Federated Inexact Alternating Direction Method of Multipliers with Mixed Error Critera for Federated Learning

Anonymous Authors[1]

Abstract

Federated learning has attracted increasing attention in the machine learning community in the past five years. In this paper, we propose a new federated learning algorithm with fast convergence guarantee to solve the machine learning models with nonsmooth regularizers. To solve this type of problem, we design an inexact federated alternating direction method of multipliers (ADMM). For each agent, we propose a new termination criterion that uses a combination of absolute and relative error criteria. We show that the resulted method has more choices of absolute error criteria while keeping the best-known complexity. Furthermore, we show that the proposed method has sequential convergence guarantees. In particular, we show that the updates at the server accumulate at the stationary point. Then, we investigate the global convergence rate of the generated sequence under the Kurdyka-Łojasiewicz (KL) assumption. This rate imposes a faster convergence rate of the distance from 0 to the subdifferential of the objective. We conduct experiments using both synthetic and real datasets to demonstrate the superiority of our new methods over existing algorithms.

1. Introduction

Federated learning (FL) is an emerging paradigm where multiple agents collaboratively solve one machine learning problem. In an FL task, each agent possesses part of the dataset and uses them to train machine learning model locally. All agents send their outputs to a central server. The server aggregates the outputs and send an update back to the agents. Most of the current works of FL focus on unconstrained smooth models (Kairouz et al., 2021; McMahan et al., 2017b; Pathak & Wainwright, 2020). A classical method for unconstrained smooth models is FedAvg (McMahan et al., 2017b). To deal with the data and system heterogeneity among the agents, variants of FedAvg are proposed in (Reddi et al., 2021; Li et al., 2020a; Karimireddy et al., 2020). For example, Li et al. (2020a) proposed FedProx that deals with the system heterogeneity by locally calculating the proximal operator of f_i. Karimireddy et al. (2020) presented SCAFFOLD that applies a variance reduction technique to deal with non-iid local data sets.

Besides unconstrained smooth problems, there are many constrained nonsmooth models in federated learning. For example, when the parameters are sparse, the model has a nonsmooth sparsity-inducing function, (Zou & Hastie, 2005; Yuan et al., 2021). Also, when parameters are low-rank matrix, the model often includes a nuclear norm, (Candès & Recht, 2009; Bao et al., 2022). In addition, if there are restrictions on the parameters, the model would have constraints. To deal with these issues, in this work, we study the nonsmooth federated learning model:

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{p} f_i(x) + g(x), \tag{1}$$

where each $f_i : \mathbb{R}^n \to \mathbb{R}$ is smooth (probably nonconvex) and ∇f_i is Lipschitz continuous with modulus L_i, $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper closed convex function. In applications, f_i's are loss functions of p local data sets and g can be ℓ_1, grouped ℓ_1, nuclear-norm regularizer (for matrix variable), or the indicator function of a convex constraint (Yuan et al., 2021; Bao et al., 2022). When $g = 0$, (1) reduces to an unconstrained smooth federated learning problem. When $g \neq 0$, the problem (1) is called the federated composite optimization in (Yuan et al., 2021). In (Yuan et al., 2021), the federated dual averaging (FedDualAvg) was proposed as an early attempt to deal with the nonsmooth g. (Bao et al., 2022) proposed a fast federated dual averaging for problem (1) with a strongly convex f.

FedAvg, FedProx, FedDualAvg, and their variants have natural intuitions of how to distribute the tasks and aggregate the local outputs, but they have bottlenecks in both theory and practice. For example, McMahan et al. (2017a) showed that FedAvg can diverge in certain cases. Even FedAvg

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

converges, as presented in (Pathak & Wainwright, 2020), the resulting fixed points may not be a stationary point of the original problem. Furthermore, in the analysis of (Yuan et al., 2021; Li et al., 2020a; Reddi et al., 2021), the dissimilarity between the agents is always assumed to be bounded, which may fail in many real applications. These deficiencies of existing methods motivate us to consider the federated splitting methods for solving (1). In general, the idea of splitting methods in federated learning is to relate (1) to the following constrained problem:

$$\min_{x} \sum_{i=1}^{p} f_i(x_i) + g(x_1) \text{ s.t. } x_1 = x_2 = \cdots = x_p, \tag{2}$$

where $x = (x_1, x_2, \ldots, x_p)$. Popular splitting methods in federated learning include FedSplit (Pathak & Wainwright, 2020), FedDR (Tran-Dinh et al., 2021), FedPD (Zhang et al., 2021), FedADMM (Gong et al., 2022), and other variants of FedADMM (Zhou & Li, 2021; Zhang et al., 2021; Yue et al., 2021; Zhou & Li, 2022a). FedSplit, FedPD, and FedADMM deal with the case where $g = 0$. FedDR considers the case where g can be nonzero. At each round, federated splitting methods calculate the proximal operator of each f_i (or the prox of f_i for short) locally and inexactly. Then, after some linear transformations, the local server sends the local update to the server. To terminate the local updates, each agent checks whether the error between the current update and the exact prox of f_i (the local error) is under a threshold. When analyzing the convergence properties of the proposed methods, this threshold either appears as an additional term in the complexity or is assumed to be summable with respect to the iterations. In particular, at round t, the i_{th} agent terminates the local updates when the local error is under a threshold ϵ_i^t. When $g = 0$ and there exists an upper bound $\epsilon_{\max}$ for $\{\epsilon_i^t\}$, Zhang et al. (2021) proved that FedPD has a complexity of $O(\epsilon^{-1}) + O(\epsilon_{\max})$ to reach an ϵ-surrogate stationary point. When $g \neq 0$, Tran-Dinh et al. (2021) showed FedDR has a complexity of $O(\epsilon^{-1})$ when assuming $\sum_{i,t} \epsilon_i^t$ is summable. When $g = 0$, Zhou & Li (2022a) presented an inexact ADMM has a complexity of $O(\epsilon^{-1})$, where they assume $\{\epsilon_i^t\}_t$ decreases exponentially, which also makes $\{\epsilon_i^t\}_t$ summable.

The above-mentioned termination criterion is usually called the absolute error criterion, as an opposite to another type of termination criterion — the relative error criterion (Rockafellar, 1976; Eckstein & Yao, 2018; Xie et al., 2017; Xie, 2018). Instead of setting a threshold for the error of the current update, the relative error criterion tests the relations between the current updates and the previous updates. For example, at iteration t, we have an approximation x^t of the exact solution $x_\star^t$ of the t_{th} subproblem. The absolute error criterion tests whether $\|x^t - x_\star^t\| \leq \epsilon$, while the relative error criterion tests whether $\|x^t - x_\star^t\| \leq \|x^t - x^{t-1}\|$, where x^{t-1} is the last update. Thus, this bound is dynamic. One advantage of methods with the relative error criterion is that they do not need summable thresholds but still obtain the same order of complexity. For example, Tran-Dinh et al. (2021) considered applying the relative error criterion to FedDR. The resulting FedDR has the complexity of $O(\epsilon^{-1})$ without assuming a sequence of thresholds to be summable. However, as far as we search, this is the only work that considers the relative error criterion in FL. The merits of relative error criteria have not been completely discovered in both nonconvex optimization and federated learning. One contribution of our work is to fill in this gap.

1.1. Our Contributions

In this work, we propose a federated inexact ADMM method with mixed errors (FIAME) for the nonconvex composed optimization problem (1). To propose FIAME, we relate (1) with an np-dimensional constrained problem. We establish a relationship between the stationary points of the np-dimensional constrained problem and the stationary points of (1). Based on the ADMM for the np-dimensional constrained problem, we develop an exact federated ADMM for (1). This is the first federated ADMM that deals with the nonsmooth regularizer in (1). Based on this exact federated ADMM, we propose its inexact variant, FIAME. To terminate the local updates, we propose a combination of absolute and relative error criteria.

To show the convergence properties of FIAME, we relate it to a new centralized inexact ADMM with mixed error criteria for the np-dementional problem mentioned in the last paragraph. We call this method as IAME. To analyze IAME, we first develop a new potential function. We show that in general stochastic cases, the function value of the potential function at each round is nonincreasing. Then we prove that IAME has a complexity of $O(\epsilon^{-1})$ and so does FIAME. The mixed criteria do not need a summable or decreasing external thresholds in the absolute criterion while maintaining the best known complexity.

We further investigate the convergence properties of IAME and FIAME in the deterministic case. In particular, we show that the sequence generated at the server of FIAME accumulates at a stationary point of (1). Then, we prove that the sequences generated by IAME and FIAME converge globally under Kurdyka-Łojasiewicz (KL) assumptions. In particular, we show the generated sequences converge finitely when the KL exponent α of the potential function is 0. The generated sequences converge linearly when $\alpha \in (0, \frac{1}{2})$. The generated sequence converges sublinearly when $\alpha \in (\frac{1}{2}, 1)$. Under the same KL assumptions, we show that the rate of 0 approaching $\sum_i \nabla f_i(y^t) + \partial g(y^t)$ is as fast as the convergence rate of the generated sequence generated by FIAME, where $\{y^t\}$ is the updates at the server. As far as we know, FIAME is the first federated learning method that has sequential convergence guarantees in the nonconvex nonsmooth settings.

To evaluate the efficiency of our method, we perform experiments on training fully-connected neural network. We compare our method with existing splitting methods and state-of-art methods. The experimental results show our methods consistently outperform the other methods in training loss, training accuracy, and testing accuracy.

1.2. Related Work

The literature of federated learning is rich. In this work, we only focus on the splitting methods in federated learning. A comparison between our method and existing splitting methods is summarized in Table 1.

Splitting methods in FL. Pathak & Wainwright (2020) proposed FedSplit that implements the Peaceman-Rachford splitting method for (2). They analyze the proposed method in the case where $g = 0$ and $\sum_i f_i$ is strongly convex. When $g \neq 0$, Tran-Dinh et al. (2021) proposed FedDR that applies the Douglas-Rachford (DR) splitting algorithms for (2). They combine the DR method with randomized block-coordinate strategies and asynchronous implementation. They estimate the complexity of FedDR with absolute error criteria and FedDR with relative error criteria respectively. When $g = 0$, Zhang et al. (2021) considered using an ADMM approach for the equivalent problem (2). When $g = 0$, Gong et al. (2022) proposed FedADMM that randomly selects agents to attend each round. The methods in (Zhang et al., 2021; Gong et al., 2022) terminated the local updates when the norm of the local gradient is under an absolute threshold. When $g = 0$ and f_i's are twice differentiable, ADMM is applied in designing a second-order FL method in (Elgabli et al., 2022). Yue et al. (2021) considered the case where g is the Bregman distance. Assuming the Hessian of f in (1) being Lipschitz continuous, Yue et al. (2021) showed any accumulation point of the generated sequence is a stationary point. When $g = 0$, Zhou & Li (2022a) proposed an inexact ADMM for federated learning problems. At round t, the i_{th} agent terminates the local updates when the norm of the local gradient is under a threshold ϵ_i^t. They assume $\{\epsilon_i^t\}_t$ decreases exponentially. They showed that the generated sequence accumulates at the stationary point. By further assuming the accumulation point of the generated sequence is isolated, they show the generated sequence converges globally. In our work, we do not assume the accumulation point of the generated sequence to be isolated. Instead, we use the KL property that is satisfied by a wild class of functions such as semi-algebraic functions.

Splitting methods with relative errors in convex optimization. The relative error criterion was early considered in deriving an inexact proximal point algorithm (PPA) for minimizing a proper closed convex function in (Rockafellar, 1976). This criterion is popular in convex optimization and designing splitting methods for the general problem $\min_x f(x) + g(AX)$, where f and g are proper closed convex and A is a linear map, (Eckstein & Yao, 2017; Xie, 2018; Xie et al., 2017; Alves et al., 2020). Using the relations between ADMM, Douglas-Rachford method (DR) and the relations between the DR method with PPA, Eckstein & Yao (2018) proposed an inexact ADMM method that finds the prox of f approximately using a relative error criterion. The global convergence of the generated sequence was established in (Eckstein & Yao, 2018) in convex settings. The method proposed in (Xie et al., 2017) solved the both prox of f and g inexactly and used a relative error criterion defined by the multipliers. Liu et al. (2021) proposed a preconditioned primal–dual hybrid gradient method for $\min f + g$ with f and g being convex. It uses the relative error criterion when solving the prox of the g^*, where g^* is the dual of g. When further assuming f is strongly convex and the generalized augmented Lagrangian of the considered problem is a KL function, Liu et al. (2021) showed the generated sequence converges to a primal-dual solution pair.

2. Preliminaries

In this paper, we denote $\mathbb{R}^n$ the n-dimensional Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and Euclidean norm $\| \cdot \|$. We denote the set of all positive numbers as $\mathbb{R}_{++}$. We denote the distance from a point a to a set $\mathcal{A}$ as $d(a, \mathcal{A})$. For a random variable ξ defined on a probability space (Ξ, Σ, P), we denote its expectation as $\mathbb{E}\xi$. Given an event A, the conditional expectation of ξ is denoted as $\mathbb{E}(\xi|A)$.

An extended-real-valued function $f : \mathbb{R}^n \to [-\infty, \infty]$ is said to be proper if $\operatorname{dom} f = \{x \in \mathbb{R}^n : f(x) < \infty\}$ is not empty and f never equals $-\infty$. We say a proper function f is closed if it is lower semicontinuous. We define the indicator function of a closed set $\mathcal{A}$ as $\delta_{\mathcal{A}}(x)$, which is zero when $x \in \mathcal{A}$ and ∞ otherwise.

Following Definition 8.3 in (Rockafellar & Wets, 1998), we define the regular subdifferential of a proper function $f : \mathbb{R}^n \to [-\infty, \infty]$ at $x \in \operatorname{dom} f$ as:

$$\hat{\partial} f(x) := \left\{ \xi \in \mathbb{R}^n \colon \liminf_{z \to x,\, z \neq x} \frac{f(z) - f(x) - \langle \xi, z - x \rangle}{\|z - x\|} \geq 0 \right\}$$

The (limiting) subdifferential of f at $x \in \operatorname{dom} f$ is defined as $\partial f(x) := \left\{ \xi \in \mathbb{R}^n \colon \exists x^k \xrightarrow{f} x, \xi^k \to \xi \text{ with } \xi^k \in \hat{\partial} f(x^k), \forall k \right\}$, where $x^k \xrightarrow{f} x$ means both $x^k \to x$ and $f(x^k) \to f(x)$. For $x \notin \operatorname{dom} f$, we define $\hat{\partial} f(x) = \partial f(x) = \emptyset$. We denote $\operatorname{dom} \partial f := \{x : \partial f(x) \neq \emptyset\}$. When f is differentiable at x, from Exercise 8.8 of (Rockafellar & Wets, 1998), we know that $\partial f(x) = \{\nabla f(x)\}$. For a differential function $h : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^l$, we denote $\nabla_x L(x, y)$ and $\nabla_y L(x, y)$ as the partial derivatives with respect to x and y correspond-

Table 1. Convergence results of our method, the federated nonsplitting methods FedAvg (Li et al., 2020b), SCAFFOLD (Karimireddy et al., 2020), FedSkip (Fan et al., 2022), and federated splitting methods FedSplit in (Pathak & Wainwright, 2020), FedPD (Zhang et al., 2021), FedDR (Tran-Dinh et al., 2021), FedADMM in (Gong et al., 2022) (denoted as FedADMM1), FedADMM in (Zhou & Li, 2022a) (denoted as FedADMM2). SC = Strongly Convex, NC = Nonconvex, NS = Nonsmooth, A = Absolute Error, R = Relative Error, G = Global Convergence, S = Subsequential Convergence, Γ = the degree of heterogeneity, $\rho \in (0, 1)$, θ is the KL exponent in Theorem 4.9, $\epsilon_{\max}$ is the upper bound of the absolute error.

Algorithm	Model		Error type	Sequential convergence		Complexity
	$\sum_i f_i$	g		G/S	Rate	
FedAvg	SC	0	–	–	–	$((O(1) + O(\Gamma)) \cdot O(1/t))$
FedSkip	SC	0	–	–	$((O(1) + O(\Gamma)) \cdot O(1/t)) + O(\rho^t)$	–
SCAFFOLD	NC	0	–	–	–	$O(1/\sqrt{t}) + O(1/t)$
FedSplit	SC	0	A(bounded)	–	$O(\rho^t) + O(\epsilon_{max})$	–
FedPD	NC	0	A(bounded)	–	–	$O(1/t) + O(\epsilon_{\max})$
FedDR	NC	NS	A(summable)	–	–	$O(1/t)$
			R	–	–	$O(1/t)$
FedADMM1	NC	0	A(bounded)	–	–	$O(1/t) + O(\epsilon_{\max})$
FedADMM2	NC	0	A(Exponentially decrease)	G	–	$O(1/t)$
FIAME (Ours)	NC	NS	A(nonincreasing)+R	G	$O(\rho^t)$ for $\theta \in (0, \frac{1}{2}]$	$O(1/t)$ in general $O(\rho^t)$ for $\theta \in (0, \frac{1}{2})$

ingly. In addition, the limiting subdifferential reduces to the classical subdifferential in convex analysis when f is convex. We defined the normal cone of a set $\mathcal{A}$ at x as $N_{\mathcal{A}}(x) := \partial \delta_{\mathcal{A}}(x)$.

For a proper function $f : \mathbb{R}^n \to [-\infty, \infty]$, we denote the proximal operator of f as $\text{Prox}_{\alpha f}(x) = \text{Arg}\min_{z \in \mathbb{R}^n} \left\{ f(z) + \frac{1}{2\alpha}\|z - x\|^2 \right\}$.

Definition 2.1. Consider a problem $\min f + g$, where f is a smooth function and g is properly closed convex. We say x is a stationary point of this problem when $0 \in \nabla f(x) + \partial g(x)$. We say x is an ε-stationary point if $d^2(0, \nabla f(x) + \partial g(x)) \leq \varepsilon$.

To analyze the convergence rate of the generated sequence, we next introduce the KL property. We let Ψ_a be defined as the set of concave functions $\psi : [0, a) \to [0, \infty)$ satisfying $\psi(0) = 0$, being continuously differentiable on $(0, a)$, and satisfying $\psi' > 0$ on $(0, a)$.

Definition 2.2 (Kurdyka-Łojasiewicz property and exponent). A proper closed function $f : \mathbb{R}^n \to (-\infty, \infty]$ is said to satisfy the Kurdyka-Łojasiewicz (KL) property at an $\hat{x} \in \text{dom}\partial f$ if there are $a \in (0, \infty]$, a neighborhood V of $\hat{x}$ and a $\varphi \in \Psi_a$ such that for any $x \in V$ with $f(\hat{x}) < f(x) < f(\hat{x}) + a$, it holds that $\psi'(f(x) - f(\hat{x}))\text{dist}(0, \partial f(x)) \geq 1$. If f satisfies the KL property at $\hat{x} \in \text{dom}\partial f$ and ψ can be chosen as $\psi(\nu) = a_0 \nu^{1-\alpha}$ for some $a_0 > 0$ and $\alpha \in [0, 1)$, then we say that f satisfies the KL property at $\hat{x}$ with exponent α. A proper closed function f satisfying the KL property with exponent $\alpha \in [0, 1)$ at every point in $\text{dom}\partial f$ is called a KL function with exponent α.

