Distributed Stochastic Consensus Optimization With Momentum for Nonconvex Nonsmooth Problems

Zhiguo Wang, Jiawei Zhang, Tsung-Hui Chang, *Senior Member, IEEE*, Jian Li, *Fellow, IEEE*, and Zhi-Quan Luo*, Fellow, IEEE*

***Abstract*—While many distributed optimization algorithms have been proposed for solving smooth or convex problems over the networks, few of them can handle non-convex and non-smooth problems. Based on a proximal primal-dual approach, this paper presents a new (stochastic) distributed algorithm with Nesterov momentum for accelerated optimization of non-convex and non-smooth problems. Theoretically, we show that the proposed algorithm can achieve an ϵ-stationary solution under a constant step size with $\mathcal{O}(1/\epsilon^2)$ computation complexity and $\mathcal{O}(1/\epsilon)$ communication complexity when the epigraph of the non-smooth term is a polyhedral set. When compared to the existing gradient tracking based methods, the proposed algorithm has the same order of computation complexity but lower order of communication complexity. To the best of our knowledge, the presented result is the first stochastic algorithm with the $\mathcal{O}(1/\epsilon)$ communication complexity for non-convex and non-smooth problems. Numerical experiments for a distributed non-convex regression problem and a deep neural network based classification problem are presented to illustrate the effectiveness of the proposed algorithms.**

***Index Terms*—Distributed optimization, stochastic optimization, momentum, non-convex and non-smooth optimization.**

I. Introduction

RECENTLY, motivated by large-scale machine learning [1] and mobile edge computing [2], many signal processing applications involve handling very large datasets [3] that are processed over networks with distributed memories and processors. Such signal processing and machine learning problems are usually formulated as a multi-agent distributed optimization problem [4]. In particular, many of the applications can be formulated as the following finite sum problem

$$\min_{x} \sum_{i=1}^{N} \left(f_i(x) + r_i(x) \right), \tag{1}$$

where N is the number of agents, $x \in \mathbb{R}^n$ contains the model parameters to be learned, $f_i(x) : \mathbb{R}^n \to \mathbb{R}$ is a closed and smooth (possibly nonconvex) loss function, and $r_i(x)$ is a convex and possibly non-smooth regularization term. Depending on how the data are acquired, there are two scenarios for problem (1) [5].

- Offline/Batch learning: the agents are assumed to have the complete local dataset. Specifically, the local cost functions can be written as

$$f_i(x) = \frac{1}{m} \sum_{j=1}^{m} f_i^j(x), i = 1, \ldots, N, \tag{2}$$

  where $f_i^j(x)$ is the cost for the j-th data sample at the i-th agent, and m is the total number of local samples. When m is not large, each agent i may compute the full gradient of $f_i(x)$ for deterministic parameter optimization.
- Online/Streaming learning: when the data samples follow certain statistical distribution and are acquired by the agents in an online/streaming fashion, one can define $f_i(x)$ as the following expected cost

$$f_i(x) = \mathbb{E}_{\xi \sim \mathcal{B}_i}[f_i(x, \xi)], i = 1, \ldots, N, \tag{3}$$

  where $\mathcal{B}_i$ denotes the data distribution at agent i, and $f_i(x, \xi)$ is the cost function of a random data sample ξ. Under the online setting, only a stochastic estimate $G_i(x, \xi)$ for $\nabla f_i(x)$ can be obtained by the agent and stochastic optimization methods can be used. Note that if the agent is not able to compute the full gradient in the batch setting, a stochastic gradient estimate by mini-batch data samples (with size $|\mathcal{I}|$) can be obtained and the problem is solved in a similar fashion by stochastic optimization.

These two settings for local cost functions are popularly used in many machine learning models including deep learning and empirical risk minimization problems [5]. For both scenarios, many distributed optimization methods have been developed for solving problems (1).

