A Provably Communication-Efficient Asynchronous Distributed Inference Method for Convex and Nonconvex Problems

Jineng Ren, *Student Member, IEEE,* and Jarvis Haupt, *Senior Member, IEEE*

***Abstract*—This paper proposes and analyzes a communication-efficient distributed optimization framework for general nonconvex nonsmooth signal processing and machine learning problems under an asynchronous protocol. At each iteration, worker machines compute gradients of a known empirical loss function using their own local data, and a master machine solves a related minimization problem to update the current estimate. We prove that for nonconvex nonsmooth problems, the proposed algorithm converges with a sublinear rate over the number of communication rounds, coinciding with the best theoretical rate that can be achieved for this class of problems. Linear convergence is established without any statistical assumptions of the local data for problems characterized by composite loss functions whose smooth parts are strongly convex. Extensive numerical experiments verify that the performance of the proposed approach indeed improves – sometimes significantly – over other state-of-the-art algorithms in terms of total communication efficiency.**

***Index Terms*—Communication-efficient, asynchronous, distributed algorithm, convergence, nonconvex, strongly convex**

I. Introduction

DUE to rapid developments in information and computing technology, modern applications often involve vast amounts of data, rendering local processing (e.g., in a single machine, or on a single processing core) computationally challenging or even prohibitive. To deal with this problem, distributed and parallel implementations are natural methods that can fully leverage multi-core computing and storage technologies. However, one drawback of distributed algorithms is that the communication cost can be very expensive in terms of raw bytes transmitted, latency, or both, as machines (i.e., computation nodes) need to frequently transmit and receive information between each other. Therefore, algorithms that require less communication are preferred in this case.

In this paper we study a general communication-efficient distributed algorithm which can be applied to a broad class of nonconvex nonsmooth inference problems. Assume that we have available some N data samples. We consider a general problem appearing frequently in signal processing and machine learning applications; we aim to solve

$$\mathbf{x}^* = \underset{\mathbf{x}\in\mathbb{R}^p}{\operatorname{argmin}} \mathbf{L}(\mathbf{x}) := \frac{1}{N}\sum_{k=1}^{N} l_k(\mathbf{x}) + h(\mathbf{x}), \tag{1}$$

where each $l_k(\mathbf{x})$ is a loss function associated with the k-th data sample, and is assumed smooth but possibly nonconvex with Lipschitz continuous gradient and $h(\mathbf{x})$ is a convex (proper and lower semi-continuous) function that is possibly nonsmooth. Problem (2) covers many important machine learning and signal processing problems such as the localization with wireless acoustic sensor networks (WASNs) [1], support vector machine (SVM) [2], the independent principal component analysis (ICA) reconstruction problem [3], and the sparse principal component analysis (PCA) problem [4].

For our distributed approach, we consider a network of m total machines having a star topology, where one node designated as the "Master" node (node 1, without loss of generality) is located at the center of the star, and the remaining $m-1$ nodes (with indices $2, 3, \ldots, m$) are the "Worker" nodes (see Figure 1). Without loss of generality, assume that the number of data samples is evenly divisible by m, i.e., $N = nm$ for some integer n, and each machine stores n unique data samples. Then (1) can be reformulated to the following problem:

$$\mathbf{x}^* = \underset{\mathbf{x}\in\mathbb{R}^p}{\operatorname{argmin}} \mathbf{L}(\mathbf{x}) := \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n} l_{ji}(\mathbf{x}) + h(\mathbf{x}), \tag{2}$$

where $l_{ji}(\mathbf{x})$ is the loss function corresponding to the i-th sample of the j-th machine.

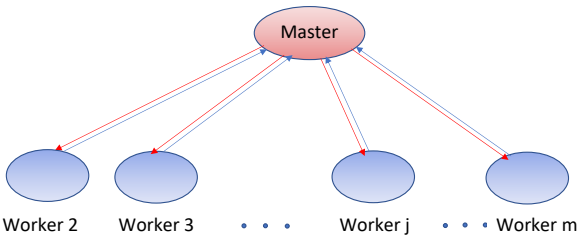


Fig. 1. m-nodes network with a star topology

A. Main Results

We propose an Efficient Distributed Algorithm for Nonconvex Nonsmooth Inference (EDANNI), and show that, for

The authors are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, 55455 USA; e-mails: {renxx282,jdhaupt}@umn.edu. Shorter preliminary versions of this work appeared at the 2018 Global Conference on Signal and Information Processing (GlobalSIP 2018).

general problems of the form of (2), EDANNI converges to the set of stationary points if the algorithm parameters are chosen appropriately according to the maximum network delay. Our results differ significantly from existing works [5]–[7] which are all developed for convex problems. Therefore, the analysis and algorithm proposed here are applicable not only to standard convex learning problems but also to important nonconvex problems. To the best of our knowledge, this is the first communication-efficient algorithm exploiting local second-order information that is guaranteed to be convergent for general nonconvex nonsmooth problems. Moreover, linear convergence is also proved in the strongly convex setting with no statistical assumption of the data stored in each local machine, which is another improvement on existing works. The synchronization inherent in previous works, including [5]–[10], slows down those methods because the master needs to wait for the slowest worker during each iteration; here, we propose an asynchronous approach that can accelerate the corresponding inference tasks significantly (as we will demonstrate in the experimental results).

B. Related Work

There is a large body of work on distributed optimization for modern data-intensive applications with varied accessibility; see, for example, [5]–[7], [11]–[23]. (Parts of the results presented here appeared in our conference paper [22] without theoretical analysis.) Early works including [6], [11], [15] mainly considered the convergence of parallelizing stochastic gradient descent schemes which stem from the idea of the seminal text by Bertsekas and Tsitsiklis [16]. Niu et al. [12] proposed a lock-free implementation of distributed SGD called Hogwild! and provided its rates of convergence for sparse learning problems. That was followed up by many variants like [24], [25]. For solving large scale problems, works including [17], [18], [19], and [26] studied distributed optimizations based on a parameter server framework and parameters partition. Chang et al. [20] studied asynchronous distributed optimizations based on the alternating direction method of multipliers (ADMM). By formulating the optimization problem as a consensus problem, the ADMM can be used to solve the consensus problem in a fully parallel fashion over networks with a star topology. One drawback of such approaches is that they can be computationally intensive, since each worker machine is required to solve a high dimensional subproblem. As we will show, these methods also converge more slowly (in terms of communication rounds) as compared to the proposed approach (see Section IV).

A growing interest on distributed algorithms also appears in the statistics community [27]–[31]. Most of these algorithms depend on the partition of data, so their work usually involves statistical assumptions that handle the correlation between the data in local machines. A popular approach in early works is averaging estimators generated locally by different machines [15], [28], [32], [33]. Yang [34], Ma et al. [35], and Jaggi et al. [36] studied distributed optimization based on stochastic dual coordinate descent, however, their communication complexity is not better than that of first-order approaches. Shamir et al. [37] and Zhang and Xiao [38] proposed truly communication-efficient distributed optimization algorithms which leveraged the local second-order information, though these approaches are only guaranteed to work for convex and smooth objectives. In a similar spirit, Wang et al. [8], Jordan et al. [9], and Ren et al. [10] developed communication-efficient algorithms for sparse learning with ℓ_1 regularization. However, each of these works needs an assumption about the strong convexity of loss functions, which may limit their approaches to only a small set of real-world applications. Here we describe an algorithm with similar flavor, but with more general applicability, and establish its convergence rate in both strongly convex and nonconvex nonsmooth settings. Moreover, unlike [8]–[10], [37], [38] where the convergence analyses rely on certain statistical assumptions on the data stored in machines, our convergence analysis is deterministic and characterizes the worst-case convergence conditions.

Notation. For a vector $\mathbf{v} = (v_1, \cdots, v_s)^\top \in \mathbb{R}^s$ and $q > 0$ we write $\|\mathbf{v}\|_q = (\sum_{i=1}^s |v_i|^q)^{1/q}$; for $q \geq 1$ this is a norm. Usually $\|\mathbf{v}\|_2$ is briefly written as $\|\mathbf{v}\|$. The set of natural numbers is denoted by $\mathbb{N}$. For an integer $m \in \mathbb{N}$, we write $[m]$ as shorthand for the set $\{1, \ldots, m\}$.

II. Algorithm

In this section, we describe our approach to computing the minimizer $\mathbf{x}^*$ of (2). Recall that we have m machines. Let us denote $t \geq 0$ as the iteration number, then $\mathcal{A}_t \subseteq [m] := \{1, 2, \cdots, m\}$ is defined as the index of a subset of worker machines from which the master receives updated gradient information during iteration t; worker i is said to be "arrived" if $i \in \mathcal{A}_t$. At iteration t, the master machine solves a subproblem to obtain an updated estimate, and communicates this to the worker machines in the subset $\mathcal{A}_t$. After receiving the updated estimate, the worker machines will compute the corresponding gradients of local empirical loss functions. These gradients are then communicated back to the master machine, and the process continues.

Formally, let

$$\mathbf{L}_j(\mathbf{x}) = \frac{1}{n} \sum_{i \in [n]} l_{ji}(\mathbf{x}), \quad j \in [m]$$

be the empirical loss at each machine. Let t_j be the latest time (in terms of the iteration count) when the worker j is arrived up to and including iteration t.

In the t-th iteration, the master (machine 1) solves the following subproblem to update $\mathbf{x}^{t+1}$

$$\mathbf{x}^{t+1} = \operatorname*{argmin}_{\mathbf{x}} \mathbf{L}_1(\mathbf{x}) + h(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{x} - \mathbf{x}^t\|^2 + \Big\langle \frac{1}{m} \sum_{j \in [m]} \nabla \mathbf{L}_j(\mathbf{x}^{t_j}) - \nabla \mathbf{L}_1(\mathbf{x}^{t_1}), \mathbf{x} - \mathbf{x}^t \Big\rangle. \quad (3)$$

This $\mathbf{x}^{t+1}$ is communicated to the worker machines that are free, where it is used to compute their local gradient $\nabla \mathbf{L}_j(\mathbf{x}^{t+1})$. Since machine 1 is assumed to be the master machine, t_1 is actually t.

Now one question is: which partial sets of worker machines (with indices in $\mathcal{A}_t$) from which the master receives updated

gradient information during iteration t are sufficient to ensure convergence of a distributed approach? Firstly, let $\tau \geq 0$ be a maximum tolerable delay, that is, the maximum number of iterations for which every worker machine can be inactive. The set $\mathcal{A}_t$ should satisfy:

Assumption II.1 (Bounded delay)**.** *For all $i \in [m]$ and iteration $t \geq 0$, it holds that $i \in \mathcal{A}_t \cup \mathcal{A}_{t-1} \cup \cdots \cup \mathcal{A}_{\max\{t-\tau,0\}}$.*

To satisfy Assumption II.1, $\mathcal{A}_t$ should contain at least the indices of the worker machines that have been inactive for longer than τ iterations. That is, the master needs to wait until those workers finish their current computation and have arrived. Note that by the definition of t_j, it holds that

$$t - \tau \leq t_j \leq t, \quad \forall j \in [m].$$

Assumption II.1 requires that every worker j is arrived at least once within the period $[t-\tau, t]$. In other words, the gradient information $\nabla \mathbf{L}_j(\mathbf{x}^{t_j})$ used by the master must be at most τ iterations old. To guarantee the bounded delay, at every iteration the master needs to wait for the workers who have not been active for τ iterations, if such workers exist. Note that, when $\tau = 0$, one has $j \in \mathcal{A}_t$ for all $j \in [m]$, which reduces to the synchronous case where the master always waits for all the workers at every iteration.

The proposed approach is presented in Algorithm 1, which specifies respectively the steps for the workers and the master. Algorithm 1 has three prominent differences compared with its synchronous alternatives. First, only the workers j in $\mathcal{A}_t$ update the gradient $\nabla \mathbf{L}_j(\mathbf{x}^t)$ and transmit it to the master machine. For the workers j in $\mathcal{A}_t^c$, the master uses their latest gradient information before t, i.e., $\nabla \mathbf{L}_j(\mathbf{x}^{t_j})$. Second, the variables d_j's are introduced to count the delays of the workers since their last updates. d_j is set to zero if worker j is arrived at the current iteration; otherwise, d_j is increased by one. Therefore, to ensure Assumption II.1 holds at each iteration, the master should wait if there exists at least one worker whose $d_j > \tau - 1$. Third, after solving subproblem (3), the master transmits the up-to-date variable $\mathbf{x}^{t+1}$ only to the arrived workers. In general both the master and fast workers in the asynchronous approach can update more frequently and have less waiting time than their synchronous counterparts.

Algorithm 1: Efficient Distributed Algorithm for Nonconvex-Nonsmooth Inference (EDANNI)

Input: Loss functions $\{l_{ji}(\cdot,\cdot)\}_{i\in[n],j\in[m]}$, parameter ρ, initial point $\mathbf{x}^0$. Set $d_1 = \cdots = d_m = 0$ and $\mathcal{A}_0 = [m]$;

for $t = 0, 1, \ldots$ **do**

Worker machines:

for $j = 2, 3, \ldots, m$ **do**

if Receive $\mathbf{x}^t$ from the master **then**

Calculate gradient $\nabla \mathbf{L}_j(\mathbf{x}^t)$ and transmit it to the master.

end

end for

Master:

Receive $\{\nabla \mathbf{L}_j(\mathbf{x}^t)\}_{j=2}^m$ from worker machines j in a set $\mathcal{A}_t$ such that $d_j \leq \tau - 1, \ \ \forall j \in \mathcal{A}_t^c$ **then**

Update

$$d_j = \begin{cases} 0 & \forall j \in \mathcal{A}_t \\ d_j + 1 & \forall j \in \mathcal{A}_t^c \end{cases}.$$

Solve the subproblem (3) with the specified ρ to obtain $\mathbf{x}^{t+1}$. Broadcast $\mathbf{x}^{t+1}$ to the worker machines j that are free.

end for

III. Theoretical Analysis

Solving subproblem (3) is inspired by the approaches of Shamir et al. [37], et al., Wang et al. [8], and Jordan et al. [9], and is designed to take advantage of both global first-order information and local higher-order information. Indeed, when $\rho = 0$ and $\mathbf{L}_j$ is quadratic, (3) has the following closed form solution:

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \nabla^2 \mathbf{L}_1(\mathbf{x}^t)^{-1} \left(\tfrac{1}{m} \textstyle\sum_{j=1}^m \nabla \mathbf{L}_j(\mathbf{x}^{t_j}) \right),$$

which is similar to a Newton updating step. The more general case has a *proximal Newton* flavor; see, e.g., [39] and the references therein. However, our method is different from their methods in the proximal term $\frac{\rho}{2}\|\mathbf{x} - \mathbf{x}^t\|^2$ as well as the first order term. Intuitively, if we have a first-order approximation

$$\mathbf{L}_1(\mathbf{x}) \approx \mathbf{L}_1(\mathbf{x}^t) + \left\langle \nabla \mathbf{L}_1(\mathbf{x}^{t_1}), \mathbf{x} - \mathbf{x}^t \right\rangle, \tag{4}$$

then (3) reduces to

$$\mathbf{x}^{t+1} = \operatorname*{argmin}_{\mathbf{x}} \left\langle \tfrac{1}{m} \textstyle\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t_j}), \mathbf{x} - \mathbf{x}^t \right\rangle + h(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{x} - \mathbf{x}^t\|^2, \tag{5}$$

which is essentially a first-order proximal gradient updating step.