KL property is satisfied by a wide range of functions including proper closed semi-algebraic functions, the quadratic loss function plus possibly nonconvex piecewise linear regularizers and some fractional functions with linear constraints (Attouch et al., 2010; Li & Pong, 2018; Attouch et al., 2013; Zeng et al., 2021). The KL property plays an important role in deducing the global convergence properties of many first-order methods, see (Borwein et al., 2017; Bolte et al., 2014; Attouch et al., 2010; Li & Pong, 2016) for examples.

3. Federated Inexact ADMM with Mixed Error Criteria

To solve (2), we consider the following np-dimensional problem:

$$\min_{X \in \mathbb{R}^{np}} F(X) + G(X), \tag{3}$$

where $X = (x_1, x_2, \ldots, x_p)$ with each $x_i \in \mathbb{R}^n$, $F(X) := \sum_{i=1}^p f_i(x_i)$ with f_i's in (1), $G(X) := g(x_1) + \delta_{\mathfrak{C}}(X)$ with $\mathfrak{C} := \{X : \; x_1 = \cdots = x_p\}$ and g in (1).

The following proposition establishes the relation between (3) and (1). The proofs are presented in Appendix A.1.

Proposition 3.1. *If $X^* = (x_1^*, \ldots, x_p^*)$ is a stationary point of* (3)*, then x_1^* is a stationary point of* (1)*. Furthermore, if $X = (x_1, \ldots, x_p)$ is an ε-stationary point of* (1)*, then x_1 is a $p\varepsilon$-stationary point of* (1)*.*

Based on this relation, we consider ADMM to solve (3). Rewrite (3) as the following equivalent problem:

$$\min_{X, Y \in \mathbb{R}^{np}} F(X) + G(Y) \text{ s. t. } X = Y. \tag{4}$$

The augmented lagrangian function of (4) is defined as:

$$L_\beta(X, Y, Z) := F(X) + G(Y) + \langle X - Y, Z \rangle + \frac{\beta}{2}\|X - Y\|^2. \tag{5}$$

Given a starting point $(X^0, Y^0, Z^0) \in \mathbb{R}^{np} \times \mathbb{R}^{np} \times \mathbb{R}^{np}$ and $\tau, \beta > 0$, the ADMM for (3) is as follows:

$$\begin{cases} X^{t+1} = \arg\min_X L_\beta(X, Y^t, Z^t), \\ Z^{t+1} = Z^t + \tau\beta(X^{t+1} - Y^t), \\ Y^{t+1} = \arg\min_Y L_\beta(X^{t+1}, Y, Z^{t+1}). \end{cases} \tag{6}$$

Now we give an equivalent form of the third equation in (6) as follows. The proof is provided in Appendix A.2.

Proposition 3.2. *Consider (3). Let $\{(X^{t+1}, Y^{t+1}, Z^{t+1})\}$ be generated by (6). Suppose $\beta > \sum_i L_i$. Then the solution of the problem in the third equation of (6) is $(y_1, \dots, y_1)$ with $y_1 = \mathrm{Prox}_{\frac{1}{\beta p} g}(\frac{1}{p}\sum_{i=1}^p (x_i^{t+1} + \frac{1}{\beta} z_i^{t+1})))$.*

On the other hand, since $F(X)$ in (3) is separable, we can write $L_\beta(X, Y, Z)$ in (5) as $L_\beta(X, Y, Z) = \sum_{i=1}^p L_{\beta,i}(x_i, y_i, z_i)$, where

$$L_{\beta,i}(x_i, y_i, z_i) := f_i(x_i) + \langle x_i - y_i, z_i \rangle + \frac{\beta}{2}\|x_i - y_i\|^2.$$

Combining this and Proposition 3.2, the ADMM in (6) can be written as a federated algorithm Algorithm 1.

Algorithm 1 Exact Federated ADMM for (1)

```
1: Input: $\tau > 0$ and $\beta > 0$. $\{(x_i^0, y_i^0, z_i^0)\}_{i=1}^p$, $\bar{z}^0 = \frac{1}{p}\sum_{i=1}^p z_i^0$, $\bar{x}^0 = \frac{1}{p}\sum_{i=1}^p x_i^0$. $t = 0$.
2: For each agent $i = 1, \dots, p$, do
       $$x_i^{t+1} = \min_{x_i} L_{\beta,i}(x_i, y_i^t, z_i^t). \tag{7}$$
   Let $\Delta_{x_i,t+1} = x_i^{t+1} - x_i^t$ and $\Delta_{z_i,t+1} = \tau\beta(x_i^{t+1} - y_i^t)$.
3: For the server: Calculate $\bar{x}^{t+1} = \bar{x}^t + \frac{1}{p}\sum_i \Delta_{x_i,t+1}$ and $\bar{z}^{t+1} = \bar{z}^t + \frac{1}{p}\sum_{i=1}^p \Delta_{z_i,t+1}$. Let
       $$y^{t+1} = \mathrm{Prox}_{\frac{1}{\beta p} g}(\bar{x}^{t+1} + \frac{1}{\beta}\bar{z}^{t+1}).$$
   If a termination criterion is not satisfied, broadcast $y^{t+1}$ to each agent, let $t = t + 1$ and go to Step 2.
```

To compute x_i^{t+1} in (7), we need to compute the prox of f_i. However, since f_i's are the loss functions of the neural networks that have complex structures, exactly computing the prox of f is not practical. Based on Algorithm 1, we propose Algorithm 2 that calculates $\min_{x_i} L_{\beta,i}(x_i, y_i^t, z_i^t)$ in (7) inexactly.[1]

When $\beta > \sum_i L_i$, the local problem (8) is minimizing a strongly convex smooth function that has Lipscihtz continuous gradient. This can be solved efficiently with many first-order methods. For example, it is well known that the stochastic gradient descent for minimizing strongly convex smooth objective has an optimal rate $O(1/\epsilon)$, (Rakhlin et al., 2012; Nemirovski et al., 2009). The following theorem shows there exist methods such that criterion (9) can be satisfied. The proofs are given in Appendix A.3.

Algorithm 2 Federated Inexact ADMM with Relative Error (FIAME) for (1)

```
1: Input: for $i = 1, \dots, p$, the $i_{\text{th}}$ agent inputs $(x_i^0, y_i^0, z_i^0)$, $x_i^{-1} \neq x_i^0$, $\beta, \tau > 0$, $\bar{x}^0 = \frac{1}{p}\sum_i x_i^0$, $\bar{z}^0 = \frac{1}{p}\sum_i z_i^0$, the $i_{\text{th}}$ agent select $\{\epsilon_i\}_i \subseteq \mathbb{R}_{++}$. Let $t = 0$.
2: For each agent $i = 1, \dots, p$:
   If $\nabla_x L_{\beta,i}(x_i^t, y_i^t, z_i^t) = 0$, let $x_i^{t+1} = x_i^t$.
   Else, find an approximate solution
       $$x_i^{t+1} \approx \min_{x_i} L_{\beta,i}(x_i, y_i^t, z_i^t) \tag{8}$$
   such that
       $$\mathbb{E}_t^i\|\nabla_x L_{\beta,i}(x_i^{t+1}, y_i^t, z_i^t)\|^2 \le \epsilon_i \min\{\|x_i^t - x_i^{t-1}\|^2, \mathbb{E}_t^i\|x_i^{t+1} - x_i^t\|^2\}. \tag{9}$$
   Let $\Delta_{x_i,t+1} = x_i^{t+1} - x_i^t$ and $\Delta_{z_i,t+1} = \tau\beta(x_i^{t+1} - y_i^t)$.
3: The server calculate $\bar{x}^{t+1} = \bar{x}^t + \frac{1}{p}\sum_i \Delta_{x_i,t+1}$ and $\bar{z}^{t+1} = \bar{z}^t + \frac{1}{p}\sum_{i=1}^p \Delta_{z_i,t+1}$. Let
       $$y^{t+1} = \mathrm{Prox}_{\frac{1}{\beta p} g}(\bar{x}^{t+1} + \frac{1}{\beta}\bar{z}^{t+1}).$$
   If a termination criterion is not satisfied, broadcast $y^{t+1}$ to each agent, let $t = t + 1$ and go to Step 2.
```

Proposition 3.3. *Let $h : \mathbb{R}^n \to \mathbb{R}$ be a μ-strongly convex and L-smooth function. Pick any $x^0 \in \mathbb{R}^n$. Let $x^{k+1} = x^k - \eta_k \bar{\nabla} h(x^k)$ with $\eta_k = \frac{1}{\mu k}$ and $\bar{\nabla} h(x^k)$ being an unbiased estimator of $\nabla h(x^k)$. Suppose $\mathbb{E}\|\bar{\nabla} h(x^k)\|^2 \le W^2$. Then for any $\epsilon', \epsilon > 0$ and $\bar{x} \neq \arg\min_x h(x)$, there exists large k such that*

$$\mathbb{E}\|\nabla h(x^k)\|^2 \le \min\{\epsilon', \epsilon\mathbb{E}\|x^k - \bar{x}\|^2\}.$$

4. Convergence Analysis of Algorithm 2

To analyze Algorithm 2, we relate it to a centralized inexact ADMM with mixed criteria for (3), which is presented as Algorithm 3.[2]

The next proposition shows Algorithm 2 is a special case of Algorithm 3. The proofs are given in Appenidx B.1.

Proposition 4.1. *Let $\{(x_1^t, \dots, x_p^t, y^t, z_1^t, \dots, z_p^t)\}_t$ be generated by Algorithm 2. Let $X^t = (x_1^t, \dots, x_p^t)$, $Z^t =$*

[1] At iteration t of Algorithm 2, we denote $x^t = (x_1^t, x_2^t, \dots, x_p^t)$, $z^t = (z_1^t, z_2^t, \dots, z_p^t)$, and $\mathcal{X}_i^t = \{x^1, x^2, \dots, x^t, y^1, \dots, y^t, z^1, \dots, z^t\}$. Then we denote $\mathbb{E}_t^i \xi = \mathbb{E}(\xi|\mathcal{X}_i^t)$ for a random variable ξ.

[2] At iteration t in Algorithm 3, we denote $\mathcal{X}^t = \{X^1, X^2, \dots, X^t, Y^1, \dots, Y^t, Z^1, \dots, Z^t\}$ and denote $\mathbb{E}_t\xi = \mathbb{E}(\xi|\mathcal{X}^t)$ the conditional expectation of a random variable ξ given $\mathcal{X}^t$.

Algorithm 3 Inexact ADMM with Mixed Criteria (IAME) for (3)

```
1: Input: $(X^0, Y^0, Z^0)$, $X^{-1} \neq X^0$, $\beta, \tau, \epsilon > 0$. Let $t = 0$.
2: If $\nabla_X L_\beta(X^t, Y^t, Z^t) = 0$, let $x^{t+1} = z^t$.
   Else, find
       $X^{t+1} \approx \min_X L_\beta(X, Y^t, Z^t) \qquad (10)$
   such that
       $\mathbb{E}_t \|\nabla_X L_\beta(X^{t+1}, Y^t, Z^t)\|^2$
       $\leq \epsilon \min\{\|X^t - X^{t-1}\|^2, \mathbb{E}_t \|X^{t+1} - X^t\|^2\}. \qquad (11)$
3: Let
       $Z^{t+1} = Z^t + \tau\beta(X^{t+1} - Y^t). \qquad (12)$
4: Find
       $Y^{t+1} = \min_Y L_\beta(X^{t+1}, Y, Z^{t+1}). \qquad (13)$
5: If a termination criterion is not satisfied, let $t = t + 1$ and go to Step 2.
```

$(z_1^t, \ldots, z_p^t)$ and $Y^t = (y^t, \ldots, y^t)$. Then $\{(X^t, Y^t, Z^t)\}$ is a sequence generated by Algorithm 3 with $\epsilon = \max_i \epsilon_i$.

Proposition 4.1 shows to analyze the convergence properties of Algorithm 2, it suffices to analyze those of Algorithm 3. For Algorithm 3, we have the following theorem that is important in establishing our main convergence properties, see Appendix B.3 for the complete statement and its proof.

Theorem 4.2. *Consider (1). Suppose f_i's and g are bounded from below. Let $\{(X^t, Y^t, Z^t)\}$ be generated by Algorithm 3. Let $L_\beta(X, Y, Z)$ be defined as in (5). Given any $\Gamma, \Upsilon > 0$, let*

$$\begin{aligned} &H(X, Y, Z, X', Z') \\ &:= L_\beta(X, Y, Z) + \frac{\Gamma}{\tau\beta}\|Z - Z'\|^2 + \Upsilon\|X - X'\|^2 \end{aligned}$$

and $H_{t+1} := \mathbb{E}H(X^{t+1}, Y^{t+1}, Z^{t+1}, X^t, Z^t)$. Suppose $\beta > \sum_i L_i := L$. Let $\tau \in (0, \frac{1+\sqrt{5}}{2})$. Then, there exist $\beta > L, \Gamma, \Upsilon, \tau, \epsilon > 0$, such that for $t \geq 1$,

$$H_{t+1} \leq H_t - \delta\mathbb{E}\|X^t - X^{t-1}\|^2 - \frac{\beta}{2}\mathbb{E}\|Y^{t+1} - Y^t\|^2.$$

In addition, the sequence $\{H_t\}$ is convergent to some H_*.

Thanks to Theorem 4.2, we have the following property with respect to the successive changes. The proofs can be found in Appendix B.4.

Corollary 4.3. *Consider (1) and let $\{(X^t, Y^t, Z^t)\}$ be generated by Algorithm 3. Suppose assumptions in Theorem 4.2 hold. Then $\lim_t \mathbb{E}\|X^t - X^{t+1}\|^2 = \lim_t \mathbb{E}\|Y^{t+1} - Y^t\|^2 = \lim_t \mathbb{E}\|Z^{t+1} - Z^t\|^2 = \lim \mathbb{E}\|Y^t - X^t\|^2 = 0$.*

Remark 4.4. Combining Corollary 4.3 with Proposition 3.1 and Proposition 4.1, we see that the expectations of successive changes of $\{(x_1^t, \ldots, x_p^t, y^t, z_1, \ldots, z_p^t)\}$ generated by FIAME also converge to 0.

Based on Theorem 4.2, Algorithm 3 has the following complexity, see Appendix B.5 for a complete statement and proofs.

Theorem 4.5. *Consider (1) and let $\{(X^t, Y^t, Z^t)\}$ be generated by Algorithm 3. Suppose assumptions in Theorem 4.2 hold and let H_* be defined as in Theorem 4.2. Then there exists $B > 0$ such that*

$$\begin{aligned} &\frac{1}{1+T}\sum_{t=0}^{T} \mathbb{E}d^2(0, \nabla F(Y^{t+1} + \partial G(Y^{t+1}))) \\ &\leq \frac{B}{T+1}\left(L_\beta(X^0, Y^0, Z^0) - H_* + \|\nabla F(X^0)\|^2 + \epsilon_0\right). \end{aligned}$$

Combining Theorem 4.5 with Proposition 3.1 and Proposition 4.1, we immediately have the following complexity of FIAME.

Corollary 4.6. *Let $\{(x_1^t, \ldots, x_p^t, y^t, z_1^t, \ldots, z_p^t\}$ be generated by Algorithm 2. Let (X^t, Y^t, Z^t) be defined as in Proposition 4.1. Suppose assumptions in Theorem 4.2 hold. Let H^* be defined as in Theorem 4.2. Then there exists $B > 0$ such that*

$$\begin{aligned} &\frac{1}{1+T}\sum_{t=0}^{T} \mathbb{E}d^2\left(0, \sum_i \nabla f_i(y^{t+1}) + \partial g(y^{t+1})\right) \\ &\leq \frac{Bp}{T+1}\left(L_\beta(X^0, Y^0, Z^0) - H_* + \|\nabla F(X^0)\|^2 + \epsilon_0\right). \end{aligned}$$

4.1. Convergence Properties in the Deterministic Case

In this section, we further investigate the convergence properties of FIAME when (9) is deterministic. Still, regarding the relationship between FIAME and IAME given in Proposition 4.1, we analyze IAME with (11) being deterministic. We first show the properties of the set of accumulation points of $\{(X^t, Y^t, Z^t, X^{t-1}, Z^{t-1})\}$. The proofs are provided in Appendix B.5.1.

Proposition 4.7. *Consider (1) and let $\{(X^t, Y^t, Z^t)\}$ be generated by Algorithm 3 with (11) being deterministic. Suppose assumptions in Theorem 4.2 hold. Suppose $\{(X^t, Y^t, Z^t)\}$ is bounded. Then any accumulation point of $\{Y^t\}$ is a stationary point of (3).*

Combining Proposition 4.7 with Proposition 3.1 and Proposition 4.1, we immediately have the subsequential convergence of the sequence generated by FIAME.

Corollary 4.8. *Let $\{(x_1^t, \ldots, x_p^t, y^t, z_1^t, \ldots, z_p^t)\}$ be generated by Algorithm 2 with (9) being deterministic. Let*

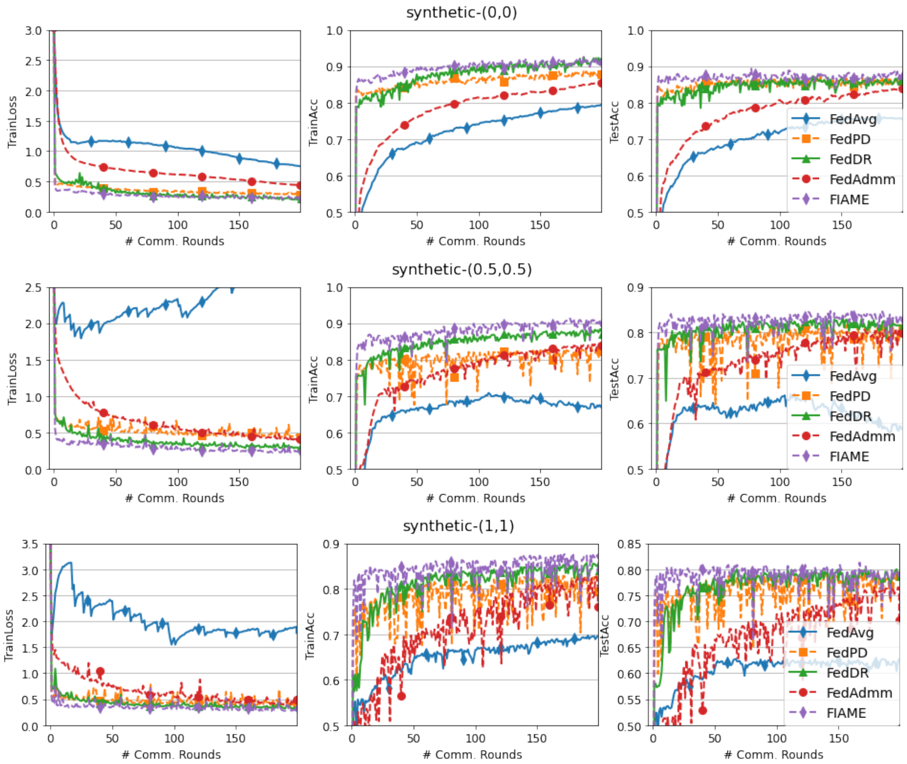

Figure 1. Results on Synthetic-{(0,0), (0.5, 0.5), (1,1)} dataset.