Specifically, for batch learning and under convex or strongly convex assumptions, algorithms such as the distributed subgradient method [6], EXTRA [7], PG-EXTRA [8] and primal-dual based methods including the alternating direction method of

Manuscript received November 19, 2020; revised May 14, 2021 and July 3, 2021; accepted July 9, 2021. Date of publication July 16, 2021; date of current version August 11, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chiara Ravazzi. The work of Zhi-Quan Luo was supported in part by the National Key R&D Program of China under Grant 2018YFB1800800, in part by the National Natural Science Foundation of China under Grant 61731018, and in part by the Guangdong Provincial Key Laboratory of Big Data Computation Theories and Methods. The work of Tsung-Hui Chang was supported in part by Shenzhen Fundamental Research Fund under Grant JCYJ20190813171003723, and in part by NSFC, China under Grants 61731018 and 62071409. *(Corresponding author: Zhi-Quan Luo.)*

Zhiguo Wang is with the College of Mathematics, Sichuan University, Chengdu, Sichuan 610064, China, and also with Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: wangzg315@126.com).

Jiawei Zhang, Tsung-Hui Chang, and Zhi-Quan Luo are with the Chinese University of Hong Kong, Shenzhen 518172, China, and also with Shenzhen Research Institute of Big Data, Shenzhen, Guangdong Province 518172, China (e-mail: jiaweizhang2@link.cuhk.edu.cn; tsunghui.chang@ieee.org; luozq@umn.edu).

Jian Li is with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: li@dsp.ufl.edu).

Digital Object Identifier 10.1109/TSP.2021.3097211

multipliers (ADMM) [1], [4], [9] and the UDA in [10] are proposed. Recently, in [11], the authors propose a general unified algorithmic framework for such a class of distributed convex problems. For non-convex problems, the authors in [12] studied the convergence of proximal decentralized gradient descent (DGD) method with a diminishing step size. Based on the successive convex approximation (SCA) technique and the gradient tracking (GT) method, the authors in [13] proposed a network successive convex approximation (NEXT) algorithm for (1), and it is extended to more general scenarios with time varying networks and stronger convergence analysis results [14], [15]. In [16], based on an inexact augmented Lagrange method, a proximal primal-dual algorithm (Prox-PDA) is developed for (1) with smooth and non-convex $f_i(x)$ and without $r_i(x)$. A near-optimal algorithm xFilter is further proposed in [17] that can achieve the computation complexity lower bound $\mathcal{O}(1/\epsilon)$ of first-order distributed optimization algorithms when the target accuracy ϵ is moderate. To handle non-convex and non-smooth problems with polyhedral constraints, the authors of [18], [19] proposed a proximal augmented Lagrangian (AL) method for solving (1) by introducing a proximal variable and an exponential averaging scheme.

For streaming learning, the stochastic proximal gradient consensus method based on ADMM is proposed in [20] to solve (1) with convex objective functions. For non-convex problems, the decentralized parallel stochastic gradient descent (D-PSGD) [21] is applied to (1) (without $r_i(x)$) for training large-scale neural networks, and the convergence rate is analyzed. The analysis of D-PSGD relies on an assumption that $\frac{1}{N}\sum_{i=1}^{N}||\nabla f_i(x) - \nabla f(x)||^2$ is bounded, which implies that the variance of data distributions across the agents should be controlled. At federated scenario, a lot of algorithms use local stochastic gradient descent method to reduce the communication burden, such as, FedAvg [22], FedProx [23]. But FedAvg may suffer from "client-drift" when the data is non-identically independent distribution (non-iid). Recently, SCAFFOLD [24] use control variates (variance reduction) to reduce influence by data heterogeneity, which does not require the assumption that $\frac{1}{N}\sum_{i=1}^{N}||\nabla f_i(x) - \nabla f(x)||^2$ is bounded. In fact, these federated learning algorithms are limited to star networks with a central coordinator, unlike the presented decentralized algorithm which can work over any connected network graphs. In [25], the authors proposed an improved D-PSGD algorithm, called D^2, which also removes such assumption and is less sensitive to the data variance across agents. However, D^2 requires a restrictive assumption on the eigenvalue of the mixing matrix. This assumption is relaxed by the GNSD algorithm in [26], which essentially is a stochastic counterpart of the GT algorithm in [15]. We should emphasize here that the algorithms in [21], [25], [26] can only handle smooth problems without constraints and regularization terms. The work [27] proposed a multi-agent projected stochastic gradient decent (PSGD) algorithm for (1) but $r_i(x)$ is limited to the indicator function of compact convex sets. Besides, there is no convergence rate analysis in [27].

