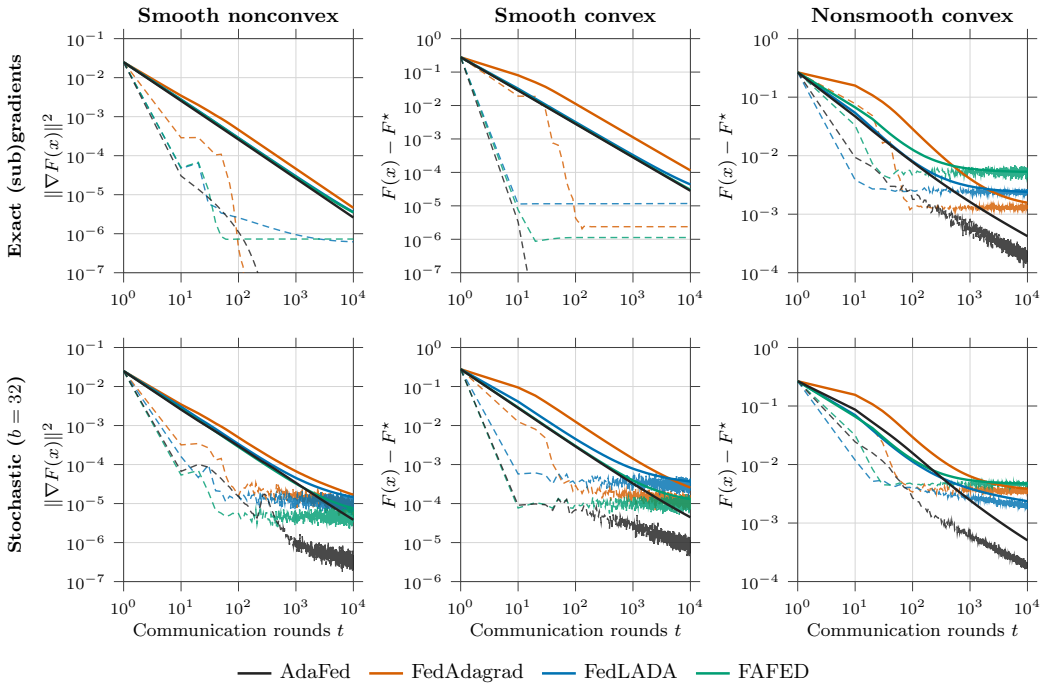

Figure 2: Performance of AdaFed compared with representative adaptive FL methods on IJCNN1 with $K = 10$ local steps. FedAdagrad, FedLADA, and FAFED represent server-, client-, and both-level adaptation, respectively. Solid and dashed curves show the uniform-random and last-iterate metrics, respectively. Stochastic results are shown as the mean over three seeds.

Additional experiments. Appendix G provides additional experiments on synthetic problems, larger minibatches, parameter sensitivity for η and γ, and partial participation. It also extends the comparison to FedAdam, FedUR, Local AdaAlter, FedSPS, and FedAda2, covering a broader range of adaptive FL methods.

6 Conclusion

Our results show that universal convergence can be achieved in federated learning (FL) by using AdaGrad-type step sizes, despite heterogeneous client objectives and multiple local updates. The key is to accumulate the server displacements, which reflect global rather than client-level stationarity from a consensus viewpoint of the FL problem. Additionally, the scaled local step controls client drift without altering the aggregated server update. From the analytical perspective, relating the local drift to server displacements and controlling the delayed accumulator yields the essential argument without bounded-iterate assumptions, which results in convergence rates comparable to centralized AdaGrad across the three problem classes. Our experiments support these theoretical findings, showing the predicted cross-regime behavior under a common parameterization and competitive performance against existing adaptive FL methods. We further provide extensions to nonuniform partial participation and Hölder-smoothness in the exact local-gradient setting in the appendix.

Several directions remain open. A natural next step is to extend both the nonuniform partial-participation and Hölder-smooth analyses to stochastic local gradients. This may be possible using moment estimates similar to those in the full-participation analysis, but the details are nontrivial and require careful investigation. Another question is whether the logarithmic factors caused by delayed accumulation can be removed through a sharper analysis or a modified update rule. Finally, combining the proposed method with acceleration may lead to improved convergence rates while preserving universality for convex problems.

AI USE STATEMENT

In this work, we used generative AI tools to assist with proof development and verification, implementing and checking numerical experiments, preparing figures, interpreting the recorded results, and drafting and editing manuscript text. We reviewed all AI-assisted content and take responsibility for the resulting claims, code, figures, and text.

REPRODUCIBILITY STATEMENT

Section 3 specifies the algorithm, and Section 4 states its assumptions and guarantees; complete proofs appear in the appendix. Section 5 gives the experimental objectives, metrics, and main results. Appendices G.1, G.2, and G.6 document data construction, client partitions, tuning, and run settings.

REFERENCES

Amit Attia and Tomer Koren. SGD with AdaGrad stepsizes: Full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Evan Chen, Shiqiang Wang, Jianing Zhang, Dong-Jun Han, Chaoyue Liu, and Christopher Brinton. Gradient correction in federated learning with adaptive optimization. *arXiv preprint arXiv:2502.02727*, 2025.

Savelii Chezhegov, Sergey Skorik, Nikolas Khachaturov, Danil Shalagin, Aram Avetisyan, Martin Takáč, Yaroslav Kholodov, and Aleksandr Beznosikov. Local methods with adaptivity via scaling. *arXiv preprint arXiv:2406.00846*, 2024.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Rick Durrett. *Probability: Theory and Examples*. Number 49 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 5 edition, 2019.

Alina Ene, Huy L. Nguyen, and Adrian Vladu. Adaptive gradient methods for constrained convex optimization and variational inequalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward. The power of adaptivity in SGD: Self-tuning step sizes with unbounded gradients and affine variance. In *Proceedings of the Thirty-Fifth Conference on Learning Theory*, 2022.

Divyansh Jhunjhunwala, Pranay Sharma, Aushim Nagarkatti, and Gauri Joshi. FedVARP: Tackling the variance due to partial client participation in federated learning. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, 2022.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. In *International Conference on Learning Representations*, 2021.

Ali Kavis, Kfir Y. Levy, Francis Bach, and Volkan Cevher. UniXGrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. In *Advances in Neural Information Processing Systems*, 2019.

Ali Kavis, Kfir Y. Levy, and Volkan Cevher. High probability bounds for a class of nonconvex algorithms with AdaGrad stepsize. In *International Conference on Learning Representations*, 2022.

Junhyung Lyle Kim, Mohammad Taha Toghani, Cesar Uribe, and Anastasios Kyrillidis. Adaptive federated learning with auto-tuned clients. In *International Conference on Learning Representations*, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Su Hyeong Lee, Sidharth Sharma, Manzil Zaheer, and Tian Li. Efficient adaptive federated optimization. In *Advances in Neural Information Processing Systems*, 2026.

Kfir Levy. Online to offline conversions, universality and adaptive minibatch sizes. In *Advances in Neural Information Processing Systems*, 2017.

Kfir Y. Levy, Alp Yurtsever, and Volkan Cevher. Online adaptive methods, universality and acceleration. In *Advances in Neural Information Processing Systems*, 2018.

Junyi Li, Feihu Huang, and Heng Huang. Fedda: Faster adaptive gradient methods for federated constrained optimization. In *The Twelfth International Conference on Learning Representations*, 2024.

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*, 2020.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 2019.

Wei Liu, Anweshit Panda, Ujwal Pandey, Haven Cook, George M Slota, Naigang Wang, Jie Chen, and Yangyang Xu. Lodadac: a unified local training-based decentralized framework with adaptive gradients and compressed communication. *arXiv preprint arXiv:2604.09970*, 2026.

Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

Sohom Mukherjee, Nicolas Loizou, and Sebastian U Stich. Locally adaptive federated learning. *arXiv preprint arXiv:2307.06306*, 2023.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Yurii Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1–2):381–404, 2015. doi: 10.1007/s10107-014-0790-0.

Francesco Orabona. Normalized gradients for all. *arXiv preprint arXiv:2308.05621*, 2023.

Sashank 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.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

Anton Rodomanov, Xiaowen Jiang, and Sebastian Stich. Universality of AdaGrad stepsizes for stochastic optimization: Inexact oracle, acceleration and variance reduction. In *Advances in Neural Information Processing Systems*, 2024.

Yan Sun, Li Shen, Hao Sun, Liang Ding, and Dacheng Tao. Efficient federated learning via local adaptive amended optimizer with linear speedup. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):14453–14464, 2023.

Shokichi Takakura, Seng Pei Liew, and Satoshi Hasegawa. FedDuA: Doubly adaptive federated learning. *arXiv preprint arXiv:2505.11126*, 2025.

Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of AdaGrad for non-convex objectives: Simple proofs and relaxed assumptions. In *Proceedings of Thirty Sixth Conference on Learning Theory*, 2023.

Chenhao Wang, Zihan Chen, Nikolaos Pappas, Howard H Yang, Tony QS Quek, and H Vincent Poor. Adaptive federated learning over the air. *IEEE Transactions on Signal Processing*, 73:3187–3202, 2025.

Jianyu Wang, Zheng Xu, Zachary Garrett, Zachary Charles, Luyang Liu, and Gauri Joshi. Local adaptivity in federated learning: Convergence and consistency. *arXiv preprint arXiv:2106.02305*, 2021.

Yujia Wang, Lu Lin, and Jinghui Chen. Communication-efficient adaptive federated learning. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Yujia Wang, Shiqiang Wang, Songtao Lu, and Jinghui Chen. FADAS: Towards federated adaptive asynchronous optimization. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Zimeng Wang and Alp Yurtsever. Universal adaptive proximal gradient methods via gradient mapping accumulation. *arXiv preprint arXiv:2605.05944*, 2026.

Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(219):1–30, 2020.

Xidong Wu, Feihu Huang, Zhengmian Hu, and Heng Huang. Faster adaptive federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Cong Xie, Oluwasanmi Koyejo, Indranil Gupta, and Haibin Lin. Local AdaAlter: Communication-efficient stochastic gradient descent with adaptive learning rates. In *OPT 2020: 12th Annual Workshop on Optimization for Machine Learning*, 2020.

Wenhao Yuan and Xuehe Wang. Aggregated gradients-based adaptive learning rate design in federated learning. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 2025.

Hengrun Zhang, Kai Zeng, and Shuai Lin. FedUR: Federated learning optimization through adaptive centralized learning optimizers. *IEEE Transactions on Signal Processing*, 71:2622–2637, 2023.

A PRELIMINARY

A.1 AUXILIARY LEMMAS

Lemma A.1. *Let $a, x > 0$ and $b \in \mathbb{R}$. If $x \le a \log x + b$, then $x \le 2a \log 2a - 2a + 2b$.*

Proof. Applying the basic inequality $\log u \le u - 1$ with $u = x/(2a)$ gives

$$\log x \le \frac{x}{2a} + \log 2a - 1,$$

which together with $x \le a \log x + b$ implies that

$$x \le \frac{x}{2} + a \log 2a - a + b.$$

We obtain the desired result after rearranging the terms. □

Lemma A.2. *Fix $\alpha \in (0, 1)$ and let $a, b, x \ge 0$. If $x \le ax^\alpha + b$, then $x \le (2a)^{1/(1-\alpha)} + 2b$.*

Proof. If $x \le 2b$, the result is immediate. Otherwise if $x > 2b$, then the given inequality implies

$$\frac{x}{2} < x - b \le ax^\alpha.$$

Therefore, $x^{1-\alpha} \le 2a$ and hence $x \le (2a)^{1/(1-\alpha)}$, which completes the proof. □

Lemma A.3. *For any $c, \eta > 0$, $0 < a < b$, and $x \ge 0$, we have*

$$x^a - cx^b \le \left(\eta^a + c^{-a/(b-a)}\right) \log\left(1 + \frac{x^a}{\eta^a}\right).$$

Proof. If $x^a - cx^b \le 0$, then the result is immediate. Otherwise we have $1 - cx^{b-a} > 0$, which implies $x < c^{-1/(b-a)}$ and hence $x^a < c^{-a/(b-a)}$. Using $\log(1 + z) \ge z/(1 + z)$ for $z \ge 0$, we obtain

$$\log\left(1 + \frac{x^a}{\eta^a}\right) \ge \frac{x^a}{\eta^a + x^a} \ge \frac{x^a}{\eta^a + c^{-a/(b-a)}} \ge \frac{x^a - cx^b}{\eta^a + c^{-a/(b-a)}},$$

which completes the proof. □

Lemma A.4 ((Wang & Yurtsever, 2026))**.** *Let $\{a_k\}_{k\ge1}$ be a nonnegative sequence and define $A_{k+1} = A_1 + \sum_{i=1}^k a_i$ with $A_1 > 0$. Then for all $t \ge 1$:*

$$2(\sqrt{A_{t+1}} - \sqrt{A_1}) \le \sum_{k=1}^t \frac{a_k}{\sqrt{A_k}}.$$

Moreover, if there exists $c > 0$ such that $a_k \le cA_k$ for all $k \ge 1$, then

$$\sum_{k=1}^t \frac{a_k}{\sqrt{A_k}} \le (1 + \sqrt{1 + c})(\sqrt{A_{t+1}} - \sqrt{A_1}),$$

$$\sum_{k=1}^t \frac{a_k}{A_k} \le (1 + c) \log \frac{A_{t+1}}{A_1}.$$

Lemma A.5. *For every $a \in \mathbb{R}$ and $p \ge 1$, the function $x \mapsto (a + \log x)^p$ is concave on $[\exp(p - 1 - a), \infty)$.*

Proof. If $p = 1$, then $a + \log x$ is concave on $(0, \infty)$. Next we assume $p > 1$. For $x \ge \exp(p-1-a)$, we have $a + \log x \ge p - 1 > 0$ and

$$\frac{d^2}{dx^2}(a + \log x)^p = \frac{p(a + \log x)^{p-2}(p - 1 - a - \log x)}{x^2} \le 0,$$

which completes the proof. □

Lemma A.6 (Conditional L^p triangle inequality)**.** *Let $p \geq 1$, let X and Y be random vectors such that $\mathbb{E}[\|X\|^p] < \infty$ and $\mathbb{E}[\|Y\|^p] < \infty$, and let $\mathcal{G}$ be a sub-σ-algebra. Then, almost surely,*

$$(\mathbb{E}[\|X+Y\|^p \mid \mathcal{G}])^{1/p} \leq (\mathbb{E}[\|X\|^p \mid \mathcal{G}])^{1/p} + (\mathbb{E}[\|Y\|^p \mid \mathcal{G}])^{1/p}.$$

Proof. For $p = 1$, the claim follows directly from $\|X+Y\| \leq \|X\| + \|Y\|$. Suppose $p > 1$ and let $q = p/(p-1)$, so that $1/p + 1/q = 1$. The pointwise triangle inequality and conditional Hölder inequality give

$$\begin{aligned}
&\mathbb{E}[\|X+Y\|^p \mid \mathcal{G}] \leq \mathbb{E}[\|X\|\|X+Y\|^{p-1} \mid \mathcal{G}] + \mathbb{E}[\|Y\|\|X+Y\|^{p-1} \mid \mathcal{G}] \\
&\leq (\mathbb{E}[\|X\|^p \mid \mathcal{G}])^{1/p} \left(\mathbb{E}[\|X+Y\|^{(p-1)q} \mid \mathcal{G}]\right)^{1/q} \\
&\quad + (\mathbb{E}[\|Y\|^p \mid \mathcal{G}])^{1/p} \left(\mathbb{E}[\|X+Y\|^{(p-1)q} \mid \mathcal{G}]\right)^{1/q} \\
&= \left((\mathbb{E}[\|X\|^p \mid \mathcal{G}])^{1/p} + (\mathbb{E}[\|Y\|^p \mid \mathcal{G}])^{1/p}\right) (\mathbb{E}[\|X+Y\|^p \mid \mathcal{G}])^{(p-1)/p},
\end{aligned}$$

which proves the result after simplification. □

Lemma A.7 (Doob's L^2 maximal inequality (Durrett, 2019, Theorem 4.4.4))**.** *Let $\{\mathcal{M}_t\}_{t=0}^T$ be a sequence of square-integrable martingale. Then*

$$\mathbb{E}\left[\max_{0 \leq t \leq T} |\mathcal{M}_t|^2\right] \leq 4\mathbb{E}[|\mathcal{M}_T|^2].$$

Lemma A.8. *Let $Z_t \geq 0$ be integrable and $\mathcal{F}_{t+1}$-measurable for all $t \geq 1$. For any constant $c > 0$ and $T \geq 1$, we have*

$$\mathbb{E}\left[\sqrt{c + \sum_{t=1}^T Z_t}\right] \leq \sqrt{2}\,\mathbb{E}\left[\sqrt{c + \sum_{t=1}^T \mathbb{E}_t[Z_t]}\right].$$

Proof. Let $a_t := \mathbb{E}_t[Z_t]$, $X_T := c + \sum_{t=1}^T Z_t$, and $A_t := c + \sum_{j=1}^t a_j$. Since $A_T \geq A_t \geq c$, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{X_T}{\sqrt{A_T}}\right] &\leq \sqrt{c} + \sum_{t=1}^T \mathbb{E}\left[\frac{Z_t}{\sqrt{A_t}}\right] = \sqrt{c} + \sum_{t=1}^T \mathbb{E}\left[\frac{a_t}{\sqrt{A_t}}\right] \\
&\leq \sqrt{c} + 2\mathbb{E}\left[\sum_{t=1}^T (\sqrt{A_t} - \sqrt{A_{t-1}})\right] = \sqrt{c} + 2\mathbb{E}[\sqrt{A_T}] - 2\sqrt{A_0} \leq 2\mathbb{E}[\sqrt{A_T}],
\end{aligned}$$

where the second inequality follows from

$$\frac{a_t}{\sqrt{A_t}} = \frac{A_t - A_{t-1}}{\sqrt{A_t}} \leq \frac{2(A_t - A_{t-1})}{\sqrt{A_t} + \sqrt{A_{t-1}}} = 2(\sqrt{A_t} - \sqrt{A_{t-1}}).$$

Therefore, Cauchy–Schwarz further yields

$$\mathbb{E}[\sqrt{X_T}] = \mathbb{E}\left[\sqrt{\frac{X_T}{\sqrt{A_T}}}\, A_T^{1/4}\right] \leq \sqrt{\mathbb{E}\left[\frac{X_T}{\sqrt{A_T}}\right]} \sqrt{\mathbb{E}[\sqrt{A_T}]} \leq \sqrt{2}\,\mathbb{E}[\sqrt{A_T}],$$

which completes the proof. □

A.2 Proof of Proposition 4.3

Proof. Note that $x^\star = 0$ and $F^\star = F(0) = 0$. Moreover, direct calculation gives

$$F'(x) = \frac{x}{\sqrt{1+x^2}}, \qquad F''(x) = \frac{1}{(1+x^2)^{3/2}}, \qquad |F'(x)| \leq 1, \qquad 0 < F''(x) \leq 1.$$

Thus every $F_i = F$ is convex, 1-smooth, and 1-Lipschitz. Since all client objectives are identical, their heterogeneity is zero.

Fix T divisible by 4, let $M := 16mK\sqrt{T}$ and all local noises $\{\xi_{i,k}^t\}$ be independent and have discrete distribution

$$\mathbb{P}(\xi = M) = \mathbb{P}(\xi = -M) = \frac{1}{2M^2}, \qquad \mathbb{P}(\xi = 0) = 1 - \frac{1}{M^2},$$

which satisfies the unbiasedness and bounded variance conditions:

$$\mathbb{E}[\xi] = M\frac{1}{2M^2} - M\frac{1}{2M^2} = 0, \qquad \mathbb{E}[\xi^2] = M^2\left(\frac{1}{2M^2} + \frac{1}{2M^2}\right) = 1.$$

Let $\mathcal{E}_T$ be the event: exactly one of the mKT noise realizations is nonzero and this realization occurs at the last local step of some client in a round $j \in \{T/2+1, \dots, 3T/4\}$. Then there are $mT/4$ possible locations for the unique nonzero noise, which implies

$$\mathbb{P}(\mathcal{E}_T) = \frac{mT}{4M^2}\left(1 - \frac{1}{M^2}\right)^{mKT-1} \geq \frac{mT}{4M^2}\left(1 - \frac{mKT-1}{M^2}\right) \geq \frac{mT}{8M^2} = \frac{1}{2048mK^2}. \tag{10}$$

Here we use $(1+x)^p \geq 1 + px$ for $x \geq -1$ and $p \geq 1$, and

$$\frac{mKT-1}{M^2} \leq \frac{mKT}{256m^2K^2T} = \frac{1}{256mK} \leq \frac{1}{2}.$$

Now let us consider the event $\mathcal{E}_T$. Since $F'(x_1) = 0$, all local gradients vanish except for the unique nonzero noise. Thus induction over the preceding local and server updates gives $x_t = x_{i,k}^t = 0$ and $q_t = 1$ until the exceptional noise occurs. In particular, it follows that $x_j = 0$ and $q_j = 1$. If the nonzero noise occurs at client i and has sign $s \in \{-1, 1\}$, then its occurrence at local step $K-1$ gives

$$\bar{g}_j^{(i)} = \frac{sM}{K}, \qquad \bar{g}_j^{(i')} = 0 \quad \forall i' \neq i.$$

Consequently, the server update (Core-FP) yields

$$\bar{g}_j = \frac{sM}{mK} = 16s\sqrt{T}, \qquad x_{j+1} = -16s\sqrt{T}, \qquad q_{j+1} = \sqrt{q_j^2 + |\bar{g}_j|^2} = \sqrt{1+256T}.$$

All later noises also vanish on $\mathcal{E}_T$. Hence, for every $t \geq j+1$,

$$|\bar{g}_t| \leq \frac{1}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}|F'(x_{i,k}^t)| \leq 1, \qquad q_t \geq q_{j+1} \geq 16\sqrt{T}.$$

Unrolling the server update now gives, for $j+1 \leq t \leq T+1$,

$$|x_t - x_{j+1}| \leq \sum_{\ell=j+1}^{t-1}\frac{|\bar{g}_\ell|}{q_\ell} \leq \frac{t-j-1}{16\sqrt{T}} \leq \frac{\sqrt{T}}{16}.$$

It follows that

$$|x_t| \geq |x_{j+1}| - |x_t - x_{j+1}| \geq 16\sqrt{T} - \frac{\sqrt{T}}{16} = \frac{255}{16}\sqrt{T} \geq 8\sqrt{T}, \tag{11}$$

which also implies that x_t has the same sign as x_{j+1}. Since $j \leq 3T/4$ and $|F'(x)|^2 = x^2/(1+x^2) \geq 1/2$ for $|x| \geq 1$, we further have

$$\frac{1}{T}\sum_{t=1}^{T}|F'(x_t)|^2 \geq \frac{1}{T}\sum_{t=j+1}^{T}|F'(x_t)|^2 \geq \frac{T-j}{T}\frac{1}{2} \geq \frac{1}{8},$$

Moreover, since $x_t = 0$ for $t \leq j$ and the remaining ones have the same sign, it follows from $j \leq 3T/4$, $F^\star = F(0) = 0$, and (11) that

$$\begin{aligned}\sum_{t=1}^{T}F(x_t) - F^\star &= \sum_{t=j+1}^{T}F(x_t) = \sum_{t=j+1}^{T}(\sqrt{1+x_t^2} - 1) \geq \sum_{t=j+1}^{T}(|x_t| - 1) \\ &\geq \sum_{t=j+1}^{T}(8\sqrt{T} - 1) \geq (T-j)4\sqrt{T} \geq T^{3/2}.\end{aligned}$$

Therefore, we conclude from (10) that

$$\mathbb{E}[|F'(\bar{x}_T)|^2] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[|F'(x_t)|^2] \geq \frac{1}{8}\mathbb{P}(\mathcal{E}_T) \geq \frac{1}{16384mK^2},$$

$$\mathbb{E}[F(\bar{x}_T) - F^\star] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F(x_t) - F^\star] \geq \sqrt{T}\,\mathbb{P}(\mathcal{E}_T) \geq \frac{\sqrt{T}}{2048mK^2},$$

which completes the proof. □

A.3 Additional notations for analysis

Write $g_t := \nabla F(x_t)$ and define

$$\hat{b}_t := \frac{1}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\big(\nabla F_i(x_{i,k}^t) - \nabla F_i(x_t)\big),$$

$$\bar{\xi}_t := \frac{1}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\xi_{i,k}^t, \qquad b_t := \bar{g}_t - g_t = \hat{b}_t + \bar{\xi}_t.$$

Fix any $\tau > 0$, we choose the local-path scale according to

$$K = 1 : \tau_t = \tau,$$
$$K \geq 2 : \tau_t \leq \begin{cases} \dfrac{\tau}{(K-1)t}, & \text{smooth nonconvex or Lipschitz convex,} \\ \dfrac{\tau}{(K-1)t^p},\ \forall p > \dfrac{1}{2}, & \text{smooth convex.} \end{cases} \tag{12}$$

In particular, we can always choose $\tau_t \leq \frac{1}{(K-1)t}$ for $K \geq 2$, which is exactly (5). Under (12), we have that $\{\tau_t\}_{t\geq 1}$ is upper bounded by τ. Let

$$c_K := \frac{1}{K}\sum_{k=0}^{K-1}k^2 = \frac{(K-1)(2K-1)}{6}, \quad \Theta := \left(1 + \frac{L\tau}{q_1}\right)^{K-1}$$

and

$$B^2 := \frac{c_K\Theta^2L^2}{m}\sum_{i=1}^{m}\|\nabla F_i(x_1)\|^2, \qquad \widetilde{B}^2 := \frac{c_K\Theta^2L^2}{m}\sum_{i=1}^{m}\big(\|\nabla F_i(x_1)\| + \sigma\big)^2.$$

For $t \geq 1$, denote the finite sums of τ_k and τ_k^2 by

$$\mathcal{T}_t := \sum_{k=1}^{t}\tau_k, \qquad \widetilde{\mathcal{T}}_t := \sum_{k=1}^{t}\tau_k^2.$$

If $K \geq 2$ and τ_t is chosen according to (12), then

$$c_K \lesssim (K-1)^2, \quad \Theta \lesssim 1, \quad B \lesssim K-1, \quad \widetilde{B} \lesssim K-1, \quad \widetilde{\mathcal{T}}_t \lesssim \frac{1}{(K-1)^2}, \tag{13}$$

where we use $(1+1/t)^t \leq e$ to obtain $\Theta \lesssim 1$. Moreover, for Lipschitz convex objectives, we have $\mathcal{T}_t \lesssim \log t/(K-1)$.

Let $\zeta_t := \bar{g}_t - \mathbb{E}_t[\bar{g}_t]$. Set

$$\mathcal{A}_T := \eta^2 + 2R_0^2 + 2R_1^2\sigma^2(\log(mKT) + 2).$$

Given $T \geq 1$, let $M_T \geq 0$ be the maximum noise norm over the first T rounds:

$$M_T := \max_{1\leq t\leq T,\, 1\leq i\leq m,\, 0\leq k<K}\|\xi_{i,k}^t\|.$$

Let $r_t := x_{t+1} - x_t$. Define the accumulated trajectory quantities:

$$\mathcal{R}_T := \sum_{t=1}^{T}\|r_t\|^2, \qquad \mathcal{G}_T := \sum_{t=1}^{T}\frac{\|g_t\|^2}{q_t}.$$

We focus on the nontrivial case $K \geq 2$ in the proofs. The special case $K = 1$ can be easily proved by noting that the coefficient of τ_t always involve $(K-1)$ (possibly through quantities c_K, B, and $\widetilde{B}$) in the analysis, which means that τ_t can be chosen arbitrarily for $K = 1$.

B SMOOTH NONCONVEX OBJECTIVES

Assume throughout this section that every F_i is L-smooth. From unbiasedness of $\xi_{i,k}^t$, we have $\mathbb{E}_t[\bar{\xi}_t] = 0$ and hence

$$\mathbb{E}_t[\bar{g}_t] = g_t + \mathbb{E}_t[\hat{b}_t]. \tag{14}$$

Additionally, the independence among the local noises $\{\xi_{i,k}^t\}$ gives

$$\mathbb{E}_t[\|\bar{\xi}_t\|^2] = \frac{1}{m^2K^2}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}_t[\|\xi_{i,k}^t\|^2] \le \frac{mK\sigma^2}{m^2K^2} = \frac{\sigma^2}{mK}. \tag{15}$$

Finally, since $\bar{g}_t = g_t + \hat{b}_t + \bar{\xi}_t$, we apply $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ to get

$$\mathbb{E}_t[\|\bar{g}_t\|^2] \le 2\mathbb{E}_t[\|g_t + \bar{\xi}_t\|^2] + 2\mathbb{E}_t[\|\hat{b}_t\|^2] = 2\|g_t\|^2 + 2\mathbb{E}_t[\|\bar{\xi}_t\|^2] + 2\mathbb{E}_t[\|\hat{b}_t\|^2],$$

where the equality follows from $\mathbb{E}_t[\langle g_t, \bar{\xi}_t\rangle] = \langle g_t, \mathbb{E}_t[\bar{\xi}_t]\rangle = 0$. This together with (15) gives

$$\mathbb{E}_t[\|\bar{g}_t\|^2] \le 2\|g_t\|^2 + 2\mathbb{E}_t[\|\hat{b}_t\|^2] + \frac{2\sigma^2}{mK} \tag{16}$$

Before presenting the main theorem and its proof, we first introduce some auxiliary lemmas that provide bounds for trajectory-related quantities.