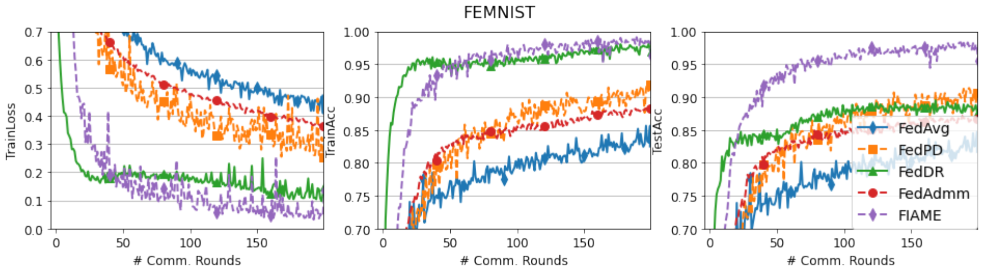

Figure 2. Results on FEMNIST dataset.

(X^t, Y^t, Z^t) *be defined as in Proposition 4.1. Suppose assumptions in Theorem 4.7 hold. Then any accumulation point of* $\{y^t\}$ *is a stationary point of* (1).

Next, we present the convergence rate of the sequence generated by Algorithm 3. The proofs are given in Appendix B.5.2.

Theorem 4.9. *Consider* (1) *and let* $\{(X^t, Y^t, Z^t)\}$ *be generated by Algorithm 3 with* (11) *being deterministic. Suppose assumptions in Theorem 4.2 hold. Let* H *be defined as in Theorem 4.2 and suppose* H *is a KL function with exponent* $\alpha \in [0, 1)$. *Then* $\{(X^t, Y^t, Z^t)\}$ *converges globally. Denoting* $(X^*, Y^*, Z^*) := \lim_t (X^t, Y^t, Z^t)$ *and* $d_s^t := \|(X^t, Y^t, Z^t) - (X^*, Y^*, Z^*)\|$, *then the followings hold.*

(i) If $\alpha = 0$, *then* $\{d_s^t\}$ *converges finitely.*

(ii) If $\alpha \in (0, \frac{1}{2}]$, *then there exist* $b > 0$, $t_1 \in \mathbb{N}$ *and* $\rho_1 \in (0, 1)$ *such that* $d_s^t \le b\rho_1^t$ *for* $t \ge t_1$.

(iii) If $\alpha \in (\frac{1}{2}, 1)$, *then there exist* t_2 *and* $c > 0$ *such that* $d_s^t \le ct^{-\frac{1}{4\alpha - 2}}$ *for* $t \ge t_2$.

Remark 4.10. Combining Theorem 4.9 with Proposition 3.1 and Proposition 4.1 immediately shows that $\{y^t\}$ generated by FIAME has the same convergence rate as $\{Y^t\}$.

Based on Theorem 4.9, we have the following convergence results on the updates at the server in FIAME. The proofs are provided in Appendix B.5.3.

Corollary 4.11. *Let* $\{(x_1^t, \ldots, x_p^t, y^t, z_1^t, \ldots, z_p^t)\}$ *be generated by Algorithm 2 with* (9) *being deterministic. Let*

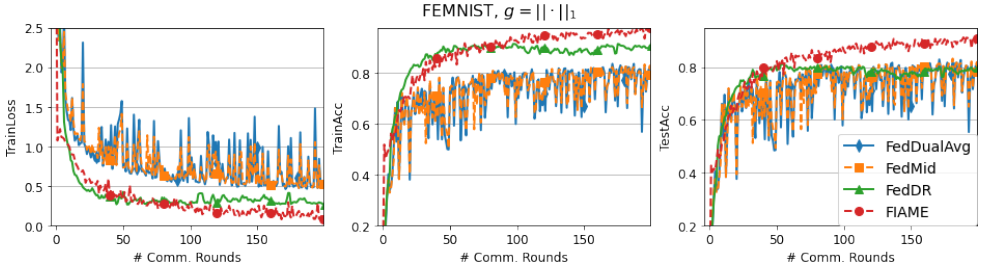

Figure 3. Results on FEMNIST dataset with $\mathbb{L}_1$-norm.

(X^t, Y^t, Z^t) be defined as in Proposition 4.1. Denote $d^t := d^2(0, \sum_i \nabla f_i(y^t) + \partial g(y^t))$. Let assumptions in Theorem 4.9 hold. Then

(i) If $\alpha = 0$, $\{d^{t+1}\}$ converges to 0 in finite iterations.

(ii) If $\alpha \in (0, \frac{1}{2}]$, there exist $a > 0$, $t_1 \in \mathbb{N}$ and $\rho_1 \in (0,1)$ such that $d^t \leq ap\rho_1^t$ for $t \geq t_1$.

(iii) If $\alpha \in (\frac{1}{2}, 1)$, there exist $t_2 \in \mathbb{N}$ and c such that $d^t \leq cpt^{-\frac{1-\alpha}{2\alpha-1}}$ for $t \geq t_2$.

5. Experimental Results

To evaluate the performance of our proposed FIAME algorithm, we conduct experiments on both realistic and synthetic datasets. When $g = 0$ in (1), we compare our algorithm with FedDR(Tran-Dinh et al., 2021), FedPD (Zhang et al., 2021), FedAvg (McMahan et al., 2017b), FedAdmm (Zhou & Li, 2022b). When $g = \lambda\|\cdot\|_1$ for some $\lambda \in \mathbb{R}_{++}$, we compare our algorithm with FedMid (Yuan et al., 2021), FedDualAvg (Yuan et al., 2021), and FedDR. For evaluation metrics, we use training loss, training accuracy, and test accuracy.

Implementation. For FedDR, FedPD, we refer to the code provided in (Tran-Dinh et al., 2021), and we also reimplement the FedAdmm based on them. All experiments are running on the Linux-based server with the configuration: 8xA6000 GPU with 48GB memory each.

Models and hyper-parameters selection. Following FedDR (Tran-Dinh et al., 2021), we choose the neural network as our model, and the details are deferred to Appendix§C. To be in accordance with the theoretical analysis, we sample all the clients to perform updates for our algorithm in each communication round. Since hyper-parameters have a large effect on the performance of different algorithms, we pick up them carefully and show the best results for each algorithm.

Results on synthetic datasets. Following the data generation process on (Li et al., 2020a; Tran-Dinh et al., 2021), we generate three datasets: `synthetic-`$\{$`(0,0), (0.5, 0.5), (1,1)`$\}$. All agents perform updates at each communication round. Our algorithm is compared using synthetic datasets in both iid and non-iid settings. The performance of five algorithms on non-iid synthetic datasets is shown as Figure 1. Our algorithm can achieve better results than FedPD, FedAdmm, FedAvg, and FedDR on all three synthetic datasets.

Results on FEMNIST dataset. FEMNIST (Cohen et al., 2017; Caldas et al., 2018) dataset is a more complex and federated extended MNIST. It has 62-class (26 upper-case and 26 lower-case letters, 10 digits) and the data is distributed to 200 devices. Figure 2 depicts the results of all 5 algorithms on FEMNIST. As it shows, FIAME can achieve comparable training accuracy and loss value with FedDR. In comparison with FedAdmm, FedPD, and FedAvg, FIAME has a significant improvement in both training accuracy and loss value. Our algorithm can also work much better with test accuracy than the other 4 algorithms, which verifies its superb generalization ability.

Results with the $\mathbb{L}_1$ norm. Following FedDR (Tran-Dinh et al., 2021), we also consider the composite setting with $g(x) := 0.01\|x\|_1$ to verify our algorithm by selecting different learning rates and the number of local SGD epochs. We conduct the experiment on the FEMNIST dataset and we show the results as Figure 3. As we can see from the training loss and training accuracy, FIAME has competitive efficiency with FedDR and outperform FedDualAvg and FedMid. In addition, in testing accuracy, FIAME outperforms all the other methods.

6. Conclusion

In this paper, we propose a federated inexact ADMM with a combination of absolute and relative criteria. For our new method, we show it has the best know complexity but uses weaker assumptions on the absolute criterion. In addition, we obtain the global convergence of the generated sequence in the deterministic case. Under KL assumptions, we show that the convergence rate of the generated sequence can be sublinear, or linear or even finite. Our experiments show the proposed method consistently outperforms the state-of-art methods, especially in testing accuracy.

References

Alves, M. M., Eckstein, J., Geremia, M., and Melo, J. G. Relative-error inertial-relaxed inexact versions of douglas-rachford and ADMM splitting algorithms. *Comput. Optim. Appl.*, 75(2):389–422, 2020.

Attouch, H. and Bolte, J. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program.*, 116(1-2):5–16, 2009.

Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-lojasiewicz inequality. *Math. Oper. Res.*, 35(2):438–457, 2010.

Attouch, H., Bolte, J., and Svaiter, B. F. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized gauss-seidel methods. *Math. Program.*, 137 (1-2):91–129, 2013.

Bao, Y., Crawshaw, M., Luo, S., and Liu, M. Fast composite optimization and statistical recovery in federated learning. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, 2022.

Bolte, J., Sabach, S., and Teboulle, M. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, 146(1-2):459–494, 2014.

Borwein, J. M., Li, G., and Tam, M. K. Convergence rate analysis for averaged fixed point iterations in common fixed point problems. *SIAM J. Optim.*, 27(1):1–33, 2017.

Caldas, S., Wu, P., Li, T., Konečný, J., McMahan, H. B., Smith, V., and Talwalkar, A. LEAF: A benchmark for federated settings. *CoRR*, abs/1812.01097, 2018.

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

Cohen, G., Afshar, S., Tapson, J., and Van Schaik, A. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.

Eckstein, J. and Yao, W. Approximate ADMM algorithms derived from lagrangian splitting. *Comput. Optim. Appl.*, 68(2):363–405, 2017.

Eckstein, J. and Yao, W. Relative-error approximate versions of douglas-rachford splitting and special cases of the ADMM. *Math. Program.*, 170(2):417–444, 2018.

Elgabli, A., Issaid, C. B., Bedi, A. S., Rajawat, K., Bennis, M., and Aggarwal, V. Fednew: A communication-efficient and privacy-preserving newton-type method for federated learning. In *International Conference on Machine Learning, ICML 2022, 17-23 July , Baltimore, Maryland, USA*, 2022.

Fan, Z., Wang, Y., Yao, J., Lyu, L., Zhang, Y., and Tian, Q. Fedskip: Combatting statistical heterogeneity with federated skip aggregation. In Zhu, X., Ranka, S., Thai, M. T., Washio, T., and Wu, X. (eds.), *IEEE International Conference on Data Mining, ICDM, Orlando, FL, USA, November 28 - Dec. 1*, pp. 131–140. IEEE, 2022.

Gong, Y., Li, Y., and Freris, N. M. Fedadmm: A robust federated deep learning framework with adaptivity to system heterogeneity. In *38th IEEE International Conference on Data Engineering, ICDE 2022, Kuala Lumpur, Malaysia, May 9-12,*, 2022.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K. A., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14(1-2):1–210, 2021.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In Frasconi, P., Landwehr, N., Manco, G., and Vreeken, J. (eds.), *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, Proceedings, Part I*, 2016.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. SCAFFOLD: stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July, Virtual Event*, 2020.

Li, G. and Pong, T. K. Douglas-rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. *Math. Program.*, 159(1-2):371–401, 2016.

Li, G. and Pong, T. K. Calculus of the exponent of kurdyka-łojasiewicz inequality and its applications to linear convergence of first-order methods. *Found. Comput. Math.*, 18(5):1199–1232, 2018.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. In Dhillon, I. S., Papailiopoulos, D. S., and Sze, V. (eds.), *Proceedings of Machine Learning and Systems 2020, MLSys 2020, Austin, TX, USA, March 2-4*, 2020a.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of fedavg on non-iid data. In *8th International Conference on Learning Representations, ICLR, Addis Ababa, Ethiopia, April 26-30*, 2020b.

Liu, Y., Xu, Y., and Yin, W. Acceleration of primal-dual methods by preconditioning and simple subproblem procedures. *J. Sci. Comput.*, 86(2):21, 2021.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April, Fort Lauderdale, FL, USA*, 2017a.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April, Fort Lauderdale, FL, USA*, 2017b.

Nemirovski, A., Juditsky, A. B., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2009.

Pathak, R. and Wainwright, M. J. FedSplit: an algorithmic framework for fast federated optimization. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12*, 2020.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1*, 2012.

Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7*, 2021.

Rockafellar, R. Monotone operators and the proximal point algorithm. *SIAM Journal of Control and Optimization*, 14:877–898, 1976.

Rockafellar, R. T. and Wets, R. J. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1998.

Tran-Dinh, Q., Pham, N. H., Phan, D. T., and Nguyen, L. M. FedDR - randomized douglas-rachford splitting algorithms for nonconvex federated composite optimization. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14*, 2021.

Xie, J. On inexact admms with relative error criteria. *Comput. Optim. Appl.*, 71(3):743–765, 2018.

Xie, J., Liao, A., and Yang, X. An inexact alternating direction method of multipliers with relative error criteria. *Optim. Lett.*, 11(3):583–596, 2017.

Yuan, H., Zaheer, M., and Reddi, S. J. Federated composite optimization. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July*, 2021.

Yue, S., Ren, J., Xin, J., Lin, S., and Zhang, J. Inexact-admm based federated meta-learning for fast and continual edge learning. In *MobiHoc '21: The Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing, Shanghai, China, 26-29 July*, 2021.

Zeng, L., Yu, P., and Pong, T. K. Analysis and algorithms for some compressed sensing models based on L1/L2 minimization. *SIAM J. Optim.*, 31(2):1576–1603, 2021.

Zhang, X., Hong, M., Dhople, S. V., Yin, W., and Liu, Y. Fedpd: A federated learning framework with adaptivity to non-iid data. *IEEE Trans. Signal Process.*, 69:6055–6070, 2021.

Zhou, S. and Li, G. Y. Communication-efficient admm-based federated learning. *CoRR*, abs/2110.15318, 2021. URL `https://arxiv.org/abs/2110.15318`.

Zhou, S. and Li, G. Y. Federated learning via inexact ADMM. *CoRR*, abs/2204.10607, 2022a.

Zhou, S. and Li, G. Y. Federated learning via inexact ADMM. *CoRR*, abs/2204.10607, 2022b. doi: 10.48550/arXiv.2204.10607. URL `https://doi.org/10.48550/arXiv.2204.10607`.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67(2):301–320, 2005.

A. Proofs of results in Section 3

A.1. Proofs of Proposition 3.1

Note that

$$\mathfrak{C} = \{(x_1, \ldots, x_p) :\ x_1 - x_2 = 0,\ x_2 - x_3 = 0,\ \ldots,\ x_{p-1} - x_p = 0\}.$$

Using Theorem 6.14 of (Rockafellar & Wets, 1998), we have

$$N_{\mathfrak{C}} = \left\{ \sum_{i=1}^{p-1} \lambda_i (0, \ldots, 0, \underbrace{\mathbf{1}}_{\text{the } i_{\text{th}} \text{ coordinate}}, -\mathbf{1}, 0, \ldots, 0) :\ (\lambda_1, \ldots, \lambda_{p-1}) \in \mathbb{R}^{p-1} \right\},$$

where $\mathbf{1}$ is the vector in $\mathbb{R}^p$ whose coordinates are all one.

This together with Corollary 10.9, Proposition 10.5 shows that for any $Y \in \mathrm{dom}\partial G$,

$$\partial G(Y) = \left\{ (\xi, 0, \ldots, 0) + \sum_{i=1}^{p-1} \lambda_i (0, \ldots, 0, \underbrace{\mathbf{1}}_{\text{the } i_{\text{th}} \text{ coordinate}}, -\mathbf{1}, 0, \ldots, 0) :\ \xi \in \partial g(y_1),\ (\lambda_1, \ldots, \lambda_{p-1}) \in \mathbb{R}^{p-1} \right\}. \tag{14}$$

Suppose $Y^* = (y_1^*, \ldots, y_p^*)$ is a stationary point of (3). Then $Y^* \in \mathrm{dom}\partial G \subseteq \mathrm{dom} G$. Thus, $y_1^* = \cdots = y_p^*$. In addition, it holds that

$$0 \in \nabla F(Y^*) + \partial G(Y^*) = (\nabla f_1(y^*), \ldots, \nabla f_p(y^*)) + (\partial g(y_1^*), 0, \ldots, 0) + \sum_{i=1}^{p} \lambda_i (0, \ldots, 0, \underbrace{\mathbf{1}}_{\text{the } i_{\text{th}} \text{ coordinate}}, -\mathbf{1}, 0, \ldots, 0),$$

where the second equality uses (14) together with Exercise 8.8 and Proposition 10.5 of (Rockafellar & Wets, 1998). The above relation is equivalent to

$$\begin{aligned}
&0 \in \nabla f_1(y^*) + \partial g(y_1^*) + \lambda_1 \mathbf{1} \\
&0 = \nabla f_2 - \lambda_1 \mathbf{1} + \lambda_2 \mathbf{1} \\
&\vdots \\
&0 = \nabla f_{p-1} - \lambda_{p-2} \mathbf{1} + \lambda_{p-1} \mathbf{1} \\
&0 = \nabla f_p(y^*) - \lambda_{p-1} \mathbf{1}.
\end{aligned} \tag{15}$$

Substituting λ_1 in (15) using the rest equality in the above relation, we have that

$$0 \in \sum_i \nabla f_i(y^*) + \partial g(y_1^*).$$

Thus y^* is a stationary point of (1).

Now, suppose $Y = (y_1, \ldots, y_p)$ is a ε-stationary point of (3). Then $Y \in \mathrm{dom}\partial G \subseteq \mathrm{dom} G$. Thus, $y_1 = \cdots = y_p$ and

$$\varepsilon \geq d^2(0, \nabla F(Y) + \partial G(Y)). \tag{16}$$

Using (14) and Proposition 10.5 of (Rockafellar & Wets, 1998), we have that

$$\begin{aligned}
&d^2(0, \nabla F(Y) + \partial G(Y)) \\
&= \min_{\xi \in \partial g(y_1), \lambda \in \mathbb{R}^{p-1}} \|\nabla f_1(y_1) + \xi + \lambda_1 \mathbf{1}\|^2 + \sum_{i=2}^{p-2} \|\nabla f_1(y_1) + \lambda_i \mathbf{1} - \lambda_{i-1} \mathbf{1}\|^2 + \|\nabla f_p(y_1) - \lambda_{p-1} \mathbf{1}\|^2 \\
&\geq \min_{\xi \in \partial g(y_1), \lambda \in \mathbb{R}^{p-1}} \frac{1}{p} \| \sum_i \nabla f_i(y_1) + \xi\|^2\|^2 = \min_{\xi \in \partial g(y_1)} \frac{1}{p} \| \sum_i \nabla f_i(y_1) + \xi\|^2\|^2 \\
&= \frac{1}{p} d^2(0, \sum_i \nabla f_i(y_1) + \partial g(y_1)).
\end{aligned} \tag{17}$$

This together with (16) shows that y_1 is a $p\varepsilon$-stationary point.