We consider the convergence of the proposed approach under the asynchronous protocol where the master has the freedom to make updates with gradients from only a partial set of worker machines. We start with introducing important conditions that are used commonly in previous work [13], [20], [40].

Assumption III.1. *The function $\mathbf{L}_j(\mathbf{x})$ is differentiable and has Lipschitz continuous gradient for all $j \in [m]$, i.e.,*

$$\|\nabla \mathbf{L}_j(\mathbf{x}) - \nabla \mathbf{L}_j(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|.$$

The proof of the linear convergence relies on the following strong convexity assumption.

Assumption III.2. *For all $j \in [m]$, the function $\mathbf{L}_j$ is strongly convex with modulus σ^2, which means that*

$$\mathbf{L}_j(x) > \mathbf{L}_j(y) + \langle \nabla \mathbf{L}_j(y), x - y \rangle + \frac{\sigma^2}{2}\|x - y\|^2,$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, $j \in [m]$.

Assumption III.3. *For all t, the parameter ρ in* (3) *is chosen large enough such that:*

I. *$\gamma(\rho) > 3L + 2L\delta\tau$ and $\rho > \frac{2L\tau}{\delta}$, for some constant*

$\delta > 0$, *where* $\gamma(\rho)$ *represents the convex modulus of the function* $h(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{x} - \mathbf{x}^t\|^2$.

II. *There exists a constant* $\underline{L}$ *such that*

$$\mathbf{L}(\mathbf{x}) > \underline{L} > -\infty \quad \forall \mathbf{x} \in \mathbb{R}^p.$$

Moreover the following concept is needed in the first part of Theorem III.1.

Definition III.1. *We say a function* $\mathcal{F}(\mathbf{x})$ *is coercive if*

$$\lim_{\|\mathbf{x}\| \to \infty} \mathcal{F}(\mathbf{x}) = +\infty.$$

Define

$$\widetilde{\nabla}_{\mathbf{x}} \mathbf{L}\left(\mathbf{x}^t\right) = \mathbf{x}^t - \mathbf{Prox}_h\Big(\mathbf{x}^t - \tfrac{1}{m} \sum_{j \in [m]} \nabla \mathbf{L}_j(\mathbf{x}^t)\Big), \tag{6}$$

where $\mathbf{Prox}_h$ is a proximal operator defined by $\mathbf{Prox}_h[z] := \operatorname{argmin}_x h(x) + \frac{1}{2}\|x - z\|^2$. Usually $\widetilde{\nabla}_{\mathbf{x}} \mathbf{L}\left(\mathbf{x}^t\right)$ is called the proximal gradient of $\mathbf{L}$; $\mathbf{x}$ is a stationary point when $\widetilde{\nabla}_{\mathbf{x}} \mathbf{L}\left(\mathbf{x}\right) = 0$.

Based on these assumptions, now we can present the main theorem.

Theorem III.1. *Suppose Assumption II.1, III.1, and III.3 are satisfied. Then we have the following claims for the sequence generated by Algorithm 1 (EDANNI).*

- **(Boundedness of Sequence).** *The gap between* $\mathbf{x}^t$ *and* $\mathbf{x}^{t+1}$ *converges to* 0*, i.e.,*

  $$\lim_{t \to \infty} \mathbf{x}^{t+1} - \mathbf{x}^t = 0.$$

  If $\mathbf{L}(\mathbf{x})$ *is coercive, then the sequence* $\{\mathbf{x}^t\}$ *generated by Algorithm 1 is bounded.*
- **(Convergence to Stationary Points).** *Every limit point of the iterates* $\{\mathbf{x}^t\}$ *generated by Algorithm 1 is a stationary point of problem* (2)*. Furthermore,* $\left\|\widetilde{\nabla}_{\mathbf{x}} \mathbf{L}(\mathbf{x}^t)\right\| \to 0$*, as* $t \to \infty$.
- **(Sublinear Convergence Rate).** *Given* $\epsilon > 0$*, let us define* T *to be the first time for the optimality gap to reach below* ϵ*, i.e.,*

  $$T := \operatorname*{argmin}_t \left\{ \left\|\widetilde{\nabla}_{\mathbf{x}} \mathbf{L}(\mathbf{x}^t)\right\| < \epsilon \right\}.$$

  Then there exists a constant $\nu > 0$ *such that*

  $$T \leq \frac{\nu}{\epsilon} + 1,$$

  where ν *equals to a positive constant times* $\left(2(2+\rho)^2 + 8L^2\tau\right) / \min\left\{\frac{\gamma(\rho)}{2} - \frac{3L}{2} - L\delta\tau, \frac{\rho}{2} - \frac{L\tau}{\delta}\right\}$ *for some* $\delta > 0$*. Therefore, the optimality gap* $\left\|\widetilde{\nabla}_{\mathbf{x}} \mathbf{L}(\mathbf{x}^t)\right\|$ *converges to* 0 *in a sublinear manner.*

Remark III.1. *The theorem suggests that the iterates* $\{\mathbf{x}^t\}$ *may or may not be bounded without the coerciveness property of* $\mathbf{L}(\mathbf{x})$*. However, it guarantees that the optimality measure* $\left\|\widetilde{\nabla}_x \mathbf{L}(\mathbf{x}^t)\right\|$ *converges to* 0 *sublinearly. We remark that [19] also analyzed the convergence of a proximal gradient method based communication-efficient algorithm for nonconvex problems, but they did not give a specific convergence rate. Note that such sublinear complexity bound is tight when applying first-order methods for nonconvex unconstrained problems (see [41], [42]).*

Let us define

$$\mathrm{F}(\mathbf{x}, \mathbf{x}^t) := \tfrac{1}{m} \textstyle\sum_{j \in [m]} \mathbf{L}_j(\mathbf{x}) + \frac{\rho}{2} \left\|\mathbf{x} - \mathbf{x}^t\right\|^2 + h(\mathbf{x}).$$

The gap between $\mathbf{x}^t$ and $\mathbf{x}^{t+1}$ is denoted by $\Delta^{(t)} := \mathbf{x}^{t+1} - \mathbf{x}^t$, for $t \in \mathbb{N}$. The proof of Theorem III.1 relies on Lemma III.1, III.2, and III.3 in the following.

Lemma III.1. *Suppose Assumption III.1 and Assumption III.3 (I) are satisfied. then the following is true for iterates* $\{\mathbf{x}^t\}$ *generated by Algorithm 1 (EDANNI)*

$$\begin{aligned}
&\frac{\rho}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 + h(\mathbf{x}^{t+1}) - \frac{\rho}{2}\|\mathbf{x}^t - \mathbf{x}^{t-1}\|^2 - h(\mathbf{x}^t) \\
&\leq -\Bigg\langle \nabla \mathbf{L}_1(\mathbf{x}^{t+1}) + \frac{1}{m}\sum_{j=1}^{m} \nabla \mathbf{L}_j(\mathbf{x}^{t_j}) - \nabla \mathbf{L}_1(\mathbf{x}^{t_1}), \\
&\Delta^{(t)} \Bigg\rangle - \frac{\gamma(\rho)}{2}\left\|\Delta^{(t)}\right\|^2 - \frac{\rho}{2}\left\|\Delta^{(t-1)}\right\|^2.
\end{aligned} \tag{7}$$

Lemma III.2. *Under the assumptions of Theorem III.1 for any* $\delta > 0$ *we have*

$$\begin{aligned}
&F(\mathbf{x}^{t+1}, \mathbf{x}^t) - F(\mathbf{x}^t, \mathbf{x}^{t-1}) \\
&\leq \left(\frac{3L}{2} - \frac{\gamma(\rho)}{2} + L\delta\tau\right)\left\|\Delta^{(t)}\right\|^2 - \frac{\rho}{2}\left\|\Delta^{(t-1)}\right\|^2 \\
&\qquad + \frac{L}{\delta}\sum_{k=1}^{\tau}\left\|\Delta^{(t-k)}\right\|^2.
\end{aligned} \tag{8}$$

Lemma III.3. *Suppose Assumption III.3 is satisfied. Then for* $\mathbf{x}^t$ *generated by (EDANNI), there exists some constants* $\underline{F}$ *and* $\bar{F}$ *such that*

$$+\infty > \bar{F} > F(\mathbf{x}^{t+1}, \mathbf{x}^t) > \underline{F} > -\infty, \quad \forall t \geq 0.$$

The proofs of these lemmata are in the Appendix. Now in the following we prove Theorem III.1.

Proof of Theorem III.1. We begin by establishing the first conclusion of the theorem. Summing inequality (8) in Lemma III.2 over t yields

$$\begin{aligned}
&\mathrm{F}(\mathbf{x}^{T+1}, \mathbf{x}^T) - \mathrm{F}(\mathbf{x}^1, \mathbf{x}^0) \\
&\leq \sum_{t=1}^{T}\left(\frac{3L}{2} - \frac{\gamma(\rho)}{2} + L\delta\tau\right)\left\|\Delta^{(t)}\right\|^2 \\
&\qquad + \sum_{t=1}^{T}\left(\frac{L\tau}{\delta} - \frac{\rho}{2}\right)\left\|\Delta^{(t-1)}\right\|^2.
\end{aligned}$$

Now define

$$c := \min\left\{\frac{\gamma(\rho)}{2} - \frac{3L}{2} - L\delta\tau, \frac{\rho}{2} - \frac{L\tau}{\delta}\right\},$$

by Assumption III.3 we have $\gamma(\rho) > 3L + 2L\delta\tau$ and $\rho > \frac{2L\tau}{\delta}$, therefore $c > 0$. It holds that

$$\mathrm{F}(\mathbf{x}^{T+1}, \mathbf{x}^T) - \mathrm{F}(\mathbf{x}^1, \mathbf{x}^0) \leq -c\sum_{t=0}^{T}\left\|\Delta^{(t)}\right\|^2. \tag{9}$$

Note that by Lemma III.3 the LHS of (9) is bounded from below. By letting $T \to \infty$, it follows that

$$\left\|\Delta^{(t)}\right\| \to 0, \qquad t \to \infty.$$

Moreover, Lemma III.3 shows that $\mathrm{F}(\mathbf{x}^{t+1}, \mathbf{x}^t)$ is bounded, but due to the coerciveness assumption

$$\lim_{\|\mathbf{x}\|\to\infty} \frac{1}{m}\sum_{j\in[m]} \mathbf{L}_j(\mathbf{x}) + h(\mathbf{x}) + \frac{\rho}{2}\left\|\mathbf{x}-\mathbf{x}^t\right\|^2 = +\infty, \quad (10)$$

so we know $\{\mathbf{x}^{t+1}\}$ is bounded. Therefore the first conclusion is proved.

We now establish the second conclusion of the Theorem. From (3), we know that

$$\mathbf{x}^{t+1} = \mathbf{Prox}_h\Big[\mathbf{x}^{t+1} - \Big(\nabla\mathbf{L}_1(\mathbf{x}^{t+1}) + \frac{1}{m}\sum_{j\in[m]}\nabla\mathbf{L}_j(\mathbf{x}^{t_j}) - \nabla\mathbf{L}_1(\mathbf{x}^{t_1}) + \rho\left(\mathbf{x}^{t+1}-\mathbf{x}^t\right)\Big)\Big],$$

where $\mathbf{Prox}_h$ is a proximal operator defined by $\mathbf{Prox}_h[z] := \operatorname{argmin}_x h(x) + \frac{1}{2}\|x-z\|^2$. This implies that

$$\begin{aligned}
&\left\|\mathbf{x}^t - \mathbf{Prox}_h\left(\mathbf{x}^t - \frac{1}{m}\sum_{j\in[m]}\nabla\mathbf{L}_j(\mathbf{x}^t)\right)\right\| \\
&\le \Bigg\|\mathbf{x}^t - \mathbf{x}^{t+1} + \mathbf{x}^{t+1} \\
&- \mathbf{Prox}_h\left(\mathbf{x}^t - \frac{1}{m}\sum_{j\in[m]}\nabla\mathbf{L}_j(\mathbf{x}^t)\right)\Bigg\| \\
&\le \left\|\mathbf{x}^t - \mathbf{x}^{t+1}\right\| \\
&+ \Bigg\|\mathbf{Prox}_h\Big[\mathbf{x}^{t+1} - \Big(\nabla\mathbf{L}_1(\mathbf{x}^{t+1}) + \frac{1}{m}\sum_{j\in[m]}\nabla\mathbf{L}_j(\mathbf{x}^{t_j}) \\
&- \nabla\mathbf{L}_1(\mathbf{x}^{t_1}) + \rho\big(\mathbf{x}^{t+1}-\mathbf{x}^t\big)\Big)\Big] \\
&- \mathbf{Prox}_h\left(\mathbf{x}^t - \frac{1}{m}\sum_{j\in[m]}\nabla\mathbf{L}_j(\mathbf{x}^t)\right)\Bigg\| \\
&\overset{(a)}{\le} \Bigg\|(1+\rho)(\mathbf{x}^{t+1}-\mathbf{x}^t) + \frac{1}{m}\sum_{j\in[m]}\nabla\mathbf{L}_j(\mathbf{x}^t) \\
&- \frac{1}{m}\sum_{j\in[m]}\nabla\mathbf{L}_j(\mathbf{x}^{t_j}) - \big(\nabla\mathbf{L}_1(\mathbf{x}^{t+1}) - \nabla\mathbf{L}_1(\mathbf{x}^{t_1})\big)\Bigg\| \\
&+ \left\|\Delta^{(t)}\right\| \\
&\le (2+\rho)\left\|\Delta^{(t)}\right\| + 2L\sum_{k=0}^{\tau}\left\|\Delta^{(t-k)}\right\| \\
&\longrightarrow 0, \qquad t \to \infty. \qquad (11)
\end{aligned}$$

Note that here inequality (a) holds because of the nonexpansiveness of the operator $\mathbf{Prox}_h$. The last inequality follows from Assumption III.1.

