TIGHTER PERFORMANCE THEORY OF FEDEXPROX

Wojciech Anyszka[*1], **Kaja Gruntkowska**[2], **Alexander Tyurin**[3,4] **& Peter Richtárik**[2]
[1] Princeton University, Princeton, United States
[2] King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
[3] AXXX, Moscow, Russia
[4] Applied AI Institute, Moscow, Russia

ABSTRACT

We revisit FedExProx–a recently proposed distributed optimization method designed to enhance convergence properties of parallel proximal algorithms via extrapolation. In the process, we uncover a surprising flaw: its known theoretical guarantees on quadratic optimization tasks are no better than those offered by the vanilla Gradient Descent (GD) method. Motivated by this observation, we develop a novel analysis framework, establishing a tighter linear convergence rate for non-strongly convex quadratic problems. By incorporating both computation and communication costs, we demonstrate that FedExProx can indeed provably outperform GD, in stark contrast to the original analysis. Furthermore, we consider partial participation scenarios and analyze two adaptive extrapolation strategies–based on gradient diversity and Polyak stepsizes—again significantly outperforming previous results. Moving beyond quadratics, we extend the applicability of our analysis to general functions satisfying the Polyak-Łojasiewicz condition, outperforming the previous strongly convex analysis while operating under weaker assumptions. Backed by empirical results, our findings point to a new and stronger potential of FedExProx, paving the way for further exploration of the benefits of extrapolation in federated learning.

1 INTRODUCTION

Federated Learning (FL) (Konečný et al., 2016; McMahan et al., 2017) is a distributed machine learning paradigm where multiple edge devices collaboratively train a global model without the need for centralized data collection. Instead of transmitting raw data, each client computes updates based on their local datasets, and these are then aggregated at a central server to update the global model. This approach ensures that no sensitive information is shared with the server or other clients, making it particularly useful in privacy-sensitive domains, such as healthcare (Rieke et al., 2020) and recommendation systems (Hard et al., 2018). Formally, FL solves the optimization problem

$$\min_{x\in\mathbb{R}^d}\left\{f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)\right\}, \tag{1}$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$ represents the empirical risk associated with the local dataset stored on client $i \in [n]$. Despite its advantages, FL introduces several nontrivial challenges, including *communication delays* caused by limited network bandwidth, and *partial participation* of clients due to random outages (Kairouz et al., 2021). Addressing these issues is essential to making FL efficient and reliable in real-world applications.

1.1 RELATED WORK

Let us briefly review several existing methods used to solve problem (1) in FL scenarios. For a detailed overview of the notation used, please refer to Table 1.

Gradient descent. The simplest and most naive approach is the Gradient Descent (GD) method (Nesterov, 2018), iterating $x_{k+1} = x_k - \gamma\nabla f(x_k) = x_k - \gamma\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_k)$, where x_0 is a

[*]The work of Wojciech Anyszka was conducted during a VSRP internship at KAUST.

starting point and $\gamma > 0$ is the stepsize. Let us introduce the key assumptions typically employed in the analysis of GD in the convex world.

Assumption 1.1. The functions f_i are proper, closed and convex for all $i \in [n]$, and the function f attains a minimum at some (potentially non-unique) point $x_* \in \mathbb{R}^d$.

Assumption 1.2. The function f is L–smooth, i.e., $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| \; \forall x, y \in \mathbb{R}^d$. Additionally, each f_i, $i \in [n]$, is differentaible and L_i–smooth, with $L_{\max} := \max_{i \in [n]} L_i$.

Under these assumptions, classical optimization theory guarantees that GD solves problem (1) and finds $\bar{x} \in \mathbb{R}^d$ such that $f(\bar{x}) - f(x_*) \leq \varepsilon$ in $\mathcal{O}(LR^2/\varepsilon)$ iterations, where $R^2 := \|x_* - x_0\|^2$. In an FL setting, finding a solution to the main problem requires each client to participate in every communication round, and thus to communicate with the server

$$\mathcal{O}\left(\frac{LR^2}{\varepsilon}\right) \tag{2}$$

times, transmitting the local gradient $\nabla f_i(x_k)$ at each iteration k. While this approach is conceptually simple, its direct application in distributed environments leads to several challenges related to communication overhead and scalability.

Partial participation. In real-world FL scenarios, not all clients remain active for the entire duration of training. Instead, a subset of clients is chosen at each iteration, typically based on practical factors such as device availability (e.g., battery life or network conditions) or statistical considerations (e.g., data heterogeneity) (Kairouz et al., 2021; Tyurin & Richtárik, 2023).

Communication bottleneck. Another key problem in distributed training is the *communication bottleneck* (Ramesh et al., 2021; Kairouz et al., 2021). Since the overall performance of a distributed algorithm is the product of the number of communication rounds K needed to find a solution and the cost C of one such round, there exist two main strategies to addressing this issue: (i) minimizing C and/or (ii) minimizing K. Objective (i) is typically achieved via compression of the information transmitted between the clients and the server (Beznosikov et al., 2020; Gruntkowska et al., 2024). Objective (ii) can be accomplished by increasing the amount of local computation performed by clients between rounds, allowing for less frequent communication.

FedAvg and modern local methods. A common strategy for ensuring communication efficiency that falls into category (ii) is *local training*. The idea is simple: allow clients to do more work before transmitting the results to the server. If executed with care, this approach leads to each communication round providing more "informative" updates, ultimately resulting in more effective changes to the global model. One of the most popular methods in this class is Federated Averaging (FedAvg), introduced by McMahan et al. (2017). Within this framework, a subset of clients is selected in each round to perform local training using a gradient-based method of their choice (e.g., GD). This is followed by an averaging step on the server, where the local updates are aggregated. Despite its popularity, the theoretical properties of FedAvg are somewhat limited (Khaled et al., 2019; Koloskova et al., 2020). A major issue is *client drift*, which can lead to slower convergence (Karimireddy et al., 2020). This problem can be mitigated using modern drift correction techniques (Karimireddy et al., 2020; Gorbunov et al., 2021; Mishchenko et al., 2022). One notable advancement in this area is FedExP, introduced by Jhunjhunwala et al. (2023), which extends FedAvg by leveraging extrapolation to better control server updates.

FedProx and proximal methods. All methods discussed so far assume that clients have access to the gradients $\{\nabla f_i\}$ of local functions. The class of *proximal methods* (Bertsekas, 2011; Ryu & Boyd, 2014; Khaled & Jin, 2022; Richtárik et al., 2024) relies on a different, more powerful oracle–the *proximal operator* $\mathrm{prox}_{\gamma f} : \mathbb{R}^d \to \mathbb{R}^d$, defined by

$$\mathrm{prox}_{\gamma f}(x) = \arg\min_{z \in \mathbb{R}^d} \left\{ f(z) + \tfrac{1}{2\gamma} \|z - x\|^2 \right\}. \tag{3}$$

Here, f is a convex function and $\gamma > 0$ plays a rule of a regularization parameter. It is important to note that computing $\mathrm{prox}_{\gamma f}(\cdot)$ is an optimization problem on its own. However, the analyses of such methods typically do not consider how this subproblem should be solved, and instead assume that computing $\mathrm{prox}_{\gamma f}(\cdot)$ is computationally cheap and can be performed using any iterative solver, e.g., GD, Newton's method or LBFGS (Nesterov, 2018; Liu & Nocedal, 1989).

There exist numerous proximal methods addressing the single-node case. A classic approach is the Proximal Point Method (PPM) by Rockafellar (1976), which iterates $x_{k+1} = \mathrm{prox}_{\gamma_k f}(x_k)$.

Stochastic versions of PPM have also been studied (Ryu & Boyd, 2014; Bianchi, 2016; Patrascu & Necoara, 2018). Recently, interest in stochastic proximal methods has surged, with promising results from several works (Condat & Richtárik, 2023; Traoré et al., 2024; Sadiev et al., 2024; Combettes & Madariaga, 2024). In the context of distributed learning, one relevant work by Li et al. (2020) builds on the principles of FedAvg to introduce its proximal variant–FedProx. The method replaces local gradient-based updates with proximal operator computations, leading to the update rule

$$x_{k+1} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{prox}_{\gamma f_i}(x_k). \tag{4}$$

By leveraging proximal operators, FedProx addresses some of the limitations of traditional GD in FL. However, simply averaging local updates at the server remains suboptimal and can be improved.

FedExProx. A recent advancement, and the main focus of this paper, is FedExProx, introduced by Li et al. (2024). This method improves on FedProx by employing the technique of *extrapolation*[1] (Combettes, 1997; Necoara et al., 2019). The algorithm operates under the *interpolation regime*:

Assumption 1.3. There exists $x_* \in \mathbb{R}^d$ such that $\nabla f_i(x_*) = 0$ for all $i \in [n]$. We denote the set of such minimizers by $\mathcal{X}_* := \{x \in \mathbb{R}^d : \nabla f_i(x) = 0\}$ and the projection of x onto $\mathcal{X}_*$ by $\Pi(x)$.

This assumption is relatively mild and is often met in practice, e.g., for over-parameterized models (Ma et al., 2018). Instead of (4), FedExProx introduces the extrapolated update

$$x_{k+1} = x_k + \alpha_k \left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{prox}_{\gamma f_i}(x_k) - x_k \right), \tag{5}$$

where α_k serves as the *extrapolation parameter*. The algorithm is analyzed by reformulating step (5) in terms of *Moreau envelopes* (Moreau, 1965): the update rule can be expressed equivalently as

$$x_{k+1} = x_k - \alpha_k \gamma \frac{1}{n}\sum_{i=1}^{n} \nabla M_{f_i}^{\gamma}(x_k), \tag{6}$$

where the Moreau envelope $M_f^{\gamma} : \mathbb{R}^d \to \mathbb{R}$ of a convex function f is defined as

$$M_f^{\gamma}(x) := \min_{z \in \mathbb{R}^d} \left\{ f(z) + \frac{1}{2\gamma}\|z - x\|^2 \right\} \quad \forall x \in \mathbb{R}^d. \tag{7}$$

Notably, when $\alpha_k = 1$, (5) reduces to (4). However, this choice is not optimal, and choosing $\alpha_k > 1$ yields better complexity. If α_k is set to be constant across iterations of the algorithm, then using the optimal value $\alpha_k \equiv \alpha = \frac{1}{\gamma L_\gamma} > 1$ guarantees that FedExProx converges after

$$\mathcal{O}\left(\frac{L_\gamma(1+\gamma L_{\max})R^2}{\varepsilon} \right) \tag{8}$$

communication rounds in the convex case, where L_γ is the smoothness constant of the function $M^{\gamma}(x) := \frac{1}{n}\sum_{i=1}^{n} M_{f_i}^{\gamma}(x)$. Li et al. (2024) show that (8) is not larger than $\mathcal{O}\left(L_{\max}R^2/\varepsilon\right)$, and can be significantly smaller in practical scenarios. Thus, FedExProx offers provably better iteration complexity compared to both FedProx and FedExP (Li et al., 2020; Jhunjhunwala et al., 2023).

2 Contributions

1. Although the iteration complexity (8) of FedExProx outperforms both FedProx and FedExP, we find that it is *not superior* to that of the simplest baseline, GD, on quadratic optimization tasks (see Remark 3.3). This raises a critical question regarding the utility of employing the proximal oracle in the first place.

2. Motivated by these pessimistic findings, we revisit the analysis of FedProx and prove a significantly more optimistic iteration complexity result. We establish a *linear* convergence rate for solving non-strongly convex quadratic distributed optimization problems of the form (1) (see Theorem 4.1) and demonstrate that our new result does indeed lead to the conclusion that FedExProx can significantly outperform GD. Our analysis assumes a realistic model that accounts for both computation

[1] Notably, the idea of extrapolation was explored by Jhunjhunwala et al. (2023) in developing FedExP.

and communication times–critical factors in real-world distributed optimization. We show that the total *time complexity* of FedExProx is never worse than that of GD, and can be strictly better when communication time dominates computation time (see Theorem 4.3), which is typically the case in FL scenarios. This stands in stark contrast to the previous analysis by Li et al. (2024), underscoring the superiority of our new findings.

3. To account for the stochastic setting that involves partial participation of clients, we complement the above results with client sampling (see Theorems 5.1, 5.2, 7.5 and F.1).

4. Beyond constant extrapolation, we establish two novel results (Theorems 6.1 and F.1) that incorporate smoothness-adaptive strategies, based on gradient diversity (FedExProx-GraDS), and Polyak stepsize (FedExProx-StoPS), again significantly improving upon the result previously established by Li et al. (2024) in the quadratic case.

5. We extend the analysis beyond quadratics to *arbitrary convex functions* satisfying the Polyak-Łojasiewicz (PŁ) condition, obtaining a linear convergence rate (see Theorems 7.2 and 7.5). In contrast, Li et al. (2024) derived a linear rate under strong convexity for the function f. Our approach is not only more general, as it relies on weaker assumptions, but also demonstrates improved dependence on problem-specific constants. Additionally, we establish a result in the PŁ setting that accounts for inexact computations of the proximal mappings (Theorem E.1).

6. The theoretical findings are validated with empirical experiments, which demonstrate the robustness and applicability of our framework.

3 Not Better than GD on Quadratics

Let us take a closer look at the complexity result (8). Suppose for now that we solve (3) using GD. Then, the number of iterations needed to find $\mathrm{prox}_{\gamma f_i}(x)$ with accuracy ε is

$$\mathcal{O}\left(\frac{L_i+1/\gamma}{1/\gamma}\log\frac{1}{\varepsilon}\right)=\mathcal{O}\left((\gamma L_i+1)\log\frac{1}{\varepsilon}\right),$$

since the function $f_i(z)+\frac{1}{2\gamma}\|z-x\|^2$ is (L_i+1/γ)–smooth and $1/\gamma$–strongly convex. Hence, γ controls the difficulty of calculating the proximal operator: the larger it is, the more difficult the problem. Therefore, to accelerate local computations, one would prefer to choose γ as small as possible. We can formalize this intuition by letting $\pi(\gamma)$ be the time per one iteration of FedExProx and making the following reasonable assumption:

Assumption 3.1. The time complexity $\pi(\gamma)$ of a FedExProx step is a non-decreasing function of γ.

Combining it with the iteration complexity result (8) established by Li et al. (2024), the total time required by FedExProx with $\alpha_k\equiv 1/\gamma L_\gamma$ to find $\bar{x}$ such that $\mathbb{E}[f(\bar{x})]-f(x_*)\le\varepsilon$ is

$$T(\gamma):=\pi(\gamma)\times\frac{L_\gamma(1+\gamma L_{\max})R^2}{\varepsilon}.$$

With this definition in place, we can now translate the main result of Li et al. (2024) (Theorem 1) to quadratic optimization problems, yielding the following pessimistic outcome:

Theorem 3.2. *Let Assumptions 1.3 and 3.1 hold. Consider solving a non-strongly convex quadratic optimization problem of the form (1), where $f_i(x)=\frac{1}{2}x^\top\boldsymbol{A}_ix-b_i^\top x$ for all $i\in[n]$, with $\boldsymbol{A}_i\in\mathrm{Sym}_+^d$ and $b_i\in\mathbb{R}^d$. Then the assumptions of Theorem 1 by Li et al. (2024) hold, and*

$$T(\gamma)=\pi(\gamma)\times\frac{L_\gamma(1+\gamma L_{\max})R^2}{\varepsilon}\ge\pi(0)\times\frac{LR^2}{\varepsilon}\tag{9}$$

for all $\gamma>0$. Moreover, when $\gamma\to 0$, then $\pi(\gamma)\times\frac{L_\gamma(1+\gamma L_{\max})R^2}{\varepsilon}\to\pi(0)\times\frac{LR^2}{\varepsilon}$, and FedExProx *effectively reduces to* GD.

Remark 3.3. In light of Theorem 3.2, GD performs no worse than FedExProx by Li et al. (2024), even if the time complexity $\pi(\gamma)$ of one FedExProx step does not depend on γ. Indeed, it holds that

$$\frac{L_\gamma(1+\gamma L_{\max})R^2}{\varepsilon}\ge\frac{LR^2}{\varepsilon}$$

for all $\gamma>0$ (see the proof of Theorem 3.2).

These results, based on the original analysis of FedExProx, lead us to a rather disappointing conclusion: from a theoretical time complexity perspective, the optimal strategy appears to be using vanilla GD, disregarding proximal oracles entirely. The question that remains is: is this an inherent limitation of the method itself or merely a consequence of suboptimal analysis by Li et al. (2024)?

4 TIGHTER AND MORE OPTIMISTIC RESULTS

Our refined analysis of FedExProx demonstrates that a tighter iteration complexity bound *can* be derived, leading to significantly more optimistic results. In the following sections, we present a novel analysis that provides an improved complexity result for FedExProx and sheds light on its true capabilities. To ensure clarity and build intuition, we begin with quadratic optimization tasks, where all relevant quantities can be explicitly derived. In Section 7, we extend these advancements to *general convex functions* satisfying the PŁ condition.

Theorem 4.1. *Fix any $\gamma > 0$ and consider solving non-strongly convex quadratic optimization problem (1) where $f_i(x) = \frac{1}{2}x^\top \boldsymbol{A}_i x - b_i^\top x$ for all $i \in [n]$, with $\boldsymbol{A}_i \in \mathrm{Sym}_+^d$ and $b_i \in \mathbb{R}^d$. Under Assumption 1.3,* FedExProx *with $\alpha = 1/\gamma L_\gamma$ finds $\bar{x}$ such that $\mathbb{E}[f(\bar{x})] - f(x_*) \leq \varepsilon$ after*

$$\mathcal{O}\left(\frac{L_\gamma}{\mu_\gamma^+} \log \frac{1}{\varepsilon}\right) \tag{10}$$

iterations, where L_γ is a smoothness constant of M^γ and μ_γ^+ is the smallest non-zero eigenvalue of the matrix $\nabla^2 M^\gamma$.

Remark 4.2. To the best of our knowledge, under the assumptions of Theorem 4.1, GD requires

$$\mathcal{O}\left(\frac{L}{\mu^+} \log \frac{1}{\varepsilon}\right) \tag{11}$$

iterations to solve the quadratic optimization problem, where μ^+ is the smallest non-zero eigenvalue of the matrix $\boldsymbol{A} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{A}_i$ (Richtárik & Takáč, 2020).

Let us now demonstrate that our new result does indeed provide a tighter bound, leading to the conclusion that FedExProx can in fact outperform GD.

4.1 ONE STEP TIME COMPLEXITY $\pi(\gamma)$ IS $\eta + \tau(\gamma L_{\max} + 1)$

As introduced in Section 1, the time per one global iteration of FedExProx has two main sources:

1. **Local computation:** In large-scale problems, each step (5) requires clients to compute $\mathrm{prox}_{\gamma f_i}(x_k)$ iteratively. One of the simplest solvers is GD, which returns a solution of subproblem i after $\widetilde{\mathcal{O}}(\gamma L_i + 1)$ local iterations[2] (see Section 3). If each gradient calculation takes τ seconds, the total time required for all clients to calculate $\mathrm{prox}_{\gamma f_i}(x_k)$ is $\widetilde{\mathcal{O}}(\tau \times (\gamma L_{\max} + 1))$ since the process is gated by the "slowest" client, associated with the problem with the highest smoothness constant.

2. **Communication:** Once the local computations are completed, clients must communicate their results before the server can execute the global step (5). We assume this communication takes η seconds, which can be huge in FL environments (Kairouz et al., 2021).

Consequently, the total time per global iteration, $\pi(\gamma)$, is proportional to $\eta + \tau(\gamma L_{\max} + 1)$ (thus satisfying Assumption 3.1), and the total time complexity of FedExProx is

$$T_\eta(\gamma) := \widetilde{\mathcal{O}}\left((\eta + \tau(\gamma L_{\max} + 1)) \times \frac{L_\gamma}{\mu_\gamma^+}\right). \tag{12}$$

Note that the total time complexity of GD is

$$T_{\mathsf{GD}} := T_\eta(0) = \widetilde{\mathcal{O}}\left((\eta + \tau) \times \frac{L}{\mu^+}\right).$$

Our next goal is to determine the value of γ that minimizes $T_\eta(\gamma)$, with the hope that this time the optimal choice does not result in $\gamma \to 0$. This is indeed the case whenever $\eta > \tau$.

Theorem 4.3. *Consider the non-strongly convex quadratic optimization problem from Theorem 4.1. Up to a constant factor, the time complexity (12) is minimized by some*

$$\gamma \in \left[\frac{1}{\max\limits_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)}, \min\left\{\frac{\frac{\eta}{\tau}-1}{\max\limits_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)}, \frac{1}{\min\limits_{i\in[n]} \lambda_{\min}^+(\boldsymbol{A}_i)}\right\}\right]$$

[2] Alternatively, an accelerated method could be employed, reducing the iteration count to $\widetilde{\mathcal{O}}\left(\sqrt{\gamma L_i + 1}\right)$. Our analysis can be extended to accommodate these faster solvers as well.

if $\frac{\eta}{\tau} \geq 2$ *and by some*

$$\gamma \in \left[0, \max\left\{0, \min\left\{\frac{\frac{\eta}{\tau}-1}{\max\limits_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)}, \frac{1}{\min\limits_{i\in[n]} \lambda^{+}_{\min}(\boldsymbol{A}_i)}\right\}\right\}\right]$$

if $\frac{\eta}{\tau} < 2$. *Moreover,* $T_\eta(\gamma) \leq T_{\mathsf{GD}}$.

The theorem above establishes that the total time complexity of FedExProx is no worse than that of GD, and can be strictly better when communication time dominates computation time, highlighting the superiority of our new analysis. Notably, this was not the case for the previous analysis by Li et al. (2024) (Theorem 3.2). The following example, where $T_\eta(\gamma) \ll T_{\mathsf{GD}}$ for some $\gamma > 0$, highlights how our complexity can significantly outperform that of GD.