A.2. Proofs of Proposition 3.2

The problem in updating Y^{t+1} in (6) is a constrained problem:

$$\begin{aligned} &\min_{Y}\, g(y_1) + \left\langle Z^t, X^{t+1} - Y \right\rangle + \frac{\beta}{2}\|X^{t+1} - Y\|^2 \\ &\text{s.t. } y_2 = y_3 = \cdots = y_p = y_1. \end{aligned} \tag{18}$$

Since $\beta > L$, the objective in the above problem is strongly convex. Thus, there exists a unique solution $(y_1, y_2, \ldots, y_p)$ to (18). Denote the Lagrange multiplier for the above problem as $W = (w_1, w_2, \ldots, w_p)$. Then the Karush–Kuhn–Tucker condition for the above problem is

$$0 \in \partial g(y_1) - z_1^{t+1} - \beta(x_1^{t+1} - y_1) - \sum_{i=2}^{p} w_i \tag{19}$$

$$0 = -z_i^{t+1} + w_i - \beta(x_i^{t+1} - y_i),\ i = 2, \ldots, p \tag{20}$$

$$y_i = y_1,\ i = 2, \ldots, p. \tag{21}$$

Combining (20) with (21) gives

$$\sum_{i=2}^{p} w_i = \beta \sum_{i=2}^{p} (x_i^{t+1} - y_i) + \sum_{i=2}^{p} z_i^{t+1} = \beta \sum_{i=2}^{p} x_i^{t+1} - (p-1)\beta y_1 + \sum_{i=2}^{p} z_i^{t+1}.$$

This together with (19) shows that

$$\beta \sum_{i=2}^{p} x_i^{t+1} - (p-1)\beta y_1 + \sum_{i=2}^{p} z_i^{t+1} + z_1^{t+1} + \beta x_1^{t+1} \in \partial g(y_1) + \beta y_1,$$

which is equivalent to

$$\frac{1}{p} \sum_{i=1}^{p} (x_i^{t+1} + \frac{1}{\beta} z_i^{t+1}) \in \frac{1}{\beta p} \partial g(y_1) + y_1.$$

This implies that $y_1 \in \mathrm{Prox}_{\frac{1}{\beta p} g}(\frac{1}{p}\sum_{i=1}^{p}(x_i^{t+1} + \frac{1}{\beta} z_i^{t+1}))$. Recalling (21), we deduce that the solution of the problem in the third equation of (6) is $(y_1, \ldots, y_1)$ with $y_1 = \mathrm{Prox}_{\frac{1}{\beta p} g}(\frac{1}{p}\sum_{i=1}^{p}(x_i^{t+1} + \frac{1}{\beta} z_i^{t+1})))$.

A.3. Proof of Proposition 3.3

Since H is strongly convex, there exists unique solution x^* for $\min h$. Using Lemma 1 of (Rakhlin et al., 2012), we have that

$$\mathbb{E}\|\nabla h(x^k)\|^2 \leq \mathbb{E} L^2 \|x^k - x^*\|^2 \leq L^2 \frac{4W^2}{\mu^2 k}. \tag{22}$$

Thus, for any $\epsilon' > 0$, there exists k such that $\mathbb{E}\|\nabla h(x^k)\|^2 < \epsilon'$.

On the other hand, since $2a^2 \geq (a+b)^2 - 2b^2$ for any vectors a and b, we have that

$$\epsilon \mathbb{E}\|x^{k+1} - \bar{x}\|^2 \geq \frac{1}{2}\epsilon\|x^* - \bar{x}\|^2 - \epsilon\mathbb{E}\|x^{k+1} - x^*\|^2 \geq \frac{1}{2}\epsilon\|x^* - \bar{x}\|^2 - \frac{4W^2}{\mu^2 k} \tag{23}$$

where the second inequality uses Lemma 1 of (Rakhlin et al., 2012) again. Thus, there exists large k such that $\frac{1}{2}\epsilon\|x^* - \bar{x}\|^2 - \frac{4W^2}{\mu^2 k} \geq \frac{1}{4}\epsilon\|x^* - \bar{x}\|^2$. Since $\bar{x} \neq z^*$, $\|x^* - \bar{x}\|^2 > 0$. This together with (22) and (23), we deduce that there exists large k such that

$$\mathbb{E}\|\nabla h(x^k)\|^2 \leq \min\{\epsilon', \frac{1}{4}\epsilon\|x^* - \bar{x}\|^2\} \leq \min\{\epsilon', \epsilon\mathbb{E}\|x^{k+1} - \bar{x}\|^2\}.$$

B. Details and proofs of results in Section 4

To prove the results in Section 4, we first present the following well known facts for strongly convex functions, see Theorem 2 in (Karimi et al., 2016) for example.

Proposition B.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a strongly convex function with modulus μ. Suppose in addition that f is smooth and has Lipschitz continuous gradient with modulus L. Then there exists unique minimizer x^* that minimize f and it holds that*

$$\|\nabla f(x)\|^2 \geq 2\mu\left(f(x) - f(x^*)\right) \geq \mu^2 \|x - x^*\|^2.$$

B.1. Proof of Proposition 4.1

Proposition A.2 shows that $Y^t = (y^t, \dots, y^t)$. It remains to show that $X^t = (x_1^t, x_2^t, \dots, x_p^t)$ satisfies (10). Sine x_i^t satisfies (9), it holds that

$$\mathbb{E}_t \|\nabla_X L_\beta(X, Y^{t+1}, Z^t)\|^2 = \sum_{i=1}^p \mathbb{E}_t \|\nabla_x L_{\beta,i}(x_i^{t+1}, y_i^t, z_i^{t+1})\|^2 \leq \sum_{i=1}^p \epsilon_i \min\{\|x_i^t - x_i^{t-1}\|^2, \mathbb{E}_t\|x_i^{t+1} - x_i^t\|^2\}$$

$$\leq \max_i \epsilon_i \min\left\{\sum_{i=1}^p \|x_i^t - x_i^{t-1}\|^2, \mathbb{E}_t \sum_{i=1}^p \|x_i^{t+1} - x_i^t\|^2\right\} = \max_i \epsilon_i \min\left\{\|X^t - X^{t-1}\|^2, \mathbb{E}_t\|X^{t+1} - X^t\|^2\right\}.$$

Thus, X^t satisfies (10). This shows that Algorithm 2 is a special case of Algorithm 3.

B.2. Details and proofs of Theorem 4.2

Before proving Theorem 4.2, we first present several properties of the subproblem (10).

Proposition B.2. *Consider (1). Let (X^t, Y^t, Z^t) be generated by Algorithm 3. Let $\beta \geq \sum_i L_i$. Denote $X_\star^{t+1} := \min_X L_\beta(X, Y^t, Z^{t+1})$.[3] Then the following statements hold:*

(i) Denote $e^t = X^{t+1} - X_\star^{t+1}$. Then there exists $\xi^{t+1} \in \partial G(Y^{t+1})$ such that

$$0 = \nabla F(X_\star^{t+1}) + Z^t + \beta(X_\star^{t+1} - Y^t) \Leftrightarrow -Z^t - \beta(X^{t+1} - e^{t+1} - Y^t) = \nabla F(X_\star^{t+1}) \tag{24}$$

and

$$0 = \xi^{t+1} - Z^{t+1} - \beta(X^{t+1} - Y^{t+1}) \tag{25}$$

(ii) It holds that

$$Z^{t+1} = (1-\tau)Z^t + \beta\tau e^{t+1} + \tau\nabla F(X_\star^{t+1}) \tag{26}$$

(iii) It holds that

$$\mathbb{E}\|e^{t+1} + e^t\|^2 \leq \frac{4}{(\beta - L)^2}\epsilon\mathbb{E}\|X^t - X^{t-1}\|^2. \tag{27}$$

Proof. (i) follows from the first optimality condition of (10) and (13). Combining (24) with (12), we have that

$$\begin{aligned}
&-Z^t - \frac{1}{\tau}(Z^{t+1} - Z^t) + \beta e^{t+1} = -Z^t - \beta(X^{t+1} - e^{t+1} - Y^t) = \nabla F(X_\star^{t+1}).\\
\Leftrightarrow& Z^{t+1} = (1-\tau)Z^t + \beta\tau e^{t+1} + \tau\nabla F(X_\star^{t+1}).
\end{aligned}$$

Now, we bound $\mathbb{E}\|e^t - e^{t+1}\|^2$ in the above inequality. Using Proposition B.1, it holds that

$$\begin{aligned}
&\mathbb{E}\|e^{t+1} + e^t\|^2 \leq (\mathbb{E}\|e^{t+1}\| + \mathbb{E}\|e^t\|)^2 \leq 2\mathbb{E}\|e^{t+1}\|^2 + 2\mathbb{E}\|e^t\|^2\\
&\leq \frac{2}{(\beta - L)^2}\mathbb{E}\|\nabla_x L_\beta(X^{t+1}, Y^t, Z^{t+1})\|^2 + \frac{2}{(\beta - L)^2}\mathbb{E}\|\nabla_x L_\beta(X^t, Y^{t-1}, Z^t)\|^2\\
&\overset{(a)}{\leq} \frac{2}{(\beta - L)^2}\left(\epsilon \min\{\mathbb{E}\|X^t - X^{t-1}\|^2, \mathbb{E}\|X^{t+1} - X^t\|^2\} + \epsilon\min\{\mathbb{E}\|X^{t-1} - X^{t-2}\|^2, \epsilon\mathbb{E}\|X^t - X^{t-1}\|^2\}\right)\\
&\leq \frac{4}{(\beta - L)^2}\epsilon\mathbb{E}\|X^t - X^{t-1}\|^2.
\end{aligned}$$

[3] The existence and uniqueness of $X_\star^{t+1}$ are thanks to $\beta \geq \sum_i L_i$ and Proposition B.1.

where (a) uses (11) and the last inequality uses the setting in the input of Algorithm 3 that $\{\epsilon^t\}$ is nonincreasing. □

Now, we are ready to give the details of Theorem 4.2.

Theorem B.3. *Consider* (1)*. Let* (X^t, Y^t, Z^t) *be generated by Algorithm* 3*. Suppose* $\beta \geq \sum_i L_i := L$*. Let* $L_\beta(X, Y, Z)$ *be defined as in* (5)*. Let* $\tau \in (0, \frac{1+\sqrt{5}}{2})$ *and* $\zeta \in (0, 1)$*. Denote* $\Gamma := \max\{\frac{1-\tau}{\tau}, \frac{\tau^2-\tau}{1-\tau^2+\tau}\}$, $\Theta = \max\{2\tau\beta^2 + 2\tau L^2(1 + \kappa^{-2}), 2\frac{\tau^3}{1-\tau^2+\tau}\beta^2 + 2\frac{\tau^4}{1-\tau^2+\tau}L^2(1+\kappa^{-2})\}$ *and* $\Lambda := \max\{2\frac{\tau^4}{1-\tau^2+\tau}L^2(1+\kappa^2), 2\tau L^2(1+\kappa^2)\}$*. Then, there exists* $\beta > L$ *great enough such that* $\frac{\beta-L}{2} - \frac{\Lambda}{\tau\beta} > 0$*. In addition, there exists* ϵ *small enough such that* $\frac{\beta-L}{2}(1-\zeta^2) - \frac{\Lambda}{\tau\beta} > \frac{1}{2}\frac{\zeta^{-2}-1}{\beta-L}\epsilon + \frac{1}{2(\beta-L)}\epsilon + \frac{\Theta}{\tau\beta}\frac{4}{(\beta-L)^2}\epsilon$*. Denote* $\Upsilon := \frac{\beta-L}{2}(1-\zeta^2) - \frac{\Lambda}{\tau\beta} - \frac{1}{2}\frac{\zeta^{-2}-1}{\beta-L}\epsilon - \frac{1}{2(\beta-L)}\epsilon - \frac{\Theta}{\tau\beta}\frac{4}{(\beta-L)^2}\epsilon$ *and* $\delta := \Upsilon - \frac{\Theta}{\tau\beta}\frac{4}{(\beta-L)^2}\epsilon$*. Given any* $\Gamma, \Upsilon > 0$*, defined*

$$H(X, Y, Z, X', Z') := L_\beta(X, Y, Z) + \frac{\Gamma}{\tau\beta}\|Z - Z'\|^2 + \Upsilon\|X - X'\|^2.$$

and $H_{t+1} := \mathbb{E}H(X^{t+1}, Y^{t+1}, Z^{t+1}, X^t, Z^t)$*. Then for* $t \geq 1$*, it holds that*

$$H_{t+1} \leq H_t - \delta\mathbb{E}\|X^t - X^{t-1}\|^2 - \frac{\beta}{2}\mathbb{E}\|Y^{t+1} - Y^t\|^2. \tag{28}$$

In addition, the sequence $\{H_t\}$ *is convergence to some* H_**.*

Proof. We first show the existence of β and ϵ satisfying the assumptions in the statement. Since we assumes that $\beta > L$, recalling the definition of Θ and Λ, we have that $\Theta = \max\{2\tau\beta^2 + 2\tau L^2(1+\kappa^{-2}), 2\frac{\tau^3}{1-\tau^2+\tau}\beta^2 + 2\frac{\tau^4}{1-\tau^2+\tau}L^2(1+\kappa^{-2})\}$ and $\Lambda := \max\{2\frac{\tau^4}{1-\tau^2+\tau}L^2(1+\kappa^2), 2\tau L^2(1+\kappa^2)\}$.

$$\frac{\Lambda}{\tau\beta} = \max\left\{2\frac{\tau^3}{1-\tau^2+\tau}\frac{L^2}{\beta}(1+\kappa^2), 2\frac{L^2}{\beta}(1+\kappa^2)\right\} = \max\left\{\frac{\tau^3}{1-\tau^2+\tau}, 1\right\}\frac{2L^2}{\beta}(1+\kappa^2).$$

This implies that there exists $\beta > L$ is great enough such that

$$\frac{\beta - L}{2} - \frac{\Lambda}{\tau\beta} > 0.$$

Recalling the definition of Θ, it holds that

$$\frac{\Theta}{\tau\beta}\frac{4}{(\beta-L)^2}\epsilon = \frac{4\epsilon}{\beta(\beta-L)^2}\max\{2\beta^2 + 2L^2(1+\kappa^{-2}), 2\frac{\tau^2}{1-\tau^2+\tau}\beta^2 + 2\frac{\tau^3}{1-\tau^2+\tau}L^2(1+\kappa^{-2})\}$$

Thus, there exists ϵ small enough such that

$$\frac{\beta-L}{2}(1-\zeta^2) - \frac{\Lambda}{\tau\beta} > \frac{1}{2}\frac{\zeta^{-2}-1}{\beta-L}\epsilon + \frac{1}{2(\beta-L)}\epsilon + \frac{\Theta}{\tau\beta}\frac{4}{(\beta-L)^2}\epsilon.$$

Recall that $\delta = \frac{\beta-L}{2}\left(1 - \zeta^2 - \frac{\epsilon(\zeta^{-2}-1)}{(\beta-L)^2}\right) - \frac{1}{2(\beta-L)}\epsilon - \frac{\Lambda}{\tau\beta} - \frac{\Theta}{\tau\beta}\frac{4}{(\beta-L)^2}\epsilon$. Then $\delta > 0$.

Now we prove (28). Since $\beta > \sum_i L_i$, we know that the objective in (10) is strongly convex with modulus $\beta - L$. Using Proposition B.1, it holds that

$$\begin{aligned}
\mathbb{E}_t L_\beta(X^{t+1}, Y^t, Z^t) &\leq L_\beta(X_\star^{t+1}, Y^t, Z^t) + \frac{1}{2(\beta-L)}\mathbb{E}_t\|\nabla_X L(X^{t+1}, Y^t, Z^t)\|^2 \\
&\leq L_\beta(X^t, Y^t, Z^t) - \frac{\beta-L}{2}\|X^t - X_\star^{t+1}\|^2 + \frac{1}{2(\beta-L)}\mathbb{E}_t\|\nabla_X L(X^{t+1}, Y^t, Z^t)\|^2 \\
&\leq L_\beta(X^t, Y^t, Z^t) - \frac{\beta-L}{2}\|X^t - X_\star^{t+1}\|^2 + \frac{1}{2(\beta-L)}\mathbb{E}_t\|\nabla_X L(X^{t+1}, Y^t, Z^t)\|^2
\end{aligned} \tag{29}$$

where the second inequality is because $X_\star^{t+1}$ is the minimizer of the strongly convex function $L(X, Y^t, Z^t)$. Now, we bound $\|X^t - X_\star^{t+1}\|^2$ in the above inequality. Note that

$$
\begin{aligned}
&\mathbb{E}_t\|X^t - X_\star^{t+1}\|^2 = \mathbb{E}_t\|X^t - X^{t+1}\|^2 - \mathbb{E}_t\left\langle X^t - X^{t+1}, X^{t+1} - X_\star^{t+1}\right\rangle + \mathbb{E}_t\|X^{t+1} - X_\star^{t+1}\|^2\\
&\geq \mathbb{E}_t\|X^t - X^{t+1}\|^2 - \mathbb{E}_t\|X^t - X^{t+1}\|\|X^{t+1} - X_\star^{t+1}\| + \mathbb{E}_t\|X^{t+1} - X_\star^{t+1}\|^2\\
&\overset{\text{(a)}}{\geq} (1-\zeta^2)\mathbb{E}_t\|X^t - X^{t+1}\|^2 - (\zeta^{-2}-1)\mathbb{E}_t\|X^{t+1} - X_\star^{t+1}\|^2\\
&\geq \left(1-\zeta^2\right)\mathbb{E}_t\|X^t - X^{t+1}\|^2 - \frac{\epsilon(\zeta^{-2}-1)}{(\frac{1}{\beta}-L)^2}\mathbb{E}_t\|X^{t+1} - X^t\|^2\\
&= \left(1-\zeta^2 - \frac{\epsilon(\zeta^{-2}-1)}{(\beta-L)^2}\right)\mathbb{E}_t\|X^t - X^{t+1}\|^2,
\end{aligned}
$$

where $\zeta \in (0,1)$, (a) uses Young's inequality for products and the last inequality is because the strong convexity of the objective of (10) with modulus $\beta - L$, Proposition B.1 and (11).

Combining this inequality with (29), we obtain

$$
\begin{aligned}
&\mathbb{E}_t L_\beta(X^{t+1}, Y^t, Z^t)\\
&\leq L_\beta(X^t, Y^t, Z^t) - \frac{\beta - L}{2}\left(1-\zeta^2 - \frac{\epsilon(\zeta^{-2}-1)}{(\beta-L)^2}\right)\mathbb{E}_t\|X^t - X^{t+1}\|^2\\
&+ \frac{1}{2(\beta-L)}\mathbb{E}_t\|\nabla_X L(X^{t+1}, Y^t, Z^t)\|^2\\
&\leq L_\beta(X^t, Y^t, Z^t) - \frac{\beta - L}{2}\left(1-\zeta^2 - \frac{\epsilon(\zeta^{-2}-1)}{(\beta-L)^2}\right)\mathbb{E}_t\|X^t - X^{t+1}\|^2\\
&+ \frac{1}{2(\beta-L)}\epsilon\mathbb{E}_t\|X^{t+1} - X^t\|^2\\
&= L_\beta(X^t, Y^t, Z^t) - \left(\frac{\beta - L}{2}\left(1-\zeta^2 - \frac{\epsilon(\zeta^{-2}-1)}{(\beta-L)^2}\right) - \frac{1}{2(\beta-L)}\epsilon\right)\mathbb{E}_t\|X^t - X^{t+1}\|^2,
\end{aligned}
\tag{30}
$$

where the second inequality uses (11).

