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

$$x^* = \arg\min_{x \in \mathbb{R}^p} L(x) := \frac{1}{N} \sum_{k=1}^{N} l_k(x) + h(x),$$

where each $l_k(x)$ is a loss function associated with the $k$-th data sample, and is assumed smooth but possibly nonconvex with Lipschitz continuous gradient and $h(x)$ 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, ..., $m$) 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:

$$x^* = \arg\min_{x \in \mathbb{R}^p} L(x) := \frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} l_{ji}(x) + h(x),$$

where $l_{ji}(x)$ 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
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 for general nonconvex nonsmooth problems. In a similar spirit, Wang et al. [8], Jordan et al. [9], and Ren et al. [10] developed communication-efficient algorithms for sparse learning problems. That was followed up by many works including [17], [18], and [26] studied 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 way, Wang et al. [8], Jordan et al. [9], and Ren et al. [10] developed communication-efficient algorithms for sparse learning with $\ell_1$ regularizations.

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.

II. ALGORITHM

In this section, we describe our approach to computing the minimizer $x^*$ of (2). Recall that we have $m$ machines. Let us denote $t \geq 0$ as the iteration number, then $A_t \subseteq [m] := \{1, 2, \ldots, m\}$ 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 $i \in A_t$. At iteration $t$, the master machine solves a subproblem to obtain an updated estimate, and communicates this to the worker machines in the subset $A_t$. 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

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

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 $x^{t+1}$

$$x^{t+1} = \arg\min_{x} L_1(x) + h(x) + \frac{\rho}{2} \|x - x^t\|^2 + \frac{1}{m} \sum_{j \in [m]} \nabla L_j(x^{t+1}) - \nabla L_1(x^{t+1}), x - x^{t}.$$ (3)

This $x^{t+1}$ is communicated to the worker machines that are free, where it is used to compute their local gradient $\nabla L_j(x^{t+1})$. 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 $A_t$) 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 $A_t$ should satisfy:

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

To satisfy Assumption II.1, $A_t$ should contain at least the indices of the worker machines that have been inactive for longer than $\tau$ 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 $t - \tau \leq t_j \leq t$, $\forall j \in [m]$.

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 $\nabla L_j(x^{t_j})$ used by the master must be at most $\tau$ iterations old. To guarantee the bounded delay, at every iteration the master needs to wait for the workers who have not been active for $\tau$ iterations, if such workers exist. Note that, when $\tau = 0$, one has $j \in A_t$ 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: Efficient Distributed Algorithm for Nonconvex-Nonsmooth Inference (EDANNI)

**Input:** Loss functions $\{L_j(x)\}_{j \in [m]}$, parameter $\rho$, initial point $x^0$. Set $d_1 = \cdots = d_m = 0$ and $A_0 = [m]$.

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

**Worker machines:**

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

if Receive $x'$ from the master then

Calculate gradient $\nabla L_j(x')$ and transmit it to the master.

end

end for

**Master:**

Receive $\{\nabla L_j(x^t)\}_{j=1}^m$ from worker machines $j$ in a set $A_t$ such that $d_j \leq \tau - 1$, $\forall j \in A_t$; then

Update

$$d_j = \begin{cases} 0 & \forall j \in A_t, \\ d_j + 1 & \forall j \in A^c_t. \end{cases}$$

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

end for

then (3) reduces to

$$x^{t+1} = \arg\min_{x} \left\{ \frac{1}{m} \sum_{j=1}^{m} \nabla L_j(x^{t_j}), x - x^t \right\} + h(x) + \frac{\rho}{2} \|x - x^t\|^2,$$

(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 $L_j(x)$ is differentiable and has Lipschitz continuous gradient for all $j \in [m]$, i.e.,

$$\|\nabla L_j(x) - \nabla L_j(y)\| \leq L\|x - y\|.$$

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

**Assumption III.2.** For all $j \in [m]$, the function $L_j$ is strongly convex with modulus $\sigma^2$, which means that

$$L_j(x) > L_j(y) + \langle \nabla L_j(y), x - y \rangle + \frac{\sigma^2}{2} \|x - y\|^2,\quad \text{for all } x, y \in \mathbb{R}^p, j \in [m].$$

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

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

III. $\gamma(\rho) > 3L + 2L\delta\tau$ and $\rho > \frac{2L}{\delta}$.
\(\delta > 0\), where \(\gamma(\rho)\) represents the convex modulus of the function \(h(x) + \frac{\rho}{2} \| x - x^t \|^2\).

II. There exists a constant \(L\) such that
\[ \text{L}(x) > L > -\infty \quad \forall x \in \mathbb{R}^n. \]

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

**Definition III.1.** We say a function \(\mathcal{F}(x)\) is coercive if
\[ \lim_{\|x\| \to \infty} \mathcal{F}(x) = +\infty. \]

Define
\[ \nabla_x \text{L}(x^t) = x^t - \text{Prox}_h \left( x^t - \frac{1}{m} \sum_{j \in [m]} \nabla L_j(x^t) \right), \quad (6) \]
where \(\text{Prox}_h\) is a proximal operator defined by \(\text{Prox}_h(z) := \text{argmin}_x h(x) + \frac{1}{2} \| x - z \|^2\). Usually \(\nabla_x \text{L}(x^t)\) is called the proximal gradient of \(\text{L}; x^t\) is a stationary point when \(\nabla_x \text{L}(x) = 0\).

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

**Theorem III.1.** Suppose Assumption \([\text{II}.1, \text{III}.1, \text{III.3}]\) are satisfied. Then the following claims for the sequence generated by Algorithm \([\text{EDANNI}]\).

- **(Boundedness of Sequence).** The gap between \(x^t\) and \(x^{t+1}\) converges to 0, i.e.,
\[ \lim_{t \to \infty} x^{t+1} - x^t = 0. \]

If \(\text{L}(x)\) is coercive, then the sequence \(\{x^t\}\) generated by Algorithm \([\text{EDANNI}]\) is bounded.