Let $\mathbf{X}^*$ be the set of stationary points of problem (2), and let

$$\operatorname{dist}\left(\mathbf{x}^t, \mathbf{X}^*\right) := \min_{\widehat{x}\in\mathbf{X}^*}\|\mathbf{x}^t - \widehat{\mathbf{x}}\|$$

denote the distance between $\mathbf{x}^t$ and the set $\mathbf{X}^*$. Now we prove

$$\lim_{t\to\infty}\operatorname{dist}\left(\mathbf{x}^t, \mathbf{X}^*\right) = 0.$$

Suppose there exists a subsequence $\{\mathbf{x}^{t_k}\}$ of $\{\mathbf{x}^t\}$ such that $\mathbf{x}^{t_k} \to \widehat{\mathbf{x}},\ k \to \infty$ but

$$\lim_{k\to\infty}\operatorname{dist}(\mathbf{x}^{t_k}, \mathbf{X}^*) \ge \gamma > 0. \quad (12)$$

Then it is obvious that $\lim_{k\to\infty}\operatorname{dist}(\mathbf{x}^{t_k}, \widehat{\mathbf{x}}) = 0$. Therefore there exists some $K(\gamma) > 0$, such that

$$\|\mathbf{x}^{t_k} - \widehat{\mathbf{x}}\| \le \frac{\gamma}{2}, \quad k > K(\gamma). \quad (13)$$

On the other hand, from (11) and the lower semi-continuity of $h(\mathbf{x})$ we have $\widehat{\mathbf{x}} \in \mathbf{X}^*$, so by the definition of the distance function we have

$$\operatorname{dist}(\mathbf{x}^{t_k}, \mathbf{X}^*) \le \operatorname{dist}(\mathbf{x}^{t_k}, \widehat{\mathbf{x}}). \quad (14)$$

Combining (13) and (14), we must have

$$\operatorname{dist}(\mathbf{x}^{t_k}, \mathbf{X}^*) \le \frac{\gamma}{2}, \quad k > K(\gamma).$$

This contradicts to (12), so the second result is proved.

We finally prove the third conclusion of the Theorem. Summing (11) over t yields

$$\begin{aligned}
&\sum_{t=0}^{T}\left\|\mathbf{x}^t - \mathbf{Prox}_h\left(\mathbf{x}^t - \frac{1}{m}\sum_{j\in[m]}\nabla\mathbf{L}_j(\mathbf{x}^t)\right)\right\|^2 \\
&\le \sum_{t=0}^{T} 2(2+\rho)^2\left\|\Delta^{(t)}\right\|^2 + 2(2L)^2\sum_{t=0}^{T}\sum_{k=0}^{\tau}\left\|\Delta^{(t-k)}\right\|^2 \\
&\le \left(2(2+\rho)^2 + 8L^2\tau\right)\sum_{t=0}^{T}\left\|\Delta^{(t)}\right\|^2. \qquad (15)
\end{aligned}$$

Combining (9) and (15) we have

$$\sum_{t=0}^{T}\left\|\widetilde{\nabla}\mathbf{L}\left(\mathbf{x}^t\right)\right\|^2 \le \frac{\mu}{c}\left(\mathrm{F}(\mathbf{x}^1, \mathbf{x}^0) - \mathrm{F}(\mathbf{x}^{T+1}, \mathbf{x}^T)\right),$$

where $\mu := \left(2(2+\rho)^2 + 8L^2\tau\right)$.

Let $T(\epsilon) := \min\left\{t \mid \left\|\widetilde{\nabla}\mathbf{L}(\mathbf{x}^t)\right\| \le \epsilon, t \ge 0\right\}$. Then the above inequality implies

$$T(\epsilon)\epsilon \le \frac{\mu}{c}\left(\mathrm{F}(\mathbf{x}^1, \mathbf{x}^0) - \mathrm{F}(\mathbf{x}^{T+1}, \mathbf{x}^T)\right).$$

Thus it follows that

$$\epsilon \le \frac{C\cdot\left(\mathrm{F}(\mathbf{x}^1, \mathbf{x}^0) - \underline{\mathrm{F}}\right)}{T(\epsilon)},$$

where $C := \frac{\mu}{c} > 0$, proving Theorem III.1. □

Besides the convergence in the nonconvex setting, in the following theorem we show that the proposed algorithm converges linearly if $\mathbf{L}_j$ is strongly convex. Quite interestingly, comparing with the results of [8]–[10], here the linear conver-

gence is established without any statistical assumption of the data stored in each local machine.

Theorem III.2. *Suppose Assumption II.1, III.1, and III.2 are satisfied. If ρ is sufficiently large such that*

$$\frac{\delta_1 L+\frac{\rho}{2}(1+\delta_1)}{\frac{\rho}{2}(1+\delta_1)+\delta_1}+\frac{3L}{2}-\frac{\rho}{2}+L\delta\tau<0$$

and

$$\begin{aligned}&\frac{\delta_1 L+\frac{\rho}{2}(1+\delta_1)}{\frac{\rho}{2}(1+\delta_1)+\delta_1}+\frac{3L}{2}-\frac{\rho}{2}+L\delta\tau\\&-\frac{\rho\eta}{2}+\Big(\frac{L}{\delta}+\frac{\frac{\delta_1}{2}L^2\tau}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\Big)\frac{\eta^\tau-1}{\eta-1}<0,\end{aligned}$$

for some $\delta>0$ and $\delta_1>(2L+\rho+1)/\sigma^2$, then it holds for the sequence generated by (EDANNI) that

$$\begin{aligned}0&\le F(\mathbf{x}^{t+1},\mathbf{x}^t)-F(\mathbf{x}^*,\mathbf{x}^*)\\&\le\frac{1}{\eta^t}(F(\mathbf{x}^1,\mathbf{x}^0)-F(\mathbf{x}^*,\mathbf{x}^*)),\end{aligned}$$

where $\eta:=1+\frac{1}{\frac{\rho}{2}(1+\delta_1)+\delta_1}$.

Note the above conditions can be satisfied when ρ is sufficiently larger than the order of L and the exponential of τ and δ_1 is larger than the order of L/σ^2. Theorem III.2 asserts that with the strongly convexity of $\mathbf{L}_j$'s, the augmented optimality gap decreases linearly to zero under these conditions. Moreover, Assumption III.2 can be replaced by only requiring each $\mathbf{L}_j$ is convex and $\frac{1}{m}\sum_{j\in[m]}\mathbf{L}_j$ is strongly convex with modulus σ^2. To prove Theorem III.2, we need the following lemma to bound the optimality gap of function F.

Lemma III.4. *Suppose Assumption II.1, III.1, and III.2 hold and $\delta_1>(2L+\rho+1)/\sigma^2$ for some $\delta_1>0$, then it follows that*

$$\begin{aligned}&\frac{1}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\big(F(\mathbf{x}^{t+1},\mathbf{x}^t)-F(\mathbf{x}^*,\mathbf{x}^*)\big)\\&\le\frac{\delta_1 L+\frac{\rho}{2}(1+\delta_1)}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\|\mathbf{x}^t-\mathbf{x}^{t+1}\|^2\\&+\frac{1}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\frac{\delta_1}{2m}L^2\sum_{j\in[m]}\|\mathbf{x}^{t_j}-\mathbf{x}^t\|^2.\end{aligned}\tag{16}$$

The proof of Lemma III.4 is in the Appendix. Now we begin to prove Theorem III.2.

Proof of Theorem III.2. We begin by defining $\widetilde{\Delta}^{(t+1)}=\mathrm{F}(\mathbf{x}^{t+1},\mathbf{x}^t)-\mathrm{F}(\mathbf{x}^*,\mathbf{x}^*)$. Then from the proof of Lemma III.2 it holds that

$$\begin{aligned}\widetilde{\Delta}^{(t+1)}&\le\widetilde{\Delta}^{(t)}+\left(\frac{3L}{2}-\frac{\rho}{2}+L\delta\tau\right)\big\|\Delta^{(t)}\big\|^2\\&-\frac{\rho}{2}\big\|\Delta^{(t-1)}\big\|^2+\left(\frac{L}{\delta}\sum_{k=1}^{\tau}\big\|\Delta^{(t-k)}\big\|^2\right)\end{aligned}\tag{17}$$

Note that from (16) of Lemma III.4 we have

$$\begin{aligned}\frac{1}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\widetilde{\Delta}^{(t+1)}&\le\frac{\delta_1 L+\frac{\rho}{2}(1+\delta_1)}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\big\|\Delta^{(t)}\big\|^2\\&+\frac{1}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\frac{\delta_1}{2m}L^2\sum_{j\in[m]}\|\mathbf{x}^{t_j}-\mathbf{x}^t\|^2\end{aligned}\tag{18}$$

By combining (18) and (17), we have the following bound of the LHS:

$$\begin{aligned}&\left(1+\frac{1}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\right)\widetilde{\Delta}^{(t+1)}\\&\le\widetilde{\Delta}^{(t)}+\Big[\frac{\delta_1 L+\frac{\rho}{2}(1+\delta_1)}{\frac{\rho}{2}(1+\delta_1)+\delta_1}+\frac{3L}{2}-\frac{\rho}{2}+L\delta\tau\Big]\big\|\Delta^{(t)}\big\|^2\\&-\frac{\rho}{2}\big\|\Delta^{(t-1)}\big\|^2+\frac{L}{\delta}\sum_{k=1}^{\tau}\big\|\Delta^{(t-k)}\big\|^2\\&+\frac{1}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\frac{\delta_1}{2m}L^2\sum_{j\in[m]}\|\mathbf{x}^{t_j}-\mathbf{x}^t\|^2.\end{aligned}\tag{19}$$

Inequality (19) gives us an relation between $\widetilde{\Delta}^{(t+1)}$ and $\widetilde{\Delta}^{(t)}$. Let us define $\eta:=1+\frac{1}{\frac{\rho}{2}(1+\delta_1)+\delta_1}$ and

$$(P3):=\frac{\delta_1 L+\frac{\rho}{2}(1+\delta_1)}{\frac{\rho}{2}(1+\delta_1)+\delta_1}+\frac{3L}{2}-\frac{\rho}{2}+L\delta\tau,$$

then by applying (19) recursively we have

$$\begin{aligned}&\widetilde{\Delta}^{(t+1)}\\&\le\frac{1}{\eta}\widetilde{\Delta}^{(t)}+\frac{1}{\eta}(P3)\big\|\Delta^{(t)}\big\|^2-\frac{\rho}{2\eta}\big\|\Delta^{(t-1)}\big\|^2\\&+\frac{L}{\delta\eta}\sum_{k=1}^{\tau}\big\|\Delta^{(t-k)}\big\|^2+\frac{1}{\eta}\frac{1}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\frac{\delta_1 L^2}{2m}\sum_{j\in[m]}\|\mathbf{x}^{t_j}-\mathbf{x}^t\|^2\\&\le\frac{1}{\eta^2}\widetilde{\Delta}^{(t-1)}+\frac{1}{\eta}\left(\frac{1}{\eta}(P3)\big\|\Delta^{(t-1)}\big\|^2-\frac{\rho}{2\eta}\big\|\Delta^{(t-2)}\big\|^2\right)\\&+\frac{1}{\eta}(P3)\big\|\Delta^{(t)}\big\|-\frac{\rho}{2\eta}\big\|\Delta^{(t-1)}\big\|^2\\&+\left(\frac{L}{\delta\eta}\sum_{k=1}^{\tau}\big\|\Delta^{(t-k)}\big\|^2+\frac{L}{\delta\eta^2}\sum_{k=1}^{\tau}\big\|\Delta^{(t-1-k)}\big\|^2\right)\\&+\frac{1}{\eta^2}\frac{1}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\frac{\delta_1}{2m}L^2\sum_{l=0}^{1}\frac{1}{\eta^{l+1}}\sum_{j\in[m]}\sum_{k=1}^{\tau}\big\|\Delta^{(t-l-k)}\big\|^2\\&\cdots\\&\le\frac{1}{\eta^t}\widetilde{\Delta}^{(1)}+\frac{1}{\eta}(P3)\big\|\Delta^{(t)}\big\|^2+\Big(\frac{1}{\eta^2}(P3)-\frac{\rho}{2\eta}\Big)\big\|\Delta^{(t-1)}\big\|^2\\&+\big(\frac{1}{\eta^3}(P3)-\frac{\rho}{2\eta^2}\big)\big\|\Delta^{(t-2)}\big\|^2+\cdots\\&+\big(\frac{1}{\eta^{t+1}}(P3)-\frac{\rho}{2\eta^t}\big)\big\|\Delta^{(0)}\big\|^2+\Big(\frac{L}{\delta\eta}\sum_{l=0}^{t}\frac{1}{\eta^l}\sum_{k=1}^{\tau}\big\|\Delta^{(t-l-k)}\big\|^2\Big)\\&+\frac{1}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\frac{\delta_1}{2m}L^2\sum_{l=0}^{t}\frac{1}{\eta^{l+1}}\sum_{j\in[m]}\sum_{k=1}^{\tau}\big\|\Delta^{(t-l-k)}\big\|^2\\&\le\frac{1}{\eta^t}\widetilde{\Delta}^{(1)}+\frac{1}{\eta}(P3)\big\|\Delta^{(t)}\big\|^2+\left(\frac{1}{\eta^2}(P3)-\frac{\rho}{2\eta}\right)\big\|\Delta^{(t-1)}\big\|^2\\&+\left(\frac{1}{\eta^3}(P3)-\frac{\rho}{2\eta^2}\right)\big\|\Delta^{(t-2)}\big\|^2+\cdots\\&+\left(\frac{1}{\eta^{t+1}}(P3)-\frac{\rho}{2\eta^t}\right)\big\|\Delta^{(0)}\big\|^2\\&+\big(\frac{L}{\delta}+\frac{\frac{\delta_1}{2}L^2\tau}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\big)\frac{\eta^\tau-1}{\eta-1}\sum_{l=1}^{t}\frac{1}{\eta^{l+1}}\big\|\Delta^{(t-l)}\big\|^2,\end{aligned}$$

where we use the fact that

$$\begin{aligned}
&\sum_{l=0}^{t}\frac{1}{\eta^{l+1}}\sum_{k=1}^{\tau}\left\|\Delta^{(t-l-k)}\right\|^2\\
&\quad=\eta^{-t-1}\sum_{l=0}^{t}\frac{\eta^t}{\eta^l}\sum_{k=1}^{\tau}\left\|\Delta^{(t-l-k)}\right\|^2\\
&\quad=\eta^{-t-1}\sum_{j=0}^{t}\eta^j\sum_{k=1}^{\tau}\left\|\Delta^{(j-k)}\right\|^2\\
&\quad\overset{(h)}{\leq}\eta^{-t-1}\sum_{j=0}^{t-1}\eta^{j+1}(1+\eta+\cdots+\eta^{\tau-1})\left\|\Delta^{(j)}\right\|^2\\
&\quad\leq\frac{\eta^\tau-1}{\eta-1}\sum_{j=0}^{t-1}\frac{1}{\eta^{t-j}}\left\|\Delta^{(j)}\right\|^2\\
&\quad\leq\frac{\eta^\tau-1}{\eta-1}\sum_{l=1}^{t}\frac{1}{\eta^{l+1}}\left\|\Delta^{(t-l)}\right\|^2.
\end{aligned}$$

The inequality (h) holds because the coefficient of $\left\|\Delta^{(j)}\right\|^2$ in the summation is less than $\eta^{j+1}(1+\eta+\cdots+\eta^{\tau-1})$.