Example 4.4. Let the matrices $\{\boldsymbol{A}_i\}$ be diagonal, i.e., $\boldsymbol{A}_i = \operatorname{diag}(a_{i1}, \ldots, a_{id})$ for all $i \in [n]$, and $a_{ij} > 0$ for all $i \in [n], j \in [d]$. Then, according to (25) and (26), we get

$$\frac{L_\gamma}{\mu^+_\gamma} = \frac{\lambda_{\max}(\gamma)}{\lambda^+_{\min}(\gamma)} = \frac{\max\limits_{j\in[d]} \sum\limits_{i=1}^{n} \frac{\gamma a_{ij}}{1+\gamma a_{ij}}}{\min\limits_{j\in[d]} \sum\limits_{i=1}^{n} \frac{\gamma a_{ij}}{1+\gamma a_{ij}}},$$

where $\lambda_{\max}(\gamma)$ and $\lambda^+_{\min}(\gamma)$ are the largest and the smallest non-zero eigenvalues of the matrix $\boldsymbol{M} := \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\gamma}(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1})$, respectively (see Section C). Taking $\gamma = 1/\min_{i\in[n],j\in[d]} a_{ij}$ and using the fact that $1 \geq \frac{x}{1+x} \geq \frac{1}{2}$ for all $x \geq 1$, we can conclude that $\frac{L_\gamma}{\mu^+_\gamma} \leq 2$. Thus,

$$T_\eta(\gamma) = \widetilde{\mathcal{O}}\left(\eta + \tau \frac{\max\limits_{i\in[n],j\in[d]} a_{ij}}{\min\limits_{i\in[n],j\in[d]} a_{ij}}\right).$$

Suppose that communication is slow. Formally, let $\eta \geq \tau \max_{i\in[n],j\in[d]} a_{ij} / \min_{i\in[n],j\in[d]} a_{ij}$. Then $T_\eta(\gamma) = \widetilde{\mathcal{O}}(\eta)$, while the total time complexity of GD is

$$T_{\mathsf{GD}} = \widetilde{\Omega}\left(\eta \times \frac{L}{\mu^+}\right) = \widetilde{\Omega}\left(\eta \times \frac{\max\limits_{j\in[d]} \sum\limits_{i=1}^{n} a_{ij}}{\min\limits_{j\in[d]} \sum\limits_{i=1}^{n} a_{ij}}\right),$$

which is at least $\max_{j\in[d]} \sum_{i=1}^{n} a_{ij} / \min_{j\in[d]} \sum_{i=1}^{n} a_{ij}$ times worse!

A similar improvement can be observed in the general case. However, the derivation is significantly more complex, so to maintain clarity, the example focuses on diagonal matrices only.

5 Partial Participation

Thus far, we have concentrated on the full participation scenario. However, as outlined in Section 1.1, practical FL settings often involve only a subset of clients participating in each training round. To address this, we supplement our theory with a convergence result in the stochastic setting. For illustration, we consider nice sampling (see Section A.2), where at each iteration, a subset $\mathcal{S}_k \subseteq [n]$ of clients is selected uniformly at random from all subsets of size S. Although we use this sampling strategy as an example, other client selection methods can also be employed. In this context, we can formulate the following complexity result:

Theorem 5.1. *Fix any* $\gamma > 0$ *and consider solving non-strongly convex quadratic optimization problem* (1)*, where* $f_i(x) = \frac{1}{2}x^\top \boldsymbol{A}_i x - b_i^\top x$ *for all* $i \in [n]$*, with* $\boldsymbol{A}_i \in \mathrm{Sym}^d_+$ *and* $b_i \in \mathbb{R}^d$*. Under Assumption 1.3,* FedExProx *with nice sampling (Algorithm 2) with* $\alpha = 1/\gamma L_{\gamma,S}$ *finds* $\bar{x}$ *such that* $\mathbb{E}[f(\bar{x})] - f(x_*) \leq \varepsilon$ *after*

$$\mathcal{O}\left(\frac{L_{\gamma,S}}{\mu^+_\gamma} \log \frac{1}{\varepsilon}\right) \tag{13}$$

iterations, where $L_{\gamma,S} := \frac{n-S}{S(n-1)} \frac{L_{\max}}{1+\gamma L_{\max}} + \frac{n(S-1)}{S(n-1)} L_\gamma$*,* L_γ *is the smoothness constant of* M^γ *and* μ^+_γ *is the smallest non-zero eigenvalue of the matrix* $\nabla^2 M^\gamma$*.*

Assuming the same time complexity model as in Section 4.1, the total time complexity of FedExProx with nice sampling is given by

$$T_\eta(\gamma, S) := \widetilde{\mathcal{O}}\left(\left(\eta + \tau\left(\gamma L_{\max} + 1\right)\right) \times \frac{L_{\gamma,S}}{\mu_\gamma^+}\right). \tag{14}$$

As it turns out, the optimal stepsize for the stochastic setting aligns with the one used in the deterministic scenario (see Theorem 4.3).

Theorem 5.2. *Up to a constant factor, the time complexity* (14) *is minimized by*

$$\gamma \in \left[\frac{1}{\max\limits_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)}, \min\left\{\frac{\frac{\eta}{\tau}-1}{\max\limits_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)}, \frac{1}{\min\limits_{i\in[n]} \lambda_{\min}^+(\boldsymbol{A}_i)}\right\}\right]$$

if $\frac{\eta}{\tau} \geq 2$ *and by*

$$\gamma \in \left[0, \max\left\{0, \min\left\{\frac{\frac{\eta}{\tau}-1}{\max\limits_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)}, \frac{1}{\min\limits_{i\in[n]} \lambda_{\min}^+(\boldsymbol{A}_i)}\right\}\right\}\right]$$

if $\frac{\eta}{\tau} < 2$.

Consequently, the conclusions from the previous section apply equally to the partial participation scenario: FedExProx performs at least as well as GD and can be strictly better when communication time exceeds computation time.

6 Adaptivity

We now turn to adaptive extrapolation strategies: *gradient diversity* (GraDS) and a variant of the classical *Polyak stepsize* (StoPS). Both were first introduced by Horváth et al. (2022) and later adapted for proximal methods by Li et al. (2024). As in the case of constant extrapolation, we refine the analysis by Li et al. (2024) for quadratic problems. Although Theorem 6.1 refers to the full participation setting, the approach can be extended to the stochastic case (see Section F.2).

Theorem 6.1. *Fix any* $\gamma > 0$ *and consider solving non-strongly convex quadratic optimization problem* (1)*, where* $f_i(x) = \frac{1}{2}x^\top \boldsymbol{A}_i x - b_i^\top x$ *for all* $i \in [n]$*, with* $\boldsymbol{A}_i \in \mathrm{Sym}_+^d$ *and* $b_i \in \mathbb{R}^d$*. Let Assumption 1.3 hold and consider two adaptive extrapolation strategies:*

1. *(*FedExProx-GraDS*) Set*

$$\alpha_k = \alpha_k^{\mathsf{GraDS}}(x_k) := \frac{\frac{1}{n}\sum_{i=1}^n \left\|\nabla M_{f_i}^\gamma(x_k)\right\|^2}{\left\|\frac{1}{n}\sum_{i=1}^n \nabla M_{f_i}^\gamma(x_k)\right\|^2} \geq 1.$$

Then, the iterates of Algorithm 1 satisfy

$$\|x_K - \Pi(x_K)\|^2 \leq (1 - C_{\mathsf{G}})^K \|x_0 - \Pi(x_0)\|^2, \tag{15}$$

where $C_{\mathsf{G}} := \min\limits_{k=0,\ldots,K-1} \alpha_k \gamma \frac{2+\gamma L_{\max}}{1+\gamma L_{\max}} \mu_\gamma^+$.

2. *(*FedExProx-StoPS*) Set*

$$\alpha_k = \alpha_k^{\mathsf{StoPS}}(x_k) := \frac{\frac{1}{n}\sum_{i=1}^n \left(M_{f_i}^\gamma(x_k) - \inf M_{f_i}^\gamma\right)}{\gamma\left\|\frac{1}{n}\sum_{i=1}^n \nabla M_{f_i}^\gamma(x_k)\right\|^2} \geq \frac{1}{2\gamma L_\gamma}.$$

Then, the iterates of Algorithm 1 satisfy

$$\|x_K - \Pi(x_K)\|^2 \leq (1 - C_{\mathsf{S}})^K \|x_0 - \Pi(x_0)\|^2, \tag{16}$$

where $C_{\mathsf{S}} := \frac{3}{2} \min\limits_{k=0,\ldots,K-1} \alpha_k \gamma \mu_\gamma^+$.

Remark 6.2. Similar to the observations by Li et al. (2024), we see that, unlike in the constant extrapolation case (see Theorem 4.1), FedExProx with adaptive extrapolation benefits from *semi-adaptivity* to the smoothness constant. Specifically, it converges for any $\gamma > 0$, and since $\gamma\mu_\gamma^+$ is bounded above (see (26)), it suffices to choose a large enough γ to achieve the optimal performance.

Remark 6.3. Bounding $\min_{k=0,\ldots,K-1} \alpha_k$ by $\frac{1}{2\gamma L_\gamma}$, inequality (16) implies that

$$\|x_K - \Pi(x_K)\|^2 \leq \left(1 - \frac{3}{4}\frac{\mu_\gamma^+}{L_\gamma}\right)^K \|x_0 - \Pi(x_0)\|^2,$$

and hence Theorem 6.1 guarantees convergence of FedExProx-StoPS in $\mathcal{O}\left({}^{L_\gamma}/_{\mu_\gamma^+} \times \log {}^{1}/_{\varepsilon}\right)$ iterations, regardless of the choice of the stepsize γ. This matches the guarantee from Theorem 4.1, but without requiring prior knowledge of the optimal extrapolation parameter $\alpha = {}^{1}/_{\gamma L_\gamma}$. The benefit comes with the trade-off of needing to know the minimum of the average of Moreau envelopes.

7 Better Theory with PŁ Condition

The result from Section 4 can be extended to solving problem (1) for general functions that satisfy the Polyak-Łojasiewicz (PŁ) condition.

Assumption 7.1. The function M^γ satisfies PŁ-condition, i.e., there exists μ_γ^+ such that

$$\tfrac{1}{2}\|\nabla M^\gamma(x)\|^2 \geq \mu_\gamma^+ \left(M^\gamma(x) - M^\gamma(x_*)\right) \qquad \forall x \in \mathbb{R}^d.$$

This assumption holds with $\mu_\gamma^+ \geq {}^{\mu^+}/_{(4(1+\gamma L_{\max}))}$ if the function f satisfies PŁ condition with a constant μ^+ (Lemma A.11). However, the choice $\mu_\gamma^+ = {}^{\mu^+}/_{(4(1+\gamma L_{\max}))}$ is loose and leads to problems described in Theorem 3.2. With this assumption in place, we can present the theorem.

Theorem 7.2. *Let Assumptions 1.1, 1.2, 1.3, and 7.1 hold. For all $\gamma > 0$,* FedExProx *(Algorithm 1) with $\alpha = {}^{1}/_{\gamma L_\gamma}$ finds $\bar{x}$ such that $\mathbb{E}[f(\bar{x})] - f(x_*) \leq \varepsilon$ in*

$$\mathcal{O}\left(\frac{L_\gamma}{\mu_\gamma^+}\log\frac{1}{\varepsilon}\right)$$

iterations, where L_γ is a smoothness constant of M^γ and μ_γ^+ is the PŁ constant.

Remark 7.3. In practice, solving the local problems exactly is often infeasible. Instead, we can only calculate an inexact proximal operator that approximates the desired quantity within a certain accuracy budget. To accommodate this, we extend the above analysis to the scenario where the clients can only compute updates $\mathrm{prox}^\delta_{\gamma f_i}(x)$ such that $\|\mathrm{prox}^\delta_{\gamma f_i}(x) - \mathrm{prox}_{\gamma f}(x)\|^2 \leq \delta$ (see Section E).

Remark 7.4. Li et al. (2024) also establish a linear rate for FedExProx. However, their result relies on the assumption that the function f is strongly convex, and hence that (1) has a unique solution. In contrast, our theorem is more general, as it allows for the possibility of multiple solutions.

A similar result can be established in the partial participation scenario.

Theorem 7.5. *Let Assumptions 1.1, 1.2, 1.3, and 7.1 hold. For all $\gamma > 0$,* FedExProx *with nice sampling (Algorithm 2) and $\alpha = {}^{1}/_{\gamma L_\gamma}$ finds $\bar{x}$ such that $\mathbb{E}[f(\bar{x})] - f(x_*) \leq \varepsilon$ in*

$$\mathcal{O}\left(\frac{L_{\gamma,S}}{\mu_\gamma^+}\log\frac{1}{\varepsilon}\right)$$

iterations, where $L_{\gamma,S} := \left(\frac{n-S}{S(n-1)}\frac{L_{\max}}{1+\gamma L_{\max}} + \frac{n(S-1)}{S(n-1)}L_\gamma\right)$, L_γ is a smoothness constant of M^γ and μ_γ^+ is the PŁ constant.

Note that the complexities in Theorems 7.2 and 7.5 are entirely analogous to those in Theorems 4.1 and 5.1, with the only distinction being the substitution of the smallest non-zero eigenvalue of $\nabla^2 M^\gamma$ with the PŁ constant.

7.1 Why do we get a tighter analysis?

The story behind our theoretical improvements is no less important than the improvements themselves, as it offers valuable insights and can be instructive for future research. To understand why the new analysis yields stronger guarantees, we need to examine the reasoning behind Theorem 3.2. When proving the result, we noticed that the main reason why (9) was true was the dependence of the complexity on the $L_{\max}$ factor. At the same time, the complexity of GD depends only on L, and there exist many examples when $L_{\max} \gg L$. The question was: why does the original analysis of FedExProx involve $L_{\max}$, and can one circumvent this dependency? The reason behind it

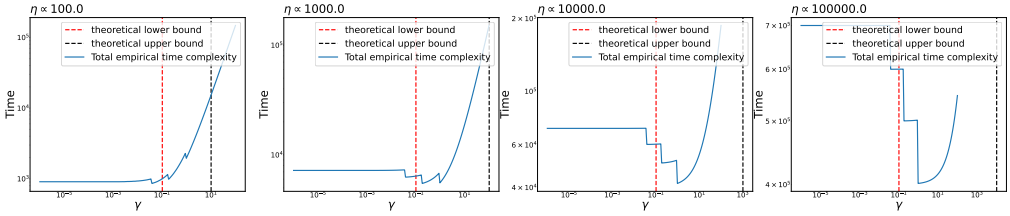

Figure 1: Empirical time complexities of FedExProx on a quadratic optimization task.

is the reliance of the proofs by Li et al. (2024) on Lemma A.9, which establishes the inequality $M^{\gamma}(x) - M^{\gamma}(x_*) \geq \frac{1}{1+\gamma L_{\max}}\left(f(x) - f(x_*)\right)$. This result appears to be essential for deriving the convergence result $\mathbb{E}[f(\bar{x})] - f(x_*) \leq \varepsilon$ (where $\bar{x} \in \mathbb{R}^d$ is the output of FedExProx). Our main idea is to instead obtain convergence in terms of *distances*, i.e., to establish that $\mathbb{E}[\|\bar{x} - x_*\|^2] \leq \varepsilon$, and then translate it to $\mathbb{E}[f(\bar{x})] - f(x_*) \leq \varepsilon$ using the L–smoothness of the function f. In this way, one can avoid using Lemma A.9, ultimately obtaining much more favourable convergence guarantees.

8 Experimental Highlights

This section presents highlights from illustrative experiments; additional details and results–including those on the ARCENE dataset (Guyon et al., 2004)–are provided in Section G. We consider quadratic objectives of the form $f(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2} x^\top \boldsymbol{A}_i x$, where $\boldsymbol{A}_i \in \mathrm{Sym}_+^7$, $i \in [n]$ are random positive semidefinite matrices with minimum eigenvalue equal to zero. Each worker $i \in [14]$ computes proximal mappings using GD with stepsize $1/\bar{L}_i$. The extrapolation parameter α is set to its optimal value from Theorem 4.1. To validate our theory, we study the empirical time complexity of FedExProx as a function of γ. In accordance with the setup described in Section 4.1, we assume that one local iteration of GD takes τ seconds (without loss of generality, $\tau = 1$). Thus, the time needed by worker i to find $\mathrm{prox}_{\gamma f_i}(x_k)$ at global iteration k is proportional to $\tau \times n_{ik}$, where n_{ik} is the number of GD iterations needed to find $\mathrm{prox}_{\gamma f_i}(x_k)$ to a given accuracy. In the full participation case, the total empirical time complexity is $\sum_{k=0}^{K-1}(\eta + \tau \max_{i\in[n]} n_{ki})$, where K is the number of global iterations needed for FedExProx to converge to the desired accuracy and η is the communication time. The results are presented in Figure 1. The dashed lines represent the theoretical bounds from Theorem 4.3, within which the optimal γ is expected to lie. One can see that when η is relatively small ($\eta \propto 100$), the best choice of γ is near 0. However, as the communication cost η increases, the best γ shifts to values greater than 0.1. A distinctive U-shape emerges, indicating the nontrivial optimal choice of γ. These observations are fully consistent with our theoretical predictions.

9 Conclusion

In this work, we revisit the extrapolated parallel proximal method FedExProx, an algorithm that has shown strong empirical performance but whose theoretical guarantees have so far lagged behind practice. Upon re-examining the state-of-the-art analysis, we find that its guarantees are overly pessimistic and that significantly stronger results are achievable. To address this gap, we develop a novel analytical framework for non-strongly convex quadratic and PŁ cases, yielding substantially improved convergence guarantees and a clearer picture of the method's true performance. To put these findings in context, we compare FedExProx with vanilla GD. This comparison is primarily pedagogical but reveals a striking insight: previous analyses could not even show an advantage over this simplest baseline. In contrast, our work is the first to rigorously demonstrate that FedExProx can outperform GD, bridging a key gap between empirical observations and theoretical understanding. Although our focus is on relatively simple quadratic problems, experiments suggest that the phenomena we describe extend well beyond this setting. These results indicate that the benefits of extrapolation in FL hold in broader optimization scenarios, motivating future research to relax our assumptions and establish similar guarantees in more general and practical settings. We view this work as a foundational stepping stone toward a deeper understanding of extrapolation in FL, paving the way for advancing both the theory and practice of federated optimization.

ACKNOWLEDGMENTS

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST): i) KAUST Baseline Research Scheme, ii) CRG Grant ORFS-CRG12-2024-6460, and iii) Center of Excellence for Generative AI, under award number 5940.