- **(Convergence to Stationary Points).** Every limit point of the iterates \(\{x^t\}\) generated by Algorithm \([\text{EDANNI}]\) is a stationary point of problem \([\text{II}]\). Furthermore, \(\| \nabla_x \text{L}(x^t) \| \to 0\), 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 \(\epsilon\), i.e.,
\[ T := \text{argmin}_t \left\{ \| \nabla_x \text{L}(x^t) \| < \epsilon \right\}. \]

Then there exists a constant \(\nu > 0\) such that
\[ T \leq \frac{\nu}{\epsilon} + 1, \]
where \(\nu\) equals to a positive constant times \(2\left(2 + \rho \right)^2 + 8L^2\delta\) / \(\min \left\{ \gamma(\rho) + \frac{L^2}{\delta} - L\delta, \frac{\rho}{2} - L\frac{\rho}{\delta} \right\}\) for some \(\delta > 0\). Therefore, the optimality gap \(\| \nabla_x \text{L}(x^t) \|\) converges to 0 in a sublinear manner.

**Remark III.1.** The theorem suggests that the iterates \(\{x^t\}\) may or may not be bounded without the coerciveness property of \(\text{L}(x)\). However, it guarantees that the optimality measure \(\| \nabla_x \text{L}(x^t) \|\) 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
\[ F(x, x^t) := \frac{1}{m} \sum_{j \in [m]} L_j(x) + \frac{\rho}{2} \| x - x^t \|^2 + h(x). \]
The gap between \(x^t\) and \(x^{t+1}\) is denoted by \(\Delta(t) := x^{t+1} - x^t\), for \(t \in \mathbb{N}\). The proof of Theorem III.1 relies on Lemma III.1 and III.2 in the following.

**Lemma III.1.** Suppose Assumption \([\text{III}.1, \text{III.3}]\) are satisfied, then the following is true for iterates \(\{x^t\}\) generated by Algorithm \([\text{EDANNI}]\)
\[ \frac{\rho}{2} \| x^{t+1} - x^t \|^2 + h(x^{t+1}) - \frac{\rho}{2} \| x^t - x^{t-1} \|^2 - h(x^t) \]
\[ \leq - \left( \nabla L_1(x^{t+1}) + \frac{1}{m} \sum_{j=1}^{m} \nabla L_j(x^{t+1}) - \nabla L_1(x^t), \right) \Delta(t) - \frac{\gamma(\rho)}{2} \| \Delta(t) \|^2 - \frac{\rho}{2} \| \Delta(t-1) \|^2. \quad (7) \]

**Lemma III.2.** Under the assumptions of Theorem III.1 for any \(\delta > 0\) we have
\[ F(x^{t+1}, x^t) - F(x^t, x^{t-1}) \]
\[ \leq \left( \frac{3L}{2} - \frac{\gamma(\rho)}{2} + L\delta\tau \right) \| \Delta(t) \|^2 - \frac{\rho}{2} \| \Delta(t-1) \|^2 \]
\[ + \frac{L}{\delta} \sum_{k=1}^{t} \| \Delta(t-k) \|^2. \quad (8) \]

**Lemma III.3.** Suppose Assumption \([\text{III.3}]\) is satisfied. Then for \(x^t\) generated by (EDANNI), there exists some constants \(F\) and \(\bar{F}\) such that
\[ +\infty > \bar{F} > F(x^{t+1}, x^t) > F > -\infty, \quad \forall t \geq 0. \]

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
\[ F(x^{T+1}, x^T) - F(x^1, x^0) \]
\[ \leq \sum_{t=1}^{T} \left( \frac{3L}{2} - \frac{\gamma(\rho)}{2} + L\delta\tau \right) \| \Delta(t) \|^2 \]
\[ + \sum_{t=1}^{T} \left( \frac{L\tau}{\delta} - \frac{\rho}{2} \right) \| \Delta(t-1) \|^2. \]

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 \(\rho > 2L\tau\), therefore \(c > 0\). It holds that
\[ F(x^{T+1}, x^T) - F(x^1, x^0) \leq -c \sum_{t=0}^{T} \| \Delta(t) \|^2. \quad (9) \]
Note that by Lemma III.3, the LHS of (9) is bounded from below. By letting $T \to \infty$, it follows that
\[
\|\Delta(t)\| \to 0, \quad t \to \infty.
\]
Moreover, Lemma III.3 shows that $F(x^{t+1}, x^t)$ is bounded, but due to the coerciveness assumption
\[
\lim_{\|x\| \to \infty} \frac{1}{m} \sum_{j \in [m]} L_j(x) + h(x) + \rho \|x - x^t\|^2 = +\infty,
\]
so we know \{$x^{t+1}\}$ is bounded. Therefore the first conclusion is proved.

We now establish the second conclusion of the Theorem. From (3), we know that
\[
x^{t+1} = \text{Prox}_h \left[ x^{t+1} - \left( \nabla L_1(x^{t+1}) + \frac{1}{m} \sum_{j \in [m]} \nabla L_j(x^t) - \nabla L_2(x^t) + \rho (x^{t+1} - x^t) \right) \right],
\]
where $\text{Prox}_h$ is a proximal operator defined by $\text{Prox}_h[z] := \arg\min_x \{h(x) + \frac{1}{2\rho} \|x - z\|^2\}$. This implies that
\[
\left\| x^t - \text{Prox}_h \left[ x^{t+1} - \left( \nabla L_1(x^{t+1}) + \frac{1}{m} \sum_{j \in [m]} \nabla L_j(x^t) - \nabla L_2(x^t) + \rho (x^{t+1} - x^t) \right) \right] \right\|
\leq \|x^t - x^{t+1} + x^{t+1} - \text{Prox}_h \left[ x^{t+1} - \left( \nabla L_1(x^{t+1}) + \frac{1}{m} \sum_{j \in [m]} \nabla L_j(x^t) - \nabla L_2(x^t) + \rho (x^{t+1} - x^t) \right) \right] \|
\leq \|x^t - x^{t+1}\|
+ \| \text{Prox}_h \left[ x^{t+1} - \left( \nabla L_1(x^{t+1}) + \frac{1}{m} \sum_{j \in [m]} \nabla L_j(x^t) - \nabla L_2(x^t) + \rho (x^{t+1} - x^t) \right) \right] \|
- \| \text{Prox}_h \left[ x^{t+1} - \left( \nabla L_1(x^{t+1}) + \frac{1}{m} \sum_{j \in [m]} \nabla L_j(x^t) \right) \right] \|
\leq (1 + \rho) \|x^{t+1} - x^t\| + \frac{1}{m} \sum_{j \in [m]} \| \nabla L_j(x^t) \|
- \| \nabla L_1(x^{t+1}) - \nabla L_1(x^{t+1}) - \nabla L_1(x^{t+1}) \|
+ \| \Delta(t) \|
\leq (2 + \rho) \| \Delta(t) \| + 2L \| \Delta(t) \|
\to 0, \quad t \to \infty.
\]