Therefore if $\rho>0$ satisfies that

$$(P3):=\frac{\delta_1 L+\frac{\rho}{2}(1+\delta_1)}{\frac{\rho}{2}(1+\delta_1)+\delta_1}+\frac{3L}{2}-\frac{\rho}{2}+L\delta\tau<0, \tag{20}$$

and

$$\begin{aligned}
&(P3)-\frac{\rho\eta}{2}+\Big(\frac{L}{\delta}+\frac{\frac{\delta_1}{2}L^2\tau}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\Big)\frac{\eta^\tau-1}{\eta-1}\\
&=\frac{\delta_1 L+\frac{\rho}{2}(1+\delta_1)}{\frac{\rho}{2}(1+\delta_1)+\delta_1}+\frac{3L}{2}-\frac{\rho}{2}+L\delta\tau\\
&-\frac{\rho\eta}{2}+\Big(\frac{L}{\delta}+\frac{\frac{\delta_1}{2}L^2\tau}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\Big)\frac{\eta^\tau-1}{\eta-1}<0,
\end{aligned}\tag{21}$$

then we have

$$0\leq\widetilde{\Delta}^{(t+1)}\leq\frac{1}{\eta^t}\widetilde{\Delta}^{(1)}.$$

The conclusion is proved. □

A. Inexactly Solving the Subproblems

In this section we discuss the case where subproblem (3) is not solved exactly. The motivation is that in some practical applications, it may not be easy to exactly minimize the objective function. The following analysis shows that the convergence still holds true when there are small errors in solving the subproblems, thus implying the robustness of the proposed algorithm. Specifically, we assume subproblem (3) is solved with some error at iteration t; that is, there is an error ϵ^t such that

$$\begin{aligned}
\epsilon^t\in\nabla\mathbf{L}_1(\mathbf{x}^{t+1})+\tfrac{1}{m}\textstyle\sum_{j\in[m]}\nabla\mathbf{L}_j(\mathbf{x}^{t_j})-\nabla\mathbf{L}_1(\mathbf{x}^{t_1})\\
+\partial h(\mathbf{x}^{t+1})+\rho\left(\mathbf{x}^{t+1}-\mathbf{x}^t\right),
\end{aligned}\tag{22}$$

which is equivalent to

$$\begin{aligned}
\mathbf{x}^{t+1}=\mathbf{Prox}_h\Big[\mathbf{x}^{t+1}-\Big(\nabla\mathbf{L}_1(\mathbf{x}^{t+1})+\frac{1}{m}\sum_{j\in[m]}\nabla\mathbf{L}_j(\mathbf{x}^{t_j})\\
-\nabla\mathbf{L}_1(\mathbf{x}^{t_1})+\rho\left(\mathbf{x}^{t+1}-\mathbf{x}^t\right)-\epsilon^t\Big)\Big].
\end{aligned}\tag{23}$$

First, we introduce the following assumption that gives the bound of the error term.

Assumption III.4. *The error term in* (22) *satisfies*

$$\|\epsilon^t\|^2<c_1\left\|\Delta^{(t-1)}\right\|^2,\ \ \textit{for}\ \ t>0.$$

This assumption requires that the error in solving the subproblem is bounded by a constant times the progress of $\mathbf{x}^t$ in the previous iteration. Note that when $\Delta^{(t-1)}:=\mathbf{x}^t-\mathbf{x}^{t-1}=0$, it holds that $\mathbf{x}^t$ is a stationary point in the nonconvex scenario and $\mathbf{x}^t=\mathbf{x}^*$ in the strongly convex scenario. Following the proof steps in Lemma III.1, we have

$$\begin{aligned}
&\frac{\rho}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2+h(\mathbf{x}^{t+1})-\frac{\rho}{2}\|\mathbf{x}^t-\mathbf{x}^{t-1}\|^2-h(\mathbf{x}^t)\\
&\leq-\Big\langle\nabla\mathbf{L}_1(\mathbf{x}^{t+1})+\frac{1}{m}\sum_{j=1}^{m}\nabla\mathbf{L}_j(\mathbf{x}^{t_j})-\nabla\mathbf{L}_1(\mathbf{x}^{t_1})-\epsilon^t,\\
&\Delta^{(t)}\Big\rangle-\frac{\gamma(\rho)}{2}\left\|\Delta^{(t)}\right\|^2-\frac{\rho}{2}\left\|\Delta^{(t-1)}\right\|^2.
\end{aligned}$$

In the second step, for the descent of function F similar to Lemma III.2 it holds that for any $\delta>0$

$$\begin{aligned}
&\mathrm{F}(\mathbf{x}^{t+1},\mathbf{x}^t)-\mathrm{F}(\mathbf{x}^t,\mathbf{x}^{t-1})\\
&\quad\leq\Big(\tfrac{3L}{2}-\tfrac{\gamma(\rho)}{2}+L\delta\tau\Big)\left\|\Delta^{(t)}\right\|^2-\tfrac{\rho}{2}\left\|\Delta^{(t-1)}\right\|^2\\
&\qquad+\tfrac{L}{\delta}\textstyle\sum_{k=1}^{\tau}\left\|\Delta^{(t-k)}\right\|^2+\langle\epsilon^t,\Delta^{(t)}\rangle.
\end{aligned}\tag{24}$$

Therefore Lemma III.3 still holds true by Assumption III.4 and (24). Now the first conclusion of Theorem III.1 can be proved. Summing up inequality (24) over t yields

$$\begin{aligned}
&\mathrm{F}(\mathbf{x}^{T+1},\mathbf{x}^T)-\mathrm{F}(\mathbf{x}^1,\mathbf{x}^0)\\
&\leq\sum_{t=1}^{T}\left(\frac{3L}{2}-\frac{\gamma(\rho)}{2}+L\delta\tau\right)\left\|\Delta^{(t)}\right\|^2\\
&+\sum_{t=1}^{T}\left(\frac{L\tau}{\delta}-\frac{\rho}{2}\right)\left\|\Delta^{(t-1)}\right\|^2+\sum_{t=1}^{T}\langle\epsilon^t,\Delta^{(t)}\rangle\\
&\leq\sum_{t=1}^{T}\left(\frac{3L}{2}-\frac{\gamma(\rho)}{2}+L\delta\tau\right)\left\|\Delta^{(t)}\right\|^2\\
&+\sum_{t=1}^{T}\left(\frac{L\tau}{\delta}-\frac{\rho}{2}\right)\left\|\Delta^{(t-1)}\right\|^2+\frac{1}{2}\sum_{t=1}^{T}(\|\epsilon^t\|^2+\left\|\Delta^{(t)}\right\|^2)\\
&\leq\sum_{t=1}^{T}\left(\frac{3L}{2}-\frac{\gamma(\rho)}{2}+L\delta\tau+\frac{1}{2}\right)\left\|\Delta^{(t)}\right\|^2\\
&+\sum_{t=1}^{T}\left(\frac{L\tau}{\delta}-\frac{\rho}{2}+\frac{1}{2}c_1\right)\left\|\Delta^{(t-1)}\right\|^2,
\end{aligned}$$

where in the last inequality we use Assumption III.4.

Now define $\widetilde{c}:=\min\left\{\frac{\gamma(\rho)}{2}-\frac{3L}{2}-L\delta\tau-\frac{1}{2},\frac{\rho}{2}-\frac{L\tau}{\delta}-\frac{1}{2}c_1\right\}$. Assume that

$$\gamma(\rho)>3L+2L\delta\tau+1\quad\text{and}\quad\rho>\frac{2L\tau}{\delta}+c_1, \tag{25}$$

then we have $\widetilde{c}>0$. Therefore

$$\mathrm{F}(\mathbf{x}^{T+1},\mathbf{x}^T)-\mathrm{F}(\mathbf{x}^1,\mathbf{x}^0)\leq-\widetilde{c}\sum_{t=0}^{T}\left\|\Delta^{(t)}\right\|^2. \tag{26}$$

Note that by Lemma III.3 the LHS of (26) is bounded from

below. It follows that

$$\left\|\Delta^{(t)}\right\| \to 0, \qquad t \to \infty.$$

We now establish the second conclusion of Theorem III.1. From (23) we know that

$$\mathbf{x}^{t+1} = \mathbf{Prox}_h\Big[\mathbf{x}^{t+1} - \Big(\nabla \mathbf{L}_1(\mathbf{x}^{t+1}) + \frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t_j}) - \nabla \mathbf{L}_1(\mathbf{x}^{t_1}) + \rho\left(\mathbf{x}^{t+1} - \mathbf{x}^t\right) + \epsilon^t\Big)\Big],$$

where $\mathbf{Prox}_h$ is a proximal operator defined by $\mathbf{Prox}_h[z] := \underset{x}{\operatorname{argmin}}\, h(x) + \frac{1}{2}\|x - z\|^2$. This implies that

$$\begin{aligned}
&\left\|\mathbf{x}^t - \mathbf{Prox}_h\left(\mathbf{x}^t - \frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^t)\right)\right\| \\
&\le \Bigg\|\mathbf{x}^t - \mathbf{x}^{t+1} + \mathbf{x}^{t+1} \\
&- \mathbf{Prox}_h\left(\mathbf{x}^t - \frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^t)\right)\Bigg\| \\
&\le \left\|\mathbf{x}^t - \mathbf{x}^{t+1}\right\| \\
&+ \Bigg\|\mathbf{Prox}_h\Bigg[\mathbf{x}^{t+1} - \Bigg(\nabla \mathbf{L}_1(\mathbf{x}^{t+1}) + \frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t_j}) \\
&- \nabla \mathbf{L}_1(\mathbf{x}^{t_1}) + \rho\big(\mathbf{x}^{t+1} - \mathbf{x}^t\big) - \epsilon^t\Bigg)\Bigg] \\
&- \mathbf{Prox}_h\left(\mathbf{x}^t - \frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^t)\right)\Bigg\| \\
&\overset{(\widetilde{a})}{\le} \Bigg\|(1+\rho)(\mathbf{x}^{t+1} - \mathbf{x}^t) - \epsilon^t + \frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^t) \\
&- \frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t_j}) - \big(\nabla \mathbf{L}_1(\mathbf{x}^{t+1}) - \nabla \mathbf{L}_1(\mathbf{x}^{t_1})\big)\Bigg\| \\
&+ \left\|\Delta^{(t)}\right\| \\
&\le (2+\rho)\left\|\Delta^{(t)}\right\| + 2L\sum_{k=0}^{\tau}\left\|\Delta^{(t-k)}\right\| + c_1^{\frac{1}{2}}\left\|\Delta^{(t-1)}\right\| \\
&\longrightarrow 0, \qquad t \to \infty.
\end{aligned} \tag{27}$$

Note that here inequality $(\widetilde{a})$ holds because of the nonexpansiveness of the operator $\mathbf{Prox}_h$. The last inequality follows from Assumption III.1. Therefore as in the proof of Theorem III.1, the second result holds.

The rest analysis is the same as that of Theorem III.1. Specifically it holds that

$$\sum_{t=0}^{T}\left\|\widetilde{\nabla}\mathbf{L}\left(\mathbf{x}^t\right)\right\|^2 \le \frac{\widetilde{\mu}}{\widetilde{c}}\left(\mathrm{F}(\mathbf{x}^1, \mathbf{x}^0) - \mathrm{F}(\mathbf{x}^{T+1}, \mathbf{x}^T)\right),$$

where $\widetilde{\mu} := 3\left(2 + \rho + 2L\tau + c_1^{\frac{1}{2}}\right)$.

Recall that $T(\epsilon) := \min\left\{t \mid \left\|\widetilde{\nabla}\mathbf{L}(\mathbf{x}^t)\right\| \le \epsilon, t \ge 0\right\}$, thus it follows that

$$\epsilon \le \frac{C \cdot \left(\mathrm{F}(\mathbf{x}^1, \mathbf{x}^0) - \underline{\mathrm{F}}\right)}{T(\epsilon)},$$

where $C := \frac{\widetilde{\mu}}{\widetilde{c}} > 0$. Therefore we have the following corollary.

