On the Convergence of FedProx with Extrapolation and Inexact Prox

Hanmin Li HANMIN.LI@KAUST.EDU.SA
Peter Richtárik PETER.RICHTARIK@KAUST.EDU.SA
GenAI CoE, KAUST

Abstract

Enhancing the FedProx federated learning algorithm [37] with server-side extrapolation, Li et al. [35] recently introduced the FedExProx method. Their theoretical analysis, however, relies on the assumption that each client computes a certain proximal operator exactly, which is impractical since this is virtually never possible to do in real settings. In this paper, we investigate the behavior of FedExProx without this exactness assumption in the smooth and globally strongly convex setting. We establish a general convergence result, showing that inexactness leads to convergence to a neighborhood of the solution. Additionally, we demonstrate that, with careful control, the adverse effects of this inexactness can be mitigated. By linking inexactness to biased compression [9], we refine our analysis, highlighting robustness of extrapolation to inexact proximal updates. We also examine the local iteration complexity required by each client to achieved the required level of inexactness using various local optimizers. Our theoretical insights are validated through comprehensive numerical experiments.

1. Introduction

Federated learning (FL) is a decentralized approach where clients collaboratively train a shared model locally, preserving privacy [33, 43]. The federated average algorithm (FedAvg), introduced by McMahan et al. [43] and Mangasarian and Solodov [41], is one of the most popular strategies for tackling federated learning problems. The algorithm comprises three essential components: client sampling, data sampling, and local training. The server samples a subset of clients to participate in each training round, where each selected client performs local training using stochastic gradient descent (SGD), with or without random reshuffling, to improve communication efficiency, as documented by Bubeck et al. [11], Gower et al. [22], Moulines and Bach [47], Sadiev et al. [65]. FedAvg has been highly successful in practice, but it suffers from client drift when data is heterogeneous [30].

Techniques like FedProx [37] have been proposed to address data heterogeneity. Instead of local SGD rounds, FedProx requires each client to compute a proximal operator, which can be treated as a local optimization problem. Proximal algorithms are effective when the proximal operators can be easily evaluated [50]. Proximal operator algorithms, like the proximal point method (PPM) [50, 62] and its stochastic extension (SPPM) [5, 8, 31, 51, 59], provide greater stability against inaccurately specified step sizes compared to gradient-based methods. This stability is especially valuable when problem-specific parameters, such as the objective function's smoothness constant, are unknown, making step size selection for SGD difficult. An excessively large step size in SGD leads to divergence, while a small step size ensures convergence but slows down the training process significantly.

Another approach to mitigating the slowdown caused by heterogeneity is the use of a server step size. In FedAvg, each client uses a local step size to minimize their individual objectives, while a server step size is applied to aggregate the 'pseudo-gradients' from each client [30, 58]. The local step size is kept small to mitigate client drift, while the server step size is larger to prevent slowdowns. However, the small local step size causes an initial training slowdown that the larger server step size cannot fully offset [24]. Building on the extrapolation technique used in parallel projection methods for solving convex feasibility problems [12, 13, 48], Jhunjhunwala et al. [24] introduced FedExP, an extension of FedAvg that incorporates adaptive extrapolation as the server step size. Extrapolation accelerates the algorithm by moving further along the line connecting the most recent iterate, x_k, and the average of its projections onto the convex sets $\mathcal{X}_i$ in the parallel projection method. In fixed point theory, this technique is also known as over-relaxation [56]. Extrapolation is a common technique used to accelerate the convergence of fixed point methods, including gradient-based and proximal splitting algorithms [14, 23]. Recently, Li et al. [35] demonstrated that combining extrapolation with FedProx improves complexity bounds. The analysis of the resulting algorithm, FedExProx, highlights the relationship between the extrapolation parameter and the step size of gradient-based methods concerning the Moreau envelope of the original objective function. However, it assumes each proximal operator is solved accurately, making it less practical and less advantageous compared to gradient-based methods.

1.1. Contributions

Our paper makes the following contributions, please refer to Appendix A for notation details.

- We provide a new analysis of FedExProx, building on Li et al. [35], focusing on the case where proximal operators are evaluated inexactly within the global strongly convex setting, eliminating the assumption of exact evaluations. By properly defining the approximation notion, we establish a general convergence guarantee to a neighborhood of the solution using biased SGD theory [16]. Specifically, our algorithm achieves a linear convergence rate of $\mathcal{O}\left(L_\gamma(1+\gamma L_{\max})/\mu\right)$ to a neighborhood of the solution, matching the rate from Li et al. [35].
- Building on our understanding of how the neighborhood arises, we propose a new method of approximation. This alternative characterization of inexactness removes the neighborhood from the previous convergence guarantee, provided the inexactness is properly bounded and the extrapolation parameter is chosen to be sufficiently small.
- By leveraging the similarity between the definitions of inexactness and compression, we enhance our analysis using the theory of biased compression [9]. The improved analysis offers a faster rate of $\mathcal{O}\left(\frac{L_\gamma(1+\gamma L_{\max})}{\mu-4\varepsilon_2 L_{\max}}\right)$[1], leading to convergence to the exact solution, provided that the inexactness is bounded in a more permissive manner. More importantly, the optimal extrapolation $1/\gamma L_\gamma$ matches the exact case. This shows that extrapolation aids convergence as long as sufficient accuracy is reached, even with inexact proximal evaluations.
- We analyze how clients can achieve these approximations, providing local iteration complexity for gradient descent (GD) and Nesterov's accelerated gradient descent (AGD). For the i-th client, the complexity is $\tilde{\mathcal{O}}\left(1+\gamma L_i\right)$ for GD and $\tilde{\mathcal{O}}\left(\sqrt{1+\gamma L_i}\right)$ for AGD. See Table 1 and Table 2 for a detailed comparison of complexities.

1. The parameter ε_2 is the parameter associated with accuracy of relative approximation as defined in Definition 4. We use the notation $\mathcal{O}\left(\cdot\right)$ to ignore constant factors and $\tilde{\mathcal{O}}\left(\cdot\right)$ when logarithmic factors are also omitted.

- Finally, we validate our theoretical findings through numerical experiments. The results show that the proposed relative approximation technique effectively eliminates bias. In some cases, the algorithm outperforms FedProx with exact updates, further proving the effectiveness of server extrapolation, even with inexact proximal updates.

2. Mathematical background

We consider the following distributed 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 $x \in \mathbb{R}^d$ is the model, $f : \mathbb{R}^d \mapsto \mathbb{R}$ is global objective, $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ is the empirical risk of model x for the i-th client.

Definition 1 (Proximal operator) *The proximal operator of an extended real-valued function $\phi : \mathbb{R}^d \mapsto \mathbb{R}\cup\{+\infty\}$ with step size $\gamma > 0$ and center $x \in \mathbb{R}^d$ is defined as*

$$\operatorname{prox}_{\gamma\phi}(x) := \arg\min_{z\in\mathbb{R}^d}\left\{\phi\{z\} + \frac{1}{2\gamma}\|z-x\|^2\right\}.$$

Definition 2 (Moreau envelope) *The Moreau envelope of an extended real-valued function $\phi : \mathbb{R}^d \mapsto \mathbb{R}\cup\{+\infty\}$ with step size $\gamma > 0$ and center $x \in \mathbb{R}^d$ is defined as*

$$M_\phi^\gamma(x) := \min_{z\in\mathbb{R}^d}\left\{\phi(z) + \frac{1}{2\gamma}\|z-x\|^2\right\}.$$

For Moreau envelope, we have

$$M_\phi^\gamma(x) = \phi\left(\operatorname{prox}_{\gamma\phi}(x)\right) + \frac{1}{2\gamma}\left\|x - \operatorname{prox}_{\gamma\phi}(x)\right\|^2.$$

Their function values are related, and for any proper, closed, convex function ϕ, the Moreau envelope is differentiable.

$$\nabla M_\phi^\gamma(x) = \frac{1}{\gamma}\left(x - \operatorname{prox}_{\gamma f}(x)\right). \tag{2}$$

This relationship plays a key role in our analysis.

Assumption 1 (Differentiability) *The function $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ in (1) is differentiable and bounded from below for all $i \in [n]$.*

Assumption 2 (Interpolation regime) *There exists $x_\star \in \mathbb{R}^d$ such that $\nabla f_i(x_\star) = 0$ for all $i \in [n]$.*

Following Li et al. [35], we assume the interpolation regime, common in modern deep learning where parameters d exceed data points [4, 45]. This assumption is motivated by parallel projection methods for solving convex feasibility problems, where the non-empty intersection of all convex sets $\mathcal{X}_i$ corresponds to the interpolation assumption that each f_i is the indicator function of $\mathcal{X}_i$. Extrapolation is known to improve these methods [48], and since $\operatorname{prox}_{\gamma f_i}(x_k)$ resembles projecting onto a level set of f_i, it is reasonable to expect extrapolation to be effective here as well.

Assumption 3 (Individual convexity) *The function $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ is convex for all $i \in [n]$. This means that for each f_i,*

$$0 \le f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y \rangle, \quad \forall x, y \in \mathbb{R}^d. \tag{3}$$

Assumption 4 (Smoothness) *The function $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ is L_i-smooth, $L_i > 0$ for all $i \in [n]$[2]. This means that for each f_i,*

$$f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y \rangle \le \frac{L_i}{2} \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d. \tag{4}$$

Assumption 5 (Global strong convexity) *The function f is μ-strongly convex, $\mu > 0$. That is*

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \ge \frac{\mu}{2} \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

We present our algorithm in Algorithm 1, with analyses under different inexactness definitions in Section 3 and Section 4. Client methods to achieve inexactness are in Appendix E, and numerical experiments in Appendix I validate our results.

3. Absolute approximation in distance

The local optimization problem for client i is given by, $\min_{z \in \mathbb{R}^d} A_{k,i}^{\gamma}(z) := f_i(z) + \frac{1}{2\gamma} \|z - x_k\|^2$, where x_k is the current iterate and $\gamma > 0$ is a constant. Since each function f_i is convex, $A_{k,i}^{\gamma}(z)$ is $1/\gamma$-strongly convex, and its unique minimizer is $\operatorname{prox}_{\gamma f_i}(x_k)$.

Definition 3 (Absolute approximation) *Given a proper, closed and convex function $\phi : \mathbb{R}^d \mapsto \mathbb{R}$, and a step size $\gamma > 0$, we say that a point $y \in \mathbb{R}^d$ is an ε_1-approximation of $\operatorname{prox}_{\gamma\phi}(x)$, if for some $\varepsilon_1 \ge 0$,*

$$\left\| y - \operatorname{prox}_{\gamma f}(x) \right\|^2 \le \varepsilon_1. \tag{5}$$

In order to analyze Algorithm 1, we first transform the update rule given in (8) in the following way,

$$\begin{aligned} x_{k+1} &= x_k + \alpha_k \left(\frac{1}{n} \sum_{i=1}^n \left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}(x_k) \right) + \frac{1}{n} \sum_{i=1}^n \operatorname{prox}_{\gamma f_i}(x_k) - x_k \right) \\ &\overset{(2)}{=} x_k - \alpha_k \cdot \underbrace{\frac{1}{n} \sum_{i=1}^n \gamma \nabla M_{f_i}^{\gamma}(x_k)}_{\text{Gradient}} + \alpha_k \cdot \underbrace{\frac{1}{n} \sum_{i=1}^n \left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}(x_k) \right)}_{\text{Bias}}. \end{aligned} \tag{6}$$

The above reformulation suggests that Algorithm 1 is in fact, SGD with respect to global objective $\gamma M^{\gamma}(x) := \frac{1}{n} \sum_{i=1}^n \gamma M_{f_i}^{\gamma}(x)$ with a biased gradient estimator. Compared to SGD with an unbiased gradient estimator, its biased counterpart is less well understood. However, we are still able to obtain the following convergence guarantee using theories for biased SGD from [16].

2. We will use $L_{\max}$ to denote $\max_{i \in [n]} L_i$.

Theorem 1 *Assume Assumption 1 (Differentiability), Assumption 2 (Interpolation Regime), Assumption 3 (Individual convexity), Assumption 4 (Smoothness) and Assumption 5 (Global strong convexity) hold. If each client only computes a ε_1-absolute approximation $\tilde{x}_{i,k+1}$ of $\operatorname{prox}_{\gamma f_i}(x_k)$, such that $\left\|\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}(x_k)\right\|^2 \leq \varepsilon_1$. Then we have the following convergence guarantee or Algorithm 1: For a constant extrapolation parameter satisfying $0 < \alpha \leq 1/4\gamma L_\gamma$, where γ is the step size of the proximal operator, $\alpha_k = \alpha$ is a constant extrapolation parameter, L_γ is the smoothness constant of M^γ. The last iterate x_K satisfy*

$$\mathcal{E}_K \leq \left(1 - \frac{\alpha\gamma\mu}{8\left(1+\gamma L_{\max}\right)}\right)^K \mathcal{E}_0 + \frac{4\varepsilon_1\left(1+\gamma L_{\max}\right)}{\mu} \cdot \left(2\alpha L_\gamma + \frac{1}{\gamma}\right),$$

where $\mathcal{E}_k = \gamma M^\gamma(x_k) - \gamma M^\gamma_{\inf}$. Specifically, when we choose $\alpha = 1/4\gamma L_\gamma$, we have

$$\Delta_K \leq \left(1 - \frac{\mu}{32 L_\gamma\left(1+\gamma L_{\max}\right)}\right)^K \frac{L_\gamma\left(1+\gamma L_{\max}\right)}{\mu} \cdot \Delta_0 + 12\varepsilon_1 \cdot \left(\frac{1/\gamma + L_{\max}}{\mu}\right)^2,$$

where $\Delta_K = \left\|x_K - x_\star\right\|^2$, $x_\star$ is a minimizer of f.

As per Fact 7, the minimizer of M^γ also minimizes f. The algorithm converges to a neighborhood around $x_\star$, with its size dependent on ε_1 and γ. A smaller γ leads to less progress per iteration, increasing the accumulated error over more iterations and, consequently, enlarging the neighborhood size. While ε_1 can be arbitrarily large, the larger neighborhood reduces the practical significance. For $\varepsilon_1 = 0$, the neighborhood disappears, yielding an iteration complexity of $\tilde{\mathcal{O}}\left(L_\gamma(1+\gamma L_{\max})/\mu\right)$[3], which recovers the result of Li et al. [35] up to a constant factor. The optimal extrapolation parameter is $\alpha_\star = 1/4\gamma L_\gamma$, 4 times smaller than that of Li et al. [35].

4. Relative approximation in distance

A key challenge in the above analysis is that without exact proximal evaluations, convergence is limited to a neighborhood of the solution. As the algorithm progresses, the gradient term in the estimator $g(x_k)$ diminishes, while the bias term remains unchanged. Based on this, we propose using a different type of approximation.

Definition 4 (Relative approximation) *Given a convex function $\phi : \mathbb{R}^d \mapsto \mathbb{R}$ and a stepsize $\gamma > 0$, we say that a point $y \in \mathbb{R}^d$ is a ε_2-relative approximation of $\operatorname{prox}_{\gamma\phi}(x)$, if for some $\varepsilon_2 \in [0, 1)$,*

$$\left\|y - \operatorname{prox}_{\gamma\phi}(x)\right\|^2 \leq \varepsilon_2 \cdot \left\|x - \operatorname{prox}_{\gamma\phi}(x)\right\|^2. \tag{7}$$

We require $\varepsilon_2 < 1$ to ensure the next iterate is no worse than the current one. If each proximal approximation meets Definition 4, both the gradient and bias terms decrease, ensuring convergence to the exact solution. Using the theory of biased SGD, we obtain the following theorem.

Theorem 2 *Assume all assumptions of Theorem 1 hold. If each client computes a ε_2-relative approximation $\tilde{x}_{i,k+1}$ with $\varepsilon_2 < \mu^2/4L^2_{\max}$, so that $\left\|\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}(x_k)\right\|^2 \leq \varepsilon_2 \cdot \left\|x_k - \operatorname{prox}_{\gamma f_i}(x_k)\right\|^2$. If we are running Algorithm 1 with $\alpha_k = \alpha$ satisfying*

$$0 < \alpha \leq \frac{1}{\gamma L_\gamma} \cdot \frac{\mu - 2\sqrt{\varepsilon_2} L_{\max}}{\mu + 4\sqrt{\varepsilon_2} L_{\max} + 4\varepsilon_2 L_{\max}}.$$

3. We leave out the log factor in $\tilde{\mathcal{O}}(\cdot)$ notation.

Then the iterates generated by Algorithm 1 satisfies

$$\mathcal{E}_K \leq \left(1 - \alpha \cdot \frac{\gamma\left(\mu - 2\sqrt{\varepsilon_2} L_{\max}\right)}{4\left(1 + \gamma L_{\max}\right)}\right)^K \mathcal{E}_0.$$

Specifically, if we choose the largest α *possible, we have*

$$\Delta_K \leq \left(1 - \frac{\mu}{4L_\gamma\left(1 + \gamma L_{\max}\right)} \cdot S\left(\varepsilon_2\right)\right)^K \cdot \frac{L\gamma\left(1 + \gamma L_{\max}\right)}{\mu} \Delta_0,$$

where $S(\varepsilon_2) := \frac{\left(\mu - 2\sqrt{\varepsilon_2} L_{\max}\right)\left(1 - 2\sqrt{\varepsilon_2}\frac{L_{\max}}{\mu}\right)}{\mu + 4\sqrt{\varepsilon_2} L_{\max} + 4\varepsilon_2 L_{\max}}$ *satisfies* $0 < S(\varepsilon_2) \leq 1$ *is the factor of slowing down due to inexact proximal operator evaluation.*

When $\varepsilon_2 = 0$, the optimal extrapolation is $\alpha = 1/\gamma L_\gamma$ with iteration complexity $\tilde{\mathcal{O}}\left(L_\gamma(1+\gamma L_{\max})/\mu\right)$, which recovers the exact result from Li et al. [35]. As ε_2 increases, both α and $S(\varepsilon_2)$ decrease, leading to a slower rate of convergence. Note that ε_2 must satisfy $\varepsilon_2 < \mu^2/4L_{\max}^2$

Definition 4 relates to the concept of compression. Indeed, we have $x_k - \mathrm{prox}_{\gamma f_i}(x_k) = \gamma \nabla M_{f_i}^{\gamma}(x_k)$, while $\tilde{x}_{i,k+1} - \mathrm{prox}_{\gamma f_i}(x_k)$ can be interpreted as the compressed gradient, that is, $\mathcal{C}(\gamma \nabla M_{f_i}^{\gamma}(x_k))$. In this case, Algorithm 1 can be viewed as compressed gradient descent with biased compressor. We obtain the following convergence guarantee based on theory provided by Beznosikov et al. [9].