(11)

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

Let $X^*$ be the set of stationary points of problem (2), and let
\[
\text{dist} (x^t, X^*) := \min_{\tilde{x} \in X^*} \| x^t - \tilde{x} \|
\]
denote the distance between $x^t$ and the set $X^*$. Now we prove
\[
\lim_{t \to \infty} \text{dist} (x^t, X^*) = 0.
\]

Suppose there exists a subsequence \{ $x^{k_t}$ \} of \{ $x^t$ \} such that
\[
x^{k_t} \to \bar{x}, \quad k_t \to \infty
\]
but
\[
\lim_{k_t \to \infty} \text{dist}(x^{k_t}, X^*) \geq \gamma > 0.
\]
Then it is obvious that $\lim_{k \to \infty} \text{dist}(x^{k}, \bar{x}) = 0$. Therefore there exists some $K(\gamma) > 0$, such that
\[
\| x^{k_t} - \bar{x} \| \leq \frac{\gamma}{2}, \quad k > K(\gamma).
\]

On the other hand, from (11) and the lower semi-continuity of $h(x)$ we have $\bar{x} \in X^*$, so by the definition of the distance function we have
\[
\text{dist} (x^{k_t}, X^*) \leq \text{dist}(x^{k_t}, \bar{x}).
\]
Combining (13) and (14), we must have
\[
\text{dist} (x^{k_t}, X^*) \leq \frac{\gamma}{2}, \quad k > K(\gamma).
\]
This contradicts to (12), so the second result is proved.

We finally prove the third conclusion of the Theorem. Summing (11) over $t$ yields
\[
\sum_{t=0}^{T} \| \nabla L(x^t) \|^2 \leq \frac{\mu}{\epsilon} \left( F(x^1, x^0) - F(x^{T+1}, x^T) \right),
\]
where $\mu := (2 + \rho)^2 + 8L^2 \tau$.

Let $T(\epsilon) := \min \{ t | \| \nabla L(x^t) \| \leq \epsilon, t \geq 0 \}$. Then the above inequality implies
\[
T(\epsilon) \epsilon \leq \frac{\mu}{\epsilon} \left( F(x^1, x^0) - F(x^{T+1}, x^T) \right).
\]
Thus it follows that
\[
\epsilon \leq C : \frac{F(x^1, x^0) - F(x^0)}{T(\epsilon)},
\]
where $C := \frac{\mu}{\epsilon} > 0$, proving Theorem III.1.

Besides the convergence in the nonconvex setting, in the following theorem we show that the proposed algorithm converges linearly if $L_j$ 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] [II.2] and [II.3] are satisfied. If \( \rho \) is sufficiently large such that
\[
\frac{\delta_1 L + \frac{\rho}{2}(1 + \delta_1)}{\frac{\delta_1}{2}(1 + \delta_1) + \delta_1} + \frac{3L}{2} - \frac{\rho}{2} + L\delta \tau < 0
\]
and
\[
\frac{\delta_1 L + \frac{\rho}{2}(1 + \delta_1)}{\frac{\delta_1}{2}(1 + \delta_1) + \delta_1} + \frac{3L}{2} - \frac{\rho}{2} + L\delta \tau
\]
\[
< 0
\]
for some \( \delta > 0 \) and \( \delta_1 > (2L + \rho + 1)/\sigma^2 \), then it holds for the sequence generated by (EDANNI) that
\[
0 \leq F(x^{t+1}, x^t) - F(x^*, x^*) \leq \frac{1}{\eta} \left( F(x^t, x^0) - F(x^*, x^*) \right),
\]
where \( \eta := 1 + \frac{1}{\delta(1 + \delta_1) + \delta_1} \).

Note the above conditions can be satisfied when \( \rho \) is sufficiently larger than the order of \( L \) and the exponential of \( \tau \) and \( \delta_1 \) is larger than the order of \( L/\sigma^2 \). Theorem III.2 asserts that with the strongly convexity of \( L_j \)'s, the augmented optimality gap decreases linearly to zero under these conditions. Moreover, Assumption [II.3] can be replaced by only requiring each \( L_j \) is convex and \( \frac{\delta_1}{2}(1 + \delta_1) + \delta_1 \) is strongly convex with modulus \( \sigma^2 \).

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] [II.2] and [II.3] hold and \( \delta_1 > (2L + \rho + 1)/\sigma^2 \) for some \( \delta_1 > 0 \), then it follows that
\[
\frac{1}{\delta_1 L + \frac{\rho}{2}(1 + \delta_1)} \left( F(x^{t+1}, x^t) - F(x^*, x^*) \right) \leq \frac{\delta_1}{2}(1 + \delta_1) + \delta_1 \left\| x^t - x^{t+1} \right\|^2 + \frac{1}{\delta_1 L + \frac{\rho}{2}(1 + \delta_1) + \delta_1} \sum_{j \in [m]} \left\| x^t_j - x^t_{j+1} \right\|^2.
\]
(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 \( \overline{\Delta}^{(t+1)} := F(x^{t+1}, x^t) - F(x^*, x^*) \). Then from the proof of Lemma III.2 it holds that
\[
\overline{\Delta}^{(t+1)} \leq \overline{\Delta}^{(t)} + \left( \frac{3L}{2} - \frac{\rho}{2} + L\delta \tau \right) \left\| \Delta^{(t-1)} \right\|^2 - \frac{\rho}{2} \left\| \Delta^{(t-1)} \right\|^2 + \left( \frac{L}{\delta} \sum_{k=1}^{\tau} \left\| \Delta^{(t-k)} \right\|^2 \right)
\]
(17)

Note that from (16) of Lemma III.4, we have
\[
\frac{1}{\delta_1 L + \frac{\rho}{2}(1 + \delta_1)} \left( \frac{\delta_1}{2}(1 + \delta_1) + \delta_1 \right) \left\| \Delta^{(t)} \right\|^2
\]
\[
+ \frac{1}{\delta_1 L + \frac{\rho}{2}(1 + \delta_1) + \delta_1} \sum_{j \in [m]} \left\| x^t_j - x^t_{j+1} \right\|^2
\]
(18)