Lemma B.1. *Let Assumptions 4.1 and 4.2 hold. Assume each F_i is L-smooth. Let $\kappa_T := \max_{t\in[T]}(t-1)\tau_t$. Then for every $T \ge 1$,*

$$\sum_{t=1}^{T}\mathbb{E}_t[\|\hat{b}_t\|^2] \le \frac{2\widetilde{B}^2\widetilde{\mathcal{T}}_T}{q_1^2} + \frac{8c_K\Theta^2L^4\kappa_T^2}{q_1^2}\mathcal{R}_T.$$

Proof. We start by using Lemma A.6 with $\mathcal{G} = \mathcal{F}_t$ and L-smoothness to obtain

$$\begin{aligned}\sqrt{\mathbb{E}_t[\|\nabla F_i(x_{i,k}^t)\|^2]} &\le \|\nabla F_i(x_t)\| + \sqrt{\mathbb{E}_t[\|\nabla F_i(x_{i,k}^t) - \nabla F_i(x_t)\|^2]} \\ &\le \|\nabla F_i(x_t)\| + L\sqrt{\mathbb{E}_t[\|x_{i,k}^t - x_t\|^2]}.\end{aligned} \tag{17}$$

Applying Lemma A.6, (17), and (2) gives

$$\begin{aligned}\sqrt{\mathbb{E}_t[\|x_{i,k+1}^t - x_t\|^2]} &= \sqrt{\mathbb{E}_t\Big[\|x_{i,k}^t - x_t - \tau_t(\nabla F_i(x_{i,k}^t) + \xi_{i,k}^t)/q_t\|^2\Big]} \\ &\le \sqrt{\mathbb{E}_t[\|x_{i,k}^t - x_t\|^2]} + \frac{\tau_t}{q_t}\sqrt{\mathbb{E}_t[\|\nabla F_i(x_{i,k}^t) + \xi_{i,k}^t\|^2]} \\ &\le \sqrt{\mathbb{E}_t[\|x_{i,k}^t - x_t\|^2]} + \frac{\tau_t}{q_t}\Big(\sqrt{\mathbb{E}_t[\|\nabla F_i(x_{i,k}^t)\|^2]} + \sqrt{\mathbb{E}_t[\|\xi_{i,k}^t\|^2]}\Big) \\ &\le \left(1 + \frac{L\tau_t}{q_t}\right)\sqrt{\mathbb{E}_t[\|x_{i,k}^t - x_t\|^2]} + \frac{\tau_t}{q_t}\big(\|\nabla F_i(x_t)\| + \sigma\big).\end{aligned}$$

Since $x_{i,0}^t = x_t$, applying the above inequality recursively and then using $\tau_t \le \tau$ and $q_t \ge q_1$ yields

$$\begin{aligned}\sqrt{\mathbb{E}_t[\|x_{i,k}^t - x_t\|^2]} &\le \frac{\tau_t}{q_t}\big(\|\nabla F_i(x_t)\| + \sigma\big)\sum_{s=0}^{k-1}\left(1 + \frac{L\tau_t}{q_t}\right)^s \\ &\le \frac{\tau_t}{q_t}\big(\|\nabla F_i(x_t)\| + \sigma\big)k\left(1 + \frac{L\tau}{q_1}\right)^{k-1} \le \frac{k\tau_t\Theta}{q_t}\big(\|\nabla F_i(x_t)\| + \sigma\big).\end{aligned} \tag{18}$$

Then it follows from the convexity of $\|\cdot\|^2$, the L-smoothness of F_i, and (18) that

$$\begin{aligned}\mathbb{E}_t[\|\hat{b}_t\|^2] &\le \frac{1}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}_t[\|\nabla F_i(x_{i,k}^t) - \nabla F_i(x_t)\|^2] \le \frac{L^2}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}_t[\|x_{i,k}^t - x_t\|^2] \\ &\le \frac{\Theta^2L^2\tau_t^2}{mq_t^2K}\sum_{i=1}^{m}\sum_{k=0}^{K-1}k^2\big(\|\nabla F_i(x_t)\| + \sigma\big)^2 = \frac{c_K\Theta^2L^2\tau_t^2}{mq_t^2}\sum_{i=1}^{m}\big(\|\nabla F_i(x_t)\| + \sigma\big)^2.\end{aligned}$$

To proceed, applying $\|\nabla F_i(x_t)\| \le \|\nabla F_i(x_1)\| + L\|x_t - x_1\|$ into the above inequality and using $(a+b)^2 \le 2a^2 + 2b^2$ yield

$$\begin{aligned}\mathbb{E}_t[\|\hat{b}_t\|^2] &\le \frac{c_K \Theta^2 L^2 \tau_t^2}{m q_t^2} \sum_{i=1}^{m} \left(\|\nabla F_i(x_1)\| + \sigma + L\|x_t - x_1\|\right)^2 \\ &\le \frac{2\tau_t^2}{q_t^2}\left(\widetilde{B}^2 + c_K \Theta^2 L^4 \|x_t - x_1\|^2\right).\end{aligned} \tag{19}$$

For $2 \le t \le T$, note that $x_t - x_1 = \sum_{j=1}^{t-1} r_j$, hence we have from $\tau_t \le \kappa_T/(t-1)$ and $q_t \ge q_j$ for $j \le t$ that

$$\begin{aligned}\sum_{t=1}^{T} \frac{\tau_t^2}{q_t^2}\|x_t - x_1\|^2 = \sum_{t=2}^{T} \frac{\tau_t^2}{q_t^2}\left\|\sum_{j=1}^{t-1} r_j\right\|^2 &\le \kappa_T^2 \sum_{t=2}^{T}\left(\frac{1}{t-1}\sum_{j=1}^{t-1}\frac{\|r_j\|}{q_j}\right)^2 \\ &\le 4\kappa_T^2 \sum_{j=1}^{T-1}\frac{\|r_j\|^2}{q_j^2} \le \frac{4\kappa_T^2}{q_1^2}\sum_{t=1}^{T}\|r_t\|^2,\end{aligned} \tag{20}$$

where the second inequality holds from Hardy's inequality $\sum_{t=1}^{T} \bar{a}_t^2 \le 4\sum_{t=1}^{T} a_t^2$ with $\bar{a}_t$ being the average of nonnegative $a_1, a_2, \ldots, a_t$. Summing (19) over $t = 1, 2, \ldots, T$ and applying (20) then gives

$$\sum_{t=1}^{T}\mathbb{E}_t[\|\hat{b}_t\|^2] \le \frac{2\widetilde{B}^2}{q_1^2}\sum_{t=1}^{T}\tau_t^2 + 2c_K\Theta^2L^4\sum_{t=1}^{T}\frac{\tau_t^2}{q_t^2}\|x_t - x_1\|^2 \le \frac{2\widetilde{B}^2\widetilde{\mathcal{T}}_T}{q_1^2} + \frac{8c_K\Theta^2L^4\kappa_T^2}{q_1^2}\sum_{t=1}^{T}\|r_t\|^2,$$

which completes the proof. □

Before introducing the next lemma, we first bound moments $\mathbb{E}[M_T^{2p}]$ for any $p \ge 1$. Suppose $\sigma > 0$, let $N := mKT$. Reindex the noises $\{\xi_{i,k}^t\}$ as $\xi_1, \ldots, \xi_N$, and denote

$$Z_j := \frac{\|\xi_j\|^2}{\sigma^2} \ge 0, \qquad A_N := \frac{1}{N}\sum_{j=1}^{N}\exp(Z_j) \ge 1.$$

Then it follows from Assumption 4.2 that $\mathbb{E}[\exp(Z_j)] \le \exp(1)$ for every j, and hence $\mathbb{E}[A_N] \le \exp(1)$. Moreover,

$$\frac{M_T^2}{\sigma^2} = \max_{1\le j\le N} Z_j \le \log\left(\sum_{j=1}^{N}\exp(Z_j)\right) = \log N + \log A_N. \tag{21}$$

For $p \ge 1$, Lemma A.5 with $a = p - 1 + \log N$ shows that the function $x \mapsto (p - 1 + \log N + \log x)^p$ is concave on $[1/N, \infty)$, hence on $[1, \infty)$. Applying Jensen's inequality on (21) therefore yields

$$\frac{\mathbb{E}[M_T^{2p}]}{\sigma^{2p}} \le \mathbb{E}\left[(p - 1 + \log N + \log A_N)^p\right] \le (p - 1 + \log N + \log \mathbb{E}[A_N])^p \le (\log(mKT) + p)^p,$$

which is equivalent to

$$\left(\mathbb{E}[M_T^{2p}]\right)^{1/p} \le \sigma^2\big(\log(mKT) + p\big), \quad \forall p \ge 1. \tag{22}$$

Note that (22) is a consequence of Assumption 4.2, which holds regardless of the problem settings.

Lemma B.2. *Let Assumption 4.2 hold. Assume each F_i is L-smooth. If $\rho_\tau := \sup_{t\ge1}\sqrt{t-1}\,\tau_t < \infty$, define deterministic constants*

$$\begin{aligned}R_0 &:= \max\left\{\frac{\|g_1\|}{q_1}, \eta\left(1 + \frac{L}{q_1}\right)\right\} + \frac{2B\tau + 2\eta\sqrt{c_K}\,\Theta L^2\rho_\tau}{q_1^2}, \\ R_1 &:= \frac{2}{q_1}\left(1 + \frac{\sqrt{c_K}\,\Theta L\tau}{q_1}\right).\end{aligned}$$

Then for every $T \geq 1$, we have $\max_{1\leq t\leq T}\|r_t\| \leq R_0 + R_1 M_T$ almost surely and

$$\mathbb{E}[\mathcal{R}_T] \leq 2\mathcal{A}_T\left(1+\log\frac{\mathbb{E}[q_{T+1}]}{q_1}\right), \tag{23}$$

$$\sqrt{\mathbb{E}[\mathcal{R}_T^2]} \leq 2\big(\mathcal{A}_T + 4R_1^2\sigma^2\big)\left(4+\log\frac{\mathbb{E}[q_{T+1}]}{q_1}\right). \tag{24}$$

Proof. We first establish a deterministic bound on $\|b_t\|$. Using smoothness and definition of M_T, we obtain

$$\begin{aligned}\|x_{i,k+1}^t - x_t\| &\leq \|x_{i,k}^t - x_t\| + \frac{\tau_t}{q_t}\big(\|\nabla F_i(x_{i,k}^t)\| + M_T\big)\\ &\leq \|x_{i,k}^t - x_t\| + \frac{\tau_t}{q_t}\big(\|\nabla F_i(x_t)\| + L\|x_{i,k}^t - x_t\| + M_T\big)\\ &= \left(1+\frac{L\tau_t}{q_t}\right)\|x_{i,k}^t - x_t\| + \frac{\tau_t}{q_t}\big(\|\nabla F_i(x_t)\| + M_T\big).\end{aligned}$$

Unrolling this recursion similarly as in the proof of Lemma B.1 gives (with σ replaced by M_T), for $t \leq T$,

$$\|x_{i,k}^t - x_t\| \leq \frac{\tau_t}{q_t}\big(\|\nabla F_i(x_t)\| + M_T\big)\sum_{s=0}^{k-1}\left(1+\frac{L\tau_t}{q_t}\right)^s \leq \frac{k\tau_t\Theta}{q_t}\big(\|\nabla F_i(x_t)\| + M_T\big),$$

$$\|\hat{b}_t\|^2 \leq \frac{c_K\Theta^2L^2\tau_t^2}{mq_t^2}\sum_{i=1}^m\big(\|\nabla F_i(x_t)\| + M_T\big)^2.$$

Then by applying Lemma A.6 and the smoothness of F_i, we further have

$$\begin{aligned}\|\hat{b}_t\| &\leq \frac{\tau_t L\Theta\sqrt{c_K}}{q_t}\Big[\frac{1}{m}\sum_{i=1}^m\big(\|\nabla F_i(x_t)\| + M_T\big)^2\Big]^{1/2}\\ &\leq \frac{\tau_t L\Theta\sqrt{c_K}}{q_t}\Big[\frac{1}{m}\sum_{i=1}^m\big(\|\nabla F_i(x_1)\| + L\|x_t - x_1\| + M_T\big)^2\Big]^{1/2}\\ &\leq \frac{\tau_t L\Theta\sqrt{c_K}}{q_t}\Big[\Big(\frac{1}{m}\sum_{i=1}^m\|\nabla F_i(x_1)\|^2\Big)^{1/2} + L\|x_t - x_1\| + M_T\Big]\\ &= \frac{\tau_t}{q_t}\left(B + \sqrt{c_K}\,\Theta L M_T + \sqrt{c_K}\,\Theta L^2\|x_t - x_1\|\right).\end{aligned} \tag{25}$$

To bound $\|x_t - x_1\|$, we use the update rule (Core-FP) and $q_t \geq q_1$ to write

$$\|x_t - x_1\|^2 = \Big\|\sum_{j=1}^{t-1} r_j\Big\|^2 \leq (t-1)\sum_{j=1}^{t-1}\|r_j\|^2 \leq \frac{t-1}{q_1^2}\sum_{j=1}^{t-1}\|\bar{g}_j\|^2 = \frac{(t-1)\eta^2}{q_1^2}(q_t^2 - q_1^2),$$

which further implies

$$\|x_t - x_1\| \leq \frac{\eta\sqrt{t-1}}{q_1}\sqrt{q_t^2 - q_1^2} \leq \frac{\eta\sqrt{t-1}}{q_1}q_t.$$

Substituting the above inequality into (25) and noting that $b_t = \hat{b}_t + \bar{\xi}_t$ and $\|\bar{\xi}_t\| \leq M_T$ gives

$$\|b_t\| \leq \|\hat{b}_t\| + \|\bar{\xi}_t\| \leq \frac{B\tau_t + \eta\sqrt{c_K}\,\Theta L^2\sqrt{t-1}\,\tau_t}{q_1} + \left(1+\frac{\sqrt{c_K}\,\Theta L\tau_t}{q_1}\right)M_T. \tag{26}$$

Next, we show that $\|r_t\| \leq R_0 + R_1 M_T$ for all $1 \leq t \leq T$. For $t = 1$, using $\bar{g}_1 = g_1 + b_1$, (26), and $\tau_t \leq \tau$ gives

$$\|r_1\| \leq \frac{\|g_1\| + \|b_1\|}{q_1} \leq \frac{\|g_1\|}{q_1} + \frac{B\tau}{q_1^2} + \frac{1}{q_1}\left(1+\frac{\sqrt{c_K}\,\Theta L\tau}{q_1}\right)M_T \leq R_0 + R_1 M_T.$$

For $t \geq 2$, note that $q_t^2 \geq \|\bar{g}_{t-1}\|^2/\eta^2$ and the smoothness of F implies:

$$\frac{\|g_t - g_{t-1}\|}{q_t} \leq \frac{L\|x_t - x_{t-1}\|}{q_t} = \frac{L\|\bar{g}_{t-1}\|}{q_{t-1}q_t} \leq \frac{L\eta}{q_1}.$$

Consequently, we have from the decomposition $\bar{g}_t = \bar{g}_{t-1} + (g_t - g_{t-1}) + (b_t - b_{t-1})$ and (26) that

$$\begin{aligned}
\|r_t\| = \frac{\|\bar{g}_t\|}{q_t} \leq& \frac{\|\bar{g}_{t-1}\| + \|g_t - g_{t-1}\| + \|b_t\| + \|b_{t-1}\|}{q_t} \\
\leq& \eta\left(1 + \frac{L}{q_1}\right) + \frac{\|b_t\| + \|b_{t-1}\|}{q_1} \\
\leq& \eta\left(1 + \frac{L}{q_1}\right) + \frac{2B\tau + 2\eta\sqrt{c_K}\,\Theta L^2\rho_\tau}{q_1^2} + \frac{2}{q_1}\left(1 + \frac{\sqrt{c_K}\,\Theta L\tau}{q_1}\right) M_T \\
\leq& R_0 + R_1 M_T.
\end{aligned}$$

We next bound the moments of $\mathcal{R}_T$. Since $\|r_t\| \leq R_0 + R_1 M_T$ for $t \leq T$ almost surely, we have

$$\frac{\|\bar{g}_t\|^2/\eta^2}{q_t^2} = \frac{\|r_t\|^2}{\eta^2} \leq \frac{(R_0 + R_1 M_T)^2}{\eta^2}.$$

Therefore, we apply the log-sum in Lemma A.4 with $a_t = \|r_t\|^2/\eta^2$, $A_t = q_t^2$, and $c = (R_0 + R_1 M_T)^2/\eta^2$ to obtain

$$\mathcal{R}_T = \eta^2 \sum_{t=1}^{T} \frac{\|\bar{g}_t\|^2/\eta^2}{q_t^2} \leq \left(\eta^2 + (R_0 + R_1 M_T)^2\right) \log \frac{q_{T+1}^2}{q_1^2}. \tag{27}$$

Taking expectations and applying Cauchy–Schwarz inequality further gives

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_T] \leq& 2\sqrt{\mathbb{E}\left[\left(\eta^2 + 2R_0^2 + 2R_1^2 M_T^2\right)^2\right]}\sqrt{\mathbb{E}\left[\left(\log \frac{q_{T+1}}{q_1}\right)^2\right]} \\
\leq& 2\left(\eta^2 + 2R_0^2 + 2R_1^2\sqrt{\mathbb{E}[M_T^4]}\right)\sqrt{\mathbb{E}\left[\left(1 + \log \frac{q_{T+1}}{q_1}\right)^2\right]} \\
\leq& 2\left[\eta^2 + 2R_0^2 + 2R_1^2\sigma^2\left(\log(mKT) + 2\right)\right]\left(1 + \log \frac{\mathbb{E}[q_{T+1}]}{q_1}\right),
\end{aligned}$$

where the second inequality uses Lemma A.6 and $q_{T+1} \geq q_1$, and the last inequality follows from Lemma A.5 with $a = 1$ and $p = 2$, as well as (22) with $p = 2$. Squaring both sides of (27) gives

$$\mathcal{R}_T^2 \leq 4\left(\eta^2 + (R_0 + R_1 M_T)^2\right)^2 \left(\log \frac{q_{T+1}}{q_1}\right)^2.$$

Taking expectations and applying Cauchy–Schwarz inequality $\mathbb{E}[XY] \leq (\mathbb{E}[X^2])^{1/2}(\mathbb{E}[Y^2])^{1/2}$ further gives

$$\begin{aligned}
\sqrt{\mathbb{E}[\mathcal{R}_T^2]} \leq& 2\left(\mathbb{E}\left[\left(\eta^2 + (R_0 + R_1 M_T)^2\right)^2 \left(\log \frac{q_{T+1}}{q_1}\right)^2\right]\right)^{1/2} \\
\leq& 2\left(\mathbb{E}\left[\left(\eta^2 + (R_0 + R_1 M_T)^2\right)^4\right]\right)^{1/4} \left(\mathbb{E}\left[\left(\log \frac{q_{T+1}}{q_1}\right)^4\right]\right)^{1/4}.
\end{aligned} \tag{28}$$

For the first term on the RHS of (28), we have from $(a+b)^2 \leq 2a^2 + 2b^2$ and Lemma A.6 with $p = 4$ that

$$\begin{aligned}
\left(\mathbb{E}\left[\left(\eta^2 + (R_0 + R_1 M_T)^2\right)^4\right]\right)^{1/4} \leq& \eta^2 + 2R_0^2 + 2R_1^2\left(\mathbb{E}[M_T^8]\right)^{1/4} \\
\leq& \mathcal{A}_T + 4R_1^2\sigma^2,
\end{aligned} \tag{29}$$

where the last inequality follows from (22) with $p = 4$. To bound the other term in (28), we apply Lemma A.5 with $a = 4$ and $p = 4$ to derive

$$\mathbb{E}\left[\left(\log \frac{q_{T+1}}{q_1}\right)^4\right] \le \mathbb{E}\left[\left(4 + \log \frac{q_{T+1}}{q_1}\right)^4\right] \le \left(4 + \log \frac{\mathbb{E}[q_{T+1}]}{q_1}\right)^4. \tag{30}$$

Substituting (29) and (30) into (28) yields (24). The proof is complete. □

Lemma B.3. *Let Assumptions 4.1 and 4.2 hold. Assume each F_i is L-smooth. Let $\kappa_T := \max_{t\in[T]}(t-1)\tau_t$. Then for every $T \ge 1$,*

$$\mathbb{E}[\mathcal{G}_T] \le 2(F(x_1) - F^\star) + \frac{2\widetilde{B}^2\widetilde{\mathcal{T}}_T}{q_1^3} + \left(L + \frac{8c_K\Theta^2L^4\kappa_T^2}{q_1^3}\right)\mathbb{E}[\mathcal{R}_T], \tag{31}$$

$$\begin{aligned}\sqrt{\mathbb{E}[\mathcal{G}_T^2]} \le{}& 2(F(x_1) - F^\star) + \frac{2\widetilde{B}^2\widetilde{\mathcal{T}}_T}{q_1^3} + \left(L + \frac{8c_K\Theta^2L^4\kappa_T^2}{q_1^3}\right)\sqrt{\mathbb{E}[\mathcal{R}_T^2]} \\ &+ \frac{2\sigma}{\sqrt{mq_1}}\left(1 + \frac{(K-1)L\tau\Theta}{2q_1}\right)\sqrt{\mathbb{E}[\mathcal{G}_T]}.\end{aligned} \tag{32}$$

Proof. Define the noiseless local quantities at round t by

$$\widetilde{x}_{i,0}^t := x_t, \qquad \widetilde{x}_{i,k+1}^t := \widetilde{x}_{i,k}^t - \frac{\tau_t}{q_t}\nabla F_i(\widetilde{x}_{i,k}^t), \qquad \widetilde{g}_t^{(i)} := \frac{1}{K}\sum_{k=0}^{K-1}\nabla F_i(\widetilde{x}_{i,k}^t).$$

Then it follows that

$$x_{i,k+1}^t - \widetilde{x}_{i,k+1}^t = x_{i,k}^t - \widetilde{x}_{i,k}^t - \frac{\tau_t}{q_t}\big(\nabla F_i(x_{i,k}^t) - \nabla F_i(\widetilde{x}_{i,k}^t)\big) - \frac{\tau_t}{q_t}\xi_{i,k}^t.$$

Lemma A.6 with $\mathcal{G} = \mathcal{F}_t$, L-smoothness, and (2) therefore give

$$\begin{aligned}\sqrt{\mathbb{E}_t[\|x_{i,k+1}^t - \widetilde{x}_{i,k+1}^t\|^2]} &\le \sqrt{\mathbb{E}_t[\|x_{i,k}^t - \widetilde{x}_{i,k}^t\|^2]} \\ &\quad + \frac{\tau_t}{q_t}\sqrt{\mathbb{E}_t[\|\nabla F_i(x_{i,k}^t) - \nabla F_i(\widetilde{x}_{i,k}^t)\|^2]} + \frac{\tau_t}{q_t}\sqrt{\mathbb{E}_t[\|\xi_{i,k}^t\|^2]} \\ &\le \left(1 + \frac{L\tau_t}{q_t}\right)\sqrt{\mathbb{E}_t[\|x_{i,k}^t - \widetilde{x}_{i,k}^t\|^2]} + \frac{\tau_t\sigma}{q_t}.\end{aligned}$$

Since $x_{i,0}^t = \widetilde{x}_{i,0}^t = x_t$, applying the above inequality recursively gives

$$\sqrt{\mathbb{E}_t[\|x_{i,k}^t - \widetilde{x}_{i,k}^t\|^2]} \le \frac{\tau_t\sigma}{q_t}\sum_{s=0}^{k-1}\left(1 + \frac{L\tau_t}{q_t}\right)^s \le \frac{k\tau_t\Theta\sigma}{q_t} \le \frac{k\tau\Theta\sigma}{q_1}. \tag{33}$$

Note that $\bar{g}_t^{(i)} - \widetilde{g}_t^{(i)} = \frac{1}{K}\sum_{k=0}^{K-1}\big(\nabla F_i(x_{i,k}^t) - \nabla F_i(\widetilde{x}_{i,k}^t) + \xi_{i,k}^t\big)$, we obtain by applying Lemma A.6, smoothness, and (33) that

$$\begin{aligned}\sqrt{\mathbb{E}_t[\|\bar{g}_t^{(i)} - \widetilde{g}_t^{(i)}\|^2]} &\le \frac{1}{K}\sum_{k=0}^{K-1}\left(L\sqrt{\mathbb{E}_t[\|x_{i,k}^t - \widetilde{x}_{i,k}^t\|^2]} + \sqrt{\mathbb{E}_t[\|\xi_{i,k}^t\|^2]}\right) \\ &\le \frac{\sigma}{K}\sum_{k=0}^{K-1}\left(1 + \frac{kL\tau\Theta}{q_1}\right) = \sigma\left(1 + \frac{(K-1)L\tau\Theta}{2q_1}\right).\end{aligned}$$

For different clients $i \ne j$, conditional independence gives

$$\mathbb{E}_t\left[\left\langle \bar{g}_t^{(i)} - \mathbb{E}_t[\bar{g}_t^{(i)}], \bar{g}_t^{(j)} - \mathbb{E}_t[\bar{g}_t^{(j)}]\right\rangle\right] = 0.$$

Since $\widetilde{g}_t^{(i)}$ is $\mathcal{F}_t$-measurable, we therefore have

$$\begin{aligned}\mathbb{E}_t[\|\zeta_t\|^2] &= \frac{1}{m^2}\sum_{i=1}^m \mathbb{E}_t\left[\|\bar{g}_t^{(i)} - \mathbb{E}_t[\bar{g}_t^{(i)}]\|^2\right] \\ &= \frac{1}{m^2}\sum_{i=1}^m\left(\mathbb{E}_t\left[\|\bar{g}_t^{(i)} - \widetilde{g}_t^{(i)}\|^2\right] - \left\|\mathbb{E}_t[\bar{g}_t^{(i)}] - \widetilde{g}_t^{(i)}\right\|^2\right) \\ &\le \frac{1}{m^2}\sum_{i=1}^m \mathbb{E}_t\left[\|\bar{g}_t^{(i)} - \widetilde{g}_t^{(i)}\|^2\right] \le \frac{\sigma^2}{m}\left(1 + \frac{(K-1)L\tau\Theta}{2q_1}\right)^2.\end{aligned} \tag{34}$$

Note $\bar{g}_t = g_t + \mathbb{E}_t[\hat{b}_t] + \zeta_t$ from (14), we have from L-smoothness of F and $r_t = -\bar{g}_t/q_t$ that

$$\begin{aligned} F(x_{t+1}) &\le F(x_t) + \langle g_t, r_t\rangle + \frac{L}{2}\|r_t\|^2 \\ &= F(x_t) - \frac{\|g_t\|^2}{q_t} - \frac{\langle g_t, \mathbb{E}_t[\hat{b}_t]\rangle}{q_t} - \frac{\langle g_t, \zeta_t\rangle}{q_t} + \frac{L}{2}\|r_t\|^2 \\ &\le F(x_t) - \frac{\|g_t\|^2}{2q_t} + \frac{\mathbb{E}_t[\|\hat{b}_t\|^2]}{2q_t} - \frac{\langle g_t, \zeta_t\rangle}{q_t} + \frac{L}{2}\|r_t\|^2, \end{aligned}$$

where the last inequality uses Young's inequality and $\|\mathbb{E}_t[\hat{b}_t]\|^2 \le \mathbb{E}_t[\|\hat{b}_t\|^2]$. Summing the above inequality from $t = 1$ to T and rearranging, we get

$$\mathcal{G}_T \le 2(F(x_1) - F^\star) + \sum_{t=1}^T \frac{\mathbb{E}_t[\|\hat{b}_t\|^2]}{q_t} + L\mathcal{R}_T - 2\sum_{t=1}^T \frac{\langle g_t, \zeta_t\rangle}{q_t}. \tag{35}$$

Applying Lemma B.1 into (35) and using $q_t \ge q_1$ then implies

$$\mathcal{G}_T \le 2(F(x_1) - F^\star) + \frac{2\widetilde{B}^2\widetilde{\mathcal{T}}_T}{q_1^3} + \left(L + \frac{8c_K\Theta^2L^4\kappa_T^2}{q_1^3}\right)\mathcal{R}_T - 2\sum_{t=1}^T \frac{\langle g_t, \zeta_t\rangle}{q_t}. \tag{36}$$

Taking expectations and using $\mathbb{E}[\langle g_t, \zeta_t\rangle] = \mathbb{E}[\langle g_t, \mathbb{E}_t[\zeta_t]\rangle] = 0$ for all t proves (31). Additionally, for any two rounds $s < t$, we have

$$\mathbb{E}[\langle g_s, \zeta_s\rangle\langle g_t, \zeta_t\rangle] = \mathbb{E}[\langle g_s, \zeta_s\rangle\langle g_t, \mathbb{E}_t[\zeta_t]\rangle] = 0.$$

Therefore, we obtain from (34) that

$$\begin{aligned} &\mathbb{E}\left[\left|2\sum_{t=1}^T \frac{\langle g_t, \zeta_t\rangle}{q_t}\right|^2\right] = 4\sum_{t=1}^T \mathbb{E}\left[\frac{\langle g_t, \zeta_t\rangle^2}{q_t^2}\right] \le \frac{4}{q_1}\sum_{t=1}^T \mathbb{E}\left[\frac{\|g_t\|^2\|\zeta_t\|^2}{q_t}\right] \\ =&\frac{4}{q_1}\sum_{t=1}^T \mathbb{E}\left[\frac{\|g_t\|^2}{q_t}\mathbb{E}_t[\|\zeta_t\|^2]\right] \le \frac{4\sigma^2}{mq_1}\left(1 + \frac{(K-1)L\tau\Theta}{2q_1}\right)^2 \mathbb{E}[\mathcal{G}_T]. \end{aligned} \tag{37}$$

Finally, taking squares, expectations, and square roots on both sides of (36), then applying Lemma A.6 with $p = 2$ and (37) yields

$$\begin{aligned} \sqrt{\mathbb{E}[\mathcal{G}_T^2]} &\le 2(F(x_1) - F^\star) + \frac{2\widetilde{B}^2\widetilde{\mathcal{T}}_T}{q_1^3} + \left(L + \frac{8c_K\Theta^2L^4\kappa_T^2}{q_1^3}\right)\sqrt{\mathbb{E}[\mathcal{R}_T^2]} \\ &\quad + \sqrt{\mathbb{E}\left[\left|2\sum_{t=1}^T \frac{\langle g_t, \zeta_t\rangle}{q_t}\right|^2\right]} \\ &\le 2(F(x_1) - F^\star) + \frac{2\widetilde{B}^2\widetilde{\mathcal{T}}_T}{q_1^3} + \left(L + \frac{8c_K\Theta^2L^4\kappa_T^2}{q_1^3}\right)\sqrt{\mathbb{E}[\mathcal{R}_T^2]} \\ &\quad + \frac{2\sigma}{\sqrt{mq_1}}\left(1 + \frac{(K-1)L\tau\Theta}{2q_1}\right)\sqrt{\mathbb{E}[\mathcal{G}_T]}, \end{aligned}$$

which proves (32). □

B.1 Main Result and Proof

Leveraging

Theorem B.4 (Smooth nonconvex convergence)**.** *Let Assumptions 4.1 and 4.2 hold. Assume each F_i is L-smooth. Let $\{x_t\}_{t\ge1}$ be the iterates generated by Algorithm 1 with τ_t chosen as in* (12) *and arbitrary $\gamma, \eta > 0$. Then for any $T \ge 1$, we have*

$$\mathbb{E}[\|\nabla F(\bar{x}_T)\|^2] \lesssim \frac{\mathcal{C}_T^2}{T} + \frac{\sigma\mathcal{C}_T}{\sqrt{mKT}},$$

where $\mathcal{C}_T > 0$ is defined in (3)*.*

Proof of Theorem B.4. With $K \geq 2$, it follows from scheme (12) that

$$\kappa_T \lesssim \frac{1}{K-1}, \qquad \widetilde{\mathcal{T}}_T \lesssim \frac{1}{(K-1)^2}, \quad \forall T \geq 1.$$

We first bound $\mathbb{E}[q_{T+1}]$. To this end, we apply Lemma A.8 with $c = q_1^2$ and $Z_t = \|\bar{g}_t\|^2/\eta^2$ to obtain

$$\begin{aligned}
\mathbb{E}[q_{T+1}] &\leq \sqrt{2}\,\mathbb{E}\left[\sqrt{q_1^2 + \frac{1}{\eta^2}\sum_{t=1}^{T}\mathbb{E}_t[\|\bar{g}_t\|^2]}\right] \\
&\leq \sqrt{2}\,\mathbb{E}\left[\sqrt{q_1^2 + \frac{2}{\eta^2}\sum_{t=1}^{T}\|g_t\|^2 + \frac{2}{\eta^2}\sum_{t=1}^{T}\mathbb{E}_t[\|\hat{b}_t\|^2] + \frac{2T\sigma^2}{\eta^2 mK}}\right] \\
&\leq \sqrt{2}q_1 + \frac{2}{\eta}\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|g_t\|^2}\right] + \frac{2}{\eta}\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\mathbb{E}_t[\|\hat{b}_t\|^2]}\right] + \frac{2\sigma}{\eta}\sqrt{\frac{T}{mK}}, \qquad (38)
\end{aligned}$$

where the second inequality uses (16), and the last follows from the basic inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$.