REFERENCES

Amir Beck. *First-Order Methods in Optimization*. SIAM-Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2017. ISBN 1611974984. (Cited on page 15)

Dimitri P Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical programming*, 129(2):163–195, 2011. (Cited on page 2)

Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *arXiv preprint arXiv:2002.12410*, 2020. (Cited on page 2)

P. Bianchi. Ergodic convergence of a stochastic proximal point algorithm. *SIAM Journal on Optimization*, 26(4):2235–2260, 2016. (Cited on page 3)

P. L. Combettes and J. I. Madariaga. Randomly activated proximal methods for nonsmooth convex minimization. preprint arXiv:2403.10673, 2024. (Cited on page 3)

Patrick L Combettes. Convex set theoretic image recovery by extrapolated iterations of parallel subgradient projections. *IEEE Transactions on Image Processing*, 6(4):493–506, 1997. (Cited on page 3)

L. Condat and P. Richtárik. RandProx: Primal-dual optimization algorithms with randomized proximal updates. In *Proc. of Int. Conf. Learning Representations (ICLR)*, 2023. (Cited on page 3)

Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. Local SGD: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pp. 3556–3564. PMLR, 2021. (Cited on page 2)

Kaja Gruntkowska, Alexander Tyurin, and Peter Richtárik. Improving the worst-case bidirectional communication complexity for nonconvex distributed optimization under function similarity. *arXiv preprint arXiv:2402.06412*, 2024. (Cited on page 2)

Isabelle Guyon, Steve Gunn, Asa Ben-Hur, and Gideon Dror. Arcene. UCI Machine Learning Repository, 2004. DOI: https://doi.org/10.24432/C58P55. (Cited on page 9 and 44)

Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018. (Cited on page 1)

Samuel Horváth, Konstantin Mishchenko, and Peter Richtárik. Adaptive learning rates for faster stochastic gradient methods. *arXiv preprint arXiv:2208.05287*, 2022. (Cited on page 7 and 37)

Divyansh Jhunjhunwala, Shiqiang Wang, and Gauri Joshi. FedExP: Speeding up federated averaging via extrapolation. *arXiv preprint arXiv:2301.09604*, 2023. (Cited on page 2 and 3)

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021. (Cited on page 1, 2, and 5)

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pp. 795–811. Springer, 2016. (Cited on page 16 and 17)

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020. (Cited on page 2)

Ahmed Khaled and Chi Jin. Faster federated optimization under second-order similarity. *arXiv preprint arXiv:2209.02257*, 2022. (Cited on page 2)

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. First analysis of local GD on heterogeneous data. *arXiv preprint arXiv:1909.04715*, 2019. (Cited on page 2)

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning*, pp. 5381–5393. PMLR, 2020. (Cited on page 2)

Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016. (Cited on page 1)

Hanmin Li, Kirill Acharya, and Peter Richtarik. The power of extrapolation in federated learning. *arXiv preprint arXiv:2405.13766*, 2024. (Cited on page 3, 4, 6, 7, 8, 9, 15, 20, 21, and 32)

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020. (Cited on page 3)

Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989. (Cited on page 2)

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *International Conference on Machine Learning*, pp. 3325–3334. PMLR, 2018. (Cited on page 3)

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017. (Cited on page 1 and 2)

Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. ProxSkip: Yes! local gradient steps provably lead to communication acceleration! Finally! In *International Conference on Machine Learning*, pp. 15750–15769. PMLR, 2022. (Cited on page 2)

J.J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965. (Cited on page 3)

Ion Necoara, Peter Richtárik, and Andrei Patrascu. Randomized projection methods for convex feasibility: Conditioning and convergence rates. *SIAM Journal on Optimization*, 29(4):2814–2852, 2019. (Cited on page 3)

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018. (Cited on page 1, 2, and 17)

A. Patrascu and I. Necoara. Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *Journal of Machine Learning Research*, 18:1–42, 2018. (Cited on page 3)

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021. (Cited on page 2)

Peter Richtárik and Martin Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156:433–484, 2016. (Cited on page 16)

Peter Richtárik and Martin Takáč. Stochastic reformulations of linear systems: algorithms and convergence theory. *SIAM Journal on Matrix Analysis and Applications*, 41(2):487–524, 2020. (Cited on page 5)

Peter Richtárik, Abdurakhmon Sadiev, and Yury Demidovich. A unified theory of stochastic proximal point methods without smoothness. *arXiv preprint arXiv:2405.15941*, 2024. (Cited on page 2)

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020. (Cited on page 1)

R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.*, 14(5):877–898, 1976. (Cited on page 2)

R. Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften. Springer Berlin, Heidelberg, 1998. (Cited on page 22)

Ernest K Ryu and Stephen Boyd. Stochastic proximal iteration: a non-asymptotic improvement upon stochastic gradient descent. *Author website, early draft*, 2014. (Cited on page 2 and 3)

Abdurakhmon Sadiev, Laurent Condat, and Peter Richtárik. Stochastic proximal point methods for monotone inclusions under expected similarity. *arXiv preprint arXiv:2405.14255*, 2024. (Cited on page 3)

Cheik Traoré, Vassilis Apidopoulos, Saverio Salzo, and Silvia Villa. Variance reduction techniques for stochastic proximal point algorithms. *Journal of Optimization Theory and Applications*, pp. 1–30, 2024. (Cited on page 3)

Alexander Tyurin and Peter Richtárik. A computation and communication efficient method for distributed nonconvex problems in the partial participation setting. *Advances in Neural Information Processing Systems*, 2023. (Cited on page 2)

Alexander Tyurin, Lukang Sun, Konstantin Burlachenko, and Peter Richtárik. Sharper rates and flexible framework for nonconvex SGD with client and data sampling. *arXiv preprint arXiv:2206.02275*, 2022. (Cited on page 17)

APPENDIX

CONTENTS

1 Introduction **1**

1.1 Related Work 1

2 Contributions **3**

3 Not Better than GD on Quadratics **4**

4 Tighter and More Optimistic Results **5**

4.1 One step time complexity $\pi(\gamma)$ is $\eta + \tau\left(\gamma L_{\max} + 1\right)$ 5

5 Partial Participation **6**

6 Adaptivity **7**

7 Better Theory with PŁ Condition **8**

7.1 Why do we get a tighter analysis? 8

8 Experimental Highlights **9**

9 Conclusion **9**

A Background **15**

A.1 Useful Facts about Moreau Envelope 15

A.2 Nice Sampling 16

A.3 Note on Other Sampling Strategies 17

A.4 Useful Lemmas 17

B FedExProx as (S)GD on Moreau Envelope Reformulation **18**

C FedExProx for Quadratics **19**

C.1 Properties of the Hessian 20

C.2 Proofs for Section 3 21

C.3 Tighter Analysis 23

C.3.1 Full Participation Case 23

C.3.2 Partial Participation Case 26

C.4 Lemmas 30

D FedExProx under PŁ Condition **32**

E FedExProx with Inexact Computations **34**

F Adaptive Extrapolation **37**
F.1 Full Participation Case 37
F.2 Partial Participation Case 39

G Experiments **42**
G.1 Experimental Setup 42
G.2 Experiments with Quadratic Optimization Problems 42
G.3 Experiments with Smooth Hinge Loss 43
G.3.1 Synthetic Data 43
G.3.2 Real Data 44

H Notation **45**

A BACKGROUND

Before discussing the contributions of this work, we provide formal definitions and essential facts that will be repeatedly referenced in the proofs. A summary of the commonly used notation is provided in Table 1.

Definition A.1 (Proximal operator)**.** The *proximal operator* $\text{prox}_{\gamma f} : \mathbb{R}^d \to \mathbb{R}^d$ of f is defined by

$$\text{prox}_{\gamma f}(x) = \arg\min_{z \in \mathbb{R}^d} \left\{ f(z) + \frac{1}{2\gamma} \|z - x\|^2 \right\},$$

where $\|\cdot\|$ is the standard Euclidean norm.

Definition A.2 (Moreau envelope)**.** The *Moreau envelope* of an extended-real-valued function $f : \mathbb{R}^d \to \mathbb{R}^d \cup \{\infty\}$ with stepsize $\gamma > 0$ is the function $M_f^\gamma : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$M_f^\gamma(x) = \min_{y \in \mathbb{R}^d} \left\{ f(y) + \frac{1}{2\gamma} \|y - x\|^2 \right\}$$

for all $x \in \mathbb{R}^d$.

A.1 USEFUL FACTS ABOUT MOREAU ENVELOPE

We will often rely on several useful properties of the Moreau envelopes, summarized below. In what follows, we denote

$$M^\gamma(x) := \frac{1}{n} \sum_{i=1}^n M_{f_i}^\gamma(x).$$

Lemma A.3 (Beck (2017) Theorem 6.60)**.** *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be proper, closed and convex. Then its Moreau envelope M_f^γ satisfies*

$$\nabla M_f^\gamma(x) = \frac{1}{\gamma}(x - \text{prox}_{\gamma f}(x))$$

for all $x \in \mathbb{R}^d$, for any $\gamma > 0$.

Lemma A.4 ((Li et al., 2024) Lemma 4)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and L–smooth. Then M_f^γ is $\frac{L}{1+\gamma L}$–smooth.*

Lemma A.5 ((Li et al., 2024) Lemma 5)**.** *Let $f : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ be a proper, closed and convex function. Then, for all $\gamma > 0$, f and $M_f^\gamma(x)$ have the same set of minimizers and minimum.*

Lemma A.6 ((Beck, 2017) Theorem 6.55)**.** *Let $f : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ be a proper, closed and convex function. Then M_f^γ is also a convex function.*

Lemma A.7 ((Li et al., 2024) Lemma 7)**.** *Let the functions $f_i : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$, $i \in [n]$, be proper, closed, convex and L_i–smooth. Then M^γ is convex and L_γ–smooth with*

$$\frac{1}{n^2} \sum_{i=1}^n \frac{L_i}{1 + \gamma L_i} \leq L_\gamma \leq \frac{1}{n} \sum_{i=1}^n \frac{L_i}{1 + \gamma L_i}.$$

Lemma A.8 ((Li et al., 2024) Lemma 8)**.** *Let the functions $f_i : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$, $i \in [n]$, be proper, closed and convex, and suppose that Assumption 1.3 holds. Then, for all $\gamma \in (0, \infty)$, the function $f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$ has the same set of minimizers and minimum as the function $M^\gamma(x)$.*

Lemma A.9. *[Li et al. (2024) Lemma 10] Let Assumptions 1.1, 1.2 and 1.3 hold. Then, for any $x_* \in \mathcal{X}_*$ and for all $x \in \mathbb{R}^d$, it holds that*

$$M_{f_i}^\gamma(x) - M_{f_i}^\gamma(x_*) \geq \frac{1}{1 + \gamma L_i} \left(f_i(x) - f_i(x_*) \right).$$

Consequently,

$$M^\gamma(x) - M^\gamma(x_*) \geq \frac{1}{1 + \gamma L_{\max}} \left(f(x) - f(x_*) \right).$$

The next lemma will play a role in the analysis of the method when assuming the Polyak-Łojasiewicz (PŁ) condition.

Assumption A.10 (Polayk-Łojasiewicz condition)**.** The function f satisfies PŁ condition, i.e., there exists $\eta > 0$ such that

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \eta\left(f(x) - f(x_*)\right)$$

for all $x \in \mathbb{R}^d$.

Lemma A.11. *Let the function f satisfy Assumption A.10 with parameter μ^+, and suppose that Assumptions 1.1, 1.2 and 1.3 hold. Then, Assumption 7.1 holds with parameter*

$$\mu_\gamma^+ \geq \frac{\mu^+}{1+\gamma L_{\max}}.$$

Proof. If f satisfies Assumption A.10, then under Assumption 1.1, Polayk-Łojasiewicz condition implies

$$f(x) - f(x_*) \geq \frac{\mu^+}{2}\|x - \Pi(x)\|^2 \quad \forall x \in \mathbb{R}^d, \tag{17}$$

(Karimi et al., 2016), where $\Pi(x)$ is the projection of x onto the solution set $\mathcal{X}_*$ of f and $x_* \in \mathcal{X}_*$. Then, using Lemma A.9, we get

$$M^\gamma(x) - M^\gamma(x_*) \geq \frac{1}{1+\gamma L_{\max}}\left(f(x) - f(x_*)\right) \overset{(17)}{\geq} \frac{\mu^+}{2(1+\gamma L_{\max})}\|x - \Pi(x)\|^2.$$

Due to Lemma A.8, M^γ and f share the set of minimizers. Thus, $\Pi(x)$ is the projection of x onto the solution set $\mathcal{X}_*$ of M^γ. Using convexity, we get

$$M^\gamma(\Pi(x)) \geq M^\gamma(x) + \langle \nabla M^\gamma(x), \Pi(x) - x\rangle$$

and

$$\begin{aligned} M^\gamma(x) - M^\gamma(\Pi(x)) &\leq \langle \nabla M^\gamma(x), x - \Pi(x)\rangle \leq \|\nabla M^\gamma(x)\|\,\|x - \Pi(x)\| \\ &\leq \|\nabla M^\gamma(x)\|\sqrt{\frac{2(1+\gamma L_{\max})}{\mu^+}}\sqrt{M^\gamma(x) - M^\gamma(\Pi(x))}. \end{aligned}$$

Therefore

$$\frac{\mu^+}{4(1+\gamma L_{\max})}\left(M^\gamma(x) - M^\gamma(\Pi(x))\right) \leq \frac{1}{2}\|\nabla M^\gamma(x)\|^2,$$

meaning that M^γ satisfies PŁ condition with parameter $\frac{\mu^+}{4(1+\gamma L_{\max})}$. □

A.2 Nice Sampling

Let us formally introduce the sampling strategy used. Fix a minibatch size $S \in [n]$ and let $\mathcal{S}$ be a random subset of $[n]$ of size S, chosen uniformly at random from all $\binom{n}{S}$ subsets of $[n]$ of this size. Such a random set $\mathcal{S}$ is known in the literature under the name *S-nice sampling* (Richtárik & Takáč, 2016).

In the proofs, we will rely on the following useful lemma:

Lemma A.12. *Fix $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_n \in \mathbb{R}^{k\times l}$ and let $\mathcal{S}$ be an S-nice sampling of the indices $[n]$. Then*

$$\begin{aligned} &\mathbb{E}\left[\left(\frac{1}{S}\sum_{i\in\mathcal{S}}\boldsymbol{B}_i\right)^\top\left(\frac{1}{S}\sum_{i\in\mathcal{S}}\boldsymbol{B}_i\right)\right] \\ &\quad= \frac{n-S}{S(n-1)}\frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_i^\top\boldsymbol{B}_i + \frac{n(S-1)}{S(n-1)}\left(\frac{1}{n}\sum_{i=1}^n\boldsymbol{B}_i\right)^\top\left(\frac{1}{n}\sum_{i=1}^n\boldsymbol{B}_i\right). \end{aligned}$$

Proof. Define

$$\chi_i := \begin{cases} 1 & i \in \mathcal{S}, \\ 0 & i \notin \mathcal{S}. \end{cases}$$

Then

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{S}\sum_{i\in\mathcal{S}} \boldsymbol{B}_i\right)^\top \left(\frac{1}{S}\sum_{i\in\mathcal{S}} \boldsymbol{B}_i\right)\right] \\
&= \mathbb{E}\left[\left(\frac{1}{S}\sum_{i=1}^{n} \chi_i \boldsymbol{B}_i\right)^\top \left(\frac{1}{S}\sum_{i=1}^{n} \chi_i \boldsymbol{B}_i\right)\right] \\
&= \frac{1}{S^2}\mathbb{E}\left[\sum_{i=1}^{n} \chi_i^2 \boldsymbol{B}_i^\top \boldsymbol{B}_i + \sum_{i\neq j} \chi_i\chi_j \boldsymbol{B}_i^\top \boldsymbol{B}_j\right] \\
&= \frac{1}{S^2}\left(\sum_{i=1}^{n} \mathbb{E}\left[\chi_i\right] \boldsymbol{B}_i^\top \boldsymbol{B}_i + \sum_{i\neq j} \mathbb{E}\left[\chi_i\chi_j\right] \boldsymbol{B}_i^\top \boldsymbol{B}_j\right) \\
&= \frac{1}{S^2}\left(\frac{S}{n}\sum_{i=1}^{n} \boldsymbol{B}_i^\top \boldsymbol{B}_i + \frac{S(S-1)}{n(n-1)}\sum_{i\neq j} \boldsymbol{B}_i^\top \boldsymbol{B}_j\right) \\
&= \frac{1}{S}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{B}_i^\top \boldsymbol{B}_i + \frac{S-1}{n(n-1)}\left(\left(\sum_{i=1}^{n} \boldsymbol{B}_i\right)^\top \left(\sum_{i=1}^{n} \boldsymbol{B}_i\right) - \sum_{i=1}^{n} \boldsymbol{B}_i^\top \boldsymbol{B}_i\right)\right) \\
&= \frac{n-S}{S(n-1)}\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{B}_i^\top \boldsymbol{B}_i + \frac{n(S-1)}{S(n-1)}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{B}_i\right)^\top \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{B}_i\right).
\end{aligned}$$

□

A.3 Note on Other Sampling Strategies

As discussed in Section 5, Algorithm 2 can handle virtually any (unbiased) sampling technique, with nice sampling used here as an illustrative example. For a comprehensive overview of other potential sampling strategies, we direct the reader to Tyurin et al. (2022), which offers a general framework for analyzing a broad range of sampling mechanisms.

A.4 Useful Lemmas

Lemma A.13 ((Karimi et al., 2016) Theorem 2)**.** *Assume that the function f is convex, L–smooth, and satisfies the Polyak-Łojasiewicz condition with parameter η. Then*

$$f(x) - f(x_*) \geq \frac{\eta}{2}\left\|x - \Pi(x)\right\|^2,$$

for all $x \in \mathbb{R}^d$, where $\Pi(x)$ is the projection of x onto the solution set $\mathcal{X}_*$ and $x_* \in \mathcal{X}_*$.

Lemma A.14 ((Nesterov, 2018))**.** *Let $D_f(x,y) := f(x) - f(y) - \langle \nabla f(y), x - y\rangle$ be the Bregman divergence of the function f. If f is convex, L–smooth and differentiable, then*

$$\frac{1}{L}\left\|\nabla f(x) - \nabla f(y)\right\|^2 \leq D_f(x,y) + D_f(y,x)$$

for any $x, y \in \mathbb{R}^d$.

Algorithm 1 FedExProx

1: **Parameters:** stepsize $\gamma > 0$, extrapolation parameter $\alpha_t > 0$, starting point $x_0 \in \mathbb{R}^d$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: $\quad x_{k+1} = x_k + \alpha_k \left(\frac{1}{n}\sum_{i=1}^n \text{prox}_{\gamma f_i}(x_k) - x_k\right)$
4: **end for**

B FEDEXPROX AS (S)GD ON MOREAU ENVELOPE REFORMULATION

The key idea behind the original analysis of FedExProx is to rewrite the step

$$x_{k+1} = x_k + \alpha_k \left(\frac{1}{n}\sum_{i=1}^n \text{prox}_{\gamma f_i}(x_k) - x_k\right)$$

in terms of the Moreau envelopes of the local functions. Indeed, if the functions f_i are proper, closed and convex for all $i \in [n]$, by Lemma A.3, the update rule of Algorithm 1 can be rewritten as

$$x_{k+1} = x_k - \alpha_k \gamma \frac{1}{n}\sum_{i=1}^n \nabla M_{f_i}^{\gamma}(x_k).$$

Analogously, in the partial participation case (Algorithm 2), we have

$$x_{k+1} = x_k - \alpha_k \gamma \frac{1}{S}\sum_{i \in \mathcal{S}_k} \nabla M_{f_i}^{\gamma}(x_k).$$

It follows that Algorithms 1 and 2 are equivalent to GD and minibatch SGD for minimizing

$$M^{\gamma}(x) := \frac{1}{n}\sum_{i=1}^n M_{f_i}^{\gamma}(x)$$

with stepsize $\alpha_k \gamma$.

C FEDEXPROX FOR QUADRATICS

We begin by formally introducing the problem setup and the notation used throughout the proofs.

We are interested in solving the quadratic optimization problem

$$\min_{x\in\mathbb{R}^n}\left\{f(x) := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}x^\top \boldsymbol{A}_i x - b_i^\top x\right)\right\}, \tag{18}$$

where $\boldsymbol{A}_i \in \mathrm{Sym}_+^d := \{\boldsymbol{X} \in \mathbb{R}^{d\times d} \,|\, \boldsymbol{X} = \boldsymbol{X}^\top, \boldsymbol{X} \succeq 0\}$ and $b_i \in \mathbb{R}^d$, assuming the interpolation regime (Assumption 1.3), which in this case states that there exists x_* such that $\boldsymbol{A}_i x_* = b_i$ for all $i \in [n]$.

We denote the set of minimizers by $\mathcal{X}_* := \{x \in \mathbb{R}^d : \frac{1}{n}\sum_{i=1}^n(\boldsymbol{A}_i x - b_i) = 0\}$ and let $\Pi(\cdot)$ be the projection onto $\mathcal{X}_*$:

$$\Pi(x) = x - \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{A}_i\right)^\dagger\left(\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{A}_i x - b_i)\right), \tag{19}$$

where $\dagger$ denotes the Moore-Penrose pseudoinverse.

Using eigendecomposition, each matrix $\boldsymbol{A}_i$, $i \in [n]$ can be decomposed as $\boldsymbol{A}_i = \boldsymbol{Q}_i\boldsymbol{\Lambda}_i\boldsymbol{Q}_i^\top$, where $\boldsymbol{Q}_i = [q_{i,1}, \ldots, q_{i,d}] \in \mathbb{R}^{d\times d}$ is the orthogonal matrix whose columns are eigenvectors of $\boldsymbol{A}_i$ and $\boldsymbol{\Lambda}_i \in \mathbb{R}^{d\times d}$ is a diagonal matrix whose diagonal elements are the corresponding eigenvalues of $\boldsymbol{A}_i$, i.e., $\boldsymbol{\Lambda}_i = \mathrm{diag}\left(\lambda_1(\boldsymbol{A}_i), \ldots, \lambda_d(\boldsymbol{A}_i)\right)$, where $0 = \lambda_1(\boldsymbol{A}_i) \le \lambda_2(\boldsymbol{A}_i) \le \ldots \le \lambda_d(\boldsymbol{A}_i)$. Let $\lambda_{\min}^+(\boldsymbol{A}_i)$ denote the smallest non-zero eigenvalue of $\boldsymbol{A}_i$ and $\lambda_{\max}(\boldsymbol{A}_i)$–the largest eigenvalue of $\boldsymbol{A}_i$.

In general, the largest eigenvalue, smallest eigenvalue, and smallest positive eigenvalue of any matrix $\boldsymbol{B}$ will be denoted by $\lambda_{\max}(\boldsymbol{B})$, $\lambda_{\min}(\boldsymbol{B})$ and $\lambda_{\min}^+(\boldsymbol{B})$, respectively.

The proximal operators in this setting are

$$\mathrm{prox}_{\gamma f_i}(x) = \left(\gamma\boldsymbol{A}_i + \boldsymbol{I}\right)^{-1}\left(x + \gamma b_i\right). \tag{20}$$

Thus, by Lemma A.3,

$$\nabla M_{f_i}^\gamma(x) = \frac{1}{\gamma}\left(x - \left(\gamma\boldsymbol{A}_i + \boldsymbol{I}\right)^{-1}\left(x + \gamma b_i\right)\right) = \frac{1}{\gamma}\left(\boldsymbol{I} - \left(\gamma\boldsymbol{A}_i + \boldsymbol{I}\right)^{-1}\right)(x - x_*) \tag{21}$$

for any $x_* \in \mathcal{X}_*$, where we use the fact that $\boldsymbol{A}_i x_* = b_i$. Furthermore, it can easily be shown that

$$\begin{aligned} M_{f_i}^\gamma(x) &= \frac{1}{2\gamma}\left(x^T(\boldsymbol{I} - (\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1})x - 2\gamma b_i^T(\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1}x - \gamma^2 b_i^T(\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1}b_i\right) \\ &= \frac{1}{2\gamma}\left((x - x_*)^T(\boldsymbol{I} - (\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1})(x - x_*) - \gamma x_*^T\boldsymbol{A}_i x_*\right), \end{aligned} \tag{22}$$

and hence

$$\begin{aligned} M^\gamma(x) &= \frac{1}{2}x^T\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\gamma}(\boldsymbol{I} - (\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1})\right)x - \left(\frac{1}{n}\sum_{i=1}^{n}(\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1}b_i\right)^T x \\ &\quad - \frac{\gamma}{2n}\sum_{i=1}^{n}b_i^T(\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1}b_i \\ &= \frac{1}{2}x^T\boldsymbol{M}x - \left(\frac{1}{n}\sum_{i=1}^{n}(\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1}b_i\right)^T x - \frac{\gamma}{2n}\sum_{i=1}^{n}b_i^T(\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1}b_i \end{aligned} \tag{23}$$

where $\boldsymbol{M} := \frac{1}{n}\sum_{i=1}^n\frac{1}{\gamma}(\boldsymbol{I} - (\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1})$ is the Hessian matrix of the function M^γ. This can equivalently be written as

$$M^\gamma(x) = \frac{1}{2}(x - x_*)^T\boldsymbol{M}(x - x_*) - x_*^T\left(\frac{1}{2n}\sum_{i=1}^{n}\boldsymbol{A}_i\right)x_* \tag{24}$$

for any $x_* \in \mathcal{X}_*$. We shall denote by $\lambda_{\min}^+(\gamma)$ and $\lambda_{\max}(\gamma)$ the smallest non-zero eigenvalue and the largest eigenvalue of $\boldsymbol{M}$, respectively.