In this paper, we develop a new distributed stochastic optimization algorithm for the non-convex and non-smooth problem (1). The proposed algorithm is inspired by the proximal AL framework in [18] and has three new features. First, the proposed algorithm is a stochastic distributed algorithm that can be used either for streaming/online learning or batch/offline learning with mini-batch stochastic gradients, however, [18] only can handle the offline learning with full gradient. Second, the proposed algorithm can handle nonconvex problem (1) with non-smooth terms that have a polyhedral epigraph, which is more general than [18], [19] and [10] only studies the strongly-convex problem with a common non-smooth term. The key step that showing the convergence of the proposed method is to bound the dual variable. Third, the proposed algorithm incorporates the Nesterov momentum technique for fast convergence. The Nesterov momentum technique has been applied for accelerating the convergence of distributed optimization. For example, in [28], [29], the distributed gradient descent methods with the Nesterov momentum are proposed, and are shown to achieve the optimal iteration complexity for convex problems. In practice, since SGD with momentum often can converge faster, it is also commonly used to train deep neural networks [30], [31]. We note that [28]–[31] are for smooth problems. To the best of our knowledge, the Nesterov momentum technique has not been used for distributed non-convex and non-smooth optimization.

Our contributions are summarized as follows.

- We propose a new stochastic proximal primal dual algorithm with momentum (SPPDM) for non-convex and non-smooth problem (1) under the online/streaming setting. For the offline/batch setting where the full gradients of the local cost functions are available, SPPDM reduces to a deterministic algorithm, named the PPDM algorithm.
- Under the assumption that the epigraph of the non-smooth functions is a polyhedron, we show that the proposed SPPDM and PPDM can achieve an ϵ-stationary solution of (1) under a constant step size with computation complexities of $\mathcal{O}(1/\epsilon^2)$ and $\mathcal{O}(1/\epsilon)$, respectively, while both have a communication complexity of $\mathcal{O}(1/\epsilon)$.
  The convergence analysis neither requires assumption on the boundedness of $\frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(x) - \nabla f(x)\|^2$ nor on the eigenvalues of the mixing matrix.
- As shown in Table I, the proposed SPPDM/PPDM algorithms have the same order of computation complexity as the existing methods and lower order of communication complexity when compared with the existing GT based methods.
- Numerical experiments for a distributed non-convex regression problem and a deep neural network (DNN) based classification problem show that the proposed algorithms outperforms the existing methods.

Notation: We denote $\mathbf{I}_n$ as the n by n identity matrix and $\mathbf{1}$ as the all-one vector, i.e., $\mathbf{1} = [1, \ldots, 1]^\top$. $\langle \mathbf{a}, \boldsymbol{b} \rangle$ represents the inner product of vectors $\mathbf{a}$ and $\boldsymbol{b}$, $\|\mathbf{a}\|$ is the Euclidean norm of vector $\mathbf{a}$ and $\|\mathbf{a}\|_1$ is the ℓ_1-norm of vector $\mathbf{a}$; $\otimes$ denotes the Kronecker product. For a matrix $\mathbf{A}$, $\sigma_A > 0$ denotes its largest singular value. $\mathrm{diag}\{a_1, \ldots, a_N\}$ denotes a diagonal matrix with $a_1, \ldots, a_N$ being the diagonal entries while $\mathrm{diag}\{\mathbf{A}_1, \ldots, \mathbf{A}_N\}$ denotes a block diagonal matrices with each $\mathbf{A}_i$ being the ith