Theorem 3 *Assume all assumptions of Theorem 1 hold. Let the approximation* $\tilde{x}_{i,k+1}$ *all satisfies Definition 4 with* $\varepsilon_2 < \mu/4L_{\max}$*, that is* $\left\|\tilde{x}_{i,k+1} - \mathrm{prox}_{\gamma f_i}(x_k)\right\|^2 \leq \varepsilon_2 \cdot \left\|x_k - \mathrm{prox}_{\gamma f_i}(x_k)\right\|^2$*. If we are running Algorithm 1 with* $\alpha_k = \alpha \in (0, 1/\gamma L_\gamma]$*, we have the iterates produced by it satisfying*

$$\mathcal{E}_K \leq \left(1 - \left(1 - \frac{4\varepsilon_2 L_{\max}}{\mu}\right) \cdot \frac{\gamma\mu}{4\left(1 + \gamma L_{\max}\right)} \cdot \alpha\right)^K \mathcal{E}_0.$$

specifically, if we take the largest extrapolation ($\alpha = 1/\gamma L_\gamma > 1$*) possible, we have*

$$\Delta_K \leq \left(1 - \left(1 - \frac{4\varepsilon_2 L_{\max}}{\mu}\right) \cdot \frac{\mu}{4L_\gamma\left(1 + \gamma L_{\max}\right)}\right)^K \cdot \frac{L_\gamma\left(1 + \gamma L_{\max}\right)}{\mu} \Delta_0.$$

The convergence guarantee is sharper, as Theorem 3 shows that if $\varepsilon_2 < \mu/4L$, we can set $\alpha = 1/\gamma L_\gamma$[4],the optimal extrapolation for exact proximal computation from Li et al. [35]. This demonstrates that extrapolation effectively accelerates the algorithm, even with inexact proximal evaluations. When $\varepsilon_2 = 0$, we recover the result from Li et al. [35].

References

[1] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, 30, 2017.

4. It is shown in Li et al. [35] that $1/\gamma L_\gamma > 1$, which justifies why α is called the extrapolation parameter.

[2] Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. *Advances in Neural Information Processing Systems*, 31, 2018.

[3] Mihai Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 10(4):1116–1135, 2000.

[4] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

[5] Hilal Asi and John C Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

[6] Heinz H Bauschke, Patrick L Combettes, and Serge G Kruk. Extrapolation algorithm for affine-convex feasibility problems. *Numerical Algorithms*, 41:239–274, 2006.

[7] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[8] Dimitri P Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129(2):163–195, 2011.

[9] Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *Journal of Machine Learning Research*, 24(276):1–50, 2023.

[10] Pascal Bianchi. Ergodic convergence of a stochastic proximal point algorithm. *SIAM Journal on Optimization*, 26(4):2235–2260, 2016.

[11] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[12] Y Censor, T Elfving, and GT Herman. Averaging strings of sequential iterations for convex feasibility problems. In *Studies in Computational Mathematics*, volume 8, pages 101–113. Elsevier, 2001.

[13] 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.

[14] Laurent Condat, Daichi Kitahara, Andrés Contreras, and Akira Hirabayashi. Proximal splitting algorithms for convex optimization: A tour of recent advances, with new twists. *SIAM Review*, 65(2):375–435, 2023.

[15] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[16] Yury Demidovich, Grigory Malinovsky, Igor Sokolov, and Peter Richtárik. A guide through the zoo of biased sgd. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[18] Pinghua Gong and Jieping Ye. Linear convergence of variance-reduced stochastic gradient without strong convexity. *arXiv preprint arXiv:1406.1102*, 2014.

[19] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent. In *International Conference on Artificial Intelligence and Statistics*, pages 680–690. PMLR, 2020.

[20] Eduard Gorbunov, Konstantin P Burlachenko, Zhize Li, and Peter Richtárik. MARINA: Faster non-convex distributed learning with compression. In *International Conference on Machine Learning*, pages 3788–3798. PMLR, 2021.

[21] Robert M Gower, Mark Schmidt, Francis Bach, and Peter Richtárik. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.

[22] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR, 2019.

[23] Franck Iutzeler and Julien M Hendrickx. A generic online acceleration scheme for optimization algorithms via relaxation and inertia. *Optimization Methods and Software*, 34(2):383–405, 2019.

[24] Divyansh Jhunjhunwala, Shiqiang Wang, and Gauri Joshi. FedExP: Speeding up federated averaging via extrapolation. In *International Conference on Learning Representations*, 2023.

[25] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

[26] Abderrahim Jourani, Lionel Thibault, and Dariusz Zagrodny. Differential properties of the moreau envelope. *Journal of Functional Analysis*, 266(3):1185–1237, 2014.

[27] Stefan Kaczmarz. Approximate solution of systems of linear equations. *International Journal of Control*, 57(6):1269–1271, 1937.

[28] Avetik Karagulyan, Egor Shulgin, Abdurakhmon Sadiev, and Peter Richtárik. Spam: Stochastic proximal point method with momentum variance reduction for non-convex cross-device federated learning. *arXiv preprint arXiv:2405.20127*, 2024.

[29] Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *International Conference on Machine Learning*, pages 3252–3261. PMLR, 2019.

[30] 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*, pages 5132–5143. PMLR, 2020.

[31] Ahmed Khaled and Chi Jin. Faster federated optimization under second-order similarity. In *The Eleventh International Conference on Learning Representations*, 2022.

[32] Sarit Khirirat, Hamid Reza Feyzmahdavian, and Mikael Johansson. Distributed learning with compressed gradients. *arXiv preprint arXiv:1806.06573*, 2018.

[33] 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*, 8, 2016.

[34] Hanmin Li, Avetik Karagulyan, and Peter Richtárik. Variance reduced distributed non-convex optimization using matrix stepsizes. *arXiv preprint arXiv:2310.04614*, 2023.

[35] Hanmin Li, Kirill Acharya, and Peter Richtárik. The power of extrapolation in federated learning. *arXiv preprint arXiv:2405.13766*, 2024.

[36] Hanmin Li, Avetik Karagulyan, and Peter Richtárik. Det-CGD: Compressed gradient descent with matrix stepsizes for non-convex optimization. In *International Conference on Learning Representations*, 2024.

[37] 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.

[38] Ji Liu and Stephen J Wright. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM Journal on Optimization*, 25(1):351–376, 2015.

[39] Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems*, 33:18261–18271, 2020.

[40] Nicolas Loizou and Peter Richtárik. Linearly convergent stochastic heavy ball method for minimizing generalization error. *arXiv preprint arXiv:1710.10737*, 2017.

[41] Olvi L Mangasarian and Mikhail V Solodov. Backpropagation convergence via deterministic nonmonotone perturbed minimization. *Advances in Neural Information Processing Systems*, 6, 1993.

[42] Bernard Martinet. *Algorithmes pour la résolution de problèmes d'optimisation et de minimax*. PhD thesis, Université Joseph-Fourier-Grenoble I, 1972.

[43] 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*, pages 1273–1282. PMLR, 2017.

[44] Konstantin Mishchenko, Slavomir Hanzely, and Peter Richtárik. Convergence of first-order algorithms for meta-learning with Moreau envelopes. *arXiv preprint arXiv:2301.06806*, 2023.

[45] Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *The Annals of Statistics*, 50(5):2816–2847, 2022.

[46] Jean-Jacques Moreau. Proximité et dualité dans un espace Hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

[47] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems*, 24, 2011.

[48] 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.

[49] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publishers, 2004.

[50] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[51] Andrei Patrascu and Ion Necoara. Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *Journal of Machine Learning Research*, 18 (198):1–42, 2018.

[52] Guy Pierra. Decomposition through formalization in a product space. *Mathematical Programming*, 28:96–115, 1984.

[53] Chayne Planiden and Xianfu Wang. Strongly convex functions, Moreau envelopes, and the generic nature of convex functions with strong minimizers. *SIAM Journal on Optimization*, 26 (2):1341–1364, 2016.

[54] Chayne Planiden and Xianfu Wang. Proximal mappings and Moreau envelopes of single-variable convex piecewise cubic functions and multivariable gauge functions. *Nonsmooth Optimization and Its Applications*, pages 89–130, 2019.

[55] Boris T Polyak. Gradient methods for solving equations and inequalities. *USSR Computational Mathematics and Mathematical Physics*, 4(6):17–32, 1964.

[56] LF Rechardson. The approximate arithmetical solution by finite difference of physical problems involving differential equations, with an application to the stresses in a masonary dam. *R. Soc. London Phil. Trans. A*, 210:307–357, 1911.

[57] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *Advances in Neural Information Processing Systems*, 24, 2011.

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

[59] Peter Richtárik and Martin Takác. Stochastic reformulations of linear systems: algorithms and convergence theory. *SIAM Journal on Matrix Analysis and Applications*, 41(2):487–524, 2020.

[60] Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. Ef21: A new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34: 4384–4396, 2021.

[61] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[62] R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[63] Ernest K Ryu and Stephen Boyd. Stochastic proximal iteration: a non-asymptotic improvement upon stochastic gradient descent. *Author website, early draft*, 2014.

[64] Abdurakhmon Sadiev, Dmitry Kovalev, and Peter Richtárik. Communication acceleration of local gradient methods via an accelerated primal-dual algorithm with an inexact prox. *Advances in Neural Information Processing Systems*, 35:21777–21791, 2022.

[65] Abdurakhmon Sadiev, Grigory Malinovsky, Eduard Gorbunov, Igor Sokolov, Ahmed Khaled, Konstantin Burlachenko, and Peter Richtárik. Federated optimization algorithms with random reshuffling and gradient compression. *arXiv preprint arXiv:2206.07021*, 2022.

[66] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Interspeech*, volume 2014, pages 1058–1062. Singapore, 2014.

[67] Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with Moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.

[68] Alexander Tyurin and Peter Richtárik. DASHA: Distributed nonconvex optimization with communication compression and optimal oracle complexity. In *International Conference on Learning Representations*, 2024.

Contents

Appendix A. Notations

Throughout the paper, we use the notation $\|\cdot\|$ to denote the standard Euclidean norm defined on $\mathbb{R}^d$ and $\langle \cdot, \cdot \rangle$ to denote the standard Euclidean inner product. Given a differentiable function $f : \mathbb{R}^d \mapsto \mathbb{R}$, its gradient is denoted as $\nabla f(x)$. We use the notation $D_f(x, y)$ to denote the Bregman

Algorithm 1 Inexact FedExProx

1: **Parameters:** extrapolation parameter $\alpha_k = \alpha > 0$, step size for the proximal operator $\gamma > 0$, starting point $x_0 \in \mathbb{R}^d$, number of clients n, total number of iterations K, proximal solution accuracy $\varepsilon \geq 0$.
2: **for** $k = 0, 1, 2 \ldots K-1$ **do**
3: The server broadcasts the current iterate x_k to each client
4: Each client computes an ε approximation of the solution $\tilde{x}_{i,k+1} \simeq \mathrm{prox}_{\gamma f_i}(x_k)$, and sends it back to the server
5: The server computes

$$x_{k+1} = x_k + \alpha_k \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{x}_{i,k+1} - x_k \right). \tag{8}$$

6: **end for**

divergence associated with a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ between x and y. The notation $\inf f$ is used to denote the minimum of a function $f : \mathbb{R}^d \mapsto \mathbb{R}$. We use $\mathrm{prox}_{\gamma\phi}(x)$ to denote the proximity operator of function $\phi : \mathbb{R}^d \mapsto \mathbb{R}$ with $\gamma > 0$ at $x \in \mathbb{R}^d$, and $M^{\gamma}_{\phi}(x)$ to denote the corresponding Moreau Envelope. We denote the average of the Moreau envelope of each local objective f_i by the notation $M^{\gamma} : \mathbb{R}^d \mapsto \mathbb{R}$. Specifically, we define $M^{\gamma}(x) = \frac{1}{n}\sum_{i=1}^{n} M^{\gamma}_{f}(x)$. Note that $M^{\gamma}(x)$ has an implicit dependence on γ, its smoothness constant is denoted by L_{γ}. We say an extended real-valued function $f : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ is proper if there exists $x \in \mathbb{R}^d$ such that $f(x) < +\infty$. We say an extended real-valued function $f : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ is closed if its epigraph is a closed set. We use the notation $\mathcal{E}_k = \gamma M^{\gamma}(x_k) - \gamma M^{\gamma}_{\mathrm{inf}}$ to denote the function value suboptimality of γM^{γ} at x_k, and $\Delta_k = \|x_k - x_\star\|^2$ to denote the squared distance. The notation $\mathcal{O}(\cdot)$ is used to describe complexity while omitting constant factors, whereas $\tilde{\mathcal{O}}(\cdot)$ is used when logarithmic factors are also omitted.

Appendix B. Related work

Arguably, stochastic gradient descent (SGD) [17, 19, 22, 61] remains one of the foundational algorithm in the field of machine learning. One can simply formulate it as

$$x_{k+1} = x_k - \eta \cdot g(x_k),$$

where $\eta > 0$ is a scalar step size, $g(x_k)$ is a possibly stochastic estimator of the true gradient $\nabla f(x_k)$. In the case when $g(x_k) = \nabla f(x_k)$, SGD becomes GD. Various extensions of SGD have been proposed since its introduction, examples include compressed gradient descent (CGD) [1, 32], SGD with momentum [39, 40], SGD with matrix step size [36] and variance reduction [20, 21, 25, 34, 68]. Gower et al. [22] presented a framework for analyzing SGD with unbiased gradient estimator in the convex case based on expected smoothness. However, in practice, sometimes the gradient estimator could be biased, examples include SGD with sparsified or delayed update [2, 57]. Beznosikov et al. [9] examined biased updates in the context of compressed gradient descent. Demidovich et al. [16] provides a framework for analyzing SGD with biased gradient estimators in the non-convex setting.

Proximal point method (PPM) was originally introduced as a method to solve variational inequalities [42, 62]. The transition to the stochastic case, driven by the need to efficiently address large-scale optimization problems, leads to the development of SPPM. Due to its stability and advantage over the gradient based methods, it has been extensively studied, as documented by [8, 10, 51]. For proximal algorithms to be practical, it is commonly assumed that the proximal operator can be solved efficiently, such as in cases where a closed-form solution is available. However, in large-scale machine learning models, it is rarely possible to find such a solution in closed form. To address this issue, most proximal algorithms assume that only an approximate solution is obtained, achieving a certain level of accuracy [28, 31, 64]. Various notions of inexactness are employed, depending on the assumptions made, the properties of the objective, and the availability of algorithms capable of efficiently finding such approximations.

Moreau envelope was first introduced to handle non-smooth functions by Moreau [46]. It is also known as the Moreau-Yosida regularization. The use of the Moreau envelope as an analytical tool to analyze proximal algorithms is not novel. Ryu and Boyd [63] noted that running a proximal algorithm on the objective is equivalent to applying gradient methods to its Moreau envelope. Davis and Drusvyatskiy [15] analyzed stochastic proximal point method (SPPM) for weakly convex and Lipschitz functions based on this finding. Recently, Li et al. [35] provided an analysis of FedProx with server-side step size in the convex case, based on the reformulation of the problem using the Moreau envelope. The role of the Moreau envelope extends beyond analyzing proximal algorithms; it has also been applied in the contexts of personalized federated learning [67] and meta-learning [44]. The mathematical properties of the Moreau envelope are relatively well understood, as documented by Jourani et al. [26], Planiden and Wang [53, 54].