By combining (18) and (17), we have the following bound of the LHS:
\[
\left( 1 + \frac{1}{\delta_1 L + \frac{\rho}{2}(1 + \delta_1) + \delta_1} \right) \overline{\Delta}^{(t+1)}
\]
\[
\leq \overline{\Delta}^{(t)} + \left( \frac{\delta_1 L + \frac{\rho}{2}(1 + \delta_1)}{\delta_1 L + \frac{\rho}{2}(1 + \delta_1) + \delta_1} + \frac{3L}{2} - \frac{\rho}{2} + L\delta \tau \right) \left\| \Delta^{(t)} \right\|^2
\]
\[
- \frac{\rho}{2} \left\| \Delta^{(t-1)} \right\|^2 + \left( \frac{L}{\delta} \sum_{k=1}^{\tau} \left\| \Delta^{(t-k)} \right\|^2 \right)
\]
\[
+ \frac{\delta_1}{(1 + \delta_1) + \delta_1} \sum_{j \in [m]} \left\| x^t_j - x^t_{j+1} \right\|^2.
\]
(19)

Inequality (19) gives us an relation between \( \overline{\Delta}^{(t+1)} \) and \( \overline{\Delta}^{(t)} \). Let us define \( \eta := 1 + \frac{1}{\delta(1 + \delta_1) + \delta_1} \) and
\[
(P3) := \frac{\delta_1 L + \frac{\rho}{2}(1 + \delta_1)}{\delta_1 L + \frac{\rho}{2}(1 + \delta_1) + \delta_1} + \frac{3L}{2} - \frac{\rho}{2} + L\delta \tau,
\]
then by applying (19) recursively we have
\[
\overline{\Delta}^{(t+1)}
\]
\[
\leq \frac{1}{\eta} \overline{\Delta}^{(t)} + \frac{1}{\eta}(P3) \left\| \Delta^{(t)} \right\|^2 - \frac{\rho}{2\eta} \left\| \Delta^{(t-1)} \right\|^2
\]
\[
+ \frac{L}{\delta \eta} \sum_{k=1}^{\tau} \left\| \Delta^{(t-1-k)} \right\|^2 + \frac{1}{\eta^2} \left( \frac{1}{\eta}(P3) \left\| \Delta^{(t-1-k)} \right\|^2 - \frac{\rho}{2\eta^2} \left\| \Delta^{(t-2-k)} \right\|^2 \right)
\]
\[
+ \frac{\delta_1 L^2}{2m} \sum_{j \in [m]} \left\| x^t_j - x^t_{j+1} \right\|^2
\]
\[
+ \frac{1}{\eta^2} \left( \frac{1}{\eta}(P3) - \frac{\rho}{2\eta^2} \right) \left\| \Delta^{(t-2-k)} \right\|^2 + \cdots
\]
\[
\leq \frac{1}{\eta^t} \overline{\Delta}^{(1)} + \frac{1}{\eta^t}(P3) \left\| \Delta^{(1)} \right\|^2 + \left( \frac{1}{\eta^t}(P3) - \frac{\rho}{2\eta^t} \right) \left\| \Delta^{(0)} \right\|^2
\]
\[
\leq \frac{1}{\eta^{t+1}} \overline{\Delta}^{(1)} + \frac{1}{\eta^{t+1}}(P3) \left\| \Delta^{(1)} \right\|^2 + \left( \frac{1}{\eta^{t+1}}(P3) - \frac{\rho}{2\eta^{t+1}} \right) \left\| \Delta^{(0)} \right\|^2
\]
\[
+ \frac{1}{\eta^{t+1}} \left( \frac{1}{\eta^{t+1}}(P3) - \frac{\rho}{2\eta^{t+1}} \right) \left\| \Delta^{(0)} \right\|^2 + \cdots
\]
\[
\leq \frac{1}{\eta^{t+1}} \overline{\Delta}^{(1)} + \frac{1}{\eta^{t+1}}(P3) \left\| \Delta^{(1)} \right\|^2 + \left( \frac{1}{\eta^{t+1}}(P3) - \frac{\rho}{2\eta^{t+1}} \right) \left\| \Delta^{(0)} \right\|^2
\]
\[
+ \frac{1}{\eta^{t+1}} \left( \frac{1}{\eta^{t+1}}(P3) - \frac{\rho}{2\eta^{t+1}} \right) \left\| \Delta^{(0)} \right\|^2 + \cdots
\]

where we use the fact that

\[
\sum_{t=0}^{\tau+1} \frac{1}{\eta^t+1} \left\| \Delta^{(t-l-k)} \right\|^2 \\
= \eta^{-t-1} \sum_{t=0}^{\tau} \eta^t \sum_{k=1}^{\tau} \left\| \Delta^{(t-l-k)} \right\|^2 \\
= \eta^{-t-1} \sum_{t=0}^{\tau} \eta^t \sum_{k=1}^{\tau} \left\| \Delta^{(t-k)} \right\|^2 \\
\leq \eta^{-t-1} \sum_{t=0}^{\tau} \eta^t \sum_{k=1}^{\tau} \left\| \Delta^{(t-k)} \right\|^2 \\
\leq \eta^{-t-1} \sum_{t=0}^{\tau} \eta^t \sum_{k=1}^{\tau} \left\| \Delta^{(t-k)} \right\|^2 \\
\leq \eta^{-t-1} \sum_{t=0}^{\tau} \eta^t \sum_{k=1}^{\tau} \left\| \Delta^{(t-k)} \right\|^2 \\
\leq \eta^{-t-1} \sum_{t=0}^{\tau} \eta^t \sum_{k=1}^{\tau} \left\| \Delta^{(t-k)} \right\|^2. 
\]

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

Therefore if \(\rho > 0\) satisfies that

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

and

\[
(P3) - \frac{m \rho}{2} + \left( \frac{L}{\delta} + \frac{\delta_1 L^2 \tau}{2 (1 + \delta_1) + \delta_1} \right) \eta^{-1} - 1 \\
= \frac{\delta_1 L + \frac{L}{\delta} (1 + \delta_1)}{2 (1 + \delta_1) + \delta_1} + \frac{3 L}{2} - \frac{\rho}{2} + L \delta \tau \\
- \frac{m \rho}{2} + \left( \frac{L}{\delta} + \frac{\delta_1 L^2 \tau}{2 (1 + \delta_1) + \delta_1} \right) \eta^{-1} - 1 < 0, \tag{21}
\]

then we have

\[
0 \leq \tilde{\Delta}^{(t+1)} \leq \frac{1}{\eta^t} \Delta^{(1)}. 
\]