C.1 Properties of the Hessian

Next, we introduce several important properties of the matrix $\boldsymbol{M}$, which will be repeatedly used in the proofs that follow.

In the main part of the paper, we stick to the notation L_γ and μ_γ^+ for easier comparison with the results by Li et al. (2024). However, in the proofs concerning the quadratic case, we adopt a more intuitive eigenvalue-based notation.

Fact 1. *Since $\boldsymbol{M} = \nabla^2 M^\gamma$, it holds that $L_\gamma = \lambda_{\max}(\gamma)$ and $\mu_\gamma^+ = \lambda_{\min}^+(\gamma)$.*

Proof. The result is an immediate consequence of (23). □

Fact 2. *Let $\lambda_{\max}(\gamma)$ and $\lambda_{\min}^+(\gamma)$ be the smallest non-zero eigenvalue and the largest eigenvalue of the matrix*

$$\boldsymbol{M} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\gamma}\left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right),$$

respectively. Then

$$\lambda_{\max}(\gamma) = \max_{\|x\|\leq 1}\frac{1}{n}\sum_{i=1}^{n} x^\top \boldsymbol{Q}_i \left[\frac{\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)}\right]_{jj} \boldsymbol{Q}_i^\top x \tag{25}$$

and

$$\lambda_{\min}^+(\gamma) = \min_{\|x\|=1, x\in\ker(\boldsymbol{A})^\perp}\frac{1}{n}\sum_{i=1}^{n} x^\top \boldsymbol{Q}_i \left[\frac{\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)}\right]_{jj} \boldsymbol{Q}_i^\top x, \tag{26}$$

where $\boldsymbol{A} := \frac{1}{n}\sum_{i=1}^n \boldsymbol{A}_i$ and $[b_j]_{jj}$ is the diagonal matrix with b_j as the jth entry.

Proof. Using the eigendecomposition, we can write $\boldsymbol{M}$ as

$$\boldsymbol{M} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\gamma}\boldsymbol{Q}_i\left(\boldsymbol{I} - (\gamma\boldsymbol{\Lambda}_i + \boldsymbol{I})^{-1}\right)\boldsymbol{Q}_i^\top = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Q}_i\left[\frac{\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)}\right]_{jj}\boldsymbol{Q}_i^\top.$$

The result follows from the identities

$$\lambda_{\max}(\gamma) = \max_{\|x\|\leq 1} x^\top \boldsymbol{M} x$$

and

$$\lambda_{\min}^+(\gamma) = \min_{\|x\|=1, x\in(\ker \boldsymbol{A})^\perp} x^\top \boldsymbol{M} x$$

(see Lemma C.7). □

Fact 3. *For any $\gamma \geq 0$, it holds that $\lambda_{\max}(\gamma) \leq \lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}^+(\gamma) \leq \lambda_{\min}^+(\boldsymbol{A})$.*

Proof. From Fact 2, we have

$$\lambda_{\max}(\gamma) \overset{(25)}{=} \max_{\|x\|\leq 1}\frac{1}{n}\sum_{i=1}^{n} x^T \boldsymbol{Q}_i \left[\frac{\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)}\right]_{jj} \boldsymbol{Q}_i^T x,$$

and similarly

$$\lambda_{\max}(\boldsymbol{A}) = \max_{\|x\|\leq 1}\frac{1}{n}\sum_{i=1}^{n} x^T \boldsymbol{Q}_i \left[\lambda_j(\boldsymbol{A}_i)\right]_{jj} \boldsymbol{Q}_i^T x.$$

Since it always holds that $\frac{\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)} \leq \lambda_j(\boldsymbol{A}_i)$, we get

$$\lambda_{\max}(\gamma) \leq \lambda_{\max}(\boldsymbol{A}).$$

The proof of the second inequality is analogous. □

C.2 PROOFS FOR SECTION 3

Theorem 3.2. *Let Assumptions 1.3 and 3.1 hold. Consider solving a non-strongly convex quadratic optimization problem of the form (1), where $f_i(x) = \frac{1}{2}x^\top \boldsymbol{A}_i x - b_i^\top x$ for all $i \in [n]$, with $\boldsymbol{A}_i \in \mathrm{Sym}_+^d$ and $b_i \in \mathbb{R}^d$. Then the assumptions of Theorem 1 by Li et al. (2024) hold, and*

$$T(\gamma) = \pi(\gamma) \times \frac{L_\gamma(1+\gamma L_{\max})R^2}{\varepsilon} \geq \pi(0) \times \frac{LR^2}{\varepsilon} \tag{9}$$

for all $\gamma > 0$. Moreover, when $\gamma \to 0$, then $\pi(\gamma) \times \frac{L_\gamma(1+\gamma L_{\max})R^2}{\varepsilon} \to \pi(0) \times \frac{LR^2}{\varepsilon}$, and FedExProx *effectively reduces to* GD.

Proof. We first prove that $L_\gamma(1+\gamma L_{\max}) \geq L = \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{A}_i\right)$ for all $\gamma \geq 0$. Since

$$\nabla^2 M_{f_i}(x) = \frac{1}{\gamma}\left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right),$$

we have

$$L_\gamma = \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \frac{1}{\gamma}\left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)\right).$$

Therefore, using the fact that $L_{\max} = \max_i \lambda_{\max}(\boldsymbol{A}_i)$, we obtain

$$\begin{aligned}(1+\gamma L_{\max})L_\gamma &= \left(1 + \gamma \max_i \lambda_{\max}(\boldsymbol{A}_i)\right)\frac{1}{\gamma}\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)\right) \\ &= \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \frac{1+\gamma \max_i \lambda_{\max}(\boldsymbol{A}_i)}{\gamma}\left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)\right),\end{aligned} \tag{27}$$

and hence, it is sufficient to show that

$$\frac{1+\gamma \max_i \lambda_{\max}(\boldsymbol{A}_i)}{\gamma}\left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right) \succeq \boldsymbol{A}_i. \tag{28}$$

Using the eigenvalue decomposition of $\boldsymbol{A}_i$, the expression can be rewritten as

$$\begin{aligned}\frac{1+\gamma \max_i \lambda_{\max}(\boldsymbol{A}_i)}{\gamma}\left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right) &= \frac{1+\gamma \max_i \lambda_{\max}(\boldsymbol{A}_i)}{\gamma}\left(\boldsymbol{I} - \left(\gamma \boldsymbol{Q}_i \boldsymbol{\Lambda}_i \boldsymbol{Q}_i^\top + \boldsymbol{I}\right)^{-1}\right) \\ &= \frac{1+\gamma \max_i \lambda_{\max}(\boldsymbol{A}_i)}{\gamma}\boldsymbol{Q}_i\left(\boldsymbol{I} - (\gamma \boldsymbol{\Lambda}_i + \boldsymbol{I})^{-1}\right)\boldsymbol{Q}_i^\top.\end{aligned}$$

Now, letting $[\boldsymbol{B}]_j$ be the jth diagonal element of matrix $\boldsymbol{B}$, we have

$$\begin{aligned}&\left[\frac{1+\gamma \max_i \lambda_{\max}(\boldsymbol{A}_i)}{\gamma}\left(\boldsymbol{I} - (\gamma \boldsymbol{\Lambda}_i + \boldsymbol{I})^{-1}\right)\right]_j = \frac{1+\gamma \max_i \lambda_{\max}(\boldsymbol{A}_i)}{\gamma}\left[\boldsymbol{I} - (\gamma \boldsymbol{\Lambda}_i + \boldsymbol{I})^{-1}\right]_j \\ &= \frac{1+\gamma \max_i \lambda_{\max}(\boldsymbol{A}_i)}{\gamma}\left(1 - \frac{1}{\gamma[\boldsymbol{\Lambda}_i]_j + 1}\right) = \frac{1+\gamma \max_i \lambda_{\max}(\boldsymbol{A}_i)}{\gamma}\frac{\gamma\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)} \geq \lambda_j(\boldsymbol{A}_i),\end{aligned}$$

and hence

$$\frac{(1+\gamma \max_i \lambda_{\max}(\boldsymbol{A}_i))}{\gamma}\left(\boldsymbol{I} - (\gamma \boldsymbol{\Lambda}_i + \boldsymbol{I})^{-1}\right) \succeq \boldsymbol{\Lambda}_i.$$

Multiplying by $\boldsymbol{Q}_i$ from the left and by $\boldsymbol{Q}_i^\top$ from the right, we see that (28) holds, and we can conclude that

$$(1+\gamma L_{\max})L_\gamma \geq \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{A}_i\right) = L$$

for all $\gamma \geq 0$. Inequality (9) follows form Assumption 3.1.

We now turn to the second part of the theorem and show that $\pi(\gamma) \times \frac{L_\gamma(1+\gamma L_{\max})R^2}{\varepsilon} \to \pi(0) \times \frac{LR^2}{\varepsilon}$ as $\gamma \to 0$. The idea behind the proof is that for small enough γ, it holds that

$$(1+\gamma L_{\max})L_\gamma \overset{(27)}{=} \left(1+\gamma \max_i \lambda_{\max}(\boldsymbol{A}_i)\right) \frac{1}{\gamma}\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)\right)$$
$$\approx \frac{1}{\gamma}\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \gamma \boldsymbol{A}_i\right) = \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{A}_i\right) = L.$$

More formally, we have

$$\boldsymbol{I} - (\boldsymbol{I} + \gamma \boldsymbol{A}_i)^{-1} = \gamma \boldsymbol{A}_i - \gamma^2 \boldsymbol{A}_i^2 + \gamma^3 \boldsymbol{A}_i^3 + \cdots = \sum_{k=1}^\infty (-1)^{k-1}\gamma^k \boldsymbol{A}_i^k,$$

so that

$$(1+\gamma L_{\max})L_\gamma = \left(\frac{1}{\gamma} + \max_i \lambda_{\max}(\boldsymbol{A}_i)\right)\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \sum_{k=1}^\infty (-1)^{k-1}\gamma^k \boldsymbol{A}_i^k\right).$$

By applying the dominated convergence theorem to exchange the limit operation and summation, we obtain

$$\lim_{\gamma\to 0}(1+\gamma L_{\max})L_\gamma = \lim_{\gamma\to 0}\left(\left(1+\gamma \max_i \lambda_{\max}(\boldsymbol{A}_i)\right)\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \sum_{k=1}^\infty (-1)^{k-1}\gamma^{k-1} \boldsymbol{A}_i^k\right)\right)$$
$$= \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \sum_{k=1}^\infty \lim_{\gamma\to 0}(-1)^{k-1}\gamma^{k-1} \boldsymbol{A}_i^k\right) = \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{A}_i\right) = L,$$

as required.

Lastly, letting $\gamma \to 0$, we have $M_{f_i}^\gamma(x) \to f_i(x)$ for all x, and consequently $M^\gamma(x) \to f(x)$ (see, e.g., (Rockafellar & Wets, 1998)). Since Algorithm 1 is equivalent to GD for minimizing $M^\gamma(x)$ with stepsize $\alpha\gamma = 1/L_\gamma$, it follows that its iterates converge to those of vanilla GD with stepsize $1/L$. □

C.3 TIGHTER ANALYSIS

C.3.1 FULL PARTICIPATION CASE

Iteration complexity. We first establish the convergence rate of the algorithm.

Theorem C.1. *Let Assumption 1.3 hold and let x_k be the iterates of* FedExProx *(Algorithm 1) applied to problem (18). Then*

$$\|x_K - \Pi(x_K)\|^2 \leq \left(1 - \alpha\gamma\left(2 - \alpha\gamma\lambda_{\max}(\gamma)\right)\lambda^+_{\min}(\gamma)\right)^K \|x_0 - \Pi(x_0)\|^2,$$

where $\lambda_{\max}(\gamma)$ and $\lambda^+_{\min}(\gamma)$ are the largest and the smallest positive eigenvalues of the matrix $\boldsymbol{M} := \frac{1}{\gamma}\left(\boldsymbol{I} - \frac{1}{n}\sum_{i=1}^n(\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)$, respectively.

The optimal choice of stepsize γ and extrapolation parameter α is $\alpha\gamma = \frac{1}{\lambda_{\max}(\gamma)}$, in which case the rate becomes

$$\|x_K - \Pi(x_K)\|^2 \leq \left(1 - \frac{\lambda^+_{\min}(\gamma)}{\lambda_{\max}(\gamma)}\right)^K \|x_0 - \Pi(x_0)\|^2.$$

Hence, the number of iterations needed to reach an ϵ-solution is

$$K \geq \frac{\lambda_{\max}(\gamma)}{\lambda^+_{\min}(\gamma)} \log\left(\frac{\|x_0 - \Pi(x_0)\|^2}{\epsilon}\right).$$

Proof of Theorem C.1. The update rule of FedExProx can be written as

$$\begin{aligned}
x_{k+1} &= x_k + \alpha\left(\frac{1}{S}\sum_{i=1}^S \mathrm{prox}_{\gamma f_i}(x_k) - x_k\right) \\
&\overset{(20)}{=} x_k + \alpha\left(\frac{1}{S}\sum_{i=1}^S(\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1}(x_k + \gamma b_i) - x_k\right) \\
&= x_k + \alpha\left(\frac{1}{S}\sum_{i=1}^S((\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1} - \boldsymbol{I})\right)x_k - \alpha\left(\frac{1}{S}\sum_{i=1}^S((\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1} - \boldsymbol{I})\right)\Pi(x_k),
\end{aligned}$$

where we use the fact that $\boldsymbol{A}_i x_* = b_i$ for all $x_* \in \mathcal{X}_*$. Therefore, using symmetry of $\boldsymbol{A}_i$

$$\begin{aligned}
\|x_{k+1} - \Pi(x_{k+1})\|^2 &\leq \|x_{k+1} - \Pi(x_k)\|^2 \\
&= \left\|\left((1-\alpha)\boldsymbol{I} + \alpha\frac{1}{n}\sum_{i=1}^n(\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)(x_k - \Pi(x_k))\right\|^2 \\
&= \left\|\left(\boldsymbol{I} - \alpha\underbrace{\left(\frac{1}{n}\sum_{i=1}^n(\boldsymbol{I} - (\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1})\right)}_{:=\gamma\boldsymbol{M}}\right)(x_k - \Pi(x_k))\right\|^2 \\
&= (x_k - \Pi(x_k))^T(\boldsymbol{I} - \alpha\gamma\boldsymbol{M})^T(\boldsymbol{I} - \alpha\gamma\boldsymbol{M})(x_k - \Pi(x_k)) \\
&= (x_k - \Pi(x_k))^T\left(\boldsymbol{I} - 2\alpha\gamma\boldsymbol{M} + \alpha^2\gamma^2\boldsymbol{M}^2\right)(x_k - \Pi(x_k)).
\end{aligned}$$

Since $\boldsymbol{M}^2 \preceq \lambda_{\max}(\gamma)\boldsymbol{M}$, it follows that

$$\begin{aligned}
\|x_{k+1} - \Pi(x_{k+1})\|^2 &\leq (x_k - \Pi(x_k))^T\left(\boldsymbol{I} - 2\alpha\gamma\boldsymbol{M} + \lambda_{\max}(\gamma)\alpha^2\gamma^2\boldsymbol{M}\right)(x_k - \Pi(x_k)) \\
&= \|x_k - \Pi(x_k)\|^2 - \alpha\gamma(2 - \lambda_{\max}(\gamma)\alpha\gamma)(x_k - \Pi(x_k))^T\boldsymbol{M}(x_k - \Pi(x_k)) \\
&\leq \|x_k - \Pi(x_k)\|^2 - \alpha\gamma(2 - \lambda_{\max}(\gamma)\alpha\gamma)\lambda^+_{\min}(\gamma)\|x_k - \Pi(x_k)\|^2
\end{aligned}$$

for any $\alpha\gamma$ such that $2 - \lambda_{\max}(\gamma)\alpha\gamma \geq 0$, where the last inequality follows from the fact that $x_k - \Pi(x_k) \in \mathrm{range}(\boldsymbol{M})$ (see Lemma C.6). Consequently,

$$\|x_{k+1} - \Pi(x_{k+1})\|^2 \leq \left(1 - \alpha\gamma(2 - \alpha\gamma\lambda_{\max}(\gamma))\lambda^+_{\min}(\gamma)\right)\|x_k - \Pi(x_k)\|^2.$$

It remains to unroll the recurrence and substitute the optimal step size $\alpha\gamma = \frac{1}{\lambda_{\max}(\gamma)}$ to obtain the rate

$$\|x_{k+1} - \Pi(x_{k+1})\|^2 \leq \left(1 - \frac{\lambda^+_{\min}(\gamma)}{\lambda_{\max}(\gamma)}\right) \|x_k - \Pi(x_k)\|^2 .$$

□

Theorem 4.1. *Fix any $\gamma > 0$ and consider solving non-strongly convex quadratic optimization problem (1) where $f_i(x) = \frac{1}{2}x^\top \boldsymbol{A}_i x - b_i^\top x$ for all $i \in [n]$, with $\boldsymbol{A}_i \in \mathrm{Sym}^d_+$ and $b_i \in \mathbb{R}^d$. Under Assumption 1.3,* FedExProx *with $\alpha = 1/\gamma L_\gamma$ finds $\bar{x}$ such that $\mathbb{E}[f(\bar{x})] - f(x_*) \leq \varepsilon$ after*

$$\mathcal{O}\left(\frac{L_\gamma}{\mu^+_\gamma} \log \frac{1}{\varepsilon}\right) \tag{10}$$

iterations, where L_γ is a smoothness constant of M^γ and μ^+_γ is the smallest non-zero eigenvalue of the matrix $\nabla^2 M^\gamma$.

Proof. The result follows directly from Theorem C.1, smoothness of f, and Fact 1 by substituting $\lambda_{\max}(\gamma) = \lambda_{\max}(\nabla^2 M^\gamma) = L_\gamma$ and $\lambda^+_{\min}(\gamma) = \lambda^+_{\min}(\nabla^2 M^\gamma) = \mu^+_\gamma$. □

Remark C.2. Recall from the proof of Fact 2 that

$$\boldsymbol{M} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{Q}_i \left[\frac{\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)}\right]_{jj} \boldsymbol{Q}_i^\top,$$

where $\frac{\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)} \in \left[\frac{\lambda_j(\boldsymbol{A}_i)}{2}, \lambda_j(\boldsymbol{A}_i)\right]$ if $\gamma \leq \frac{1}{\lambda_j(\boldsymbol{A}_i)}$. Consequently, for $\gamma \leq \frac{1}{\max_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)}$, we have

$$\boldsymbol{M} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{Q}_i \left[\frac{\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)}\right]_{jj} \boldsymbol{Q}_i^\top \preceq \frac{1}{n}\sum_{i=1}^n \boldsymbol{Q}_i \left[\lambda_j(\boldsymbol{A}_i)\right]_{jj} \boldsymbol{Q}_i^\top = \frac{1}{n}\sum_{i=1}^n \boldsymbol{A}_i = \boldsymbol{A},$$

and analogously, $\boldsymbol{M} \succeq \frac{1}{2}\boldsymbol{A}$. Using (25) and (26) gives

$$\frac{1}{2}\lambda_{\max}(\boldsymbol{A}) = \frac{1}{2}\max_{\|x\|\leq 1} x^\top \boldsymbol{A} x \leq \lambda_{\max}(\gamma) = \max_{\|x\|\leq 1} x^\top \boldsymbol{M} x \leq \max_{\|x\|\leq 1} x^\top \boldsymbol{A} x = \lambda_{\max}(\boldsymbol{A}),$$

and similarly

$$\frac{1}{2}\lambda^+_{\min}(\boldsymbol{A}) \leq \lambda^+_{\min}(\gamma) = \min_{\|x\|=1, x\in(\ker \boldsymbol{A})^\perp} x^\top \boldsymbol{M} x \leq \lambda^+_{\min}(\boldsymbol{A}).$$

Noting that $L_\gamma = \lambda_{\max}(\gamma)$, $\mu^+_\gamma = \lambda^+_{\min}(\gamma)$, $L = \lambda_{\max}(\boldsymbol{A})$ and $\mu^+ = \lambda^+_{\min}(\boldsymbol{A})$, it follows that the rates (10) of FedExProx and (30) of GD coincide up to a constant factor when γ is sufficiently small. This should come as no surprise, as Theorem 3.2 shows that FedExProx effectively reduces to GD as $\gamma \to 0$.