Projection methods initially emerged as an effective tool for solving systems of linear equations or inequalities [27] and were later generalized to solve the convex feasibility problem [13]. The parallel version of this approach involves averaging the projections of the current iterates onto all existing convex sets $\mathcal{X}_i$ to obtain the next iterate, a process that is empirically known to be accelerated by extrapolation. Numerous heuristic rules have been proposed to adaptively set the extrapolation parameter, such as those by Bauschke et al. [6] and Pierra [52]. Only recently, the mechanism behind constant extrapolation was uncovered by Necoara et al. [48], who developed the corresponding theoretical framework. Additionally, Li et al. [35] provides explanations for the effectiveness of adaptive rules, revealing the connection between the extrapolation parameter and the step size of SGD when using the Moreau envelope as the global objective.

Appendix C. Facts and lemmas

Fact 1 (Young's inequality) *For any two vectors $x, y \in \mathbb{R}^d$, the following inequality holds,*

$$\|x + y\|^2 \leq 2\|x\|^2 + 2\|y\|^2 . \tag{9}$$

Fact 2 (Property of convex smooth functions) *Let $\phi : \mathbb{R}^d \mapsto \mathbb{R}$ be differentiable. The following statements are equivalent:*

1. *ϕ is convex and L-smooth.*
2. *$0 \leq 2D_{\phi}(x, y) \leq L\|x - y\|^2$ for all $x, y \in \mathbb{R}^d$.*
3. *$\frac{1}{L}\|\nabla\phi(x) - \nabla\phi(y)\|^2 \leq 2D_{\phi}(x, y)$ for all $x, y \in \mathbb{R}^d$.*

Table 1: Comparison of FedExProx [35] and our proposed inexact versions of the algorithms using different approximations. In the convergence column, we present the rate at which each algorithm converges to either the solution or a neighborhood in the global strongly convex setting. Here, L_γ represents the smoothness constant of M^γ as defined before Theorem 1. The neighborhood column indicates the size of the neighborhood, while the optimal extrapolation column suggests the best choice of α for each algorithm. The final column outlines the conditions on the inexactness. All quantities are presented with constant factors omitted, K is the number of total iterations, γ is the local step size for the proximal operator, $S(\varepsilon_2)$ defined in Theorem 2 is a factor of slowing down due to inexactness in $(0,1]$. For relative approximation, we first present the original theory in the third row and then place the sharper analysis in the following row for comparison.

Algorithm	**Convergence**	**Neighborhood**	**Optimal Extrapolation**	**Bound on Inexactness**
FedExProx	$\exp\left(-\frac{K\mu}{L_\gamma(1+\gamma L_{\max})}\right)$	0	$\frac{1}{\gamma L_\gamma}$	NA
(NEW) FedExProx with ε_1 approximation	$\exp\left(-\frac{K\mu}{L_\gamma(1+\gamma L_{\max})}\right)$	$\varepsilon_1\left(\frac{\frac{1}{\gamma}+L_{\max}}{\mu}\right)^2$ [(a)]	$\frac{1}{4\gamma L_\gamma}$	NA [(b)]
(NEW) FedExProx with ε_2 relative approximation by biased SGD	$\exp\left(-\frac{K\mu S(\varepsilon_2)}{L_\gamma(1+\gamma L_{\max})}\right)$ [(c)]	0	$< \frac{1}{\gamma L_\gamma}$	$< \frac{\mu^2}{4L_{\max}^2}$
(NEW) FedExProx with ε_2 relative approximation by biased compression	$\exp\left(-\frac{K(\mu-4\varepsilon_2 L_{\max})}{L_\gamma(1+\gamma L_{\max})}\right)$	0	$\frac{1}{\gamma L_\gamma}$ [(d)]	$< \frac{\mu}{4L_{\max}}$

[(a)] Note that when $\varepsilon_1 = 0$, i.e., when the proximal operators are evaluated exactly, the neighborhood diminishes, and we recover the result of FedExProx by Li et al. [35], up to a constant factor.
[(b)] Unlike relative approximations, the convergence guarantee here is more general, allowing for the analysis of unbounded inexactness. However, as the inexactness increases, the neighborhood grows correspondingly, rendering the result practically useless.
[(c)] Refer to Theorem 2 for the definition of $S(\varepsilon_2)$ and the corresponding optimal extrapolation parameter. The theory indicates that inexactness will adversely affect the algorithm's convergence.
[(d)] Surprisingly, our sharper analysis reveals that the optimal extrapolation parameter in this case remains the same as in the exact setting, highlighting the effectiveness of extrapolation even when the proximal operators are evaluated inexactly.

The notation $D_\phi(x,y)$ denotes the Bregman divergence associate with ϕ at $x, y \in R^d$, defined as

$$D_\phi(x,y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x-y\rangle .$$

The following two facts establish that the convexity and smoothness of a function $\phi : \mathbb{R}^d \mapsto \mathbb{R}$ ensure the convexity and smoothness of its Moreau envelope.

Fact 3 (Convexity of Moreau envelope) *[7, Theorem 6.55] Let $\phi : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ be a proper and convex function. Then M_ϕ^γ is a convex function.*

Fact 4 (Smoothness of Moreau envelope) *[35, Lemma 4] Let $\phi : \mathbb{R}^d \mapsto \mathbb{R}$ be a convex and L-smooth function. Then M_ϕ^γ is $\frac{L}{1+\gamma L}$-smooth.*

Table 2: Comparison of local iteration complexities of each client in order to obtain an approximation using either GD or AGD [49]. We use the i-th client as an example, where the local objective $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ is L_i-smooth and convex, $i \in \{1, 2, \dots, n\}$.

Algorithm	ε_1 **absolute approximation**	ε_2 **relative approximation**
Gradient descent	$\mathcal{O}\left((1+\gamma L_i)\log\left(\frac{\left\|x_k - \mathrm{prox}_{\gamma f_i}(x_k)\right\|^2}{\varepsilon_1}\right)\right)$ [(a)]	$\mathcal{O}\left((1+\gamma L_i)\log\left(\frac{1}{\varepsilon_2}\right)\right)$
Accelerate gradient descent	$\mathcal{O}\left(\sqrt{1+\gamma L_i}\log\left(\frac{\left\|x_k - \mathrm{prox}_{\gamma f_i}(x_k)\right\|^2}{\varepsilon_1}\right)\right)$	$\mathcal{O}\left(\sqrt{1+\gamma L_i}\log\left(\frac{1}{\varepsilon_2}\right)\right)$

[(a)] We can easily provide an upper bound of $\left\|x_k - \mathrm{prox}_{\gamma f_i}(x_k)\right\|^2$ for determining the number of local computations needed.

The following fact illustrates the relationship between the minimizer of a function ϕ and its Moreau envelope M_ϕ^γ.

Fact 5 (Minimizer equivalence) *[35, Lemma 5] Let $\phi : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ be a proper, closed and convex function. Then for any $\gamma > 0$, ϕ and M_ϕ^γ has the same set of minimizers.*

In our case, we assume each f_i from (1) is convex and L_i-smooth. Therefore by Fact 3 and Fact 4, we know that each $M_{f_i}^\gamma$ is also convex and $\frac{L_i}{1+\gamma L_i}$-smooth. This means that $M_\gamma = \frac{1}{n}\sum_{i=1}^n M_{f_i}^\gamma$ is also convex and smooth. We denote its smoothness constant as L_γ, and the following fact provides a range for this constant.

Fact 6 (Global convexity and smoothness) *[35, Lemma 7] Let each f_i be proper, closed convex and L_i-smooth. Then M^γ is convex and L_γ-smooth with*

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

The following fact establishes that the minimizer of f and M^γ are the same.

Fact 7 (Global minimizer equivalence) *[35, Lemma 8] If we let every $f_i : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ be proper, closed and convex, then $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ has the same set of minimizers and minimum as*

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

if we are in the interpolation regime and $0 < \gamma < \infty$.

The above fact demonstrates that running SGD on the objective M^γ will lead us to the correct destination, as the minimizers of M^γ and f are identical in our setting. In problem (1), if we assume that f is strongly convex, then we have M^γ satisfies the following star strong convexity inequality.

Fact 8 (Star strong convexity) *[35, Lemma 11] Assume Assumption 1 (Differentiability), Assumption 2 (Interpolation Regime), Assumption 3 (Individual convexity), Assumption 4 (Smoothness) and Assumption 5 (Global strong convexity) hold, then the convex function $M^{\gamma}(x)$ satisfies the following inequality,*

$$M^{\gamma}(x) - M^{\gamma}_{\text{inf}} \geq \frac{\mu}{1+\gamma L_{\max}} \cdot \frac{1}{2} \|x - x_{\star}\|^2 ,$$

for any $x \in \mathbb{R}^d$ and a minimizer $x_{\star}$ of $M^{\gamma}(x)$.

The above fact implies that the strong convexity of f translates to the star strong convexity of M^{γ}. Star strong convexity is also known as quadratic growth (QG) condition [3]. In the case of a convex function, it is also known as optimal strong convexity [38] and semi-strong convexity [18]. It is known that for a convex function satisfying quadratic growth condition, it also satisfies the Polyak-Lojasiewicz inequality [55] which is described by the following lemma. Notice that since Algorithm 1 can be viewed as running SGD with objective γM^{γ} and a fixed step size $\alpha_k = \alpha$, we describe the inequality based on γM^{γ} in the following lemma.

Lemma 1 (PL-inequality) *Assume that Assumption 1 (Differentiability), Assumption 2 (Interpolation Regime), Assumption 3 (Individual convexity), Assumption 4 (Smoothness) and Assumption 5 (Global strong convexity) hold, then $\gamma M^{\gamma}(x)$ satisfies the following Polyak-Lojasiewicz inequality,*

$$\|\gamma \nabla M^{\gamma}(x)\|^2 \geq 2 \cdot \frac{\gamma\mu}{4(1+\gamma L_{\max})} \left(\gamma M^{\gamma}(x) - \gamma M^{\gamma}_{\text{inf}}\right), \tag{10}$$

where $x \in \mathbb{R}^d$ is an arbitrary vector and $x_{\star}$ is a minimizer of $M^{\gamma}(x)$.

Appendix D. Theory of biased SGD

For completeness, we provide the theory of biased SGD we used to analyze our algorithm in this paper. It is adapted from Demidovich et al. [16], which offers a comprehensive study of various assumptions employed in the analysis of SGD with biased gradient updates. In addition, the authors introduced a new set of assumptions, referred to as the Biased ABC assumption, which are less restrictive than all previous assumptions. The authors provided convergence guarantees for SGD with biased gradient updates in the non-convex and convex setting. Specifically, they considered the case of minimizing a function $f : \mathbb{R}^d \mapsto \mathbb{R}$,

$$\min_{x \in \mathbb{R}^d} f(x),$$

with

$$x_{k+1} = x_k - \eta g(x_k), \tag{biased SGD}$$

where $\eta > 0$ is the stepsize, $g(x_k)$ is a possibly stochastic and biased gradient estimator. They introduced the biased ABC assumption,

Assumption 6 (Biased-ABC) *[16, Assumption 9] There exists constants $A, B, C, b, c \geq 0$ such that the gradient estimator $g(x)$ for every $x \in \mathbb{R}^d$ satisfies*

$$\begin{aligned} \langle \nabla f(x), \mathbb{E}[g(x)] \rangle &\geq b \|\nabla f(x)\|^2 - c \\ \mathbb{E}\left[\|g(x)\|^2\right] &\leq 2A(f(x) - f_{\text{inf}}) + B\|\nabla f(x)\|^2 + C. \end{aligned}$$

A convergence guarantee was provided for biased SGD under Assumption 6 given that f is $\widehat{L}$-smooth and $\widehat{\mu}$-PL, that is, there exists $\widehat{\mu} > 0$, such that

$$\|\nabla f(x)\|^2 \geq 2\widehat{\mu}\left(f(x) - f_{\inf}\right),$$

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

Theorem 4 (Theory of biased SGD) *[16, Theorem 4] Let f be $\widehat{L}$-smooth and $\widehat{\mu}$-PL and Assumption 6 hold. If we choose a step size η satisfying*

$$0 < \eta < \min\left\{\frac{\widehat{\mu}b}{\widehat{L}\left(A + \widehat{\mu}B\right)}, \frac{1}{\widehat{\mu}b}\right\}. \tag{11}$$

Then we have

$$\mathbb{E}\left[f(x_k) - f_{\inf}\right] \leq \left(1 - \eta\widehat{\mu}b\right)^k \left(f(x_0) - f_{\inf}\right) + \frac{LC\eta}{2\widehat{\mu}b} + \frac{c}{\widehat{\mu}b}.$$

Under the special case of

$$\frac{\widehat{\mu}b}{\widehat{L}\left(A + \widehat{\mu}B\right)} < \frac{1}{\widehat{\mu}b},$$

The range of the step size can be simplified to

$$0 < \eta \leq \frac{\widehat{\mu}b}{\widehat{L}\left(A + \widehat{\mu}B\right)},$$

and if we take the largest possible step size, we have

$$\mathbb{E}\left[f(x_k) - f_{\inf}\right] \leq \left(1 - \frac{\widehat{\mu}^2 b^2}{\widehat{L}\left(A + \widehat{\mu}B\right)}\right)^k \left(f(x_0) - f_{\inf}\right) + \frac{LC}{2\widehat{L}\left(A + \widehat{\mu}B\right)} + \frac{c}{\widehat{\mu}b}.$$

The constants C, c determine whether the algorithm is converging to the exact solution or just a neighborhood. For $g(x) = \nabla f(x)$, clearly we have $A = 0, B = 1, b = 1, C = 0, c = 0$, and there is no neighborhood. This is expected because the algorithm reduces to standard GD The iteration complexity is give by $\tilde{\mathcal{O}}\left(\frac{\widehat{L}}{\widehat{\mu}}\right)$, which is also expected for GD.

Appendix E. Achieving the level of inexactness

To fully comprehend the overall complexity of Algorithm 1, it is essential to examine whether the inexactness in evaluating the proximal operators can be effectively achieved. Since each $\mathrm{prox}_{\gamma f_i}(x_k)$ is computed locally by the corresponding client, the client has access to all the necessary data points for the computation. Thus, the most straightforward approach is to have each client perform GD.

Theorem 5 (Local computation via GD) *Assume Assumption 1 (Differentiability), Assumption 3 (Individual convexity) and Assumption 4 (Smoothness) hold. The iteration complexity for the i-th client to provide an approximation using GD in the k-th iteration with local step size $\eta_i = \frac{\gamma}{1+\gamma L_i}$, satisfying Definition 3 is $\mathcal{O}\left(\left(1 + \gamma L_i\right)\log\left(\left\|x_k - \mathrm{prox}_{\gamma f_i}(x_k)\right\|^2 / \varepsilon_1\right)\right)$, and for Definition 4, it is $\mathcal{O}\left(\left(1 + \gamma L_i\right)\log\left(1/\varepsilon_2\right)\right)$.*

Note that there are no constraints on ε_1, and since $\left\|x_k - \mathrm{prox}_{\gamma f_i}(x_k)\right\|^2 \leq \|\gamma \nabla f(x_k)\|^2$ by (39), it is straightforward to adjust GD to optimize the approximation. However, for ε_2, we require $\varepsilon_2 < \frac{\mu}{4L_{\max}}$. In practice, ε_2 can be set to a sufficiently small value to satisfy this condition, though this will increase the number of local iterations performed by each client. The complexity bounds also indicate that as the local step size γ increases, it becomes more challenging to compute the approximation. We can use the accelerated gradient descent (AGD) of Nesterov [49] to obtain a better iteration complexity for each client.

Theorem 6 (Local computation via AGD) *Assume all assumptions mentioned in Theorem 5 hold. The iteration complexities for the i-th client to provide an approximation in the k-the iteration using AGD with local step size $\eta_i = \frac{\gamma}{1+\gamma L_i}$ and momentum parameter $\alpha_i = \frac{\sqrt{1+\gamma L_i}-1}{\sqrt{1+\gamma L_i}+1}$, satisfying Definition 3, Definition 4 are*

$$\mathcal{O}\left(\sqrt{1+\gamma L_i}\log\left(\frac{(1+\gamma L_i)\cdot\left\|x_k - \mathrm{prox}_{\gamma f_i}(x_k)\right\|^2}{\varepsilon_1}\right)\right); \quad \mathcal{O}\left(\sqrt{1+\gamma L_i}\log\left(\frac{1+\gamma L_i}{\varepsilon_2}\right)\right),$$

respectively.

Appendix F. Theory of biased compression

In this section, we present the theory of SGD with biased compression. The theory is adapted from Beznosikov et al. [9]. The authors introduced theory for analyzing compressed gradient descent (CGD) with biased compressor, both in the single node case and in the distributed case when the objective function is assumed to be strongly convex. Here, we are only concerned with the single node case because distributed compressed gradient descent (DCGD) with biased compressor may fail to converge. To address this issue, error feedback mechanism [29, 60, 66] is needed. In the single node case, the authors considered solving

$$\min_{x\in\mathbb{R}^d} f(x),$$

where $f : \mathbb{R}^d \mapsto \mathbb{R}$ is $\widehat{L}$-smooth and $\widehat{\mu}$-strongly convex, with the following compressed gradient descent algorithm

$$x_{k+1} = x_k - \eta\mathcal{C}\left(\nabla f(x_k)\right), \tag{CGD}$$

where $\mathcal{C} : \mathbb{R}^d \mapsto \mathbb{R}$ are potentially biased compression operators, $\eta > 0$ is a step size. The author proved that if certain conditions on $\mathcal{C}$ is satisfied, a corresponding convergence guarantee can then be established. Three classes of compressor/mapping were introduced.