TABLE I
COMPARISONS OF DIFFERENT ALGORITHMS

Algorithm	objective function	non-smooth $r(\mathbf{x})$	gradient	stepsize	momentum	computational	communication
D-PSGD [21]	$f(\mathbf{x})$	✗	stochastic	decreasing	✗	$\mathcal{O}(\frac{N}{\epsilon^2})$	$\mathcal{O}(\frac{1}{\epsilon^2})$
D^2 [25]	$f(\mathbf{x})$	✗	stochastic	decreasing	✗	$\mathcal{O}(\frac{N}{\epsilon^2})$	$\mathcal{O}(\frac{1}{\epsilon^2})$
GNSD [26]	$f(\mathbf{x})$	✗	stochastic	decreasing	✗	$\mathcal{O}(\frac{N}{\epsilon^2})$	$\mathcal{O}(\frac{1}{\epsilon^2})$
PR-SGD-M [31]	$f(\mathbf{x})$	✗	stochastic	decreasing	✓	$\mathcal{O}(\frac{N}{\epsilon^2})$	$\mathcal{O}(\frac{1}{\epsilon^2})$
PSGD [27]	$f(\mathbf{x})+r(\mathbf{x})$	compact convex set	stochastic	decreasing	✗	✗	✗
NEXT[13]	$f(\mathbf{x})+r(\mathbf{x})$	convex	full	decreasing	✗	✗	✗
Prox-PDA [16]	$f(\mathbf{x})$	✗	full	fixed	✗	$\mathcal{O}(\frac{mN}{\epsilon})$	$\mathcal{O}(\frac{1}{\epsilon})$
Prox-DGD [12]	$f(\mathbf{x})+r(\mathbf{x})$	nonconvex	full	decreasing	✗	✗	✗
Prox-ADMM [18]	$f(\mathbf{x})+r(\mathbf{x})$	box set	full	fixed	✗	$\mathcal{O}(\frac{mN}{\epsilon})$	$\mathcal{O}(\frac{1}{\epsilon})$
Proposed[†]	$f(\mathbf{x})+r(\mathbf{x})$	epigraph is polyhedral	full	fixed	✓	$\mathcal{O}(\frac{mN}{\epsilon})$	$\mathcal{O}(\frac{1}{\epsilon})$
			stochastic	fixed	✓	$\mathcal{O}(\frac{N}{\epsilon^2})$	$\mathcal{O}(\frac{1}{\epsilon})$

[*] To obtain a fair comparison, we need to establish the relation between the proposed measure and the other measures in Remark 4.

[†] The proposed method requires that the batch size $|\mathcal{I}|$ is proportional to ϵ.

block diagonal matrix. $[\mathbf{A}]_{ij}$ represents the element of $\mathbf{A}$ in the ith row and jth column.

For problem (1), we denote $\mathbf{x} = [x_1^\top, \ldots, x_N^\top]^\top \in \mathbb{R}^{Nn}$, $f(\mathbf{x}) = \sum_{i=1}^N f_i(x_i)$, and $r(\mathbf{x}) = \sum_{i=1}^N r_i(x_i)$. The gradient of $f(\cdot)$ at $\mathbf{x}$ is denoted by

$$\nabla f(\mathbf{x}) = [(\nabla f_1(x_1))^\top, \ldots, (\nabla f_N(x_N))^\top]^\top,$$

where $\nabla f_i(x_i)$ is the gradient of f_i at x_i. In the online/streaming setting, we denote the stochastic gradient estimates of agents as

$$G(\mathbf{x}, \boldsymbol{\xi}) = [(G_i(x_1, \xi_1))^\top, \ldots, (G_N(x_N, \xi_N))^\top]^\top,$$

where $\boldsymbol{\xi} = [\xi_1^\top, , \ldots, \xi_N^\top]$. Lastly, we define the following proximal operator of r_i

$$\mathrm{prox}_{r_i}^{\alpha}(x) = \arg\min_u \frac{\alpha}{2}\|x-u\|^2 + r_i(u),$$

where α is a parameter.

Synopsis: In Section II, the proposed SPPDM and PPDM algorithms are presented and their connections with existing methods are discussed. Based on an inexact stochastic primal-dual framework, it is shown how the SPPDM and PPDM algorithms are devised. Section III presents the theoretical results of the convergence conditions and convergence rate of the SPPDM and PPDM algorithms. The performance of the SPPDM and PPDM algorithms are illustrated in Section IV. Lastly, the conclusion is given in Section V.