Next, we bound the two expectation terms in (38). Since $q_t \leq q_{T+1}$ for all $t \leq T+1$, we apply Cauchy–Schwarz inequality to obtain

$$\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|g_t\|^2}\right] \leq \mathbb{E}\left[\sqrt{q_{T+1}\sum_{t=1}^{T}\frac{\|g_t\|^2}{q_t}}\right] = \mathbb{E}\left[\sqrt{q_{T+1}\mathcal{G}_T}\right] \leq \sqrt{\mathbb{E}[q_{T+1}]}\sqrt{\mathbb{E}[\mathcal{G}_T]}. \qquad (39)$$

For the other term, applying Lemma B.1 and concavity of $\sqrt{\cdot}$ gives

$$\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\mathbb{E}_t[\|\hat{b}_t\|^2]}\right] \leq \sqrt{\frac{2\widetilde{B}^2\widetilde{\mathcal{T}}_T}{q_1^2} + \frac{8c_K\Theta^2L^4\kappa_T^2}{q_1^2}\mathbb{E}[\mathcal{R}_T]}.$$

Substituting the above two bounds into (38), and applying $2\sqrt{ab} \leq a+b$ and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ yields

$$\begin{aligned}
\mathbb{E}[q_{T+1}] \leq& \sqrt{2}q_1 + \frac{2}{\eta}\sqrt{\mathbb{E}[q_{T+1}]\mathbb{E}[\mathcal{G}_T]} + \frac{2}{\eta}\sqrt{\frac{2\widetilde{B}^2\widetilde{\mathcal{T}}_T}{q_1^2} + \frac{8c_K\Theta^2L^4\kappa_T^2}{q_1^2}\mathbb{E}[\mathcal{R}_T]} + \frac{2\sigma}{\eta}\sqrt{\frac{T}{mK}} \\
\leq& \sqrt{2}q_1 + \frac{\mathbb{E}[q_{T+1}]}{2} + \frac{2}{\eta^2}\mathbb{E}[\mathcal{G}_T] + \frac{2\sqrt{2}\widetilde{B}\sqrt{\widetilde{\mathcal{T}}_T}}{\eta q_1} + \frac{4\sqrt{2c_K}\,\Theta L^2\kappa_T}{\eta q_1}\sqrt{\mathbb{E}[\mathcal{R}_T]} + \frac{2\sigma}{\eta}\sqrt{\frac{T}{mK}},
\end{aligned}$$

which together with $2\sqrt{\mathbb{E}[\mathcal{R}_T]} \leq 1 + \mathbb{E}[\mathcal{R}_T]$ further simplifies to

$$\mathbb{E}[q_{T+1}] \leq 2\sqrt{2}q_1 + \frac{4}{\eta^2}\mathbb{E}[\mathcal{G}_T] + \frac{4\sqrt{2}\widetilde{B}\sqrt{\widetilde{\mathcal{T}}_T}}{\eta q_1} + \frac{4\sqrt{2c_K}\,\Theta L^2\kappa_T}{\eta q_1}\left(1 + \mathbb{E}[\mathcal{R}_T]\right) + \frac{4\sigma}{\eta}\sqrt{\frac{T}{mK}}, \qquad (40)$$

Plugging (31) and (23) into (40) and rearranging the terms gives

$$\mathbb{E}[q_{T+1}] \leq c_1 + \frac{4\sigma}{\eta}\sqrt{\frac{T}{mK}} + 2c_2\mathcal{A}_T\left(1 + \log\frac{\mathbb{E}[q_{T+1}]}{q_1}\right), \qquad (41)$$

where the positive constants c_1 and c_2 are defined as (we use (13))

$$\begin{aligned}
c_1 &:= 2\sqrt{2}q_1 + \frac{8(F(x_1) - F^\star)}{\eta^2} + \frac{8\widetilde{B}^2\widetilde{\mathcal{T}}_T}{\eta^2 q_1^3} + \frac{4\sqrt{2}\widetilde{B}\sqrt{\widetilde{\mathcal{T}}_T}}{\eta q_1} + \frac{4\sqrt{2c_K}\,\Theta L^2\kappa_T}{\eta q_1} \lesssim 1, \\
c_2 &:= \frac{4}{\eta^2}\left(L + \frac{8c_K\Theta^2L^4\kappa_T^2}{q_1^3}\right) + \frac{4\sqrt{2c_K}\,\Theta L^2\kappa_T}{\eta q_1} \lesssim 1.
\end{aligned}$$

Solving the inequality (41) for $\mathbb{E}[q_{T+1}]$ using Lemma A.1 gives

$$\mathbb{E}[q_{T+1}] \le 2c_1 + \frac{8\sigma}{\eta}\sqrt{\frac{T}{mK}} + 4c_2\mathcal{A}_T \log\frac{4c_2\mathcal{A}_T}{q_1}. \tag{42}$$

Since $R_0 \lesssim 1$ and $R_1 \lesssim 1$ from (13), we have $\mathcal{A}_T \lesssim 1 + \sigma^2(1+\log(mKT))$ and hence it follows from (42) that

$$\begin{aligned}\mathbb{E}[q_{T+1}] \lesssim& 1 + \sigma\sqrt{\frac{T}{mK}} + \big(1+\sigma^2(1+\log(mKT))\big)\log\big(\exp(1)+\sigma^2(1+\log(mKT))\big)\\ \lesssim& 1 + \sigma\sqrt{\frac{T}{mK}} + \sigma^2(1+\log(mKT))\log\big(1+\sigma^2(1+\log(mKT))\big) = \mathcal{U}_T.\end{aligned} \tag{43}$$

Applying the bound (43) in (23) then gives

$$\mathbb{E}[\mathcal{R}_T] \le 2\mathcal{A}_T\left(1+\log\frac{\mathbb{E}[q_{T+1}]}{q_1}\right) \lesssim \big(1+\sigma^2(1+\log(mKT))\big)(1+\log\mathcal{U}_T) = \mathcal{C}_T,$$

applying which into (24) and (31) respectively yields

$$\begin{aligned}\sqrt{\mathbb{E}[\mathcal{R}_T^2]} \le& \big(\mathcal{A}_T + 4R_1^2\sigma^2\big)\left(4+\log\frac{\mathbb{E}[q_{T+1}]}{q_1}\right) \lesssim \mathcal{A}_T\left(1+\log\frac{\mathbb{E}[q_{T+1}]}{q_1}\right) \lesssim \mathcal{C}_T,\\ \mathbb{E}[\mathcal{G}_T] \lesssim& 1 + \mathbb{E}[\mathcal{R}_T] \lesssim 1 + \mathcal{C}_T \lesssim \mathcal{C}_T.\end{aligned} \tag{44}$$

Applying the above two bounds into (32) gives

$$\sqrt{\mathbb{E}[\mathcal{G}_T^2]} \lesssim 1 + \sqrt{\mathbb{E}[\mathcal{R}_T^2]} + \frac{\sigma}{\sqrt{m}}\sqrt{\mathbb{E}[\mathcal{G}_T]} \lesssim 1 + \mathcal{C}_T + \frac{\sigma}{\sqrt{m}}\sqrt{\mathcal{C}_T} \lesssim \mathcal{C}_T + \sigma\sqrt{\frac{\mathcal{C}_T}{m}} \lesssim \mathcal{C}_T,$$

where the last inequality uses $\sigma^2 \lesssim \mathcal{C}_T$ and $m > 1$.

To proceed, we obtain from the monotonicity of q_t and Cauchy–Schwarz inequality that

$$\sum_{t=1}^{T}\mathbb{E}[\|g_t\|^2] \le \mathbb{E}[q_{T+1}\mathcal{G}_T] \le \sqrt{\mathbb{E}[q_{T+1}^2]}\sqrt{\mathbb{E}[\mathcal{G}_T^2]} \lesssim \mathcal{C}_T\sqrt{\mathbb{E}[q_{T+1}^2]}. \tag{45}$$

It remains to control $\mathbb{E}[q_{T+1}^2]$. Combining (Core-FP) and (16) gives

$$\mathbb{E}[q_{T+1}^2] = q_1^2 + \frac{1}{\eta^2}\sum_{t=1}^{T}\mathbb{E}[\|\bar{g}_t\|^2] \le q_1^2 + \frac{2}{\eta^2}\sum_{t=1}^{T}\mathbb{E}[\|g_t\|^2] + \frac{2}{\eta^2}\sum_{t=1}^{T}\mathbb{E}[\|\hat{b}_t\|^2] + \frac{2T\sigma^2}{\eta^2 mK}. \tag{46}$$

Applying Lemma B.1 and using $\mathbb{E}[\mathcal{R}_T] \lesssim \mathcal{C}_T$ leads to

$$\sum_{t=1}^{T}\mathbb{E}[\|\hat{b}_t\|^2] \lesssim 1 + \mathbb{E}[\mathcal{R}_T] \lesssim \mathcal{C}_T.$$

Substituting the above inequality and (45) into (46) and noting that $1 \lesssim \sqrt{\mathbb{E}[q_{T+1}^2]}$ gives

$$\mathbb{E}[q_{T+1}^2] \lesssim 1 + \mathcal{C}_T\sqrt{\mathbb{E}[q_{T+1}^2]} + \mathcal{C}_T + \frac{2T\sigma^2}{\eta^2 mK} \lesssim \mathcal{C}_T\sqrt{\mathbb{E}[q_{T+1}^2]} + \frac{2T\sigma^2}{\eta^2 mK}.$$

Consequently, there exist constants $C > 0$ such that

$$\mathbb{E}[q_{T+1}^2] \le C\mathcal{C}_T\sqrt{\mathbb{E}[q_{T+1}^2]} + \frac{2T\sigma^2}{\eta^2 mK} \le \frac{\mathbb{E}[q_{T+1}^2]}{2} + \frac{C^2}{2}\mathcal{C}_T^2 + \frac{2T\sigma^2}{\eta^2 mK},$$

which simplifies to

$$\mathbb{E}[q_{T+1}^2] \lesssim \mathcal{C}_T^2 + \frac{T\sigma^2}{mK}. \tag{47}$$

Finally, substituting (47) into (45) and dividing by T gives

$$\mathbb{E}[\|\nabla F(\bar{x}_T)\|^2] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|g_t\|^2] \lesssim \frac{\mathcal{C}_T}{T}\sqrt{\mathbb{E}[q_{T+1}^2]} \le \frac{\mathcal{C}_T}{T}\Big(\mathcal{C}_T + \sigma\sqrt{\frac{T}{mK}}\Big) = \frac{\mathcal{C}_T^2}{T} + \frac{\sigma\mathcal{C}_T}{\sqrt{mKT}},$$

which completes the proof. □

Corollary B.5 (Deterministic smooth nonconvex convergence)**.** *Consider the same setting as in Theorem B.4, except that the local oracle is exact, i.e., $\sigma = 0$. Then for any $T \geq 1$, we have*

$$\mathbb{E}[\|\nabla F(\bar{x}_T)\|^2] = \frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\|^2 \lesssim \frac{1}{T}.$$

Proof. When $\sigma = 0$, we have from the definition (3) that

$$\mathcal{C}_T = 1 + \log \mathcal{U}_T = 1.$$

Then the result follows immediately from Theorem B.4 by plugging in $\mathcal{C}_T = 1$ and $\sigma = 0$. □

C SMOOTH CONVEX OBJECTIVES

Assume in this subsection that every F_i is convex and smooth, and let $x^\star$ minimize F. Thus all results from the smooth nonconvex analysis remain valid.

For convex objectives, define

$$\epsilon_t := \frac{1}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}_t\left[F_i(x_t) - F_i(x_{i,k}^t) - \langle \nabla F_i(x_{i,k}^t), x_t - x_{i,k}^t\rangle\right] \geq 0, \tag{48}$$

where $\nabla F_i(x_{i,k}^t) \in \partial F_i(x_{i,k}^t)$ denotes a subgradient if F_i is nonsmooth. We also denote

$$d_t := x_t - x^\star, \qquad D_T := \max_{1\leq t\leq T+1}\|d_t\|^2.$$

Lemma C.1. *Let Assumptions 4.1 and 4.2 hold. Assume each F_i is convex. Then, for every $t \geq 1$,*

$$F(y) \geq F(x_t) + \langle \mathbb{E}_t[\bar{g}_t], y - x_t\rangle - \epsilon_t, \quad \forall y \in \mathbb{R}^d. \tag{49}$$

If additionally each F_i is L-smooth, then

$$\epsilon_t \leq \frac{c_K\Theta^2 L\tau_t^2}{q_t^2}\left[\frac{1}{m}\sum_{i=1}^{m}\big(\|\nabla F_i(x^\star)\| + \sigma\big)^2 + L^2\|d_t\|^2\right]. \tag{50}$$

Proof. For each client, convexity of F_i at $x_{i,k}^t$ implies

$$\begin{aligned} F_i(y) &\geq F_i(x_{i,k}^t) + \langle \nabla F_i(x_{i,k}^t), y - x_{i,k}^t\rangle \\ &= F_i(x_t) + \langle \nabla F_i(x_{i,k}^t), y - x_t\rangle - \left[F_i(x_t) - F_i(x_{i,k}^t) - \langle \nabla F_i(x_{i,k}^t), x_t - x_{i,k}^t\rangle\right]. \end{aligned}$$

Then (49) follows by averaging this inequality over i, k and taking $\mathbb{E}_t$, and noting that

$$\frac{1}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}_t[\nabla F_i(x_{i,k}^t)] = \frac{1}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}_t[g_{i,k}^t - \xi_{i,k}^t] = \mathbb{E}_t[\bar{g}_t].$$

To prove (50), we use convexity and smoothness to obtain

$$0 \leq F_i(x_t) - F_i(x_{i,k}^t) - \langle \nabla F_i(x_{i,k}^t), x_t - x_{i,k}^t\rangle \leq \frac{L}{2}\|x_t - x_{i,k}^t\|^2,$$

which together with (18) and $K^{-1}\sum_{k=0}^{K-1}k^2 = c_K$ yields

$$0 \leq \epsilon_t \leq \frac{L}{2mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}_t[\|x_t - x_{i,k}^t\|^2] \leq \frac{c_K\Theta^2 L\tau_t^2}{2mq_t^2}\sum_{i=1}^{m}\big(\|\nabla F_i(x_t)\| + \sigma\big)^2.$$

Then (50) follows by plugging in the following inequality due to smoothness of F_i around $x^\star$ and $(a+b)^2 \leq 2a^2 + 2b^2$

$$\big(\|\nabla F_i(x_t)\| + \sigma\big)^2 \leq 2\big(\|\nabla F_i(x^\star)\| + \sigma\big)^2 + 2L^2\|x_t - x^\star\|^2.$$

The proof is complete. □

Lemma C.2. *Let Assumptions 4.1 and 4.2 hold. Assume each F_i is convex and L-smooth. If τ_t is chosen as in* (12)*, then for any $T \geq 1$, we have*

$$\sqrt{\mathbb{E}[D_T^2]} \lesssim \mathcal{C}_T,$$

where $\mathcal{C}_T > 0$ is defined in (3)*.*

Proof. Note that $d_{t+1} = d_t - \bar{g}_t/q_t$ from the update rules (Core-FP) and also $\bar{g}_t = \mathbb{E}_t[\bar{g}_t] + \zeta_t$, hence we have

$$\|d_{t+1}\|^2 = \left\| d_t - \frac{\bar{g}_t}{q_t} \right\|^2 = \|d_t\|^2 - \frac{2}{q_t}\langle \mathbb{E}_t[\bar{g}_t], d_t \rangle - \frac{2}{q_t}\langle \zeta_t, d_t \rangle + \|r_t\|^2.$$

Adding $2(F(x_t) - F^\star)/q_t$ to both sides of the above equality and applying $F(x_t) - F^\star \leq \langle \mathbb{E}_t[\bar{g}_t], d_t \rangle + \epsilon_t$ (which is (49) with $y = x^\star$) yields

$$\|d_{t+1}\|^2 + \frac{2(F(x_t) - F^\star)}{q_t} \leq \|d_t\|^2 + \|r_t\|^2 - \frac{2}{q_t}\langle \zeta_t, d_t \rangle + \frac{2\epsilon_t}{q_t}. \tag{51}$$

To bound the term $2\epsilon_t/q_t$, we apply (50) from Lemma C.1 and note $q_t \geq q_1$ to obtain

$$\begin{aligned} \frac{2\epsilon_t}{q_t} &\leq \frac{2c_K\Theta^2 L\tau_t^2}{q_t^3}\left[\frac{1}{m}\sum_{i=1}^m \big(\|\nabla F_i(x^\star)\| + \sigma\big)^2 + L^2\|d_t\|^2\right] \\ &\leq \frac{2c_K\Theta^2 L^3\tau_t^2}{q_1^3}\|d_t\|^2 + \frac{2c_K\Theta^2 L\tau_t^2}{mq_1^3}\sum_{i=1}^m \big(\|\nabla F_i(x^\star)\| + \sigma\big)^2. \end{aligned} \tag{52}$$

For ease of notation, denote the coefficients as

$$\alpha_t := \frac{2c_K\Theta^2 L^3\tau_t^2}{q_1^3}, \qquad \beta_t := \frac{2c_K\Theta^2 L\tau_t^2}{mq_1^3}\sum_{i=1}^m \big(\|\nabla F_i(x^\star)\| + \sigma\big)^2.$$

Then substituting (52) into (51) yields

$$\|d_{t+1}\|^2 + \frac{2(F(x_t) - F^\star)}{q_t} \leq (1+\alpha_t)\|d_t\|^2 + \|r_t\|^2 + \beta_t - \frac{2}{q_t}\langle \zeta_t, d_t \rangle. \tag{53}$$

To proceed, let $P_t := \prod_{j=1}^t (1+\alpha_j)$ with $P_0 := 1$, and

$$\mathcal{M}_t := -2\sum_{s=1}^t \frac{\langle \zeta_s, d_s \rangle}{P_s q_s}, \qquad 1 \leq t \leq T.$$

Since $\mathbb{E}_t[\zeta_t] = 0$, we have

$$\mathbb{E}_t[\mathcal{M}_t] = \mathcal{M}_{t-1} - \frac{2}{P_t q_t}\langle \mathbb{E}_t[\zeta_t], d_t \rangle = \mathcal{M}_{t-1},$$

which implies that $\{\mathcal{M}_t\}_{t=1}^T$ is a martingale. Since $(1+\alpha_t)/P_t = 1/P_{t-1}$, dividing both sides of (53) by P_t and unrolling gives

$$\frac{\|d_{t+1}\|^2}{P_t} \leq \frac{\|d_t\|^2}{P_{t-1}} + \frac{\|r_t\|^2 + \beta_t}{P_t} - \frac{2\langle \zeta_t, d_t \rangle}{P_t q_t} \leq \|d_1\|^2 + \sum_{s=1}^t \frac{\|r_s\|^2 + \beta_s}{P_s} + \mathcal{M}_t.$$

Then it follows from $1 \leq P_t \leq P_T$ that

$$D_T = \max_{1 \leq t \leq T+1} \|d_t\|^2 \leq P_T\left(\|d_1\|^2 + \sum_{t=1}^T \beta_t + \mathcal{R}_T + \max_{1 \leq t \leq T} |\mathcal{M}_t|\right),$$

which together with Lemma A.6 yields

$$\sqrt{\mathbb{E}[D_T^2]} \leq P_T\left(\|d_1\|^2 + \sum_{t=1}^T \beta_t + \sqrt{\mathbb{E}[\mathcal{R}_T^2]} + \sqrt{\mathbb{E}\left[\max_{1 \leq t \leq T} |\mathcal{M}_t|^2\right]}\right). \tag{54}$$

Next, we bound the last term in (54). Since $\zeta_t/q_t = -(r_t - \mathbb{E}_t[r_t])$, we have

$$\frac{\mathbb{E}_t[\|\zeta_t\|^2]}{q_t^2} = \mathbb{E}_t[\|r_t - \mathbb{E}_t[r_t]\|^2] = \mathbb{E}_t[\|r_t\|^2] - \|\mathbb{E}_t[r_t]\|^2 \le \mathbb{E}_t[\|r_t\|^2],$$

which together with Lemma A.7 and $\mathbb{E}[\langle g_s, \zeta_s\rangle\langle g_t, \zeta_t\rangle] = \mathbb{E}[\langle g_s, \zeta_s\rangle\langle g_t, \mathbb{E}_t[\zeta_t]\rangle] = 0,\ \forall s < t$ yield

$$\begin{aligned}
\mathbb{E}\left[\max_{1\le t\le T}|\mathcal{M}_t|^2\right] &\le 4\mathbb{E}[|\mathcal{M}_T|^2] = 16\sum_{t=1}^T \mathbb{E}\left[\frac{\langle\zeta_t, d_t\rangle^2}{P_t^2 q_t^2}\right] \le 16\sum_{t=1}^T \mathbb{E}\left[\frac{\|\zeta_t\|^2\|d_t\|^2}{q_t^2}\right] \\
&\le 16\sum_{t=1}^T \mathbb{E}\left[\|d_t\|^2\mathbb{E}_t[\|r_t\|^2]\right] = 16\sum_{t=1}^T \mathbb{E}\left[\|d_t\|^2\|r_t\|^2\right] \\
&\le 16\mathbb{E}[D_T\mathcal{R}_T] \le 16\sqrt{\mathbb{E}[D_T^2]}\sqrt{\mathbb{E}[\mathcal{R}_T^2]}. \qquad (55)
\end{aligned}$$

Applying (55) into (54) and using Young's inequality gives

$$\begin{aligned}
\sqrt{\mathbb{E}[D_T^2]} \le& P_T\left(\|d_1\|^2 + \sum_{t=1}^T \beta_t + \sqrt{\mathbb{E}[\mathcal{R}_T^2]} + 4\sqrt{\sqrt{\mathbb{E}[D_T^2]}\sqrt{\mathbb{E}[\mathcal{R}_T^2]}}\right) \\
\le& P_T\left(\|d_1\|^2 + \sum_{t=1}^T \beta_t\right) + P_T\sqrt{\mathbb{E}[\mathcal{R}_T^2]} + \frac{1}{2}\sqrt{\mathbb{E}[D_T^2]} + 8P_T^2\sqrt{\mathbb{E}[\mathcal{R}_T^2]}.
\end{aligned}$$

Additionally, we have from $1 + x \le e^x$, and $\sum_{t=1}^T \tau_t^2 = \widetilde{\mathcal{T}}_T$, and (13) that

$$\begin{aligned}
P_T &\le \exp\Big(\sum_{t=1}^T \alpha_t\Big) = \exp\left(\frac{2c_K\Theta^2L^3}{q_1^3}\widetilde{\mathcal{T}}_T\right) \lesssim 1, \\
\sum_{t=1}^T \beta_t &= \frac{2c_K\Theta^2L\widetilde{\mathcal{T}}_T}{mq_1^3}\sum_{i=1}^m\big(\|\nabla F_i(x^\star)\| + \sigma\big)^2 \lesssim 1.
\end{aligned}$$

Therefore, we conclude that

$$\sqrt{\mathbb{E}[D_T^2]} \le 2P_T\left(\|d_1\|^2 + \sum_{t=1}^T \beta_t\right) + (2P_T + 16P_T^2)\sqrt{\mathbb{E}[\mathcal{R}_T^2]} \lesssim 1 + \sqrt{\mathbb{E}[\mathcal{R}_T^2]} \lesssim \sqrt{\mathbb{E}[\mathcal{R}_T^2]}. \quad (56)$$

It remains to establish a refined bound on $\mathbb{E}[\mathcal{R}_T^2]$. Note that we may not directly apply the result (44) from the proof of Theorem B.4, since it relies on the stronger scheme $\tau_t \lesssim \frac{1}{(K-1)t}$. Instead, we start by plugging (39) into (38) and applying Young's inequality to obtain

$$\begin{aligned}
\mathbb{E}[q_{T+1}] \le& \sqrt{2}q_1 + \frac{2}{\eta}\sqrt{\mathbb{E}[q_{T+1}]\mathbb{E}[\mathcal{G}_T]} + \frac{2}{\eta}\mathbb{E}\left[\sqrt{\sum_{t=1}^T \mathbb{E}_t[\|\hat{b}_t\|^2]}\right] + \frac{2\sigma}{\eta}\sqrt{\frac{T}{mK}} \\
\le& \sqrt{2}q_1 + \frac{\mathbb{E}[q_{T+1}]}{2} + \frac{2}{\eta^2}\mathbb{E}[\mathcal{G}_T] + 1 + \frac{1}{\eta^2}\sum_{t=1}^T \mathbb{E}[\|\hat{b}_t\|^2] + \frac{2\sigma}{\eta}\sqrt{\frac{T}{mK}},
\end{aligned}$$

which simplifies to

$$\mathbb{E}[q_{T+1}] \le 2 + 2\sqrt{2}q_1 + \frac{4}{\eta^2}\mathbb{E}[\mathcal{G}_T] + \frac{2}{\eta^2}\sum_{t=1}^T \mathbb{E}[\|\hat{b}_t\|^2] + \frac{4\sigma}{\eta}\sqrt{\frac{T}{mK}}. \quad (57)$$

We next bound the two expectation terms on the RHS of (57). To bound $\sum_{t=1}^T \mathbb{E}[\|\hat{b}_t\|^2]$, we replace x_1 with $x^\star$ in (19) to obtain

$$\begin{aligned}
\sum_{t=1}^T \mathbb{E}[\|\hat{b}_t\|^2] \le& \sum_{t=1}^T \frac{c_K\Theta^2L^2\tau_t^2}{mq_t^2}\sum_{i=1}^m \mathbb{E}[(\|\nabla F_i(x_1)\| + \sigma + L\|x_t - x^\star\|)^2] \\
\le& \frac{2c_K\Theta^2L^2\widetilde{\mathcal{T}}_T}{q_1^2}\left[\frac{1}{m}\sum_{i=1}^m\big(\|\nabla F_i(x^\star)\| + \sigma\big)^2 + L^2\sqrt{\mathbb{E}[D_T^2]}\right] \\
\lesssim& 1 + \sqrt{\mathbb{E}[D_T^2]} \lesssim \sqrt{\mathbb{E}[\mathcal{R}_T^2]}, \qquad (58)
\end{aligned}$$

where the second inequality holds from $(a+b)^2 \le 2a^2 + 2b^2$ and $q_t \ge q_1$. Moreover, taking expectations on both sides of (35), then using (58) and $\mathbb{E}[\mathcal{R}_T] \le \sqrt{\mathbb{E}[\mathcal{R}_T^2]}$ gives

$$\mathbb{E}[\mathcal{G}_T] \le 2(F(x_1) - F^\star) + \frac{1}{q_1}\sum_{t=1}^{T}\mathbb{E}[\|\hat{b}_t\|^2] + L\mathbb{E}[\mathcal{R}_T] \lesssim 1 + \sqrt{\mathbb{E}[\mathcal{R}_T^2]} + \mathbb{E}[\mathcal{R}_T] \lesssim \sqrt{\mathbb{E}[\mathcal{R}_T^2]}. \quad (59)$$

Now substituting (58) and (59) into (57) gives

$$\mathbb{E}[q_{T+1}] \lesssim 1 + \sqrt{\mathbb{E}[\mathcal{R}_T^2]} + \sigma\sqrt{\frac{T}{mK}}. \quad (60)$$

Plugging (60) into (24) and noting $\mathcal{A}_T = \eta^2 + 2R_0^2 + 2R_1^2\sigma^2(\log(mKT)+2) \lesssim 1 + \sigma^2(\log(mKT)+1)$ implies

$$\begin{aligned}\sqrt{\mathbb{E}[\mathcal{R}_T^2]} &\le 2\big(\mathcal{A}_T + 4R_1^2\sigma^2\big)\left(4 + \log\frac{\mathbb{E}[q_{T+1}]}{q_1}\right)\\ &\lesssim \big(1 + \sigma^2(\log(mKT)+1)\big)\left(1 + \log\big(1 + \sqrt{\mathbb{E}[\mathcal{R}_T^2]} + \sigma\sqrt{T/(mK)}\big)\right).\end{aligned}$$

Let $y := \sqrt{\mathbb{E}[\mathcal{R}_T^2]}$, $s_T := \sigma\sqrt{T/(mK)}$, and $Q_T := 1 + \sigma^2(\log(mKT)+1)$. Then there exists a constant $C > 0$ such that

$$y \le CQ_T\big(1 + \log(1 + y + s_T)\big).$$

Since $1 + y + s_T \le (1+y)(1+s_T)$, we have

$$1 + y \le CQ_T\log(1+y) + 1 + CQ_T\big(1 + \log(1+s_T)\big).$$

Applying Lemma A.1 with $x = 1 + y$, $a = CQ_T$, and $b = 1 + CQ_T(1 + \log(1+s_T))$ gives

$$\begin{aligned}1 + y &\le 2CQ_T\log(2CQ_T) - 2CQ_T + 2 + 2CQ_T\big(1 + \log(1+s_T)\big)\\ &= 2 + 2CQ_T\big(\log(2CQ_T) + \log(1+s_T)\big),\end{aligned}$$

and hence

$$\sqrt{\mathbb{E}[\mathcal{R}_T^2]} = y \lesssim Q_T\big(1 + \log Q_T + \log(1+s_T)\big) \lesssim Q_T(1 + \log\mathcal{U}_T) = \mathcal{C}_T.$$

Finally, substituting this bound into (56) yields the desired result. □

C.1 Main Result and Proof

Theorem C.3 (Smooth convex convergence)**.** *Let Assumptions 4.1 and 4.2 hold. Assume each F_i is L-smooth and convex. Let $\{x_t\}_{t\ge1}$ be the iterates generated by Algorithm 1 with τ_t chosen as in* (12) *and arbitrary $\gamma, \eta > 0$. Then for any $T \ge 1$, we have*

$$\mathbb{E}[F(\bar{x}_T) - F^\star] \lesssim \frac{\mathcal{C}_T^2}{T} + \frac{\sigma\mathcal{C}_T}{\sqrt{mKT}},$$

where $\mathcal{C}_T > 0$ is defined in (3)*.*

Proof of Theorem C.3. Let $\Delta_t := F(x_t) - F^\star$. Rearranging (51) gives

$$\Delta_t \le \frac{q_t}{2}\big(\|d_t\|^2 - \|d_{t+1}\|^2\big) + \frac{q_t}{2}\|r_t\|^2 + \epsilon_t - \langle\zeta_t, d_t\rangle. \quad (61)$$

We proceed by bounding the summation of each term on the RHS of (61) in expectation. Moreover, we have from the monotonicity of q_t that

$$\sum_{t=1}^{T} q_t\big(\|d_t\|^2 - \|d_{t+1}\|^2\big) = q_1\|d_1\|^2 + \sum_{t=2}^{T}(q_t - q_{t-1})\|d_t\|^2 - q_T\|d_{T+1}\|^2 \le q_T D_T,$$

which together with Cauchy–Schwarz inequality and Lemma C.2 implies

$$\mathbb{E}\Big[\sum_{t=1}^{T} q_t\big(\|d_t\|^2 - \|d_{t+1}\|^2\big)\Big] \le \sqrt{\mathbb{E}[q_{T+1}^2]}\sqrt{\mathbb{E}[D_T^2]} \lesssim \mathcal{C}_T\sqrt{\mathbb{E}[q_{T+1}^2]}. \quad (62)$$

Since $(\|\bar{g}_t\|^2/\eta^2)/q_t^2 = \|r_t\|^2/\eta^2 \leq (R_0 + R_1 M_T)^2/\eta^2$ from Lemma B.2, we obtain from Lemma A.4 that

$$\sum_{t=1}^{T} q_t \|r_t\|^2 = \eta^2 \sum_{t=1}^{T} \frac{\|\bar{g}_t\|^2/\eta^2}{q_t} \leq \eta^2 \left(1 + \sqrt{1 + (R_0 + R_1 M_T)^2/\eta^2}\right)(q_{T+1} - q_1). \tag{63}$$

Let $B_T := 1 + \sqrt{1 + (R_0 + R_1 M_T)^2/\eta^2}$, then applying $(a+b)^2 \leq 2(a^2+b^2)$ twice implies

$$B_T^2 \leq 4 + 2(R_0 + R_1 M_T)^2/\eta^2 \leq 4 + \frac{4R_0^2}{\eta^2} + \frac{4R_1^2 M_T^2}{\eta^2}.$$

Note that $\mathbb{E}[M_T^2] \leq \sigma^2(\log(mKT) + 1)$ from (22). Therefore, it follows from $R_0 \lesssim 1$ and $R_1 \lesssim 1$ that

$$\sqrt{\mathbb{E}[B_T^2]} \leq \left(4 + \frac{4R_0^2}{\eta^2} + \frac{4R_1^2\sigma^2(\log(mKT)+1)}{\eta^2}\right)^{1/2} \lesssim 1 + \sigma\sqrt{\log(mKT)+1} \lesssim \mathcal{C}_T.$$