The conclusion is proved. \(\square\)

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 \(\epsilon^t\) such that

\[
\epsilon^t \in \nabla L_1(x^{t+1}) + \frac{1}{m} \sum_{j=0}^{m} \nabla L_j(x^{t}) - \nabla L_1(x^{t+1}) + \rho \left( x^{t+1} - x^t \right), \tag{22}
\]

which is equivalent to

\[
x^{t+1} = \text{Prox}_{h} \left[ x^{t+1} - \left( \nabla L_1(x^{t+1}) + \frac{1}{m} \sum_{j=0}^{m} \nabla L_j(x^{t}) + \rho \left( x^{t+1} - x^t \right) - \epsilon^t \right] \right]. \tag{23}
\]

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

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

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

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

\[
\frac{\rho}{2} \left\| x^{t+1} - x^t \right\|^2 + h(x^{t+1}) - \frac{\rho}{2} \left\| x^t - x^{t-1} \right\|^2 - h(x^t) \\
\leq -\left( \nabla L_1(x^{t+1}) + \frac{1}{m} \sum_{j=1}^{m} \nabla L_j(x^{t}) - \nabla L_1(x^{t+1}) - \epsilon^t, \Delta^{(t)} \right) - \frac{\gamma(\rho)}{2} \left\| \Delta^{(t)} \right\|^2 - \frac{\rho}{2} \left\| \Delta^{(t-1)} \right\|^2.
\]

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

\[
F(x^{T+1}, x^T) - F(x^1, x^0) \leq \left( \frac{3 L}{2} - \frac{\gamma(\rho)}{2} + L \delta \tau \right) \left\| \Delta^{(t)} \right\|^2 + \frac{\tau \rho}{2} \sum_{i=1}^{T} \left\| \Delta^{(i-1)} \right\|^2 + \langle \epsilon^t, \Delta^{(i)} \rangle, \tag{24}\]

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

\[
F(x^{T+1}, x^T) - F(x^1, x^0) \leq \sum_{t=1}^{T} \left( \frac{3 L}{2} - \frac{\gamma(\rho)}{2} + L \delta \tau \right) \left\| \Delta^{(t)} \right\|^2 + \frac{\tau \rho}{2} \sum_{i=1}^{T} \left\| \Delta^{(i-1)} \right\|^2 + \langle \epsilon^t, \Delta^{(t)} \rangle \leq \sum_{t=1}^{T} \left( \frac{3 L}{2} - \frac{\gamma(\rho)}{2} + L \delta \tau \right) \left\| \Delta^{(t)} \right\|^2 + \frac{\tau \rho}{2} \sum_{i=1}^{T} \left( \left\| \epsilon^t \right\|^2 + \left\| \Delta^{(i)} \right\|^2 \right) \leq \sum_{t=1}^{T} \left( \frac{3 L}{2} - \frac{\gamma(\rho)}{2} + L \delta \tau + \frac{1}{2} c_1 \right) \left\| \Delta^{(t-1)} \right\|^2 + \frac{\tau \rho}{2} c_1 \sum_{i=1}^{T} \left\| \epsilon^t \right\|^2, \tag{26}\]

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

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

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

then we have \(\bar{c} > 0\). Therefore

\[
F(x^{T+1}, x^T) - F(x^1, x^0) \leq -\bar{c} \sum_{t=0}^{T} \left\| \Delta^{(t)} \right\|^2. \tag{26}\]

Note that by Lemma III.3 the LHS of \((26)\) is bounded from
below. It follows that
\[ \| \Delta^{(t)} \| \to 0, \quad t \to \infty. \]

We now establish the second conclusion of Theorem III.1. From (23) we know that
\[ x^{t+1} = \text{Prox}_h \left[ x^{t+1} - \left( \nabla L_1(x^{t+1}) + \frac{1}{m} \sum_{j \in [m]} \nabla L_j(x^t) \right) \right], \]
where Prox$_h$ is a proximal operator defined by \( \text{Prox}_h[z] := \arg \min_x h(x) + \frac{1}{2} \| x - z \|_2^2 \). This implies that
\[
\left\| x^t - \text{Prox}_h \left( x^t - \frac{1}{m} \sum_{j \in [m]} \nabla L_j(x^t) \right) \right\| \lesssim |t-t^*| + \frac{1}{2} \left( \| x_t^* - x_t \|_2 + \epsilon_t \right) \delta_t \| \Delta^{(t)} \| + \frac{1}{2} \| \Delta^{(t)} \|_2.
\]

By a recursive argument similar to the proof of Theorem III.1, one can prove that if \( \rho > 0 \) satisfies that
\[
\delta_t L + \frac{\rho}{2} (1 + \delta_t) + \frac{1}{2} \left( \| x_{t+1}^* - x_t^* \|_2^2 + \epsilon_t \right) \leq \frac{\rho}{2} (1 + \delta_t) + \delta_t \| \Delta^{(t)} \|_2.
\]
and

\[
\begin{aligned}
&\frac{\delta_1 L + \frac{\rho}{2} (1 + \delta_1) + \frac{\rho}{2}}{\rho (1 + \delta_1) + \delta_1} + 3 L - \frac{\rho - 1}{2} + L \delta \tau - \frac{(\rho - c_1) \eta}{2} \\
&+ \left( \frac{L}{\delta} + \frac{\delta_1 L^2 \tau}{\rho (1 + \delta_1) + \delta_1} \right) \eta - 1 + \frac{c_1 \eta}{\rho (1 + \delta_1) + \delta_1} < 0,
\end{aligned}
\]  

(31)

then we have

\[
0 \leq \tilde{\Delta}^{(t+1)} \leq \frac{1}{\eta} \tilde{\Delta}^{(1)}.
\]

In summary we have the following corollary.

**Corollary III.2.** Suppose Assumption II.1, III.1, III.2, and III.4 are satisfied. If \( \rho \) 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

\[
0 \leq F(x^{t+1}, x^t) - F(x^*, x^*) \\
\leq \frac{1}{\eta_1} F(x^t, x^0) - F(x^*, x^*),
\]

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

### 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

\[
\argmin_{w \in \mathbb{R}^p} \frac{1}{2mn} \sum_{j \in [m]} \sum_{i \in [n]} \|x_{ji}^T w - y_{ji}\|^2 + \theta \|w\|_1,
\]

(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 \( w \) instead of \( x \).