Corollary III.1. *Let Assumptions II.1 III.1, III.3(II), and III.4 hold, and suppose ρ satisfies* (25)*. Then all conclusions in Theorem III.1 hold true for the sequence generated by (EDANNI) with subproblems being solved inexactly (as quantified above).*

Results corresponding to Theorem III.2 also holds in the scenario of solving the subproblems inexactly. Similar to Lemma III.4, the optimality gap of function F can be bounded by

$$\begin{aligned}
&\frac{1}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\left(\mathrm{F}(\mathbf{x}^{t+1}, \mathbf{x}^t) - \mathrm{F}(\mathbf{x}^*, \mathbf{x}^*)\right) \\
&\le \frac{\delta_1 L}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\|\mathbf{x}^t - \mathbf{x}^{t+1}\|^2 \\
&+ \frac{\frac{\rho}{2}(1+\delta_1)}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\|\mathbf{x}^t - \mathbf{x}^{t+1}\|^2 \\
&+ \frac{1}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\frac{\delta_1}{2m}L^2\sum_{j\in[m]}\|\mathbf{x}^{t_j} - \mathbf{x}^t\|^2 \\
&+ \left(\frac{1}{2}\|\epsilon^t\|^2 + \frac{1}{2}\left\|\Delta^{(t)}\right\|^2\right)\frac{1}{\frac{\rho}{2}(1+\delta_1)+\delta_1}.
\end{aligned} \tag{28}$$

Following the steps of Lemma III.2, we have

$$\begin{aligned}
&\mathrm{F}(\mathbf{x}^{t+1}, \mathbf{x}^t) - \mathrm{F}(\mathbf{x}^t, \mathbf{x}^{t-1}) \\
&\le \left(\frac{3L}{2} - \frac{\gamma(\rho)}{2} + L\delta\tau\right)\left\|\Delta^{(t)}\right\|^2 - \frac{\rho}{2}\left\|\Delta^{(t-1)}\right\|^2 \\
&\qquad + \frac{L}{\delta}\sum_{k=1}^{\tau}\left\|\Delta^{(t-k)}\right\|^2 + \langle\epsilon^t, \Delta^{(t)}\rangle.
\end{aligned} \tag{29}$$

Combining (28) and (29) and then applying the Assumption III.4 leads to

$$\begin{aligned}
&\left(1 + \frac{1}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\right)\widetilde{\Delta}^{(t+1)} \\
&\le \widetilde{\Delta}^{(t)} - \frac{\rho - c_1}{2}\left\|\Delta^{(t-1)}\right\|^2 + \frac{L}{\delta}\sum_{k=1}^{\tau}\left\|\Delta^{(t-k)}\right\|^2 \\
&+ \left[\frac{\delta_1 L + \frac{\rho}{2}(1+\delta_1) + \frac{1}{2}}{\frac{\rho}{2}(1+\delta_1)+\delta_1} + \frac{3L}{2} - \frac{\rho-1}{2} + L\delta\tau\right]\left\|\Delta^{(t)}\right\|^2 \\
&+ \frac{1}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\frac{\delta_1}{2m}L^2\sum_{j\in[m]}\|\mathbf{x}^{t_j} - \mathbf{x}^t\|^2 \\
&+ \frac{c_1}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\left\|\Delta^{(t-1)}\right\|^2.
\end{aligned}$$

By a recursive argument similar to the proof of Theorem III.2, one can prove that if $\rho > 0$ satisfies that

$$\frac{\delta_1 L + \frac{\rho}{2}(1+\delta_1) + \frac{1}{2}}{\frac{\rho}{2}(1+\delta_1)+\delta_1} + \frac{3L}{2} - \frac{\rho-1}{2} + L\delta\tau < 0 \tag{30}$$

and

$$\frac{\delta_1 L+\frac{\rho}{2}(1+\delta_1)+\frac{1}{2}}{\frac{\rho}{2}(1+\delta_1)+\delta_1}+\frac{3L}{2}-\frac{\rho-1}{2}+L\delta\tau-\frac{(\rho-c_1)\eta}{2}$$
$$+\left(\frac{L}{\delta}+\frac{\frac{\delta_1}{2}L^2\tau}{\frac{\rho}{2}(1+\delta_1)+\delta_1}\right)\frac{\eta^\tau-1}{\eta-1}+\frac{c_1\eta}{\frac{\rho}{2}(1+\delta_1)+\delta_1}<0, \tag{31}$$

then we have

$$0\le\widetilde{\Delta}^{(t+1)}\le\frac{1}{\eta^t}\widetilde{\Delta}^{(1)}.$$

In summary we have the following corollary.

Corollary III.2. *Suppose Assumption II.1, III.1, III.2, and III.4 are satisfied. If ρ satisfies* (30) *and* (31) *for some $\delta>0$ and $\delta_1>(2L+\rho+1)/\sigma^2$, then for the sequence generated by (EDANNI) with subproblems being solved inexactly we have*

$$0\le F(\mathbf{x}^{t+1},\mathbf{x}^t)-F(\mathbf{x}^*,\mathbf{x}^*)$$
$$\le\frac{1}{\eta^t}\Big(F(\mathbf{x}^1,\mathbf{x}^0)-F(\mathbf{x}^*,\mathbf{x}^*)\Big),$$

where $\eta:=1+\frac{1}{\frac{\rho}{2}(1+\delta_1)+\delta_1}$.

IV. Experiments

Now we test our algorithm on both a convex application (LASSO) and a nonconvex application (Sparse PCA). In both settings, we compare with various advanced algorithms:

(1) Efficient Distributed Learning with the Parameter Server (Parameter Server): the state-of-the-art proximal gradient descent based framework with the parameter server proposed in [19].
(2) Asynchronous Distributed ADMM (AD-ADMM): the ADMM based asynchronous algorithm proposed in [20].
(3) Efficient Distributed Algorithm for Nonconvex-Nonsmooth Inference (EDANNI): the proposed approach in this paper.

We first compare their communication cost, that is, the total number of transmissions between the master and the workers. Then the effects of the asynchrony on the working time and the idle time are examined.

A. LASSO

In this example, to demonstrate the convergence performance of the above algorithms in terms of communication rounds, we consider the following LASSO problem

$$\underset{\mathbf{w}\in\mathbb{R}^p}{\operatorname{argmin}}\ \frac{1}{2mn}\sum_{j\in[m]}\sum_{i\in[n]}\left\|\mathbf{x}_{ji}^T\mathbf{w}-y_{ji}\right\|^2+\theta\|\mathbf{w}\|_1, \tag{32}$$

where $\theta>0$ is the coefficient of the regularizer. Note that from now on we switch the notation to let the unknown quantity be denoted by $\mathbf{w}$ instead of $\mathbf{x}$.

The data $\{\mathbf{x}_{ji}\}_{i\in[n],j\in[m]}$ is independently sampled from a multivariate Gaussian distribution with zero mean and covariance matrix Σ. For $r\in[p],t\in[p]$, the rt-th entry of covariance matrix is set to be: $|\Sigma_{rt}|=0.5^{|r-t|}$. The corresponding y_{ji} is constructed by

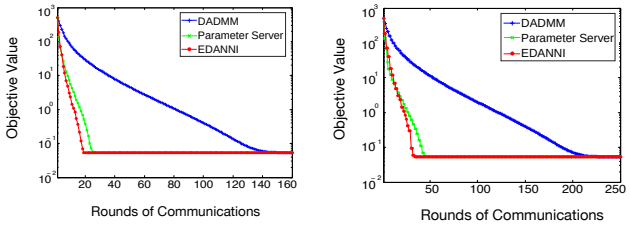
$$y_{ji}=\mathbf{x}_{ji}^T\mathbf{w}^*+\epsilon_{ji},\quad\forall j\in[m],i\in[n],$$

where noise ϵ_{ji} is a zero mean Gaussian random variable with variance 0.01. The true parameter $\mathbf{w}^*$ is s-sparse where all the entries are zero except that the first s entries are i.i.d random variables from a uniform distribution in [0,1]. The sparsity s is set to be $0.01\times p$, where p is the dimension of data $\mathbf{x}_{ji}$.

Those algorithms are compared in the setting where $m=20$, $n=500$, $p=1000$, $s=10$, and $\theta=0.01$. Even though Theorem III.1 suggests that ρ should be a larger value, we find that $\rho=0$ works well for this case. Both the synchronous scenario and the asynchronous scenario are considered. To simulate the asynchronous case, in each iteration half of the workers are assumed to be arrived with probability 0.2 and the other workers are assumed be arrived with probability 0.5. Moreover, the maximum tolerable delay τ is set be 3 and the master will update the variables once the workers j with $d_j>\tau-1$ have arrived.

TABLE I
Comparison of the communication cost for LASSO

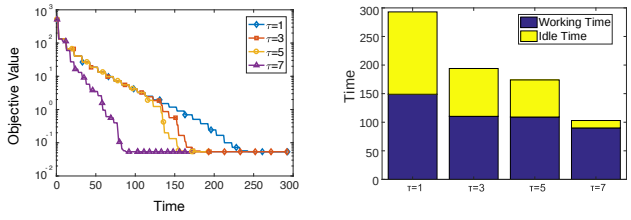
Type	Parameters (m,n,p,s,θ)	Method	Communication
Synchronous	(10,1000,500,5,0.01)	**AD-ADMM**	21.9
		PS	1.7
		EDANNI	**1**
	(20,500,500,5,0.01)	**AD-ADMM**	34.1
		PS	1.3
		EDANNI	**1**
	(20,500,1000,10,0.01)	**AD-ADMM**	8.3
		PS	1.4
		EDANNI	**1**
Asynchronous ($\tau=10$)	(10,1000,500,5,0.01)	**AD-ADMM**	20.6
		PS	1.2
		EDANNI	**1**
	(20,500,500,5,0.01)	**AD-ADMM**	35.1
		PS	1.3
		EDANNI	**1**
	(20,500,1000,10,0.01)	**AD-ADMM**	6.7
		PS	1.5
		EDANNI	**1**



(a) Synchronous, **(7.7, 1.3, 1)** (b) Asynchronous, **(6.7, 1.5, 1)**

Fig. 2. Comparison of candidate algorithms in LASSO when $m=20$, $n=500$, $p=1000$, $s=10$, and $\mathbf{X}\sim\mathcal{N}(0,\Sigma)$.

One can observe from Figure 2 that the proposed algorithm indeed converges faster than AD-ADMM and Parameter Server in terms of communication rounds, in both the synchronous scenario and the asynchronous scenario. The triple of numbers in each figure's caption indicates the communication cost needed for AD-ADMM, Parameter Server, and EDANNI to attain the minimum objective value with error less than 10^{-6}. For simplicity, we scale the communication complexity of EDANNI to 1. The results for other settings of m,n,p,s are summarized in Table I. It is shown in these results that



(a) Convergence for different τ's (b) Time table for different τ's

Fig. 3. Comparison of EDANNI in LASSO with different settings of τ when $\rho = 0$.

EDANNI is the most communication-efficient among the three algorithms.

Figure 3 shows the performance of the proposed approach when we choose different maximum tolerable delay τ. It can be observed from Figure 3 (a) that the convergence rate varies much with different values of τ. Basically, EDANNI converges faster when τ is larger, and converges relatively slower when $\tau = 0$ (i.e., the synchronous case). The results in Figure 3 (b) shows that the ratio of the computing time over the idle time increases when the delay bound τ becomes larger, therefore speeding up the convergence. Here the distributed implementation is simulated on a single machine by randomly setting the computation speed for each node from a uniform$[1, 10]$ distribution.

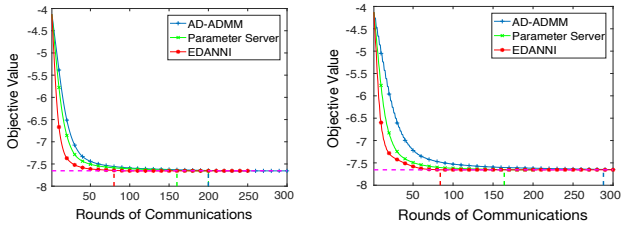
B. Sparse PCA

To verify the convergence conclusion of Theorem III.1 for nonconvex nonsmooth problems, we consider the following sparse PCA problem [4]:

$$\begin{aligned} &\underset{\mathbf{w}\in\mathbb{R}^p}{\operatorname{argmin}} -\frac{1}{mn}\sum_{j\in[m]}\sum_{i\in[n]}\mathbf{w}^T B_{ji}B_{ji}^T\mathbf{w} + \theta\|\mathbf{w}\|_1, \\ &\text{s.t.} \qquad \|\mathbf{w}\| \le 1 \end{aligned} \tag{33}$$

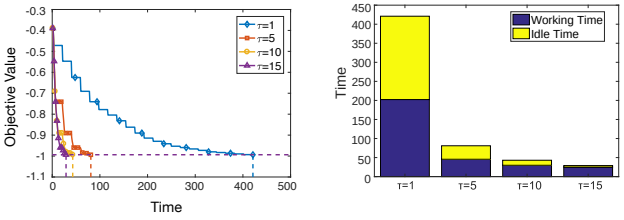
where $B_{ji} \in \mathbb{R}^{p\times q}$ is a sparse matrix, $\forall j \in [m], \quad i \in [n]$, and the regularization coefficient $\theta > 0$. Note that this is not a convex problem. In this example, we set $m = 3$, $n = 20$ $p = 500$, $q = 1000$, and $\theta = 0.1$. Each matrix $B_{ji} \in \mathbb{R}^{500\times 1000}$ is a sparse random matrix with nearly $s = 3000$ non-zero entries. The parameter ρ in (3) is set to $\rho = 2\lambda_{\max}\left(\sum_{j\in[m]} B_{ji}B_{ji}^T\right)$. The candidate algorithms are compared in both the synchronous scenario and the asynchronous scenario. One can see from Figure 4 (a) and Figure 4 (b) that the proposed approach converges much faster than AD-ADMM and Parameter Server with much less communication cost. The results for other settings of m, n, p, q, s in Table II also verify such a conclusion.

The performance of the proposed approach with different maximum tolerable delay τ is summarized in Figure 5. Here the distributed implementation is simulated in the same way as in the LASSO case. In Figure 5 we set $m = 6$, $n = 20$ $p = 100$, $q = 1000$. One can observe from Figure 5 (a) that in this example the convergence rate in terms of time is indeed affected by values of τ. The running time when $\tau = 15$ is much less than that when τ is small. Similar to the LASSO case, Figure 5 (b) shows that the ratio of the computing time in the overall running time increases closely to 1 when the delay bound τ becomes 15, therefore speeding up the convergence. Such results can be observed generally regardless of the choice of parameters m, n, p, q, s, θ.