Next, using (12), we have

$$
L_\beta(X^{t+1}, Y^t, Z^{t+1}) - L_\beta(X^{t+1}, Y^t, Z^t) = \frac{1}{\tau\beta}\|Z^{t+1} - Z^t\|^2 \tag{31}
$$

When $\tau \in (0,1)$, combining (26) and the convexity of $\|\cdot\|^2$, we have that

$$
\begin{aligned}
&\|Z^{t+1} - Z^t\|^2 \leq (1-\tau)\|Z^t - Z^{t-1}\|^2 + \tau\|\beta(e^{t+1} - e^t) + \nabla(F(X_\star^{t+1}) - F(X_\star^t))\|^2\\
&\leq (1-\tau)\|Z^t - Z^{t-1}\|^2 + 2\tau\beta^2\|e^{t+1} - e^t\|^2 + 2\tau\|\nabla(F(X_\star^{t+1}) - F(X_\star^t))\|^2\\
&\leq (1-\tau)\|Z^t - Z^{t-1}\|^2 + 2\tau\beta^2\|e^{t+1} - e^t\|^2 + 2\tau L^2\|X_\star^{t+1} - X_\star^t\|^2,
\end{aligned}
$$

where the second inequality uses the Young's inequality for product, and the last inequality uses the Lipschitz continuity of ∇F. Rearranging the above inequality, we have that

$$
\begin{aligned}
&\|Z^{t+1} - Z^t\|^2\\
&\leq \frac{1-\tau}{\tau}\left(\|Z^t - Z^{t-1}\|^2 - \|Z^{t+1} - Z^t\|^2\right) + +2\tau\beta^2\|e^{t+1} - e^t\|^2 + 2\tau L^2\|X_\star^{t+1} - X_\star^t\|^2\\
&\leq \frac{1-\tau}{\tau}\left(\|Z^t - Z^{t-1}\|^2 - \|Z^{t+1} - Z^t\|^2\right) + 2\tau\beta^2\|e^{t+1} - e^t\|^2\\
&+ 2\tau L^2\left((1+\kappa^2)\|X^{t+1} - X^t\|^2 + (1+\kappa^{-2})\|e^{t+1} - e^t\|^2\right)\\
&= \frac{1-\tau}{\tau}\left(\|Z^t - Z^{t-1}\|^2 - \|Z^{t+1} - Z^t\|^2\right) + \left(2\tau\beta^2 + 2\tau L^2(1+\kappa^{-2})\right)\|e^{t+1} - e^t\|^2\\
&+ 2\tau L^2(1+\kappa^2)\|X^{t+1} - X^t\|^2,
\end{aligned}
\tag{32}
$$

where $\kappa > 0$ and the last inequality uses the definition of e^{t+1} and Young's inequality for products.

When $\tau \in (1, \frac{1+\sqrt{5}}{2})$, dividing both sides of (26) with τ, we have that

$$\frac{1}{\tau} Z^{t+1} = \frac{1-\tau}{\tau} Z^t + \beta e^{t+1} + \nabla F(X_\star^{t+1}).$$

This implies that

$$\begin{aligned}
&\left\| \frac{1}{\tau}(Z^{t+1} - Z^t) \right\|^2 = \left\| \frac{1-\tau}{\tau}(Z^t - Z^{t-1}) + \beta(e^{t+1} - e^t) + (\nabla F(X_\star^{t+1}) - \nabla F(X_\star^t)) \right\|^2 \\
&= \left\| (1 - \frac{1}{\tau})(Z^{t-1} - Z^t) + \beta(e^t - e^{t+1}) + (\nabla F(X_\star^t) - \nabla F(X_\star^{t+1})) \right\|^2 \\
&\overset{(a)}{\le} (1 - \frac{1}{\tau}) \left\| Z^{t-1} - Z^t \| \right|^2 + \frac{1}{\tau} \left\| \tau\beta(e^t - e^{t+1}) + \tau(\nabla F(X_\star^t) - \nabla F(X_\star^{t+1})) \right\|^2 \\
&\overset{\mathrm{b}}{\le} (1 - \frac{1}{\tau}) \left\| Z^{t-1} - Z^t \| \right|^2 + 2\tau\beta^2 \|e^t - e^{t+1}\|^2 + 2\tau^2 \left\| \nabla F(X_\star^t) - \nabla F(X_\star^{t+1}) \right\|^2 \\
&\le (1 - \frac{1}{\tau}) \left\| Z^{t-1} - Z^t \| \right|^2 + 2\tau\beta^2 \|e^t - e^{t+1}\|^2 + 2\tau^2 L^2 \left\| X_\star^t - X_\star^{t+1} \right\|^2,
\end{aligned}$$

where (a) is because $\frac{1}{\tau} \in (0, 1)$ and $\| \cdot \|^2$ is convex, (b) uses the young's inequality for product and the last inequality uses the Lipschitz continuity of F. Since $\tau \in (1, \frac{1+\sqrt{5}}{2})$, we have $1 - \tau^2 + \tau > 0$. Rearranging the above inequality and divide both sides with $1 - \tau^2 + \tau$, we have that

$$\begin{aligned}
&\left\| Z^{t+1} - Z^t \right\|^2 \\
&\le \frac{\tau^2 - \tau}{1 - \tau^2 + \tau} \left(\|Z^{t-1} - Z^t\|^2 - \|Z^{t+1} - Z^t\|^2 \right) + 2\frac{\tau^3}{1 - \tau^2 + \tau} \beta^2 \|e^t - e^{t+1}\|^2 \\
&+ 2\frac{\tau^4}{1 - \tau^2 + \tau} L^2 \left\| X_\star^t - X_\star^{t+1} \right\|^2 \\
&\le \frac{\tau^2 - \tau}{1 - \tau^2 + \tau} \left(\|Z^{t-1} - Z^t\|^2 - \|Z^{t+1} - Z^t\|^2 \right) + 2\frac{\tau^3}{1 - \tau^2 + \tau} \beta^2 \|e^t - e^{t+1}\|^2 \\
&+ 2\frac{\tau^4}{1 - \tau^2 + \tau} L^2 \left((1 + \kappa^2) \left\| X^t - X^{t+1} \right\|^2 + (1 + \kappa^{-2}) \|e^{t+1} - e^t\|^2) \right) \\
&= \frac{\tau^2 - \tau}{1 - \tau^2 + \tau} \left(\|Z^{t-1} - Z^t\|^2 - \|Z^{t+1} - Z^t\|^2 \right) + \left(2\frac{\tau^3}{1 - \tau^2 + \tau} \beta^2 + 2\frac{\tau^4}{1 - \tau^2 + \tau} L^2 (1 + \kappa^{-2}) \right) \|e^t - e^{t+1}\|^2 \\
&+ 2\frac{\tau^4}{1 - \tau^2 + \tau} L^2 (1 + \kappa^2) \left\| X^t - X^{t+1} \right\|^2,
\end{aligned} \tag{33}$$

where $\kappa > 0$ and the last inequality uses the definition of e^{t+1} and Young's inequality for products.

Combining (32) and (33), we have

$$\left\| Z^{t+1} - Z^t \right\|^2 \le \Gamma \left(\|Z^{t-1} - Z^t\|^2 - \|Z^{t+1} - Z^t\|^2 \right) + \Theta \|e^t - e^{t+1}\|^2 + \Lambda \left\| X^t - X^{t+1} \right\|^2, \tag{34}$$

where Γ, Θ and Λ are defined in the statement.

Now, combining (30), (31) and (34), we obtain that

$$
\begin{aligned}
&\mathbb{E}_t L_\beta(X^{t+1}, Y^t, Z^{t+1}) \\
&\leq L_\beta(X^t, Y^t, Z^t) - \left(\frac{\beta - L}{2} \left(1 - \zeta^2 - \frac{\epsilon(\zeta^{-2} - 1)}{(\beta - L)^2} \right) - \frac{1}{2(\beta - L)}\epsilon - \frac{\Lambda}{\tau\beta} \right) \mathbb{E}_t \|X^t - X^{t+1}\|^2 \\
&+ \frac{\Gamma}{\tau\beta} \left(\|Z^{t-1} - Z^t\|^2 - \|Z^{t+1} - Z^t\|^2 \right) + \frac{\Theta}{\tau\beta} \mathbb{E}_t \|e^t - e^{t+1}\|^2 + \frac{\Lambda}{\tau\beta} \mathbb{E}_t \left\|X^t - X^{t+1}\right\|^2 \\
&= L_\beta(X^t, Y^t, Z^t) - \left(\frac{\beta - L}{2} \left(1 - \zeta^2 - \frac{\epsilon(\zeta^{-2} - 1)}{(\beta - L)^2} \right) - \frac{1}{2(\beta - L)}\epsilon - \frac{\Lambda}{\tau\beta} \right) \mathbb{E}_t \|X^t - X^{t+1}\|^2 \\
&+ \frac{\Gamma}{\tau\beta} \left(\|Z^{t-1} - Z^t\|^2 - \mathbb{E}_t \|Z^{t+1} - Z^t\|^2 \right) + \frac{\Theta}{\tau\beta} \mathbb{E}_t \|e^t - e^{t+1}\|^2.
\end{aligned}
$$

Taking expectations with respect to $\mathcal{X}^t$, the above inequality induces

$$
\begin{aligned}
&\mathbb{E} L_\beta(X^{t+1}, Y^t, Z^{t+1}) \leq \mathbb{E} L_\beta(X^t, Y^t, Z^t) - \left(\frac{\beta - L}{2} \left(1 - \zeta^2 - \frac{\epsilon(\zeta^{-2} - 1)}{(\beta - L)^2} \right) - \frac{1}{2(\beta - L)}\epsilon - \frac{\Lambda}{\tau\beta} \right) \mathbb{E} \|X^t - X^{t+1}\|^2 \\
&+ \frac{\Gamma}{\tau\beta} \left(\mathbb{E} \|Z^{t-1} - Z^t\|^2 - \mathbb{E} \|Z^{t+1} - Z^t\|^2 \right) + \frac{\Theta}{\tau\beta} \mathbb{E} \|e^t - e^{t+1}\|^2.
\end{aligned} \tag{35}
$$

Combining (27) with (35), we obtain that

$$
\begin{aligned}
&\epsilon \mathbb{E} L_\beta(X^{t+1}, Y^t, Z^{t+1}) \\
&\leq \mathbb{E} L_\beta(X^t, Y^t, Z^t) - \left(\frac{\beta - L}{2} \left(1 - \zeta^2 - \frac{\epsilon(\zeta^{-2} - 1)}{(\beta - L)^2} \right) - \frac{1}{2(\beta - L)}\epsilon - \frac{\Lambda}{\tau\beta} \right) \mathbb{E} \|X^t - X^{t+1}\|^2 \\
&+ \frac{\Gamma}{\tau\beta} \left(\mathbb{E} \|Z^{t-1} - Z^t\|^2 - \mathbb{E} \|Z^{t+1} - Z^t\|^2 \right) + \frac{\Theta}{\tau\beta} \frac{4}{(\beta - L)^2} \epsilon \mathbb{E} \|X^t - X^{t-1}\|^2.
\end{aligned} \tag{36}
$$

Finally, using the definition of δ and Υ, (36) can be further passed to

$$
\begin{aligned}
&\mathbb{E} L_\beta(X^{t+1}, Y^t, Z^{t+1}) \\
&\leq \mathbb{E} L_\beta(X^t, Y^t, Z^t) - \delta \|X^t - X^{t-1}\|^2 \\
&+ \frac{\Gamma}{\tau\beta} \left(\mathbb{E} \|Z^{t-1} - Z^t\|^2 - \mathbb{E} \|Z^{t+1} - Z^t\|^2 \right) \\
&+ \Upsilon \left(\mathbb{E} \|X^t - X^{t-1}\|^2 - \mathbb{E} \|X^{t+1} - X^t\|^2 \right).
\end{aligned} \tag{37}
$$

Next, noting that Y^{t+1} is the minimizer of (13) and the objective of (13) is strongly convex, it holds that

$$
\mathbb{E} L_\beta(X^{t+1}, Y^{t+1}, Z^{t+1}) \leq \mathbb{E} L_\beta(X^{t+1}, Y^t, Z^{t+1}) - \frac{\beta}{2} \mathbb{E} \|Y^{t+1} - Y^t\|^2. \tag{38}
$$

Summing (38) and (37), we have that

$$
\begin{aligned}
&\mathbb{E} L_\beta(X^{t+1}, Y^{t+1}, Z^{t+1}) \\
&\leq \mathbb{E} L_\beta(X^t, Y^t, Z^t) - \delta \mathbb{E} \|X^t - X^{t-1}\|^2 + \frac{\Gamma}{\tau\beta} \left(\mathbb{E} \|Z^{t-1} - Z^t\|^2 - \mathbb{E} \|Z^{t+1} - Z^t\|^2 \right) \\
&+ \frac{\Theta}{\tau\beta} \frac{4}{(\beta - L)^2} \epsilon \left(\mathbb{E} \|X^t - X^{t-1}\|^2 - \mathbb{E} \|X^{t+1} - X^t\|^2 \right) - \frac{\beta}{2} \mathbb{E} \|Y^{t+1} - Y^t\|^2.
\end{aligned}
$$

Rearranging the above inequality and recalling the definition of $H(X, Y, Z, X', Z')$, we have that

$$
\mathbb{E} H(X^{t+1}, Y^{t+1}, Z^{t+1}, X^t, Z^t) \leq \mathbb{E} H(X^t, Y^t, Z^t, X^{t-1}, Z^{t-1}) - \delta \mathbb{E} \|X^t - X^{t-1}\|^2 - \frac{\beta}{2} \mathbb{E} \|Y^{t+1} - Y^t\|^2.
$$

Now we prove $\{H_t\}$ is convergent. Inequality (28) implies that $\{H_t\}$ is nonincreasing. Since F and G are bounded from below, we denote $W = \inf F + \inf G$. Now we show that $H_t \geq W$ for all t. Suppose to the contrary that there exists t_0 such that $H_{t_0} < W$. Since (28) implies H_t is nonincreasing, it hold that

$$\sum_{t \geq t_0}^{T} (H_t - W) \leq \sum_{t=1}^{t_0 - 1} (H_t - W) + (T - t_0 + 1)(H_{t_0} - W).$$

Thus

$$\lim_{T \to \infty} \sum_{t \geq t_0}^{T} (H_t - W) = -\infty. \tag{39}$$

On the other hand, using (37), for $t \geq 1$, it holds that

$$\begin{aligned}
&H_t - W \geq \mathbb{E} H(X^{t+1}, Y^{t+1}, Z^{t+1}, X^t, Z^t) - W \overset{\text{(a)}}{\geq} \mathbb{E} L_\beta(X^{t+1}, Y^t, Z^{t+1}) - W \\
&\geq \mathbb{E} F(X^{t+1}) + G(Y^t) + \left\langle X^{t+1} - Y^t, Z^{t+1} \right\rangle - W \\
&\geq \mathbb{E} \left\langle X^{t+1} - Y^t, Z^{t+1} \right\rangle \overset{\text{(b)}}{=} \mathbb{E} \left\langle Z^{t+1} - Z^t, Z^{t+1} \right\rangle = \mathbb{E} \|Z^{t+1}\|^2 - \mathbb{E} \|Z^t\|^2 + \mathbb{E} \|Z^{t+1} - Z^t\|^2 \\
&\geq \mathbb{E} \|Z^{t+1}\|^2 - \mathbb{E} \|Z^t\|^2.
\end{aligned}$$

where (a) makes use of the definition of H_t and L_β, (b) uses (12). Summing the above inequality from $t = 0$ to T and take T to the infinity, we have that

$$\lim_{T \to \infty} \sum_{t=1}^{T} (H_t - W) \geq \lim_{T \to \infty} \sum_{t=1}^{T} (\|Z^{t+1}\|^2 - \|Z^t\|^2) = \lim_{T \to \infty} (\mathbb{E} \|Z^{T+1}\|^2 - \mathbb{E} \|Z^0\|^2) \geq -\|Z^0\|^2 > -\infty,$$

which contradicts with (39). Therefore, H_t is bounded from below. This together with (28) gives that $\{H_t\}$ is convergent. □

B.3. Details and proofs of Corollary 4.3

Thanks to Theorem B.3, we have the following properties with respect to the successive changes.

Corollary B.4. *Consider (1) and let $\{(X^t, Y^t, Z^t)\}$ be generated by Algorithm 3. Suppose assumptions in Theorem B.3 hold. Then the following statements hold.*

(i) It holds that

$$\sum_{t=0}^{T} \mathbb{E} \|X^t - X^{t+1}\|^2 + \sum_{t=0}^{T} \mathbb{E} \|Y^{t+1} - Y^t\|^2 \leq \frac{L_\beta(X^0, Y^0, Z^0) + C - H_*}{\min\{\delta, \frac{\beta - L}{2}, \frac{\beta}{2}\}}. \tag{40}$$

and

$$\sum_{t=0}^{T} \mathbb{E} \|Z^t - Z^{t+1}\|^2 \leq \tau \beta C + \max\left\{ \Theta \frac{4}{(\beta - L)^2} \epsilon, \Gamma \right\} \frac{L_\beta(X^0, Y^0, Z^0) + C - H_*}{\min\{\delta, \frac{\beta - L}{2}, \frac{\beta}{2}\}}, \tag{41}$$

where $C := \frac{\Gamma}{\tau \beta} \left(3\beta^2 \tau^2 + 3\tau^2 L^2\right) \epsilon_0^2 + 3\tau^2 \|\nabla F(X^0)\|^2 + \frac{\Theta}{\tau \beta} (2 + 2 \frac{1}{(\beta - L)^2}) \epsilon_0^2$. with Θ and Γ being defined as in Theorem B.3.