Definition 7 (Class $\mathbb{B}^1$) *We say a mapping $\mathcal{C} \in \mathbb{B}^1(\alpha, \beta)$ for some $\alpha, \beta > 0$ if*

$$\alpha\|x\|^2 \leq \mathbb{E}\left[\|\mathcal{C}(x)\|^2\right] \leq \beta\left\langle \mathbb{E}\left[\mathcal{C}(x)\right], x\right\rangle, \qquad \forall x \in \mathbb{R}^d.$$

Definition 8 (Class $\mathbb{B}^2$) *We say a mapping $\mathcal{C} \in \mathbb{B}^2(\xi, \beta)$ for some $\xi, \beta > 0$ if*

$$\max\left\{\xi\|x\|^2, \frac{1}{\beta}\mathbb{E}\left[\|\mathcal{C}(x)\|^2\right]\right\} \leq \left\langle \mathbb{E}\left[\mathcal{C}(x)\right], x\right\rangle, \qquad \forall x \in \mathbb{R}^d.$$

Definition 9 (Class $\mathbb{B}^3$) *We say a mapping $\mathcal{C} \in \mathbb{B}^3(\delta)$ for some $\delta > 0$, if*

$$\mathbb{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq \left(1 - \frac{1}{\delta}\right)\|x\|^2.$$

The authors proved the following theorem about the convergence of the algorithm, the notation $\mathcal{F}_k$ is used to denote $\mathbb{E}[f(x_k)] - f_{\inf}$, with $\mathcal{F}_0 = f(x_0) - f_{\inf}$,

Theorem 10 *Let $\mathcal{C} \in \mathbb{B}^1(\alpha, \beta)$. Then we have $\mathcal{F}_k \leq \left(1 - \alpha/\beta\eta\widehat{\mu}\left(2 - \eta\beta\widehat{L}\right)\right)\mathcal{F}_{k-1}$, as long as $0 \leq \eta \leq \frac{2}{\beta\widehat{L}}$. If we choose $\eta = \frac{1}{\beta\widehat{L}}$, we have*

$$\mathcal{F}_k \leq \left(1 - \frac{\alpha}{\beta^2} \cdot \frac{\widehat{\mu}}{\widehat{L}}\right)^K \mathcal{F}_0. \tag{12}$$

Let $\mathcal{C} \in \mathbb{B}^2(\xi, \beta)$. Then we have $\mathcal{F}_k \leq \left(1 - \xi\eta(2 - \eta\beta)\widehat{L}\right)\mathcal{F}_{k-1}$, as long as $0 \leq \eta \leq \frac{2}{\beta\widehat{L}}$. If we choose $\eta = \frac{1}{\beta\widehat{L}}$, we have

$$\mathcal{F}_k \leq \left(1 - \frac{\xi}{\beta} \cdot \frac{\widehat{\mu}}{\widehat{L}}\right)^k \mathcal{F}_0. \tag{13}$$

Let $\mathcal{C} \in \mathbb{B}^3(\delta)$. Then we have $\mathcal{F}_k \leq \left(1 - \frac{1}{\delta}\eta\widehat{\mu}\right)\mathcal{F}_{k-1}$, as long as $0 \leq \eta \leq \frac{1}{\widehat{L}}$. If we choose $\eta = \frac{1}{\widehat{L}}$, we have

$$\mathcal{F}_k \leq \left(1 - \frac{1}{\delta} \cdot \frac{\widehat{\mu}}{\widehat{L}}\right)^k \mathcal{F}_0. \tag{14}$$

Notice that when $\mathcal{C}(x) = x$, that is, when no compression happens, we have $\alpha = \beta = \xi = \delta = 1$. In this case, the iteration complexity of CGD is given by $\tilde{\mathcal{O}}\left(\frac{\widehat{L}}{\widehat{\mu}}\right)$ and we recover the result of GD. It is worth noting that Theorem 10 remains valid if the condition of f being $\widehat{\mu}$-strongly convex is replaced with f being $\widehat{\mu}$-PL.

Appendix G. Analysis of inexact FedExProx in the client sampling setting

In this section, we will discuss the case where we do client sampling in algorithm 1, we first formulate the algorithm as below. For the sake of simplicity, we use τ-nice sampling as an example.

G.1. Relative approximation in distance

The failure of biased compression theory: Similar to Theorem 10, we initially apply the theory from Beznosikov et al. [9], as it provides improved results in the full-batch scenario. We first define the compressing mapping $\mathcal{C}_\tau$ in this case,

$$\mathcal{C}_\tau(\gamma\nabla M^\gamma(x_k)) = \frac{1}{\tau}\sum_{i\in S_k}\left(\gamma\nabla M^\gamma_{f_i}(x_k) - \left(\tilde{x}_{i,k+1} - \text{prox}_{\gamma f_i}(x_k)\right)\right). \tag{16}$$

One can verify for every x_k and ε_2-approximation $\tilde{x}_{i,k+1}$ of $\text{prox}_{\gamma f_i}(x_k)$, we have

$$\mathcal{C}_\tau \in \mathbb{B}^3\left(\delta = \frac{\mu}{\mu - 4\varepsilon_2 L_{\max} - \frac{n-\tau}{\tau(n-1)}\left[4(2+\varepsilon_2)L_{\max} - 2\mu\right]}\right)$$

Algorithm 2 Inexact FedExProx with τ-nice sampling

Parameters: extrapolation parameter $\alpha_k = \alpha > 0$, step size for the proximal operator $\gamma > 0$, starting point $x_0 \in \mathbb{R}^d$, number of clients n, size of minibatch τ, total number of iterations K, proximal solution accuracy $\varepsilon_2 \geq 0$.
for $k = 0, 1, 2 \ldots K-1$ **do**
  The server broadcasts the current iterate x_k to a selected set of client S_k of size τ
  Each selected client computes a ε approximation of the solution $\tilde{x}_{i,k+1} \simeq \operatorname{prox}_{\gamma f_i}(x_k)$, and sends it back to the server
  The server computes

$$x_{k+1} = x_k + \alpha_k \left(\frac{1}{\tau} \sum_{i \in S_k} \tilde{x}_{i,k+1} - x_k \right). \tag{15}$$

end for

In the case of $\tau = n$, we have $\mathcal{C}_n \in \mathbb{B}^3\left(\frac{\mu}{\mu - 4\varepsilon_2 L_{\max}}\right)$, which recovers the result of (37). When $\tau = 1, \varepsilon_2 = 0$, however, this is problematic, as $\mathcal{C}_1 \in \mathbb{B}^3\left(\delta = \frac{\mu}{3\mu - 8L_{\max}}\right)$. Notice that we require $\delta > 0$, so we require $3\mu > 8L_{\max}$ which only holds in a very restrictive setting. This is due to the stochasticity contained in (16), which arises from client sampling.

Theory of biased SGD: The algorithm does converge, however, and one can use the theory of Demidovich et al. [16] to obtain a convergence guarantee.

Theorem 11 *Assume Assumption 1 (Differentiability), Assumption 2 (Interpolation regime), Assumption 3 (Individual convexity), Assumption 4 (Smoothness) and Assumption 5 (Global strong convexity) hold. Let the approximation $\tilde{x}_{i,k+1}$ all satisfies Definition 4 with $\varepsilon_2 < \frac{\mu^2}{4L_{\max}^2}$, that is*

$$\left\| \tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}(x_k) \right\|^2 \leq \varepsilon_2 \cdot \left\| x_k - \operatorname{prox}_{\gamma f_i}(x_k) \right\|^2,$$

holds for all client i at iteration k. If we are running Algorithm 2 with minibatch size τ and extrapolation parameter $\alpha_k = \alpha > 0$ satisfying

$$\alpha \leq \frac{1}{\gamma L_\gamma} \cdot \frac{\mu - 2\sqrt{\varepsilon_2} L_{\max}}{\mu + 4\varepsilon_2 L_{\max} + 4\sqrt{\varepsilon_2} L_{\max} + \frac{n-\tau}{\tau(n-1)} \cdot \left(4L_{\max} + 4\sqrt{\varepsilon_2} L_{\max} - \mu\right)}$$

Then the iterates generated by Algorithm 2 satisfies

$$\mathbb{E}\left[\mathcal{E}_K\right] \leq \left(1 - \alpha \cdot \frac{\gamma\left(\mu - 2\sqrt{\varepsilon_2} L_{\max}\right)}{4\left(1 + \gamma L_{\max}\right)}\right)^K \mathcal{E}_0. \tag{17}$$

Specifically, if we choose the largest α possible, we have

$$\mathbb{E}\left[\Delta_K\right] \leq \left(1 - \frac{\mu}{4L_\gamma\left(1 + \gamma L_{\max}\right)} \cdot S\left(\varepsilon_2, \tau\right)\right)^K \cdot \frac{L\gamma\left(1 + \gamma L_{\max}\right)}{\mu} \Delta_0,$$

where $S(\varepsilon_2, \tau)$ *is defined as*

$$S(\varepsilon_2, \tau) := \frac{\left(\mu - 2\sqrt{\varepsilon_2} L_{\max}\right)\left(1 - 2\sqrt{\varepsilon_2}\frac{L_{\max}}{\mu}\right)}{\mu + 4\varepsilon_2 L_{\max} + 4\sqrt{\varepsilon_2} L_{\max} + \frac{n-\tau}{\tau(n-1)} \cdot \left(4L_{\max} + 4\sqrt{\varepsilon_2} L_{\max} - \mu\right)},$$

satisfying

$$0 < S(\varepsilon_2, \tau) \leq 1.$$

Notice that we have $S(\varepsilon_2, \tau = n) = S(\varepsilon_2)$, which appears in Theorem 2. For the special case when $\varepsilon_2 = 0$, every proximal operator is solved exactly. The range of α becomes,

$$0 < \alpha \leq \frac{1}{\gamma L_\gamma} \cdot \frac{\mu}{\frac{n-\tau}{\tau(n-1)} \cdot 4L_{\max} + \frac{n(\tau-1)}{\tau(n-1)}\mu}.$$

According to Li et al. [35],

$$0 < \alpha \leq \frac{1}{\gamma L_\gamma} \cdot \frac{L_\gamma(1 + \gamma L_{\max})}{\frac{n-\tau}{\tau(n-1)} L_{\max} + \frac{n(\tau-1)}{\tau(n-1)} \cdot L_\gamma(1 + \gamma L_{\max})}.$$

Clearly the bound we obtain here is suboptimal, since we have $\mu \leq L_\gamma(1 + \gamma L_{\max})$ according to (22). This is due to the previously mentioned issue: the nature of biased compression. When client sampling is used together with biased compressors, it does not necessarily guarantee any benefits. To solve this, the modification of the algorithm itself may be needed, which we consider as a future work direction.

G.2. Absolute approximation in distance

Similarly to Theorem 11, by applying the theory of biased SGD [16], we can derive a convergence guarantee for the minibatch case, though with a suboptimal convergence rate. For brevity and clarity, we do not include the details here.

Appendix H. Proof of theorems and lemmas

H.1. Proof of Lemma 1

Using Fact 8, we have

$$M^\gamma(x) - M^\gamma_{\inf} \geq \frac{\mu}{1 + \gamma L_{\max}} \cdot \frac{1}{2}\|x - x_\star\|^2, \tag{18}$$

where $x \in \mathbb{R}^d$ is any vector, $x_\star$ is a minimizer of M^γ, by Fact 5, it is also a minimizer of f. Since we assume each function f_i is convex, by Fact 3, we know that $M^\gamma_{f_i}$ is also convex. As a result, the average of $M^\gamma_{f_i}$, M^γ is also a convex function. Utilizing the convexity of M^γ, we have,

$$M^\gamma_{\inf} \geq M^\gamma(x) + \langle \nabla M^\gamma(x), x_\star - x\rangle.$$

Rearranging terms we get,

$$\langle \nabla M^\gamma(x), x - x_\star\rangle \geq M^\gamma(x) - M^\gamma_{\inf}. \tag{19}$$

As a result, we have

$$\langle \nabla M^{\gamma}(x), x - x_{\star} \rangle \overset{(18)+(19)}{\geq} \frac{\mu}{1+\gamma L_{\max}} \cdot \frac{1}{2} \|x - x_{\star}\|^2 .$$

Using Cauchy-Schwarz inequality, we have

$$\|\nabla M^{\gamma}(x)\| \|x - x_{\star}\| \geq \langle \nabla M^{\gamma}(x), x - x_{\star} \rangle \geq \frac{\mu}{1+\gamma L_{\max}} \cdot \frac{1}{2} \|x - x_{\star}\|^2 .$$

When $\|x - x_{\star}\| > 0$, the above inequality leads to

$$\|\nabla M^{\gamma}(x)\| \geq \frac{\mu}{2(1+\gamma L_{\max})} \cdot \|x - x_{\star}\| , \tag{20}$$

which also holds when $\|x - x_{\star}\| = 0$. Now using (19) and (20), we obtain

$$\begin{aligned} M^{\gamma}(x) - M^{\gamma}_{\inf} &\overset{(19)}{\leq} \langle \nabla M^{\gamma}(x), x - x_{\star} \rangle \\ &\leq \|\nabla M^{\gamma}(x)\| \|x - x_{\star}\| \\ &\overset{(20)}{\leq} \frac{2(1+\gamma L_{\max})}{\mu} \|\nabla M^{\gamma}(x)\|^2 . \end{aligned}$$

A simple rearranging of terms result in

$$\|\gamma \nabla M^{\gamma}(x)\|^2 \geq 2 \cdot \frac{\gamma\mu}{4(1+\gamma L_{\max})} \left(\gamma M^{\gamma}(x) - \gamma M^{\gamma}_{\inf}\right) .$$

Up till here we have already proved the statement in the lemma, but we want to look at the strongly constant μ of f a little bit. In order to provide an upper bound of μ, we notice that due to Fact 4, each $M^{\gamma}_{f_i}$ is $\frac{L_i}{1+\gamma L_i}$-smooth and therefore M^{γ} is smooth. We use the notation L_{γ} to denote its smoothness constant. Applying the smoothness of $M^{\gamma}(x)$, we have

$$M^{\gamma}(x) \leq M^{\gamma}(x_{\star}) + \langle \nabla M^{\gamma}(x_{\star}), x - x_{\star} \rangle + \frac{L_{\gamma}}{2} \|x - x^{\star}\|^2 .$$

Utilizing the fact that $\nabla M^{\gamma}(x_{\star}) = 0$, we have

$$M^{\gamma}(x) - M^{\gamma}_{\inf} \leq \frac{L_{\gamma}}{2} \|x - x_{\star}\|^2 \tag{21}$$

Combining (21) and (18), we can deduce that

$$\frac{\mu}{1+\gamma L_{\max}} \cdot \frac{1}{2} \|x - x_{\star}\|^2 \leq M^{\gamma}(x) - M^{\gamma}_{\inf} \leq \frac{L_{\gamma}}{2} \|x - x_{\star}\|^2 .$$

which results in the estimate that

$$\mu \leq L_{\gamma}(1+\gamma L_{\max}) . \tag{22}$$

H.2. Proof of Theorem 1

Let us first recall that after reformulation, Algorithm 1 can be written as

$$x_{k+1} = x_k - \alpha \cdot g(x_k),$$

where $g(x_k)$ is defined as

$$g(x_k) := \frac{1}{n}\sum_{i=1}^{n} \gamma \nabla M^{\gamma}_{f_i}(x_k) - \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{x}_{i,k+1} - \mathrm{prox}_{\gamma f_i}(x_k)\right).$$

We view this as running full batch biased SGD with stepsize α and global objective $\gamma M^{\gamma}(x)$. We first examine if Assumption 6 (Biased-ABC) holds for arbitrary x_k. Since we are in the full batch case, it is easy to see that

$$\mathbb{E}\left[g(x_k)\right] = g(x_k).$$

Since our objective now is $\gamma M^{\gamma}(x)$, we have that

$$\begin{aligned}
\left\langle \gamma \nabla M^{\gamma}(x_k), g(x_k)\right\rangle &= \left\langle \gamma \nabla M^{\gamma}(x_k), \gamma \nabla M^{\gamma}(x_k) - \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{x}_{i,k+1} - \mathrm{prox}_{\gamma f_i}(x_k)\right)\right\rangle \\
&= \left\|\gamma \nabla M^{\gamma}(x_k)\right\|^2 - \underbrace{\left\langle \gamma \nabla M^{\gamma}(x_k), \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{x}_{i,k+1} - \mathrm{prox}_{\gamma f_i}(x_k)\right)\right\rangle}_{:=P_1}.
\end{aligned}$$

Now let us focus on P_1, we have the following upper bound,

$$\begin{aligned}
P_1 &\leq \frac{1}{2}\left\|\gamma \nabla M^{\gamma}(x_k)\right\|^2 + \frac{1}{2}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{x}_{i,k+1} - \mathrm{prox}_{\gamma f_i}(x_k)\right)\right\|^2 \\
&\overset{(5)}{\leq} \frac{1}{2}\left\|\gamma \nabla M^{\gamma}(x_k)\right\|^2 + \frac{\varepsilon_1}{2}.
\end{aligned}$$