Time complexity. Let us now prove the result from Section 4.1. We assume that each worker i computes the proximal operator using an iterative method, with the running time proportional to the condition number of the subproblem. Given that $\lambda_{\min}(\boldsymbol{A}_i) = 0$ for all $i \in [n]$ (otherwise the problem becomes trivial), the time required for all clients to compute $\mathrm{prox}_{\gamma f_i}(x_k)$ at each global iteration k is

$$\tau \times \left(1 + \gamma \max_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)\right)$$

seconds, where τ is the time per one iteration of solving the subproblem. Then, it takes η seconds to aggregate the updates at the server and move on to the next step. Since, according to Theorem C.1, the number of iterations needed to reach an ϵ-solution is

$$K = \frac{\lambda_{\max}(\gamma)}{\lambda^+_{\min}(\gamma)} \log\left(\frac{\|x_0 - \Pi(x_0)\|^2}{\epsilon}\right),$$

the total time required to solve the global problem is

$$T_\eta(\gamma) := \frac{\lambda_{\max}(\gamma)}{\lambda^+_{\min}(\gamma)} \log\left(\frac{\|x_0 - \Pi(x_0)\|^2}{\epsilon}\right) \times \left(\eta + \tau \times \left(1 + \gamma \max_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)\right)\right). \tag{29}$$

Theorem 4.3. *Consider the non-strongly convex quadratic optimization problem from Theorem 4.1. Up to a constant factor, the time complexity (12) is minimized by some*

$$\gamma \in \left[\frac{1}{\max\limits_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)}, \min\left\{\frac{\frac{\eta}{\tau}-1}{\max\limits_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)}, \frac{1}{\min\limits_{i\in[n]} \lambda^{+}_{\min}(\boldsymbol{A}_i)}\right\}\right]$$

if $\frac{\eta}{\tau} \geq 2$ *and by some*

$$\gamma \in \left[0, \max\left\{0, \min\left\{\frac{\frac{\eta}{\tau}-1}{\max\limits_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)}, \frac{1}{\min\limits_{i\in[n]} \lambda^{+}_{\min}(\boldsymbol{A}_i)}\right\}\right\}\right]$$

if $\frac{\eta}{\tau} < 2$. *Moreover,* $T_\eta(\gamma) \leq T_{\mathsf{GD}}$.

Proof. Recall that $\lambda_{\max}(\gamma)$ and $\lambda^{+}_{\min}(\gamma)$ are the eigenvalues of the matrix

$$\boldsymbol{M} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\gamma}\left(\boldsymbol{I} - (\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1}\right) \overset{(2)}{=} \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Q}_i\left[\frac{\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)}\right]_{jj}\boldsymbol{Q}_i^\top.$$

Since if $\gamma \geq \frac{1}{\lambda_j(\boldsymbol{A}_i)}$ and $\lambda_j(\boldsymbol{A}_i) > 0$, then $\frac{\gamma\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)} \in \left[\frac{1}{2}, 1\right)$, and if $\gamma \leq \frac{1}{\lambda_j(\boldsymbol{A}_i)}$, then $\frac{\gamma\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)} \in \left[\frac{\gamma\lambda_j(\boldsymbol{A}_i)}{2}, \gamma\lambda_j(\boldsymbol{A}_i)\right)$, it follows that

$$\frac{\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)} = \begin{cases}\Theta\left(\frac{1}{\gamma}\right), & \gamma \geq \frac{1}{\lambda_j(\boldsymbol{A}_i)},\\ \Theta\left(\lambda_j(\boldsymbol{A}_i)\right), & \gamma < \frac{1}{\lambda_j(\boldsymbol{A}_i)}.\end{cases} \tag{30}$$

Moreover, the identities (25) and (26) from Fact 2 tell us that

$$\lambda_{\max}(\gamma) = \max_{\|x\|\leq 1}\frac{1}{n}\sum_{i=1}^{n}x^\top\boldsymbol{Q}_i\left[\frac{\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)}\right]_{jj}\boldsymbol{Q}_i^\top x$$

and

$$\lambda^{+}_{\min}(\gamma) = \min_{\|x\|=1, x\in\ker(\boldsymbol{A})^\perp}\frac{1}{n}\sum_{i=1}^{n}x^\top\boldsymbol{Q}_i\left[\frac{\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)}\right]_{jj}\boldsymbol{Q}_i^\top x,$$

where $\boldsymbol{A} := \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{A}_i$. Due to (25), (26) and (30), the function $T_\eta(\gamma)$ is approximately constant for all $\gamma \leq \frac{1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}$. If $\gamma \geq \frac{\frac{\eta}{\tau}-1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}$ then $\tau(1+\gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)) \geq \eta$, and consequently,

$$\begin{aligned}T_\eta(\gamma) &= \frac{\lambda_{\max}(\gamma)}{\lambda^{+}_{\min}(\gamma)}\log\left(\frac{\|x_0-\Pi(x_0)\|^2}{\epsilon}\right)\times\left(\eta+\tau\left(1+\gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)\right)\right)\\ &\leq 2\tau\frac{\lambda_{\max}(\gamma)}{\lambda^{+}_{\min}(\gamma)}\log\left(\frac{\|x_0-\Pi(x_0)\|^2}{\epsilon}\right)\times\left(1+\gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)\right) := 2\bar{g}(\gamma).\end{aligned}$$

Hence, $\bar{g}(\gamma) \leq T_\eta(\gamma) \leq 2\bar{g}(\gamma)$, and the term to be minimized is

$$\bar{T}_\eta(\gamma) := \frac{\lambda_{\max}(\gamma)}{\lambda^{+}_{\min}(\gamma)}\times\left(1+\gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)\right).$$

Using (25) and (26), this can be written as

$$\bar{T}_\eta(\gamma) = \frac{\max_{\|x\|\leq 1}\frac{1}{n}\sum_{i=1}^{n}x^\top\boldsymbol{Q}_i\left[\frac{1+\gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)}\lambda_j(\boldsymbol{A}_i)\right]_{jj}\boldsymbol{Q}_i^\top x}{\min_{\|x\|=1, x\in\ker(\boldsymbol{A})^\perp}\frac{1}{n}\sum_{i=1}^{n}x^\top\boldsymbol{Q}_i\left[\frac{\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)}\right]_{jj}\boldsymbol{Q}_i^\top x},$$

where the numerator is non-decreasing and the denominator is non-increasing as a function of γ. It follows that $\bar{g}(\gamma)$ is non-decreasing for $\gamma \geq \frac{\frac{\eta}{\tau}-1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}$. Hence, when $\frac{\eta}{\tau} \geq 2$, the optimal (up to a multiplicative factor) γ belongs to the interval

$$\left[\frac{1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}, \frac{\frac{\eta}{\tau}-1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}\right]. \tag{31}$$

Algorithm 2 FedExProx with partial participation

1: **Parameters:** stepsize $\gamma > 0$, extrapolation parameter $\alpha_k > 0$, starting point $x_0 \in \mathbb{R}^d$, batch size S
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Sample a minibatch $\mathcal{S}_k \subseteq [n]$ uniformly from all subsets of cardinality S
4: $x_{k+1} = x_k + \alpha_k \left(\frac{1}{S}\sum_{i\in\mathcal{S}_k} \mathrm{prox}_{\gamma f_i}(x_k) - x_k\right)$
5: **end for**

When $\frac{\eta}{\tau} < 2$, the time-complexity is non-decreasing for $\gamma \geq \frac{\frac{\eta}{\tau}-1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}$ and approximately constant otherwise, so the choice

$$\gamma \in \left[0, \max\left\{0, \frac{\frac{\eta}{\tau}-1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}\right\}\right] \tag{32}$$

is optimal (again, up to a constant factor).

Now, suppose that $\gamma \geq \frac{1}{\min_{i\in[n]}\lambda^+_{\min}(\boldsymbol{A}_i)}$. In this case, the diagonal entries of the matrix $\left[\frac{\gamma\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)}\right]_{jj}$ are either to 0 (when $\lambda_j(\boldsymbol{A}_i) = 0$) or lie within the interval $\left[\frac{1}{2}, 1\right)$. Therefore, the ratio

$$\frac{\lambda_{\max}(\gamma)}{\lambda^+_{\min}(\gamma)} = \frac{\max_{\|x\|\leq 1}\frac{1}{n}\sum_{i=1}^n x^\top \boldsymbol{Q}_i \left[\frac{\gamma\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)}\right]_{jj} \boldsymbol{Q}_i^\top x}{\min_{\|x\|=1, x\in\ker(\boldsymbol{A})^\perp}\frac{1}{n}\sum_{i=1}^n x^\top \boldsymbol{Q}_i \left[\frac{\gamma\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)}\right]_{jj} \boldsymbol{Q}_i^\top x}$$

is approximately constant in γ. Consequently, the time

$$T_\eta(\gamma) := \frac{\lambda_{\max}(\gamma)}{\lambda^+_{\min}(\gamma)} \log\left(\frac{\|x_0 - \Pi(x_0)\|^2}{\epsilon}\right) \times \left(\eta + \tau\left(1 + \gamma \max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)\right)\right)$$

is an increasing function of γ for $\gamma \geq \frac{1}{\min_{i\in[n]}\lambda^+_{\min}(\boldsymbol{A}_i)}$, and the optimal range for γ can be tightened from the previously established bounds (31) and (32) to

$$\left[\frac{1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}, \min\left\{\frac{\frac{\eta}{\tau}-1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}, \frac{1}{\min_{i\in[n]}\lambda^+_{\min}(\boldsymbol{A}_i)}\right\}\right]$$

when $\frac{\eta}{\tau} \geq 2$ and

$$\gamma \in \left[0, \max\left\{0, \min\left\{\frac{\frac{\eta}{\tau}-1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}, \frac{1}{\min_{i\in[n]}\lambda^+_{\min}(\boldsymbol{A}_i)}\right\}\right\}\right]$$

when $\frac{\eta}{\tau} < 2$. □

C.3.2 Partial Participation Case

Iteration complexity. We again start with establishing the iteration complexity of the algorithm.

Theorem C.3. *Let Assumption 1.3 hold and let x_k be the iterates of minibatch* FedExProx *(Algorithm 2) applied to problem* (18)*. Then*

$$\mathbb{E}\left[\|x_K - \Pi(x_K)\|^2\right] \leq \left(1 - \alpha\gamma\left(2 - \alpha\gamma L_{\gamma,S}\right)\lambda^+_{\min}(\gamma)\right)^K \|x_0 - \Pi(x_0)\|^2,$$

where $\lambda^+_{\min}(\gamma)$ is the smallest positive eigenvalue of the matrix $\boldsymbol{M} := \frac{1}{\gamma}\left(\boldsymbol{I} - \frac{1}{n}\sum_{i=1}^n(\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)$ and $L_{\gamma,S} := \frac{n-S}{S(n-1)}\frac{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}{1+\gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)} + \frac{n(S-1)}{S(n-1)}\lambda_{\max}(\gamma)$.

The optimal choice of stepsize γ and extrapolation parameter α is $\alpha\gamma = \frac{1}{L_{\gamma,S}}$, in which case the rate becomes

$$\mathbb{E}\left[\|x_K - \Pi(x_K)\|^2\right] \leq \left(1 - \frac{\lambda^+_{\min}(\gamma)}{L_{\gamma,S}}\right)^K \|x_0 - \Pi(x_0)\|^2.$$

Hence, the number of iterations needed to reach an ϵ-solution is

$$K \geq \frac{L_{\gamma,S}}{\lambda_{\min}^{+}(\gamma)} \log\left(\frac{\|x_0 - \Pi(x_0)\|^2}{\epsilon}\right). \tag{33}$$

Remark C.4. Similar to the proof of Theorem 4.1, the result of Theorem 5.1 follows directly from Theorem C.3, smoothness of f, and Fact 1 by noting that $\lambda_{\max}(\gamma) = \lambda_{\max}(\nabla^2 M^\gamma) = L_\gamma$ and $\lambda_{\min}^{+}(\gamma) = \lambda_{\min}^{+}(\nabla^2 M^\gamma) = \mu_\gamma^{+}$.

Proof of Theorem C.3. The update rule of FedExProx can be written as

$$\begin{aligned}
x_{k+1} &= x_k + \alpha\left(\frac{1}{S}\sum_{i\in\mathcal{S}_k} \mathrm{prox}_{\gamma f_i}(x_k) - x_k\right) \\
&= x_k + \alpha\left(\frac{1}{S}\sum_{i\in\mathcal{S}_k} (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}(x_k + \gamma b_i) - x_k\right) \\
&= x_k + \alpha\left(\frac{1}{S}\sum_{i\in\mathcal{S}_k} ((\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1} - \boldsymbol{I})\right) x_k - \alpha\left(\frac{1}{S}\sum_{i\in\mathcal{S}_k} ((\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1} - \boldsymbol{I})\right)\Pi(x_k),
\end{aligned}$$

where we use the fact that $\boldsymbol{A}_i x_* = b_i$ for all $x_* \in \mathcal{X}_*$. Therefore

$$\begin{aligned}
&\|x_{k+1} - \Pi(x_{k+1})\|^2 \\
&\leq \|x_{k+1} - \Pi(x_k)\|^2 \\
&= \left\|\left((1-\alpha)\boldsymbol{I} + \alpha\frac{1}{S}\sum_{i\in\mathcal{S}_k}(\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)(x_k - \Pi(x_k))\right\|^2 \\
&= \|x_k - \Pi(x_k)\|^2 - 2\alpha(x_k - \Pi(x_k))^T\left(\boldsymbol{I} - \frac{1}{S}\sum_{i\in\mathcal{S}_k}(\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)(x_k - \Pi(x_k)) \\
&\quad + \alpha^2(x_k - \Pi(x_k))^T\left(\boldsymbol{I} - \frac{1}{S}\sum_{i\in\mathcal{S}_k}(\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)^T\left(\boldsymbol{I} - \frac{1}{S}\sum_{i\in\mathcal{S}_k}(\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)(x_k - \Pi(x_k)) \\
&= \|x_k - \Pi(x_k)\|^2 - 2\alpha\gamma(x_k - \Pi(x_k))^T \boldsymbol{M}_k(x_k - \Pi(x_k)) \\
&\quad + \alpha^2\gamma^2(x_k - \Pi(x_k))^T \boldsymbol{M}_k^T \boldsymbol{M}_k(x_k - \Pi(x_k)),
\end{aligned} \tag{34}$$

where $\boldsymbol{M}_k := \frac{1}{\gamma}\left(\boldsymbol{I} - \frac{1}{S}\sum_{i\in\mathcal{S}_k}(\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)$. Next, using the fact that $\boldsymbol{M}^2 \preceq \lambda_{\max}(\gamma)\boldsymbol{M}$ and applying Lemma A.12 with $\boldsymbol{B}_i = \frac{1}{\gamma}\left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)$, the expectation of $\boldsymbol{M}_k^T \boldsymbol{M}_k$ is

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{M}_k^T \boldsymbol{M}_k\right] &= \frac{1}{\gamma^2}\frac{n-S}{S(n-1)}\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)^\top\left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right) + \frac{n(S-1)}{S(n-1)}\boldsymbol{M}^\top \boldsymbol{M} \\
&= \frac{1}{\gamma^2}\frac{n-S}{S(n-1)}\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)^2 + \frac{n(S-1)}{S(n-1)}\boldsymbol{M}^2 \\
&\leq \frac{1}{\gamma}\frac{n-S}{S(n-1)}\max_{i\in[n]}\lambda_{\max}\left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\gamma}\left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right) \\
&\quad + \frac{n(S-1)}{S(n-1)}\lambda_{\max}(\gamma)\boldsymbol{M} \\
&= \frac{1}{\gamma}\frac{n-S}{S(n-1)}\max_{i\in[n]}\left(\frac{\gamma\lambda_{\max}(\boldsymbol{A}_i)}{1+\gamma\lambda_{\max}(\boldsymbol{A}_i)}\right)\boldsymbol{M} + \frac{n(S-1)}{S(n-1)}\lambda_{\max}(\gamma)\boldsymbol{M} \\
&= \underbrace{\left(\frac{n-S}{S(n-1)}\frac{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}{1+\gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)} + \frac{n(S-1)}{S(n-1)}\lambda_{\max}(\gamma)\right)}_{:=L_{\gamma,S}}\boldsymbol{M}.
\end{aligned} \tag{35}$$

Hence, taking expectation conditioned on x_k in (34) gives

$$\begin{aligned}
&\mathbb{E}_k\left[\|x_{k+1}-\Pi(x_{k+1})\|^2\right] \\
&\quad\leq \|x_k-\Pi(x_k)\|^2+\alpha^2\gamma^2(x_k-\Pi(x_k))^T\mathbb{E}_k\left[\boldsymbol{M}_k^T\boldsymbol{M}_k\right](x_k-\Pi(x_k)) \\
&\qquad -2\alpha\gamma(x_k-\Pi(x_k))^T\mathbb{E}_k\left[\boldsymbol{M}_k\right](x_k-\Pi(x_k)) \\
&\quad\overset{(35)}{=} \|x_k-\Pi(x_k)\|^2+\alpha^2\gamma^2L_{\gamma,S}(x_k-\Pi(x_k))^T\boldsymbol{M}(x_k-\Pi(x_k)) \\
&\qquad -2\alpha\gamma(x_k-\Pi(x_k))^T\boldsymbol{M}(x_k-\Pi(x_k)) \\
&\quad= \|x_k-\Pi(x_k)\|^2-\alpha\gamma\left(2-\alpha\gamma L_{\gamma,S}\right)(x_k-\Pi(x_k))^T\boldsymbol{M}(x_k-\Pi(x_k)) \\
&\quad\leq \|x_k-\Pi(x_k)\|^2-\alpha\gamma\left(2-\alpha\gamma L_{\gamma,S}\right)\lambda_{\min}^+(\gamma)(x_k-\Pi(x_k))^T(x_k-\Pi(x_k)),
\end{aligned}$$

where the last inequality follows from the fact that $x_k-\Pi(x_k)\in\operatorname{range}(\boldsymbol{M})$ (see Lemma C.6). Taking expectation again,

$$\mathbb{E}\left[\|x_{k+1}-\Pi(x_{k+1})\|^2\right]\leq\left(1-\alpha\gamma\left(2-\alpha\gamma L_{\gamma,S}\right)\lambda_{\min}^+(\gamma)\right)\mathbb{E}\left[\|x_k-\Pi(x_k)\|^2\right].$$

It remains to unroll the recurrence. □

Time complexity. Let us consider the same setup as in Section C.3.1, i.e., at each iteration k, each client computes a proximal operator $\operatorname{prox}_{\gamma f_i}(x_k)$ in at most

$$\tau\times\left(1+\gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)\right)$$

seconds, where τ is the time per one iteration of solving the subproblem, and it takes η seconds to aggregate the updates at the server and move on to the next step. According to Theorem C.3, the number of iterations needed to reach an ϵ-solution is

$$K=\frac{L_{\gamma,S}}{\lambda_{\min}^+(\gamma)}\log\left(\frac{\|x_0-\Pi(x_0)\|^2}{\epsilon}\right),$$

where $L_{\gamma,S}:=\frac{n-S}{S(n-1)}\frac{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}{1+\gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}+\frac{n(S-1)}{S(n-1)}\lambda_{\max}(\gamma)$. Hence, the total time required to solve the global problem is at most

$$T_\eta(\gamma,S):=\frac{L_{\gamma,S}}{\lambda_{\min}^+(\gamma)}\log\left(\frac{\|x_0-\Pi(x_0)\|^2}{\epsilon}\right)\times\left(\eta+\tau\times\left(1+\gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)\right)\right). \tag{36}$$

Theorem 5.2. *Up to a constant factor, the time complexity* (14) *is minimized by*

$$\gamma\in\left[\frac{1}{\max\limits_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)},\min\left\{\frac{\frac{\eta}{\tau}-1}{\max\limits_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)},\frac{1}{\min\limits_{i\in[n]}\lambda_{\min}^+(\boldsymbol{A}_i)}\right\}\right]$$

if $\frac{\eta}{\tau}\geq 2$ *and by*

$$\gamma\in\left[0,\max\left\{0,\min\left\{\frac{\frac{\eta}{\tau}-1}{\max\limits_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)},\frac{1}{\min\limits_{i\in[n]}\lambda_{\min}^+(\boldsymbol{A}_i)}\right\}\right\}\right]$$

if $\frac{\eta}{\tau}<2$.

Proof. Fix some $S\in[n]$. Then, we are interested in minimizing

$$\begin{aligned}
\bar{T}_\eta(\gamma)\;=\;&\frac{n-S}{S(n-1)}\underbrace{\frac{1}{\lambda_{\min}^+(\gamma)}\frac{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}{1+\gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}\left(\eta+\tau\left(1+\gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)\right)\right)}_{:=\bar{g}_1(\gamma)} \\
&+\frac{n(S-1)}{S(n-1)}\underbrace{\frac{\lambda_{\max}(\gamma)}{\lambda_{\min}^+(\gamma)}\left(\eta+\tau\left(1+\gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)\right)\right)}_{:=\bar{g}_2(\gamma)}
\end{aligned}$$

$$:= \frac{n-S}{S(n-1)}\bar{g}_1(\gamma) + \frac{n(S-1)}{S(n-1)}\bar{g}_2(\gamma).$$

We know from Theorem 4.3 that the second term of this expression, $\bar{g}_2(\gamma)$, is minimized by

$$\left[\frac{1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}, \min\left\{\frac{\frac{\eta}{\tau}-1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}, \frac{1}{\min_{i\in[n]}\lambda^+_{\min}(\boldsymbol{A}_i)}\right\}\right]$$

when $\frac{\eta}{\tau} \geq 2$ and

$$\gamma \in \left[0, \max\left\{0, \min\left\{\frac{\frac{\eta}{\tau}-1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}, \frac{1}{\min_{i\in[n]}\lambda^+_{\min}(\boldsymbol{A}_i)}\right\}\right\}\right]$$

when $\frac{\eta}{\tau} < 2$.