Taking expectations on both sides of (63) and plugging in the above bound gives

$$\mathbb{E}\left[\sum_{t=1}^{T} q_t \|r_t\|^2\right] \leq \eta^2 \sqrt{\mathbb{E}[B_T^2]}\sqrt{\mathbb{E}[(q_{T+1} - q_1)^2]} \lesssim \mathcal{C}_T \sqrt{\mathbb{E}[q_{T+1}^2]}. \tag{64}$$

Moreover, we have from (50), $q_t \geq q_1$, $\mathbb{E}[\|d_t\|^2] \leq \mathbb{E}[D_T] \leq \sqrt{\mathbb{E}[D_T^2]}$, and (13) that

$$\sum_{t=1}^{T} \mathbb{E}[\epsilon_t] \leq \frac{c_K \Theta^2 L \widetilde{\mathcal{T}}_T}{q_1^2}\left[\frac{1}{m}\sum_{i=1}^{m}\left(\|\nabla F_i(x^\star)\| + \sigma\right)^2 + L^2\sqrt{\mathbb{E}[D_T^2]}\right] \lesssim \mathcal{C}_T, \tag{65}$$

where the last inequality uses Lemma C.2. Note that the last term in (61) has zero expectation since $\mathbb{E}[\langle \zeta_t, d_t\rangle] = \mathbb{E}[\langle \mathbb{E}_t[\zeta_t], d_t\rangle] = 0$. Taking expectations on both sides of (61) and plugging in (62), (64), and (65) gives

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}[\Delta_t] &\leq \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T} q_t(\|d_t\|^2 - \|d_{t+1}\|^2)\right] + \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T} q_t\|r_t\|^2\right] + \sum_{t=1}^{T}\mathbb{E}[\epsilon_t] \\ &\lesssim \mathcal{C}_T\sqrt{\mathbb{E}[q_{T+1}^2]} + \mathcal{C}_T \lesssim \mathcal{C}_T\sqrt{\mathbb{E}[q_{T+1}^2]}.\end{aligned} \tag{66}$$

It remains to control $\mathbb{E}[q_{T+1}^2]$. To this end, we obtain from the update rule for q_{t+1} and (16) that

$$\begin{aligned}\mathbb{E}[q_{T+1}^2] &= q_1^2 + \frac{1}{\eta^2}\sum_{t=1}^{T}\mathbb{E}[\|\bar{g}_t\|^2] \\ &\leq q_1^2 + \frac{2}{\eta^2}\sum_{t=1}^{T}\mathbb{E}[\|\nabla F(x_t)\|^2] + \frac{2}{\eta^2}\sum_{t=1}^{T}\mathbb{E}[\|\hat{b}_t\|^2] + \frac{2T\sigma^2}{\eta^2 mK} \\ &\leq q_1^2 + \frac{4L}{\eta^2}\sum_{t=1}^{T}\mathbb{E}[\Delta_t] + \frac{2}{\eta^2}\sum_{t=1}^{T}\mathbb{E}[\|\hat{b}_t\|^2] + \frac{2T\sigma^2}{\eta^2 mK},\end{aligned} \tag{67}$$

where the last inequality uses the fact that $\|\nabla F(x_t)\|^2 \leq 2L\Delta_t$ for smooth convex functions. To bound $\sum_{t=1}^{T}\mathbb{E}[\|\hat{b}_t\|^2]$, we replace x_1 with $x^\star$ in (19) to obtain

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}[\|\hat{b}_t\|^2] &\leq \sum_{t=1}^{T}\frac{c_K\Theta^2 L^2\tau_t^2}{m q_t^2}\sum_{i=1}^{m}\left(\|\nabla F_i(x_1)\| + \sigma + L\|x_t - x^\star\|\right)^2 \\ &\leq \frac{2c_K\Theta^2L^2\widetilde{\mathcal{T}}_T}{q_1^2}\left[\frac{1}{m}\sum_{i=1}^{m}\left(\|\nabla F_i(x^\star)\| + \sigma\right)^2 + L^2\sqrt{\mathbb{E}[D_T^2]}\right] \lesssim \mathcal{C}_T,\end{aligned}$$

where the second inequality holds from $(a+b)^2 \leq 2a^2 + 2b^2$, $q_t \geq q_1$ and the last inequality uses Lemma C.2. Substituting the above bound and (66) into (67) leads to

$$\mathbb{E}[q_{T+1}^2] \lesssim 1 + \mathcal{C}_T\sqrt{\mathbb{E}[q_{T+1}^2]} + \mathcal{C}_T + \frac{2T\sigma^2}{\eta^2 mK} \lesssim \mathcal{C}_T\sqrt{\mathbb{E}[q_{T+1}^2]} + \frac{2T\sigma^2}{\eta^2 mK}.$$

Therefore, there exists a constant $C > 0$ such that

$$\mathbb{E}[q_{T+1}^2] \le C\mathcal{C}_T\sqrt{\mathbb{E}[q_{T+1}^2]} + \frac{2T\sigma^2}{\eta^2 mK} \le \frac{\mathbb{E}[q_{T+1}^2]}{2} + \frac{C^2\mathcal{C}_T^2}{2} + \frac{2T\sigma^2}{\eta^2 mK},$$

which further implies

$$\sqrt{\mathbb{E}[q_{T+1}^2]} \le \sqrt{C^2\mathcal{C}_T^2 + \frac{4T\sigma^2}{\eta^2 mK}} \lesssim \mathcal{C}_T + \sigma\sqrt{\frac{T}{mK}}.$$

Plugging the above inequality into (66) gives

$$\sum_{t=1}^{T}\mathbb{E}[\Delta_t] \lesssim \mathcal{C}_T^2 + \sigma\mathcal{C}_T\sqrt{\frac{T}{mK}}.$$

Then the proof is complete after dividing both sides of the above inequality by T. □

Corollary C.4 (Deterministic smooth convex convergence)**.** *Consider the same setting as in Theorem C.3, except that the local oracle is exact, i.e., $\sigma = 0$. Then for any $T \ge 1$, we have*

$$\mathbb{E}[F(\bar{x}_T) - F^\star] = \frac{1}{T}\sum_{t=1}^{T}F(x_t) - F^\star \lesssim \frac{1}{T}.$$

Proof. The result follows immediately from Theorem C.3 by plugging in $\mathcal{C}_T = 1$ and $\sigma = 0$. □

D LIPSCHITZ CONVEX OBJECTIVES

Assume throughout this subsection that every F_i is convex and G-Lipschitz, and write $\nabla F_i(x) \in \partial F_i(x)$ for a measurable subgradient selection satisfying $\|\nabla F_i(x)\| \le G$.

Note that

$$\Big\|\frac{1}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{i,k}^t)\Big\| \le \frac{1}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\|\nabla F_i(x_{i,k}^t)\| \le G. \tag{68}$$

Lemma A.6 and (15) therefore yields

$$\sqrt{\mathbb{E}_t[\|\bar{g}_t\|^2]} \le \sqrt{\mathbb{E}_t\Big[\Big\|\frac{1}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{i,k}^t)\Big\|^2\Big]} + \sqrt{\mathbb{E}_t[\|\bar{\xi}_t\|^2]} \le G + \frac{\sigma}{\sqrt{mK}},$$

and equivalently,

$$\mathbb{E}_t[\|\bar{g}_t\|^2] \le \left(G + \frac{\sigma}{\sqrt{mK}}\right)^2. \tag{69}$$

Lemma D.1. *Let Assumption 4.2 hold. Assume each F_i is convex and G-Lipschitz. Let ϵ_t be defined in* (48) *with $\nabla F_i(x_{i,k}^t) \in \partial F_i(x_{i,k}^t)$. Then, for every $t \ge 1$,*

$$\epsilon_t \le \frac{(K-1)G\sqrt{G^2+\sigma^2}}{q_t}\tau_t.$$

Proof. We first use G-Lipschitz continuity and $\|\nabla F_i(x_{i,k}^t)\| \le G$ to get

$$\begin{aligned} &F_i(x_t) - F_i(x_{i,k}^t) - \langle\nabla F_i(x_{i,k}^t), x_t - x_{i,k}^t\rangle \\ \le& G\|x_t - x_{i,k}^t\| + \|\nabla F_i(x_{i,k}^t)\|\|x_t - x_{i,k}^t\| \le 2G\|x_t - x_{i,k}^t\|. \end{aligned} \tag{70}$$

On the other hand, we have from (2) and unbiasedness of $\xi_{i,k}^t$ that

$$\mathbb{E}[\|g_{i,k}^t\|^2] = \mathbb{E}[\|\nabla F_i(x_{i,k}^t)\|^2] + 2\mathbb{E}[\langle\nabla F_i(x_{i,k}^t), \mathbb{E}[\xi_{i,k}^t \mid \mathcal{F}_{t,k}^{(i)}]\rangle] + \mathbb{E}[\|\xi_{i,k}^t\|^2] \le G^2 + \sigma^2.$$

Recall that $x_{i,k}^t - x_t = -(\tau_t/q_t)\sum_{s=0}^{k-1} g_{i,s}^t$, we thus obtain

$$\mathbb{E}_t[\|x_{i,k}^t - x_t\|] \le \frac{\tau_t}{q_t}\sum_{s=0}^{k-1}\mathbb{E}_t[\|g_{i,s}^t\|] \le \frac{\tau_t}{q_t}\sum_{s=0}^{k-1}\sqrt{\mathbb{E}_t[\|g_{i,s}^t\|^2]} \le \frac{k\tau_t}{q_t}\sqrt{G^2+\sigma^2},$$

which together with (70) gives

$$\epsilon_t \le \frac{2G}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}_t[\|x_{i,k}^t - x_t\|] \le \frac{2G\tau_t\sqrt{G^2+\sigma^2}}{q_t K}\sum_{k=0}^{K-1} k = \frac{(K-1)G\sqrt{G^2+\sigma^2}}{q_t}\tau_t.$$

The proof is complete. □

Lemma D.2. *Let Assumption 4.2 hold. Assume each F_i is convex and G-Lipschitz. Then for every $T \ge 1$, we have $\max_{1\le t\le T}\|r_t\| \le (G+M_T)/q_1$ almost surely, and*

$$\mathbb{E}[\mathcal{R}_T] \le 2\left(\eta^2 + \frac{2G^2}{q_1^2} + \frac{2\sigma^2(\log(mKT)+2)}{q_1^2}\right)\left(1+\log\frac{\mathbb{E}[q_{T+1}]}{q_1}\right), \tag{71}$$

$$\sqrt{\mathbb{E}[\mathcal{R}_T^2]} \le 2\left(\eta^2 + \frac{2G^2}{q_1^2} + \frac{2\sigma^2(\log(mKT)+4)}{q_1^2}\right)\left(4+\log\frac{\mathbb{E}[q_{T+1}]}{q_1}\right). \tag{72}$$

Proof. Firstly, it follows from (68) and $\|\bar{\xi}_t\| \le M_T$ that

$$\|r_t\| = \frac{\|\bar{g}_t\|}{q_t} \le \frac{G+M_T}{q_t} \le \frac{G+M_T}{q_1},\ \forall 1\le t\le T,$$

which proves $\max_{1\le t\le T}\|r_t\| \le (G+M_T)/q_1$ almost surely.

We proceed to bound $\mathbb{E}[\mathcal{R}_T]$. Since $(\|\bar{g}_t\|^2/\eta^2)/q_t^2 = \|r_t\|^2/\eta^2 \le (G+M_T)^2/(\eta^2 q_1^2)$, we apply Lemma A.4 with $A_t = q_t^2$, $a_t = \|\bar{g}_t\|^2/\eta^2$, and $c = (G+M_T)^2/(\eta^2 q_1^2)$ to get

$$\mathcal{R}_T = \sum_{t=1}^{T}\|r_t\|^2 = \eta^2\sum_{t=1}^{T}\frac{\|\bar{g}_t\|^2/\eta^2}{q_t^2} \le 2\left(\eta^2 + \frac{(G+M_T)^2}{q_1^2}\right)\log\frac{q_{T+1}}{q_1}. \tag{73}$$

Taking expectations on both sides of (73) and applying Cauchy–Schwarz inequality further gives

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_T] \le& 2\sqrt{\mathbb{E}\Big[\Big(\eta^2 + \frac{2G^2}{q_1^2} + \frac{2M_T^2}{q_1^2}\Big)^2\Big]}\sqrt{\mathbb{E}\Big[\log^2\frac{q_{T+1}}{q_1}\Big]} \\
\le& 2\left(\eta^2 + \frac{2G^2}{q_1^2} + \frac{2\sqrt{\mathbb{E}[M_T^4]}}{q_1^2}\right)\sqrt{\mathbb{E}\left[\left(1+\log\frac{q_{T+1}}{q_1}\right)^2\right]} \\
\le& 2\left(\eta^2 + \frac{2G^2}{q_1^2} + \frac{2\sigma^2(\log(mKT)+2)}{q_1^2}\right)\left(1+\log\frac{\mathbb{E}[q_{T+1}]}{q_1}\right),
\end{aligned}$$

where the second inequality uses Lemma A.6 and $q_{T+1} \ge q_1$, and the last one uses Lemma A.5 with $a=1$ and $p=2$ and $\sqrt{\mathbb{E}[M_T^4]} \le \sigma^2(\log(mKT)+2)$ from (22). This proves (71).

Using similar techniques, we square both sides of (73) and take expectations to get

$$\begin{aligned}
\sqrt{\mathbb{E}[\mathcal{R}_T^2]} \le& 2\left(\mathbb{E}\Big[\Big(\eta^2 + \frac{(G+M_T)^2}{q_1^2}\Big)^2\log^2\frac{q_{T+1}}{q_1}\Big]\right)^{1/2} \\
\le& 2\left(\mathbb{E}\Big[\Big(\eta^2 + \frac{2G^2}{q_1^2} + \frac{2M_T^2}{q_1^2}\Big)^4\Big]\right)^{1/4}\left(\mathbb{E}\Big[\log^4\frac{q_{T+1}}{q_1}\Big]\right)^{1/4} \\
\le& 2\left(\eta^2 + \frac{2G^2}{q_1^2} + \frac{2(\mathbb{E}[M_T^8])^{1/4}}{q_1^2}\right)\left(\mathbb{E}\Big[\left(4+\log\frac{q_{T+1}}{q_1}\right)^4\Big]\right)^{1/4} \\
\le& 2\left(\eta^2 + \frac{2G^2}{q_1^2} + \frac{2\sigma^2(\log(mKT)+4)}{q_1^2}\right)\left(4+\log\frac{\mathbb{E}[q_{T+1}]}{q_1}\right),
\end{aligned}$$

where the last inequality uses Lemma A.5 with $a=4$ and $p=4$ and $(\mathbb{E}[M_T^8])^{1/4} \le \sigma^2(\log(mKT)+4)$ from (22). The proof is complete. □

$$\mathcal{V}_T := \left(1+\sigma^2(\log(mKT)+1)\right)\left(1+\log\left(\sqrt{T}+\sigma\sqrt{T/(mK)}\right)\right)+(1+\sigma)\log T.$$

Lemma D.3. *Let Assumptions 4.1 and 4.2 hold. Assume each F_i is convex and G-Lipschitz. If $\mathcal{T}_T \lesssim \log(T)/(K-1)$, then for every $T \geq 1$,*

$$\sqrt{\mathbb{E}[\mathcal{R}_T^2]} \lesssim \mathcal{V}_T, \qquad \sqrt{\mathbb{E}[D_T^2]} \lesssim \mathcal{V}_T,$$

where $\mathcal{V}_T > 0$ is defined in (4)*.*

Proof. We first bound $\mathbb{E}[q_{T+1}]$. Note from the update rule and (69) that

$$\mathbb{E}[q_{T+1}^2] = q_1^2 + \frac{1}{\eta^2}\sum_{t=1}^{T}\mathbb{E}[\|\bar{g}_t\|^2] \leq q_1^2 + \frac{T}{\eta^2}\left(G+\frac{\sigma}{\sqrt{mK}}\right)^2 \lesssim T\left(1+\frac{\sigma}{\sqrt{mK}}\right)^2, \tag{74}$$

which implies

$$\mathbb{E}[q_{T+1}] \leq \sqrt{\mathbb{E}[q_{T+1}^2]} \lesssim \sqrt{T}\left(1+\frac{\sigma}{\sqrt{mK}}\right) = \sqrt{T}+\sigma\sqrt{\frac{T}{mK}}.$$

Plugging the above inequality into (72) gives

$$\sqrt{\mathbb{E}[\mathcal{R}_T^2]} \lesssim \left(1+\sigma^2(\log(mKT)+1)\right)\left(1+\log\left(\sqrt{T}+\sigma\sqrt{T/(mK)}\right)\right) \lesssim \mathcal{V}_T.$$

To bound $\mathbb{E}[D_T^2]$, we start from (51) to get

$$\|d_{t+1}\|^2 + \frac{2(F(x_t)-F^\star)}{q_t} \leq \|d_t\|^2 + \|r_t\|^2 - \frac{2}{q_t}\langle\zeta_t, d_t\rangle + \frac{2\epsilon_t}{q_t}. \tag{75}$$

From Lemma D.1 and $q_t \geq q_1$, we obtain

$$\frac{2\epsilon_t}{q_t} \leq \frac{2(K-1)G\sqrt{G^2+\sigma^2}}{q_t^2}\tau_t \leq \frac{2(K-1)G\sqrt{G^2+\sigma^2}}{q_1^2}\tau_t =: \tilde{\beta}_t. \tag{76}$$

For $1 \leq t \leq T$, let $\widetilde{\mathcal{M}}_t := -2\sum_{s=1}^{t}\langle\zeta_s, d_s\rangle/q_s$. Since $\mathbb{E}_t[\zeta_t]=0$, we have

$$\mathbb{E}_t[\widetilde{\mathcal{M}}_t] = \widetilde{\mathcal{M}}_{t-1} - \frac{2}{q_t}\langle\mathbb{E}_t[\zeta_t], d_t\rangle = \widetilde{\mathcal{M}}_{t-1},$$

which shows that $\{\widetilde{\mathcal{M}}_t\}_{t=1}^T$ is a martingale.

Applying(76) on (75) and unrolling yields

$$\|d_{t+1}\|^2 \leq \|d_1\|^2 + \sum_{s=1}^{t}(\|r_s\|^2+\tilde{\beta}_s) + \widetilde{\mathcal{M}}_t,$$

and hence

$$D_T \leq \|d_1\|^2 + \mathcal{R}_T + \sum_{t=1}^{T}\tilde{\beta}_t + \max_{1\leq s\leq T}|\widetilde{\mathcal{M}}_s|. \tag{77}$$

Using similar arguments as (55) in the proof of Lemma C.2, we have

$$\mathbb{E}\left[\max_{1\leq s\leq T}|\widetilde{\mathcal{M}}_s|^2\right] \leq 16\sqrt{\mathbb{E}[D_T^2]}\sqrt{\mathbb{E}[\mathcal{R}_T^2]}.$$

Using Lemma A.6 and the above inequality, then it follows from (77) that

$$\begin{aligned}
\sqrt{\mathbb{E}[D_T^2]} &\leq \|d_1\|^2 + \sum_{t=1}^{T}\tilde{\beta}_t + \sqrt{\mathbb{E}[\mathcal{R}_T^2]} + \sqrt{\mathbb{E}\left[\max_{1\leq s\leq T}|\widetilde{\mathcal{M}}_s|^2\right]} \\
&\leq \|d_1\|^2 + \sum_{t=1}^{T}\tilde{\beta}_t + \sqrt{\mathbb{E}[\mathcal{R}_T^2]} + 4\sqrt{\sqrt{\mathbb{E}[D_T^2]}\sqrt{\mathbb{E}[\mathcal{R}_T^2]}} \\
&\leq \|d_1\|^2 + \sum_{t=1}^{T}\tilde{\beta}_t + \frac{1}{2}\sqrt{\mathbb{E}[D_T^2]} + 9\sqrt{\mathbb{E}[\mathcal{R}_T^2]},
\end{aligned}$$

where the last inequality uses Young's inequality. Simplifying the above inequality and noting (76) gives

$$\begin{aligned}\sqrt{\mathbb{E}[D_T^2]} \leq & 2\left(\|d_1\|^2 + \frac{2(K-1)G(G+\sigma)}{q_1^2}\mathcal{T}_T\right) + 18\sqrt{\mathbb{E}[\mathcal{R}_T^2]} \\ \lesssim & 1 + (1+\sigma)\log T + \mathcal{V}_T \lesssim \mathcal{V}_T,\end{aligned}$$

where the second inequality uses $\mathcal{T}_T \lesssim \log(T)/(K-1)$. The proof is complete. □

D.1 Main Result and Proof

Theorem D.4 (Convex Lipschitz convergence)**.** *Let Assumptions 4.1 and 4.2 hold. Assume each F_i is convex and G-Lipschitz. Let $\{x_t\}_{t\geq 1}$ be the iterates generated by Algorithm 1 with τ_t chosen as in* (12) *and arbitrary $\gamma, \eta > 0$. Then for any $T \geq 1$, we have*

$$\mathbb{E}[F(\bar{x}_T) - F^\star] \lesssim \Big(\frac{1}{\sqrt{T}} + \frac{\sigma}{\sqrt{mKT}}\Big)\mathcal{V}_T,$$

where $\mathcal{V}_T > 0$ is defined in (4)*.*

Proof of Theorem D.4. Let $\Delta_t := F(x_t) - F^\star$. Rearranging (75) gives

$$\Delta_t \leq \frac{q_t}{2}\big(\|d_t\|^2 - \|d_{t+1}\|^2\big) + \frac{q_t}{2}\|r_t\|^2 + \epsilon_t - \langle \zeta_t, d_t\rangle. \tag{78}$$

We bound the sum of each term on the RHS of (78) in expectation. First, monotonicity of q_t gives

$$\sum_{t=1}^{T} q_t\big(\|d_t\|^2 - \|d_{t+1}\|^2\big) = q_1\|d_1\|^2 + \sum_{t=2}^{T}(q_t - q_{t-1})\|d_t\|^2 - q_T\|d_{T+1}\|^2 \leq q_T D_T. \tag{79}$$

Therefore, Cauchy–Schwarz and Lemma D.3 yield

$$\mathbb{E}\left[\sum_{t=1}^{T} q_t\big(\|d_t\|^2 - \|d_{t+1}\|^2\big)\right] \leq \mathbb{E}[q_{T+1}D_T] \leq \sqrt{\mathbb{E}[q_{T+1}^2]}\sqrt{\mathbb{E}[D_T^2]} \lesssim \mathcal{V}_T\sqrt{\mathbb{E}[q_{T+1}^2]}. \tag{80}$$

On the other hand, since $(\|\bar{g}_t\|^2/\eta^2)/q_t^2 = \|r_t\|^2/\eta^2 \leq (G+M_T)^2/(\eta^2 q_1^2)$ from Lemma D.2, we apply Lemma A.4 with $A_t = q_t^2$, $a_t = \|\bar{g}_t\|^2/\eta^2$, and $c = (G+M_T)^2/(\eta^2 q_1^2)$ to get

$$\sum_{t=1}^{T} q_t\|r_t\|^2 \leq \eta^2\left(1 + \sqrt{1 + \frac{(G+M_T)^2}{\eta^2 q_1^2}}\right)(q_{T+1} - q_1). \tag{81}$$

Denote $B_T := \Big(1 + \sqrt{1 + (G+M_T)^2/(\eta^2 q_1^2)}\Big)$. Then it follows from $(a+b)^2 \leq 2a^2 + 2b^2$ and (22) with $p = 1$ that

$$\begin{aligned}\sqrt{\mathbb{E}[B_T^2]} \leq & \sqrt{4 + \frac{4G^2}{\eta^2 q_1^2} + \frac{4\mathbb{E}[M_T^2]}{\eta^2 q_1^2}} \leq \left(4 + \frac{4G^2}{\eta^2 q_1^2} + \frac{4\sigma^2(\log(mKT)+1)}{\eta^2 q_1^2}\right)^{1/2} \\ \lesssim & \sqrt{1 + \sigma^2(\log(mKT)+1)} \leq 1 + \sigma\sqrt{\log(mKT)+1} \lesssim \mathcal{V}_T.\end{aligned}$$

Taking expectations on (81), we obtain from the above inequality and applying Cauchy–Schwarz inequality that

$$\mathbb{E}\left[\sum_{t=1}^{T} q_t\|r_t\|^2\right] \leq \eta^2\sqrt{\mathbb{E}[B_T^2]}\sqrt{\mathbb{E}[(q_{T+1} - q_1)^2]} \lesssim \mathcal{V}_T\sqrt{\mathbb{E}[q_{T+1}^2]}. \tag{82}$$

Additionally, Lemma D.1 and $\mathcal{T}_t \lesssim \log t/(K-1)$ gives

$$\sum_{t=1}^{T}\mathbb{E}[\epsilon_t] \leq \frac{(K-1)G\sqrt{G^2+\sigma^2}}{q_1}\mathcal{T}_T \lesssim (1+\sigma)\log T \lesssim \mathcal{V}_T. \tag{83}$$

Note that the last term in (78) has zero expectation since $\mathbb{E}[\langle \zeta_t, d_t \rangle] = \mathbb{E}[\langle \mathbb{E}_t[\zeta_t], d_t \rangle] = 0$. Now summing (78) over $t = 1, 2, \ldots, T$, taking expectations, and applying (80), (82), (83) yield

$$\sum_{t=1}^{T} \mathbb{E}[\Delta_t] \leq \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T} q_t(\|d_t\|^2 - \|d_{t+1}\|^2)\right] + \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T} q_t \|r_t\|^2\right] + \sum_{t=1}^{T} \mathbb{E}[\epsilon_t]$$
$$\lesssim \mathcal{V}_T \left(1 + \sqrt{\mathbb{E}[q_{T+1}^2]}\right) \lesssim \mathcal{V}_T \sqrt{\mathbb{E}[q_{T+1}^2]}. \tag{84}$$

Finally, plugging the bound on $\sqrt{\mathbb{E}[q_{T+1}^2]}$ (74) into (84) leads to

$$\sum_{t=1}^{T} \mathbb{E}[\Delta_t] \lesssim \sqrt{T}\left(1 + \frac{\sigma}{\sqrt{mK}}\right)\mathcal{V}_T.$$

Then the proof is complete after dividing both sides of the above inequality by T. □

Corollary D.5 (Deterministic Lipschitz convex convergence)**.** *Consider the same setting as in Theorem D.4, except that the local oracle is exact, i.e., $\sigma = 0$. Then, for every $T \geq 1$,*

$$\mathbb{E}[F(\bar{x}_T) - F^\star] = \frac{1}{T}\sum_{t=1}^{T} F(x_t) - F^\star \lesssim \frac{\log(1+T)}{\sqrt{T}}.$$

Proof. When $\sigma = 0$, we obtain from the definition of $\mathcal{V}_T$ (4) that

$$\mathcal{V}_T = 1 + \log\sqrt{T} + \log T \lesssim \log(1+T).$$

Then the result follows immediately from Theorem D.4 by plugging in $\sigma = 0$. □

E Extension I: Partial Participation

Algorithm 2 Adaptive Federated Learning Method (AdaFed) with Partial Participation

Require: initial model $x_1 \in \mathbb{R}^d$; local-step count $K \geq 1$; server iterations $T \geq 1$; inclusion probabilities $0 < p_i \leq 1$; prescribed local scales $\tau_t > 0$; parameters $\gamma, \eta > 0$.
Initialize the step size $\eta_1 := \eta/\gamma$.
for $t = 1, 2, \ldots, T$ **do**
  Sample active clients $\mathcal{S}_t \subset \{1, \ldots, m\}$ independently with probabilities $p_1, \ldots, p_m$.
  Server sends (x_t, η_t, τ_t) to every **client** $i \in \mathcal{S}_t$.
  for each **client** $i \in \mathcal{S}_t$ **in parallel do**
    Initialize $x_{i,0}^t := x_t$.
    for $k = 0, \ldots, K-1$ **do**
      Update $x_{i,k+1}^t := x_{i,k}^t - \tau_t \eta_t g_{i,k}^t$ with $g_{i,k}^t \approx \nabla F_i(x_{i,k}^t)$.
    end for
    Set $\Delta_i^t := x_{i,K}^t - x_t$ and send Δ_i^t to the server.
  end for
  Server updates the model and the step size:

$$x_{t+1} := x_t + \frac{1}{mK\tau_t}\sum_{i \in \mathcal{S}_t} \frac{\Delta_i^t}{p_i}, \qquad \frac{1}{\eta_{t+1}^2} := \frac{1}{\eta_t^2}\left(1 + \frac{\|x_{t+1} - x_t\|^2}{\eta^2}\right).$$

end for
Output: $\bar{x}_T$ chosen uniformly at random from $\{x_t\}_{t=1}^T$.

Conditionally on $\mathcal{F}_t$, the inclusion indicators $\chi_t^{(i)} := \mathbf{1}\{i \in \mathcal{S}_t\}$ are independent with $\mathbb{E}_t[\chi_t^{(i)}] = p_i \in (0, 1]$. Since $\Delta_i^t = -K\tau_t \bar{g}_t^{(i)}/q_t$, the server update in Algorithm 2 can be rewritten as

$$\widehat{g}_t := \frac{1}{m}\sum_{i=1}^{m} \frac{\chi_t^{(i)}}{p_i}\bar{g}_t^{(i)}, \qquad x_{t+1} = x_t - \frac{\widehat{g}_t}{q_t}, \qquad q_{t+1}^2 = q_t^2 + \frac{\|\widehat{g}_t\|^2}{\eta^2}. \tag{Core-PP}$$

Note that scaling $\bar{g}_t^{(i)}$ by $1/p_i$ for each active client removes the bias arising from unequal participation probabilities. Indeed, due to the independence of the inclusion indicators conditionally on $\mathcal{F}_t$, we have

$$\mathbb{E}_t[\widehat{g}_t] = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_t\left[\frac{\chi_t^{(i)}}{p_i}\right] \mathbb{E}_t[\bar{g}_t^{(i)}] = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_t[\bar{g}_t^{(i)}] = \mathbb{E}_t[\bar{g}_t].$$

E.1 Analysis Setup

For simplicity, we analyze Algorithm 2 with exact local gradients $g_{i,k}^t = \nabla F_i(x_{i,k}^t)$ and τ_t chosen as in (5). Thus $\kappa_\tau := \sup_{t\geq 2}(t-1)\tau_t < \infty$ and $\widetilde{\mathcal{T}}_T < \infty$. For analysis, define the counterfactual local path for every client (including inactive ones) by

$$x_{i,0}^t = x_t, \qquad x_{i,k+1}^t = x_{i,k}^t - \frac{\tau_t}{q_t} g_{i,k}^t, \qquad \bar{g}_t^{(i)} := \frac{1}{K}\sum_{k=0}^{K-1} g_{i,k}^t. \tag{86}$$

Recall $g_t = \nabla F(x_t)$, $b_t^{(i)} = \bar{g}_t^{(i)} - \nabla F_i(x_t)$, and $b_t := \frac{1}{m}\sum_{i=1}^m b_t^{(i)} = \bar{g}_t - g_t$ from full-participation analysis. We also define the participation noise via

$$\nu_t := \widehat{g}_t - \bar{g}_t = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\chi_t^{(i)}}{p_i} - 1\right)\bar{g}_t^{(i)}.$$

We first bound the conditional variance of $\widehat{g}_t$ in the following lemma.