As a result, we have

$$\left\langle \gamma \nabla M^{\gamma}(x_k), g(x_k)\right\rangle \geq \frac{1}{2}\left\|\gamma \nabla M^{\gamma}(x_k)\right\| - \frac{\varepsilon_1}{2},$$

which holds for arbitrary x_k. This suggests that $b = \frac{1}{2}, c = \frac{\varepsilon_1}{2}$. On the other hand,

$$\begin{aligned}
\mathbb{E}\left[\left\|g(x_k)\right\|^2\right] &= \left\|\gamma \nabla M^{\gamma}(x_k) + \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{x}_{i,k+1} - \mathrm{prox}_{\gamma f_i}(x_k)\right)\right\|^2 \\
&\overset{(9)}{\leq} 2\left\|\gamma \nabla M^{\gamma}(x_k)\right\|^2 + 2\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{x}_{i,k+1} - \mathrm{prox}_{\gamma f_i}(x_k)\right)\right\|^2 \\
&\overset{(5)}{\leq} 2\left\|\gamma \nabla M^{\gamma}(x_k)\right\|^2 + 2\varepsilon_1.
\end{aligned}$$

Thus, we can choose $A = 0, B = 2, C = 2\varepsilon_1$. Since we have assumed Assumption 3 (Individual convexity) and Assumption 4 (Smoothness), it is easy to see that M^γ is smooth, and we denote its smoothness constant as L_γ. It is therefore straightforward to see that our global objective γM^γ is γL_γ-smooth. We also assume f is μ-strongly convex, which by Fact 8 indicates that M^γ is $\frac{\mu}{1+\gamma L_{\max}}$ star strongly convex. We immediately obtain using Lemma 1 that γM^γ is $\frac{\gamma\mu}{4(1+\gamma L_{\max})}$-PL. Now, we have validated all the assumptions for using Theorem 4. Applying Theorem 4, we obtain that when the extrapolation parameter satisfies

$$0 < \alpha < \frac{1}{4} \cdot \min\left\{\frac{1}{\gamma L_\gamma}, \frac{2\left(1+\gamma L_{\max}\right)}{\gamma\mu}\right\},$$

the last iterate x_K of Algorithm 1 with each proximal operator solved inexactly according to Definition 1 satisfies

$$\mathcal{E}_K \le \left(1 - \frac{\alpha\gamma\mu}{8\left(1+\gamma L_{\max}\right)}\right)^K \mathcal{E}_0 + \frac{8\varepsilon_1\alpha L_\gamma\left(1+\gamma L_{\max}\right)}{\mu} + \frac{4\varepsilon_1\left(1+\gamma L_{\max}\right)}{\gamma\mu},$$

where $\mathcal{E}_k = \gamma M^\gamma\left(x_k\right) - M^\gamma_{\text{inf}}$. Let us now prove that

$$\frac{1}{\gamma L_\gamma} < \frac{2\left(1+\gamma L_{\max}\right)}{\gamma\mu}.$$

This is equivalent to prove

$$\mu < 2L_\gamma\left(1+\gamma L_{\max}\right),$$

which is always true since (22) holds. As a result, we can simplify the range of the extrapolation parameter to

$$0 < \alpha \le \frac{1}{4\gamma L_\gamma}.$$

If we pick the largest possible α, we have

$$\mathcal{E}_K \le \left(1 - \frac{\mu}{32 L_\gamma\left(1+\gamma L_{\max}\right)}\right)^K \mathcal{E}_0 + \frac{6\varepsilon_1\left(1+\gamma L_{\max}\right)}{\gamma\mu}.$$

This result is not directly comparable to that of Li et al. [35]. However, using smoothness of γL_γ, if we denote $\Delta_k = \left\|x_k - x_\star\right\|^2$ where $x_\star$ is a minimizer of both M^γ and f since we assume we are in the interpolation regime (Assumption 2), we have

$$\mathcal{E}_0 \le \frac{\gamma L_\gamma}{2}\Delta_0.$$

Using star strong convexity, we have

$$\mathcal{E}_K \ge \frac{\gamma\mu}{2\left(1+\gamma L_{\max}\right)}\Delta_K.$$

As a result, we can transform the above convergence guarantee into

$$\Delta_K \le \left(1 - \frac{\mu}{32 L_\gamma\left(1+\gamma L_{\max}\right)}\right)^K \frac{L_\gamma\left(1+\gamma L_{\max}\right)}{\mu} \cdot \Delta_0 + 12\varepsilon_1 \cdot \left(\frac{1/\gamma + L_{\max}}{\mu}\right)^2.$$

This completes the proof.

H.3. Proof of Theorem 2

Since we based our analysis on the theory of biased SGD, we first verify the validity of Assumption 6.

Finding b and c: Let us start with finding a lower bound on $\langle \gamma \nabla M^{\gamma}(x_k), \mathbb{E}[g(x_k)]\rangle$. We have

$$\begin{aligned}\langle \gamma M^{\gamma}(x_k), \mathbb{E}[g(x_k)]\rangle &= \left\langle \gamma M^{\gamma}(x_k), \gamma M^{\gamma}(x_k) - \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{x}_{i,k+1} - \mathrm{prox}_{\gamma f_i}(x_k)\right)\right\rangle \\ &= \|\gamma M^{\gamma}(x_k)\|^2 - \left\langle \gamma M^{\gamma}(x_k), \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{x}_{i,k+1} - \mathrm{prox}_{\gamma f_i}(x_k)\right)\right\rangle \\ &\geq \|\gamma M^{\gamma}(x_k)\|^2 - \|\gamma M^{\gamma}(x_k)\| \cdot \left\|\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{x}_{i,k+1} - \mathrm{prox}_{\gamma f_i}(x_k)\right)\right\|,\end{aligned}$$

where the last inequality is obtained using Cauchy-Schwarz inequality. We then utilize the convexity of $\|\cdot\|$ and obtain,

$$\begin{aligned}\langle \gamma M^{\gamma}(x_k), \mathbb{E}[g(x_k)]\rangle &\geq \|\gamma M^{\gamma}(x_k)\|^2 - \|\gamma M^{\gamma}(x_k)\| \cdot \frac{1}{n}\sum_{i=1}^{n}\left\|\left(\tilde{x}_{i,k+1} - \mathrm{prox}_{\gamma f_i}(x_k)\right)\right\| \\ &\overset{(7)}{\geq} \|\gamma M^{\gamma}(x_k)\|^2 - \sqrt{\varepsilon_2}\,\|\gamma M^{\gamma}(x_k)\| \cdot \frac{1}{n}\sum_{i=1}^{n}\left\|x_k - \mathrm{prox}_{\gamma f_i}(x_k)\right\| \\ &= \|\gamma M^{\gamma}(x_k)\|^2 - \sqrt{\varepsilon_2}\,\|\gamma M^{\gamma}(x_k)\| \cdot \frac{1}{n}\sum_{i=1}^{n}\left\|\gamma \nabla M^{\gamma}_{f_i}(x_k)\right\|.\end{aligned}$$

Notice that

$$\left\|\gamma \nabla M^{\gamma}_{f_i}(x_k)\right\| = \left\|\gamma \nabla M^{\gamma}_{f_i}(x_k) - \gamma \nabla M^{\gamma}_{f_i}(x_\star)\right\|,$$

holds for any $x_\star$ that is a minimizer of $M^{\gamma}(x)$ due to interpolation regime assumption. As a result, we can provide an upper bound based on smoothness of each individual $\gamma M^{\gamma}_{f_i}(x)$ using Fact 2,

$$\left\|\gamma \nabla M^{\gamma}_{f_i}(x_k) - \gamma \nabla M^{\gamma}_{f_i}(x_\star)\right\| \leq \frac{\gamma L_i}{1+\gamma L_i}\|x_k - x_\star\|. \tag{23}$$

Thus,

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\gamma \nabla M^{\gamma}_{f_i}(x_k)\right\| \leq \frac{1}{n}\sum_{i=1}^{n}\frac{\gamma L_i}{1+\gamma L_i}\|x_k - x_\star\| \leq \frac{\gamma L_{\max}}{1+\gamma L_{\max}} \cdot \|x_k - x_\star\|.$$

In addition, we have due to Cauchy-Schwarz inequality and the convexity of $M^{\gamma}(x)$

$$\|\nabla M^{\gamma}(x_k)\| \cdot \|x_k - x_\star\| \geq \langle \nabla M^{\gamma}(x_k), x_k - x_\star\rangle \geq M^{\gamma}(x_k) - M^{\gamma}_{\inf}, \tag{24}$$

and due to quadratic growth condition that

$$M^{\gamma}(x_k) - M^{\gamma}_{\inf} \geq \frac{\mu}{1+\gamma L_{\max}} \cdot \frac{1}{2}\|x_k - x_\star\|^2. \tag{25}$$

Combining (24) and (25), we have

$$\frac{\mu}{2\left(1+\gamma L_{\max}\right)} \cdot \left\|x_k - x_\star\right\|^2 \overset{(24)+(25)}{\leq} \left\|\nabla M^\gamma\left(x_k\right)\right\| \cdot \left\|x_k - x_\star\right\|.$$

This indicates that

$$\left\|x_k - x_\star\right\| \leq \frac{2\left(1+\gamma L_{\max}\right)}{\mu}\left\|\nabla M^\gamma\left(x_k\right)\right\|. \tag{26}$$

Combining (23) and (26), we generate the following lower bound

$$\begin{aligned}\left\langle \gamma M^\gamma\left(x_k\right), \mathbb{E}\left[g(x_k)\right]\right\rangle &\overset{(23)}{\geq} \left\|\gamma M^\gamma\left(x_k\right)\right\|^2 - \sqrt{\varepsilon_2}\left\|\gamma M^\gamma\left(x_k\right)\right\| \cdot \frac{\gamma L_{\max}}{1+\gamma L_{\max}}\left\|x_k - x_\star\right\| \\ &\overset{(26)}{\geq} \left\|\gamma M^\gamma\left(x_k\right)\right\|^2 - \sqrt{\varepsilon_2} \cdot \frac{L_{\max}}{1+\gamma L_{\max}} \cdot \frac{2\left(1+\gamma L_{\max}\right)}{\mu}\left\|\gamma M^\gamma\left(x_k\right)\right\|^2 \\ &= \left(1 - \sqrt{\varepsilon_2} \cdot \frac{2L_{\max}}{\mu}\right) \cdot \left\|\gamma M^\gamma\left(x_k\right)\right\|^2.\end{aligned}$$

Thus, as long as $\varepsilon_2 < \frac{\mu^2}{4L_{\max}^2}$, we have $b = 1 - \sqrt{\varepsilon_2} \cdot \frac{2L_{\max}}{\mu}$, and $c = 0$.

Finding A, B and C: We start with expanding $\left\|g(x_k)\right\|^2$,

$$\begin{aligned}\mathbb{E}\left[\left\|g(x_k)\right\|^2\right] &= \left\|\gamma M^\gamma\left(x_k\right) - \frac{1}{n}\sum_{i=1}^n\left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}\left(x_k\right)\right)\right\|^2 \\ &= \left\|\gamma M^\gamma\left(x_k\right)\right\|^2 + \underbrace{\left\|\frac{1}{n}\sum_{i=1}^n\left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}\left(x_k\right)\right)\right\|^2}_{:=T_2} \\ &\quad \underbrace{-2\left\langle \gamma M^\gamma\left(x_k\right), \frac{1}{n}\sum_{i=1}^n\left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}\left(x_k\right)\right)\right\rangle}_{:=T_3}.\end{aligned} \tag{27}$$

It is easy to bound T_2 utilizing the convexity of $\left\|\cdot\right\|^2$,

$$\begin{aligned}T_2 &\leq \frac{1}{n}\sum_{i=1}^n\left\|\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}\left(x_k\right)\right\|^2 \\ &\overset{(7)}{\leq} \frac{\varepsilon_2}{n}\sum_{i=1}^n\left\|x_k - \operatorname{prox}_{\gamma f_i}\left(x_k\right)\right\|^2 = \frac{\varepsilon_2}{n}\sum_{i=1}^n\left\|\gamma M_{f_i}^\gamma\left(x_k\right)\right\|^2.\end{aligned}$$

Let $x_\star$ be a minimizer of M^γ, since we assume Assumption 2 holds, it is also a minimizer of each $M_{f_i}^\gamma$. As a result,

$$\begin{aligned}T_2 &\leq \frac{\varepsilon_2}{n}\sum_{i=1}^n\left\|\gamma M_{f_i}^\gamma\left(x_k\right) - \gamma M_{f_i}^\gamma\left(x_\star\right)\right\|^2 \\ &\leq \frac{\varepsilon_2}{n}\sum_{i=1}^n\frac{2\gamma L_i}{1+\gamma L_i}\left(\gamma M_{f_i}^\gamma\left(x_k\right) - \gamma M_{f_i}^\gamma\left(x_\star\right)\right) \leq \frac{2\varepsilon_2\gamma L_{\max}}{1+\gamma L_{\max}} \cdot \left(\gamma M^\gamma\left(x_k\right) - \gamma M_{\inf}^\gamma\right).\end{aligned} \tag{28}$$

We then consider T_3, and start with applying Cauchy-Schwarz inequality

$$T_3 \leq 2 \left\| \gamma \nabla M^{\gamma}\left(x_k\right) \right\| \left\| \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}\left(x_k\right) \right) \right\|. \tag{29}$$

Using the convexity of $\|\cdot\|$, we have

$$\begin{aligned}
\left\| \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}\left(x_k\right) \right) \right\| &\leq \frac{1}{n} \sum_{i=1}^{n} \left\| \tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}\left(x_k\right) \right\| \\
&\overset{(7)}{\leq} \frac{\sqrt{\varepsilon_2}}{n} \sum_{i=1}^{n} \left\| x_k - \operatorname{prox}_{\gamma f_i}\left(x_k\right) \right\| \\
&\overset{(2)}{=} \frac{\sqrt{\varepsilon_2}}{n} \sum_{i=1}^{n} \left\| \gamma \nabla M_{f_i}^{\gamma}\left(x_k\right) - \gamma \nabla M_{f_i}^{\gamma}\left(x_\star\right) \right\| \\
&\overset{\text{Fact 2}}{\leq} \frac{\sqrt{\varepsilon_2}}{n} \sum_{i=1}^{n} \frac{\gamma L_i}{1+\gamma L_i} \left\| x_k - x_\star \right\| \\
&\leq \frac{\sqrt{\varepsilon_2}\gamma L_{\max}}{1+\gamma L_{\max}} \cdot \left\| x_k - x_\star \right\|.
\end{aligned}$$

Utilizing (26), we have

$$\begin{aligned}
\left\| \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}\left(x_k\right) \right) \right\| &\leq \frac{\sqrt{\varepsilon_2}\gamma L_{\max}}{1+\gamma L_{\max}} \cdot \frac{2\left(1+\gamma L_{\max}\right)}{\mu} \left\| \nabla M^{\gamma}\left(x_k\right) \right\| \\
&= \frac{2\sqrt{\varepsilon_2} L_{\max}}{\mu} \cdot \left\| \gamma \nabla M^{\gamma}\left(x_k\right) \right\|.
\end{aligned} \tag{30}$$

Plug the above inequality into (29), we have

$$T_3 \leq \frac{4\sqrt{\varepsilon_2} L_{\max}}{\mu} \cdot \left\| \gamma \nabla M^{\gamma}\left(x_k\right) \right\|^2. \tag{31}$$

Combining (31) and (28), plug them into (27), we have

$$\mathbb{E}\left[\left\| g\left(x_k\right) \right\|^2 \right] \leq \frac{2\varepsilon_2 \gamma L_{\max}}{1+\gamma L_{\max}} \cdot \left(\gamma M^{\gamma}\left(x_k\right) - \gamma M_{\inf}^{\gamma} \right) + \left(1 + \frac{4\sqrt{\varepsilon_2} L_{\max}}{\mu} \right) \cdot \left\| \gamma \nabla M^{\gamma}\left(x_k\right) \right\|^2.$$

Thus, we have

$$A = \frac{\varepsilon_2 \gamma L_{\max}}{1+\gamma L_{\max}}, \quad B = \frac{\mu + 4\sqrt{\varepsilon_2} L_{\max}}{\mu}, \quad C = 0.$$

Applying Theorem 4: First, we list our the values appeared respectively,

$$\begin{aligned}
&A = \frac{\varepsilon_2 \gamma L_{\max}}{1+\gamma L_{\max}}, \quad B = \frac{\mu + 4\sqrt{\varepsilon_2} L_{\max}}{\mu}, \quad b = \frac{\mu - 2\sqrt{\varepsilon_2} L_{\max}}{\mu}, \\
&C = c = 0.
\end{aligned}$$

We know that the PL constant of γM^γ is given by $\frac{\gamma\mu}{4(1+\gamma L_{\max})}$ and the corresponding smoothness constant is γL_γ. Applying Theorem 4, the range of α is given by

$$0 < \alpha < \min\left\{\underbrace{\frac{1}{\gamma L_\gamma}\cdot\frac{\mu - 2\sqrt{\varepsilon_2}L_{\max}}{\mu + 4\sqrt{\varepsilon_2}L_{\max} + 4\varepsilon_2 L_{\max}}}_{:=B_1}, \underbrace{\frac{4\left(1+\gamma L_{\max}\right)}{\gamma\left(\mu - 2\sqrt{\varepsilon_2}L_{\max}\right)}}_{:=B_2}\right\}. \tag{32}$$

Now notice that actually we can prove that for $\varepsilon_2 < \frac{\mu^2}{4L_{\max}^2}$, we have $B_2 > B_1$, and we can simplify the range of α to

$$0 < \alpha \leq \frac{1}{\gamma L_\gamma}\cdot\frac{\mu - 2\sqrt{\varepsilon_2}L_{\max}}{\mu + 4\sqrt{\varepsilon_2}L_{\max} + 4\varepsilon_2 L_{\max}}.$$

Proof of $B_2 > B_1$: It is easy to verify that the above inequality ($B_2 > B_1$) can be equivalently written as

$$4L_\gamma\left(1+\gamma L_{\max}\right)\left(\mu + 4\sqrt{\varepsilon_2}L_{\max} + 4\varepsilon_2 L_{\max}\right) > \left(\mu - 2\sqrt{\varepsilon_2}L_{\max}\right)^2,$$

since when $\sqrt{\varepsilon_2} < \frac{\mu}{2L_{\max}}$, we have $\mu - 2\sqrt{\varepsilon_2}L_{\max} > 0$. We expand the right-hand side and obtain:

$$\left(\mu - 2\sqrt{\varepsilon_2}L_{\max}\right)^2 = \mu^2 - 4\sqrt{\varepsilon_2}L_{\max} + 4\varepsilon_2 L_{\max}^2 < 2\mu^2 - 4\sqrt{\varepsilon_2}L_{\max} < 2\mu^2.$$

For the left-hand side, as we have already shown in 22, we have

$$4L_\gamma\left(1+\gamma L_{\max}\right)\left(\mu + 4\sqrt{\varepsilon_2}L_{\max} + 4\varepsilon_2 L_{\max}\right) \geq 4\mu\left(\mu + 4\sqrt{\varepsilon_2}L_{\max} + 2\varepsilon_2 L_{\max}\right) > 4\mu^2.$$

Combining the above inequality we arrive at $B_2 > B_1$.