The same argument can be applied to the term $\bar{g}_1(\gamma)$. To be more precise, due to (25), (26), and (30), the function $\bar{g}_1(\gamma)$ is approximately constant for all $\gamma \leq \frac{1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}$. For $\gamma > \frac{\frac{\eta}{\tau}-1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}$, we have $\frac{\eta}{\tau} + 1 + \gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i) < 2(1+\gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i))$. Consequently, letting $\bar{g}(\gamma) := 2\tau\frac{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}{\lambda^+_{\min}(\gamma)}$, $\bar{g}_1(\gamma)$ can be bounded as

$$\frac{1}{2}\bar{g}(\gamma) < \bar{g}_1(\gamma) < \bar{g}(\gamma).$$

Since $\lambda^+_{\min}(\gamma)$ is a non-increasing function of γ, $\bar{g}(\gamma)$ is non-decreasing. It follows that for $\gamma > \frac{\frac{\eta}{\tau}-1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}$, $\bar{g}_1(\gamma)$ is non-decreasing in γ (up to a constant factor), and we arrive at the same conclusions as in the case of $\bar{g}_2(\gamma)$: $\bar{g}_1(\gamma)$ is minimized by $\gamma \in \left[\frac{1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}, \frac{\frac{\eta}{\tau}-1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}\right]$ if $\frac{\eta}{\tau} \geq 2$ and by $\gamma \in \left[0, \max\left\{0, \frac{\frac{\eta}{\tau}-1}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}\right\}\right]$ if $\frac{\eta}{\tau} < 2$.

On the other hand, by following a similar argument as in the proof of Theorem 4.3, the expression

$$\frac{1}{\lambda^+_{\min}(\gamma)}\frac{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}{1+\gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)} = \frac{\frac{\gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}{1+\gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}}{\min_{\|x\|=1, x\in\ker(\boldsymbol{A})^\perp}\frac{1}{n}\sum_{i=1}^n x^\top \boldsymbol{Q}_i\left[\frac{\gamma\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)}\right]_{jj}\boldsymbol{Q}_i^\top x}$$

is approximately constant for $\gamma \geq \frac{1}{\min_{i\in[n]}\lambda^+_{\min}(\boldsymbol{A}_i)}$. This implies that $\bar{g}_1(\gamma)$ is increasing, and hence the intervals can be bounded above by $\frac{1}{\min_{i\in[n]}\lambda^+_{\min}(\boldsymbol{A}_i)}$.

The conclusion follows from the fact that the minimum of a convex combination of $\bar{g}_1(\gamma)$ and $\bar{g}_2(\gamma)$ must lie within the same interval. □

Remark C.5. The only term in (33) and (36) that depends on S is

$$L_{\gamma,S} := \frac{n-S}{S(n-1)}\frac{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}{1+\gamma\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)} + \frac{n(S-1)}{S(n-1)}\lambda_{\max}(\gamma).$$

Recall that $\lambda_{\max}(\gamma)$ is the largest eigenvalue of the matrix $\boldsymbol{M} = \frac{1}{n}\sum_{i=1}^n \frac{1}{2\gamma}(\boldsymbol{I} - (\boldsymbol{I}+\gamma\boldsymbol{A}_i)^{-1})$, so

$$\lambda_{\max}(\gamma) \leq \frac{1}{\gamma}\frac{1}{n}\sum_{i=1}^n \frac{\lambda_{\max}(\boldsymbol{A}_i)\gamma}{\lambda_{\max}(\boldsymbol{A}_i)\gamma+1} \leq \frac{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)}{\max_{i\in[n]}\lambda_{\max}(\boldsymbol{A}_i)\gamma+1}.$$

As a result, since $\frac{n-S}{S(n-1)}$ is decreasing, and $\frac{n(S-1)}{S(n-1)}$ is increasing in S, both the iteration and time complexities are increasing functions of S. This underscores the advantage of involving a larger number of clients in the training process.

C.4 LEMMAS

Lemma C.6. *Let x_k be the iterates of* SGD *applied to the problem $\min_x \left\{ f(x) := \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{2}x^T \boldsymbol{B}_i x + c_i^T x + d_i\right)\right\}$, where the matrices $\boldsymbol{B}_i$ are symmetric. Then $x_k - \Pi(x_k) \in \operatorname{range}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_i)$ for all k.*

Proof. By definition of Π, we have

$$\Pi(x) = x - \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_i\right)^\dagger \left(\frac{1}{n}\sum_{i=1}^n (\boldsymbol{B}_i x + c_i)\right).$$

Hence, using the identity $\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^\dagger = \boldsymbol{A}^\dagger$,

$$\begin{aligned} x_k - \Pi(x_k) &= \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_i\right)^\dagger \left(\frac{1}{n}\sum_{i=1}^n (\boldsymbol{B}_i x_k + c_i)\right) \\ &= \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_i\right)^T \left(\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_i\right)\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_i\right)^T\right)^\dagger \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_i \mathbb{E}\left[x_k\right] - c_i\right), \end{aligned}$$

and so, by symmetry of $\boldsymbol{B}_i$, we get $x_k - \Pi(x_k) \in \operatorname{range}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_i)$. □

Lemma C.7. *Let $\boldsymbol{M} := \frac{1}{n}\sum_{i=1}^n \frac{1}{\gamma}(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1})$, where $\gamma > 0$ and $\boldsymbol{A}_i \in \operatorname{Sym}_+^d$ for all $i \in [n]$. The smallest positive eigenvalue of $\boldsymbol{M}$ is given by*

$$\lambda_{\min}^+(\gamma) := \min_{\|x\|=1, x \in (\ker \boldsymbol{A})^\perp} \frac{1}{n}\sum_{i=1}^n x^\top \boldsymbol{Q}_i \left[\frac{\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)}\right]_{jj} \boldsymbol{Q}_i^\top x,$$

where $\boldsymbol{A} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{A}_i$ and $[b_j]_{jj}$ denotes a diagonal matrix with b_j as the jth entry.

Proof. First, observe that the matrices $\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}$ are symmetric, and their eigenvalues are given by

$$1 - \frac{1}{1+\gamma\lambda_j(\boldsymbol{A}_i)} = \frac{\gamma\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)}.$$

Consequently, $\boldsymbol{M}$ is a sum of symmetric positive definite matrices and is therefore also symmetric positive definite. We now claim that

$$\lambda_{\min}^+(\gamma) = \min_{\|x\|=1, x \in (\ker \boldsymbol{M})^\perp} x^\top \boldsymbol{M} x.$$

First, choosing x to be a multiple of the eigenvector of $\boldsymbol{M}$ corresponding to $\lambda_{\min}^+(\gamma)$, we see that

$$\lambda_{\min}^+(\gamma) \geq \min_{\|x\|=1, x \in (\ker \boldsymbol{M})^\perp} x^\top \boldsymbol{M} x.$$

To establish the reverse inequality, let $\{e_i\}$ be an orthonormal eigenbasis of $\boldsymbol{M}$ and let x be such that $\|x\| = 1$. Then we can write $x = \sum_{i=1}^d \alpha_i e_i$, and since $x \in (\ker \boldsymbol{M})^\perp$, all coefficients corresponding to an eigenvalue 0 vanish. Thus

$$x^\top \boldsymbol{M} x = \sum_{i=1}^d \alpha_i^2 \lambda_i \geq \lambda_{\min}^+(\gamma) \sum_{i=1}^d \alpha_i^2 = \lambda_{\min}^+(\gamma)\|x\|^2 = \lambda_{\min}^+(\gamma).$$

This proves that

$$\lambda_{\min}^+(\gamma) = \min_{\|x\|=1, x \in (\ker \boldsymbol{M})^\perp} x^\top \boldsymbol{M} x.$$

Now, each matrix $\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}$ can be decomposed as $\boldsymbol{Q}_i \left[\frac{\lambda_j(\boldsymbol{A}_i)}{1+\gamma\lambda_j(\boldsymbol{A}_i)}\right]_{jj} \boldsymbol{Q}_i^\top$. Therefore what remains to be proven is that $\ker(\boldsymbol{M}) = \ker(\boldsymbol{A})$. To this end, take $x \in \ker(\boldsymbol{M})$. Then $\boldsymbol{M}x = 0$, and since $\gamma > 0$, we have

$$\frac{1}{n}\sum_{i=1}^{n}(\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}x = x. \tag{37}$$

Next, observe that

$$\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n}(\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right) \leq \frac{1}{n}\sum_{i=1}^{n}\lambda_{\max}\left((\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\gamma\lambda_{\min}(\boldsymbol{A}_i)}.$$

Now consider two cases.

First, if there exists $j \in [n]$ such that $\lambda_{\min}(\boldsymbol{A}_j) > 0$, the above upper bound is strictly less than 1. This implies that there exists no nonzero x that satisfies equation (37), and so $\ker(\boldsymbol{M}) = \{0\}$. But $\lambda_{\min}(\boldsymbol{A}_j) > 0$ also implies that

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{A}_i\right) \geq \frac{1}{n}\sum_{i=1}^{n}\lambda_{\min}(\boldsymbol{A}_i) > 0.$$

As a result, $\ker(\boldsymbol{M}) = \{0\} = \ker(\boldsymbol{A})$.

Now, let us suppose that $\lambda_{\min}(\boldsymbol{A}_i) = 0$ for all $i \in [n]$. In this case, $\lambda_{\max}\left((\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}\right) = 1$ for all $i \in [n]$. Since all matrices $(\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}$ are symmetric positive definite with maximum eigenvalue equal to 1, it follows that equation (37) holds if and only if

$$(\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1}x = x$$

for all $i \in [n]$. This is equivalent to $\gamma \boldsymbol{A}_i x = 0$ for all $i \in [n]$. Since $\gamma > 0$ and $\boldsymbol{A}_i$ are symmetric positive definite, we can, in turn, equivalently express it as $\boldsymbol{A}x = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{A}_i x = 0$.

Consequently, (37) holds (i.e., $x \in \ker(\boldsymbol{M})$) if and only if $x \in \ker(\boldsymbol{A})$ and hence $\ker(\boldsymbol{A}) = \ker(\boldsymbol{M})$.

The final expression in the statement of the Lemma follows from eigendecomposition (see Fact 2). □

D FEDEXPROX UNDER PŁ CONDITION

Theorem D.1. *Let Assumptions 1.1, 1.2, 1.3, and 7.1 hold, and assume that $\alpha\gamma \leq \frac{2}{L_{\gamma,S}}$. Then the iterates of Algorithm 2 satisfy*

$$\mathbb{E}\left[\|x_K - \Pi(x_K)\|^2\right] \leq \left(1 - \alpha\gamma\left(2 - \alpha\gamma L_{\gamma,S}\right)\frac{\mu_\gamma^+}{2}\right)^K \mathbb{E}\left[\|x_0 - \Pi(x_0)\|^2\right]$$

and hence

$$\mathbb{E}\left[f(x_K) - f(\Pi(x_K))\right] \leq \frac{L}{2}\left(1 - \alpha\gamma\left(2 - \alpha\gamma L_{\gamma,S}\right)\frac{\mu_\gamma^+}{2}\right)^K \mathbb{E}\left[\|x_0 - \Pi(x_0)\|^2\right],$$

where $L_{\gamma,S} := \left(\frac{n-S}{S(n-1)}\frac{L_{\max}}{1+\gamma L_{\max}} + \frac{n(S-1)}{S(n-1)}L_\gamma\right)$. For the optimal choice of stepsize and extrapolation parameter $\alpha\gamma = 1/L_{\gamma,S}$, these rates become

$$\mathbb{E}\left[\|x_K - \Pi(x_K)\|^2\right] \leq \left(1 - \frac{\mu_\gamma^+}{2L_{\gamma,S}}\right)^K \mathbb{E}\left[\|x_0 - \Pi(x_0)\|^2\right]$$

and

$$\mathbb{E}\left[f(x_K) - f(\Pi(x_K))\right] \leq \frac{L}{2}\left(1 - \frac{\mu_\gamma^+}{2L_{\gamma,S}}\right)^K \mathbb{E}\left[\|x_0 - \Pi(x_0)\|^2\right].$$

Remark D.2. The above theorem proves Theorem 7.5. Furthermore, by setting $S = n$, we recover the result from Theorem 7.2.

Proof. The proof closely follows the proof of Theorem 3 from Li et al. (2024). Recall from Section B that Algorithm 2 is equivalent to SGD for minimizing $M^\gamma(x) := \frac{1}{n}\sum_{i=1}^n \nabla M_{f_i}^\gamma(x)$ with stepsize $\alpha\gamma$, and its updates can be written as

$$x_{k+1} = x_k - \alpha\gamma\frac{1}{S}\sum_{i\in\mathcal{S}_k}\nabla M_{f_i}^\gamma(x_k).$$

Then, for any $x_* \in \mathcal{X}_*$

$$\begin{aligned}
&\mathbb{E}_k\left[\|x_{k+1} - x_*\|^2\right] \\
&= \|x_k - x_*\|^2 - 2\alpha\gamma\left\langle x_k - x_*, \mathbb{E}_k\left[\frac{1}{S}\sum_{i\in\mathcal{S}_k}\nabla M_{f_i}^\gamma(x_k)\right]\right\rangle + \alpha^2\gamma^2\mathbb{E}_k\left[\left\|\frac{1}{S}\sum_{i\in\mathcal{S}_k}\nabla M_{f_i}^\gamma(x_k)\right\|^2\right] \\
&\overset{(A.8)}{=} \|x_k - x_*\|^2 - 2\alpha\gamma\left\langle x_k - x_*, \nabla M^\gamma(x_k) - \nabla M^\gamma(x_*)\right\rangle + \alpha^2\gamma^2\mathbb{E}_k\left[\left\|\frac{1}{S}\sum_{i\in\mathcal{S}_k}\nabla M_{f_i}^\gamma(x_k)\right\|^2\right] \\
&= \|x_k - x_*\|^2 - 2\alpha\gamma\left(D_{M^\gamma}(x_k, x_*) + D_{M^\gamma}(x_*, x_k)\right) + \alpha^2\gamma^2\mathbb{E}_k\left[\left\|\frac{1}{S}\sum_{i\in\mathcal{S}_k}\nabla M_{f_i}^\gamma(x_k)\right\|^2\right], \quad (38)
\end{aligned}$$

where $D_{M^\gamma}(x,y) := M^\gamma(x) - M^\gamma(y) - \langle\nabla M^\gamma(y), x - y\rangle$. Next, applying Lemma A.12 with $\boldsymbol{B}_i = \nabla M_{f_i}^\gamma(x_k) - \nabla M_{f_i}^\gamma(x_*) \in \mathbb{R}^{d\times 1}$, the last term in the above inequality can be written as

$$\begin{aligned}
\mathbb{E}_k\left[\left\|\frac{1}{S}\sum_{i\in\mathcal{S}_k}\nabla M_{f_i}^\gamma(x_k)\right\|^2\right] &\overset{(A.5)}{=} \mathbb{E}_k\left[\left\|\frac{1}{S}\sum_{i\in\mathcal{S}_k}\left(\nabla M_{f_i}^\gamma(x_k) - \nabla M_{f_i}^\gamma(x_*)\right)\right\|^2\right] \\
&\overset{(A.12)}{=} \frac{n-S}{S(n-1)}\frac{1}{n}\sum_{i=1}^n\left\|\nabla M_{f_i}^\gamma(x_k) - \nabla M_{f_i}^\gamma(x_*)\right\|^2 \\
&\quad + \frac{n(S-1)}{S(n-1)}\left\|\frac{1}{n}\sum_{i=1}^n\left(\nabla M_{f_i}^\gamma(x_k) - \nabla M_{f_i}^\gamma(x_*)\right)\right\|^2.
\end{aligned}$$

Looking at the first term of the inequality above, using convexity and smoothness of the functions $M_{f_i}^{\gamma}$ (Lemmas A.4 and A.6), we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla M_{f_i}^{\gamma}(x_k)-\nabla M_{f_i}^{\gamma}(x_*)\right\|^2 &\overset{(A.14)}{\leq} \frac{1}{n}\sum_{i=1}^{n}\frac{L_i}{1+\gamma L_i}\left(D_{M_{f_i}^{\gamma}}(x_k,x_*)+D_{M_{f_i}^{\gamma}}(x_*,x_k)\right)\\ &\leq \frac{L_{\max}}{1+\gamma L_{\max}}\frac{1}{n}\sum_{i=1}^{n}\left(D_{M_{f_i}^{\gamma}}(x_k,x_*)+D_{M_{f_i}^{\gamma}}(x_*,x_k)\right)\\ &= \frac{L_{\max}}{1+\gamma L_{\max}}\left(D_{M^{\gamma}}(x_k,x_*)+D_{M^{\gamma}}(x_*,x_k)\right).\end{aligned}$$

Next, since by Lemma A.7, the function M^{γ} is convex and smooth, the second term can be bounded as

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla M_{f_i}^{\gamma}(x_k)-\nabla M_{f_i}^{\gamma}(x_*)\right)\right\|^2 &= \left\|\nabla M^{\gamma}(x_k)-\nabla M^{\gamma}(x_*)\right\|^2\\ &\overset{(A.14)}{\leq} L_{\gamma}\left(D_{M^{\gamma}}(x_k,x_*)+D_{M^{\gamma}}(x_*,x_k)\right).\end{aligned}$$

Applying these bounds in (38) gives

$$\begin{aligned}&\mathbb{E}_k\left[\|x_{k+1}-x_*\|^2\right]\\ &\leq \|x_k-x_*\|^2-2\alpha\gamma\left(D_{M^{\gamma}}(x_k,x_*)+D_{M^{\gamma}}(x_*,x_k)\right)\\ &\quad+\alpha^2\gamma^2\underbrace{\left(\frac{n-S}{S(n-1)}\frac{L_{\max}}{1+\gamma L_{\max}}+\frac{n(S-1)}{S(n-1)}L_{\gamma}\right)}_{:=L_{\gamma,S}}\left(D_{M^{\gamma}}(x_k,x_*)+D_{M^{\gamma}}(x_*,x_k)\right)\\ &= \|x_k-x_*\|^2-\alpha\gamma\left(2-\alpha\gamma L_{\gamma,S}\right)\left(D_{M^{\gamma}}(x_k,x_*)+D_{M^{\gamma}}(x_*,x_k)\right).\end{aligned}$$

By Lemma A.7, M^{γ} is convex and smooth, so $D_{M^{\gamma}}(x_*,x_k)\geq 0$, and by Assumption 7.1 and Lemma A.13, we have

$$D_{M^{\gamma}}(x_k,x_*)=M^{\gamma}(x_k)-M^{\gamma}(x_*)\geq\frac{\mu_{\gamma}^{+}}{2}\|x_k-\Pi(x_k)\|^2.$$

Therefore, for $\alpha\gamma L_{\gamma,S}\leq 2$

$$\begin{aligned}\mathbb{E}_k\left[\|x_{k+1}-x_*\|^2\right] &\leq \|x_k-x_*\|^2-\alpha\gamma\left(2-\alpha\gamma L_{\gamma,S}\right)\left(D_{M^{\gamma}}(x_k,x_*)+D_{M^{\gamma}}(x_*,x_k)\right)\\ &\leq \|x_k-x_*\|^2-\alpha\gamma\left(2-\alpha\gamma L_{\gamma,S}\right)\frac{\mu_{\gamma}^{+}}{2}\|x_k-\Pi(x_k)\|^2.\end{aligned}$$

Taking expectation and letting $x_*=\Pi(x_k)$ gives

$$\begin{aligned}\mathbb{E}\left[\|x_{k+1}-\Pi(x_{k+1})\|^2\right] &\leq \mathbb{E}\left[\|x_{k+1}-\Pi(x_k)\|^2\right]\\ &\leq \mathbb{E}\left[\|x_k-\Pi(x_k)\|^2\right]-\alpha\gamma\left(2-\alpha\gamma L_{\gamma,S}\right)\frac{\mu_{\gamma}^{+}}{2}\mathbb{E}\left[\|x_k-\Pi(x_k)\|^2\right].\end{aligned}$$

Unrolling the recurrence, we obtain the first result

$$\mathbb{E}\left[\|x_K-\Pi(x_K)\|^2\right]\leq\left(1-\alpha\gamma\left(2-\alpha\gamma L_{\gamma,S}\right)\frac{\mu_{\gamma}^{+}}{2}\right)^K\mathbb{E}\left[\|x_0-\Pi(x_0)\|^2\right].$$

Lastly, using L–smoothness of f, it follows that

$$\begin{aligned}\mathbb{E}\left[f(x_K)-f(\Pi(x_K))\right] &\leq \frac{L}{2}\mathbb{E}\left[\|x_K-\Pi(x_K)\|^2\right]\\ &\leq \frac{L}{2}\left(1-\alpha\gamma\left(2-\alpha\gamma L_{\gamma,S}\right)\frac{\mu_{\gamma}^{+}}{2}\right)^K\mathbb{E}\left[\|x_0-\Pi(x_0)\|^2\right].\end{aligned}$$

Substituting $\alpha\gamma={}^{1}/_{L_{\gamma,S}}$, which minimizes the expression $\alpha\gamma\left(2-\alpha\gamma L_{\gamma,S}\right)$, finishes the proof. □

E FEDEXPROX WITH INEXACT COMPUTATIONS

In practice, solving (3) exactly is often infeasible, and we can only find a vector $\mathrm{prox}_{\gamma f}^{\delta}(x)$ such that

$$\left\|\mathrm{prox}_{\gamma f}^{\delta}(x) - \mathrm{prox}_{\gamma f}(x)\right\|^2 \leq \delta,$$

where δ represents the accuracy of the approximate solution to (3). As a result, we can only calculate an inexact gradient of the Moreau envelope, defined as

$$\nabla M_f^{\gamma,\delta}(x) := \frac{1}{\gamma}(x - \mathrm{prox}_{\gamma f}^{\delta}(x)).$$

One can easily show that

$$\left\|\nabla M_f^{\gamma,\delta}(x) - \nabla M_f^{\gamma}(x)\right\|^2 \leq \frac{\delta}{\gamma^2}.$$

With these inexact updates, Algorithm 1 iterates

$$x_{k+1} = x_k + \alpha_k \left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{prox}_{\gamma f_i}^{\delta}(x_k) - x_k\right) = x_k - \alpha_k \gamma \underbrace{\frac{1}{n}\sum_{i=1}^{n} \nabla M_{f_i}^{\gamma,\delta}(x_k)}_{:=\nabla M^{\gamma,\delta}(x_k)}, \tag{39}$$

where

$$\begin{aligned}
\left\|\nabla M^{\gamma,\delta}(x) - \nabla M^{\gamma}(x)\right\|^2 &= \left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla M_{f_i}^{\gamma,\delta}(x) - \nabla M_{f_i}^{\gamma}(x)\right)\right\|^2 \\
&= \frac{1}{\gamma^2}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{prox}_{\gamma f}^{\delta}(x) - \mathrm{prox}_{\gamma f}(x)\right)\right\|^2 \\
&\leq \frac{1}{\gamma^2}\frac{1}{n}\sum_{i=1}^{n}\left\|\mathrm{prox}_{\gamma f}^{\delta}(x) - \mathrm{prox}_{\gamma f}(x)\right\|^2 \leq \frac{\delta}{\gamma^2}.
\end{aligned} \tag{40}$$

With this in mind, we proceed to our analysis.