The data \( \{x_{ji} \}_{i \in [n], j \in [m]} \) is independently sampled from a multivariate Gaussian distribution with zero mean and covariance matrix \( \Sigma \). 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

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

where noise \( \epsilon_{ji} \) is a zero mean Gaussian random variable with variance 0.01. The true parameter \( w^* \) 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 \( x_{ji} \).

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 \( \rho \) 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 0.5 and the other workers are assumed be arrived with probability 0.5. Moreover, the maximum tolerable delay \( \tau \) is set be 3 and the master will update the variables once the workers \( j \) with \( d_j > \tau - 1 \) have arrived.

![Fig. 2. Comparison of candidate algorithms in LASSO when \( m = 20, n = 500, p = 1000, s = 10, \) and \( \Sigma \sim N(0, \Sigma) \).
](image)

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 1. It is shown in these results that...
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 $\tau$. It can be observed from Figure 3(a) that the convergence rate varies much with different values of $\tau$. Basically, EDANNI converges faster when $\tau$ 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 $\tau$ 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:

$$
\arg\min_{\mathbf{w}} - \frac{1}{m} \sum_{j \in [m]} \sum_{i \in [n]} \mathbf{w}^T B_{ji} B_{ji}^T \mathbf{w} + \theta \| \mathbf{w} \|_1,
\text{s.t. } \| \mathbf{w} \| \leq 1
$$

(33)

where $B_{ji} \in \mathbb{R}^{p \times q}$ is a sparse matrix, $\forall j \in [m], \ i \in [n]$, 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 $\rho$ in (6) 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 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 $\tau$ 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 $\tau$. The running time when $\tau = 15$ is much less than that when $\tau$ 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 $\tau$ becomes 15, therefore speeding up the convergence. Such results can be observed generally regardless of the choice of parameters $m, n, p, q, s, \theta$.

![Graph](image1.png)

(a) Convergence for different $\tau$’s

![Graph](image2.png)

(b) Time table for different $\tau$’s

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

![Graph](image3.png)

(a) Convergence for different $\tau$’s

![Graph](image4.png)

(b) Asynchronous, (2.5, 2.0, 1)

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

![Graph](image5.png)

(a) Convergence for different $\tau$’s

![Graph](image6.png)

(b) Time table for different $\tau$’s

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

| Type            | Parameters                  | Method       | Communication |
|-----------------|-----------------------------|--------------|---------------|
| Synchronous     | $(20,100,50,100,50,0.1)$    | AD-ADMM      | 11.5          |
| $(30,200,50,100,50,0.1)$ | PS                          | 2.3          |
| $(3,20,500,1000,500,1)$ | EDANNI                     | 1.1          |
| Asynchronous    | $(20,100,50,100,50,0.1)$    | AD-ADMM      | 12.5          |
| $(30,200,50,100,50,0.1)$ | PS                          | 4.1          |
| $(3,20,500,1000,500,1)$ | EDANNI                     | 4.1          |

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.
APPENDIX A
PROOFS OF LEMMATA

A. Proof of Lemma III.1

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

$$-\left[\nabla L_1(x^{t+1}) + \frac{1}{m} \sum_{j=1}^{m} \nabla L_j(x^t) - \nabla L_1(x^t)\right] \in \partial \left[h(x) + \frac{\rho}{2} \|x - x^t\|^2\right].$$

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

$$\frac{\rho}{2}\|x^{t+1} - x^t\|^2 + h(x^{t+1}) - \left(\frac{\rho}{2}\|x^t - x^t\|^2 + h(x^t)\right) \leq -\left[\nabla L_1(x^{t+1}) + \frac{1}{m} \sum_{j=1}^{m} \nabla L_j(x^t) - \nabla L_1(x^t)\right] \Delta(t) - \frac{\gamma(\rho)}{2} \|\Delta(t)\|^2,$$

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

$$\frac{\rho}{2}\|x^{t+1} - x^t\|^2 + h(x^{t+1}) - \frac{\rho}{2}\|x^t - x^t\|^2 - h(x^t) \leq \frac{\rho}{2}\|x^{t+1} - x^t\|^2 + h(x^{t+1}) - \frac{\rho}{2}\|x^t - x^t\|^2 - h(x^t) + \frac{\rho}{2}\|x^t - x^t\|^2 + h(x^t) - \frac{\rho}{2}\|x^t - x^t\|^2 + h(x^t)

\leq -\left[\nabla L_1(x^{t+1}) + \frac{1}{m} \sum_{j=1}^{m} \nabla L_j(x^t) - \nabla L_1(x^t)\right] \Delta(t) - \frac{\gamma(\rho)}{2} \|\Delta(t)\|^2 - \frac{\rho}{2} \|\Delta(t-1)\|^2,$$

proving Lemma III.1

B. Proof of Lemma III.2

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

$$\frac{1}{m} \sum_{j=1}^{m} L_j(x^{t+1}) - \frac{1}{m} \sum_{j=1}^{m} L_j(x^t) \leq\left\langle x^{t+1} - x^t, \frac{1}{m} \sum_{j=1}^{m} \nabla L_j(x^t)\right\rangle + \frac{L}{2}\|x^{t+1} - x^t\|^2$$

$$=\left\langle \Delta(t), \frac{1}{m} \sum_{j=1}^{m} \nabla L_j(x^t)\right\rangle + \frac{L}{2}\|\Delta(t)\|^2.$$

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

$$F(x^{t+1}, x^t) - F(x^t, x^{t-1}) \leq \left\langle \nabla L_1(x^{t+1}) + \frac{1}{m} \sum_{j=1}^{m} \nabla L_j(x^t) - \nabla L_1(x^t), \Delta(t)\right\rangle - \frac{\gamma(\rho)}{2} \|\Delta(t)\|^2 - \frac{\rho}{2} \|\Delta(t-1)\|^2$$

$$+\left\langle \nabla L_1(x^{t+1}) + \frac{1}{m} \sum_{j=1}^{m} \nabla L_j(x^t) - \nabla L_1(x^t), \Delta(t)\right\rangle - \frac{\gamma(\rho)}{2} \|\Delta(t)\|^2 - \frac{\rho}{2} \|\Delta(t-1)\|^2$$

$$=\left\langle \nabla L_1(x^{t+1}) + \frac{1}{m} \sum_{j=1}^{m} \nabla L_j(x^t) - \nabla L_1(x^t), \Delta(t)\right\rangle - \frac{\gamma(\rho)}{2} \|\Delta(t)\|^2 - \frac{\rho}{2} \|\Delta(t-1)\|^2$$

$$\leq \left\langle \nabla L_1(x^{t+1}) - \nabla L_1(x^t), \Delta(t)\right\rangle - \frac{\gamma(\rho)}{2} \|\Delta(t)\|^2 - \frac{\rho}{2} \|\Delta(t-1)\|^2$$

$$+\left\langle \nabla L_1(x^{t+1}) - \nabla L_1(x^t), \Delta(t)\right\rangle - \frac{\gamma(\rho)}{2} \|\Delta(t)\|^2 - \frac{\rho}{2} \|\Delta(t-1)\|^2$$