The convergence guarantee : Given that we select α properly, we have

$$\mathcal{E}_K \leq \left(1 - \alpha\cdot\frac{\gamma\left(\mu - 2\sqrt{\varepsilon_2}L_{\max}\right)}{4\left(1+\gamma L_{\max}\right)}\right)^K \mathcal{E}_0,$$

where $\mathcal{E}_k = \gamma M^\gamma\left(x_k\right) - \gamma M^\gamma_{\inf}$. We do not have expectation here since we are in the full batch case. Specifically, if we choose the largest α possible, we have

$$\mathcal{E}_K \leq \left(1 - \frac{\mu}{4L_\gamma\left(1+\gamma L_{\max}\right)}\cdot S\left(\varepsilon_2\right)\right)^k \mathcal{E}_0,$$

where

$$S(\varepsilon_2) = \frac{\left(\mu - 2\sqrt{\varepsilon_2}L_{\max}\right)\left(1 - 2\sqrt{\varepsilon_2}\frac{L_{\max}}{\mu}\right)}{\mu + 4\sqrt{\varepsilon_2}L_{\max} + 4\varepsilon_2 L_{\max}},$$

satisfies $0 < S(\varepsilon_2) \leq 1$ is the factor of slowing down due to inexact proximity operator evaluation. Using smoothness of γL_γ, if we denote $\Delta_k = \|x_k - x_\star\|^2$ where $x_\star$ is a minimizer of both M^γ and f since we assume we are in the interpolation regime (Assumption 2), we have

$$\mathcal{E}_0 \leq \frac{\gamma L_\gamma}{2}\Delta_0.$$

Using star strong convexity (quadratic growth property), we have

$$\mathcal{E}_K \geq \frac{\gamma\mu}{2\left(1+\gamma L_{\max}\right)}\Delta_K.$$

As a result, we can transform the above convergence guarantee into

$$\Delta_K \leq \left(1 - \frac{\mu}{4L_\gamma\left(1+\gamma L_{\max}\right)} \cdot S\left(\varepsilon_2\right)\right)^K \cdot \frac{L\gamma\left(1+\gamma L_{\max}\right)}{\mu}\Delta_0.$$

This completes the proof.

H.4. Proof of Theorem 3

We start with formalizing the problem. We can write the update rule of Algorithm 1 as

$$x_{k+1} = x_k - \alpha \cdot \left(\frac{1}{n}\sum_{i=1}^{n}\gamma\nabla M_{f_i}^{\gamma}\left(x_k\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}\left(x_k\right)\right)\right) \tag{33}$$

Since by Definition 4, we have $\left\|\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}\left(x_k\right)\right\|^2 \leq \varepsilon_2 \left\|\gamma\nabla M_{f_i}^{\gamma}\left(x_k\right)\right\|^2$, we can view the left hand side as a compressed version of the true gradient. Specifically, there are two possible perspectives:

(I). Let $\mathcal{C}_i\left(\cdot\right)$ be the compressing mapping with the i-th client, $i \in \{1, 2, \ldots, n\}$, defined as

$$\mathcal{C}_i\left(\gamma\nabla M_{f_i}^{\gamma}\left(x_k\right)\right) := \gamma\nabla M_{f_i}^{\gamma}\left(x_k\right) - \left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}\left(x_k\right)\right).$$

In this way, we reformulate (33) as

$$x_{k+1} = x_k - \alpha \cdot \frac{1}{n}\sum_{i=1}^{n}\mathcal{C}_i\left(\gamma\nabla M_{f_i}^{\gamma}\left(x_k\right)\right). \tag{34}$$

(34) is exactly DCGD with biased compression. We can easily prove that

$$\begin{aligned}
&\mathcal{C}_i \in \mathbb{B}^1\left(\alpha = 1 - 2\sqrt{\varepsilon_2}, \beta = \frac{1-\sqrt{\varepsilon_2}}{1+\varepsilon_2}\right)\\
&\mathcal{C}_i \in \mathbb{B}^2\left(\xi = 1 - \sqrt{\varepsilon_2}, \beta = \frac{1-\sqrt{\varepsilon_2}}{1+\varepsilon_2}\right)\\
&\mathcal{C}_i \in \mathbb{B}^3\left(\delta = \frac{1}{1-\varepsilon_2}\right).
\end{aligned}$$

However, DCGD with biased compression may fail to converge even if the above formulation of compression mapping seems quite nice. For an example of such failure, we refer the readers to Beznosikov et al. [9, Example 1]. This limitation can be circumvented by employing an error feedback mechanism; however, this approach requires modifications to the original algorithm. We therefore leave it as a future research direction.

(II). We can also view it as if we are in the single node case. Let $\mathcal{C}(\cdot)$ be the compressing mapping defined as

$$\begin{aligned}\mathcal{C}\left(\nabla M^{\gamma}\left(x_k\right)\right) &:= \frac{1}{n}\sum_{i=1}^{n}\gamma\nabla M_{f_i}^{\gamma}\left(x_k\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}\left(x_k\right)\right) \\ &= \gamma\nabla M^{\gamma}\left(x_k\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}\left(x_k\right)\right). \end{aligned} \tag{35}$$

This formulation leads us to the convergence guarantee appeared in Theorem 3, as we illustrate below.

Let us first analyze $\mathcal{C}$ defined in (35). We will verify it belongs to $\mathbb{B}^3(\delta)$. The inequality we want to prove can be written equivalently as

$$\left\|\gamma\nabla M^{\gamma}\left(x_k\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}\left(x_k\right)\right) - \gamma\nabla M^{\gamma}\left(x_k\right)\right\|^2 \leq \left(1-\frac{1}{\delta}\right)\left\|\gamma\nabla M^{\gamma}\left(x_k\right)\right\|^2, \tag{36}$$

which is exactly

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}\left(x_k\right)\right)\right\|^2 \leq \left\|\gamma\nabla M^{\gamma}\left(x_k\right)\right\|^2$$

For the left-hand side, using the convexity of $\|\cdot\|^2$ in combination with Definition 4, we obtain

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}\left(x_k\right)\right)\right\|^2 &\leq \frac{1}{n}\sum_{i=1}^{n}\left\|\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}\left(x_k\right)\right\|^2 \\ &\leq \frac{\varepsilon_2}{n}\sum_{i=1}^{n}\left\|x_k - \operatorname{prox}_{\gamma f_i}\left(x_k\right)\right\|^2.\end{aligned}$$

Let $x_\star$ be a minimizer of f, since we assume Assumption 2 holds, by Fact 7, it is also a minimizer of γM^{γ},

$$\begin{aligned}\frac{\varepsilon_2}{n}\sum_{i=1}^{n}\left\|x_k - \operatorname{prox}_{\gamma f_i}\left(x_k\right)\right\|^2 &\overset{(2)}{=} \frac{\varepsilon_2}{n}\sum_{i=1}^{n}\left\|\gamma\nabla M_{f_i}^{\gamma}\left(x_k\right)\right\|^2 \\ &= \frac{\varepsilon_2}{n}\sum_{i=1}^{n}\left\|\gamma\nabla M_{f_i}^{\gamma}\left(x_k\right) - \gamma\nabla M_{f_i}^{\gamma}\left(x_\star\right)\right\|^2 \\ &\overset{\text{Fact 2}}{\leq} \frac{2\varepsilon_2}{n}\sum_{i=1}^{n}\frac{\gamma L_i}{1+\gamma L_i}\left(\gamma M_{f_i}^{\gamma}\left(x_k\right) - \gamma M_{f_i}^{\gamma}\left(x_\star\right)\right) \\ &\leq \frac{2\varepsilon_2\gamma L_{\max}}{1+\gamma L_{\max}}\left(\gamma M^{\gamma}\left(x_k\right) - \gamma M^{\gamma}\left(x_\star\right)\right).\end{aligned}$$

We then notice that as it is illustrated by Lemma 1, we have

$$\left(1-\frac{1}{\delta}\right)\left\|\gamma\nabla M^{\gamma}\left(x_k\right)\right\|^2 \geq \left(1-\frac{1}{\delta}\right)\frac{\gamma\mu}{2\left(1+\gamma L_{\max}\right)}\left(\gamma M^{\gamma}\left(x_k\right) - \gamma M^{\gamma}\left(x_\star\right)\right).$$

Combining the above two inequalities, we know that the following inequality is a sufficient condition for (36),

$$\frac{2\varepsilon_2\gamma L_{\max}}{1+\gamma L_{\max}}\left(\gamma M^{\gamma}\left(x_k\right)-\gamma M^{\gamma}\left(x_\star\right)\right)\leq\left(1-\frac{1}{\delta}\right)\frac{\gamma\mu}{2\left(1+\gamma L_{\max}\right)}\left(\gamma M^{\gamma}\left(x_k\right)-\gamma M^{\gamma}\left(x_\star\right)\right).$$

It is easy to check that if we pick

$$\delta=\frac{\mu}{\mu-4\varepsilon_2 L_{\max}}>0, \tag{37}$$

the condition is met. However, for this to hold, we must ensure that $\varepsilon_2<\frac{\mu}{4L_{\max}}$.

As we mentioned in Appendix F, Beznosikov et al. [9] provided the theory of CGD with biased compressor belongs to $\mathbb{B}^3\left(\delta\right)$. We have already shown that $\mathcal{C}\in\mathbb{B}^3\left(\delta=\frac{\mu}{\mu-4\varepsilon_2 L_{\max}}\right)$, when $\varepsilon_2<\frac{4L_{\max}}{\mu}$. Notice that our objective γM^{γ} is γL_{γ}-smooth and $\frac{\gamma\mu}{1+\gamma L_{\max}}$-PL.[5] Therefore, as long as $0<\alpha\leq\frac{1}{\gamma L_{\gamma}}$ and $\varepsilon_2<\frac{\mu}{4L_{\max}}$, we have

$$\mathcal{E}_K\leq\left(1-\frac{\mu-4\varepsilon_2 L_{\max}}{\mu}\cdot\frac{\gamma\mu}{4\left(1+\gamma L_{\max}\right)}\cdot\alpha\right)^K\mathcal{E}_0,$$

Taking $\alpha=\frac{1}{\gamma L_{\gamma}}$, which is the largest step size possible, we can further simplify the above convergence into

$$M^{\gamma}\left(x_k\right)-M^{\gamma}_{\star}\leq\left(1-\left(1-\frac{4\varepsilon_2 L_{\max}}{\mu}\right)\cdot\frac{\mu}{4L_{\gamma}\left(1+\gamma L_{\max}\right)}\right)^K\left(M^{\gamma}\left(x_0\right)-M^{\gamma_\star}\right).$$

Using smoothness of γL_{γ}, if we denote $\Delta_k=\|x_k-x_\star\|^2$ where $x_\star$ is a minimizer of both M^{γ} and f since we assume we are in the interpolation regime (Assumption 2), we have

$$\mathcal{E}_0\leq\frac{\gamma L_{\gamma}}{2}\Delta_0.$$

Using star strong convexity (quadratic growth property), we have

$$\mathcal{E}_K\geq\frac{\gamma\mu}{2\left(1+\gamma L_{\max}\right)}\Delta_K.$$

As a result, we can transform the above convergence guarantee into

$$\Delta_K\leq\left(1-\left(1-\frac{4\varepsilon_2 L_{\max}}{\mu}\right)\cdot\frac{\mu}{4L_{\gamma}\left(1+\gamma L_{\max}\right)}\right)^K\cdot\frac{L_{\gamma}\left(1+\gamma L_{\max}\right)}{\mu}\Delta_0.$$

This completes the proof.

H.5. Proof of Theorem 5

Notice that we assume each f_i is L_i-smooth and convex. The local optimization of each client can be written as

$$\min_{z\in\mathbb{R}^d}\left\{A^{\gamma}_{k,i}\left(z\right)=f_i\left(z\right)+\frac{1}{2\gamma}\left\|z-x_k\right\|^2\right\},$$

It is easy to see that $A^{\gamma}_{k,i}\left(z\right)$ is $L_i+\frac{1}{\gamma}$-smooth and $\frac{1}{\gamma}$-strongly convex. We first provide the convergence theory of GD for reference.

5. Theorem 10 remains valid if we replace f being strongly convex with PL.

Theory of GD: For a $\widehat{\mu}$-strongly convex, $\widehat{L}$-smooth function ϕ, the algorithm can be formulated as

$$z_{t+1} = z_t - \eta \nabla \phi(z_t), \tag{GD}$$

where z_t is the iterate in the t-th iteration, and $\eta > 0$ is the step size. GD with step size $\eta \in (0, \frac{1}{\widehat{L}}]$ generates iterates that satisfy

$$\|z_t - z_\star\|^2 \leq (1 - \eta\widehat{\mu})^t \|z_0 - z_\star\|^2,$$

where $z_\star$ is a minimizer of ϕ, t is the number of iterations (number of gradient evaluations).

Approximation satisfying Definition 3: Notice that $\mathrm{prox}_{\gamma f_i}(x_k)$ is the minimizer of $A_{k,i}^{\gamma}(z)$ and $z_0 = x_k$. As a result, if we run GD with the largest step size $\frac{\gamma}{1+\gamma L_i}$,

$$\left\|z_t - \mathrm{prox}_{\gamma f_i}(x_k)\right\|^2 \leq \left(1 - \frac{1}{1+\gamma L_i}\right)^t \left\|x_k - \mathrm{prox}_{\gamma f_i}(x_k)\right\|^2 \tag{38}$$

We have

$$t = \mathcal{O}\left((1+\gamma L_i)\log\left(\frac{\left\|x_k - \mathrm{prox}_{\gamma f_i}(x_k)\right\|^2}{\varepsilon_1}\right)\right).$$

The unknown term $\left\|x_k - \mathrm{prox}_{\gamma f_i}(x_k)\right\|^2$ within the log can be bounded by

$$\begin{aligned}\left\|x_k - \mathrm{prox}_{\gamma f_i}(x_k)\right\|^2 &= \|z_0 - z_\star\|^2 \\ &\leq \gamma^2 \left\|\nabla A_{k,i}^{\gamma}(z_0) - \nabla A_{k,i}^{\gamma}(z_\star)\right\|^2 = \|\gamma \nabla f_i(x_k)\|^2,\end{aligned} \tag{39}$$

which can be easily calculated.

Approximation satisfying Definition 4: According to (38), we have

$$t = \mathcal{O}\left((1+\gamma L_i)\log\left(\frac{1}{\varepsilon_2}\right)\right).$$

This completes the proof.

H.6. Proof of Theorem 6

We first provide the theory of AGD [49].

Theory of AGD: For a $\widehat{\mu}$-strongly convex, $\widehat{L}$-smooth function ϕ, the algorithm can be formulated as

$$\begin{aligned} y_{t+1} &= z_t + \alpha(z_t - z_{t-1}) \\ z_{t+1} &= y_{t+1} - \eta\nabla\phi(y_{t+1}), \end{aligned} \tag{AGD}$$

where z_t, y_t are iterates, $\eta > 0$ is the step size, $\alpha > 0$ is the momentum parameter. AGD with step size $\eta = \frac{1}{\widehat{L}}$, momentum $\alpha = \frac{\sqrt{\widehat{L}}-\sqrt{\widehat{\mu}}}{\sqrt{\widehat{L}}+\sqrt{\widehat{\mu}}}$ generates iterates that satisfy

$$\|z_t - z_\star\|^2 \leq \frac{2\widehat{L}}{\widehat{\mu}} \cdot \left(1 - \sqrt{\frac{\widehat{\mu}}{\widehat{L}}}\right)^t \|z_0 - z_\star\|^2,$$

where $z_\star$ is a minimizer of ϕ, t is the number of iterations (number of gradient evaluations).