II. ALGORITHM DEVELOPMENT

A. Network Model and Consensus Formulation

Let us denote the multi-agent network as a graph $\mathcal{G}$, which contains a node set $V := \{1, \ldots, N\}$ and an edge set $\mathcal{E}$ with cardinality $|\mathcal{E}|$ and it does not have repeat edge. For each agent i, it has neighboring agents in the subset $\mathcal{N}_i := \{j \in V | (i,j) \in \mathcal{E}\}$ with size $d_i \geq 1$. It is assumed that each agent i can communicate with its neighborhood $\mathcal{N}_i$. We also assume that the graph $\mathcal{G}$ is undirected and is connected in the sense that for any of two agents in the network there is a path connecting them through the edge links. Thus, problem (1) can be equivalently written as

$$\min_{\substack{x_i \\ i=1,\ldots,N}} \sum_{i=1}^N (f_i(x_i) + r_i(x_i)) \tag{4a}$$

$$\text{s.t. } x_i = x_j, \forall (i,j) \in \mathcal{E}. \tag{4b}$$

Let us introduce the incidence matrix $\tilde{\mathbf{A}} \in \mathbb{R}^{|\mathcal{E}| \times N}$ which has $\tilde{\mathbf{A}}(\ell, i) = 1$ and $\tilde{\mathbf{A}}(\ell, j) = -1$ if $(i,j) \in \mathcal{E}$ with $j > i$, and zero otherwise, for $\ell = 1, \ldots, |\mathcal{E}|$. Define the extended incidence matrix as $\mathbf{A} := \tilde{\mathbf{A}} \otimes \mathbf{I}_n$. Then (4) is equivalent to

$$\min_{\mathbf{x}} f(\mathbf{x}) + r(\mathbf{x}) \tag{5a}$$

$$\text{s.t. } \mathbf{A}\mathbf{x} = \mathbf{0}. \tag{5b}$$

B. Proposed SPPDM and PPDM Algorithm

In this section, we present the proposed SPPDM algorithm for solving (5) under the online/streaming setting in (3). The algorithm steps are outlined in Algorithm 1. Before showing how the algorithm is developed in Section II-C, let us make a few comments about SPPDM.

In Algorithm 1, $\alpha, \beta, \gamma, c, \kappa, \eta$ are some positive constant parameters that depend on the problem instance (such as the Lipschitz constants of $\{\nabla f_i\}$) and the graph Laplacian matrix). Equations (7)-(10) are the updates performed by each agent i within the kth communication round, for $k = 1, 2, \ldots,$ and $i = 1, \ldots, N$. Specifically, step (7) is the introduced Nesterov momentum term s_i^k for accelerating the algorithm convergence, where η is the extrapolation coefficient at iteration k. Step (8) shows how the neighboring variables $\{x_j\}_{j \in \mathcal{N}_i}$ are used for local gradient update. Note here that in SPPDM the agent uses the sample average $\frac{1}{|\mathcal{I}|}\sum_{j=1}^{|\mathcal{I}|} G_i(s_i^k, \xi_{ij}^k)$ to approximate $\nabla f_i(s_i^k)$, where $\xi_{ij}^k \sim \mathcal{B}_i, j = 1, \ldots, |\mathcal{I}|$, denotes the samples drawn by agent i in the kth iteration. Besides, in (8), both approximate gradients at s_i^k and s_i^{k-1} are used. Step (9) performs the proximal gradient update with respect to the regularization term $r_i(x)$. In step (8), the variable $\{z_i^k\}$ is a "proximal" variable introduced