Lemma E.1. *For all $t \geq 1$, we have $\mathbb{E}_t[\nu_t] = 0$ and*

$$\mathbb{E}_t[\|\nu_t\|^2] = \frac{1}{m^2}\sum_{i=1}^{m}\frac{1-p_i}{p_i}\|\bar{g}_t^{(i)}\|^2.$$

Proof. Conditionally on $\mathcal{F}_t$, every $\bar{g}_t^{(i)}$ is fixed. Since $\mathbb{E}_t[\chi_t^{(i)}] = p_i$, we have

$$\mathbb{E}_t[\nu_t] = \mathbb{E}_t\left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{\chi_t^{(i)}}{p_i} - 1\right)\bar{g}_t^{(i)}\right] = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\mathbb{E}_t[\chi_t^{(i)}]}{p_i} - 1\right)\bar{g}_t^{(i)} = 0.$$

Moreover, note that for any $i \in [m]$,

$$\mathbb{E}_t\left[\left(\frac{\chi_t^{(i)}}{p_i} - 1\right)^2\right] = p_i\left(\frac{1}{p_i} - 1\right)^2 + (1-p_i) = \frac{1-p_i}{p_i}.$$

For $i \neq j$, conditional independence gives

$$\mathbb{E}_t\left[\left(\frac{\chi_t^{(i)}}{p_i} - 1\right)\left(\frac{\chi_t^{(j)}}{p_j} - 1\right)\right] = 0.$$

Expanding the squared norm therefore yields

$$\mathbb{E}_t[\|\nu_t\|^2] = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbb{E}_t\left[\left(\frac{\chi_t^{(i)}}{p_i} - 1\right)\left(\frac{\chi_t^{(j)}}{p_j} - 1\right)\right]\left\langle \bar{g}_t^{(i)}, \bar{g}_t^{(j)}\right\rangle = \frac{1}{m^2}\sum_{i=1}^{m}\frac{1-p_i}{p_i}\|\bar{g}_t^{(i)}\|^2,$$

which completes the proof. □

Using Lemma E.1, we have from $\|\widehat{g}_t\|^2 = \|\bar{g}_t + \nu_t\|^2$ and $\mathbb{E}_t[\nu_t] = 0$ that

$$\mathbb{E}_t[\|\widehat{g}_t\|^2] = \|\bar{g}_t\|^2 + \frac{1}{m^2}\sum_{i=1}^{m}\frac{1-p_i}{p_i}\|\bar{g}_t^{(i)}\|^2. \tag{87}$$

As indicated in Lemma E.1, partial participation creates conditional sampling noise even though every active client uses exact gradients. In particular, the sampling variance can remain positive at a stationary point of F when the individual client gradients do not vanish there. The accumulator q_t can therefore grow continuously with t, unlike in the full-participation case. To control the magnitude of the participation variance, we introduce the following bounded-heterogeneity condition.

Assumption E.2 (Bounded heterogeneity)**.** There exists a constant $\delta \geq 0$ such that

$$\frac{1}{m}\sum_{i=1}^{m}\|\nabla F_i(x) - \nabla F(x)\|^2 \leq \delta^2, \qquad x \in \mathbb{R}^d.$$

Note that Assumption E.2 is only required for the analysis of smooth objectives. It is not needed for the Lipschitz convex result. For ease of notation, we define

$$A_p := 2 + \frac{2}{m}\max_{i\in[m]}\frac{1-p_i}{p_i}, \qquad \sigma_p^2 := \frac{2\delta^2}{m}\max_{i\in[m]}\frac{1-p_i}{p_i}.$$

Note that $\sigma_p = 0$ under full participation or zero heterogeneity.

Lemma E.3. *Let Assumption E.2 hold and assume each F_i is L-smooth. Then there exist nonnegative constants M_p, R_p, and B_p indenpendent of t, such that for all $t \geq 1$,*

$$\frac{\|g_t\|}{q_t} \leq M_p, \qquad \|r_t\| \leq R_p, \qquad \frac{1}{m}\sum_{i=1}^{m}\|b_t^{(i)}\|^2 \leq B_p^2\tau_t^2, \qquad \|b_t\| \leq B_p\tau_t.$$

Proof. Note that the counterfactual path (86) overlaps with the full-participation local path with exact local gradients. Hence (18) holds clientwise with $\sigma = 0$. Averaging the distance over k and using smoothness gives

$$\|x_{i,k}^t - x_t\| \leq \frac{k\tau_t\Theta}{q_t}\|\nabla F_i(x_t)\|, \qquad \|\bar{g}_t^{(i)}\| \leq \Theta\|\nabla F_i(x_t)\|, \qquad \|b_t^{(i)}\| \leq \frac{(K-1)L\Theta\tau_t}{2q_t}\|\nabla F_i(x_t)\|. \tag{88}$$

On the other hand, Assumption E.2 implies

$$\|\nabla F_i(x_t) - g_t\| \leq \sqrt{m}\,\delta, \qquad \frac{1}{m}\sum_{i=1}^{m}\|\nabla F_i(x_t)\|^2 \leq \|g_t\|^2 + \delta^2. \tag{89}$$

Consequently, (88) and (89) yield

$$\|\widehat{g}_t\| \leq \frac{1}{m}\sum_{i=1}^{m}\frac{1}{p_i}\|\bar{g}_t^{(i)}\| \leq \frac{\Theta}{m}\sum_{i=1}^{m}\frac{1}{p_i}(\|g_t\| + \sqrt{m}\,\delta). \tag{90}$$

We first show that $\|g_t\|/q_t$ is bounded. Set $n_0 := \max\{2, 1 + \lceil (K-1)L\Theta\kappa_\tau/q_1\rceil\}$ and $\beta_t := (K-1)L\Theta\tau_t/(2q_t)$. Since $q_t \geq q_1$ and $\tau_t \leq \kappa_\tau/(t-1)$ for $t \geq 2$, we have for any $t \geq n_0$ that

$$\beta_t \leq \frac{(K-1)L\Theta}{2q_t}\frac{\kappa_\tau}{\lceil (K-1)L\Theta\kappa_\tau/q_1\rceil} \leq 1/2.$$

We focus on nonempty rounds, since otherwise $x_{t+1} = x_t$ and hence $g_{t+1} = g_t$ and $q_{t+1} = q_t$. For a nonempty round, we obtain from (88) and (89) that

$$\|\bar{g}_t^{(i)} - g_t\| = \|b_t^{(i)} + \nabla F_i(x_t) - g_t\| \leq \beta_t\|g_t\| + (1+\beta_t)\sqrt{m}\,\delta.$$

Let $\omega_t := m^{-1}\sum_{i\in\mathcal{S}_t} p_i^{-1} \geq 1/m$. Then it follows from the above inequality that

$$\begin{aligned}
\|\widehat{g}_t\| &= \left\|\omega_t g_t + \frac{1}{m}\sum_{i\in\mathcal{S}_t}\frac{\bar{g}_t^{(i)} - g_t}{p_i}\right\| \\
&\geq \omega_t\|g_t\| - \frac{1}{m}\sum_{i\in\mathcal{S}_t}\frac{\|\bar{g}_t^{(i)} - g_t\|}{p_i} \\
&\geq \omega_t\big((1-\beta_t)\|g_t\| - (1+\beta_t)\sqrt{m}\,\delta\big).
\end{aligned}$$

For $t \geq n_0$, using $\beta_t \leq 1/2$ and $\omega_t^{-1} \leq m$ then gives

$$\|g_t\| \leq \frac{2}{\omega_t}\|\widehat{g}_t\| + 3\sqrt{m}\,\delta \leq 2m\|\widehat{g}_t\| + 3\sqrt{m}\,\delta. \tag{91}$$

To bound $\|g_{t+1}\|$, we use smoothness, (Core-PP), and plugging in (91) to get

$$\|g_{t+1}\| \le \|g_t\| + L\|x_{t+1} - x_t\| = \|g_t\| + \frac{L}{q_t}\|\widehat{g}_t\| \le \left(2m + \frac{L}{q_t}\right)\|\widehat{g}_t\| + 3\sqrt{m}\,\delta.$$

Note that $q_{t+1} \ge q_1$ and $q_{t+1} \ge \|\widehat{g}_t\|/\eta$, dividing both sides of the above inequality by q_{t+1} yields

$$\frac{\|g_{t+1}\|}{q_{t+1}} \le \frac{(2m + L/q_t)\|\widehat{g}_t\| + 3\sqrt{m}\,\delta}{q_{t+1}} \le \eta(2m + L/q_1) + 3\sqrt{m}\,\delta/q_1.$$

For $t < n_0$, combining (90) and $\|g_{t+1}\| \le \|g_t\| + L\|\widehat{g}_t\|/q_t$ implies

$$\|g_{t+1}\| + \sqrt{m}\,\delta \le \left(1 + \frac{L\Theta}{mq_1}\sum_{i=1}^{m}\frac{1}{p_i}\right)(\|g_t\| + \sqrt{m}\,\delta). \tag{92}$$

Iterating (92) for at most $n_0 - 1$ rounds and using $q_t \ge q_1$ then bounds $\|g_{t+1}\|/q_{t+1}$ for $t < n_0$. Now define the nonnegative constants M_p, R_p, and B_p via:

$$M_p := \max\left\{\left(1 + \frac{L\Theta}{mq_1}\sum_{i=1}^{m}\frac{1}{p_i}\right)^{n_0-1}\frac{\|g_1\| + \sqrt{m}\,\delta}{q_1},\ \eta(2m + L/q_1) + \frac{3\sqrt{m}\,\delta}{q_1}\right\},$$

$$R_p := \frac{\Theta}{m}\sum_{i=1}^{m}\frac{1}{p_i}\left(M_p + \frac{\sqrt{m}\,\delta}{q_1}\right), \qquad B_p := \frac{(K-1)L\Theta}{2}\sqrt{M_p^2 + \delta^2/q_1^2}.$$

It then follows that $\|g_t\|/q_t \le M_p$ for all $t \ge 1$.

To verify the remaining bounds, divide (90) by q_t and use $q_t \ge q_1$ to obtain

$$\|r_t\| = \frac{\|\widehat{g}_t\|}{q_t} \le \frac{\Theta}{m}\sum_{i=1}^{m}\frac{1}{p_i}\left(\frac{\|g_t\|}{q_t} + \frac{\sqrt{m}\,\delta}{q_t}\right) \le R_p.$$

Moreover, squaring and averaging the last inequality in (88), then using (89), yields

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\|b_t^{(i)}\|^2 &\le \frac{(K-1)^2L^2\Theta^2\tau_t^2}{4q_t^2}\frac{1}{m}\sum_{i=1}^{m}\|\nabla F_i(x_t)\|^2\\ &\le \frac{(K-1)^2L^2\Theta^2\tau_t^2}{4}\left(\frac{\|g_t\|^2}{q_t^2} + \frac{\delta^2}{q_t^2}\right) \le B_p^2\tau_t^2.\end{aligned}$$

Convexity of $\|\cdot\|^2$ then gives $\|b_t\|^2 \le m^{-1}\sum_{i=1}^{m}\|b_t^{(i)}\|^2 \le B_p^2\tau_t^2$, which completes the proof. □

Using $\|\bar{g}_t^{(i)}\|^2 \le 2\|\nabla F_i(x_t)\|^2 + 2\|b_t^{(i)}\|^2$, then it follows from Lemma E.1 and Lemma E.3 that

$$\mathbb{E}_t[\|\nu_t\|^2] \le \frac{2}{m}\max_{i\in[m]}\frac{1-p_i}{p_i}(\|g_t\|^2 + \delta^2 + B_p^2\tau_t^2),$$

which together with $\widehat{g}_t = g_t + b_t + \nu_t$ implies

$$\begin{aligned}&\mathbb{E}_t[\|\widehat{g}_t\|^2] = \|g_t + b_t\|^2 + \mathbb{E}_t[\|\nu_t\|^2]\\ \le&2\|g_t\|^2 + 2B_p^2\tau_t^2 + \frac{2}{m}\max_{i\in[m]}\frac{1-p_i}{p_i}(\|g_t\|^2 + \delta^2 + B_p^2\tau_t^2) = A_p\|g_t\|^2 + \sigma_p^2 + A_pB_p^2\tau_t^2.\end{aligned} \tag{93}$$

In the analysis, we will use the following bounds on trajectory quantities:

$$\sum_{t=1}^{T}\|\widehat{g}_t\|^2 = \eta^2(q_{T+1}^2 - q_1^2), \tag{94}$$

$$\sum_{t=1}^{T}\|r_t\|^2 \le (\eta^2 + R_p^2)\log\frac{q_{T+1}^2}{q_1^2}, \tag{95}$$

$$\sum_{t=1}^{T}\frac{\|\widehat{g}_t\|^2}{q_t} \le \eta^2\left(1 + \sqrt{1 + R_p^2/\eta^2}\right)(q_{T+1} - q_1). \tag{96}$$

The first identity follows directly from (Core-PP). The remaining two bounds are obtained by applying Lemma A.4 with $A_t = q_t^2$, $a_t = \|\widehat{g}_t\|^2/\eta^2$, and $c = R_p^2/\eta^2$ due to $a_t/A_t = \|r_t\|^2/\eta^2 \le R_p^2/\eta^2$.

E.2 SMOOTH NONCONVEX CONVERGENCE

For ease of presentation, we define

$$\mathcal{P}_T := 1 + \log(1 + \sigma_p\sqrt{T}).$$

Thus $\mathcal{P}_T = 1$ when $\sigma_p = 0$, while $\mathcal{P}_T \lesssim \log(1+T)$ for $\sigma_p > 0$.

Theorem E.4 (Smooth nonconvex convergence)**.** *Let Assumptions 4.1and E.2 hold and assume each F_i is L-smooth. Let $\{x_t\}_{t\geq 1}$ be the iterates generated by Algorithm 2 with τ_t chosen as in* (5) *and arbitrary $\gamma, \eta > 0$. Then for any $T \geq 1$, we have*

$$\mathbb{E}[\|\nabla F(\bar{x}_T)\|^2] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla F(x_t)\|^2] \lesssim \frac{\mathcal{P}_T^2}{T} + \frac{\sigma_p \mathcal{P}_T}{\sqrt{T}}.$$

Proof of Theorem E.4. We first estimate $\mathbb{E}[q_{T+1}]$ and $\mathbb{E}[\mathcal{G}_T]$. Since q_t is $\mathcal{F}_t$-measurable, smoothness of F gives

$$\begin{aligned} F(x_{t+1}) \leq & F(x_t) - \frac{\langle g_t, \widehat{g}_t\rangle}{q_t} + \frac{L}{2}\|r_t\|^2 \\ = & F(x_t) - \frac{\langle g_t, b_t\rangle}{q_t} - \frac{\langle g_t, \nu_t\rangle}{q_t} - \frac{\|g_t\|^2}{q_t} + \frac{L}{2}\|r_t\|^2 \\ \leq & F(x_t) - \frac{\|g_t\|^2}{2q_t} + \frac{\|b_t\|^2}{2q_t} - \frac{\langle g_t, \nu_t\rangle}{q_t} + \frac{L}{2}\|r_t\|^2. \end{aligned} \tag{97}$$

Taking conditional expectations on both sides of (97) and using $\mathbb{E}_t[\nu_t] = 0$ gives

$$\mathbb{E}_t[F(x_{t+1})] \leq F(x_t) - \frac{\|g_t\|^2}{2q_t} + \frac{\|b_t\|^2}{2q_t} + \frac{L}{2}\mathbb{E}_t[\|r_t\|^2].$$

Summing the above inequality, we obtain from $q_t \geq q_1$, $\|b_t\| \leq B_p\tau_t$, $\widetilde{\mathcal{T}}_T \lesssim 1$ under (12), and (95) that

$$\mathbb{E}[\mathcal{G}_T] \leq 2\big(F(x_1) - F^\star\big) + \frac{B_p^2}{q_1}\widetilde{\mathcal{T}}_T + L(\eta^2 + R_p^2)\mathbb{E}\left[\log\frac{q_{T+1}^2}{q_1^2}\right] \lesssim 1 + \log\mathbb{E}[q_{T+1}]. \tag{98}$$

Applying Lemma A.8 with $c = q_1^2$ and $Z_t = \|\widehat{g}_t\|^2/\eta^2$, together with (94), (93), and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we get

$$\begin{aligned} \mathbb{E}[q_{T+1}] = \mathbb{E}\left[\sqrt{q_1^2 + \frac{1}{\eta^2}\sum_{t=1}^{T}\|\widehat{g}_t\|^2}\right] &\leq \sqrt{2}\,\mathbb{E}\left[\sqrt{q_1^2 + \frac{1}{\eta^2}\sum_{t=1}^{T}\mathbb{E}_t[\|\widehat{g}_t\|^2]}\right] \\ &\leq \sqrt{2}\,\mathbb{E}\left[\sqrt{q_1^2 + \frac{A_p}{\eta^2}\sum_{t=1}^{T}\|g_t\|^2 + \frac{\sigma_p^2 T}{\eta^2} + \frac{A_p B_p^2}{\eta^2}\widetilde{\mathcal{T}}_T}\right] \\ &\lesssim 1 + \mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|g_t\|^2}\right] + \sigma_p\sqrt{T}. \end{aligned} \tag{99}$$

On the other hand, monotonicity of q_t and then Cauchy–Schwarz yield

$$\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|g_t\|^2}\right] \leq \mathbb{E}\left[\sqrt{q_{T+1}\sum_{t=1}^{T}\frac{\|g_t\|^2}{q_t}}\right] \leq \sqrt{\mathbb{E}[q_{T+1}]\mathbb{E}[\mathcal{G}_T]}.$$

Substituting the above bound into (99) therefore gives

$$\mathbb{E}[q_{T+1}] \lesssim 1 + \sqrt{\mathbb{E}[q_{T+1}]\mathbb{E}[\mathcal{G}_T]} + \sigma_p\sqrt{T},$$

which further imply that for a constant C,

$$\mathbb{E}[q_{T+1}] \leq \frac{\mathbb{E}[q_{T+1}]}{2} + C(1+\sigma_p\sqrt{T}) + \frac{C^2}{2}\mathbb{E}[\mathcal{G}_T].$$

Simplifying the above and applying (98) then gives

$$\mathbb{E}[q_{T+1}] \lesssim 1 + \sigma_p\sqrt{T} + \log \mathbb{E}[q_{T+1}],$$

solving which with Lemma A.1 implies

$$\mathbb{E}[q_{T+1}] \lesssim 1 + \sigma_p\sqrt{T}, \qquad \mathbb{E}[\mathcal{G}_T] \lesssim 1 + \log(1+\sigma_p\sqrt{T}) = \mathcal{P}_T. \tag{100}$$

We next obtain the second moment $\mathbb{E}[q_{T+1}^2]$ and $\mathbb{E}[\mathcal{G}_T^2]$. For $0 \leq t \leq T$, define

$$\mathcal{Z}_t := 2\big(F(x_{t+1}) - F^\star\big) + \sum_{s=1}^{t} \frac{\|g_s\|^2}{q_s}, \quad \mathcal{Z}_T^* := \max_{0\leq t\leq T} \mathcal{Z}_t, \quad \mathcal{N}_t := -2\sum_{s=1}^{t} \frac{\langle g_s, \nu_s\rangle}{q_s},$$

The process $\{\mathcal{N}_t\}$ is a martingale since $\mathbb{E}_t[\nu_t] = 0$. Summing (97) over $s = 1$ to t, then maximizing over t and using $\sum_{t=1}^{T} \|b_t\|^2 \lesssim \widetilde{\mathcal{T}}_T \lesssim 1$ yields

$$\mathcal{Z}_T^* \lesssim 1 + \mathcal{R}_T + \max_{0\leq t\leq T} |\mathcal{N}_t|. \tag{101}$$

Squaring (95) and then plugging (100) imply

$$\begin{aligned}\sqrt{\mathbb{E}[\mathcal{R}_T^2]} &\leq 2(\eta^2 + R_p^2)\sqrt{\mathbb{E}\left[\log^2 \frac{q_{T+1}}{q_1}\right]} \\ &\leq 2(\eta^2 + R_p^2)\left(1 + \log\frac{\mathbb{E}[q_{T+1}]}{q_1}\right) \lesssim \mathcal{P}_T,\end{aligned} \tag{102}$$

where the second inequality follows from concavity of the function $u \mapsto (1+\log u)^2$ on $[1,\infty)$. Additionally, by smoothness, $\|g_t\|^2 \leq 2L\big(F(x_t) - F^\star\big) \leq L\mathcal{Z}_T^*$.

$$\begin{aligned}\mathbb{E}\left[\max_{0\leq t\leq T} |\mathcal{N}_t|^2\right] &\leq 4\mathbb{E}[|\mathcal{N}_T|^2] \leq 16\sum_{t=1}^{T}\mathbb{E}\left[\frac{\langle g_t,\nu_t\rangle^2}{q_t^2}\right] \leq 16\sum_{t=1}^{T}\mathbb{E}\left[\frac{\|g_t\|^2\|\nu_t\|^2}{q_t^2}\right] \\ &\leq 16L\mathbb{E}[\mathcal{Z}_T^*\mathcal{R}_T] \leq 16L\sqrt{\mathbb{E}[(\mathcal{Z}_T^*)^2]}\sqrt{\mathbb{E}[\mathcal{R}_T^2]} \lesssim \sqrt{\mathbb{E}[(\mathcal{Z}_T^*)^2]}\mathcal{P}_T,\end{aligned} \tag{103}$$

where the fourth inequality follows from $\mathbb{E}_t[\|\nu_t/q_t\|^2] \leq \mathbb{E}_t[\|r_t\|^2]$ due to $\nu_t/q_t = -(r_t - \mathbb{E}_t[r_t])$. Taking L^2 norms in (101) and using (103) then gives

$$\sqrt{\mathbb{E}[(\mathcal{Z}_T^*)^2]} \lesssim 1 + \mathcal{P}_T + \sqrt{\mathcal{P}_T\sqrt{\mathbb{E}[(\mathcal{Z}_T^*)^2]}} \leq \frac{1}{2}\sqrt{\mathbb{E}[(\mathcal{Z}_T^*)^2]} + C\mathcal{P}_T$$

for some constant $C > 0$. Rearranging the above yields $\sqrt{\mathbb{E}[(\mathcal{Z}_T^*)^2]} \lesssim \mathcal{P}_T$, which together with $\mathcal{G}_T \leq \mathcal{Z}_T \leq \mathcal{Z}_T^*$ implies that

$$\sqrt{\mathbb{E}[\mathcal{G}_T^2]} \lesssim \mathcal{P}_T. \tag{104}$$

It remains to bound $\mathbb{E}[q_{T+1}^2]$. By monotonicity of q_t and (104), we have

$$\sum_{t=1}^{T}\mathbb{E}[\|g_t\|^2] \leq \mathbb{E}[q_{T+1}\mathcal{G}_T] \leq \sqrt{\mathbb{E}[q_{T+1}^2]}\sqrt{\mathbb{E}[\mathcal{G}_T^2]} \lesssim \mathcal{P}_T\sqrt{\mathbb{E}[q_{T+1}^2]}. \tag{105}$$

Therefore, it follows from (94), (93), and $\widetilde{\mathcal{T}}_T \lesssim 1$ that

$$\begin{aligned}\mathbb{E}[q_{T+1}^2] &\leq q_1^2 + \frac{A_p}{\eta^2}\sum_{t=1}^{T}\mathbb{E}[\|g_t\|^2] + \frac{\sigma_p^2 T}{\eta^2} + \frac{A_p B_p^2}{\eta^2}\widetilde{\mathcal{T}}_T \\ &\lesssim 1 + \mathcal{P}_T\sqrt{\mathbb{E}[q_{T+1}^2]} + \sigma_p^2 T,\end{aligned}$$

which yields

$$\sqrt{\mathbb{E}[q_{T+1}^2]} \lesssim \mathcal{P}_T + \sigma_p\sqrt{T}. \tag{106}$$

Therefore, we obtain from (105) that

$$\sum_{t=1}^{T} \mathbb{E}[\|g_t\|^2] \lesssim \mathcal{P}_T^2 + \sigma_p \mathcal{P}_T\sqrt{T}.$$

Finally, since $\bar{x}_T$ is chosen uniformly from $\{x_t\}_{t=1}^T$, it follows that

$$\mathbb{E}[\|\nabla F(\bar{x}_T)\|^2] = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|g_t\|^2] \lesssim \frac{\mathcal{P}_T^2}{T} + \frac{\sigma_p \mathcal{P}_T}{\sqrt{T}},$$

which completes the proof. □

E.3 Smooth convex convergence

For convex problems, we similarly define the counterfactual local error

$$\epsilon_t := \frac{1}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1} \left[F_i(x_t) - F_i(x_{i,k}^t) - \langle \nabla F_i(x_{i,k}^t), x_t - x_{i,k}^t\rangle\right] \geq 0.$$

Then the same convexity calculation as in (49) gives

$$F(y) \geq F(x_t) + \langle \bar{g}_t, y - x_t\rangle - \epsilon_t, \qquad y \in \mathbb{R}^d. \tag{107}$$

Theorem E.5 (Smooth convex convergence)**.** *Let Assumptions 4.1 and E.2 hold and assume each F_i is L-smooth and convex. Let $\{x_t\}_{t\geq 1}$ be the iterates generated by Algorithm 2 with τ_t chosen as in* (5) *and arbitrary $\gamma, \eta > 0$. Then for any $T \geq 1$, we have*

$$\mathbb{E}[F(\bar{x}_T) - F(x^\star)] \lesssim \frac{\mathcal{P}_T^2}{T} + \frac{\sigma_p \mathcal{P}_T}{\sqrt{T}}.$$

Proof of Theorem E.5. Let $d_t = x_t - x^\star$, $D_T := \max_{1\leq t\leq T+1}\|d_t\|^2$, and $\Delta_t := F(x_t) - F(x^\star)$. Since Lemma C.1 applies to the exact counterfactual paths, we apply it and then use (89) and Lemma E.3 to obtain

$$\epsilon_t \leq \frac{c_K L\Theta^2\tau_t^2}{2mq_t^2}\sum_{i=1}^{m}\|\nabla F_i(x_t)\|^2 \leq \frac{c_K L\Theta^2\tau_t^2}{2}\left(\frac{\|g_t\|^2}{q_t^2} + \frac{\delta^2}{q_t^2}\right) \lesssim \tau_t^2. \tag{108}$$

Expand $\mathbb{E}_t[\|d_{t+1}\|^2]$, then it follows from $\mathbb{E}_t[\widehat{g}_t] = \bar{g}_t$, (107) with $y = x^\star$ that

$$\begin{aligned}\mathbb{E}_t[\|d_{t+1}\|^2] &= \|d_t\|^2 - \frac{2}{q_t}\langle \bar{g}_t, d_t\rangle + \mathbb{E}_t[\|r_t\|^2] \\ &\leq \|d_t\|^2 - \frac{2}{q_t}\Delta_t + \mathbb{E}_t[\|r_t\|^2] + \frac{2\epsilon_t}{q_t}.\end{aligned} \tag{109}$$

Summing (109) over $t = 1$ to T, and applying (108), (100), (95), and $\widetilde{\mathcal{T}}_T \lesssim 1$ give

$$\mathbb{E}[\|d_{T+1}\|^2] + 2\sum_{t=1}^{T}\mathbb{E}\left[\Delta_t/q_t\right] \lesssim \mathcal{P}_T.$$

To proceed, consider the martingale

$$\mathcal{M}_s := -2\sum_{t=1}^{s}\frac{\langle \nu_t, d_t\rangle}{q_t}, \qquad 0 \leq s \leq T.$$

Expanding $\|d_{t+1}\|^2$ and using (107) leads to

$$\|d_{t+1}\|^2 + \frac{2\Delta_t}{q_t} \leq \|d_t\|^2 + \|r_t\|^2 + \frac{2\epsilon_t}{q_t} - \frac{2\langle \nu_t, d_t\rangle}{q_t}. \tag{110}$$

Sum this inequality to any $s \le T$ and maximize over s, then it follows from (108), $\Delta_t \ge 0$, and $\widetilde{\mathcal{T}}_T \lesssim 1$ that

$$D_T \lesssim 1 + \mathcal{R}_T + \max_{0\le s\le T} |\mathcal{M}_s|. \tag{111}$$

On the other hand, using similar arguments as in (103), we have

$$\mathbb{E}\left[\max_{0\le s\le T} |\mathcal{M}_s|^2\right] \le 16 \sum_{t=1}^{T} \mathbb{E}\left[\|d_t\|^2 \mathbb{E}_t[\|r_t\|^2]\right] \le 16\mathbb{E}[D_T \mathcal{R}_T] \lesssim \mathcal{P}_T \sqrt{\mathbb{E}[D_T^2]}. \tag{112}$$

Therefore, squaring (111) and using (112) and (102) give

$$\sqrt{\mathbb{E}[D_T^2]} \lesssim 1 + \mathcal{P}_T + \sqrt{\mathcal{P}_T \sqrt{\mathbb{E}[D_T^2]}} \le \frac{1}{2}\sqrt{\mathbb{E}[D_T^2]} + C\mathcal{P}_T, \qquad \sqrt{\mathbb{E}[D_T^2]} \lesssim \mathcal{P}_T.$$

It remains to bound $\sum_{t=1}^{T} \mathbb{E}[\Delta_t]$. Rearranging (110) and plugging in $\|r_t\| = \|\widehat{g}_t/q_t\| s$ gives

$$\Delta_t \le \frac{q_t}{2}\left(\|d_t\|^2 - \|d_{t+1}\|^2\right) + \frac{\|\widehat{g}_t\|^2}{2q_t} + \epsilon_t - \langle \nu_t, d_t \rangle. \tag{113}$$

Note that the last term in (113) has zero expectation. For the sum of the first term, we have

$$\sum_{t=1}^{T} q_t \left(\|d_t\|^2 - \|d_{t+1}\|^2\right) = q_1\|d_1\|^2 + \sum_{t=2}^{T} (q_t - q_{t-1})\|d_t\|^2 - q_T\|d_{T+1}\|^2 \le q_T D_T.$$

Consequently, sum (113) over t, then use the above inequality, (96), and (106) from the proof of Theorem E.4, we derive

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}[\Delta_t] \lesssim& \mathbb{E}[q_T D_T] + \mathbb{E}[q_{T+1}] + 1 \lesssim \sqrt{\mathbb{E}[D_T^2]}\sqrt{\mathbb{E}[q_{T+1}^2]} + \sqrt{\mathbb{E}[q_{T+1}^2]} \\ \lesssim& \mathcal{P}_T \sqrt{\mathbb{E}[q_{T+1}^2]} \lesssim \mathcal{P}_T^2 + \sigma_p \mathcal{P}_T \sqrt{T},\end{aligned}$$

which completes the proof after dividing both sides by T. □

E.4 Convex Lipschitz convergence

Note that in this subsection, we do not require the bounded heterogeneity assumption in Assumption E.2.