Theorem E.1. *Consider the inexact* FedExProx *method in the full participation setting defined in* (39)*. Let Assumptions 1.1, 1.2, 1.3, and 7.1 hold, and set* $\alpha\gamma = \frac{1}{2L_\gamma}$*. Then, the iterates of the algorithm satisfy*

$$\|x_K - \Pi(x_K)\|^2 \leq \left(1 - \frac{\mu_\gamma^+}{8L_\gamma}\right)^K \|x_0 - \Pi(x_0)\|^2 + \frac{20\delta}{\left(\mu_\gamma^+\right)^2\gamma^2},$$

and hence

$$f(x_K) - f(\Pi(x_K)) \leq \frac{L}{2}\left(1 - \frac{\mu_\gamma^+}{8L_\gamma}\right)^K \|x_0 - \Pi(x_0)\|^2 + \frac{10L\delta}{\left(\mu_\gamma^+\right)^2\gamma^2}. \tag{41}$$

Remark E.2. Note that the price one has to pay for inexactness is minimal. To illustrate this, suppose that clients solve the local problems using GD. Then, taking

$$\delta = \frac{\left(\mu_\gamma^+\right)^2\gamma^2\varepsilon}{20L},$$

we ensure that the last term in (41) is small:

$$\frac{10L\delta}{\left(\mu_\gamma^+\right)^2\gamma^2} \leq \frac{\varepsilon}{2}.$$

Since problem (3) is strongly convex, the local complexity is proportional to

$$\mathcal{O}\left(\kappa \times \log\frac{1}{\delta}\right) = \mathcal{O}\left(\kappa \times \log\left(\frac{20L}{\left(\mu_\gamma^+\right)^2\gamma^2\varepsilon}\right)\right),$$

where $\kappa \gg 1$ is the condition number. In practice, this logarithmic factor can be ignored and treated as a constant.

Proof of Theorem E.1. The updates can be written as

$$x_{k+1} = x_k - \alpha\gamma\nabla M^{\gamma,\delta}(x_k).$$

Thus

$$\begin{aligned}
\|x_{k+1} - \Pi(x_{k+1})\|^2 &\leq \|x_{k+1} - \Pi(x_k)\|^2 \\
&= \|x_k - \Pi(x_k)\|^2 - 2\alpha\gamma \left\langle x_k - \Pi(x_k), \nabla M^{\gamma,\delta}(x_k)\right\rangle + \alpha^2\gamma^2 \left\|\nabla M^{\gamma,\delta}(x_k)\right\|^2 \\
&= \|x_k - \Pi(x_k)\|^2 - 2\alpha\gamma \left\langle x_k - \Pi(x_k), \nabla M^{\gamma}(x_k)\right\rangle \\
&\quad - 2\alpha\gamma \left\langle x_k - \Pi(x_k), \nabla M^{\gamma,\delta}(x_k) - \nabla M^{\gamma}(x_k)\right\rangle \\
&\quad + \alpha^2\gamma^2 \left\|\nabla M^{\gamma}(x_k) + (\nabla M^{\gamma,\delta}(x_k) - \nabla M^{\gamma}(x_k))\right\|^2.
\end{aligned}$$

Using Jensen's inequality, we get

$$\begin{aligned}
\|x_{k+1} - \Pi(x_{k+1})\|^2 \leq \|x_k - \Pi(x_k)\|^2 &- 2\alpha\gamma \left\langle x_k - \Pi(x_k), \nabla M^{\gamma}(x_k)\right\rangle \\
&- 2\alpha\gamma \left\langle x_k - \Pi(x_k), \nabla M^{\gamma,\delta}(x_k) - \nabla M^{\gamma}(x_k)\right\rangle \\
&+ 2\alpha^2\gamma^2 \|\nabla M^{\gamma}(x_k)\|^2 + 2\alpha^2\gamma^2 \left\|\nabla M^{\gamma,\delta}(x_k) - \nabla M^{\gamma}(x_k)\right\|^2. \qquad (42)
\end{aligned}$$

Following the same steps as in the proof of Theorem D.1 gives

$$\begin{aligned}
&\|x_k - \Pi(x_k)\|^2 - 2\alpha\gamma \left\langle x_k - \Pi(x_k), \nabla M^{\gamma}(x_k)\right\rangle + 2\alpha^2\gamma^2 \|\nabla M^{\gamma}(x_k)\|^2 \\
&\quad \leq \left(1 - \alpha\gamma(1 - \alpha\gamma L_\gamma)\mu_\gamma^+\right) \|x_k - \Pi(x_k)\|^2 \\
&\quad \leq \left(1 - \alpha\gamma\frac{\mu_\gamma^+}{2}\right) \|x_k - \Pi(x_k)\|^2
\end{aligned}$$

for $\alpha\gamma \leq \frac{1}{2L_\gamma}$, and hence

$$\begin{aligned}
\|x_{k+1} - \Pi(x_{k+1})\|^2 \overset{(42)}{\leq} &\left(1 - \alpha\gamma\frac{\mu_\gamma^+}{2}\right) \|x_k - \Pi(x_k)\|^2 \\
&-2\alpha\gamma \left\langle x_k - \Pi(x_k), \nabla M^{\gamma,\delta}(x_k) - \nabla M^{\gamma}(x_k)\right\rangle \\
&+2\alpha^2\gamma^2 \left\|\nabla M^{\gamma,\delta}(x_k) - \nabla M^{\gamma}(x_k)\right\|^2.
\end{aligned}$$

Young's inequality yields

$$\begin{aligned}
\|x_{k+1} - \Pi(x_{k+1})\|^2 \leq &\left(1 - \alpha\gamma\frac{\mu_\gamma^+}{2}\right) \|x_k - \Pi(x_k)\|^2 + \alpha\gamma\lambda \|x_k - \Pi(x_k)\|^2 \\
&+ \frac{\alpha\gamma}{\lambda} \left\|\nabla M^{\gamma,\delta}(x_k) - \nabla M^{\gamma}(x_k)\right\|^2 + 2\alpha^2\gamma^2 \left\|\nabla M^{\gamma,\delta}(x_k) - \nabla M^{\gamma}(x_k)\right\|^2,
\end{aligned}$$

where $\lambda > 0$ is a free parameter. Taking $\lambda = \frac{\mu_\gamma^+}{4}$, we obtain

$$\begin{aligned}
\|x_{k+1} - \Pi(x_{k+1})\|^2 &\leq \left(1 - \alpha\gamma\frac{\mu_\gamma^+}{4}\right) \|x_k - \Pi(x_k)\|^2 + \frac{4\alpha\gamma}{\mu_\gamma^+} \left\|\nabla M^{\gamma,\delta}(x_k) - \nabla M^{\gamma}(x_k)\right\|^2 \\
&\quad + 2\alpha^2\gamma^2 \left\|\nabla M^{\gamma,\delta}(x_k) - \nabla M^{\gamma}(x_k)\right\|^2 \\
&\leq \left(1 - \alpha\gamma\frac{\mu_\gamma^+}{4}\right) \|x_k - \Pi(x_k)\|^2 + \frac{5\alpha\gamma}{\mu_\gamma^+} \left\|\nabla M^{\gamma,\delta}(x_k) - \nabla M^{\gamma}(x_k)\right\|^2,
\end{aligned}$$

where the last inequality due to $\alpha\gamma \leq \frac{1}{2L_\gamma}$ and $L_\gamma \geq \mu_\gamma^+$. Lastly, due to (40), we have

$$\|x_{k+1} - \Pi(x_{k+1})\|^2 \leq \left(1 - \alpha\gamma\frac{\mu_\gamma^+}{4}\right) \|x_k - \Pi(x_k)\|^2 + \frac{5\alpha\delta}{\mu_\gamma^+\gamma}.$$

Unrolling the recursion, one can show that

$$\|x_K - \Pi(x_K)\|^2 \leq \left(1 - \alpha\gamma\frac{\mu_\gamma^+}{4}\right)^K \|x_0 - \Pi(x_0)\|^2 + \frac{20\delta}{\left(\mu_\gamma^+\right)^2\gamma^2},$$

and hence

$$
\begin{aligned}
f(x_K) - f(\Pi(x_K)) &\leq \frac{L}{2} \left\| x_K - \Pi(x_K) \right\|^2 \\
&\leq \frac{L}{2} \left(1 - \alpha\gamma \frac{\mu_\gamma^+}{4} \right)^K \left\| x_0 - \Pi(x_0) \right\|^2 + \frac{10L\delta}{\left(\mu_\gamma^+\right)^2 \gamma^2}.
\end{aligned}
$$

□

F ADAPTIVE EXTRAPOLATION

The optimal extrapolation parameter values in Theorems 4.1, 5.1, 7.2 and 7.5 depend on L_γ (or $L_{\gamma,S}$). This dependency can be avoided by employing *adaptive* extrapolation strategies (Horváth et al., 2022). Specifically, we analyze methods based on gradient diversity (FedExProx-GraDS) and the Polyak stepsize (FedExProx-StoPS).

F.1 FULL PARTICIPATION CASE

Theorem 6.1. *Fix any $\gamma > 0$ and consider solving non-strongly convex quadratic optimization problem (1), where $f_i(x) = \frac{1}{2}x^\top \boldsymbol{A}_i x - b_i^\top x$ for all $i \in [n]$, with $\boldsymbol{A}_i \in \mathrm{Sym}_+^d$ and $b_i \in \mathbb{R}^d$. Let Assumption 1.3 hold and consider two adaptive extrapolation strategies:*

1. *(FedExProx-GraDS) Set*

$$\alpha_k = \alpha_k^{\mathsf{GraDS}}(x_k) := \frac{\frac{1}{n}\sum_{i=1}^n \left\|\nabla M_{f_i}^\gamma(x_k)\right\|^2}{\left\|\frac{1}{n}\sum_{i=1}^n \nabla M_{f_i}^\gamma(x_k)\right\|^2} \geq 1.$$

Then, the iterates of Algorithm 1 satisfy

$$\|x_K - \Pi(x_K)\|^2 \leq (1 - C_{\mathsf{G}})^K \|x_0 - \Pi(x_0)\|^2, \tag{15}$$

where $C_{\mathsf{G}} := \min_{k=0,\ldots,K-1} \alpha_k \gamma \frac{2+\gamma L_{\max}}{1+\gamma L_{\max}} \mu_\gamma^+$.

2. *(FedExProx-StoPS) Set*

$$\alpha_k = \alpha_k^{\mathsf{StoPS}}(x_k) := \frac{\frac{1}{n}\sum_{i=1}^n \left(M_{f_i}^\gamma(x_k) - \inf M_{f_i}^\gamma\right)}{\gamma\left\|\frac{1}{n}\sum_{i=1}^n \nabla M_{f_i}^\gamma(x_k)\right\|^2} \geq \frac{1}{2\gamma L_\gamma}.$$

Then, the iterates of Algorithm 1 satisfy

$$\|x_K - \Pi(x_K)\|^2 \leq (1 - C_{\mathsf{S}})^K \|x_0 - \Pi(x_0)\|^2, \tag{16}$$

where $C_{\mathsf{S}} := \frac{3}{2} \min_{k=0,\ldots,K-1} \alpha_k \gamma \mu_\gamma^+$.

Proof. We start with the standard decomposition

$$\begin{aligned}
\|x_{k+1} - \Pi(x_{k+1})\|^2 &\leq \|x_{k+1} - \Pi(x_k)\|^2 \\
&= \|x_k - \Pi(x_k)\|^2 - 2\alpha_k\gamma \left\langle \frac{1}{n}\sum_{i=1}^n \nabla M_{f_i}^\gamma(x_k), x_k - \Pi(x_k) \right\rangle \\
&\quad + \alpha_k^2\gamma^2 \left\| \frac{1}{n}\sum_{i=1}^n \nabla M_{f_i}^\gamma(x_k) \right\|^2 \\
&\overset{(21)}{=} \|x_k - \Pi(x_k)\|^2 - 2\alpha_k\gamma (x_k - \Pi(x_k))^\top \boldsymbol{M}(x_k - \Pi(x_k)) \\
&\quad + \alpha_k^2\gamma^2 \left\| \frac{1}{n}\sum_{i=1}^n \nabla M_{f_i}^\gamma(x_k) \right\|^2.
\end{aligned} \tag{43}$$

Now, let us consider the two adaptive extrapolation strategies.

1. (FedExProx-GraDS) In this case, the last term of (43) can be rewritten as

$$\begin{aligned}
&\alpha_k^2\gamma^2 \left\| \frac{1}{n}\sum_{i=1}^n \nabla M_{f_i}^\gamma(x_k) \right\|^2 \\
&= \alpha_k\gamma^2 \frac{1}{n}\sum_{i=1}^n \left\| \nabla M_{f_i}^\gamma(x_k) \right\|^2
\end{aligned}$$

$$
\begin{aligned}
&\overset{(21)}{=} \alpha_k \gamma^2 \frac{1}{n} \sum_{i=1}^{n} \left\| \frac{1}{\gamma} \left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1} \right) (x_k - \Pi(x_k)) \right\|^2 \\
&= \alpha_k (x_k - \Pi(x_k))^\top \left(\frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1} \right)^\top \left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1} \right) \right) (x_k - \Pi(x_k)) \\
&\leq \alpha_k \max_{i \in [n]} \lambda_{\max} \left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1} \right) (x_k - \Pi(x_k))^\top \left(\frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1} \right) \right) (x_k - \Pi(x_k)) \\
&= \alpha_k \gamma \frac{\gamma \max_{i \in [n]} \lambda_{\max}(\boldsymbol{A}_i)}{1 + \gamma \max_{i \in [n]} \lambda_{\max}(\boldsymbol{A}_i)} (x_k - \Pi(x_k))^\top \boldsymbol{M} (x_k - \Pi(x_k)).
\end{aligned}
$$

Substituting this in (43) gives

$$
\begin{aligned}
&\|x_{k+1} - \Pi(x_{k+1})\|^2 \\
&\leq \|x_k - \Pi(x_k)\|^2 - 2\alpha_k \gamma (x_k - \Pi(x_k))^\top \boldsymbol{M} (x_k - \Pi(x_k)) \\
&\quad + \alpha_k \gamma \frac{\gamma \max_{i \in [n]} \lambda_{\max}(\boldsymbol{A}_i)}{1 + \gamma \max_{i \in [n]} \lambda_{\max}(\boldsymbol{A}_i)} (x_k - \Pi(x_k))^\top \boldsymbol{M} (x_k - \Pi(x_k)) \\
&= \|x_k - \Pi(x_k)\|^2 - \alpha_k \gamma \frac{2 + \gamma \max_{i \in [n]} \lambda_{\max}(\boldsymbol{A}_i)}{1 + \gamma \max_{i \in [n]} \lambda_{\max}(\boldsymbol{A}_i)} (x_k - \Pi(x_k))^\top \boldsymbol{M} (x_k - \Pi(x_k)) \\
&\leq \left(1 - \alpha_k \gamma \frac{2 + \gamma \max_{i \in [n]} \lambda_{\max}(\boldsymbol{A}_i)}{1 + \gamma \max_{i \in [n]} \lambda_{\max}(\boldsymbol{A}_i)} \lambda_{\min}^+(\gamma) \right) \|x_k - \Pi(x_k)\|^2,
\end{aligned}
$$

where the last inequality follows from the fact that $x_k - \Pi(x_k) \in \operatorname{range}(\boldsymbol{M})$ (see Lemma C.6). Applying this bound iteratively for $k = 0, \ldots, K-1$ gives

$$
\begin{aligned}
&\|x_K - \Pi(x_K)\|^2 \\
&\quad \leq \left(1 - \gamma \min_{k=0,\ldots,K-1} \alpha_k \frac{2 + \gamma \max_{i \in [n]} \lambda_{\max}(\boldsymbol{A}_i)}{1 + \gamma \max_{i \in [n]} \lambda_{\max}(\boldsymbol{A}_i)} \lambda_{\min}^+(\gamma) \right)^K \|x_0 - \Pi(x_0)\|^2.
\end{aligned}
$$

2. (FedExProx-StoPS) First, note that L_γ–smoothness of M^γ implies that

$$
M^\gamma(x_k) - \inf M^\gamma \geq \frac{1}{2L_\gamma} \|\nabla M^\gamma(x_k)\|^2,
$$

and hence $\alpha_{k,S}(x_k) \geq \frac{1}{2\gamma L_\gamma}$.

Now, the last term of (43) is

$$
\begin{aligned}
\alpha_k^2 \gamma^2 \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla M_{f_i}^\gamma(x_k) \right\|^2 &= \alpha_k \gamma \frac{1}{n} \sum_{i=1}^{n} \left(M_{f_i}^\gamma(x_k) - \inf M_{f_i}^\gamma \right) \\
&= \alpha_k \gamma \left(M^\gamma(x_k) - M^\gamma(\Pi(x_k)) \right) \\
&\overset{(24)}{=} \frac{1}{2} \alpha_k \gamma (x_k - \Pi(x_k))^T \boldsymbol{M} (x_k - \Pi(x_k)).
\end{aligned}
$$

Therefore, using Lemma C.6, we get

$$
\begin{aligned}
\|x_{k+1} - \Pi(x_{k+1})\|^2 &\overset{(43)}{\leq} \|x_k - \Pi(x_k)\|^2 - 2\alpha_k \gamma (x_k - \Pi(x_k))^\top \boldsymbol{M} (x_k - \Pi(x_k)) \\
&\quad + \frac{1}{2} \alpha_k \gamma (x - \Pi(x_k))^T \boldsymbol{M} (x - \Pi(x_k)) \\
&\leq \left(1 - \frac{3}{2} \alpha_k \gamma \lambda_{\min}^+(\gamma) \right) \|x_k - \Pi(x_k)\|^2.
\end{aligned}
$$

Applying this bound iteratively for $k = 0, \ldots, K-1$ gives

$$
\|x_K - \Pi(x_K)\|^2 \leq \left(1 - \frac{3}{2} \min_{k=0,\ldots,K-1} \alpha_k \gamma \lambda_{\min}^+(\gamma) \right)^K \|x_0 - \Pi(x_0)\|^2.
$$

□

F.2 Partial Participation Case

Theorem F.1. *Fix any $\gamma > 0$ and consider solving non-strongly convex quadratic optimization problem (1), where $f_i(x) = \frac{1}{2}x^\top \boldsymbol{A}_i x - b_i^\top x$ for all $i \in [n]$, with $\boldsymbol{A}_i \in \mathrm{Sym}_+^d$ and $b_i \in \mathbb{R}^d$. Let Assumption 1.3 hold and consider two adaptive extrapolation strategies:*

1. *(FedExProx-GraDS) Set*

$$\alpha_k = \alpha_{k,S}^{\mathsf{GraDS}}(x_k, \mathcal{S}_k) := \frac{\frac{1}{S}\sum_{i\in\mathcal{S}^k}\left\|\nabla M_{f_i}^{\gamma}(x_k)\right\|^2}{\left\|\frac{1}{S}\sum_{i\in\mathcal{S}^k}\nabla M_{f_i}^{\gamma}(x_k)\right\|^2} \geq 1.$$

Then, the iterates of Algorithm 2 satisfy

$$\mathbb{E}\left[\|x_K - \Pi(x_K)\|^2\right] \leq \left(1 - \inf \alpha_k \gamma \frac{2+\gamma L_{\max}}{1+\gamma L_{\max}} \lambda_{\min}^+(\gamma)\right)^K \|x_0 - \Pi(x_0)\|^2, \quad (44)$$

where

$$\inf \alpha_k := \inf_{x\in\mathbb{R}^d, \mathcal{S}\subseteq[n], |\mathcal{S}|=S} \alpha_{k,S}^{\mathsf{GraDS}}(x, \mathcal{S}).$$

2. *(FedExProx-StoPS) Set*

$$\alpha_k = \alpha_{k,S}^{\mathsf{StoPS}}(x_k, \mathcal{S}_k) := \frac{\frac{1}{S}\sum_{i\in\mathcal{S}^k}\left(M_{f_i}^{\gamma}(x_k) - \inf M_{f_i}^{\gamma}\right)}{\gamma\left\|\frac{1}{n}\sum_{i\in\mathcal{S}^k}\nabla M_{f_i}^{\gamma}(x_k)\right\|^2}.$$

Then, the iterates of Algorithm 2 satisfy

$$\mathbb{E}\left[\|x_K - \Pi(x_K)\|^2\right] \leq \left(1 - \frac{3}{2}\inf \alpha_k \gamma \lambda_{\min}^+(\gamma)\right)^K \|x_0 - \Pi(x_0)\|^2, \quad (45)$$

where

$$\inf \alpha_k := \inf_{x\in\mathbb{R}^d, \mathcal{S}\subseteq[n], |\mathcal{S}|=S} \alpha_{k,S}^{\mathsf{StoPS}}(x, \mathcal{S}).$$

Remark F.2. In the single node setting ($S = 1$),

$$\inf_{x\in\mathbb{R}^d, \mathcal{S}\subseteq[n], |\mathcal{S}|=1} \alpha_{k,1}^{\mathsf{GraDS}}(x, \mathcal{S}) = 1.$$

However, as more clients participate, the extrapolation parameter may exceed 1, resulting in improved performance of FedExProx-GraDS.