Algorithm 1: Proposed SPPDM Algorithm.

```
Given parameters $\alpha, \beta, \gamma, c, \kappa, \eta_k$ and initial values of $x_i^0$
 and $z_i^0 = x_i^0$, $i = 1, \ldots, N$. Let
        $\psi_i = \gamma + 2cd_i + \kappa$        (6)
and set $s_i^0 = x_i^0$, $i = 1, \ldots, N$. Do
    $x_i^{\frac{1}{2}} = (\gamma + cd_i + \kappa)\frac{x_i^0}{\psi_i} + \frac{c}{\psi_i}\sum_{j\in\mathcal{N}_i} x_j^0 - \frac{1}{\psi_i}\nabla f_i(x_i^0),$
    $x_i^1 = \mathrm{prox}_{r_i}^{\alpha_i}\left(x_i^{\frac{1}{2}}\right), i = 1, \ldots, N.$
for communication round $k = 1, 2, \ldots$ do
  for agent $i = 1, 2, \ldots, N$ (in parallel) do
      $s_i^k = x_i^k + \eta(x_i^k - x_i^{k-1}),$        (7)
      $x_i^{k+\frac{1}{2}} = x_i^{k-1+\frac{1}{2}} + \frac{d_i}{\psi_i}((c-\alpha)x_i^k - cx_i^{k-1}),$        (8)
          $+ \frac{1}{\psi_i}\sum_{j\in\mathcal{N}_i}((c+\alpha)x_j^k - cx_j^{k-1})$
          $+ \frac{1}{\psi_i}\left(\gamma(s_i^k - s_i^{k-1}) + \kappa(z_i^k - z_i^{k-1})\right)$
          $- \frac{1}{\psi_i|\mathcal{I}|}\sum_{j=1}^{|\mathcal{I}|}(G_i(s_i^k, \xi_{ij}^k) - G_i(s_i^{k-1}, \xi_{ij}^{k-1})),$
      $x_i^{k+1} = \mathrm{prox}_{r_i}^{\psi_i}\left(x_i^{k+\frac{1}{2}}\right),$        (9)
      $z_i^{k+1} = z_i^k + \beta(x_i^{k+1} - z_i^k).$        (10)
  end for
end for
```

for overcoming the non-convexity of f_i (see (19)), and is updated as in step (10).

By stacking the variables for all $i = 1, \ldots, N$, one can write (7)-(10) in a vector form. Specifically, step (8) for $i = 1, \ldots, N$, can be expressed compactly as

$$\begin{aligned}\mathbf{x}^{k+\frac{1}{2}} &= \mathbf{x}^{k-1+\frac{1}{2}} + \mathbf{U}\mathbf{x}^k - \tilde{\mathbf{U}}\mathbf{x}^{k-1}\\ &\quad + \gamma\boldsymbol{\Psi}^{-1}(\mathbf{s}^k - \mathbf{s}^{k-1}) + \kappa\boldsymbol{\Psi}^{-1}(\mathbf{z}^k - \mathbf{z}^{k-1})\\ &\quad - \boldsymbol{\Psi}^{-1}(\bar{G}(\mathbf{s}^k, \boldsymbol{\xi}^k) - \bar{G}(\mathbf{s}^{k-1}, \boldsymbol{\xi}^{k-1})), \end{aligned} \tag{11}$$

where $\mathbf{U} = \mathbf{U}^1 \otimes \mathbf{I}_n \in \mathbb{R}^{Nn\times Nn}$ and $\tilde{\mathbf{U}} = \tilde{\mathbf{U}}^1 \otimes \mathbf{I}_n \in \mathbb{R}^{Nn\times Nn}$ are two matrices satisfying

$$[\mathbf{U}]_{ij}^1 = \begin{cases} \frac{d_i}{\psi_i}(c-\alpha), & i = j,\\ \frac{c+\alpha}{\psi_i}, & i \neq j \text{ and } (i,j)\in\mathcal{E},\\ 0, & \text{otherwise.}\end{cases} \tag{12}$$

$$[\tilde{\mathbf{U}}]_{ij}^1 = \begin{cases} \frac{d_i c}{\psi_i}, & i = j,\\ \frac{c}{\psi_i}, & i \neq j \text{ and } (i,j)\in\mathcal{E},\\ 0 & \text{otherwise.}\end{cases} \tag{13}$$

for all $i, j = 1, \ldots, N$, $\boldsymbol{\Psi} = \boldsymbol{\Psi}^1 \otimes \mathbf{I}_n \in \mathbb{R}^{Nn\times Nn}$ is a diagonal matrix and the ith element of $\boldsymbol{\Psi}^1$ being $\psi_i := \gamma + 2cd_i + \kappa$ for $i = 1, \ldots, N$, and

$$\bar{G}(\mathbf{s}^k, \boldsymbol{\xi}^k) := \frac{1}{|\mathcal{I}|}\sum_{j=1}^{|\mathcal{I}|} G(\mathbf{s}^k, \boldsymbol{\xi}_j^k). \tag{14}$$

When the full gradients ∇f_i are available under the offline/batch setting, the approximate gradient G_i in (8) and (11) can be replaced by ∇f_i. Then, the SPPDM algorithm reduces to the PPDM algorithm.