Theorem E.6 (Lipschitz convex convergence)**.** *Let Assumption 4.1 hold. Assume each F_i is G-Lipschitz and convex. Let $\{x_t\}_{t\ge 1}$ be the iterates generated by Algorithm 2 with τ_t chosen as in* (5) *and arbitrary $\gamma, \eta > 0$. Then for any $T \ge 1$, we have*

$$\mathbb{E}[F(\bar{x}_T) - F(x^\star)] \lesssim \frac{\log(1+T)}{\sqrt{T}}.$$

Proof of Theorem E.6. Let $d_t = x_t - x^\star$ and $\Delta_t := F(x_t) - F(x^\star)$. For this proof only, set

$$\widetilde{G} := \frac{G}{m}\sum_{i=1}^{m} \frac{1}{p_i}, \qquad M_G := G^2\left(1 + \frac{1}{m^2}\sum_{i=1}^{m} \frac{1-p_i}{p_i}\right).$$

Then it follows from the definition of $\widehat{g}_t$ and (87) that

$$\|\widehat{g}_t\| \le \frac{G}{m}\sum_{i\in\mathcal{S}_t} p_i^{-1} \le \widetilde{G}, \qquad \mathbb{E}_t[\|\widehat{g}_t\|^2] = \|\bar{g}_t\|^2 + \frac{1}{m^2}\sum_{i=1}^{m} \frac{1-p_i}{p_i}\|\bar{g}_t^{(i)}\|^2 \le M_G. \tag{114}$$

Consequently, (94) and (114) give

$$q_{T+1} \le \sqrt{q_1^2 + \frac{T\widetilde{G}^2}{\eta^2}} \lesssim \sqrt{T}, \qquad \mathbb{E}[q_{T+1}^2] \le q_1^2 + \frac{TM_G}{\eta^2}.$$

Since $\|r_t\| = \|\widehat{g}_t\|/q_t \le \widetilde{G}/q_1$, we further obtain from (95) with $R_p = \widetilde{G}/q_1$ that

$$\sum_{t=1}^{T} \mathbb{E}[\|r_t\|^2] \le \left(\eta^2 + \frac{\widetilde{G}^2}{q_1^2}\right) \mathbb{E}\left[\log \frac{q_{T+1}^2}{q_1^2}\right] \lesssim \log(1+T), \tag{115}$$

On the other hand, rearranging (109) and noting that $\epsilon_t \le (K-1)G^2\tau_t/q_t$ from Lemma D.1 with $\sigma = 0$ yields

$$\mathbb{E}_t[\|d_{t+1}\|^2] + \frac{2\Delta_t}{q_t} \le \|d_t\|^2 + \mathbb{E}_t[\|r_t\|^2] + \frac{2\epsilon_t}{q_t} \le \|d_t\|^2 + \mathbb{E}_t[\|r_t\|^2] + \frac{2(K-1)G^2\tau_t}{q_1^2}.$$

Summing the above inequality from $t = 1$ to T, applying (115), and using $\mathcal{T}_T \lesssim \log(1+T)$ give

$$\mathbb{E}[\|d_{T+1}\|^2] + 2\sum_{t=1}^{T} \mathbb{E}\left[\frac{\Delta_t}{q_t}\right] \lesssim \log(1+T).$$

Finally, it follows from $q_t \lesssim \sqrt{T}$ that

$$\mathbb{E}[F(\bar{x}_T) - F(x^\star)] = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\Delta_t] \lesssim \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbb{E}\left[\frac{\Delta_t}{q_t}\right] \lesssim \frac{\log(1+T)}{\sqrt{T}},$$

which completes the proof. □

F Extension II: Adaptation to Hölder smoothness

For convex objectives, Orabona (2023) showed that the standard AdaGrad with exact gradients can adapt to Hölder smoothness, which interpolates between smoothness and Lipschitzness. Inspired by their work, we extend the convex analysis for Algorithm 1 under the Hölder-smooth condition. Similar to (Orabona, 2023) and for simplicity, we focus on Algorithm 1 with exact local gradients, i.e., $g_{i,k}^t = \nabla F_i(x_{i,k}^t)$.

F.1 Auxiliary Results

Recall that a function $F : \mathbb{R}^d \to \mathbb{R}$ is (L_α, α)-Hölder smooth with $L_\alpha > 0$ and $\alpha \in (0, 1]$ if

$$\|\nabla F(x) - \nabla F(y)\| \le L_\alpha \|x - y\|^\alpha, \qquad x, y \in \mathbb{R}^d.$$

Note that the Hölder smoothness condition also implies (Orabona, 2023, Theorem 1):

$$F(y) \le F(x) + \langle \nabla F(x), y - x\rangle + \frac{L_\alpha}{1+\alpha}\|y - x\|^{1+\alpha}, \qquad x, y \in \mathbb{R}^d. \tag{116}$$

Then $\alpha = 1$ corresponds to the standard L-smoothness, while $\alpha = 0$ corresponds to Lipschitzness. Since we have already established analysis for these two cases separately, in the remainder of this subsection, we mainly focus on $\alpha \in (0, 1)$.

We first invoke the following lemma that converts gradient norms to function values.

Lemma F.1 (Hölder self-bounding property (Orabona, 2023))**.** *Suppose F is convex and (L_α, α)-Hölder smooth with $\alpha \in (0, 1]$. Then for any $x \in \mathbb{R}^d$, we have*

$$\|\nabla F(x)\|^{(1+\alpha)/\alpha} \le \left(1 + \frac{1}{\alpha}\right) L_\alpha^{1/\alpha}\big(F(x) - F^\star\big).$$

Recall $b_t := \bar{g}_t - g_t$, $d_t := x_t - x^\star$, and ϵ_t defined in (48). We next bound the local errors $\|b_t\|$ and ϵ_t in the following lemma.

Lemma F.2. *Let Assumption 4.1 hold. Assume each F_i is convex and (L_α, α)-Hölder smooth for some $\alpha \in (0, 1)$. Consider Algorithm 1 with exact local gradients $g_{i,k}^t = \nabla F_i(x_{i,k}^t)$ and τ_t chosen as* (5)*. Then*

$$\sum_{t=1}^{T} \left(\|b_t\|^{(1+\alpha)/\alpha} + \epsilon_t\right) \lesssim 1 + D_T^{\alpha(1+\alpha)/2}.$$

Proof. We focus on $K \geq 2$ since the claims for $K = 1$ are immediate due to $b_t = \epsilon_t = 0$. For any $k < K$, the local updates and (L_α, α)-Hölder smoothness of F_i give

$$\|x_{i,k}^t - x_t\| \leq \frac{\tau_t}{q_t} \sum_{s=0}^{k-1} \|\nabla F_i(x_{i,s}^t)\| \leq \frac{K\tau_t}{q_t} \left(\|\nabla F_i(x_t)\| + L_\alpha \Big(\max_{0 \leq s < K} \|x_{i,s}^t - x_t\| \Big)^\alpha \right),$$

which further implies

$$\max_{0 \leq k < K} \|x_{i,k}^t - x_t\| \leq \frac{K\tau_t}{q_t} \|\nabla F_i(x_t)\| + \frac{KL_\alpha \tau_t}{q_t} \Big(\max_{0 \leq k < K} \|x_{i,k}^t - x_t\| \Big)^\alpha.$$

Solving the above inequality using Lemma A.2 gives

$$\max_{0 \leq k < K} \|x_{i,k}^t - x_t\| \leq \frac{2K\tau_t}{q_t} \|\nabla F_i(x_t)\| + \Big(\frac{2KL_\alpha \tau_t}{q_t} \Big)^{\frac{1}{1-\alpha}}. \tag{117}$$

Moreover, using Hölder smoothness, (117), and $(a+b)^\alpha \leq a^\alpha + b^\alpha$ for $\alpha \in (0,1)$, we obtain

$$\|b_t\| \leq \frac{L_\alpha}{mK} \sum_{i=1}^m \sum_{k=0}^{K-1} \|x_{i,k}^t - x_t\|^\alpha \lesssim \left(\frac{K\tau_t}{q_t} \right)^\alpha \left(1 + \|d_t\|^{\alpha^2} \right) + \left(\frac{K\tau_t}{q_t} \right)^{\frac{\alpha}{1-\alpha}}. \tag{118}$$

where we used $\|\nabla F_i(x_t)\| \leq \|\nabla F_i(x^\star)\| + L_\alpha \|d_t\|^\alpha$ in the last inequality. Similarly, we have from (116), (117), and $(a+b)^c \leq 2^{c-1}(a^c + b^c)$ for $c \geq 1$ that

$$\epsilon_t \leq \frac{L_\alpha}{(1+\alpha)mK} \sum_{i=1}^m \sum_{k=0}^{K-1} \|x_{i,k}^t - x_t\|^{1+\alpha} \lesssim \left(\frac{K\tau_t}{q_t} \right)^{1+\alpha} \left(1 + \|d_t\|^{\alpha(1+\alpha)} \right) + \left(\frac{K\tau_t}{q_t} \right)^{\frac{1+\alpha}{1-\alpha}}. \tag{119}$$

Then raising (118) to power $(1+\alpha)/\alpha$ and adding (119) leads to

$$\|b_t\|^{(1+\alpha)/\alpha} + \epsilon_t \lesssim \left(\frac{K\tau_t}{q_t} \right)^{1+\alpha} \left(1 + \|d_t\|^{\alpha(1+\alpha)} \right) + \left(\frac{K\tau_t}{q_t} \right)^{\frac{1+\alpha}{1-\alpha}}. \tag{120}$$

Finally, note that $q_t \geq q_1$, and $\tau_t \lesssim 1/((K-1)t)$ satisfies

$$\sum_{t=1}^T (K\tau_t)^{1+\alpha} \lesssim 1, \qquad \sum_{t=1}^T (K\tau_t)^{(1+\alpha)/(1-\alpha)} \lesssim 1, \quad \forall \alpha \in (0,1).$$

Summing (120) and using $\|d_t\|^{\alpha(1+\alpha)} \leq D_T^{\alpha(1+\alpha)/2}$ thus completes the proof. □

Lemma F.3. *Consider the same setting as in Lemma F.2.and recall $D_T := \max_{1 \leq t \leq T+1} \|d_t\|^2$. Then, for every $T \geq 1$, we have*

$$D_T \lesssim 1 + \log(q_{T+1}/q_1), \tag{121}$$

$$\sum_{t=1}^T q_t \|r_t\|^2 \lesssim q_{T+1} \left(1 + \log(q_{T+1}/q_1) \right). \tag{122}$$

Proof. Let $\Delta_t := F(x_t) - F^\star$ and $p := (1+\alpha)/\alpha > 2$. Since $\bar{g}_t = g_t + b_t$ and $\|\bar{g}_t\| = q_t \|r_t\|$, we have from $(a+b)^c \leq 2^{c-1}(a^c + b^c)$ for $c \geq 1$ that

$$\|g_t\|^p \geq 2^{1-p} \|\bar{g}_t\|^p - \|b_t\|^p = 2^{1-p} q_t^p \|r_t\|^p - \|b_t\|^p,$$

which together with Lemma F.1 gives

$$-\frac{2\Delta_t}{q_t} \leq -\frac{2\alpha L_\alpha^{-1/\alpha}}{1+\alpha} \frac{\|g_t\|^p}{q_t} \leq -\frac{2^{2-p}\alpha L_\alpha^{-1/\alpha}}{1+\alpha} q_t^{p-1} \|r_t\|^p + \frac{2\alpha L_\alpha^{-1/\alpha}}{1+\alpha} \frac{\|b_t\|^p}{q_t}.$$

Therefore, setting

$$c_\alpha := \frac{2^{2-p}\alpha L_\alpha^{-1/\alpha}}{1+\alpha}, \qquad C_\alpha := \frac{2\alpha L_\alpha^{-1/\alpha}}{1+\alpha},$$

we obtain from (61) with $\zeta = 0$ and $q_t \geq q_1$ that

$$\|d_{t+1}\|^2 - \|d_t\|^2 \leq \|r_t\|^2 - \frac{2\Delta_t}{q_t} + \frac{2\epsilon_t}{q_t} \leq \|r_t\|^2 - c_\alpha q_1^{p-1}\|r_t\|^p + \frac{C_\alpha \|b_t\|^p + 2\epsilon_t}{q_t}.$$

Applying Lemma A.3 to bound $\|r_t\|^2 - c_\alpha q_1^{p-1}\|r_t\|^p$, we obtain

$$\begin{aligned}\|d_{t+1}\|^2 - \|d_t\|^2 &\leq \left(\eta^2 + \left(c_\alpha q_1^{p-1}\right)^{-2/(p-2)}\right)\log\left(1 + \frac{\|r_t\|^2}{\eta^2}\right) + \frac{C_\alpha \|b_t\|^p + 2\epsilon_t}{q_t} \\ &\lesssim \log\left(1 + \frac{\|r_t\|^2}{\eta^2}\right) + \frac{\|b_t\|^p + \epsilon_t}{q_t}. \end{aligned} \tag{123}$$

The definitions of r_t and q_{t+1} give

$$\frac{\|r_t\|^2}{\eta^2} = \frac{\|\bar{g}_t\|^2}{\eta^2 q_t^2} = \frac{q_{t+1}^2 - q_t^2}{q_t^2}.$$

Consequently, for every $1 \leq s \leq T$,

$$\sum_{t=1}^{s} \log\left(1 + \frac{\|r_t\|^2}{\eta^2}\right) = \sum_{t=1}^{s} \log \frac{q_{t+1}^2}{q_t^2} = 2\log\frac{q_{s+1}}{q_1} \leq 2\log\frac{q_{T+1}}{q_1}.$$

Moreover, $q_t \geq q_1$ and Lemma F.2 imply

$$\sum_{t=1}^{s} \frac{\|b_t\|^p + \epsilon_t}{q_t} \leq \frac{1}{q_1}\sum_{t=1}^{T}\left(\|b_t\|^p + \epsilon_t\right) \lesssim 1 + D_T^{\alpha(1+\alpha)/2}.$$

Therefore, summing (123) from $t = 1$ to s and maximizing over $s \in \{1, \ldots, T\}$ give

$$D_T \lesssim 1 + \log(q_{T+1}/q_1) + D_T^{\alpha(1+\alpha)/2}.$$

Since $\alpha(1+\alpha)/2 < 1$, solving the above inequality yields (121).

We proceed to prove (122). Since $\nabla F(x^\star) = 0$, the global Hölder condition gives

$$\|g_t\|^2 = \|\nabla F(x_t) - \nabla F(x^\star)\|^2 \leq L_\alpha^2 \|d_t\|^{2\alpha} \leq L_\alpha^2 D_T^\alpha.$$

Moreover, since $K\tau_t/q_t \lesssim 1$, (118) therefore yields

$$\|b_t\|^2 \lesssim 1 + \|d_t\|^{2\alpha^2} \lesssim 1 + D_T^\alpha.$$

Thus, using $\bar{g}_t = g_t + b_t$ and (121), we further obtain

$$\frac{\|\bar{g}_t\|^2/\eta^2}{q_t^2} \leq \frac{2}{\eta^2 q_1^2}\left(\|g_t\|^2 + \|b_t\|^2\right) \lesssim 1 + D_T^\alpha \lesssim \left(1 + \log(q_{T+1}/q_1)\right)^\alpha.$$

Applying Lemma A.4 to $A_t = q_t^2$, $a_t = \|\bar{g}_t\|^2/\eta^2$, and $c = \left(1 + \log(q_{T+1}/q_1)\right)^\alpha$ yields

$$\begin{aligned}\sum_{t=1}^{T} q_t \|r_t\|^2 = \eta^2 \sum_{t=1}^{T} \frac{\|\bar{g}_t\|^2/\eta^2}{q_t} &\leq \eta^2(1 + \sqrt{1+c})(q_{T+1} - q_1) \\ &\lesssim q_{T+1}\left(1 + \left(1 + \log\frac{q_{T+1}}{q_1}\right)^{\alpha/2}\right) \lesssim q_{T+1}\left(1 + \log\frac{q_{T+1}}{q_1}\right),\end{aligned}$$

which is (122), hence completing the proof. □

F.2 Main Theorem

Theorem F.4 (Hölder-smooth convex convergence)**.** *Let Assumption 4.1 hold. Assume each F_i is convex and (L_α, α)-Hölder smooth for some $\alpha \in [0, 1]$. Let $\{x_t\}_{t\geq 1}$ be generated by Algorithm 1 with exact local gradients $g_{i,k}^t = \nabla F_i(x_{i,k}^t)$, τ_t chosen as* (5)*, and arbitrary $\gamma, \eta > 0$. Then, for every $T \geq 1$,*

$$\mathbb{E}[F(\bar{x}_T) - F^\star] \lesssim \begin{cases} \dfrac{(\log(1+T))^{1+\alpha}}{T^{(1+\alpha)/2}}, & 0 \leq \alpha < 1, \\ \dfrac{1}{T}, & \alpha = 1. \end{cases}$$

Proof of Theorem F.4. The special cases $\alpha = 1$ and $\alpha = 0$ are covered by Corollaries C.4 and D.5, respectively. It remains to consider $\alpha \in (0, 1)$. Let $\Delta_t := F(x_t) - F^\star$. We start from (61) with $\zeta = 0$ to obtain

$$\Delta_t \le \frac{q_t}{2}\big(\|d_t\|^2 - \|d_{t+1}\|^2\big) + \frac{q_t}{2}\|r_t\|^2 + \epsilon_t.$$

Summing this inequality over $t = 1, \dots, T$, plugging in (79), and applying Lemmas F.2 and F.3 then yield

$$\sum_{t=1}^{T} \Delta_t \le q_T D_T + \frac{1}{2}\sum_{t=1}^{T} q_t\|r_t\|^2 + \sum_{t=1}^{T}\epsilon_t \lesssim q_{T+1}\left(1 + \log\frac{q_{T+1}}{q_1}\right). \tag{124}$$

It remains to bound q_{T+1}. Applying Hölder's inequality to $\{1\}_{t=1}^T$ and $\{\|g_t\|^2\}_{t=1}^T$ with conjugate exponents $p := \frac{1+\alpha}{1-\alpha}$ and $q := \frac{1+\alpha}{2\alpha}$ and then using Lemma F.1, we obtain

$$\sum_{t=1}^{T}\|g_t\|^2 \le T^{\frac{1-\alpha}{1+\alpha}}\left(\sum_{t=1}^{T}\|g_t\|^{\frac{1+\alpha}{\alpha}}\right)^{\frac{2\alpha}{1+\alpha}} \lesssim T^{\frac{1-\alpha}{1+\alpha}}\left(\sum_{t=1}^{T}\Delta_t\right)^{\frac{2\alpha}{1+\alpha}}.$$

Moreover, applying Hölder's inequality with the same exponents to $\{1\}_{t=1}^T$ and $\{\|b_t\|^2\}_{t=1}^T$, and then using Lemma F.2 and (121), gives

$$\sum_{t=1}^{T}\|b_t\|^2 \le T^{\frac{1-\alpha}{1+\alpha}}\left(\sum_{t=1}^{T}\|b_t\|^{\frac{1+\alpha}{\alpha}}\right)^{\frac{2\alpha}{1+\alpha}} \lesssim T^{\frac{1-\alpha}{1+\alpha}}\left(1 + D_T^{\frac{\alpha(1+\alpha)}{2}}\right)^{\frac{2\alpha}{1+\alpha}} \lesssim T^{\frac{1-\alpha}{1+\alpha}}\left(1 + \log\frac{q_{T+1}}{q_1}\right)^{\alpha^2}.$$

Using $\|\bar g_t\|^2 \le 2\|g_t\|^2 + 2\|b_t\|^2$, the above two estimates, and (124), we obtain

$$\begin{aligned} q_{T+1}^2 &= q_1^2 + \frac{1}{\eta^2}\sum_{t=1}^{T}\|\bar g_t\|^2 \lesssim \sum_{t=1}^{T}\|g_t\|^2 + \sum_{t=1}^{T}\|b_t\|^2 \\ &\lesssim T^{\frac{1-\alpha}{1+\alpha}} q_{T+1}^{\frac{2\alpha}{1+\alpha}}\left(1 + \log\frac{q_{T+1}}{q_1}\right)^{\frac{2\alpha}{1+\alpha}} + T^{\frac{1-\alpha}{1+\alpha}}\left(1 + \log\frac{q_{T+1}}{q_1}\right)^{\alpha^2}. \end{aligned}$$

Taking square roots and using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, then applying to the first term the Young's inequality

$$ab \le \frac{\lambda^p}{p}a^p + \frac{\lambda^{-q}}{q}b^q, \qquad \frac{1}{p} + \frac{1}{q} = 1$$

with $p := (1+\alpha)/\alpha$, $q := 1 + \alpha$, $\lambda := (p/2)^{1/p}$ to isolate $q_{T+1}/2$, we derive for some constant $C > 0$ that

$$\begin{aligned} q_{T+1} \le& C q_{T+1}^{\frac{\alpha}{1+\alpha}} T^{\frac{1-\alpha}{2(1+\alpha)}}\left(1 + \log\frac{q_{T+1}}{q_1}\right)^{\frac{\alpha}{1+\alpha}} + C T^{\frac{1-\alpha}{2(1+\alpha)}}\left(1 + \log\frac{q_{T+1}}{q_1}\right)^{\alpha^2/2} \\ \le& \frac{q_{T+1}}{2} + \frac{\lambda^{-q}}{q} C^{1+\alpha} T^{\frac{1-\alpha}{2}}\left(1 + \log\frac{q_{T+1}}{q_1}\right)^{\alpha} + C T^{\frac{1-\alpha}{2(1+\alpha)}}\left(1 + \log\frac{q_{T+1}}{q_1}\right)^{\alpha^2/2}. \end{aligned} \tag{125}$$

Note that the last term in (125) is bounded by the second term up to a constant since $\alpha \ge \alpha^2/2$ and $1 + \alpha \ge 1$. Therefore, simplifying (125) gives

$$\frac{q_{T+1}}{q_1} \lesssim T^{\frac{1-\alpha}{2}}\left(1 + \log\frac{q_{T+1}}{q_1}\right)^{\alpha}.$$

Let $x := (q_{T+1}/q_1)^{1/\alpha}$, then it follows from $1 + \log(q_{T+1}/q_1) \ge 0$ that

$$x \lesssim T^{\frac{1-\alpha}{2\alpha}}(1 + \alpha\log x),$$

solving which using Lemma A.1 and recovering to q_{T+1} gives

$$q_{T+1} \lesssim T^{\frac{1-\alpha}{2}}(1 + \log T)^{\alpha}.$$

Substituting it into (124) yields

$$\sum_{t=1}^{T}\Delta_t \lesssim T^{\frac{1-\alpha}{2}}(1 + \log T)^{1+\alpha},$$

which completes the proof by noting that $\mathbb{E}[F(\bar x_T) - F^\star] = T^{-1}\sum_{t=1}^{T}\Delta_t$. □

G Additional Experiments

G.1 Data construction and reference metrics

The synthetic logistic problem has $m = 32$ clients, $n_i = 64$ examples per client ($n = 2048$), and dimension $d = 32$. A deterministic Halton design is mapped through the standard-normal quantile, centered, and orthogonalized. Its singular values make the eigenvalues of $A^\top A/n$ a geometric sequence from 1 to $1/\kappa$, where $\kappa \in \{1, 10, 100\}$. Labels are deterministic Bernoulli-logistic realizations from $x_{\mathrm{true}} = 2e_1 - e_{11} + 0.5e_{22}$. Examples are sorted by planted margin, divided into blocks of 32, and dealt round-robin to clients to create fixed heterogeneity.

For IJCNN1, we take the first 32,000 rows of the LIBSVM binary-classification dataset, with $d = 22$ features and $n_i = 1000$ examples on each of 32 clients. We do not append an intercept or rescale rows or columns. Examples are sorted by margin under a high-accuracy centralized ℓ_1-logistic reference solution, divided into blocks of 50, and dealt round-robin to clients. The three penalties and metrics are defined in Section 5. We compute $F^\star$ to high precision using damped Newton for the squared-ℓ_2 problem and FISTA followed by active-set Newton polishing for the ℓ_1 problem.

G.2 Experimental setup and tuning

Run protocol. Exact local gradients use all n_i examples. Stochastic gradients sample independently with replacement for every (round, client, local-step) tuple. The primary experiment uses minibatch size $b = 1$; the additional panels use $b = 8$ for synthetic clients and $b = 32$ for IJCNN1 clients. Stochastic curves are means over seeds 2027–2036. Exact-gradient trajectories are deterministic and shared by the figure sets. The solid curve averages metric values along the trajectory, which represents expected value for a uniformly random iterate chosen from up to that iteration. The dashed last-iterate curves.

Step-size selection. We fix $\gamma = 1$ in all experiments unless specified otherwise. For each dataset family, regime, oracle type, and stochastic batch size, we tune η at $K = 10$ and then hold it fixed across K. The grid consists of integer powers of ten. If the best value is on the boundary, we extend the grid until the selected point is interior. Synthetic tuning is performed only at $\kappa = 10$ and reused at $\kappa \in \{1, 100\}$. The selected values are:

Data	Regime	Exact	Stochastic $b = 1$	Additional stochastic
Synthetic	Smooth nonconvex	10	1	10 ($b = 8$)
Synthetic	Smooth convex	10	1	1 ($b = 8$)
Synthetic	Nonsmooth convex	1	1	1 ($b = 8$)
IJCNN1	Smooth nonconvex	10	10	10 ($b = 32$)
IJCNN1	Smooth convex	10	10	10 ($b = 32$)
IJCNN1	Nonsmooth convex	10	10	10 ($b = 32$)

All selections are interior grid points. Because η is retuned for each batch size, the primary and additional panels are separately tuned demonstrations rather than a causal batch-size ablation.

G.3 AdaFed rate study on synthetic data

The synthetic rate study uses $T = 2000$ reported rounds, full participation, $\gamma = 1$, and $K \in \{1, 10, 100\}$. The local schedule is $\tau_t = 1$ for $K = 1$ and $\tau_t = 1/((K-1)t)$ otherwise. Exact gradients use all local examples; stochastic gradients use minibatches of size $b = 1$ and pointwise means over seeds 2027–2036. The corresponding IJCNN1 study in Figure 1 uses $T = 10^4$ rounds.

With exact gradients, the two smooth metrics decay approximately as $1/T$, while the nonsmooth convex gap exhibits a slower trend consistent with $1/\sqrt{T}$. The exact-gradient curves nearly coincide across K. With $b = 1$, increasing K from 1 to 100 improves the final synthetic $\kappa = 10$ uniform-random metric by factors 3.63, 4.76, and 3.57 for smooth nonconvex, smooth convex, and nonsmooth convex objectives.

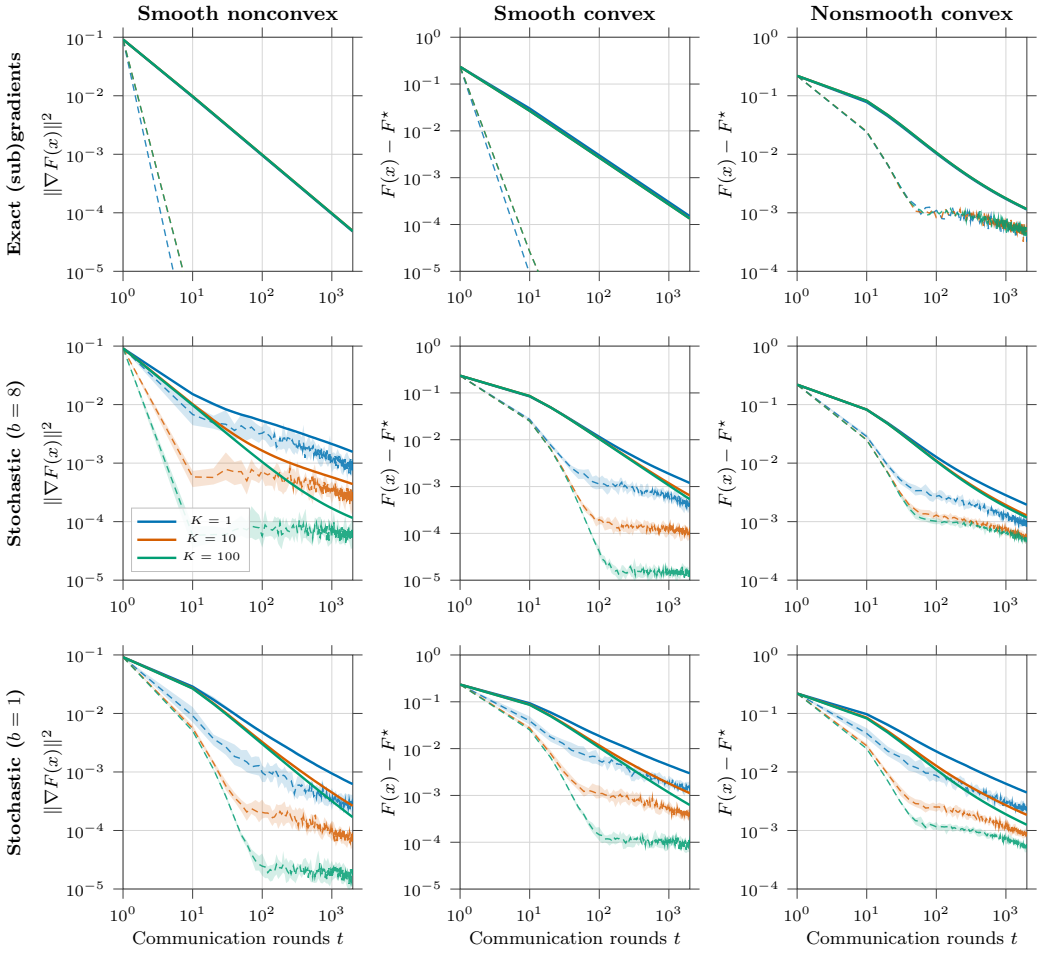

Figure 3: Performance of AdaFed on synthetic logistic problems with Gram condition number $\kappa = 10$ across the three problem classes and different numbers of local steps K. Rows show exact gradients and stochastic gradients with minibatch sizes $b = 1$ and $b = 8$, respectively. Solid and dashed curves show the uniform-random and last-iterate metrics, respectively. Stochastic results show the mean over ten seeds, with shaded regions indicating the upward and downward RMS deviations.

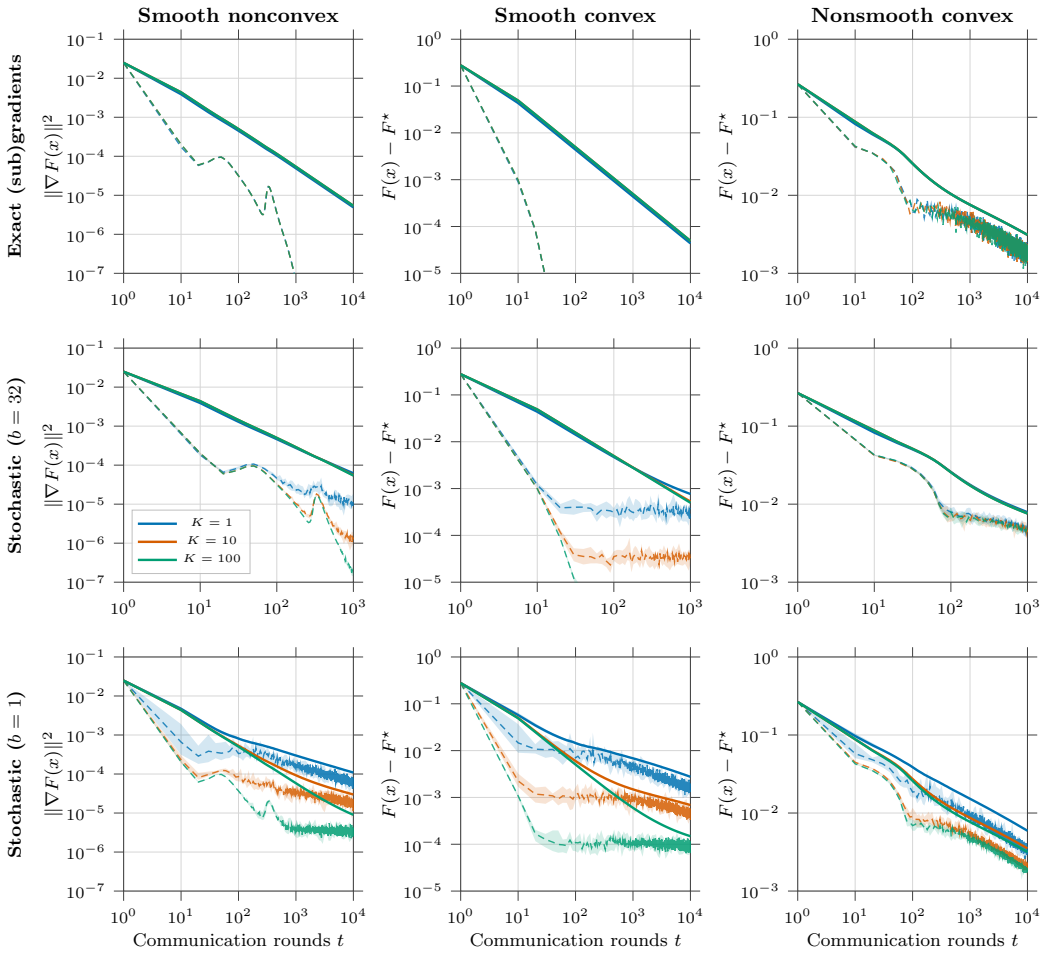

Figure 4: Performance of AdaFed on IJCNN1 across the three problem classes and different numbers of local steps K. Rows show exact gradients and stochastic gradients with minibatch sizes $b = 1$ and $b = 32$, respectively. Solid and dashed curves show the uniform-random and last-iterate metrics, respectively. Stochastic results show the mean over ten seeds, with shaded regions indicating the upward and downward RMS deviations.