Remark F.3. In the single node setting ($S = 1$), Lemma A.4 implies that

$$\inf_{x\in\mathbb{R}^d, \mathcal{S}\subseteq[n], |\mathcal{S}|=1} \alpha_{k,1}^{\mathsf{StoPS}}(x, \mathcal{S}) = \frac{1}{2}\frac{1+\gamma L_{\max}}{\gamma L_{\max}}.$$

As the number of participating clients increases, the extrapolation parameter may exceed this bound, leading to better performance of FedExProx-StoPS.

Proof of Theorem F.1. As in the full batch case, we start with the decomposition

$$\begin{aligned}
\|x_{k+1} - \Pi(x_{k+1})\|^2 &\leq \|x_{k+1} - \Pi(x_k)\|^2 \\
&= \|x_k - \Pi(x_k)\|^2 - 2\alpha_k\gamma\left\langle \frac{1}{S}\sum_{i\in\mathcal{S}^k}\nabla M_{f_i}^{\gamma}(x_k), x_k - \Pi(x_k)\right\rangle \\
&\quad + \alpha_k^2\gamma^2\left\|\frac{1}{S}\sum_{i\in\mathcal{S}^k}\nabla M_{f_i}^{\gamma}(x_k)\right\|^2
\end{aligned}$$

$$\overset{(21)}{=} \|x_k - \Pi(x_k)\|^2 - 2\alpha_k\gamma(x_k - \Pi(x_k))^\top \boldsymbol{M}_k(x_k - \Pi(x_k)) + \alpha_k^2\gamma^2 \left\| \frac{1}{S}\sum_{i\in\mathcal{S}^k} \nabla M_{f_i}^\gamma(x_k) \right\|^2, \tag{46}$$

where $\boldsymbol{M}_k := \frac{1}{\gamma}\left(\boldsymbol{I} - \frac{1}{S}\sum_{i\in\mathcal{S}_k}(\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)$. Now, let us consider the two adaptive extrapolation strategies.

1. (FedExProx-GraDS) In this case, the last term of (46) can be rewritten as

$$\begin{aligned}
&\alpha_k^2\gamma^2 \left\| \frac{1}{S}\sum_{i\in\mathcal{S}^k} \nabla M_{f_i}^\gamma(x_k) \right\|^2 \\
&= \alpha_k\gamma^2 \frac{1}{S}\sum_{i\in\mathcal{S}^k} \left\| \nabla M_{f_i}^\gamma(x_k) \right\|^2 \\
&\overset{(21)}{=} \alpha_k\gamma^2 \frac{1}{S}\sum_{i\in\mathcal{S}^k} \left\| \frac{1}{\gamma}\left(\boldsymbol{I} - (\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)(x_k - \Pi(x_k)) \right\|^2 \\
&= \alpha_k (x_k - \Pi(x_k))^\top \left(\frac{1}{S}\sum_{i\in\mathcal{S}^k} \left(\boldsymbol{I} - (\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)^\top \left(\boldsymbol{I} - (\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1}\right) \right)(x_k - \Pi(x_k)) \\
&\leq \alpha_k \max_{i\in[n]} \lambda_{\max}\left(\boldsymbol{I} - (\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1}\right)(x_k - \Pi(x_k))^\top \left(\frac{1}{S}\sum_{i\in\mathcal{S}^k} \left(\boldsymbol{I} - (\gamma\boldsymbol{A}_i + \boldsymbol{I})^{-1}\right) \right)(x_k - \Pi(x_k)) \\
&= \alpha_k\gamma \frac{\gamma \max_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)}{1 + \gamma \max_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)} (x_k - \Pi(x_k))^\top \boldsymbol{M}_k (x_k - \Pi(x_k)).
\end{aligned}$$

Substituting this in (46) gives

$$\begin{aligned}
&\|x_{k+1} - \Pi(x_{k+1})\|^2 \\
&\leq \|x_k - \Pi(x_k)\|^2 - 2\alpha_k\gamma(x_k - \Pi(x_k))^\top \boldsymbol{M}_k(x_k - \Pi(x_k)) \\
&\quad + \alpha_k\gamma \frac{\gamma \max_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)}{1 + \gamma \max_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)} (x_k - \Pi(x_k))^\top \boldsymbol{M}_k (x_k - \Pi(x_k)) \\
&= \|x_k - \Pi(x_k)\|^2 - \alpha_k\gamma \frac{2 + \gamma \max_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)}{1 + \gamma \max_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)} (x_k - \Pi(x_k))^\top \boldsymbol{M}_k (x_k - \Pi(x_k)) \\
&\leq \|x_k - \Pi(x_k)\|^2 - \inf \alpha_k\gamma \frac{2 + \gamma \max_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)}{1 + \gamma \max_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)} (x_k - \Pi(x_k))^\top \boldsymbol{M}_k (x_k - \Pi(x_k)),
\end{aligned}$$

where we view α_k as a function of x an $\mathcal{S}$. Taking expectation conditioned on x_k, we obtain

$$\begin{aligned}
&\mathbb{E}_k\left[\|x_{k+1} - \Pi(x_{k+1})\|^2\right] \\
&\leq \|x_k - \Pi(x_k)\|^2 \\
&\quad - \inf \alpha_k\gamma \frac{2 + \gamma \max_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)}{1 + \gamma \max_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)} (x_k - \Pi(x_k))^\top \mathbb{E}_k\left[\boldsymbol{M}_k\right](x_k - \Pi(x_k)) \\
&= \|x_k - \Pi(x_k)\|^2 - \inf \alpha_k\gamma \frac{2 + \gamma \max_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)}{1 + \gamma \max_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)} (x_k - \Pi(x_k))^\top \boldsymbol{M} (x_k - \Pi(x_k)) \\
&\leq \left(1 - \inf \alpha_k\gamma \frac{2 + \gamma \max_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)}{1 + \gamma \max_{i\in[n]} \lambda_{\max}(\boldsymbol{A}_i)} \lambda_{\min}^+(\gamma)\right) \|x_k - \Pi(x_k)\|^2,
\end{aligned}$$

where the last inequality follows from the fact that $x_k - \Pi(x_k) \in \operatorname{range}(\boldsymbol{M})$ (see Lemma C.6). Taking full expectation and applying this bound iteratively for $k = 0, \ldots, K-1$ gives the final result.

2. (FedExProx-StoPS) In this case, the last term of (46) is

$$
\begin{aligned}
&\alpha_k^2\gamma^2 \left\| \frac{1}{S}\sum_{i\in\mathcal{S}^k} \nabla M_{f_i}^{\gamma}(x_k) \right\|^2 \\
&\quad = \alpha_k\gamma \frac{1}{S}\sum_{i\in\mathcal{S}^k} \left(M_{f_i}^{\gamma}(x_k) - \inf M_{f_i}^{\gamma} \right) \\
&\quad \overset{(22)}{=} \frac{1}{2}\alpha_k \frac{1}{S}\sum_{i\in\mathcal{S}^k} (x_k - \Pi(x_k))^T (\boldsymbol{I} - (\gamma \boldsymbol{A}_i + \boldsymbol{I})^{-1})(x_k - \Pi(x_k)) \\
&\quad = \frac{1}{2}\alpha_k\gamma (x_k - \Pi(x_k))^T \boldsymbol{M}_k (x_k - \Pi(x_k)).
\end{aligned}
$$

Therefore, (46) gives

$$
\begin{aligned}
\|x_{k+1} - \Pi(x_{k+1})\|^2 &\leq \|x_k - \Pi(x_k)\|^2 - 2\alpha_k\gamma (x_k - \Pi(x_k))^\top \boldsymbol{M}_k (x_k - \Pi(x_k)) \\
&\quad + \frac{1}{2}\alpha_k\gamma (x_k - \Pi(x_k))^T \boldsymbol{M}_k (x_k - \Pi(x_k)) \\
&\leq \|x_k - \Pi(x_k)\|^2 - \frac{3}{2}\inf \alpha_k\gamma (x_k - \Pi(x_k))^T \boldsymbol{M}_k (x_k - \Pi(x_k)).
\end{aligned}
$$

Taking full expectation and applying Lemma C.6, we obtain

$$
\mathbb{E}\left[\|x_{k+1} - \Pi(x_{k+1})\|^2 \right] \leq \left(1 - \frac{3}{2}\inf \alpha_k\gamma\lambda_{\min}^{+}(\gamma) \right) \|x_k - \Pi(x_k)\|^2 ,
$$

and hence

$$
\mathbb{E}\left[\|x_K - \Pi(x_K)\|^2 \right] \leq \left(1 - \frac{3}{2}\inf \alpha_k\gamma\lambda_{\min}^{+}(\gamma) \right)^K \|x_0 - \Pi(x_0)\|^2 .
$$

□

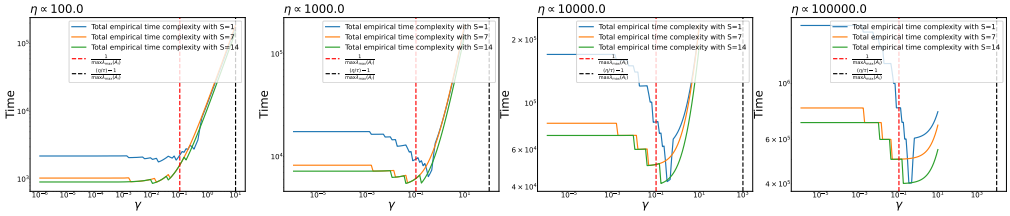

Figure 2: Empirical time complexities of FedExProx with partial client participation on a quadratic optimization task for $S \in \{1, 7, 14\}$ clients participating in each round.

G Experiments

G.1 Experimental Setup

To verify our theoretical results, we assess the empirical time complexity of FedExProx as a function of γ. We consider two types of optimization problems: synthetic quadratic optimization tasks and classification problems with a smooth hinge loss function. In both cases, local updates are computed using GD. Below, we first outline the methodology common to both experiments, followed by a detailed description of the specific settings and their corresponding results.

Total empirical time complexity. As described in Section 4.1, we assume that one local iteration of GD takes τ seconds. Thus, the time needed by worker i to find $\mathrm{prox}_{\gamma f_i}(x_k)$ at global iteration k is proportional to $\tau \times n_{ik}$, where n_{ik} is the number of GD iterations needed to find $\mathrm{prox}_{\gamma f_i}(x_k)$ to a given fixed accuracy. In the full participation case, the total empirical time complexity is given by

$$\sum_{k=0}^{K-1} \left(\eta + \tau \max_{i \in [n]} n_{ki} \right) \tag{47}$$

where K is the number of global iterations needed for FedExProx to converge to the desired accuracy and η is the communication cost. In experiments involving client sampling, we consider the complexity

$$\sum_{k=0}^{K-1} \left(\eta + \tau \max_{i \in \mathcal{S}_k} n_{ki} \right) \tag{48}$$

instead, where $\mathcal{S}_k$ is the random subset of clients sampled at global iteration k. Notice that we take the maximum over $n_{\cdot i}$ because the overall time is determined by the slowest client. In all experiments, we fix $\tau = 1$ and vary η, effectively considering the ratio η/τ.

To ensure robust experiments, we count the number of GD steps and communications K and apply formula (47) (or (48)). This proxy for real wall-clock time is robust, reliable, reflects reality, and allows controlled experiments. The plots report the empirical time complexity required to find x^k such that $f(x^k) - \inf_{x \in \mathbb{R}^d} f(x) \leq 10^{-3}$.

One of experimental tasks is to verify that this empirical complexity aligns with the theoretical complexity derived in (14) (see Section G.3.1).

G.2 Experiments with Quadratic Optimization Problems

In the first set of experiments, we consider quadratic optimization problems of the form

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} x^\top \boldsymbol{A}_i x,$$

where $\boldsymbol{A}_i \in \mathrm{Sym}_+^7$, $i \in [n]$ are random positive semidefinite matrices with the smallest eigenvalue equal to zero. In particular, each worker's matrix is generated as $\boldsymbol{A}_i = \boldsymbol{Q}_i \boldsymbol{D} \boldsymbol{Q}_i^T$, where $\boldsymbol{Q}_i$ is a random orthogonal matrix from QR decomposition of a $d \times d$ matrix with standard normal entries. The diagonal matrix $\boldsymbol{D}$ is constructed by uniformly sampling eigenvalues between 0 and a maximum

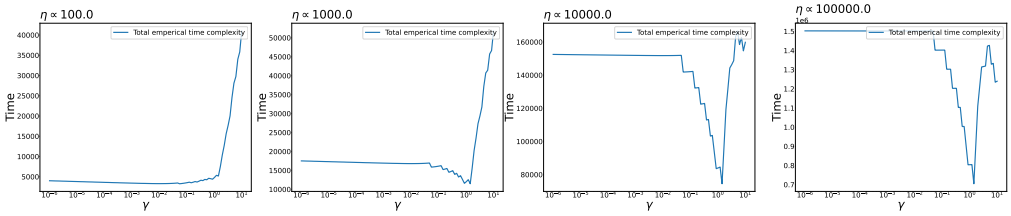

Figure 3: Empirical time complexities of FedExProx on a synthetic classification task with smooth hinge loss.

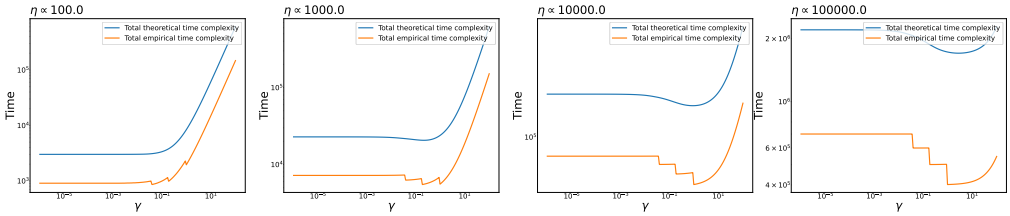

Figure 4: Comparison of theoretical time complexity (12) and empirical time complexity (47).

drawn from the interval $[5, 10]$. Each worker i computes the proximal mappings using GD with a stepsize of $1/\bar{L}_i$, where $\bar{L}_i = \lambda_{\max}(\boldsymbol{A}_i) + 1/\gamma$ is the smoothness constant of the local subproblem (3). The extrapolation parameter α is set to its optimal value from Theorem 4.1, using an explicitly computed value of L_γ. The data is distributed across $n = 14$ workers.

Figures 1 and 2 display the results for the full and partial participation settings, respectively. The dashed lines indicate the theoretical bounds derived in Theorems 4.3 and 5.2, within which the optimal value of γ is expected to lie. When the communication cost η is relatively low ($\eta \propto 100$), the optimal γ is close to zero. However, as η increases, the optimal γ shifts to values exceeding 0.1. This results in a characteristic U-shaped curve, reflecting the nontrivial optimal choice of γ. These observations are fully consistent with our theoretical predictions.

G.3 Experiments with Smooth Hinge Loss

G.3.1 Synthetic Data

We now turn to a classification problem involving linearly separable data $\{(x_{ij}, y_{ij})\}_{i\in[n], j\in[m]}$, ensuring the interpolation regime. Specifically, we construct a synthetic binary classification problem by independently sampling $n \times m$ data points $x_{ij} \in \mathbb{R}^d$ and a ground-truth weight vector $w^* \in \mathbb{R}^d$ from the standard normal distribution. Labels are then assigned as $y_{ij} = \text{sign}(x_{ij}^T w^*)$.

The function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w) := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\ell_{ij}(w),$$

where $(x_{ij}, y_{ij}) \in \mathbb{R}^d \times \{-1, 1\}$ for all $i \in [n], j \in [m]$ and ℓ_{ij} is the smooth hinge loss, defined by

$$\ell_{ij}(w) = \begin{cases} 0, & \text{if } y_{ij} w^T x_{ij} \geq 1, \\ \frac{1}{2}(1 - y_{ij} w^T x_{ij})^2, & \text{if } 0 < y_{ij} w^T x_{ij} < 1, \\ 1 - y_{ij} w^T x_{ij}, & \text{if } y_{ij} w^T x_{ij} \leq 0. \end{cases}$$

Each client $i \in [n]$ computes the local proximal mappings using GD with stepsize $1/\bar{L}_i$, where $\bar{L}_i = \frac{1}{m}\sum_{j=1}^{m}\|x_{ij}\|^2$ is an upper bound on the smoothness constant of the function f_i. The extrapolation parameter α is determined via grid search within the theoretical range given by Lemma A.7. We use $n = 4$ workers, each holding $m = 4$ datapoints, and the number of parameters is $d = 3$.

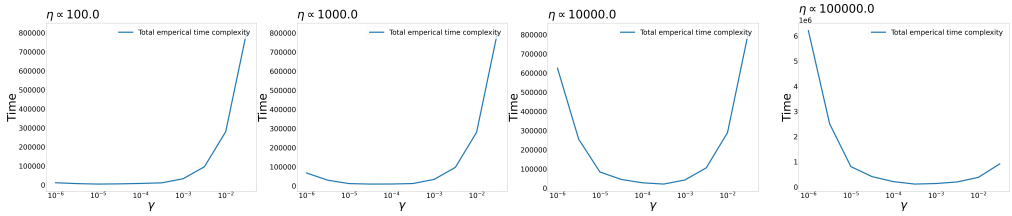

Figure 5: Empirical time complexities of FedExProx on the ARCENE classification task with smooth hinge loss.

The results are presented in Figure 3. As in Section G.2, these experiments again suggest a non-zero optimal choice of γ when communication is slow. Indeed, the values of γ minimizing the U-shapes in the last three plots are around 1. These experiments provide further support for our theory.

Comparing theoretical and empirical time complexities. Next, we compare the theoretical and empirical time complexities (12) and (47). Both are plotted on the same scale in Figure 4. We see that our theoretical model reflects the empirical dependencies (up to a multiplicative factor).

G.3.2 Real Data

We further validate our findings on real-world data by repeating the experiment on the ARCENE dataset (Guyon et al., 2004). We subsample 20 random data points and restrict attention to 30 randomly chosen features. The resulting plot in Figure 5 again exhibits the characteristic U-shape behavior as the communication cost increases, mirroring the synthetic case and reinforcing the robustness of our observations

H NOTATION

Notation	
$\|\cdot\|$	Standard Euclidean norm
$\langle\cdot,\cdot\rangle$	Standard Euclidean inner product
$[k]$	$:= \{1,\ldots,k\}$
$\mathbb{E}_k[\cdot]$	Expectation conditioned on the first k iterations
n	Number of workers/nodes/clients/devices
d	Dimensionality of the problem
γ	Stepsize
α_k	Extrapolation parameter
L, L_i, $L_{\max}$	Smoothness constants (Assumption 1.2)
$\mathrm{prox}_{\gamma f}(x)$	Proximal operator of the function f with parameter γ (see (3))
M_f^γ	Moreau envelope of the function f with parameter γ (see (7))
M^γ	$:= \frac{1}{n}\sum_{i=1}^n M_{f_i}^\gamma$
L_γ	Smoothness constant of M^γ
μ_γ^+	The smallest non-zero eigenvalue of the matrix $\nabla^2 M^\gamma$ (see Theorems 4.1, 5.1) / PŁ constant (see Theorems 7.2, 7.5)
$\pi(\gamma)$	Total time per global iteration of Algorithm 1 / 2
$T_\eta(\gamma)$	Time complexity of Algorithm 1
$T_\eta(\gamma,S)$	Time complexity of Algorithm 2
$\mathcal{X}_*$	$:= \{x\in\mathbb{R}^d : \nabla f_i(x)=0\}$
$\Pi(x)$	Projection of x onto the solution set $\mathcal{X}_*$
R^2	$:= \|x_*-x_0\|^2$
$D_f(x,y)$	Bregman divergence between x and y associated with a function $f:\mathbb{R}^d\to\mathbb{R}$
$\boldsymbol{I}$	Identity matrix
$\boldsymbol{A}_i$	Local data matrix stored on worker i, with spectral decomposition $\boldsymbol{A}_i=\boldsymbol{Q}_i\boldsymbol{\Lambda}_i\boldsymbol{Q}_i^\top$
$\boldsymbol{A}$	$:= \frac{1}{n}\sum_{i=1}^n \boldsymbol{A}_i$
Sym_+^d	$:= \{\boldsymbol{X}\in\mathbb{R}^{d\times d} \mid \boldsymbol{X}=\boldsymbol{X}^\top, \boldsymbol{X}\succeq 0\}$ - the set of symmetric positive semidefinite matrices
$\ker \boldsymbol{A}$	Kernel of a matrix $\boldsymbol{A}$
$\boldsymbol{M}_k$	$:= \frac{1}{\gamma}\left(\boldsymbol{I}-\frac{1}{S}\sum_{i\in\mathcal{S}_k}(\gamma\boldsymbol{A}_i+\boldsymbol{I})^{-1}\right)$
$\boldsymbol{M}$	$:= \frac{1}{\gamma}\left(\boldsymbol{I}-\frac{1}{n}\sum_{i=1}^n(\gamma\boldsymbol{A}_i+\boldsymbol{I})^{-1}\right)$
$\lambda_{\max}(\boldsymbol{B})$	The largest eigenvalue of matrix $\boldsymbol{B}$
$\lambda_{\min}(\boldsymbol{B})$	The smallest eigenvalue of matrix $\boldsymbol{B}$
$\lambda_{\min}^+(\boldsymbol{B})$	The smallest positive eigenvalue of matrix $\boldsymbol{B}$
$[\boldsymbol{B}]_j$	The jth diagonal element of a diagonal matrix $\boldsymbol{B}$
$[b_j]_{jj}$	A diagonal matrix with b_j as the jth entry

Table 1: Frequently used notation.

Note on LLM Usage. Large Language Models were used to assist in polishing the writing of the manuscript. LLM assistance did not contribute to the scientific content of the paper.