(a) Synchronous, **(2.5, 2.0, 1)** (b) Asynchronous, **(3.2, 2.0, 1)**

Fig. 4. Comparison of candidate algorithms in sparse PCA.



(a) Convergence for different τ's (b) Time table for different τ's

Fig. 5. Comparison of EDANNI in sparse PCA with different settings of τ when $\rho = 2$.

TABLE II
COMPARISON OF THE COMMUNICATION COST FOR SPCA

Type	Parameters (m,n,p,q,s,θ)	Method	Communication
Synchronous	(20,100,50,100,50,0.1)	**AD-ADMM**	11.5
		PS	2.4
		EDANNI	**1**
	(30,200,50,100,50,0.1)	**AD-ADMM**	30.0
		PS	2.5
		EDANNI	**1**
	(3,20,500,1000,500,0.1)	**AD-ADMM**	5.2
		PS	1.7
		EDANNI	**1**
Asynchronous	(20,100,50,100,50,0.1) ($\tau = 3$)	**AD-ADMM**	12.5
		PS	4.1
		EDANNI	**1**
	(30,200,50,100,50,0.1) ($\tau = 10$)	**AD-ADMM**	31.1
		PS	7.2
		EDANNI	**1**
	(3,20,500,1000,500,0.1) ($\tau = 3$)	**AD-ADMM**	6.3
		PS	2.0
		EDANNI	**1**

V. CONCLUSION

This paper proposes a communication-efficient distributed algorithm (EDANNI) solving a general problem (2) in signal processing and machine learning under an asynchronous protocol. Theoretically, we prove the proposed algorithm converges to a stationary point in a sublinear rate, even in nonconvex nonsmooth scenarios. Moreover, unlike the previous work, linear convergence rate is established in strongly convex scenarios without any statistical assumptions of the local data. In experiments, we compare EDANNI with other state-of-the-art distributed algorithms in different applications, and the results show the superior performance of the proposed algorithm in terms of communication efficiency and the speed up caused by the asynchrony.

REFERENCES

[1] S. Meguerdichian, S. Slijepcevic, V. Karayan, and M. Potkonjak, "Localized algorithms in wireless ad-hoc networks: location discovery and sensor exposure," in *Proceedings of the 2nd ACM international symposium on Mobile ad hoc networking & computing*. ACM, 2001, pp. 106–116.

[2] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1, Springer series in statistics New York, 2001.

[3] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng, "ICA with reconstruction cost for efficient overcomplete feature learning," in *Advances in Neural Information Processing Systems*, 2011, pp. 1017–1025.

[4] P. Richtárik, M. Takáč, and S. D. Ahipaşaoğlu, "Alternating maximization: Unifying framework for 8 sparse PCA formulations and efficient parallel codes," *arXiv preprint arXiv:1212.4137*, 2012.

[5] M. Rabbat and R. Nowak, "Distributed optimization in sensor networks," in *Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks*. ACM, 2004, pp. 20–27.

[6] A. Agarwal and J. C. Duchi, "Distributed delayed stochastic optimization," in *Advances in Neural Information Processing Systems*, 2011, pp. 873–881.

[7] J. K. Bradley, A. Kyrola, D. Bickson, and C. Guestrin, "Parallel coordinate descent for l1-regularized loss minimization," *arXiv preprint arXiv:1105.5379*, 2011.

[8] J. Wang, M. Kolar, N. Srebro, and T. Zhang, "Efficient distributed learning with sparsity," in *Proceedings of the International Conference on Machine Learning*, 2017, pp. 3636–3645.

[9] M. I. Jordan, J. D. Lee, and Y. Yang, "Communication-efficient distributed statistical inference," *arXiv preprint arXiv: 1605.07689*, 2016.

[10] J. Ren, X. Li, and J. Haupt, "Communication-efficient distributed optimization for sparse learning via two-way truncation," in *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2017.

[11] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 803–812, 1986.

[12] B. Recht, C. Re, S. Wright, and F. Niu, "Hogwild: A lock-free approach to parallelizing stochastic gradient descent," in *Advances in Neural Information Processing Systems*, 2011, pp. 693–701.

[13] M. Hong, "A distributed, asynchronous and incremental algorithm for nonconvex optimization: An ADMM based approach," *arXiv preprint arXiv:1412.6058*, 2014.

[14] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[15] M. Zinkevich, M. Weimer, L. Li, and A. J. Smola, "Parallelized stochastic gradient descent," in *Advances in Neural Information Processing Systems*, 2010, pp. 2595–2603.

[16] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: Numerical methods*, vol. 23, Prentice hall Englewood Cliffs, NJ, 1989.

[17] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le, et al., "Large scale distributed deep networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1223–1231.

[18] M. Li, D. G. Andersen, J. W. Park, A. J. Smola, A. Ahmed, V. Josifovski, J. Long, E. J. Shekita, and B. Su, "Scaling distributed machine learning with the parameter server.," in *OSDI*, 2014, vol. 14, pp. 583–598.

[19] M. Li, D. G. Andersen, A. J. Smola, and K. Yu, "Communication efficient distributed machine learning with the parameter server," in *Advances in Neural Information Processing Systems*, 2014, pp. 19–27.

[20] T. Chang, M. Hong, W. Liao, and X. Wang, "Asynchronous distributed admm for large-scale optimization Part I: Algorithm and convergence analysis," *IEEE Transactions on Signal Processing*, vol. 64, no. 12, pp. 3118–3130, 2016.

[21] T. Chang, W. Liao, M. Hong, and X. Wang, "Asynchronous distributed admm for large-scale optimization Part II: Linear convergence analysis and numerical performance," *IEEE Transactions on Signal Processing*, vol. 64, no. 12, pp. 3131–3144, 2016.

[22] J. Ren and J. Haupt, "Provably communication-efficient asynchronous distributed inference for convex and nonconvex problems," in *IEEE Global Conference on Signal and Information Processing*, 2018.

[23] M. Ma, J. Ren, G. B. Giannakis, and J. Haupt, "Fast asynchronous decentralized optimization: allowing multiple masters," in *IEEE Global Conference on Signal and Information Processing*, 2018.

[24] L. Cannelli, F. Facchinei, V. Kungurtsev, and G. Scutari, "Asynchronous parallel algorithms for nonconvex big-data optimization. part I: Model and convergence," *arXiv preprint arXiv:1607.04818*, 2016.

[25] J. Liu and S. J. Wright, "Asynchronous stochastic coordinate descent: Parallelism and convergence properties," *SIAM Journal on Optimization*, vol. 25, no. 1, pp. 351–376, 2015.

[26] A. Aytekin, H. R. Feyzmahdavian, and M. Johansson, "Analysis and implementation of an asynchronous optimization algorithm for the parameter server," *arXiv preprint arXiv:1610.05507*, 2016.

[27] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao, "Optimal distributed online prediction using mini-batches," *Journal of Machine Learning Research*, vol. 13, no. Jan, pp. 165–202, 2012.

[28] Y. Zhang, M. J. Wainwright, and J. C. Duchi, "Communication-efficient algorithms for statistical optimization," in *Advances in Neural Information Processing Systems*, 2012, pp. 1502–1510.

[29] O. Shamir and N. Srebro, "Distributed stochastic optimization and learning," in *52nd Annual Allerton Conference on Communication, Control, and Computing*, 2014, pp. 850–857.

[30] Y. Arjevani and O. Shamir, "Communication complexity of distributed convex learning and optimization," in *Advances in Neural Information Processing Systems*, 2015, pp. 1756–1764.

[31] J. D. Lee, Q. Lin, T. Ma, and T. Yang, "Distributed stochastic variance reduced gradient methods and a lower bound for communication complexity," *arXiv preprint arXiv:1507.07595*, 2015.

[32] R. Mcdonald, M. Mohri, N. Silberman, D. Walker, and G. S. Mann, "Efficient large-scale distributed training of conditional maximum entropy models," in *Advances in Neural Information Processing Systems*, 2009, pp. 1231–1239.

[33] C. Huang and X. Huo, "A distributed one-step estimator," *arXiv preprint arXiv:1511.01443*, 2015.

[34] T. Yang, "Trading computation for communication: Distributed stochastic dual coordinate ascent," in *Advances in Neural Information Processing Systems*, 2013, pp. 629–637.

[35] C. Ma, V. Smith, M. Jaggi, M. I. Jordan, P. Richtárik, and M. Takáč, "Adding vs. averaging in distributed primal-dual optimization," *arXiv preprint arXiv:1502.03508*, 2015.

[36] M. Jaggi, V. Smith, M. Takác, J. Terhorst, S. Krishnan, T. Hofmann, and M. I. Jordan, "Communication-efficient distributed dual coordinate ascent," in *Advances in Neural Information Processing Systems*, 2014, pp. 3068–3076.

[37] O. Shamir, N. Srebro, and T. Zhang, "Communication-efficient distributed optimization using an approximate Newton-type method.," in *Proceedings of the International Conference on Machine Learning*, 2014, vol. 32, pp. 1000–1008.

[38] Y. Zhang and X. Lin, "Disco: Distributed optimization for self-concordant empirical loss.," in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 362–370.

[39] J. D. Lee, Y. Sun, and M. A. Saunders, "Proximal Newton-type methods for minimizing composite functions," *SIAM Journal on Optimization*, vol. 24, no. 3, pp. 1420–1443, 2014.

[40] M. Hong, Z. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 337–364, 2016.

[41] S. Ghadimi, G. Lan, and H. Zhang, "Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization," *Mathematical Programming*, vol. 155, no. 1-2, pp. 267–305, 2016.

[42] C. Cartis, N. I. Gould, and P. L. Toint, "On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems," *SIAM Journal on Optimization*, vol. 20, no. 6, pp. 2833–2852, 2010.

APPENDIX A
PROOFS OF LEMMATA

A. Proof of Lemma III.1

Using optimality of $\mathbf{x}^{t+1}$ in the update (3), we have

$$-\left[\nabla \mathbf{L}_1(\mathbf{x}^{t+1}) + \frac{1}{m}\sum_{j=1}^{m}\nabla \mathbf{L}_j(\mathbf{x}^{t_j}) - \nabla \mathbf{L}_1(\mathbf{x}^{t_1})\right] \in \partial\left[h(\mathbf{x}) + \tfrac{\rho}{2}\left\|\mathbf{x}-\mathbf{x}^t\right\|^2\right]. \tag{34}$$

Recall that in Assumption III.3 (I), we define the convex modulus of $\frac{\rho}{2}\|\mathbf{x}-\mathbf{x}^t\|^2 + h(\mathbf{x})$ by $\gamma(\rho)$. It follows that

$$\begin{aligned}
&\tfrac{\rho}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2 + h(\mathbf{x}^{t+1}) - \left(\tfrac{\rho}{2}\|\mathbf{x}^t-\mathbf{x}^t\|^2 + h(\mathbf{x}^t)\right)\\
&\le \left\langle -\Big[\nabla \mathbf{L}_1(\mathbf{x}^{t+1}) + \frac{1}{m}\sum_{j=1}^{m}\nabla \mathbf{L}_j(\mathbf{x}^{t_j}) - \nabla \mathbf{L}_1(\mathbf{x}^{t_1})\Big],\right.\\
&\quad \left.\mathbf{x}^{t+1}-\mathbf{x}^t\right\rangle - \frac{\gamma(\rho)}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2\\
&= -\left\langle \nabla \mathbf{L}_1(\mathbf{x}^{t+1}) + \frac{1}{m}\sum_{j=1}^{m}\nabla \mathbf{L}_j(\mathbf{x}^{t_j}) - \nabla \mathbf{L}_1(\mathbf{x}^{t_1}),\right.\\
&\quad \left.\Delta^{(t)}\right\rangle - \frac{\gamma(\rho)}{2}\left\|\Delta^{(t)}\right\|^2,
\end{aligned}$$

where we define $\Delta^{(t)} := \mathbf{x}^{t+1}-\mathbf{x}^t$. Therefore

$$\begin{aligned}
&\frac{\rho}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2 + h(\mathbf{x}^{t+1}) - \frac{\rho}{2}\|\mathbf{x}^t-\mathbf{x}^{t-1}\|^2 - h(\mathbf{x}^t)\\
&= \frac{\rho}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2 + h(\mathbf{x}^{t+1}) - \frac{\rho}{2}\|\mathbf{x}^t-\mathbf{x}^t\|^2 - h(\mathbf{x}^t)\\
&+ \frac{\rho}{2}\|\mathbf{x}^t-\mathbf{x}^t\|^2 + h(\mathbf{x}^t) - \frac{\rho}{2}\|\mathbf{x}^t-\mathbf{x}^{t-1}\|^2 - h(\mathbf{x}^t)\\
&\le -\left\langle \nabla \mathbf{L}_1(\mathbf{x}^{t+1}) + \frac{1}{m}\sum_{j=1}^{m}\nabla \mathbf{L}_j(\mathbf{x}^{t_j}) - \nabla \mathbf{L}_1(\mathbf{x}^{t_1}),\right.\\
&\left.\Delta^{(t)}\right\rangle - \frac{\gamma(\rho)}{2}\left\|\Delta^{(t)}\right\|^2 - \frac{\rho}{2}\left\|\Delta^{(t-1)}\right\|^2,
\end{aligned} \tag{35}$$

proving Lemma III.1.