(ii) It holds that

$$\lim_t \mathbb{E} \|X^t - X^{t+1}\|^2 = \lim_t \mathbb{E} \|Y^{t+1} - Y^t\|^2 = \lim_t \mathbb{E} \|Z^{t+1} - Z^t\|^2 = \lim \mathbb{E} \|Y^t - X^t\|^2 = 0. \tag{42}$$

Proof. Summing (28) from $t = 1$ to T, it holds that

$$\begin{aligned}
H_T &\leq H_1 - \delta \sum_{t=1}^{T} \mathbb{E} \|X^t - X^{t-1}\|^2 - \frac{\beta}{2} \sum_{t=1}^{T} \mathbb{E} \|Y^{t+1} - Y^t\|^2 \\
&\leq H_1 - \delta \sum_{t=1}^{T-1} \mathbb{E} \|X^t - X^{t+1}\|^2 - \frac{\beta}{2} \sum_{t=1}^{T-1} \mathbb{E} \|Y^{t+1} - Y^t\|^2
\end{aligned} \tag{43}$$

Now we bound H_1. Set $\epsilon_{-1} > \epsilon_0$. Let d be any vector in the unit ball of $\mathbb{R}^{np}$. Set $X^0_\star := X^0 - \epsilon_0 d$. Define $Z^{-1} = \frac{1}{1-\tau}(\beta\tau e^0 + \tau\nabla F(X^0_\star) - Z^0)$. Then $e^0 = \epsilon_0 d$ and $Z^0 = Z^{-1} + \beta\tau e^0 + \tau\nabla F(X^0_\star)$. The later relation means (26) also holds with $t = -1$. Using the same procedures in the proofs of Theorem B.3, we deduce that (35) holds when $t = 0$, i.e.,

$$\begin{aligned}
&\mathbb{E}L_\beta(X^1, Y^1, Z^1)\\
&\leq L_\beta(X^0, Y^0, Z^0) - \left(\frac{\beta - L}{2} - \frac{\sqrt{\epsilon}}{2} - \frac{1}{2(\beta - L)}\epsilon - \frac{\Lambda}{\tau\beta}\right)\mathbb{E}\|X^0 - X^1\|^2\\
&+ \frac{\Gamma}{\tau\beta}\left(\|Z^{-1} - Z^0\|^2 - \mathbb{E}\|Z^1 - Z^0\|^2\right) + \frac{\Theta}{\tau\beta}\mathbb{E}\|e^0 - e^1\|^2
\end{aligned}$$

Rearranging the above inequality and recalling the definition of H_T, we have that

$$H_1 \leq L_\beta(X^0, Y^0, Z^0) + \frac{\Gamma}{\tau\beta}\left(\|Z^{-1} - Z^0\|^2\right) + \frac{\Theta}{\tau\beta}\mathbb{E}_0\|e^0 - e^1\|^2. \tag{44}$$

Now we bound $\|e^0 - e^1\|^2$. Note that

$$\|e^0 - e^1\|^2 \leq 2\|e^0\|^2 + 2\|e^1\|^2 \leq 2\|e^0\|^2 + 2\frac{1}{(\beta - L)^2}\epsilon_0^2 \leq (2 + 2\frac{1}{(\beta - L)^2})\epsilon_0^2, \tag{45}$$

where the second inequality is because Proposition B.1 and (11) and the third equality uses the definitoin of e^0. Next we bound $\|Z^{-1} - Z^0\|^2$. Using the definintion of Z^{-1}, we have that

$$\begin{aligned}
&\|Z^{-1} - Z^0\|^2 \leq \|\beta\tau e^0 + \tau\nabla F(X^0_\star)\|^2\\
&\leq 3\beta^2\tau^2\|e^0\|^2 + 3\tau^2\|\nabla F(X^0_\star) - \nabla F(x^0)\|^2 + 3\tau^2\|\nabla F(X^0)\|^2\\
&\leq \left(3\beta^2\tau^2 + 3\tau^2 L^2\right)\epsilon_0^2 + 3\tau^2\|\nabla F(X^0)\|^2,
\end{aligned} \tag{46}$$

where the second inequality uses Cauchy-Schwarz inequality and the third inequality uses the Lipschitz continuity of ∇F and the definition of $X^0_\star$ and e^0. Using this inequality and (45), (44) can be further passed to

$$H_1 \leq L_\beta(X^0, Y^0, Z^0) + C. \tag{47}$$

where $C = \frac{\Gamma}{\tau\beta}\left(3\beta^2\tau^2 + 3\tau^2 L^2\right)\epsilon_0^2 + 3\tau^2\|\nabla F(X^0)\|^2 + \frac{\Theta}{\tau\beta}(2 + 2\frac{1}{(\beta - L)^2})\epsilon_0^2$.

Thus, summing (43) and (47), we have

$$\begin{aligned}
&H_T \leq L_\beta(X^0, Y^0, Z^0) + C\\
&- \delta\sum_{t=1}^{T}\mathbb{E}\|X^t - X^{t-1}\|^2 - \frac{\beta - L}{2}\mathbb{E}\sum_{t=1}^{T}\|X^{t+1} - X^{t+1}_\star\|^2 - \frac{\beta}{2}\sum_{t=1}^{T}\mathbb{E}\|Y^{t+1} - Y^t\|^2\\
&\leq H_1 - \delta\sum_{t=1}^{T-1}\mathbb{E}\|X^t - X^{t+1}\|^2 - \frac{\beta - L}{2}\mathbb{E}\sum_{t=0}^{T}\|X^{t+1} - X^{t+1}_\star\|^2 - \frac{\beta}{2}\sum_{t=1}^{T-1}\mathbb{E}\|Y^{t+1} - Y^t\|^2
\end{aligned}$$

Rearranging the above inequality, we have that

$$\begin{aligned}
&\delta\sum_{t=1}^{T-1}\mathbb{E}\|X^t - X^{t+1}\|^2 + \frac{\beta}{2}\sum_{t=1}^{T-1}\mathbb{E}\|Y^{t+1} - Y^t\|^2\\
&\leq L_\beta(X^0, Y^0, Z^0) + C - H_T \leq L_\beta(X^0, Y^0, Z^0) + C - H_*,
\end{aligned}$$

where the second inequality is because $\{H_t\}$ is nonincreasing and convergent. This implies (40).

Taking T in the above inequality to infinity, we deduce that

$$\delta\sum_{t=0}^{\infty}\mathbb{E}\|X^t - X^{t+1}\|^2 + \frac{\beta}{2}\sum_{t=0}^{\infty}\mathbb{E}\|Y^{t+1} - Y^t\|^2 < \infty.$$

where the last inequality is because $\{H_t\}$ is convergent. Therefore, we have $\{\mathbb{E}\|X^t - X^{t+1}\|^2\}$, and $\lim_t \mathbb{E}\|Y^{t+1} - Y^t\|^2$ are summable and

$$\lim_t \mathbb{E}\|X^t - X^{t+1}\|^2 = \lim_t \mathbb{E}\|Y^{t+1} - Y^t\|^2 = 0. \tag{48}$$

In addition, summing (34) from $t = 0$ to T, we have that

$$\begin{aligned}
&\sum_{t=0}^{T} \mathbb{E}\|Z^t - Z^{t+1}\|^2 \le \Gamma\|Z^{-1} - Z^0\|^2 + \Theta\sum_{t=1}^{T}\mathbb{E}\|e^t - e^{t+1}\|^2 + \mathbb{E}\|e^0 - e^1\|^2 + \Gamma\sum_{t=0}^{T}\mathbb{E}\|X^t - X^{t+1}\|^2\\
&\le \Gamma\|Z^{-1} - Z^0\|^2 + \Theta\mathbb{E}\|e^0 - e^1\|^2 + \Theta\frac{4}{(\beta - L)^2}\epsilon\sum_{t=1}^{T}\mathbb{E}\|X^t - X^{t-1}\|^2 + \Gamma\sum_{t=0}^{T}\mathbb{E}\|X^t - X^{t+1}\|^2\\
&\le \Gamma\|Z^{-1} - Z^0\|^2 + \Theta\mathbb{E}\|e^0 - e^1\|^2 + \max\left\{\Theta\frac{4}{(\beta - L)^2}\epsilon, \Gamma\right\}\sum_{t=0}^{T}\mathbb{E}\|X^t - X^{t+1}\|^2\\
&\le \tau\beta C + \max\left\{\Theta\frac{4}{(\beta - L)^2}\epsilon, \Gamma\right\}\frac{L_\beta(X^0, Y^0, Z^0) + C - H_*}{\min\{\delta, \frac{\beta - L}{2}, \frac{\beta}{2}\}},
\end{aligned}$$

where the second inequality uses (27), the last inequality uses (45), (46) and (40). Taking T in the above inequality to infinity we deduce that $\{\mathbb{E}\|Z^t - Z^{t+1}\|^2\}$ is summable and using (12), we have that

$$\lim \mathbb{E}\|Y^t - X^{t+1}\|^2 = \lim_t \mathbb{E}\|Z^t - Z^{t+1}\|^2 = 0.$$

This together with (48) gives that

$$\lim \mathbb{E}\|Y^t - X^t\|^2 = 0.$$

□

B.4. Details and proofs of Theorem 4.5

Here, we prove the complexity of Algorithm 3 in Theorem 4.5.

Theorem B.5. *Consider* (1) *and let* $\{(X^t, Y^t, Z^t)\}$ *be generated by Algorithm* 3*. Suppose assumptions in Theorem* B.3 *hold. Then the following statements hold.*

(i) There exists $E > 0$ *such that*

$$\|\nabla F(Y^{t+1}) + \xi^{t+1}\| \le E\left(\|X^{t+1} - X^t\| + \|Z^{t+1} - Z^t\| + \|Y^t - Y^{t+1}\|\right). \tag{49}$$

where $\xi \in \partial F(Y^{t+1})$.

(ii) It holds that

$$\begin{aligned}
&\frac{1}{1+T}\sum_{t=0}^{T}\mathbb{E}d^2(0, \nabla F(Y^{t+1} + \partial G(Y^{t+1}))\\
&\le \frac{1}{T+1}\Upsilon\left(\tau\beta C + \left(\max\left\{\Theta\frac{4}{(\beta - L)^2}\epsilon, \Gamma, 1\right\}\frac{L_\beta(X^0, Y^0, Z^0) + C - H_*}{\min\{\delta, \frac{\beta - L}{2}, \frac{\beta}{2}\}}\right)\right),
\end{aligned}$$

where Γ *and* Θ *are defined in Theorem* B.3*,* H_* *and* C *is defined in Theorem* B.3 *and Corollary* B.4 *respectively,* $\Upsilon = \max\{3(L+\beta)^2\frac{\epsilon}{(\beta - L)^2}, \left(\frac{L}{\tau\beta} + 1\right)^2, (L+\beta)^2\}$

Proof. Using (24), it hold that

$$0 = \nabla F(Y^{t+1}) + \nabla F(X_\star^{t+1}) - \nabla F(Y^{t+1}) + Z^t + \beta(X_\star^{t+1} - Y^t).$$

Summing this with (25), we have that

$$0 = \nabla F(Y^{t+1}) + \xi^{t+1} + \nabla F(X_\star^{t+1}) - \nabla F(Y^{t+1}) + Z^t - Z^{t+1} + \beta(X_\star^{t+1} - X^{t+1}) - \beta(Y^{t+1} - Y^t).$$

This implies that

$$\begin{aligned}
&\|\nabla F(Y^{t+1}) + \xi^{t+1}\| \\
&\le \|\nabla F(X_\star^{t+1}) - \nabla F(Y^{t+1})\| + \|Z^t - Z^{t+1}\| + \beta\|X_\star^{t+1} - X^{t+1}\| + \beta\|Y^{t+1} - Y^t\| \\
&\le L\|X_\star^{t+1} - Y^{t+1}\| + \|Z^t - Z^{t+1}\| + \beta\|X_\star^{t+1} - X^{t+1}\| + \beta\|Y^{t+1} - Y^t\| \\
&\le L\|X_\star^{t+1} - X^{t+1}\| + L\|X^{t+1} - Y^t\| + (L+\beta)\|Y^t - Y^{t+1}\| + \|Z^t - Z^{t+1}\| + \beta\|X_\star^{t+1} - X^{t+1}\| \\
&= (L+\beta)\|X_\star^{t+1} - X^{t+1}\| + \left(\frac{L}{\tau\beta} + 1\right)\|Z^{t+1} - Z^t\| + (L+\beta)\|Y^t - Y^{t+1}\|,
\end{aligned} \tag{50}$$

where the last equality uses (12). Using Proposition B.1 and (11), we have that $\|X_\star^{t+1} - X^{t+1}\| \le \frac{\sqrt{\epsilon}}{\beta - L}\|X^{t+1} - X^t\|$. Using this, (50) can be further passed to

$$\|\nabla F(Y^{t+1}) + \xi^{t+1}\| \le (L+\beta)\frac{\sqrt{\epsilon}}{\beta - L}\|X^{t+1} - X^t\| + \left(\frac{L}{\tau\beta} + 1\right)\|Z^{t+1} - Z^t\| + (L+\beta)\|Y^t - Y^{t+1}\|.$$

This together with Cauchy-Schwarz inequality, we have that

$$\|\nabla F(Y^{t+1}) + \xi^{t+1}\|^2 \le 3(L+\beta)^2\frac{\epsilon}{(\beta - L)^2}\|X^{t+1} - X^t\|^2 + \left(\frac{L}{\tau\beta} + 1\right)^2\|Z^{t+1} - Z^t\|^2 + (L+\beta)^2\|Y^t - Y^{t+1}\|^2. \tag{51}$$

This proves (49).

Next, taking expectations on both sides of (51), we have that

$$\begin{aligned}
&\mathbb{E}\|\nabla F(Y^{t+1}) + \xi^{t+1}\|^2 \\
&\le 3(L+\beta)^2\frac{\epsilon}{(\beta - L)^2}\mathbb{E}\|X^{t+1} - X^t\|^2 + \left(\frac{L}{\tau\beta} + 1\right)^2\mathbb{E}\|Z^{t+1} - Z^t\|^2 + (L+\beta)^2\mathbb{E}\|Y^t - Y^{t+1}\|^2.
\end{aligned}$$

Summing the above inequality from $t = 0$ to T, it holds that

$$\begin{aligned}
&\sum_{t=0}^{T}\mathbb{E}\|\nabla F(Y^{t+1}) + \xi^{t+1}\|^2 \\
&\le 3(L+\beta)^2\frac{\epsilon}{(\beta - L)^2}\sum_{t=0}^{T}\mathbb{E}\|X^{t+1} - X^t\|^2 + \left(\frac{L}{\tau\beta} + 1\right)^2\sum_{t=0}^{T}\mathbb{E}\|Z^{t+1} - Z^t\|^2 + (L+\beta)^2\sum_{t=0}^{T}\mathbb{E}\|Y^t - Y^{t+1}\|^2 \\
&\le \max\{3(L+\beta)^2\frac{\epsilon}{(\beta - L)^2}, \left(\frac{L}{\tau\beta} + 1\right)^2, (L+\beta)^2\}\left(\sum_{t=0}^{T}\mathbb{E}\|X^{t+1} - X^t\|^2 + \|Y^t - Y^{t+1}\|^2 + \|Z^{t+1} - Z^t\|^2\right) \\
&\le \max\{3(L+\beta)^2\frac{\epsilon}{(\beta - L)^2}, \left(\frac{L}{\tau\beta} + 1\right)^2, (L+\beta)^2\}\tau\beta C \\
&+ \max\{3(L+\beta)^2\frac{\epsilon}{(\beta - L)^2}, \left(\frac{L}{\tau\beta} + 1\right)^2, (L+\beta)^2\}\left(\max\left\{\Theta\frac{4}{(\beta - L)^2}\epsilon, \Gamma, 1\right\}\frac{L_\beta(X^0, Y^0, Z^0) + C - H_*}{\min\{\delta, \frac{\beta - L}{2}, \frac{\beta}{2}\}}\right),
\end{aligned} \tag{52}$$

where the last inequality uses (40) and (41). Dividing both sides with $T + 1$ and recalling $\xi^{t+1} \in \partial G(Y^{t+1})$, we have the conclusion. □

B.5. Details and proofs in Section 4.1

B.5.1. Proofs of Proposition 4.7

We provide the detailed version of Proposition 4.7 as follows.

Proposition B.6. *Consider* (1) *and let* $\{(X^t, Y^t, Z^t)\}$ *be generated by Algorithm 3 with* (11) *is deterministic. Suppose assumptions in Theorem B.3 hold. Suppose* $\{(X^t, Y^t, Z^t)\}$ *is bounded and denote the set of accumulation points of* $\{(X^t, Y^t, Z^t, X^{t-1}, Z^{t-1})\}$ *as* Ω*. The following statements hold:*

(i) $\lim_t d((X^t, Y^t, Z^t, X^{t-1}, Z^{t-1})), \Omega) = 0$.

(ii) Any accumulation point of $\{Y^t\}$ *is a stationary point of* (1).

(iii) $H \equiv H_*$ *on* Ω.

Proof. For (i), let Y^* be an accumulation point of $\{Y^t\}$ with $Y^{t_i} \to Y^*$. Using (24) and (25), there exists $\xi^{t_i} \in G(Y^{t_i})$ such that

$$0 = \nabla F(X_\star^{t_i}) + Z^{t_i-1} + \beta(X_\star^{t_i} - Y^{t_i-1}) = \nabla F(Y^t) + \nabla F(X_\star^{t_i}) - \nabla F(Y^t) + Z^{t_i-1} + \beta(X_\star^{t_i} - Y^{t_i-1}).$$

and

$$0 = \xi^{t_i} - Z^{t_i} - \beta(X^{t_i} - Y^{t_i}).$$

The above relations shows that

$$\begin{aligned} 0 &= \nabla F(Y^t) + \xi^{t_i} + \nabla F(X_\star^{t_i}) - \nabla F(Y^t) + Z^{t_i-1} - Z^{t_i} + \beta(X_\star^{t_i} - Y^{t_i-1}) - \beta(X^{t_i} - Y^{t_i}) \\ &= \nabla F(Y^t) + \xi^{t_i} + \nabla F(X_\star^{t_i}) - \nabla F(Y^t) + \tau\beta(X^{t_i} - Y^{t_i-1}) + \beta(X_\star^{t_i} - Y^{t_i-1}) - \beta(X^{t_i} - Y^{t_i}) \end{aligned} \tag{53}$$

where the equality makes uses of (12). Now we show that $\lim_i \|X_\star^t - X^t\| = 0$. Using Proposition B.1 and (11), we have that

$$\|e^t\|^2 = \|X_\star^t - X^t\|^2 \le \frac{\epsilon}{(\beta - L)^2}\|X^t - X^{t-1}\|^2.$$

Since $\lim_t \|X^t - X^{t-1}\| = 0$, we have that

$$\lim_i \|X_\star^t - X^t\| = 0. \tag{54}$$

Next, we show that $\lim_i \|X^t - Y^{t-1}\| = 0$. Using (12), it holds that

$$\begin{aligned} &\left\|Z^t - Z^{t-1}\right\|^2 \\ &\le \Gamma\left(\|Z^{t-2} - Z^{t-1}\|^2 - \|Z^t - Z^{t-1}\|^2\right) + \Theta\|e^{t-1} - e^t\|^2 + \Lambda\left\|X^{t-1} - X^t\right\|^2 \\ &\le \Gamma\left(\|Z^{t-2} - Z^{t-1}\|^2 - \|Z^t - Z^{t-1}\|^2\right) + \Theta\frac{4}{(\beta - L)^2}\epsilon\|X^{t-1} - X^{t-2}\|^2 + \Lambda\left\|X^{t-1} - X^t\right\|^2 \end{aligned}$$

where the first inequality uses (34) and the second inequality is due to (27). Summing the above inequality from $t = 1$ to T, we have that

$$\begin{aligned} &\sum_{1=1}^{T}\left\|Z^t - Z^{t-1}\right\|^2 \le \Gamma\left(\|Z^{t_1-2} - Z^{t_1-1}\|^2 - \|Z^{t_K} - Z^{t_K-1}\|^2\right) \\ &+ \frac{1}{\tau\beta}\Theta\frac{4}{(\beta - L)^2}\epsilon\sum_{1=1}^{T}\|X^{t-1} - X^{t-2}\|^2 + \Lambda\sum_{1=1}^{T}\left\|X^{t-1} - X^t\right\|^2 \\ &\le \Gamma\left(\|Z^{t_1-2} - Z^{t_1-1}\|^2 - \|Z^{t_K} - Z^{t_K-1}\|^2\right) + \Theta\frac{4}{(\beta - L)^2}\epsilon\sum_{i=1}^{K}\|X^{t-1} - X^{t-2}\|^2 + \Lambda\sum_{i=1}^{K}\left\|X^{t-1} - X^t\right\|^2 \\ &\le \Gamma\|Z^{t_1-2} - Z^{t_1-1}\|^2 + \Theta\frac{4}{(\beta - L)^2}\epsilon\sum_{1=1}^{T}\|X^{t-1} - X^{t-2}\|^2 + \Lambda\sum_{1=1}^{T}\left\|X^{t-1} - X^t\right\|^2. \end{aligned}$$

Taking K in the above inequality to infinity and recalling that $\left\|X^{t-1} - X^t\right\|^2$ is summable, we deduce that $\sum_{1=1}^{T}\|Z^t - Z^{t-1}\|^2 < \infty$. This together with (12) show that

$$\lim_t \|X^t - Y^{t-1}\| = \frac{1}{\tau\beta}\lim_t \|Z^t - Z^{t-1}\| = 0. \tag{55}$$

Next, we show that $\lim_t \|Y^t - Y^{t-1}\| = 0$. Using (12) again, we have that

$$Y^t - Y^{t-1} = X^{t+1} - X^t - \frac{1}{\tau\beta}(Z^{t+1} - Z^t) - \frac{1}{\tau\beta}(Z^t - Z^{t-1}).$$

This together with the fact that $\lim_t \|X^t - X^{t-1}\| = \lim_t \|Z^t - Z^{t-1}\| = 0$ implies that $\lim_t \|Y^t - Y^{t-1}\| = 0$. Since $Y^{t_i} \to Y^*$, combining (54), (55) and (42), we have that

$$\lim_i Y^{t_i-1} = \lim_i X^{t_i} = \lim_i X_\star^{t_i} = \lim_i Y^{t_i} = Y^*.$$

This together with the continuity of ∇F, the closedness of ∂G and (53) shows that

$$0 \in \nabla F(Y^*) + \partial G(Y^*).$$

This completes the proof.