Approximation satisfying Definition 3: Notice that $\mathrm{prox}_{\gamma f_i}(x_k)$ is the minimizer of $A_{k,i}^{\gamma}(z)$ and $z_0 = x_k$. As a result, if we run AGD with the step size $\frac{\gamma}{1+\gamma L_i}$ and momentum $\alpha = \frac{\sqrt{1+\gamma L_i}-1}{\sqrt{1+\gamma L_i}+1}$,

$$\left\| z_t - \mathrm{prox}_{\gamma f_i}(x_k) \right\|^2 \leq 2 \cdot (1+\gamma L_i) \left(1 - \frac{1}{\sqrt{1+\gamma L_i}}\right)^t \left\| x_k - \mathrm{prox}_{\gamma f_i}(x_k) \right\|^2. \tag{40}$$

We have

$$t = \mathcal{O}\left(\sqrt{1+\gamma L_i} \log\left(\frac{(1+\gamma L_i) \cdot \left\| x_k - \mathrm{prox}_{\gamma f_i}(x_k) \right\|^2}{\varepsilon_1}\right)\right)$$

Similar to the proof of Theorem 5, since we have according to (39),

$$\left\| x_k - \mathrm{prox}_{\gamma f_i}(x_k) \right\|^2 \leq \left\| \gamma \nabla f_i(x_k) \right\|^2,$$

it is straightforward to determine the number of local iterations needed.

Approximation satisfying Definition 4: Using (40), we have

$$t = \mathcal{O}\left(\sqrt{1+\gamma L_i} \log\left(\frac{1+\gamma L_i}{\varepsilon_2}\right)\right).$$

H.7. Proof of Theorem 11

In this case, the gradient estimator is defined as

$$g(x_k) = \frac{1}{\tau} \sum_{i \in S_k} \left(\gamma \nabla M_{f_i}^{\gamma}(x_k) - \left(\tilde{x}_{i,k+1} - \mathrm{prox}_{\gamma f_i}(x_k)\right)\right).$$

Notice that we have

$$\begin{aligned}
&\left\langle \gamma \nabla M^{\gamma}(x_k), \mathbb{E}\left[g(x_k)\right] \right\rangle \\
&= \left\langle \gamma \nabla M^{\gamma}(x_k), \mathbb{E}\left[\frac{1}{\tau} \sum_{i \in S_k} \gamma \nabla M_{f_i}^{\gamma}(x_k) - \frac{1}{\tau} \sum_{i \in S_k} \left(\tilde{x}_{i,k+1} - \mathrm{prox}_{\gamma f_i}(x_k)\right)\right] \right\rangle \\
&= \left\langle \gamma \nabla M^{\gamma}(x_k), \gamma \nabla M^{\gamma}(x_k) - \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{x}_{i,k+1} - \mathrm{prox}_{\gamma f_i}(x_k)\right) \right\rangle.
\end{aligned}$$

Using the same technique in the proof of Theorem 2, we are able to obtain that

$$\left\langle \gamma \nabla M^{\gamma}(x_k), \mathbb{E}\left[g(x_k)\right] \right\rangle \geq \left(1 - \frac{2\sqrt{\varepsilon_2} L_{\max}}{\mu}\right) \cdot \left\| \gamma \nabla M^{\gamma}(x_k) \right\|^2.$$

Thus, as long as we pick $\varepsilon_2 < \frac{\mu^2}{4L_{\max}^2}$, we can pick $b = 1 - \sqrt{\varepsilon_2} \cdot \frac{2L_{\max}}{\mu}$ and $c = 0$. We then compute $\mathbb{E}\left[\|g(x_k)\|^2\right]$,

$$\begin{aligned}
\mathbb{E}\left[\|g(x_k)\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{\tau}\sum_{i\in S_k}\gamma\nabla M_{f_i}^{\gamma}(x_k) - \frac{1}{\tau}\sum_{i\in S_k}\left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}(x_k)\right)\right\|^2\right] \\
&= \underbrace{\mathbb{E}\left[\left\|\frac{1}{\tau}\sum_{i\in S_k}\gamma\nabla M_{f_i}^{\gamma}(x_k)\right\|^2\right]}_{:=T_1} + \underbrace{\mathbb{E}\left[\left\|\frac{1}{\tau}\sum_{i\in S_k}\left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}(x_k)\right)\right\|^2\right]}_{:=T_2} \\
&\quad \underbrace{-2\mathbb{E}\left[\left\langle \frac{1}{\tau}\sum_{i\in S_k}\gamma\nabla M_{f_i}^{\gamma}(x_k), \frac{1}{\tau}\sum_{i\in S_k}\left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}(x_k)\right)\right\rangle\right]}_{:=T_3}.
\end{aligned}$$

We try to provide upper bounds for those terms separately.

Term T_1: We have

$$T_1 = \frac{n-\tau}{\tau(n-1)} \cdot \frac{1}{n}\sum_{i=1}^{n}\left\|\gamma\nabla M_{f_i}^{\gamma}(x_k)\right\|^2 + \frac{n(\tau-1)}{\tau(n-1)} \cdot \left\|\gamma\nabla M^{\gamma}(x_k)\right\|^2.$$

Using smoothness of $\gamma M_{f_i}^{\gamma}$ and the fact that we are in the interpolation regime, we have

$$\begin{aligned}
T_1 &= \frac{n-\tau}{\tau(n-1)} \cdot \frac{1}{n}\sum_{i=1}^{n}\left\|\gamma\nabla M_{f_i}^{\gamma}(x_k) - \gamma\nabla M_{f_i}^{\gamma}(x_\star)\right\|^2 + \frac{n(\tau-1)}{\tau(n-1)} \cdot \left\|\gamma\nabla M^{\gamma}(x_k)\right\|^2 \\
&\leq \frac{n-\tau}{\tau(n-1)} \cdot \frac{1}{n}\sum_{i=1}^{n}\frac{2\gamma L_i}{1+\gamma L_i} \cdot \left(\gamma M_{f_i}^{\gamma}(x_k) - \gamma\left(M_{f_i}^{\gamma}\right)_{\text{inf}}\right) + \frac{n(\tau-1)}{\tau(n-1)} \cdot \left\|\gamma\nabla M^{\gamma}(x_k)\right\|^2 \\
&\leq \frac{n-\tau}{\tau(n-1)} \cdot \frac{2\gamma L_{\max}}{1+\gamma L_{\max}} \cdot \left(\gamma M^{\gamma}(x_k) - \gamma M_{\text{inf}}^{\gamma}\right) + \frac{n(\tau-1)}{\tau(n-1)} \cdot \left\|\gamma\nabla M^{\gamma}(x_k)\right\|^2. \qquad (41)
\end{aligned}$$

Term T_2: It is easy to see that using convexity of the squared Euclidean norm, we have

$$\begin{aligned}
T_2 &\leq \mathbb{E}\left[\frac{1}{\tau}\sum_{i\in S_k}\left\|\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}(x_k)\right\|^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\|\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}(x_k)\right\|^2 \overset{(7)}{\leq} \frac{\varepsilon_2}{n}\sum_{i=1}^{n}\left\|\gamma\nabla M_{f_i}^{\gamma}(x_k)\right\|^2.
\end{aligned}$$

Using smoothness of each individual $\gamma M_{f_i}^{\gamma}(x_k)$ and the fact we are in the interpolation regime, we have

$$T_2 \leq \frac{2\varepsilon_2\gamma L_{\max}}{1+\gamma L_{\max}}\left(\gamma M^{\gamma}(x_k) - \gamma M_{\text{inf}}^{\gamma}\right). \qquad (42)$$

Term T_3: We have

$$T_3 = -2 \cdot \frac{n-\tau}{\tau(n-1)} \cdot \frac{1}{n} \sum_{i=1}^{n} \left\langle \gamma \nabla M_{f_i}^{\gamma}(x_k), \tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}(x_k) \right\rangle$$
$$- 2 \cdot \frac{n(\tau-1)}{\tau(n-1)} \cdot \left\langle \gamma \nabla M^{\gamma}(x_k), \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}(x_k) \right) \right\rangle.$$

Using Cauchy-Schwarz inequality and convexity, we further obtain

$$T_3 \leq 2 \cdot \frac{n-\tau}{\tau(n-1)} \cdot \frac{1}{n} \sum_{i=1}^{n} \left\| \gamma \nabla M_{f_i}^{\gamma}(x_k) \right\| \left\| \tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}(x_k) \right\|$$
$$+ 2 \cdot \frac{n(\tau-1)}{\tau(n-1)} \left\| \gamma \nabla M^{\gamma}(x_k) \right\| \cdot \frac{1}{n} \sum_{i=1}^{n} \left\| \tilde{x}_{i,k+1} - \operatorname{prox}_{\gamma f_i}(x_k) \right\|.$$

Using similar approaches in the previous paragraphs, we have

$$\begin{aligned}
& T_3 \\
& \overset{(7)}{\leq} \frac{2(n-\tau)}{\tau(n-1)} \cdot \frac{\sqrt{\varepsilon_2}}{n} \sum_{i=1}^{n} \left\| \gamma \nabla M_{f_i}^{\gamma}(x_k) \right\|^2 + \frac{2n(\tau-1)}{\tau(n-1)} \left\| \gamma M^{\gamma}(x_k) \right\| \frac{\sqrt{\varepsilon_2}}{n} \cdot \sum_{i=1}^{n} \left\| \gamma \nabla M_{f_i}^{\gamma}(x_k) \right\| \\
& \leq \frac{2(n-\tau)}{\tau(n-1)} \cdot \frac{\sqrt{\varepsilon_2}}{n} \sum_{i=1}^{n} \left\| \gamma \nabla M_{f_i}^{\gamma}(x_k) - \gamma \nabla M_{f_i}^{\gamma}(x_\star) \right\|^2 \\
& \quad + \frac{2n(\tau-1)}{\tau(n-1)} \left\| \gamma M^{\gamma}(x_k) \right\| \frac{\sqrt{\varepsilon_2}}{n} \cdot \sum_{i=1}^{n} \left\| \gamma \nabla M_{f_i}^{\gamma}(x_k) - \gamma \nabla M_{f_i}^{\gamma}(x_k) \right\| \\
& \leq \frac{4\sqrt{\varepsilon_2}(n-\tau)}{\tau(n-1)} \cdot \frac{\gamma L_{\max}}{1+\gamma L_{\max}} \left(\gamma M^{\gamma}(x_k) - \gamma M_{\inf}^{\gamma} \right) \\
& \quad + \frac{4\sqrt{\varepsilon_2} n(\tau-1)}{\tau(n-1)} \cdot \frac{\gamma L_{\max}}{1+\gamma L_{\max}} \left\| x_k - x_\star \right\| \left\| \gamma \nabla M^{\gamma}(x_k) \right\| \\
& \overset{(20)}{\leq} \frac{4\sqrt{\varepsilon_2}(n-\tau)}{\tau(n-1)} \cdot \frac{\gamma L_{\max}}{1+\gamma L_{\max}} \left(\gamma M^{\gamma}(x_k) - \gamma M_{\inf}^{\gamma} \right) \\
& \quad + \frac{4\sqrt{\varepsilon_2} n(\tau-1)}{\tau(n-1)} \cdot \frac{L_{\max}}{\mu} \left\| \gamma \nabla M^{\gamma}(x_k) \right\|^2. \qquad (43)
\end{aligned}$$

Combining (41), (42) and (43), we have

$$\sum_{i=1}^{3} T_i \leq 2 \left(\varepsilon_2 + \frac{2\sqrt{\varepsilon_2}(n-\tau)}{\tau(n-1)} + \frac{(n-\tau)}{\tau(n-1)} \right) \cdot \frac{\gamma L_{\max}}{1+\gamma L_{\max}} \cdot \left(\gamma M^{\gamma}(x_k) - \gamma M_{\inf}^{\gamma} \right)$$
$$+ \left(\frac{n(\tau-1)}{\tau(n-1)} + \frac{4\sqrt{\varepsilon_2} n(\tau-1)}{\tau(n-1)} \right) \cdot \frac{L_{\max}}{\mu} \cdot \left\| \gamma M^{\gamma}(x_k) \right\|^2. \qquad (44)$$

Therefore, it is easy to see that we can pick

$$A = \left(\varepsilon_2 + \frac{2\sqrt{\varepsilon_2}\,(n-\tau)}{\tau\,(n-1)} + \frac{(n-\tau)}{\tau\,(n-1)}\right) \cdot \frac{\gamma L_{\max}}{1+\gamma L_{\max}}$$
$$B = \left(\frac{n\,(\tau-1)}{\tau\,(n-1)} + \frac{4\sqrt{\varepsilon_2}n\,(\tau-1)}{\tau\,(n-1)}\right) \cdot \frac{L_{\max}}{\mu}, \qquad C = 0.$$

Applying Theorem 4 of [16], we list the corresponding values of $A, B, C, b, c \geq 0$ below,

$$A = \frac{\gamma L_{\max}}{1+\gamma L_{\max}} \left(\varepsilon_2 + \frac{2\sqrt{\varepsilon_2}\,(n-\tau)}{\tau\,(n-1)} + \frac{(n-\tau)}{\tau\,(n-1)}\right)$$
$$B = \frac{n\,(\tau-1)}{\tau\,(n-1)} \left(1 + \frac{4\sqrt{\varepsilon_2}L_{\max}}{\mu}\right), \quad C = 0$$
$$b = \frac{\mu - 2\sqrt{\varepsilon_2}L_{\max}}{\mu}, \quad c = 0.$$

We know that the PL constant of γM^{γ} is given by $\frac{\gamma\mu}{4(1+\gamma L_{\max})}$ and the corresponding smoothness constant is γL_{γ}. As a result, when $\alpha > 0$ satisfies

$$\alpha < \underbrace{\frac{1}{\gamma L_{\gamma}} \cdot \frac{\mu - 2\sqrt{\varepsilon_2}L_{\max}}{\mu + 4\varepsilon_2 L_{\max} + 4\sqrt{\varepsilon_2}L_{\max} + \frac{n-\tau}{\tau(n-1)} \cdot \left(4L_{\max} + 4\sqrt{\varepsilon_2}L_{\max} - \mu\right)}}_{:=B_1'},$$

and

$$\alpha < \underbrace{\frac{4\,(1+\gamma L_{\max})}{\gamma\left(\mu - 2\sqrt{\varepsilon_2}L_{\max}\right)}}_{=B_2},$$

we can obtain a convergence guarantee for the algorithm. Notice that $B_1' \leq B_1 < B_2$[6], thus we can further simplify the range of α to

$$\alpha \leq \underbrace{\frac{1}{\gamma L_{\gamma}} \cdot \frac{\mu - 2\sqrt{\varepsilon_2}L_{\max}}{\mu + 4\varepsilon_2 L_{\max} + 4\sqrt{\varepsilon_2}L_{\max} + \frac{n-\tau}{\tau(n-1)} \cdot \left(4L_{\max} + 4\sqrt{\varepsilon_2}L_{\max} - \mu\right)}}_{:=B_1'}.$$

Given that we select α properly, we have

$$\mathbb{E}\left[\mathcal{E}_K\right] \leq \left(1 - \alpha \cdot \frac{\gamma\left(\mu - 2\sqrt{\varepsilon_2}L_{\max}\right)}{4\,(1+\gamma L_{\max})}\right)^K \mathcal{E}_0.$$

Specifically, if we choose the largest α possible, we have

$$\mathbb{E}\left[\mathcal{E}_K\right] \leq \left(1 - \frac{\mu}{4L_{\gamma}\,(1+\gamma L_{\max})} \cdot S\left(\varepsilon_2, \tau\right)\right)^K \mathcal{E}_0,$$

6. The definition of B_1 is given in (32)

where $S(\varepsilon_2, \tau)$ is defined as

$$S(\varepsilon_2, \tau) = \frac{\left(\mu - 2\sqrt{\varepsilon_2} L_{\max}\right)\left(1 - 2\sqrt{\varepsilon_2}\frac{L_{\max}}{\mu}\right)}{\mu + 4\varepsilon_2 L_{\max} + 4\sqrt{\varepsilon_2} L_{\max} + \frac{n-\tau}{\tau(n-1)} \cdot \left(4L_{\max} + 4\sqrt{\varepsilon_2} L_{\max} - \mu\right)},$$

satisfying

$$0 < S(\varepsilon_2, \tau) \leq 1.$$

Using smoothness of γL_γ, if we denote $\Delta_k = \|x_k - x_\star\|^2$ where $x_\star$ is a minimizer of both M^γ and f since we assume we are in the interpolation regime (Assumption 2), we have

$$\mathcal{E}_0 \leq \frac{\gamma L_\gamma}{2} \Delta_0.$$

Using star strong convexity (quadratic growth property), we have

$$\mathcal{E}_K \geq \frac{\gamma \mu}{2\left(1 + \gamma L_{\max}\right)} \Delta_K.$$