B. Proof of Lemma III.2

It follows from Assumption III.1 that $\nabla \mathbf{L}_j(\mathbf{x})$ is Lipschitz continuous with constant L. Therefore we have

$$\begin{aligned}
&\frac{1}{m}\sum_{j\in[m]}\mathbf{L}_j(\mathbf{x}^{t+1}) - \frac{1}{m}\sum_{j\in[m]}\mathbf{L}_j(\mathbf{x}^t)\\
&\le \left\langle \mathbf{x}^{t+1}-\mathbf{x}^t, \frac{1}{m}\sum_{j\in[m]}\nabla \mathbf{L}_j(\mathbf{x}^t)\right\rangle + \frac{L}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2\\
&= \left\langle \Delta^{(t)}, \frac{1}{m}\sum_{j\in[m]}\nabla \mathbf{L}_j(\mathbf{x}^t)\right\rangle + \frac{L}{2}\left\|\Delta^{(t)}\right\|^2.
\end{aligned} \tag{36}$$

By the above definition of fucntion F, combining (35) and (36) results in

$$\begin{aligned}
&\mathrm{F}(\mathbf{x}^{t+1},\mathbf{x}^t) - \mathrm{F}(\mathbf{x}^t,\mathbf{x}^{t-1})\\
&\overset{(b)}{\le} -\left\langle \nabla \mathbf{L}_1(\mathbf{x}^{t+1}) + \frac{1}{m}\sum_{j\in[m]}\nabla \mathbf{L}_j(\mathbf{x}^{t_j}) - \nabla \mathbf{L}_1(\mathbf{x}^{t_1}),\right.\\
&\quad \left.\Delta^{(t)}\right\rangle - \frac{\gamma(\rho)}{2}\left\|\Delta^{(t)}\right\|^2 - \frac{\rho}{2}\left\|\Delta^{(t-1)}\right\|^2\\
&+ \left\langle \Delta^{(t)}, \frac{1}{m}\sum_{j\in[m]}\nabla \mathbf{L}_j(\mathbf{x}^t)\right\rangle + \frac{L}{2}\left\|\Delta^{(t)}\right\|^2\\
&= -\left\langle \nabla \mathbf{L}_1(\mathbf{x}^{t+1}) + \frac{1}{m}\sum_{j\in[m]}\nabla \mathbf{L}_j(\mathbf{x}^t) - \nabla \mathbf{L}_1(\mathbf{x}^t),\right.\\
&\quad \left.\Delta^{(t)}\right\rangle - \frac{\gamma(\rho)}{2}\left\|\Delta^{(t)}\right\|^2 - \frac{\rho}{2}\left\|\Delta^{(t-1)}\right\|^2\\
&+ \left\langle \Delta^{(t)}, \frac{1}{m}\sum_{j\in[m]}\nabla \mathbf{L}_j(\mathbf{x}^t)\right\rangle + \frac{L}{2}\left\|\Delta^{(t)}\right\|^2\\
&+ \left\langle \frac{1}{m}\sum_{j\in[m]}\nabla \mathbf{L}_j(\mathbf{x}^t) - \frac{1}{m}\sum_{j\in[m]}\nabla \mathbf{L}_j(\mathbf{x}^{t_j}), \Delta^{(t)}\right\rangle\\
&+ \left\langle \nabla \mathbf{L}_1(\mathbf{x}^{t_1}) - \nabla \mathbf{L}_1(\mathbf{x}^t), \Delta^{(t)}\right\rangle\\
&\le \left(\frac{3L}{2} - \frac{\gamma(\rho)}{2}\right)\left\|\Delta^{(t)}\right\|^2 - \frac{\rho}{2}\left\|\Delta^{(t-1)}\right\|^2\\
&+ \underbrace{\left\langle \frac{1}{m}\sum_{j\in[m]}\nabla \mathbf{L}_j(\mathbf{x}^t) - \frac{1}{m}\sum_{j\in[m]}\nabla \mathbf{L}_j(\mathbf{x}^{t_j}), \Delta^{(t)}\right\rangle}_{(P1)}\\
&\underbrace{+ \left\langle \nabla \mathbf{L}_1(\mathbf{x}^{t_1}) - \nabla \mathbf{L}_1(\mathbf{x}^t), \Delta^{(t)}\right\rangle}_{(P1)},
\end{aligned} \tag{37}$$

where inequality (b) is due to Lemma III.1 and Assumption III.1.

Note that

$$\nabla \mathbf{L}_1(\mathbf{x}^{t_1}) - \nabla \mathbf{L}_1(\mathbf{x}^t) = \sum_{k=1}^{t-t_1}\left(\nabla \mathbf{L}_1(\mathbf{x}^{t-k}) - \nabla \mathbf{L}_1(\mathbf{x}^{t-k+1})\right),$$

which implies

$$\begin{aligned}
&\|\nabla \mathbf{L}_1(\mathbf{x}^{t_1}) - \nabla \mathbf{L}_1(\mathbf{x}^t)\|\\
&\le \sum_{k=1}^{t-t_1}\left\|\nabla \mathbf{L}_1(\mathbf{x}^{t-k}) - \nabla \mathbf{L}_1(\mathbf{x}^{t-k+1})\right\|\\
&\le \sum_{k=1}^{t-t_1} L\left\|\mathbf{x}^{t-k} - \mathbf{x}^{t-k+1}\right\|\\
&\le \sum_{k=1}^{\tau} L\left\|\mathbf{x}^{t-k} - \mathbf{x}^{t-k+1}\right\|\\
&= \sum_{k=1}^{\tau} L\left\|\Delta^{(t-k)}\right\|.
\end{aligned}$$

Similarly, we can see

$$\left\|\tfrac{1}{m}\sum_{j\in[m]}\nabla\mathbf{L}_j(\mathbf{x}^t)-\tfrac{1}{m}\sum_{j\in[m]}\nabla\mathbf{L}_j(\mathbf{x}^{t_j})\right\| \leq \sum_{k=1}^{\tau} L\|\Delta^{(t-k)}\|.$$

These two inequalities result in

$$\begin{aligned}(P1) &\leq 2L\sum_{k=1}^{\tau}\|\Delta^{(t-k)}\|\|\Delta^{(t)}\| \\ &\leq L\sum_{k=1}^{\tau}\left(\frac{1}{\delta}\|\Delta^{(t-k)}\|^2+\delta\|\Delta^{(t)}\|^2\right) \\ &\leq \frac{L}{\delta}\sum_{k=1}^{\tau}\|\Delta^{(t-k)}\|^2+L\delta\tau\|\Delta^{(t)}\|^2, \end{aligned} \tag{38}$$

where in the second inequality we apply the fact that

$$a\cdot b\leq\frac{1}{2}\left(\frac{1}{\delta}a^2+\delta b^2\right)$$

for any $a,b,\delta>0$. By inserting (38) into (37) we have

$$\begin{aligned}&\mathrm{F}(\mathbf{x}^{t+1},\mathbf{x}^t)-\mathrm{F}(\mathbf{x}^t,\mathbf{x}^{t-1}) \\ &\leq\left(\frac{3L}{2}-\frac{\gamma(\rho)}{2}+L\delta\tau\right)\|\Delta^{(t)}\|^2-\frac{\rho}{2}\|\Delta^{(t-1)}\|^2 \\ &\quad+\frac{L}{\delta}\sum_{k=1}^{\tau}\|\Delta^{(t-k)}\|^2,\end{aligned}$$

proving the conclusion of Lemma III.2.

C. Proof of Lemma III.3

First of all, summing the above inequality (8) of Lemma III.2 over t yields

$$\begin{aligned}&\mathrm{F}(\mathbf{x}^{T+1},\mathbf{x}^T)-\mathrm{F}(\mathbf{x}^1,\mathbf{x}^0) \\ &\leq\sum_{t=0}^{T}\left(\frac{3L}{2}-\frac{\gamma(\rho)}{2}+L\delta\tau\right)\|\Delta^{(t)}\|^2 \\ &\quad+\sum_{t=0}^{T}\left(\frac{L\tau}{\delta}-\frac{\rho}{2}\right)\|\Delta^{(t-1)}\|^2.\end{aligned}$$

If ρ satisfies Assumption III.3, then it holds that

$$\mathrm{F}(\mathbf{x}^{T+1},\mathbf{x}^T)-\mathrm{F}(\mathbf{x}^1,\mathbf{x}^0)<0.$$

By taking $\mathbf{x}^T$ as the initial point, similarly we have

$$\mathrm{F}(\mathbf{x}^{2T+1},\mathbf{x}^{2T})-\mathrm{F}(\mathbf{x}^{T+1},\mathbf{x}^T)<0.$$

Continuing this process we get a decreasing subsequence $\left\{\mathrm{F}(\mathbf{x}^{kT+1},\mathbf{x}^{kT})\right\}_{k=0,1,\cdots}$. Therefore there exists a constant $\bar{F}_0$ such that

$$\mathrm{F}(\mathbf{x}^{kT+1},\mathbf{x}^{kT})\leq\bar{F}_0. \tag{39}$$

When starting with $\mathbf{x}^1,\cdots,\mathbf{x}^{T-1}$, with similar analysis we can prove that there exists constants $\bar{F}_1,\cdots,\bar{F}_{T-1}$ such that

$$\mathrm{F}(\mathbf{x}^{kT+l+1},\mathbf{x}^{kT+l})\leq\bar{F}_l, \tag{40}$$

for $l=1,2,\cdots,T-1$. Define $\bar{\mathrm{F}}:=\max\left\{\bar{F}_0,\cdots,\bar{F}_{T-1}\right\}$, then

$$\mathrm{F}(\mathbf{x}^{t+1},\mathbf{x}^t)<\bar{\mathrm{F}}<+\infty,\quad\forall t\in\mathbb{N}.$$

On the other hand, let $\underline{\mathrm{F}}:=\underline{L}$, then by the definition of $\mathrm{F}(\mathbf{x}^{t+1},\mathbf{x}^t)$ and Assumption III.3, we have

$$\begin{aligned}&\mathrm{F}(\mathbf{x}^{t+1},\mathbf{x}^t) \\ &=\frac{1}{m}\sum_{j\in[m]}\mathbf{L}_j(\mathbf{x}^{t+1})+\frac{\rho}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2+h(\mathbf{x}^{t+1}) \\ &\geq\frac{1}{m}\sum_{j\in[m]}\mathbf{L}_j(\mathbf{x}^{t+1})+h(\mathbf{x}^{t+1}) \\ &=\mathbf{L}(\mathbf{x}^{t+1})>\underline{L}=\underline{\mathrm{F}}>-\infty,\end{aligned}$$

for any $t\in\mathbb{N}$. Therefore the boundedness of function F in Lemma III.3 is proved.

D. Proof of Lemma III.4

To prove the convergence rate, we first need to bound $\left(\mathrm{F}(\mathbf{x}^{t+1},\mathbf{x}^t)-\mathrm{F}(\mathbf{x}^*,\mathbf{x}^*)\right)$, where $\mathrm{F}(\mathbf{x},\mathbf{x}^t)=\frac{1}{m}\sum_{j\in[m]}\mathbf{L}_j(\mathbf{x})+\frac{\rho}{2}\|\mathbf{x}-\mathbf{x}^t\|^2+h(\mathbf{x})\geq\mathrm{F}(\mathbf{x}^*,\mathbf{x}^*)=\frac{1}{m}\sum_{j\in[m]}\mathbf{L}_j(\mathbf{x}^*)+h(\mathbf{x}^*)$.

By the optimality of $\mathbf{x}^{t+1}$ in the update (3), we have

$$\begin{aligned}&\Big(\nabla\mathbf{L}_1(\mathbf{x}^{t+1})+\frac{1}{m}\sum_{j\in[m]}\nabla\mathbf{L}_j(\mathbf{x}^{t_j})-\nabla\mathbf{L}_1(\mathbf{x}^t) \\ &+\partial h(\mathbf{x}^{t+1})+\rho(\mathbf{x}^{t+1}-\mathbf{x}^t)\Big)^\top(\mathbf{x}^{t+1}-\mathbf{x})\leq 0,\end{aligned} \tag{41}$$

for all $\mathbf{x}\in\mathbb{R}^p$. Letting $\mathbf{x}=\mathbf{x}^*$ implies

$$\begin{aligned}&\Big(\nabla\mathbf{L}_1(\mathbf{x}^{t+1})+\frac{1}{m}\sum_{j\in[m]}\nabla\mathbf{L}_j(\mathbf{x}^{t_j})-\nabla\mathbf{L}_1(\mathbf{x}^t) \\ &+\partial h(\mathbf{x}^{t+1})+\rho(\mathbf{x}^{t+1}-\mathbf{x}^t)\Big)^\top(\mathbf{x}^{t+1}-\mathbf{x}^*)\leq 0.\end{aligned} \tag{42}$$

By the strong convexity of $\mathbf{L}_j$ one has

$$\mathbf{L}_j(\mathbf{y})\geq\mathbf{L}_j(\mathbf{x})+(\nabla\mathbf{L}_j(\mathbf{x}))^\top(\mathbf{y}-\mathbf{x})+\frac{\sigma^2}{2}\|\mathbf{y}-\mathbf{x}\|^2,\quad\forall\mathbf{x},\ \mathbf{y}\in\mathbb{R}^p. \tag{43}$$

Setting $\mathbf{y}=\mathbf{x}^*$, $\mathbf{x}=\mathbf{x}^{t+1}$ in (43) we have

$$\begin{aligned}\mathbf{L}_j(\mathbf{x}^*)&\geq\mathbf{L}_j(\mathbf{x}^{t+1})+\left(\nabla\mathbf{L}_j(\mathbf{x}^{t+1})\right)^\top(\mathbf{x}^*-\mathbf{x}^{t+1}) \\ &\quad+\frac{\sigma^2}{2}\|\mathbf{x}^*-\mathbf{x}^{t+1}\|^2,\end{aligned}$$

which further implies that

$$\begin{aligned}&\left(\nabla\mathbf{L}_j(\mathbf{x}^{t+1})\right)^\top(\mathbf{x}^{t+1}-\mathbf{x}^*) \\ &\quad\geq\mathbf{L}_j(\mathbf{x}^{t+1})-\mathbf{L}_j(\mathbf{x}^*)+\frac{\sigma^2}{2}\|\mathbf{x}^*-\mathbf{x}^{t+1}\|^2.\end{aligned} \tag{44}$$

Note that (42) implies that

$$\Big(\nabla \mathbf{L}_1(\mathbf{x}^{t+1}) - \nabla \mathbf{L}_1(\mathbf{x}^t) + \Big(\frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t_j}) - \frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t+1})\Big) + \frac{1}{m}\sum_{j} \nabla \mathbf{L}_j(\mathbf{x}^{t+1}) + \partial h(\mathbf{x}^{t+1}) + \rho(\mathbf{x}^{t+1} - \mathbf{x}^t)\Big)^\top (\mathbf{x}^{t+1} - \mathbf{x}^*) \leq 0. \tag{45}$$

Summing (44) over $j \in [m]$ gives that

$$\frac{1}{m}\sum_{j\in[m]} \big(\nabla \mathbf{L}_j(\mathbf{x}^{t+1})\big)^\top (\mathbf{x}^{t+1} - \mathbf{x}^*) \geq \frac{1}{m}\sum_{j\in[m]} \mathbf{L}_j(\mathbf{x}^{t+1}) - \frac{1}{m}\sum_{j\in[m]} \mathbf{L}_j(\mathbf{x}^*) + \frac{\sigma^2}{2}\|\mathbf{x}^* - \mathbf{x}^{t+1}\|^2. \tag{46}$$