Now we prove (ii). Fix any $(X^*, Y^*, Z^*, \bar{X}^*, \bar{Z}^*) \in \Omega$. Then there exists $\{t_i\}_i$ such that $(X^{t_i}, Y^{t_i}, Z^{t_i}, X^{t_i-1}, Y^{t_i-1})$ converges to $(X^*, Y^*, Z^*, \bar{X}^*, \bar{Z}^*)$. Thanks to Theorem B.3 (ii), we know that

$$H_* = \lim_i H(X^{t_i}, Y^{t_i}, Z^{t_i}, X^{t_i-1}, Y^{t_i-1}) \tag{56}$$

and

$$H(X^*, Y^*, Z^*, \bar{X}^*, \bar{Z}^*) = L_\beta(X^*, Y^*, Z^*) = F(X^*) + G(Y^*) + \langle X^* - Y^*, Z^* \rangle + \frac{\beta}{2}\|X^* - Y^*\|^2. \tag{57}$$

Since Y^t is the minimizer of (13), it holds that

$$G(Y^{t_i}) + \left\langle X^{t_i} - Y^{t_i}, Z^{t_i} \right\rangle + \frac{\beta}{2}\|X^{t_i} - Y^{t_i}\|^2 \le G(Y^*) + \left\langle X^{t_i} - Y^*, Z^{t_i} \right\rangle + \frac{\beta}{2}\|X^{t_i} - Y^*\|^2.$$

Taking the above inequality to infinity, we have that

$$\begin{aligned}
&\limsup_i G(Y^{t_i}) + \langle X^* - Y^*, Z^* \rangle + \frac{\beta}{2}\|X^* - Y^*\|^2 \\
&= \limsup_i G(Y^{t_i}) + \left\langle X^{t_i} - Y^{t_i}, Z^{t_i} \right\rangle + \frac{\beta}{2}\|X^{t_i} - Y^{t_i}\|^2 \le G(Y^*) + \langle X^* - Y^*, Z^* \rangle + \frac{\beta}{2}\|X^* - Y^*\|^2.
\end{aligned}$$

This together with the closedness of G shows that $\lim_i G(Y^{t_i}) = G(Y^*)$. This together with the continuity of F, Corollary B.4 (ii) and (56) gives that

$$\begin{aligned}
H_* &= \lim_i H(X^{t_i}, Y^{t_i}, Z^{t_i}, X^{t_i-1}, Y^{t_i-1}) \\
&= F(X^*) + G(Y^*) + \langle X^* - Y^*, Z^* \rangle + \frac{\beta}{2}\|X^* - Y^*\|^2 = H(X^*, Y^*, Z^*, \bar{X}^*, \bar{Z}^*),
\end{aligned}$$

where the second equality uses (57). □

B.5.2. Details and proofs for Theorem 4.9

To show the global convergence of the generated sequence, we first need to bound the subdifferential of $\partial H(X^{t+1}, Y^{t+1}, Z^{t+1}, X^t, Z^t)$.

Lemma B.7. *Consider* (1) *and let* $\{(X^t, Y^t, Z^t)\}$ *be generated by Algorithm* 3 *with* (11) *is deterministic. Suppose assumptions in Theorem* B.3 *hold. There exists* $D > 0$ *such that*

$$d(0, \partial H(X^{t+1}, Y^{t+1}, Z^{t+1}, X^t, Z^t)) \le D\left(\|X^{t+1} - X^t\| + \|Y^{t+1} - Y^t\| + \|Z^{t+1} - Z^t\|\right).$$

Proof. Using Exercise 8.8, Proposition 10.5 and Corollary 10.9 of RockWets98, it holds that

$$\partial H(X, Y, Z, X', Z') \supseteq \begin{pmatrix} \nabla F(X) \\ \partial G(Y) \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} Z + \beta(X - Y) + \frac{\Theta}{\tau\beta}\frac{8}{(\beta-L)^2}\epsilon(X - X') \\ -Z - \beta(X - Y) \\ X - Y + \frac{2\Gamma}{\tau\beta}(Z - Z') \\ -\frac{\Theta}{\tau\beta}\frac{8}{(\beta-L)^2}\epsilon(X - X') \\ -\frac{2\Gamma}{\tau\beta}(Z - Z'). \end{pmatrix}$$

Thus,

$$
\begin{aligned}
\partial H(X^{t+1}, Y^{t+1}, Z^{t+1}, X^t, Z^t) &\supseteq \begin{pmatrix} \nabla F(X^{t+1}) + Z^{t+1} + \beta(X^{t+1} - Y^{t+1}) + \frac{\Theta}{\tau\beta}\frac{8}{(\beta-L)^2}\epsilon(X^{t+1} - X^t) \\ \partial G(Y^{t+1}) - Z^{t+1} - \beta(X^{t+1} - Y^{t+1}) \\ X^{t+1} - Y^{t+1} + \frac{2\Gamma}{\tau\beta}(Z^{t+1} - Z^t) \\ -\frac{\Theta}{\tau\beta}\frac{8}{(\beta-L)^2}\epsilon(X^{t+1} - X^t) \\ -\frac{2\Gamma}{\tau\beta}(Z^{t+1} - Z^t) \end{pmatrix} \\
&\supseteq \begin{pmatrix} \nabla F(X^{t+1}) + Z^{t+1} + \beta(X^{t+1} - Y^{t+1}) + \frac{\Theta}{\tau\beta}\frac{8}{(\beta-L)^2}\epsilon(X^{t+1} - X^t) \\ 0 \\ X^{t+1} - Y^{t+1} + \frac{2\Gamma}{\tau\beta}(Z^{t+1} - Z^t) \\ -\frac{\Theta}{\tau\beta}\frac{8}{(\beta-L)^2}\epsilon(X^{t+1} - X^t) \\ -\frac{2\Gamma}{\tau\beta}(Z^{t+1} - Z^t) \end{pmatrix}
\end{aligned}
\tag{58}
$$

where the seconde inclusion follows from (25).

Now, we bound each coordinate in the right hand side of the relation. For the first one, we denote $\mathcal{A}^{t+1} := \nabla F(X^{t+1}) + Z^{t+1} + \beta(X^{t+1} - Y^{t+1}) + \frac{\Theta}{\tau\beta}\frac{8}{(\beta-L)^2}\epsilon(X^{t+1} - X^t)$. Using (24), we have that

$$
\mathcal{A}^{t+1} \ni \nabla F(X^{t+1}) - \nabla F(X_\star^{t+1}) + (Z^{t+1} - Z^t) + \beta(X^{t+1} - Y^{t+1} - X_\star^{t+1} + Y^t) + \frac{\Theta}{\tau\beta}\frac{8}{(\beta-L)^2}\epsilon(X^{t+1} - X^t).
$$

Thus, we deduce that $d^2(0, \mathcal{A}^{t+1})$ is bounded above by

$$
\begin{aligned}
&\left(\|\nabla F(X^{t+1}) - \nabla F(X_\star^{t+1})\| + \|Z^{t+1} - Z^t\| + \beta\|X^{t+1} - X_\star^{t+1}\| + \beta\|Y^t - Y^{t+1}\| + \frac{\Theta}{\tau\beta}\frac{8}{(\beta-L)^2}\epsilon\|X^{t+1} - X^t\| \right)^2 \\
&\leq 4(L+\beta)^2\|X^{t+1} - X_\star^{t+1}\|^2 + 4\|Z^{t+1} - Z^t\|^2 + 4\beta^2\|Y^t - Y^{t+1}\|^2 + \frac{4\Theta^2}{\tau^2\beta^2}\frac{64}{(\beta-L)^4}\epsilon^2\|X^{t+1} - X^t\|^2
\end{aligned}
\tag{59}
$$

where the second inequality uses the Lipscitz continuity of ∇F.

For the third coordinate in (58), using (12), it holds that

$$
\begin{aligned}
&\left\| X^{t+1} - Y^{t+1} + \frac{2\Gamma}{\tau\beta}(Z^{t+1} - Z^t) \right\|^2 = \left\| \frac{1}{\tau\beta}(Z^{t+1} - Z^t) + Y^t - Y^{t+1} + \frac{2\Gamma}{\tau\beta}(Z^{t+1} - Z^t) \right\|^2 \\
&\leq 2\|Y^t - Y^{t+1}\|^2 + \frac{(1+2\Gamma)^2}{\tau^2\beta^2}\|Z^{t+1} - Z^t\|^2
\end{aligned}
$$

This together with (58) and (59), we deduce that

$$
\begin{aligned}
&d^2(0, \partial H(X^{t+1}, Y^{t+1}, Z^{t+1}, X^t, Z^t)) \\
&\leq 4(L+\beta)^2\|X^{t+1} - X_\star^{t+1}\|^2 + 4\|Z^{t+1} - Z^t\|^2 + 4\beta^2\|Y^t - Y^{t+1}\|^2 + \frac{4\Theta^2}{\tau^2\beta^2}\frac{64}{(\beta-L)^4}\epsilon^2\|X^{t+1} - X^t\|^2 \\
&+ 2\|Y^t - Y^{t+1}\|^2 + \frac{(1+2\Gamma)^2}{\tau^2\beta^2}\|Z^{t+1} - Z^t)\|^2 + \frac{\Theta^2}{\tau^2\beta^2}\frac{64}{(\beta-L)^4}\epsilon^2\|X^{t+1} - X^t\|^2 + \frac{4\Gamma^2}{\tau^2\beta^2}\|Z^{t+1} - Z^t\|^2.
\end{aligned}
\tag{60}
$$

Note that using strong convexity of the objective in (10) and Proposition B.1, we have that

$$
\|X^{t+1} - X_\star^{t+1}\|^2 \leq \frac{1}{(\beta-L)^2}\epsilon\|X^{t+1} - X^t\|^2. \tag{61}
$$

Combining (60) with (61), we have that

$$
\begin{aligned}
&d^2(0, \partial H(X^{t+1}, Y^{t+1}, Z^{t+1}, X^t, Z^t)) \\
&\leq \frac{4(L+\beta)^2\epsilon}{(\beta-L)^2}\|X^{t+1} - X^t\|^2 + 4\|Z^{t+1} - Z^t\|^2 + 4\beta^2\|Y^t - Y^{t+1}\|^2 + \frac{4\Theta^2}{\tau^2\beta^2}\frac{64}{(\beta-L)^4}\epsilon^2\|X^{t+1} - X^t\|^2 \\
&+ 2\|Y^t - Y^{t+1}\|^2 + \frac{(1+2\Gamma)^2}{\tau^2\beta^2}\|Z^{t+1} - Z^t)\|^2 + \frac{\Theta^2}{\tau^2\beta^2}\frac{64}{(\beta-L)^4}\epsilon^2\|X^{t+1} - X^t\|^2 + \frac{4\Gamma^2}{\tau^2\beta^2}\|Z^{t+1} - Z^t\|^2 \\
&= D'(\|X^{t+1} - X^t\|^2 + \|Y^t - Y^{t+1}\|^2 + \|Z^{t+1} - Z^t\|^2),
\end{aligned}
$$

where D is the maximum of the coordinates of $\|X^{t+1}-X^t\|^2$, $\|Y^t-Y^{t+1}\|$ and $\|Z^{t+1}-Z^t\|^2$ on the right hand side of the above inequality. Finally, using the fact that $\sum_i^3 s_i^2 \le (\sum_i^3 a_i)^2$ for any $a_1, a_2, a_3 \ge 0$, the above inequality can be further passed to

$$d^2(0, \partial H(X^{t+1}, Y^{t+1}, Z^{t+1}, X^t, Z^t)) \le D'(\|X^{t+1}-X^t\| + \|Y^t - Y^{t+1}\| + \|Z^{t+1}-Z^t\|).$$

Taking square root on both sides of the above inequality we have the conclusion. □

Now we are ready to prove Theorem 4.9. In fact, we already show the key properties that will be needed. They are Theorem B.3, Corollary B.4, Proposition B.5.1 and Lemma B.7. The rest steps are routine. We follow the proofs in (Borwein et al., 2017; Bolte et al., 2014; Li & Pong, 2016) and include it only for completeness.

Theorem B.8. *Consider* (1) *and let* $\{(X^t, Y^t, Z^t)\}$ *be generated by Algorithm 3 with* (11) *is deterministic. Suppose assumptions in Theorem B.3 hold. Let* H *be defines as in Theorem 4.2 and suppose* H *is a KL function with exponent* $\theta \in [0, 1)$. *Denoting* $(X^*, Y^*, Z^*) := \lim_t (X^t, Y^t, Z^t)$, *then the followings hold.*

(i) If $\alpha = 0$, *then* $\{(x^t, w^t, z^t)\}$ *converges finitely.*

(ii) If $\alpha \in (0, \frac{1}{2}]$, *then there exist* $b > 0$ *and* $\rho_1 \in (0, 1)$ *such that* $\max\{\|w^t - w^*\|, \|x^t - x^*\|, \|z^t - z^*\|, \|y^t - y^*\|\} \le b\rho_1^t$ *for large* t. *Furthermore,*

(iii) If $\alpha \in (\frac{1}{2}, 1)$, *then there exists* $c > 0$ *such that* $\max\{\|w^t - w^*\|, \|x^t - x^*\|, \|z^t - z^*\|, \|y^t - y^*\|\} \le ct^{-\frac{1}{4\alpha-2}}$ *for large* t.

Proof. We first show that $\{(X^t, Y^t, Z^t)\}$ is convergent. If there exists t_0 such that $H_{t_0} = H_*$. Since $\{H_t\}$ is nonincreasing thanks to (28), we deduce that $H_t = H_*$ for all $t \ge t_0$. Using (28) again we have that for all $t \ge t_0$, it holds that $X^t = X^{t-1} = \cdots = X^{t_0-1}$ and $Y^t = Y^{t-1} = \cdots = Y^{t_0}$. Recalling in (42) we have that $\lim_t (X^t - Y^t) = 0$, we have that $Y^{t_0} = X^{t_0-1}$. Thus, $X^{t+1} - Y^t = Y^{t_0} - X^{t_0-1} = 0$ for all $t \ge t_0$. This together with (12), we deduce that $Z^{t+1} = Z^t = \cdots = Z^{t_0}$ for all $t \ge t_0$. Therefore, when there exists t_0 such that $H_{t_0} = H_*$, $\{(X^t, Y^t, Z^t)\}$ converge finitely.

Next, we consider the case where $H_t > H_*$ for all t. Thanks to Proposition B.5.1 (iii), using Lemma 6 of (Bolte et al., 2014), there exists $\epsilon > 0$, $a > 0$ and $\psi \in \Psi_a$ such that

$$\psi'(H(X, Y, Z, X'Z') - H_*) d(0, \partial H(X, Y, Z, X', Z')) \ge 1$$

when $d((X, Y, Z, X', Z'), \Omega) \le \epsilon$ and $H_* < H(X, Y, Z, X', Z') < H_* + a$. Thanks to Corollary B.4 and Theorem B.3, we know that there exists t_1 such that when $t > t_1$, $d((X^t, Y^t, Z^t, X^{t-1}, Z^{t-1}), \Omega) \le \epsilon$ and $H_* < H(X^t, Y^t, Z^t, X^{t-1}, Z^{t-1}) < H_* + a$. Thus, it holds that

$$\psi'(H((X^t, Y^t, Z^t, X^{t-1}, Z^{t-1}) - H_*) d(0, \partial H((X^t, Y^t, Z^t, X^{t-1}, Z^{t-1})) \ge 1. \tag{62}$$

Recaling (28), we have that Since ψ is concave, using the above inequality we have that

$$\begin{aligned}
&\delta\|X^{t+1}-X^t\|^2 + \frac{\beta}{2}\|Y^{t+1}-Y^t\|^2 \le H_t - H_{t+1} \\
&\le \psi'(H_t - H_*) d(0, \partial H(X^t, Y^t, Z^t, X^{t-1}, Z^{t-1}))\,(H_t - H_{t+1}) \\
&\le (\psi(H_t - H_*) - \psi(H_{t+1} - H_*))\, d(0, \partial H(X^t, Y^t, Z^t, X^{t-1}, Z^{t-1}))
\end{aligned} \tag{63}$$

where the second inequality uses (62) and the last inequality uses the concavity of ψ. Using Lemma B.7, we have from (63) that

$$\begin{aligned}
&\frac{1}{2}\min\{\delta, \frac{\beta}{2}\}\left(\|X^{t+1}-X^t\| + \|Y^{t+1}-Y^t\|\right)^2 \le \min\{\delta, \frac{\beta}{2}\}\left(\|X^{t+1}-X^t\|^2 + \|Y^{t+1}-Y^t\|^2\right) \\
&\le \delta\|X^{t+1}-X^t\|^2 + \frac{\beta}{2}\|Y^{t+1}-Y^t\|^2 \\
&\le (\psi(H_t - H_*) - \psi(H_{t+1} - H_*))\, D\left(\|X^t - X^{t-1}\| + \|Y^t - Y^{t-1}\| + \|Z^t - Z^{t-1}\|\right)
\end{aligned} \tag{64}$$

where the first inequality uses the fact that $\frac{1}{2}(a+b)^2 \le a^2 + b^2$ for any $a, b \in \mathbb{R}$.