G.4 SENSITIVITY TO η AND γ

The two scalar parameters play different roles. The parameter η sets the numerator of the adaptive shared step, while γ initializes its accumulator; in particular, $\eta_1 = \eta/\gamma$. We vary one parameter on an absolute power-of-ten grid while fixing the other, with $K = 10$ and full participation. Exact-gradient curves are deterministic. Stochastic curves use $b = 1$ and are pointwise means over seeds 2027–2036.

The η panels verify that the selected values are interior to the tested grids rather than artifacts of a truncated search. The response to γ is nonmonotone because changing γ also changes the initial step η/γ. On IJCNN1, $\gamma = 1$ gives the smallest final random-output metric in five of the six row–regime combinations; the exact nonconvex row favors 0.1. On the synthetic stochastic rows, 0.1 has the smallest final random-output metric, although $\gamma = 1$ remains close and is used throughout the controlled rate study. Large γ consistently slows the initial transient.

G.5 PARTIAL PARTICIPATION

For an empirical experiment, each client participates independently with probability $p \in \{0.25, 0.5, 1\}$. The server uses the same Horvitz–Thompson-scaled update as the implementation, so its conditional mean matches the full-participation update. We fix $K = 10$, $\gamma = 1$, and the corresponding tuned η. Whenever participation is random, curves are pointwise means over seeds 2027–2036; the full-participation exact curve is deterministic.

Reducing participation increases the stochastic error level in every regime. At the final checkpoint, $p = 0.25$ raises the stochastic random-output metric relative to $p = 1$ by factors between 1.66 and 1.79 on synthetic data and between 1.35 and 2.54 on IJCNN1. The effect is much larger for the exact-gradient smooth synthetic problems because client subsampling becomes their only source of randomness.

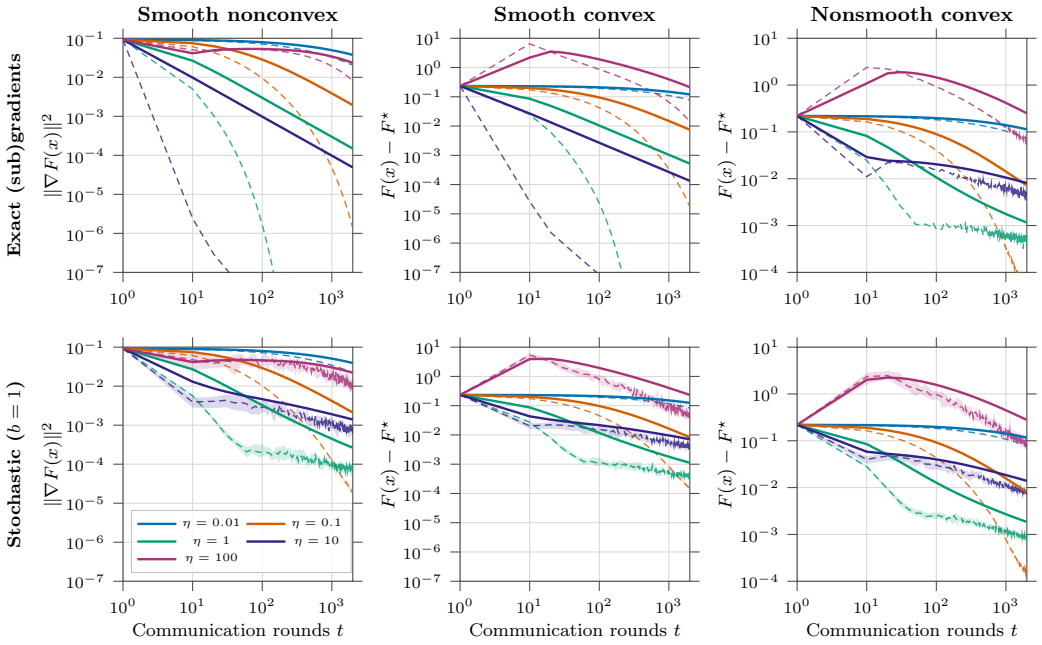

Figure 5: Sensitivity of AdaFed to η on synthetic logistic problems with Gram condition number $\kappa = 10$, with $\gamma = 1$ and $K = 10$. Solid and dashed curves show the uniform-random and last-iterate metrics, respectively. Stochastic results show the mean over ten seeds, with shaded regions indicating the upward and downward RMS deviations.

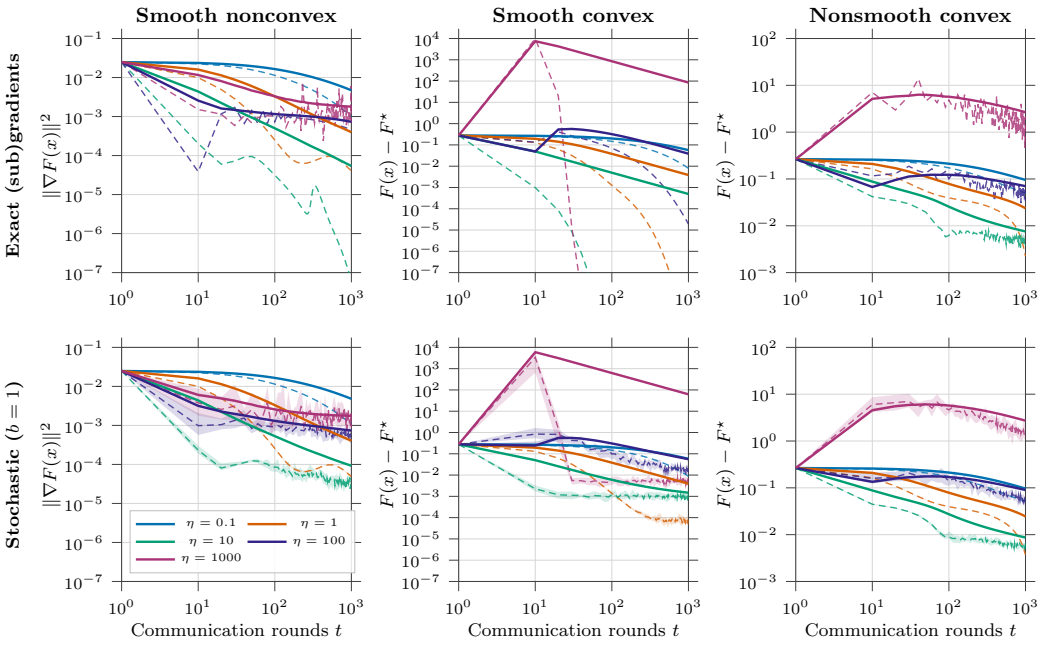

Figure 6: Sensitivity of AdaFed to η on IJCNN1, with $\gamma = 1$ and $K = 10$. Solid and dashed curves show the uniform-random and last-iterate metrics, respectively. Stochastic results show the mean over ten seeds, with shaded regions indicating the upward and downward RMS deviations

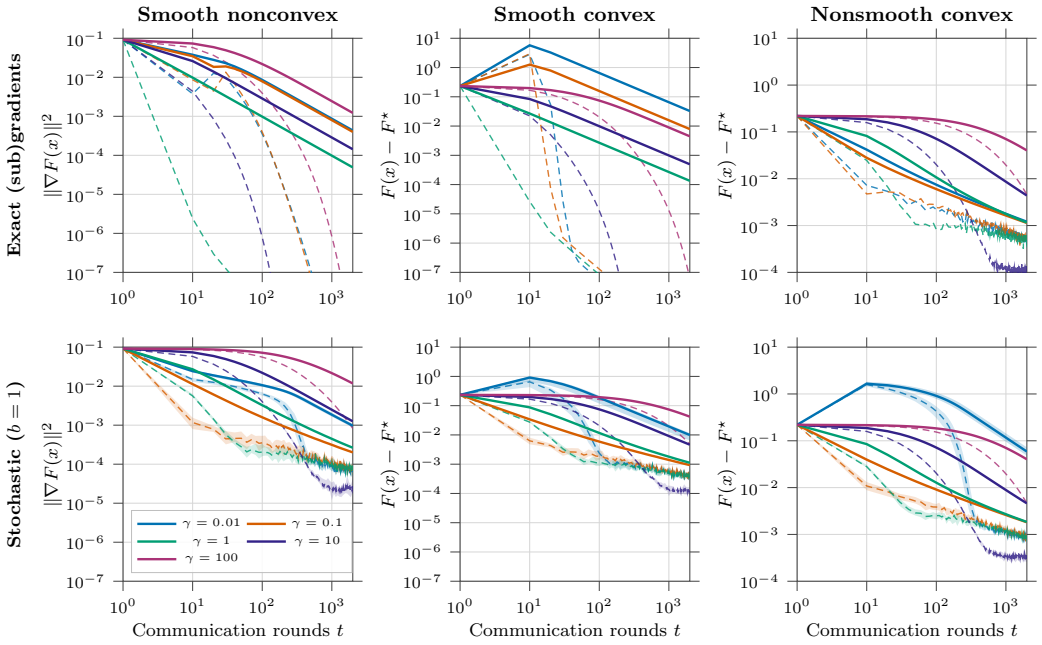

Figure 7: Sensitivity of AdaFed to γ on synthetic logistic problems with Gram condition number $\kappa = 10$, with $\eta = 1$ and $K = 10$. Solid and dashed curves show the uniform-random and last-iterate metrics, respectively. Stochastic results show the mean over ten seeds, with shaded regions indicating the upward and downward RMS deviations.

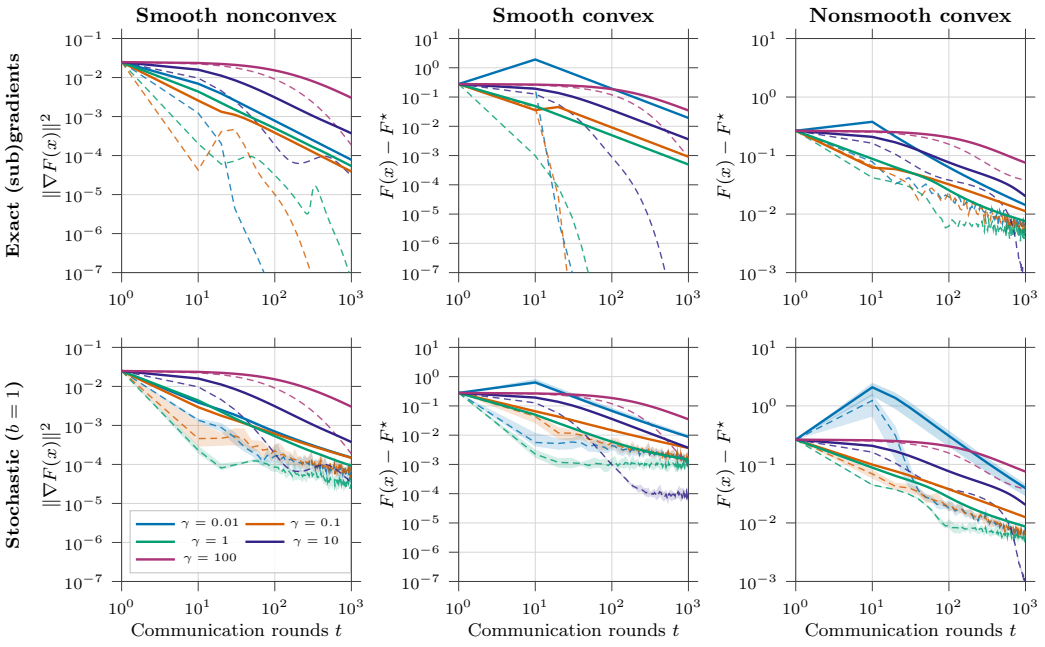

Figure 8: Sensitivity of AdaFed to γ on IJCNN1, with $\eta = 1$ and $K = 10$. Solid and dashed curves show the uniform-random and last-iterate metrics, respectively. Stochastic results show the mean over ten seeds, with shaded regions indicating the upward and downward RMS deviations.

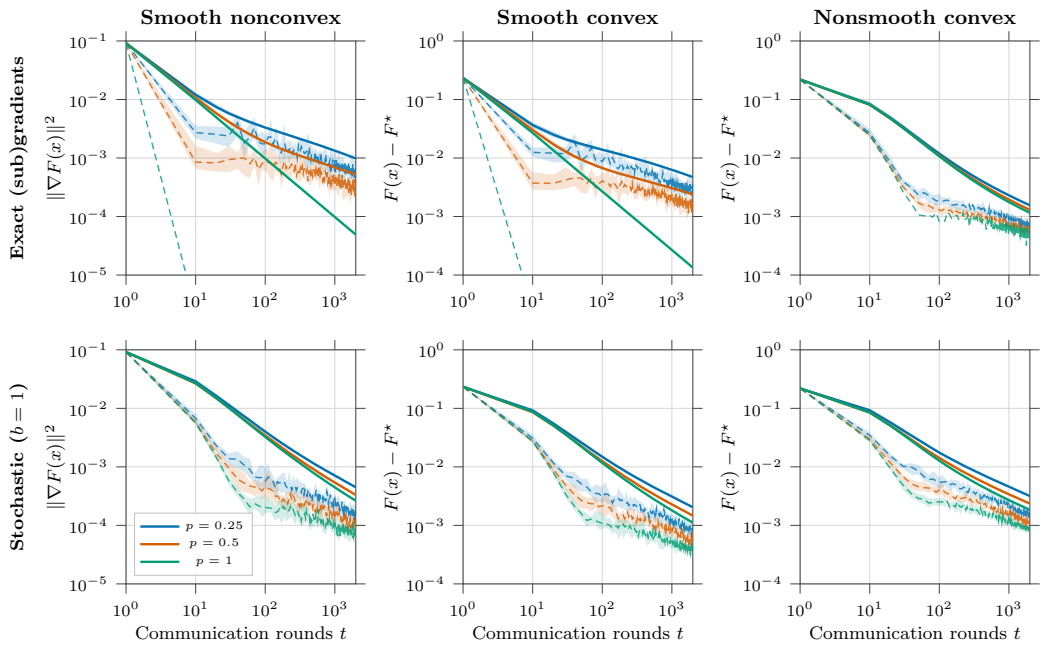

Figure 9: Performance of AdaFed under partial participation on synthetic logistic problems with Gram condition number $\kappa = 10$ and different participation rates p. Solid and dashed curves show the uniform-random and last-iterate metrics, respectively. Stochastic results show the mean over ten seeds, with shaded regions indicating the upward and downward RMS deviations.

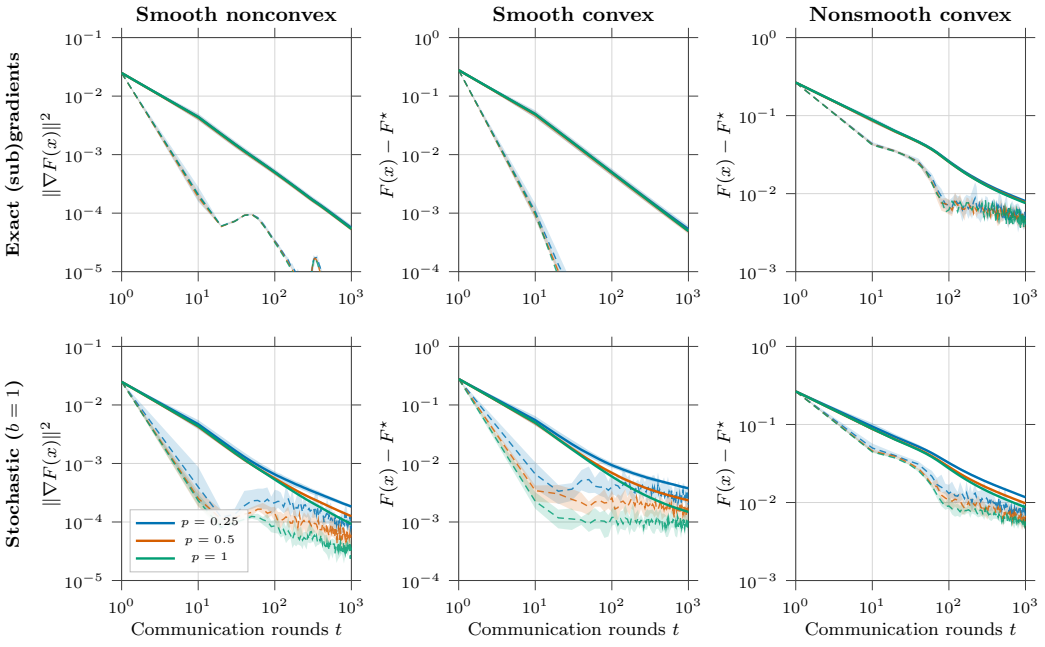

Figure 10: Performance of AdaFed under partial participation on IJCNN1 with different participation rates p. Solid and dashed curves show the uniform-random and last-iterate metrics, respectively. Stochastic results show the mean over ten seeds, with shaded regions indicating the upward and downward RMS deviations.

G.6 ADAPTIVE-METHOD COMPARISONS

These experiments use full participation of all 32 clients, $K = 10$ local steps, and $T = 10^4$ rounds on synthetic data with $\kappa = 1$ and on IJCNN1. Exact local gradients use each client's entire sample. Stochastic gradients use minibatches sampled independently with replacement, of size $b = 16$ for synthetic data and $b = 32$ for IJCNN1. The exact trajectories use seed 2027; stochastic curves are pointwise means of seeds 2027–2029. All methods start from the zero vector and are evaluated using the same full-objective metrics as in Section 5. The solid and dashed line meanings match Figure 2.

Each method is tuned separately in each dataset–objective–oracle cell. Candidate runs use seed 1901 and 300 reported rounds on synthetic data or 150 on IJCNN1. The tuning criterion is the final cumulative-mean metric, namely the objective gap for convex problems and squared full pooled gradient norm for the smooth nonconvex problem. Initial Cartesian grids are extended by a factor of ten along an axis when the selected value lies at its edge, subject to the finite bounds in Table 2. The resulting settings are held fixed throughout the 10,000-round runs; they are short-horizon selections rather than parameters optimized for the full horizon. The selected four-method comparison and the wider adaptation-level comparisons reuse the same AdaFed, FedAdagrad, FedLADA, and FAFED runs. AdaFed jointly tunes its η and accumulator initialization γ; its local schedule remains $\tau_t = 1/((K-1)t)$. For baselines, c/L denotes a local learning rate using the pooled logistic smoothness constant L.

Table 2: Tuned quantities and finite search bounds for the adaptive-method comparisons. η_g is a server rate, c or c_0 is a local scale, g is a FedSPS cap scale, and ϵ is Local AdaAlter's placeholder.

Method	Tuned quantities	Finite search bounds
AdaFed	η, γ	$\eta \in [10^{-3}, 10^{3}]$; $\gamma \in [10^{-3}, 10^{3}]$
FedAdagrad	c, η_g	$c \in [1/128000, 20]$, $\eta_g \in [10^{-3}, 100]$
FedAdam	c, η_g	$c \in [1/128000, 20]$, $\eta_g \in [10^{-3}, 100]$
FedUR	c_0, α	$c_0 \in [10^{-4}, 10]$, $\alpha \in [0.1, 1]$
Local AdaAlter	c, ϵ	$c \in [10^{-5}, 100]$, $\epsilon \in [10^{-4}, 10]$
FedSPS	c, g	$c \in [10^{-3}, 1000]$, $g \in [10^{-4}, 100]$
FedLADA	c, η_g	$c \in [10^{-5}, 10]$, $\eta_g \in [10^{-2}, 10]$
FAFED	c	$c \in [10^{-5}, 0.1]$; $\eta_g = 1$ fixed
FedAda^2	c, η_g	$c \in [10^{-6}, 10]$, $\eta_g \in [10^{-3}, 100]$

Initial discrete grids. The tested starting grids, in the order of the tuned quantities in Table 2, are as follows:

- AdaFed: $\eta \in \{0.01, 0.1, 1, 10, 100\}$ and $\gamma \in \{0.01, 0.1, 1, 10, 100\}$.
- FedAdagrad: $c \in \{1/128, 1/32, 1/8, 1/2, 2\}$ and $\eta_g \in \{0.001, 0.01, 0.1, 1, 10\}$.
- FedAdam: $c \in \{1/128, 1/8, 2\}$ and $\eta_g \in \{0.01, 0.1, 1, 10\}$.
- FedUR: $c_0 \in \{0.001, 0.01, 0.1, 1\}$ and $\alpha \in \{0.1, 0.5, 0.9, 1\}$.
- Local AdaAlter: $c \in \{0.001, 0.01, 0.1, 1, 10\}$ and $\epsilon \in \{0.01, 0.1, 1\}$.
- FedSPS: $c \in \{0.1, 1, 10, 100\}$ and $g \in \{0.01, 0.1, 1, 10\}$.
- FedLADA: $c \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and $\eta_g \in \{0.1, 1, 10\}$.
- FAFED: $c \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$.
- FedAda^2: $c \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ and $\eta_g \in \{0.01, 0.1, 1, 10\}$.

Selected methods on synthetic data. Figure 11 is the synthetic counterpart of the main-text IJCNN1 comparison. AdaFed has the smallest final cumulative-mean metric in all six synthetic settings among this four-method set. The nonsmooth runs of FedLADA and FAFED use the selected $\mathrm{sign}(0) = 0$ subgradient and are empirical extensions beyond the smooth-objective assumptions in their cited analyses.

Server adaptation. Figures 12 and 13 add FedAdam and FedUR to FedAdagrad as server-side comparators. AdaFed has the lowest final cumulative-mean metric in all 12 dataset–objective–oracle

cells of this group. This is an empirical comparison at equal communication rounds; different methods may incur different local computational costs.

Client adaptation. Figures 14 and 15 compare AdaFed with Local AdaAlter, FedSPS, and FedLADA. AdaFed has the smallest final cumulative-mean metric in 10 of the 12 cells. Local AdaAlter is lower in the synthetic stochastic smooth nonconvex and smooth convex cells, with values 3.009×10^{-5} versus 5.401×10^{-5} and 9.025×10^{-5} versus 1.647×10^{-4}, respectively.

Adaptation at both levels. Figures 16 and 17 compare AdaFed with FAFED and FedAda^2. AdaFed has the smallest final cumulative-mean metric in all 12 cells. FAFED uses about two local gradient evaluations per tracker step; FedAda^2 uses coordinatewise AdaGrad at both levels with client accumulators restarted each round. Their nonsmooth results, like those of the other baselines, should be read as empirical tests of the implemented subgradient extensions.

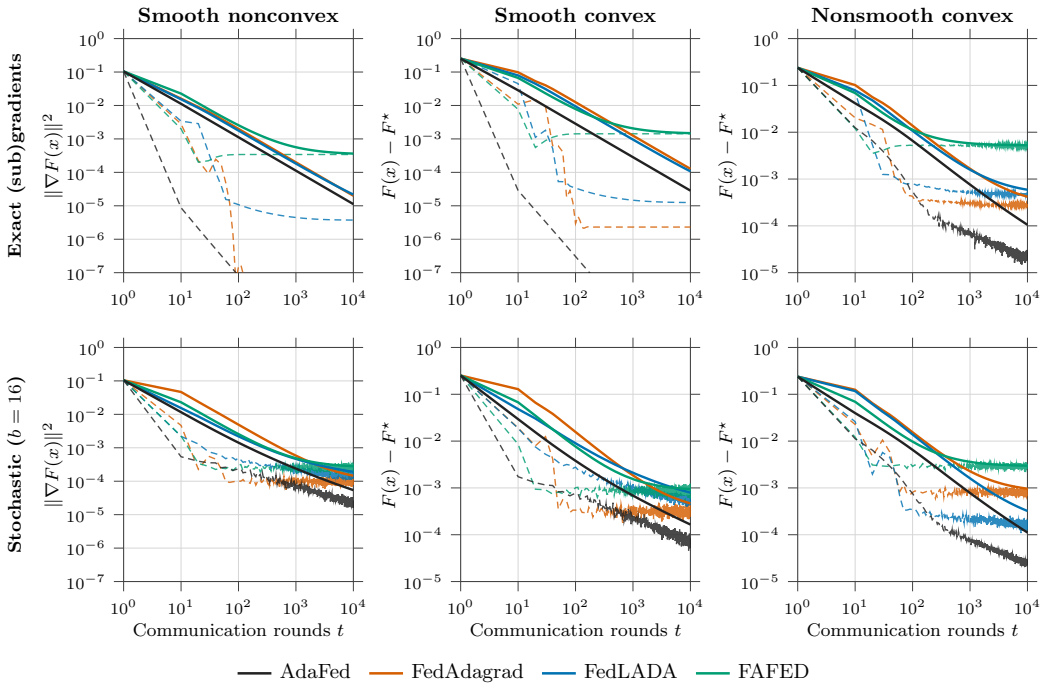

Figure 11: Performance of AdaFed compared with representative adaptive FL methods on synthetic data with $\kappa = 1$ and $K = 10$ local steps. FedAdagrad, FedLADA, and FAFED represent server-, client-, and both-level adaptation, respectively. Solid and dashed curves show the uniform-random and last-iterate metrics, respectively. Stochastic results are shown as the mean over three seeds.

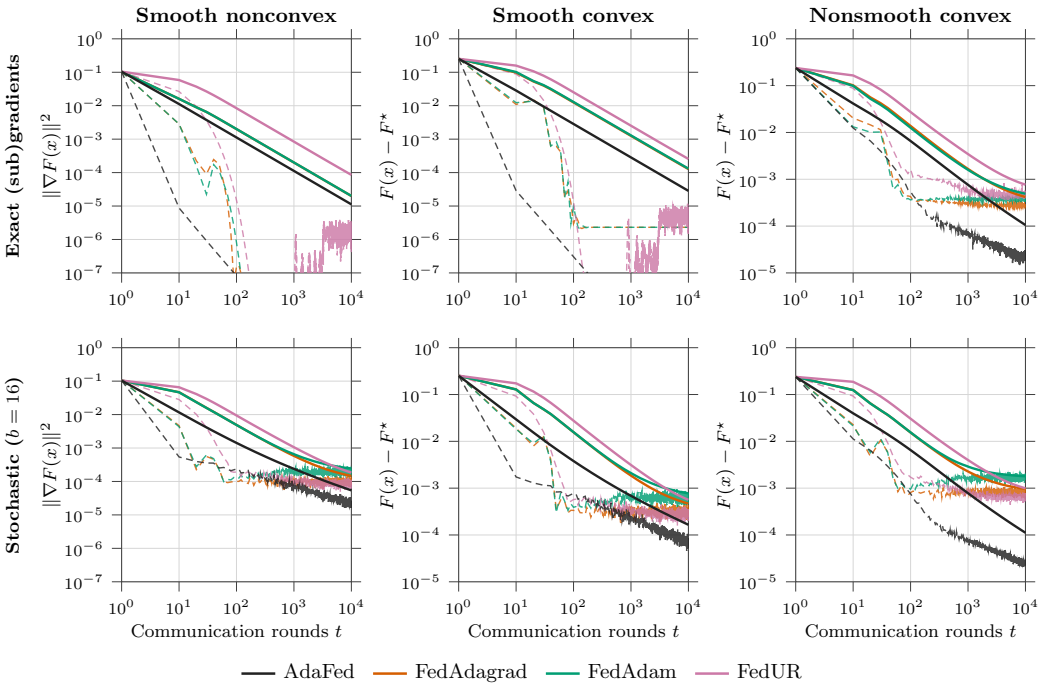

Figure 12: Performance of AdaFed compared with server-adaptive FL methods on synthetic data with $\kappa = 1$ and $K = 10$ local steps. FedAdagrad, FedAdam, and FedUR are server-level adaptive methods. Solid and dashed curves show the uniform-random and last-iterate metrics, respectively. Stochastic results are shown as the mean over three seeds.

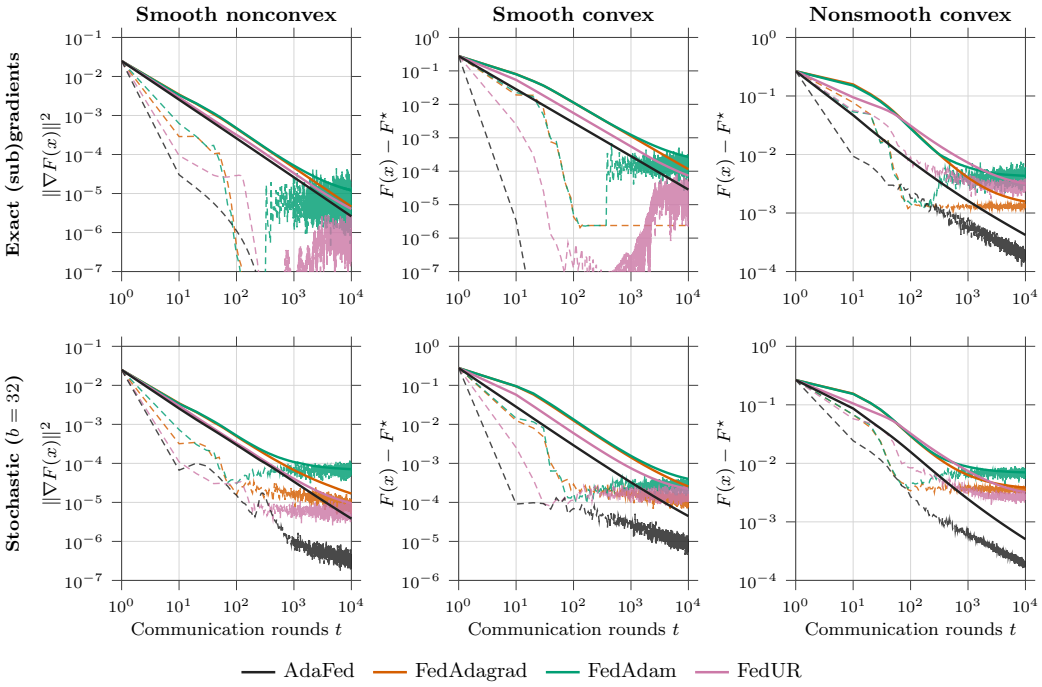

Figure 13: Performance of AdaFed compared with server-adaptive FL methods on IJCNN1 with $K = 10$ local steps. FedAdagrad, FedAdam, and FedUR represent server-level adaptation. Solid and dashed curves show the uniform-random and last-iterate metrics, respectively. Stochastic results are shown as the mean over three seeds.

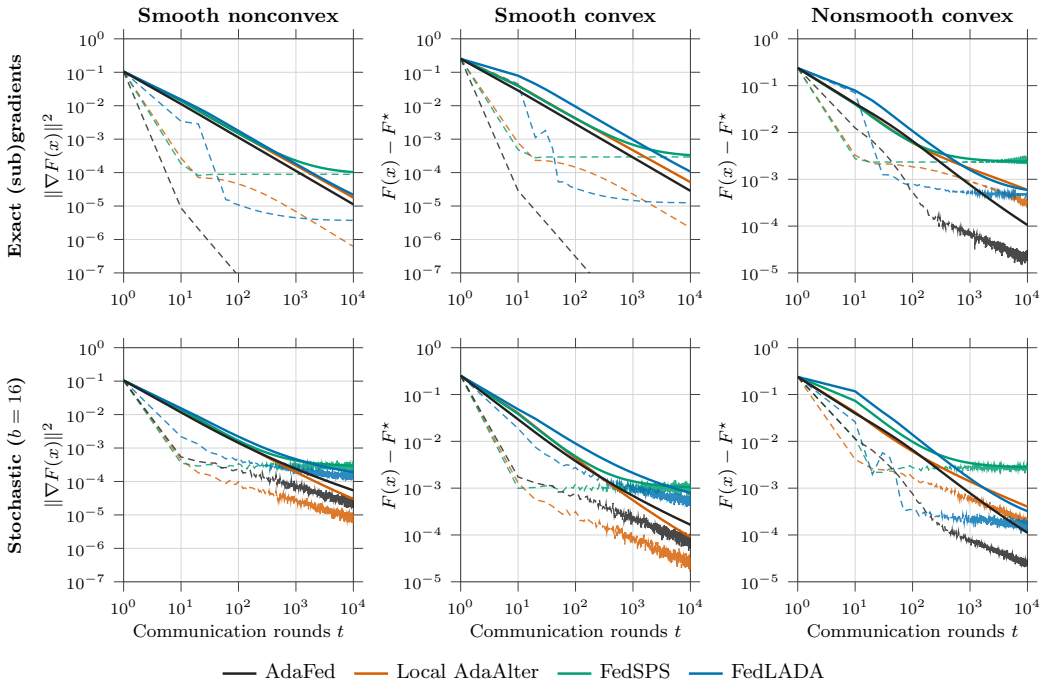

Figure 14: Performance of AdaFed compared with client-adaptive FL methods on synthetic data with $\kappa = 1$ and $K = 10$ local steps. Local AdaAlter, FedSPS, and FedLADA represent client-level adaptation. Solid and dashed curves show the uniform-random and last-iterate metrics, respectively. Stochastic results are shown as the mean over three seeds.