As a result, we can transform the above convergence guarantee into

$$\mathbb{E}\left[\Delta_K\right] \leq \left(1 - \frac{\mu}{4L_\gamma\left(1 + \gamma L_{\max}\right)} \cdot S(\varepsilon_2, \tau)\right)^K \cdot \frac{L\gamma\left(1 + \gamma L_{\max}\right)}{\mu} \Delta_0.$$

Appendix I. Experiments

We describe the settings for the numerical experiments and the corresponding results to validate our theoretical findings. We are interested in the following optimization problem in the distributed setting,

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

Here n denotes the number of clients, d is the dimension, each function $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ has the following form

$$f_i(x) = \frac{1}{2} x^\top \boldsymbol{A}_i x + b_i^\top x + c_i,$$

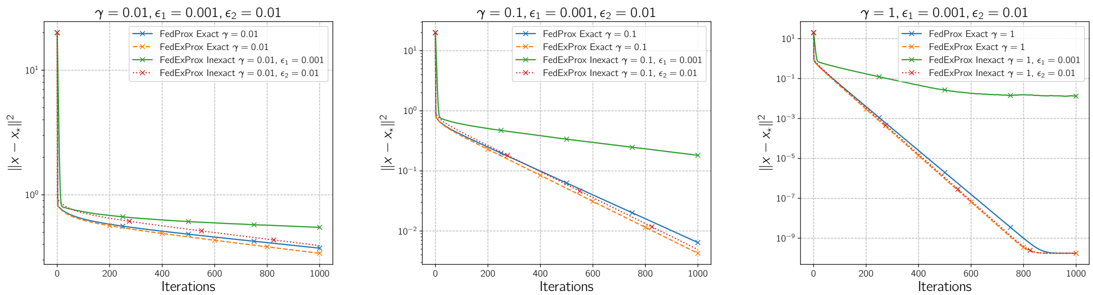
where $\boldsymbol{A}_i \in \mathbb{S}_+^d, b_i \in \mathbb{R}^d, c_i \in \mathbb{R}$. Specifically, we pick $n = 20$ and $d = 300$ for the experiments. Notice that we have

$$\nabla f_i(x) = \boldsymbol{A}_i x - b_i; \qquad \nabla^2 f_i(x) = \boldsymbol{A}_i \succeq \boldsymbol{O}_d,$$

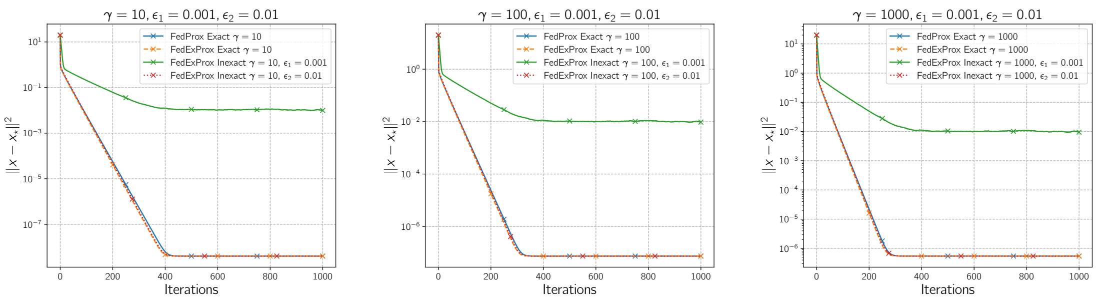
which suggests that each f_i is convex and smooth. We can easily compute that in this case, we have

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

All experiment codes were implemented in Python 3.11 using the NumPy and SciPy libraries. The computations were performed on a system powered by an AMD Ryzen 9 5900HX processor with Radeon Graphics, featuring 8 cores and 16 threads, running at 3.3 GHz. Code availability: https://anonymous.4open.science/r/Inexact-FedExProx-code-E783/



(*a*)



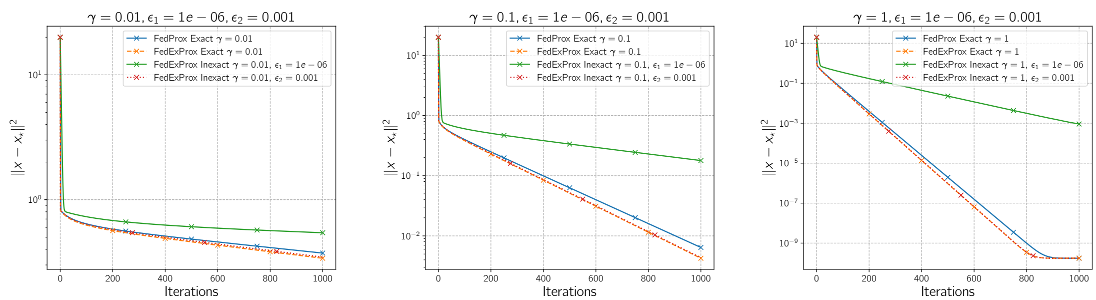
(*b*)

Figure 1: Comparison of FedProx, FedExProx with exact proximal evaluations, FedExProx with ε_1-absolute approximation and FedExProx with ε_2-relative approximation. In this case, we fix $\varepsilon_1 = 0.001$, $\varepsilon_2 = 0.01$ respectively and pick the local step size $\gamma \in \{1000, 100, 10, 1, 0.1.0.01\}$. The y-axis is the squared distance to the minimizer of f, and the x-axis denotes the iterations.

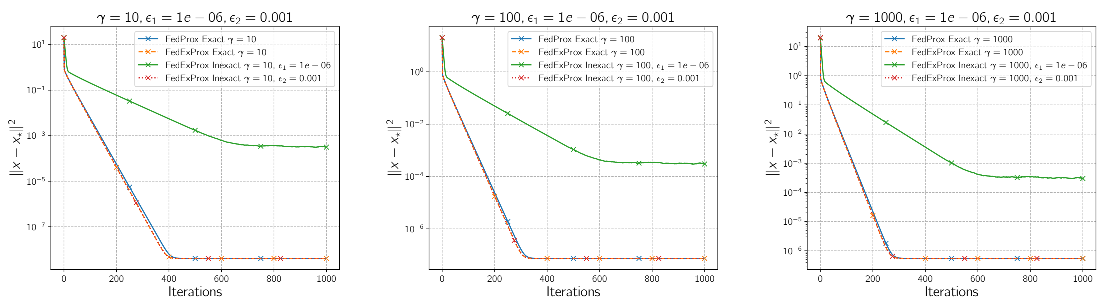
I.1. Comparison of FedProx, FedExProx, FedExProx with absolute approximation and relative approximation

In this section, we compare the convergence of FedProx, FedExProx and FedExProx with absolute approximation and relative approximation. For FedProx, we simply set the server extrapolation to be 1 while for FedExProx, we set its extrapolation parameter to be $\frac{1}{\gamma L_\gamma}$. We assume exact proximal evaluation for the above two algorithms. For FedExProx with approximations, we fix ε_1 and ε_2 to be reasonable values, respectively. We then set their extrapolation parameter to be the optimal value under the specific setting. Throughout the experiment, we vary the value of the local step size γ to see its effect on all the algorithms. Specifically, we select γ from the set $\{1000, 100, 10, 1, 0.1.0.01\}$, and we fix $\varepsilon_1 = 0.001$, $\varepsilon_2 = 0.01$ first, then we set them to $\varepsilon_1 = 1e-6$, $\varepsilon_2 = 0.001$.

Notably in Figure 1 and Figure 2, in all cases, FedExProx with absolute approximation exhibits the poorest performance and converges only to a neighborhood of the solution. This is expected, since the bias in this case does not go to zero as the algorithm progresses. It is worth mentioning that as the local step size γ increases, the size of the neighborhood decreases, which supports our claim in Theorem 1. As anticipated, in all cases, FedExProx outperforms FedProx due to server extrapolation. However, as γ increases, the performance gap between them diminishes. The performance

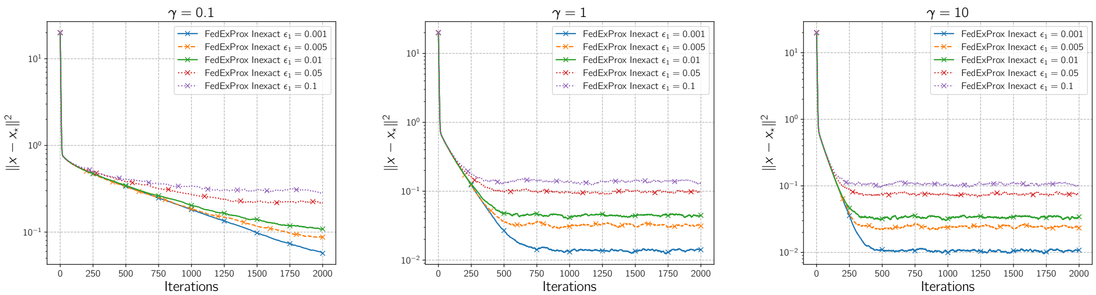


(a)



(b)

Figure 2: Comparison of FedProx, FedExProx with exact proximal evaluations, FedExProx with ε_1-absolute approximation and FedExProx with ε_2-relative approximation. In this case, we choose $\varepsilon_1 = 1e - 6$, $\varepsilon_2 = 0.001$ and pick the local step size $\gamma \in \{1000, 100, 10, 1, 0.1.0.01\}$. The y-axis is the squared distance to the minimizer of f, and the x-axis denotes the iterations.



(*a*)

Figure 3: Comparison of FedExProx with ε_1-absolute approximation under different level of inexactness. We select γ from the set $\{0.1, 1, 10\}$ and for each choice of γ, we select ε_1 from the set $\{0.001, 0.005, 0.01, 0.05, 0.1\}$. The y-axis denotes the squared distance to the minimizer and the x-axis is the number of iterations.

of FedExProx with relative approximation is surprisingly good, outperforming FedProx in several cases. This suggests the effectiveness of server extrapolation even when the proximal evaluations are inexact.

I.2. Comparison of FedExProx with absolute approximation under different inaccuracies

In this section, we compare FedExProx with absolute approximations under different level of inaccuracies. We fix the local step size γ to be a reasonable value, and we vary the level of inexactness for the algorithm. Specifically, we select γ from the set $\{0.1, 1, 10\}$ and for each choice of γ, we select ε_1 from the set $\{0.001, 0.005, 0.01, 0.05, 0.1\}$.

As observed in Figure 3, the size of the neighborhood increases with ε_1, further corroborating our theoretical findings in Theorem 1. Before reaching the neighborhood, the convergence rates of FedExProx with different level of inexactness are similar, which is expected.

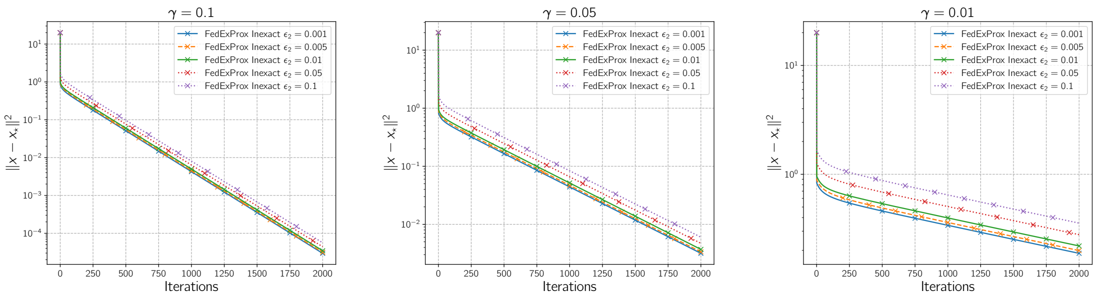
I.3. Comparison of FedExProx with relative approximation under different inaccuracies

In this section, we compare FedExProx with relative approximations under different level of relative inaccuracies. We fix the local step size γ to be a reasonable value, and we vary the level of inexactness for the algorithm. Specifically, we select γ from the set $\{0.1, 0.05, 0.01\}$ and for each choice of γ, we select ε_2 from the set $\{0.001, 0.005, 0.01, 0.05, 0.1\}$.

As observed in Figure 4, in all cases, a smaller ε_2 corresponds to faster convergence of the algorithm. This supports the claim of Theorem 3. All the tested algorithm converges to the exact solution linearly, which validates the effectiveness of the proposed technique of relative approximation to reduce the bias term.

I.4. Adaptive extrapolation for inexact proximal evaluations

In this section, we study the possibility of applying adaptive extrapolation to FedExProx with relative approximations. We do not consider the case of absolute approximation since it converges only



(*a*)

Figure 4: Comparison of FedExProx with ε_2-relative approximation under different level of inexactness. We select γ from the set $\{0.01, 0.05, 0.1\}$ and for each choice of γ, we select ε_2 from the set $\{0.001, 0.005, 0.01, 0.05, 0.1\}$. The y-axis denotes the squared distance to the minimizer and the x-axis is the number of iterations.

to a neighborhood, which causes problems when combined with adaptive step sizes such as gradient diversity and Polyak step size.

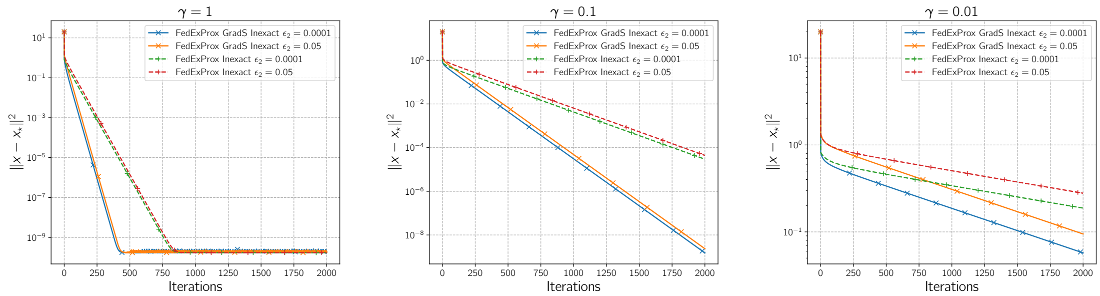
We are using the following definition of gradient diversity based extrapolation,

$$\alpha_k = \alpha_{k,G} := \frac{1+\gamma L_{\max}}{\gamma L_{\max}} \cdot \frac{\frac{1}{n}\sum_{i=1}^{n}\left\|x_k - \operatorname{prox}_{\gamma f_i}(x_k)\right\|^2}{\left\|\frac{1}{n}\sum_{i=1}^{n}\left(x_k - \operatorname{prox}_{\gamma f_i}(x_k)\right)\right\|^2}.$$

for Polyak type extrapolation, we use

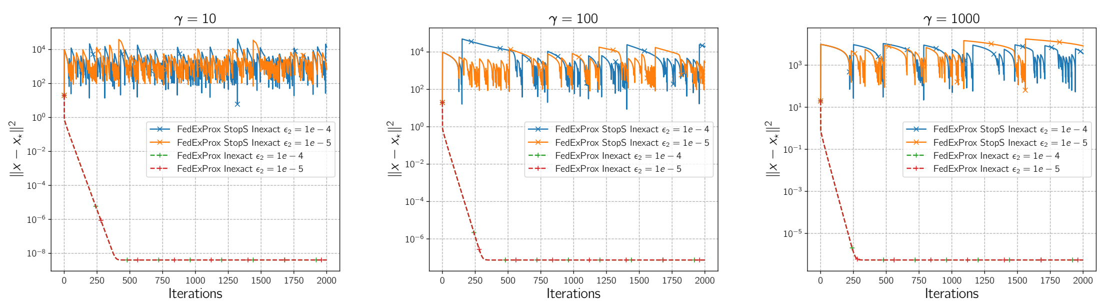
$$\alpha_k = \alpha_{k,S} := \frac{\frac{1}{n}\sum_{i=1}^{n}\left(M_{f_i}^{\gamma}(x_k) - \inf M_{f_i}^{\gamma}\right)}{\gamma\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla M_{f_i}^{\gamma}(x_k)\right\|^2}.$$

As it can be observed from Figure 5, in all cases, the use of a gradient diversity based adaptive extrapolation results in faster convergence of the algorithm. This suggests the possibility of developing an adaptive extrapolation for our methods. However, as we can see from Figure 6, a direct implementation of Polyak step size type extrapolation results in divergence of the algorithm, indicating that the challenge may be more complex than anticipated. In our case, this is equivalent to designing adaptive step sizes for SGD with biased updates or CGD with biased compression. To the best of our knowledge, this field remains open and requires further investigation, as biased updates are quite common in practice.



(*a*)

Figure 5: Comparison of FedExProx with ε_2-relative approximation under different level of inexactness using gradient diversity based extrapolation. we select γ from the set $\{1, 0.1, 0.01\}$ and for each choice of γ, we select ε_2 from the set $\{0.0001, 0.05\}$. The y-axis denotes the squared distance to the minimizer and the x-axis is the number of iterations.



(*a*)

Figure 6: Comparison of FedExProx with ε_2-relative approximation under different level of inexactness using Polyak step size based extrapolation. we select γ from the set $\{10, 100, 1000\}$ and for each choice of γ, we select ε_2 from the set $\{1e-4, 1e-5\}$. The y-axis denotes the squared distance to the minimizer and the x-axis is the number of iterations.