Putting (46) into (45), we have

$$\begin{aligned}&\Big(\nabla \mathbf{L}_1(\mathbf{x}^{t+1}) - \nabla \mathbf{L}_1(\mathbf{x}^t) + \Big(\frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t_j}) - \frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t+1})\Big)\Big)^\top (\mathbf{x}^{t+1} - \mathbf{x}^*) \\ &+ \Big(\frac{1}{m}\sum_{j\in[m]} \mathbf{L}_j(\mathbf{x}^{t+1}) - \frac{1}{m}\sum_{j\in[m]} \mathbf{L}_j(\mathbf{x}^*)\Big) + \frac{\sigma^2}{2}\|\mathbf{x}^* - \mathbf{x}^{t+1}\|^2 + \partial h(\mathbf{x}^{t+1})(\mathbf{x}^{t+1} - \mathbf{x}^*) \\ &+ \rho(\mathbf{x}^{t+1} - \mathbf{x}^t)^\top (\mathbf{x}^{t+1} - \mathbf{x}^*) \leq 0.\end{aligned} \tag{47}$$

Since $h(\mathbf{x})$ is convex, we have

$$h(\mathbf{x}^{t+1}) - h(\mathbf{x}^*) \leq \partial h(\mathbf{x}^{t+1})(\mathbf{x}^{t+1} - \mathbf{x}^*).$$

Putting it into (47), we have

$$\begin{aligned}&\Big(\nabla \mathbf{L}_1(\mathbf{x}^{t+1}) - \nabla \mathbf{L}_1(\mathbf{x}^t) + \Big(\frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t_j}) - \frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t+1})\Big)\Big)^\top (\mathbf{x}^{t+1} - \mathbf{x}^*) \\ &+ \Big(\frac{1}{m}\sum_{j\in[m]} \mathbf{L}_j(\mathbf{x}^{t+1}) - \frac{1}{m}\sum_{j\in[m]} \mathbf{L}_j(\mathbf{x}^*)\Big) + \frac{\sigma^2}{2}\|\mathbf{x}^* - \mathbf{x}^{t+1}\|^2 + h(\mathbf{x}^{t+1}) - h(\mathbf{x}^*) \\ &+ \rho(\mathbf{x}^{t+1} - \mathbf{x}^t)^\top (\mathbf{x}^{t+1} - \mathbf{x}^*) \leq 0.\end{aligned} \tag{48}$$

Note that

$$\rho(\mathbf{x}^{t+1} - \mathbf{x}^t)^\top (\mathbf{x}^{t+1} - \mathbf{x}^*) = \frac{\rho}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 + \frac{\rho}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 - \frac{\rho}{2}\|\mathbf{x}^t - \mathbf{x}^*\|^2. \tag{49}$$

Then putting (49) into (48), one obtains

$$\begin{aligned}&\Big(\nabla \mathbf{L}_1(\mathbf{x}^{t+1}) - \nabla \mathbf{L}_1(\mathbf{x}^t) + \Big(\frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t_j}) - \frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t+1})\Big)\Big)^\top (\mathbf{x}^{t+1} - \mathbf{x}^*) \\ &+ \Big(\frac{1}{m}\sum_{j\in[m]} \mathbf{L}_j(\mathbf{x}^{t+1}) - \frac{1}{m}\sum_{j\in[m]} \mathbf{L}_j(\mathbf{x}^*)\Big) + \frac{\sigma^2}{2}\|\mathbf{x}^* - \mathbf{x}^{t+1}\|^2 + h(\mathbf{x}^{t+1}) - h(\mathbf{x}^*) \\ &+ \frac{\rho}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 + \frac{\rho}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 - \frac{\rho}{2}\|\mathbf{x}^t - \mathbf{x}^*\|^2 \leq 0,\end{aligned}$$

which is equivalent to

$$\begin{aligned}\mathrm{F}(\mathbf{x}^{t+1}, \mathbf{x}^t) - \mathrm{F}(\mathbf{x}^*, \mathbf{x}^*) \leq &-\frac{\sigma^2}{2}\|\mathbf{x}^* - \mathbf{x}^{t+1}\|^2 \\ &- \Big(\nabla \mathbf{L}_1(\mathbf{x}^{t+1}) - \nabla \mathbf{L}_1(\mathbf{x}^t) + \Big(\frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t_j}) - \frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t+1})\Big)\Big)^\top (\mathbf{x}^{t+1} - \mathbf{x}^*) \\ &- \frac{\rho}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 + \frac{\rho}{2}\|\mathbf{x}^t - \mathbf{x}^*\|^2.\end{aligned} \tag{50}$$

Now, note that

$$\begin{aligned}&\Big(\nabla \mathbf{L}_1(\mathbf{x}^{t+1}) - \nabla \mathbf{L}_1(\mathbf{x}^t) + \Big(\frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t_j}) - \frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t+1})\Big)\Big)^\top (\mathbf{x}^{t+1} - \mathbf{x}^*) \\ &= \Big(\nabla \mathbf{L}_1(\mathbf{x}^{t+1}) - \nabla \mathbf{L}_1(\mathbf{x}^t) + \frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^t) - \frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t+1})\Big)^\top (\mathbf{x}^{t+1} - \mathbf{x}^*) \\ &\underbrace{+\Big(\frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t_j}) - \frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^t)\Big)^\top (\mathbf{x}^{t+1} - \mathbf{x}^*)}_{(P5)}.\end{aligned} \tag{51}$$

By the Mean Value Theorem one has

$$\begin{aligned}&\Big(\nabla \mathbf{L}_1(\mathbf{x}^{t+1}) - \nabla \mathbf{L}_1(\mathbf{x}^t) + \Big(\frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^t) - \frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t+1})\Big)\Big)^\top (\mathbf{x}^{t+1} - \mathbf{x}^*) \\ &= (\mathbf{x}^{t+1} - \mathbf{x}^t)^\top \Big[\nabla^2 \mathbf{L}_1(\xi) - \frac{1}{m}\sum_{j\in[m]} \nabla^2 \mathbf{L}_j(\xi)\Big] \cdot (\mathbf{x}^{t+1} - \mathbf{x}^*) \\ &= \frac{1}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_\Sigma^2 - \frac{1}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^{t+1}\|_\Sigma^2 \\ &+ \frac{1}{2}\|\mathbf{x}^t - \mathbf{x}^{t+1}\|_\Sigma^2 - \frac{1}{2}\|\mathbf{x}^t - \mathbf{x}^*\|_\Sigma^2,\end{aligned}$$

where $\Sigma := \nabla^2 \mathbf{L}_1(\xi) - \frac{1}{m}\sum_{j\in[m]} \nabla^2 \mathbf{L}_j(\xi)$. It follows that

$$\begin{aligned}
&\Big(\nabla \mathbf{L}_1(\mathbf{x}^{t+1}) - \nabla \mathbf{L}_1(\mathbf{x}^t) + \big(\frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^t) \\
&- \frac{1}{m}\sum_{j\in[m]} \nabla \mathbf{L}_j(\mathbf{x}^{t+1})\big)\Big)^\top (\mathbf{x}^{t+1} - \mathbf{x}^*) \\
&= \frac{1}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_\Sigma^2 + \frac{1}{2}\|\mathbf{x}^t - \mathbf{x}^{t+1}\|_\Sigma^2 \\
&- \frac{1}{2}\|\mathbf{x}^t - \mathbf{x}^{t+1} + \mathbf{x}^{t+1} - \mathbf{x}^*\|_\Sigma^2 \\
&\geq \frac{1}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_\Sigma^2 + \frac{1}{2}\|\mathbf{x}^t - \mathbf{x}^{t+1]}\|_\Sigma^2 \\
&- \frac{1}{2}(1+\delta_1)\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_\Sigma^2 - \frac{1}{2}(1+1/\delta_1)\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_\Sigma^2 \\
&\geq -\frac{1}{2}\delta_1\|\mathbf{x}^t - \mathbf{x}^{t+1}\|_\Sigma^2 - \frac{1}{2\delta_1}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_\Sigma^2. \qquad (52)
\end{aligned}$$

Note that

$$\begin{aligned}
&\frac{\rho}{2}\|\mathbf{x}^t - \mathbf{x}^*\|^2 = \frac{\rho}{2}\|\mathbf{x}^t - \mathbf{x}^{t+1} + \mathbf{x}^{t+1} - \mathbf{x}^*\|^2 \\
&\leq \frac{\rho}{2}(1+\delta_1)\|\mathbf{x}^t - \mathbf{x}^{t+1}\|^2 + \frac{\rho}{2}(1+1/\delta_1)\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2,
\end{aligned}$$

which implies that

$$\begin{aligned}
&-\frac{\rho}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 + \frac{\rho}{2}\|\mathbf{x}^t - \mathbf{x}^*\|^2 \\
&\leq \frac{\rho}{2}(1+\delta_1)\|\mathbf{x}^t - \mathbf{x}^{t+1}\|^2 + \frac{\rho}{2}\frac{1}{\delta_1}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2. \qquad (53)
\end{aligned}$$

Putting (51), (52), and (53) into (50), we can bound the optimality gap of function F by

$$\begin{aligned}
&\mathrm{F}(\mathbf{x}^{t+1}, \mathbf{x}^t) - \mathrm{F}(\mathbf{x}^*, \mathbf{x}^*) \leq -\frac{\sigma^2}{2}\|\mathbf{x}^* - \mathbf{x}^{t+1}\|^2 \\
&+ \frac{1}{2}\delta_1\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_\Sigma^2 + \frac{1}{2\delta_1}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_\Sigma^2 \\
&- \frac{\rho}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 + \frac{\rho}{2}\|\mathbf{x}^t - \mathbf{x}^*\|^2 - (P5) \\
&= -\frac{\sigma^2}{2}\|\mathbf{x}^* - \mathbf{x}^{t+1}\|^2 + \frac{1}{2\delta_1}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_\Sigma^2 \\
&+ \frac{\rho}{2\delta_1}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 + \frac{1}{2}\delta_1\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_\Sigma^2 \\
&+ \frac{\rho}{2}(1+\delta_1)\|\mathbf{x}^t - \mathbf{x}^{t+1}\|^2 - (P5). \qquad (54)
\end{aligned}$$

Now we bound (P5) on the RHS of (54). For some $\delta_1 > 0$ one has

$$\begin{aligned}
&(P5) \\
&\geq -\frac{\delta_1}{2m}\sum_{j\in[m]} \|\nabla \mathbf{L}_j(\mathbf{x}^{t_j}) - \nabla \mathbf{L}_j(\mathbf{x}^t)\|^2 - \frac{1}{2\delta_1}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 \\
&\geq -\frac{\delta_1}{2m}L^2\sum_{j\in[m]} \|\mathbf{x}^{t_j} - \mathbf{x}^t\|^2 - \frac{1}{2\delta_1}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2.
\end{aligned}$$

Therefore we have

$$\begin{aligned}
&\mathrm{F}(\mathbf{x}^{t+1}, \mathbf{x}^t) - \mathrm{F}(\mathbf{x}^*, \mathbf{x}^*) \\
&\leq -\frac{\sigma^2}{2}\|\mathbf{x}^* - \mathbf{x}^{t+1}\|^2 + \frac{1}{2\delta_1}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_\Sigma^2 \\
&+ \frac{\rho}{2\delta_1}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 + \frac{1}{2}\delta_1\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_\Sigma^2 \\
&+ \frac{\rho}{2}(1+\delta_1)\|\mathbf{x}^t - \mathbf{x}^{t+1}\|^2 \\
&+ \frac{\delta_1}{2m}L^2\sum_{j\in[m]} \|\mathbf{x}^{t_j} - \mathbf{x}^t\|^2 + \frac{1}{2\delta_1}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2. \qquad (55)
\end{aligned}$$

Let $\frac{\sigma^2}{2} > \frac{2L}{2\delta_1} + \frac{\rho}{2\delta_1} + \frac{1}{2\delta_1}$, i.e., $\sigma^2 > \frac{2L}{\delta_1} + \frac{\rho}{\delta_1} + \frac{1}{\delta_1}$, then

$$\begin{aligned}
&\mathrm{F}(\mathbf{x}^{t+1}, \mathbf{x}^t) - \mathrm{F}(\mathbf{x}^*, \mathbf{x}^*) \\
&\leq \frac{1}{2}\delta_1\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_\Sigma^2 + \frac{\rho}{2}(1+\delta_1)\|\mathbf{x}^t - \mathbf{x}^{t+1}\|^2 \\
&+ \frac{\delta_1}{2m}L^2\sum_{j\in[m]} \|\mathbf{x}^{t_j} - \mathbf{x}^t\|^2 \\
&\leq \delta_1 L\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 + \frac{\rho}{2}(1+\delta_1)\|\mathbf{x}^t - \mathbf{x}^{t+1}\|^2 \\
&+ \frac{\delta_1}{2m}L^2\sum_{j\in[m]} \|\mathbf{x}^{t_j} - \mathbf{x}^t\|^2, \qquad (56)
\end{aligned}$$

from which we have

$$\begin{aligned}
&\frac{1}{\frac{\rho}{2}(1+\delta_1) + \delta_1}\big(\mathrm{F}(\mathbf{x}^{t+1}, \mathbf{x}^t) - \mathrm{F}(\mathbf{x}^*, \mathbf{x}^*)\big) \\
&\leq \frac{\delta_1 L}{\frac{\rho}{2}(1+\delta_1) + \delta_1}\|\mathbf{x}^t - \mathbf{x}^{t+1}\|^2 \\
&+ \frac{\frac{\rho}{2}(1+\delta_1)}{\frac{\rho}{2}(1+\delta_1) + \delta_1}\|\mathbf{x}^t - \mathbf{x}^{t+1}\|^2 \\
&+ \frac{1}{\frac{\rho}{2}(1+\delta_1) + \delta_1}\frac{\delta_1}{2m}L^2\sum_{j\in[m]} \|\mathbf{x}^{t_j} - \mathbf{x}^t\|^2,
\end{aligned}$$

therefore proving Lemma III.4.