Now we bound $\|Z^t - Z^{t-1}\|$. Using (26), we have that

$$\begin{aligned}
\|Z^{t+1} - Z^t\| &= |1-\tau|\|Z^t - Z^{t-1}\| + \beta\tau\|e^{t+1} - e^t\| + \tau\|\nabla F(X_\star^{t+1}) - \nabla F(X_\star^t)\| \\
&\le |1-\tau|\|Z^t - Z^{t-1}\| + \beta\tau\|e^{t+1} - e^t\| + \tau L\|X_\star^{t+1} - X_\star^t\| \\
&\le |1-\tau|\|Z^t - Z^{t-1}\| + (\beta+L)\tau\|e^{t+1} - e^t\| + \tau L\|X^{t+1} - X^t\| \\
&\le |1-\tau|\|Z^t - Z^{t-1}\| + (\beta+L)\tau\frac{4}{(\beta-L)^2}\|X^t - X^{t-1}\| + \tau L\|X^{t+1} - X^t\|
\end{aligned}$$

where the second inequality uses the definition of e^t and last inequality uses (27). Rearranging the above inequality, it holds that

$$\begin{aligned}
\|Z^t - Z^{t-1}\| \le{}& \frac{1+|1-\tau|}{1-|1-\tau|}\left(\|Z^t - Z^{t-1}\| - \|Z^t - Z^{t+1}\|\right) - \|Z^t - Z^{t+1}\| \\
&+ \frac{2}{1-|1-\tau|}(\beta+L)\tau\frac{4}{(\beta-L)^2}\|X^t - X^{t-1}\| + \frac{2}{1-|1-\tau|}\tau L\|X^{t+1} - X^t\|.
\end{aligned}$$

Plugging this bound into (64), we have that

$$\begin{aligned}
&\frac{1}{2}\min\{\delta, \frac{\beta}{2}\}\left(\|X^{t+1} - X^t\| + \|Y^{t+1} - Y^t\|\right)^2 \\
&\le \left(\psi(H_t - H_*) - \psi(H_{t+1} - H_*)\right) D\left(\|X^t - X^{t-1}\| + \|Y^t - Y^{t-1}\|\right) \\
&+ \left(\psi(H_t - H_*) - \psi(H_{t+1} - H_*)\right) D\left(\frac{1+|1-\tau|}{1-|1-\tau|}\left(\|Z^t - Z^{t-1}\| - \|Z^t - Z^{t+1}\|\right) - \|Z^t - Z^{t+1}\|\right) \\
&+ \left(\psi(H_t - H_*) - \psi(H_{t+1} - H_*)\right) D\left(\frac{2(\beta+L)\tau}{1-|1-\tau|}\frac{4}{(\beta-L)^2}\|X^t - X^{t-1}\| + \frac{2\tau L}{1-|1-\tau|}\|X^{t+1} - X^t\|\right) \\
&\le \left(\psi(H_t - H_*) - \psi(H_{t+1} - H_*)\right) DD_1\left(\Delta_t^1 + \Delta_t^2\right),
\end{aligned}$$

where $D_1 := \max\{1 + \frac{2(\beta+L)\tau}{1-|1-\tau|}\frac{4}{(\beta-L)^2}, \frac{2\tau L}{1-|1-\tau|}, 1, \frac{1+|1-\tau|}{1-|1-\tau|}\}$, $\Delta_t := \|X^t - X^{t-1}\| + \|X^{t+1} - X^t\| + \|Y^t - Y^{t-1}\|$ and $\Delta_t^2 := \left(\|Z^t - Z^{t-1}\| - \|Z^t - Z^{t+1}\|\right) - \|Z^t - Z^{t+1}\|$. Rearranging the above inequality and taking square toot on both sides, we obtain that

$$\begin{aligned}
\|X^{t+1} - X^t\| + \|Y^{t+1} - Y^t\| &\le \sqrt{\frac{2}{\min\{\delta, \frac{\beta}{2}\}}\left(\psi(H_t - H_*) - \psi(H_{t+1} - H_*)\right) DD_1\left(\Delta_t^1 + \Delta_t^2\right)} \\
&\le \frac{2}{\min\{\delta, \frac{\beta}{2}\}}\left(\psi(H_t - H_*) - \psi(H_{t+1} - H_*)\right) DD_1 + \frac{1}{4}\left(\Delta_t^1 + \Delta_t^2\right)
\end{aligned}$$

where the second inequality uses the fact that $\sqrt{ab} \le \frac{1}{2}(a+b)$ for any $a, b > 0$. Recalling the definitions of Δ_t^1 and Δ_t^2, and rearranging the above inequality, we have that

$$\begin{aligned}
\|X^{t+1} - X^t\| + \|Y^{t+1} - Y^t\| &\le \sqrt{\frac{2}{\min\{\delta, \frac{\beta}{2}\}}\left(\psi(H_t - H_*) - \psi(H_{t+1} - H_*)\right) DD_1\Delta} \\
&\le \frac{2}{\min\{\delta, \frac{\beta}{2}\}}\left(\psi(H_t - H_*) - \psi(H_{t+1} - H_*)\right) DD_1 \\
&+ \frac{1}{4}\left(\|X^t - X^{t-1}\| + \|X^{t+1} - X^t\| + \|Y^t - Y^{t-1}\| + \|Z^t - Z^{t-1}\| - \|Z^t - Z^{t+1}\| - \|Z^t - Z^{t+1}\|\right)
\end{aligned}$$

Further rearranging the above inequality, we have

$$\begin{aligned}
\frac{1}{4}\|X^{t+1} - X^t\| + \frac{3}{4}\|Y^{t+1} - Y^t\| + \frac{1}{4}\|Z^t - Z^{t+1}\| &\le \frac{2}{\min\{\delta, \frac{\beta}{2}\}}\left(\psi(H_t - H_*) - \psi(H_{t+1} - H_*)\right) DD_1 \\
+ \frac{1}{4}\left(\|X^t - X^{t-1}\| - \|X^{t+1} - X^t\| + \|Y^t - Y^{t-1}\| - \|Y^t - Y^{t+1}\| + \|Z^t - Z^{t-1}\| - \|Z^t - Z^{t+1}\|\right)
\end{aligned} \tag{65}$$

Then, denoting $\Delta_{t+1} := \|X^{t+1} - X^t\| + \|Y^{t+1} - Y^t\| + D_2\|Z^{t+1} - Z^t\|$ (65) can be further passed to

$$\frac{1}{4}\Delta_{t+1} \leq \frac{2}{\min\{\delta, \frac{\beta}{2}\}} \left(\psi(H_t - H_*) - \psi(H_{t+1} - H_*)\right) DD_1 + \frac{1}{4}\left(\Delta_t - \Delta_{t+1}\right) \tag{66}$$

Summing the above inequality from $t = t_1 + 1$ to T, we have that

$$\begin{aligned}\frac{1}{4}\sum_{t=t_1+1}^{T} \Delta_{t+1} &\leq \frac{2}{\min\{\delta, \frac{\beta}{2}\}} \left(\psi(H_t - H_*) - \psi(H_{t+1} - H_*)\right) DD_1 + \frac{1}{4}\left(\Delta_{t_1+1} - \Delta_{T+1}\right) \\ &\leq \frac{2}{\min\{\delta, \frac{\beta}{2}\}} \psi(H_t - H_*) DD_1 + \frac{1}{4}\Delta_{t_1+1}\end{aligned}$$

where the last inequality uses the fact that $\psi > 0$. Taking T in the above inequality to infinity, we see that $\sum_{t=t_1+1}^{\infty} \Delta_{t+1} < \infty$. Thus $\{(X^t, Y^t, Z^t)\}$ is convergent.

Next, we show the convergence rate of the generated sequence. Denote the limit of (X^t, Y^t, Z^t) as (X^*, Y^*, Z^*). Define $S_t = \sum_{i=t+1}^{\infty} \Delta_i$. Noting that $\|X^* - X^t\| + \|Y^* - Y^t\| + \|Z^t - Z^*\| \leq \sum_{i=t}^{\infty} \Delta_i = S_t$, it suffices to show the convergence rate of S_t. Using (66), there exists $D_2 > 0$ such that

$$\begin{aligned}S_t = \sum_{i=t}^{\infty} \Delta_i &\leq D_2\left(\psi(H_t - H_*) - \psi(H_{t+1} - H_*)\right) + \left(\Delta_t - \Delta_{t+1}\right) \\ &\leq D_2\psi(H_t - H_*) + \Delta_t = D_2\psi(H_t - H_*) + (S_{t-1} - S_t)\,.\end{aligned} \tag{67}$$

Now we bound $\psi(H_t - H_*)$. From the KL assumption, $\psi(w) = cw^{1-\theta}$ with some $c >$. Thanks to Theorem B.3 (ii) and (28), we have from the KL inequality, it holds that

$$c(1-\theta)d(0, \partial H(X^t, Y^t, Z^t, X^{t-1}, Z^{t-1})) \geq (H_t - H_*)^{\theta}. \tag{68}$$

Combining this with (B.7), we have that

$$c(1-\theta)D(S_{t-1} - S_t) \geq (H_t - H_*)^{\theta}.$$

This is equivalent to

$$c\left(c(1-\theta)D(S_{t-1} - S_t)\right)^{\frac{1-\theta}{\theta}} \geq c(H_t - H_*)^{1-\theta} = \psi(H_t - H_*).$$

Using this (67) can be further passed to

$$S_t \leq D_3(S_{t-1} - S_t)^{\frac{1-\theta}{\theta}} + (S_{t-1} - S_t)\,, \tag{69}$$

where $D_3 := D_2 c\left(c(1-\theta)D\right)^{\frac{1-\theta}{\theta}}$. Now we claim

1. When $\theta = 0$, $\{(X^t, Y^t, Z^t)\}$ converges finitely.
2. When $\theta \in (0, \frac{1}{2}]$, there exist $a > 0$ and $\rho_1 \in (0, 1)$ such that $S_t \leq a\rho_1^t$.
3. When $\theta \in (\frac{1}{2}, 1)$, there exists D_4 such that $S_t \leq ct^{-\frac{1-\theta}{2\theta-1}}$ for large t.

When $\theta = 0$, we claim that there exists t such that $H_t = H_*$. Suppose to the contrary that $H_t > H_*$ for all t. Then, for large t, (68) holds, i.e., $d(0, \partial H(X^t, Y^t, Z^t, X^{t-1}, Z^{t-1})) \geq \frac{1}{c(1-\theta)} > 0$. However, thanks to B.7 and Corollary B.4, we know that $\lim_t d(0, \partial H(X^t, Y^t, Z^t, X^{t-1}, Z^{t-1})) = 0$, a contradiction. Therefore, there exists t such that $H_t = H_*$. From the argument in the beginning of this proof, we see that $\{(X^t, Y^t, Z^t)\}$ converges finitely.

When $\theta \in (0, \frac{1}{2}]$, we have $\frac{1-\theta}{\theta} \geq 1$. Thanks to Corollary B.4, we know that there exists t_2 such that $S_t - S_{t-1} < 1$. Thus, (69) can be further passed to $S_t \leq D_3(S_{t-1} - S_t) + (S_{t-1} - S_t)$. This implies that

$$S_t \leq \frac{D_3 + 1}{D_3 + 2} S_{t-1}.$$

Dataset	Size(Input x FC layer x Output)
Synthetic	60 x 32 x 10
MNIST	784 x 128 x 10
FEMNIST	784 x 128 x 26

Table 2. The details of the neural networks in our numerical experiments.

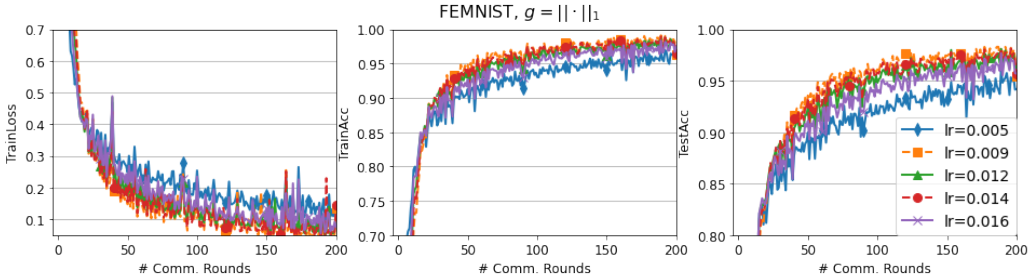

Figure 4. Results of our algorithm on FEMNIST dataset with different learning rates. ($\mathbb{L}_1$-norm regularize.)

Thus there exist $a > 0$ and $\rho_1 \in (0, 1)$ such that $S_t \le a\rho_1^t$.

When $\theta \in (\frac{1}{2}, 1)$, it holds that $\frac{1-\theta}{\theta} < 1$. From the last case, we know that $S_t - S_{t-1} < 1$ when $t > t_2$. Using (69), we have that $S_t \le D_3(S_{t-1} - S_t)^{\frac{1-\theta}{\theta}} + (S_{t-1} - S_t)^{\frac{1-\theta}{\theta}} = (D_3 + 1)\,(S_{t-1} - S_t)^{\frac{1-\theta}{\theta}}$. This implies that

$$S_t^{\frac{\theta}{1-\theta}} \le D_3^{\frac{\theta}{1+\theta}}(S_{t-1} - S_t).$$

With this inequality, following the arguments in Theorem 2 of (Attouch & Bolte, 2009) starting from Equation (13) in (Attouch & Bolte, 2009), there exists $c > 0$ such that $S_t \le ct^{-\frac{1-\theta}{2\theta-1}}$ for large t. Thus, $\{S^t\}$ converges sublinearly. □

B.5.3. Proofs of Corollary 4.11

Using (49) together with the definition of S_t in Theorem B.5.2, there exists E_1 and t',

$$d^2(\nabla F(Y^{t+1}) + \partial G(Y^{t+1})) \le E_1 S_t^2 \text{ for } t > t'. \tag{70}$$

Combining this with Proposition A.1, we know that

$$d^2(0, \sum_i \nabla f_i(y^{t+1}) + \partial g(y^{t+1})) \le E_1^2 p S_t^2 \text{ for } t > t'.$$

Using the convergence of $\{S_t\}$ and $\{(X^t, Y^t, Z^t)\}$ shown in the proof of Theorem B.5.2, we reach the conclusion.

C. Supplement for Experiment

The details of the training models. For all datasets, we apply neural networks with only Fully-connected (FC) layers as training models. The size of the models is shown as Table 2.

Hyperparameter choosing. The learning rates are 0.012 for synthetic datasets, and 0.009 for FEMNIST. For FedPD, FedDR, and FedProx, we follow (Tran-Dinh et al., 2021) to select the hyper-parameters, including μ for FedProx, η for FedPD, and η, α for FedDR. As for FedMid (Yuan et al., 2021) and FedDualAvg (Yuan et al., 2021), we also select the hyper-parameters working best for plotting the performance and comparison.

Additional Results with Different Learning Rates Figure 4 shows how different learning rates affect the performance of our FIAME on the FEMNIST dataset.

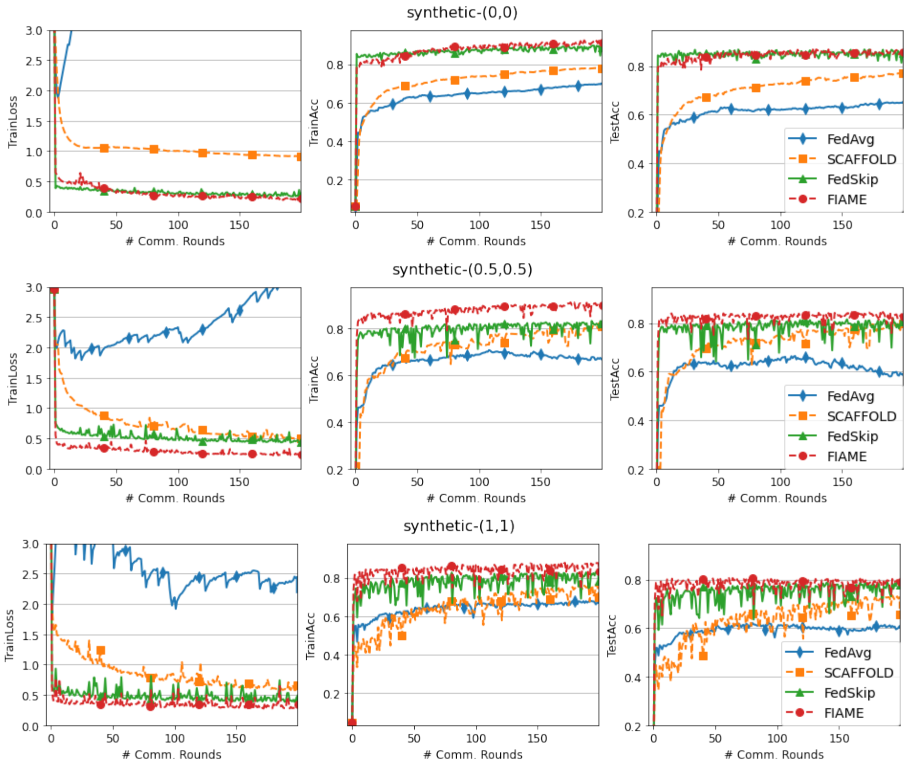

Figure 5. Results on Synthetic-{(0,0), (0.5, 0.5), (1,1)} dataset.

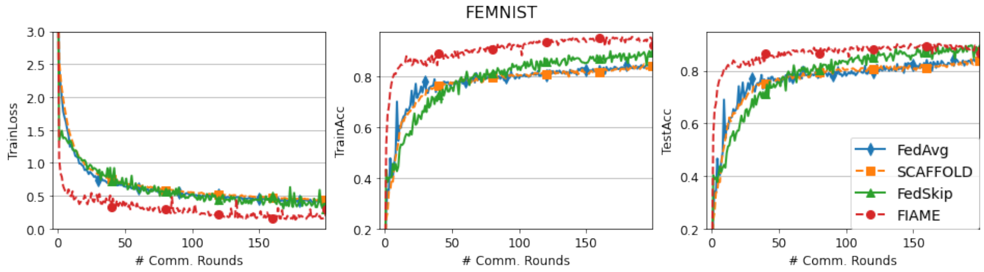

Figure 6. Results on FEMNIST dataset.

C.1. Additional results comparing FIAME with non-ADMM based FL algorithms

We compare our method with FedAvg (Li et al., 2020b), SCAFFOLD (Karimireddy et al., 2020), FedSkip (Fan et al., 2022).

Results on synthetic datasets. Following the data generation process on (Li et al., 2020a; Tran-Dinh et al., 2021), we generate three datasets: `synthetic-{(0,0), (0.5, 0.5), (1,1)}`. All agents perform updates at each communication round. Our algorithm is compared using synthetic datasets in both iid and non-iid settings. The performance of 4 algorithms on non-iid synthetic datasets is shown as Figure 5. Our algorithm can achieve better results than FedAvg, SCAFFOLD, FedSkip on all three synthetic datasets.

Results on FEMNIST dataset. FEMNIST (Cohen et al., 2017; Caldas et al., 2018) dataset is a more complex and federated extended MNIST. It has 62-class (26 upper-case and 26 lower-case letters, 10 digits) and the data is distributed to 200 devices. Figure 6 depicts the results of all 4 algorithms on FEMNIST. As it shows, compared with the other 3 methods, FIAME has a significant improvement in both training accuracy and loss value. Our algorithm can also work much better with test accuracy than the other 3 algorithms.