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

Note that

$$\nabla L_1(x^{t+1}) - \nabla L_1(x^t) \leq \sum_{k=1}^{t+1} (\nabla L_1(x^{t-k}) - \nabla L_1(x^{t-k+1})),$$

which implies

$$\|\nabla L_1(x^{t+1}) - \nabla L_1(x^t)\|$$

$$\leq \sum_{k=1}^{t+1} \|\nabla L_1(x^{t-k}) - \nabla L_1(x^{t-k+1})\|$$

$$\leq \sum_{k=1}^{t} L \|x^{t-k} - x^{t-k+1}\|$$

$$\leq \sum_{k=1}^{t} L \|\Delta(t-k)\|.$$
Similarly, we can see
\[ \left\| \frac{1}{m} \sum_{j \in [m]} \nabla L_j(x^t) - \frac{1}{m} \sum_{j \in [m]} \nabla L_j(x^{t_j}) \right\| \leq \sum_{k=1}^T L \Delta^{(t-k)} \|. \]

These two inequalities result in
\[ \begin{align*}
(P1) & \leq 2L \sum_{k=1}^T \Delta^{(t-k)} \| \Delta^{(t)} \|
\leq L \sum_{k=1}^T \left( \frac{1}{\delta} \| \Delta^{(t-k)} \|^2 + \delta \| \Delta^{(t)} \|^2 \right)
\leq \frac{L}{\delta} \sum_{k=1}^T \| \Delta^{(t-k)} \|^2 + L\delta \tau \| \Delta^{(t)} \|^2 ,
\end{align*} \]

(38)

where in the second inequality we apply the fact that
\[ a \cdot b \leq \frac{1}{2} \left( \frac{1}{\delta} a^2 + 2\delta^2 b^2 \right), \]

for any \( a, b, \delta > 0 \). By inserting (38) into (37) we have
\[ F(x^{t+1}, x^t) - F(x^t, x^{t-1}) \leq \left( \frac{3L}{2} - \frac{\gamma(\rho)}{2} + L\delta \tau \right) \| \Delta^{(t)} \|^2 - \frac{\rho}{2} \| \Delta^{(t-1)} \|^2 \]

\[ + \frac{L}{\delta} \sum_{k=1}^T \| \Delta^{(t-k)} \|^2, \]

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
\[ F(x^{T+1}, x^T) - F(x^1, x^0) \leq \sum_{t=0}^T \left( \frac{3L}{2} - \frac{\gamma(\rho)}{2} + L\delta \tau \right) \| \Delta^{(t)} \|^2 \]

\[ + \sum_{t=0}^T \left( \frac{L\tau}{\delta} - \frac{\rho}{2} \right) \| \Delta^{(t-1)} \|^2. \]

If \( \rho \) satisfies Assumption III.3 then it holds that
\[ F(x^{T+1}, x^T) - F(x^1, x^0) < 0. \]

By taking \( x^T \) as the initial point, similarly we have
\[ F(x^{T+1}, x^{T+1}) - F(x^T+1, x^T) < 0. \]

Continuing this process we get a decreasing subsequence \( \{F(x^{kT+1}, x^{kT})\}_{k=0,1,\cdots} \). Therefore there exists a constant \( F_0 \) such that
\[ F(x^{kT+1}, x^{kT}) \leq F_0. \]

(39)

When starting with \( x^1, \cdots, x^{T-1} \), with similar analysis we can prove that there exists constants \( F_1, \cdots, F_{T-1} \) such that
\[ F(x^{kT+1}, x^{kT+1}) \leq F_t, \]

(40)

for \( l = 1, 2, \cdots, T - 1 \). Define \( \tilde{F} := \max \{F_0, \cdots, F_{T-1}\} \), then
\[ F(x^{t+1}, x^t) < \tilde{F} < +\infty, \quad \forall t \in \mathbb{N}. \]

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

\[ = \frac{1}{m} \sum_{j \in [m]} \Delta_j(x^{t+1}) + \frac{\rho}{2} \| x^{t+1} - x^t \|^2 + h(x^{t+1}) \]

\[ \geq \frac{1}{m} \sum_{j \in [m]} \Delta_j(x^{t+1}) + h(x^{t+1}) \]

\[ = L(x^{t+1}) > L = \mathbb{F} > -\infty, \]

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
\[ F(x^{t+1}, x^t) = F(x^*, x^t) - F(x^*, x^{t+1}) \]

\[ + \sum_{j \in [m]} \Delta_j(x^{t+1}) + \frac{\rho}{2} \| x^{t+1} - x^t \|^2 + h(x^{t+1}) \]

\[ \geq 0, \]

for all \( x \in \mathbb{R}^p \). Setting \( x = x^* \) implies
\[ \left( \frac{\rho}{2} \| x^{t+1} - x^t \|^2 + h(x^{t+1}) \right) \leq 0. \]

(41)

By the strong convexity of \( L_j \) one has
\[ L_j(y) \geq L_j(x) + \sum_{j \in [m]} \Delta_j(x^{t+1}) \]

\[ + \partial h(x^{t+1}) + \rho (x^t - x^{t+1})^\top (x^{t+1} - x) \leq 0. \]

(42)

By the strong convexity of \( L_j \) one has
\[ L_j(y) \geq L_j(x^*) + \sum_{j \in [m]} \Delta_j(x^{t+1}) + \frac{\rho}{2} \| y - x \|^2, \]

\[ \forall x, y \in \mathbb{R}^p. \]

(43)