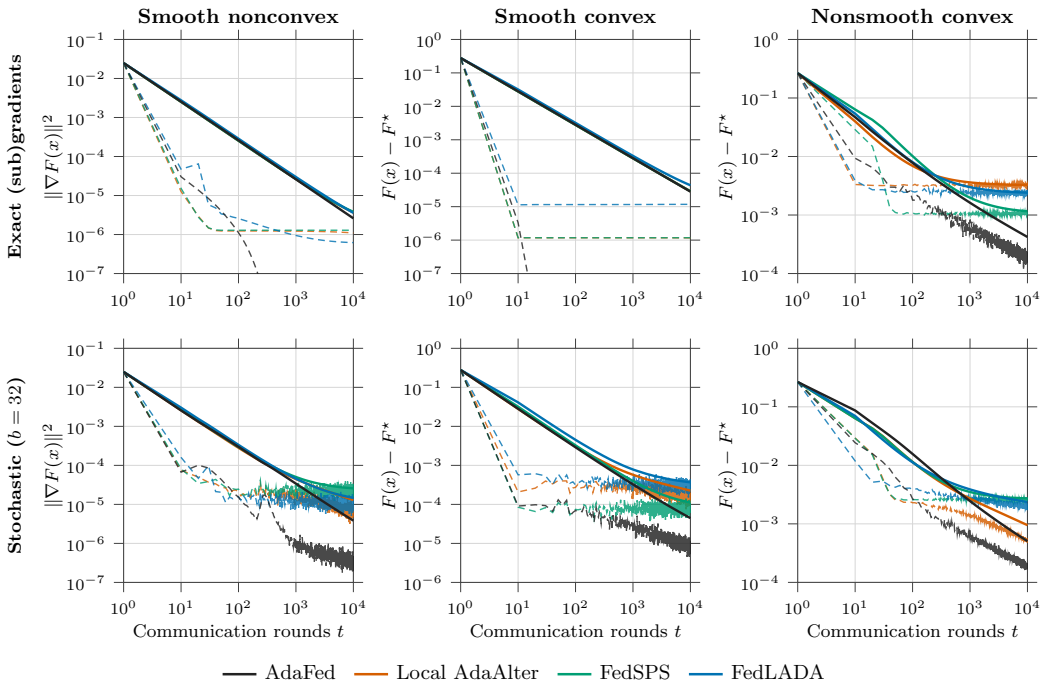

Figure 15: Performance of AdaFed compared with client-adaptive FL methods on IJCNN1 with $K = 10$ local steps. Local AdaAlter, FedSPS, and FedLADA represent client-level adaptation. Solid and dashed curves show the uniform-random and last-iterate metrics, respectively. Stochastic results are shown as the mean over three seeds.

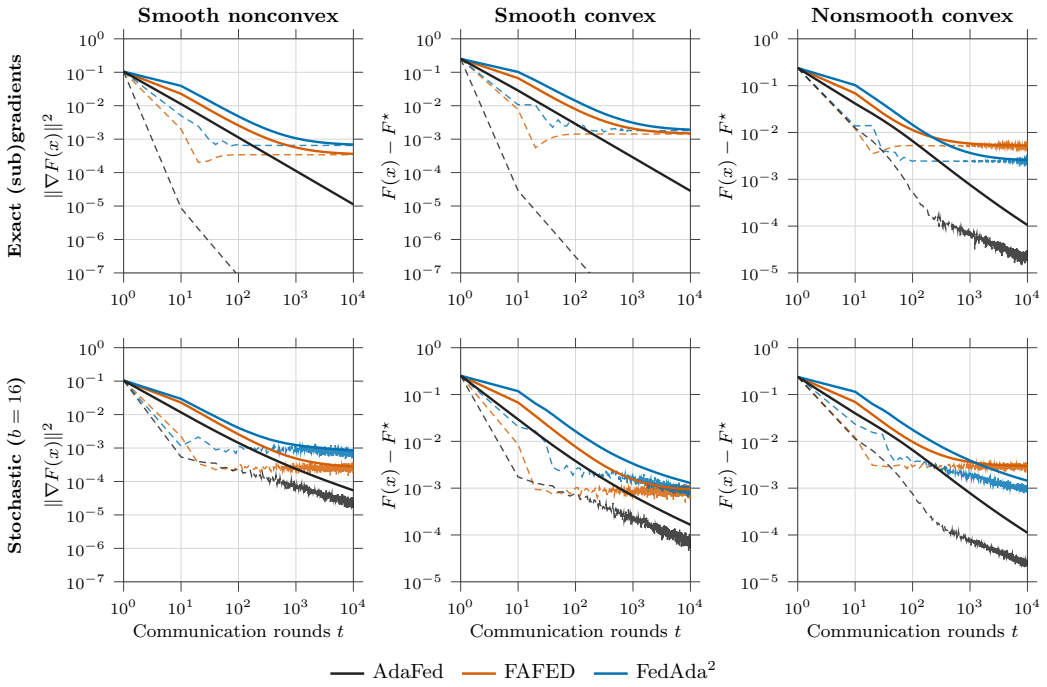

Figure 16: Performance of AdaFed compared with FL methods using both server- and client-level adaptation on synthetic data with $\kappa = 1$ and $K = 10$ local steps. FAFED and FedAda^2 represent both-level adaptation. Solid and dashed curves show the uniform-random and last-iterate metrics, respectively. Stochastic results are shown as the mean over three seeds.

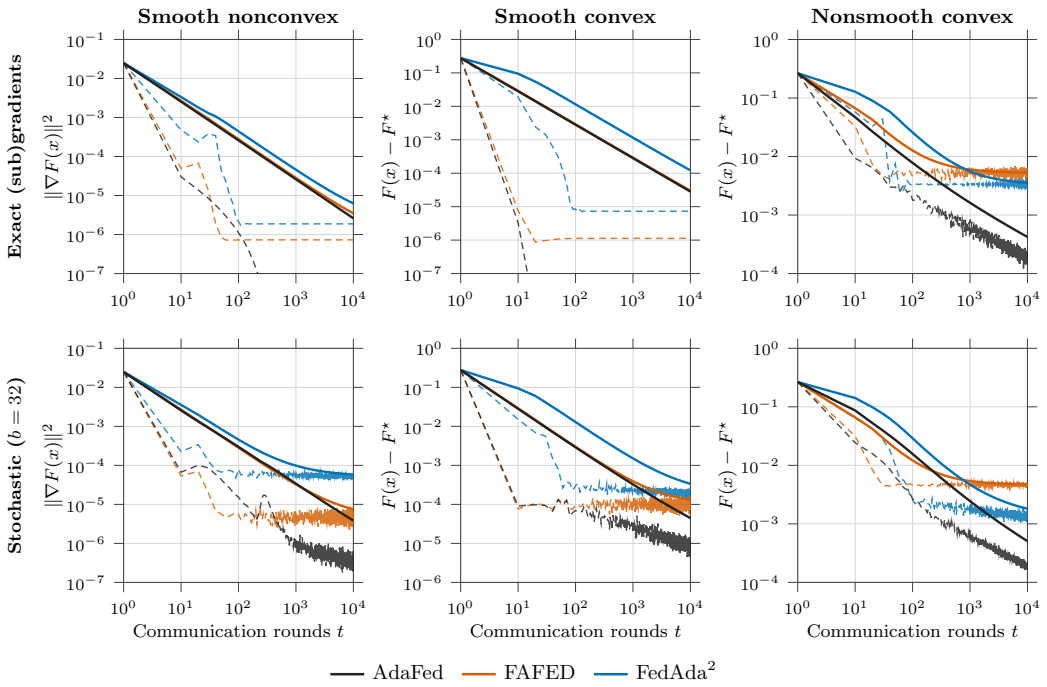

Figure 17: Performance of AdaFed compared with FL methods using both server- and client-level adaptation on IJCNN1 with $K = 10$ local steps. FAFED and FedAda^2 represent both-level adaptation. Solid and dashed curves show the uniform-random and last-iterate metrics, respectively. Stochastic results are shown as the mean over three seeds.

H Comparison of Adaptive FL Assumptions and Guarantees

Table 3 complements the related-work discussion with a detailed comparison of adaptive federated optimization methods. For each result, we summarize the objective and stochastic assumptions, the participation and local-update settings, where step-size adaptation is performed, and the main theoretical guarantees.

Table 3: Detailed comparison of assumptions, adaptive mechanisms, and theoretical guarantees in related federated optimization methods.

Paper / result	Objective assumptions	Stochastic / heterogeneity	Method setting	Main result / notes
(Wang et al., 2025) Thm. 1	**Class:** nonconvex **Smoothness:** L-smooth in $\|\cdot\|_\alpha$ **Strong conv.:** ✗ **Grad. bound:** $\|\nabla f_n(w)\|_\infty \le C$	**Variance/noise:** $\mathbb{E}\|\xi_t\|_\alpha^\alpha \le G$ **Heterogeneity:** ✗	**Participation:** full participation **Adaptation: In server** **Local updates:** one local step	Thm. 1: AdaGrad-OTA; rate $O(\log(T)/T^{1-1/\alpha})$
(Wang et al., 2025) Thm. 2	**Class:** nonconvex **Smoothness:** L-smooth in $\|\cdot\|_\alpha$ **Strong conv.:** ✗ **Grad. bound:** $\|\nabla f_n(w)\|_\infty \le C$	**Variance/noise:** $\mathbb{E}\|\xi_t\|_\alpha^\alpha \le G$ **Heterogeneity:** ✗	**Participation:** full participation **Adaptation: In server** **Local updates:** one local step	Thm. 2: Adam-OTA; rate $O(1/T)$
(Chen et al., 2025) Thm. 5.2	**Class:** nonconvex **Smoothness:** each f_i is L-smooth **Strong conv.:** ✗ **Grad. bound:** ✗	**Variance/noise:** $\mathbb{E}\|n_i^{(t,k)}\|^2 \le \sigma^2$ **Heterogeneity:** ✗	**Participation:** client sampling S^t; minibatch **Adaptation: In client** **Local updates:** K local steps	Thm. 5.2: FAdamGC, $\beta_2 = \epsilon = 0$; rate $O\Big(\sqrt{\frac{LF\sigma^2}{nKT}} + \frac{LF}{T} + \frac{LK}{T} + \frac{K\sigma}{T}\Big)$
(Chen et al., 2025) Thm. 5.4	**Class:** nonconvex **Smoothness:** each f_i is L-smooth **Strong conv.:** ✗ **Grad. bound:** $\|g_i^{(t)}\| \le G$	**Variance/noise:** $\mathbb{E}\|n_i^{(t,k)}\|^2 \le \sigma^2$ **Heterogeneity:** ✗	**Participation:** client sampling S^t; minibatch **Adaptation: In client** **Local updates:** K local steps	Thm. 5.4: FAdamGC, general Adam case; rate $O\Big(\sqrt{\frac{LF\sigma^2}{nKT}} + \frac{LF}{T} + \frac{KG^6}{\epsilon^2 T} + \frac{K(\sigma^2+(1+\epsilon^2)G^2)}{\epsilon^2 T}\Big)$
(Chen et al., 2025) Thm. C.2	**Class:** nonconvex **Smoothness:** each f_i is L-smooth **Strong conv.:** ✗ **Grad. bound:** $\|g_i^{(t)}\| \le G$	**Variance/noise:** $\mathbb{E}\|n_i^{(t,k)}\|^2 \le \sigma^2$ **Heterogeneity:** ✗	**Participation:** client sampling S^t; minibatch **Adaptation: In client** **Local updates:** K local steps	Thm. C.2: FA-NT, general Adam case; rate has worse K^2 final term
(Chen et al., 2025) Thm. C.3	**Class:** nonconvex **Smoothness:** each f_i is L-smooth **Strong conv.:** ✗ **Grad. bound:** ✗	**Variance/noise:** $\mathbb{E}\|n_i^{(t,k)}\|^2 \le \sigma^2$ **Heterogeneity:** $\|\nabla f_i(x) - \nabla f(x)\| \le B$	**Participation:** client sampling S^t; minibatch **Adaptation: In client** **Local updates:** K local steps	Thm. C.3: FA-NT, $\beta_2 = \epsilon = 0$; requires bounded data heterogeneity
(Chezhegov et al., 2024) Thm. 1 / Cor. 2	**Class:** strongly convex **Smoothness:** each f_m is L-smooth **Strong conv.:** μ-strongly convex, $\mu > 0$ **Grad. bound:** ✗	**Variance/noise:** unbiased stochastic gradients; variance $\le \sigma^2$ **Heterogeneity:** ✗	**Participation:** full participation **Adaptation: Server-like** All clients use the same $\widehat{D}^{t_p}$. **Local updates:** local steps; max gap H	Thm. 1 / Cor. 2: SAVIC for identical data; Local SGD with adaptive scaling/preconditioning
(Chezhegov et al., 2024) Thm. 2 / Cor. 3	**Class:** strongly convex **Smoothness:** each $f_m(\cdot, z)$ is L-smooth **Strong conv.:** μ-strongly convex **Grad. bound:** ✗	**Variance/noise:** via $\sigma_{\text{dif}}^2 = \frac{1}{M}\sum_m \mathbb{E}\|\nabla f_m(x^\star, z_m)\|^2$ **Heterogeneity:** heterogeneity measured at $x^\star$	**Participation:** full participation **Adaptation: Server-like** All clients use the same $\widehat{D}^{t_p}$. **Local updates:** local steps; max gap H	Thm. 2 / Cor. 3: SAVIC for heterogeneous data; Local SGD with adaptive scaling/preconditioning
(Wang et al., 2021) Thm. 1	**Class:** strongly convex **Smoothness:** each F_i is L_i-smooth **Strong conv.:** each F_i is μ_i-strongly convex **Grad. bound:** ✗	**Variance/noise:** bounded cumulative variance $q_i(k)\sigma^2$ **Heterogeneity:** ✗	**Participation:** full participation **Adaptation: In client** **Local updates:** τ_i local steps	Thm. 1: FEDOPT with local adaptive CLIENTOPT; convergence to fixed point $\tilde{x}$, not necessarily global minimizer.

Continued on next page

Table 3 – continued from previous page

Paper / result	Objective assumptions	Stochastic / heterogeneity	Method setting	Main result / notes
(Wang et al., 2021) Thm. 2	**Class:** quadratic convex **Smoothness:** quadratic smooth **Strong conv.:** $H_i \succ 0$ **Grad. bound:** ✗	**Variance/noise:** ✗ deterministic **Heterogeneity:** ✗	**Participation:** full participation **Adaptation: In client** **Local updates:** τ_i local steps	Thm. 2: preconditioned local GD; non-vanishing bias if η_i, τ_i, P_i differ across clients. The fixed point $\tilde{x}$ of the algo is not necessarily the global minimizer.
(Wang et al., 2021) Thm. 3	**Class:** nonconvex **Smoothness:** each F_i is L-smooth **Strong conv.:** ✗ **Grad. bound:** preconditioned gradients bounded: $\|B_i^{(t,k)} \nabla F_i(x_i^{(t,k)})\| \le G$	**Variance/noise:** ✗ deterministic **Heterogeneity:** ✗	**Participation:** full participation **Adaptation: In client** **Local updates:** same τ local steps	Thm. 3: local correction; removes non-vanishing bias; rate $O\Big(\frac{1}{T} + \frac{\Delta^{1/3} G^{2/3}}{T^{2/3}}\Big)$
(Kim et al., 2024) Thm. 1	**Class:** nonconvex **Smoothness:** local smoothness $\tilde{L}$ **Strong conv.:** ✗ **Grad. bound:** $\|\nabla f_i(x)\| \le G$	**Variance/noise:** $\mathbb{E}\|\nabla F_i(x,z) - \nabla f_i(x)\|^2 \le \sigma^2$ **Heterogeneity:** strong-growth dissimilarity	**Participation:** partial participation **Adaptation: In client** **Local updates:** K local steps	Thm. 1: Δ-SGD; client-specific locality-adaptive step sizes; rate $O(T^{-1/2})$.
(Kim et al., 2024) Thm. 5	**Class:** convex **Smoothness:** locally Lipschitz gradients **Strong conv.:** ✗ **Grad. bound:** ✗	**Variance/noise:** ✗ deterministic **Heterogeneity:** common minimizer	**Participation:** full participation **Adaptation: In client** **Local updates:** $K = 1$	Thm. 5: convex full-batch Δ-SGD; Lyapunov convergence.
(Lee et al., 2026) Thm. 4.1 / Cor. 4.3	**Class:** nonconvex **Smoothness:** L-smooth local objectives **Strong conv.:** ✗ **Grad. bound:** $\|[\nabla F_i(x)]_j\| \le G$	**Variance/noise:** ✗ deterministic **Heterogeneity:** ✗	**Participation:** partial participation **Adaptation: In server and client** **Local updates:** K local steps	Thm. 4.1 / Cor. 4.3: FedAda^2 and FedAda^{2++}; efficient joint adaptivity without transmitting preconditioners; FedAda^{2++} uses memory-efficient client optimizer; rate $O(T^{-1/2})$.
(Lee et al., 2026) Thm. E.1	**Class:** nonconvex **Smoothness:** L-smooth local objectives **Strong conv.:** ✗ **Grad. bound:** $\|[\nabla F_i(x)]_j\| \le G$	**Variance/noise:** ✗ deterministic **Heterogeneity:** ✗	**Participation:** partial participation **Adaptation: In server and client** **Local updates:** variable $K(O_i^\ell)$ local steps	Thm. E.1: federated blended optimization; allows heterogeneous client optimizer strategies such as SGD, AdaGrad, Adam, and SM3 variants, with bounded optimizer updates.
(Li et al., 2024) Thm. 5.6 / Cor. 5.7	**Class:** nonconvex **Smoothness:** sample mean-squared L-smoothness **Strong conv.:** ✗ **Grad. bound:** ✗	**Variance/noise:** unbiased stochastic gradients; variance $\le \sigma^2$ **Heterogeneity:** $\|\nabla f^{(k)}(x) - \nabla f^{(\ell)}(x)\|^2 \le \zeta^2$	**Participation:** full participation **Adaptation: In server/client** **Local updates:** I local dual-averaging steps	Thm. 5.6 / Cor. 5.7: FedDA-MVR for **federated constrained optimization**; adaptive matrix fixed during local steps and updated at server; sample complexity $\widetilde{O}(K^{-1}\epsilon^{-1.5})$, communication complexity $\widetilde{O}(K^{-0.25}\epsilon^{-1.25})$. $\frac{1}{T}\sum_1^T \mathbb{E}[\mathcal{G}_t] = \tilde{\mathcal{O}}\left(T^{-2/3}K^{-2/3}\right)$
(Liu et al., 2026) Thm. 3.2	**Class:** nonconvex **Smoothness:** L-smooth local objectives **Strong conv.:** ✗ **Grad. bound:** $\|g_i^t\| \le B$, $\|g_i^t\|_\infty \le B_\infty$	**Variance/noise:** unbiased stochastic gradients; bounded gradients **Heterogeneity:** ✗	**Participation:** all agents; decentralized neighbor communication **Adaptation: In local agents** **Local updates:** K local steps per communication	Thm. 3.2: LoDAdaC; decentralized adaptive local training with compressed communication. Covers SGD, momentum SGD, Adam, AMSGrad, AdaGrad, and Adam-Mini; rate $O(1/\sqrt{nTK})$ under suitable conditions, with communication cost $O((1-\eta)/(nK\epsilon^4))$ for an ϵ-stationary point.
(Mukherjee et al., 2023) Thm. 3(a)	**Class:** convex **Smoothness:** $F_i(\cdot,\xi)$ is L-smooth **Strong conv.:** ✗ **Grad. bound:** ✗	**Variance/noise:** stochastic gradients **Heterogeneity:** finite $\sigma_f^2 = f^\star - \frac{1}{n}\sum_i \ell_i^\star$	**Participation:** full participation **Adaptation: In client** **Local updates:** τ local steps	Thm. 3(a): FedSPS with fully local stochastic Polyak step-sizes; sublinear convergence to a neighborhood for convex objectives.

Continued on next page

Table 3 – continued from previous page

Paper / result	Objective assumptions	Stochastic / heterogeneity	Method setting	Main result / notes
(Mukherjee et al., 2023) Thm. 3(b) / Cor. 5	**Class:** strongly convex **Smoothness:** $F_i(\cdot,\xi)$ is L-smooth **Strong conv.:** μ-strongly convex **Grad. bound:** ✗	**Variance/noise:** stochastic gradients; finite σ_f^2 **Heterogeneity:** $\sigma_f^2 = f^\star - \frac{1}{n}\sum_i \ell_i^\star$	**Participation:** full participation **Adaptation: In client** **Local updates:** τ local steps	Thm. 3(b) / Cor. 5: FedSPS; linear convergence under interpolation, i.e., $\sigma_f^2 = 0$.
(Mukherjee et al., 2023) Thm. 18	**Class:** ✗ nonconvex **Smoothness:** $F_i(\cdot,\xi)$ is L-smooth **Strong conv.:** ✗ **Grad. bound:** ✗	**Variance/noise:** $\mathbb{E}\|\nabla F_i(x,\xi) - \nabla f_i(x)\|^2 \le \sigma^2$ **Heterogeneity:** $\frac{1}{n}\sum_i \|\nabla f_i(x) - \nabla f(x)\|^2 \le \zeta^2$	**Participation:** full participation **Adaptation: In client** **Local updates:** τ local steps	Thm. 18: small-stepsize FedSPS/FedAvg for nonconvex objectives; rate for $\min_t \mathbb{E}\|\nabla f(\bar{x}_t)\|^2$.
(Reddi et al., 2021) Thm. 1 / Cor. 1	**Class:** nonconvex **Smoothness:** each F_i is L-smooth **Strong conv.:** ✗ **Grad. bound:** $\|[\nabla f_i(x,z)]_j\| \le G$	**Variance/noise:** local variance $\sigma_{l,j}^2$ **Heterogeneity:** $\frac{1}{m}\sum_i \|[\nabla F_i(x)]_j - [\nabla f(x)]_j\|^2 \le \sigma_{g,j}^2$	**Participation:** full participation **Adaptation: In server** **Local updates:** K local SGD steps	Thm. 1 / Cor. 1: FedAdagrad with SGD as client optimizer and adaptive server optimizer; rate essentially $O(1/\sqrt{mKT})$ for suitable η_l, η, τ.
(Reddi et al., 2021) Thm. 2 / Cor. 2	**Class:** nonconvex **Smoothness:** each F_i is L-smooth **Strong conv.:** ✗ **Grad. bound:** $\|[\nabla f_i(x,z)]_j\| \le G$	**Variance/noise:** local variance $\sigma_{l,j}^2$ **Heterogeneity:** $\frac{1}{m}\sum_i \|[\nabla F_i(x)]_j - [\nabla f(x)]_j\|^2 \le \sigma_{g,j}^2$	**Participation:** full participation **Adaptation: In server** **Local updates:** K local SGD steps	Thm. 2 / Cor. 2: FedAdam; FedYogi follows similarly. Server uses Adam/Yogi-style adaptive preconditioning on the pseudo-gradient from client model differences.
(Sun et al., 2023) Thm. 3.5 / Cor. 3.6	**Class:** nonconvex **Smoothness:** F_i is L-smooth **Strong conv.:** ✗ **Grad. bound:** $\|g_i^t\|_\infty \le G_g$	**Variance/noise:** unbiased stochastic gradients; variance $\le \sigma_l^2$ **Heterogeneity:** $\|\nabla F_i(x) - \nabla F(x)\|^2 \le \sigma_g^2$	**Participation:** partial participation; S clients **Adaptation: In client** **Local updates:** K local adaptive steps	Thm. 3.5 / Cor. 3.6: FedLADA; local adaptive amended optimizer with global offset correction g_a^t; rate $O\Big(\frac{1}{\sqrt{SKT}} + \frac{1}{KT} + \frac{1}{T} + \frac{1}{T^{3/2}}\Big)$.
(Takakura et al., 2025) Thm. 4.3 / Cor. 4.4	**Class:** convex **Smoothness:** F_i is L-smooth **Strong conv.:** ✗ **Grad. bound:** ✗	**Variance/noise:** ✗ deterministic **Heterogeneity:** $\frac{1}{M}\sum_i \|\nabla F_i(w^\star)\|^2 \le \sigma_\star^2$	**Participation:** full participation **Adaptation: In server** **Local updates:** τ local SGD steps	Thm. 4.3 / Cor. 4.4: FedDuA / FedDuAdagrad; server-side doubly adaptive global step-size based on both inter-client and coordinate-wise heterogeneity; convergence for convex objectives with full-batch local gradient steps.
(Wu et al., 2023) Thm. 2 / Rem. 1	**Class:** nonconvex **Smoothness:** $f_i(\cdot;\xi)$ is L-smooth **Strong conv.:** ✗ **Grad. bound:** $\|\nabla f_i(x)\| \le G$	**Variance/noise:** unbiased stochastic gradients; variance $\le \sigma^2$ **Heterogeneity:** $\|\nabla f_i(x) - \nabla f_j(x)\|^2 \le \zeta^2$	**Participation:** full participation; cross-silo **Adaptation: In client/server** **Local updates:** periodic local adaptive steps; sync every q steps	Thm. 2 / Rem. 1: FAFED with shared adaptive learning rates and momentum-based variance reduction; adaptive matrix $A_t = \mathrm{diag}(\sqrt{v_t} + \rho)$; sample complexity $\widetilde{O}(\epsilon^{-3})$ and communication complexity $\widetilde{O}(\epsilon^{-2})$.
(Xie et al., 2020) Thm. 1	**Class:** nonconvex **Smoothness:** F and F_i are L-smooth **Strong conv.:** ✗ **Grad. bound:** $\|G_{i,t}\|_\infty \le \rho$	**Variance/noise:** ✗ no explicit variance bound **Heterogeneity:** ✗	**Participation:** full participation **Adaptation: In client** **Local updates:** H local adaptive steps	Thm. 1: Local AdaAlter; AdaGrad-style local SGD with lazy denominator updates and synchronization every H iterations; convergence rate $O(1/(\eta\sqrt{T})) + O(\eta^2 H^2 \log(T)/\sqrt{T}) + O(\eta \log(T)/(n\sqrt{T}))$.
(Wang et al., 2022) Thm. 4.4 / 4.9	**Class:** nonconvex **Smoothness:** F_i is L-smooth **Strong conv.:** ✗ **Grad. bound:** $\|\nabla f_i(x,\xi)\| \le G$	**Variance/noise:** local variance $\le \sigma_l^2$ **Heterogeneity:** $\frac{1}{m}\sum_{i=1}^m \|\nabla F_i(x) - \nabla f(x)\|^2 \le \sigma_g^2$	**Participation:** full/partial participation **Adaptation: In server** **Local updates:** K local SGD steps	Thm. 4.4 / 4.9: FedAMS; AMSGrad-type server adaptive optimizer with max stabilization; rate $O(1/\sqrt{TKm})$ under full participation and partial-participation extension.

Continued on next page

Table 3 – continued from previous page

Paper / result	Objective assumptions	Stochastic / heterogeneity	Method setting	Main result / notes
(Wang et al., 2022) Thm. 4.17 / Thm. B.1	**Class:** nonconvex **Smoothness:** F_i is L-smooth **Strong conv.:** ✗ **Grad. bound:** $\|\nabla f_i(x,\xi)\| \le G$	**Variance/noise:** local variance $\le \sigma_l^2$ **Heterogeneity:** $\frac{1}{m}\sum_{i=1}^m \|\nabla F_i(x) - \nabla f(x)\|^2 \le \sigma_g^2$	**Participation:** full/partial participation **Adaptation: In server** **Local updates:** K local SGD steps	Thm. 4.17 / Thm. B.1: FedCAMS; communication-compressed FedAMS with error feedback and biased compressor $\mathbb{E}\|\mathcal{C}(x) - x\| \le q\|x\|$; same full-participation rate $O(1/\sqrt{TKm})$ as FedAMS.
(Wang et al., 2024) Thm. 5.6 / Cor. 5.7	**Class:** nonconvex **Smoothness:** F_i is L-smooth **Strong conv.:** ✗ **Grad. bound:** $\|\nabla F_i(x;\xi)\| \le G$	**Variance/noise:** unbiased stochastic gradients; variance $\le \sigma^2$ **Heterogeneity:** $\frac{1}{N}\sum_{i=1}^N \|\nabla F_i(x) - \nabla f(x)\|^2 \le \sigma_g^2$	**Participation:** asynchronous; buffer size M; uniform arrivals **Adaptation: In server** **Local updates:** K local SGD steps	Thm. 5.6 / Cor. 5.7: FADAS; asynchronous adaptive FL with AMSGrad-type server update; rate $O\left(\frac{1}{\sqrt{TM}} + \frac{\tau_{\max}\tau_{\text{avg}}}{T}\right)$.
(Wang et al., 2024) Thm. 5.9 / Cor. 5.10	**Class:** nonconvex **Smoothness:** F_i is L-smooth **Strong conv.:** ✗ **Grad. bound:** $\|\nabla F_i(x;\xi)\| \le G$	**Variance/noise:** unbiased stochastic gradients; variance $\le \sigma^2$ **Heterogeneity:** $\frac{1}{N}\sum_{i=1}^N \|\nabla F_i(x) - \nabla f(x)\|^2 \le \sigma_g^2$	**Participation:** asynchronous; buffer size M; uniform arrivals **Adaptation: In server; delay-adaptive** **Local updates:** K local SGD steps	Thm. 5.9 / Cor. 5.10: delay-adaptive FADAS; server learning rate adjusted by maximum delay $\tau_t^{\max}$; reduces dependence on worst-case delay $\tau_{\max}$.
(Yuan & Wang, 2025) Thm. 1–3	**Class:** nonconvex **Smoothness:** β-Lipschitz smooth **Strong conv.:** ✗ **Grad. bound:** $\|\nabla F_i\|_2 \le P$	**Variance/noise:** ✗ deterministic **Heterogeneity:** ✗	**Participation:** full participation **Adaptation: In client** **Local updates:** ✗ one local update in theory	Thm. 1–3: FedAgile; aggregated-gradient adaptive learning rate with mean-field terms and JS-distance refinement; convergence rate $\widetilde{O}(T^{-1})$.
(Zhang et al., 2023) Thm. IV.1–IV.3	**Class:** convex **Smoothness:** β-smooth **Strong conv.:** ✗ **Grad. bound:** F is ρ-Lipschitz	**Variance/noise:** ✗ no explicit variance bound **Heterogeneity:** $\|G_k(w) - G(w)\| \le \delta_k$	**Participation:** partial participation; K selected devices **Adaptation: In server** **Local updates:** τ local iterations	Thm. IV.1–IV.3: FedUR; adaptive centralized-learning optimizer for FL with client-side momentum and piecewise adaptive learning rate in global aggregation; analyzes communication-frequency and convergence-rate bounds.