Setting \( y = x^* \), \( x = x^{t+1} \) in (43) we have
\[ L_j(x^*) \geq L_j(x^{t+1}) + \sum_{j \in [m]} \Delta_j(x^{t+1}) + \frac{\rho}{2} \| x^* - x^{t+1} \|^2, \]

which further implies that
\[ (\nabla L_j(x^{t+1})^\top (x^{t+1} - x^*) \]

\[ \geq L_j(x^{t+1}) - L_j(x^*) + \frac{\rho}{2} \| x^* - x^{t+1} \|^2. \]

(44)
Note that (42) implies that
\[
\left( \nabla L_1(x^{t+1}) - \nabla L_1(x^t) + \left( \frac{1}{m} \sum_{j \in [m]} \nabla L_j(x^{t+1}) \right) \right) \nabla \rho(x^{t+1}) + \rho(x^{t+1} - x^t)^T (x^{t+1} - x^*) \leq 0.
\]

Summing (44) over \( j \in [m] \) gives that
\[
\frac{1}{m} \sum_{j \in [m]} \left( \nabla L_j(x^{t+1}) \right)^T (x^{t+1} - x^*) \geq \frac{1}{m} \sum_{j \in [m]} L_j(x^{t+1}) - \frac{1}{m} \sum_{j \in [m]} L_j(x^*) + \frac{1}{2} \|x^* - x^{t+1}\|^2.
\]

Putting (46) into (45), we have
\[
\left( \nabla L_1(x^{t+1}) - \nabla L_1(x^t) + \left( \frac{1}{m} \sum_{j \in [m]} \nabla L_j(x^{t+1}) \right) \right) \nabla \rho(x^{t+1}) + \rho(x^{t+1} - x^t)^T (x^{t+1} - x^*) \leq 0.
\]

Then putting (49) into (48), one obtains
\[
\left( \nabla L_1(x^{t+1}) - \nabla L_1(x^t) + \left( \frac{1}{m} \sum_{j \in [m]} \nabla L_j(x^{t+1}) \right) \right)^T (x^{t+1} - x^*) \leq 0.
\]

Putting (46) into (45), we have
\[
\frac{1}{m} \sum_{j \in [m]} \left( \nabla L_j(x^{t+1}) \right)^T (x^{t+1} - x^*) \geq \frac{1}{m} \sum_{j \in [m]} L_j(x^{t+1}) - \frac{1}{m} \sum_{j \in [m]} L_j(x^*) + \frac{1}{2} \|x^* - x^{t+1}\|^2.
\]

Putting it into (47), we have
\[
\left( \nabla L_1(x^{t+1}) - \nabla L_1(x^t) + \left( \frac{1}{m} \sum_{j \in [m]} \nabla L_j(x^{t+1}) \right) \right) \nabla \rho(x^{t+1}) + \rho(x^{t+1} - x^t)^T (x^{t+1} - x^*) \leq 0.
\]

Now, note that
\[
\frac{1}{m} \sum_{j \in [m]} \nabla L_j(x^{t+1}) \nabla \rho(x^{t+1}) + \rho(x^{t+1} - x^t)^T (x^{t+1} - x^*) \leq 0.
\]

By the Mean Value Theorem one has
\[
\left( \nabla L_1(x^{t+1}) - \nabla L_1(x^t) + \left( \frac{1}{m} \sum_{j \in [m]} \nabla L_j(x^t) \right) \right) \nabla \rho(x^{t+1}) + \rho(x^{t+1} - x^t)^T (x^{t+1} - x^*) \leq 0.
\]

Note that
\[
\rho(x^{t+1} - x^t)^T (x^{t+1} - x^*) = \frac{\rho}{2} \|x^{t+1} - x^*\|^2 + \frac{\rho}{2} \|x^{t+1} - x^t\|^2 - \frac{\rho}{2} \|x^t - x^*\|^2.
\]
Note that which implies that
The optimality gap of function $F$ by

$$\left( \nabla L_1(x^{t+1}) - \nabla L_1(x^t) + \left( \frac{1}{m} \sum_{j \in [m]} \nabla L_j(x^t) \right)^\top (x^{t+1} - x^t) \right) \geq \frac{1}{2} \|x^{t+1} - x^*\|^2 + \frac{1}{2} \|x^t - x^{t+1}\|^2
\geq \frac{1}{2} \|x^t - x^{t+1}\|^2 + \frac{1}{2} \|x^t - x^*\|^2
\geq \frac{1}{2} \|x^t - x^*\|^2 + \frac{1}{2} \|x^t - x^{t+1}\|^2$$

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

$$F(x^{t+1}, x^t) - F(x^*, x^*) \leq -\frac{\sigma^2}{2} \|x^* - x^{t+1}\|^2
+ \frac{1}{2} \|x^{t+1} - x^*\|^2 + \frac{1}{2} \|x^t - x^{t+1}\|^2
- \frac{\rho}{2} \|x^{t+1} - x^*\|^2 + \frac{1}{2} \|x^t - x^*\|^2
- \frac{\delta_1}{2} \|x^t - x^{t+1}\|^2 + \frac{1}{2} \|x^t - x^*\|^2
+ \frac{\delta_1}{2m} \sum_{j \in [m]} \|x^t - x^*\|^2$$

Therefore we have

$$F(x^{t+1}, x^t) - F(x^*, x^*) \leq -\frac{\sigma^2}{2} \|x^* - x^{t+1}\|^2 + \frac{1}{2\delta_1} \|x^{t+1} - x^*\|^2
+ \frac{\rho}{2(1 + \delta_1)} \|x^t - x^{t+1}\|^2$$

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

$$F(x^{t+1}, x^t) - F(x^*, x^*) \leq \frac{1}{2} \delta_1 L \|x^t - x^{t+1}\|^2 + \frac{\delta_1}{2m} \sum_{j \in [m]} \|x^t - x^*\|^2$$

from which we have

$$\frac{1}{2} \delta_1 L \|x^t - x^{t+1}\|^2 + \frac{1}{2m} \sum_{j \in [m]} \|x^t - x^*\|^2 \leq \frac{\delta_1 L}{2(1 + \delta_1)} \|x^t - x^{t+1}\|^2$$

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

$$\frac{1}{2m} \sum_{j \in [m]} \|\nabla L_j(x^t) - \nabla L_j(x^{t+1})\|^2 \geq \frac{1}{2m} \sum_{j \in [m]} \|x^t - x^{t+1}\|^2$$

Therefore we have

$$\frac{1}{2m} \sum_{j \in [m]} \|x^t - x^{t+1}\|^2 \geq \frac{1}{2m} \sum_{j \in [m]} \|x^t - x^*\|^2$$