Remark 1: We show that the PPDM algorithm can have a close connection with the PG-EXTRA algorithm in [8]. Specifically, let us set $\eta = 0$ (no Nesterov momentum) and $\beta = 1$ (no proximal variable). Then, we have $s_i^k = z_i^k = x_i^k$ for all k, i, and the momentum and proximal variable update in (7) and (10) can be removed. As a result, (11) reduces to

$$\begin{aligned}\mathbf{x}^{k+\frac{1}{2}} &= \mathbf{x}^{k-1+\frac{1}{2}} + \mathbf{W}\mathbf{x}^k - \tilde{\mathbf{W}}\mathbf{x}^{k-1}\\ &\quad - \boldsymbol{\Psi}^{-1}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1})), \end{aligned} \tag{15}$$

where $\mathbf{W} = \mathbf{U} + (\gamma + \kappa)\boldsymbol{\Psi}^{-1}$ and $\tilde{\mathbf{W}} = \tilde{\mathbf{U}} + (\gamma + \kappa)\boldsymbol{\Psi}^{-1}$. One can see that (15) and (9) have an identical form as the PG-EXTRA algorithm in [8, Eqn. (3a)-(3b)]. Therefore, the proposed PPDM algorithm can be regarded as an accelerated version of the PG-EXTRA with extra capability to handle non-convex problems. One should note that, unlike (12) and (13), the PG-EXTRA allows a more flexible choice of the mixing matrix $\mathbf{W}$, and thus it is also closely related to the GT based methods [5].

Remark 2: The PPDM algorithm also has a close connection with the distributed Nesterov gradient (D-NG) algorithm in [28]. Specifically, let us set $\alpha = c$ and $\beta = 1$ (no proximal variable) and remove the non-smooth regularization term $r(\mathbf{x})$. Then, we have $z_i^k = x_i^k$ for all k, i, and the proximal gradient update (9) and the proximal variable update (10) can be removed. Under the setting, as shown in [32], one can write (11) of the PPDM algorithm as

$$\mathbf{x}^{k+1} = \tilde{\mathbf{W}}\mathbf{s}^k - \boldsymbol{\Psi}^{-1}\nabla f(\mathbf{s}^k) + \mathbf{C}^k, \tag{16}$$

where $\mathbf{C}^k = (\tilde{\mathbf{U}}(\mathbf{x}^k - \mathbf{s}^k) + \kappa(\mathbf{x}^k - \mathbf{s}^k)\boldsymbol{\Psi}^{-1}) - \sum_{t=0}^k(\mathbf{I} - \tilde{\mathbf{W}})\mathbf{x}^t$ can regarded as a cumulative correction term. Note that the D-NG algorithm in [28, Eqn. (2)-(3)] is

$$\mathbf{s}^k = \mathbf{x}^k + \eta(\mathbf{x}^k - \mathbf{x}^{k-1}), \tag{17}$$

$$\mathbf{x}^{k+1} = \tilde{\mathbf{W}}\mathbf{s}^k - \boldsymbol{\Psi}^{-1}\nabla f(\mathbf{s}^k). \tag{18}$$

One can see that (18) and (16) have a similar form except for the correction term. Note that the convergence of the D-NG algorithm is proved in [28] only for convex problems with a diminishing step size. Therefore, the proposed PPDM algorithm is an enhanced counterpart of the D-NG algorithm with the ability to handle non-convex and non-smooth problems.

C. Algorithm Development

In this subsection, let us elaborate how the SPPDM algorithm is devised. Our proposed algorithm is inspired by the proximal

AL framework in [18]. First, we introduce a proximal term $\mathbf{z}$ to (5) as

$$\min_{\mathbf{x},\mathbf{z}} f(\mathbf{x}) + r(\mathbf{x}) + \frac{\kappa}{2}\|\mathbf{x}-\mathbf{z}\|^2 \tag{19a}$$

$$\text{s.t. } \mathbf{A}\mathbf{x} = 0, \tag{19b}$$

where $\kappa > 0$ is a parameter. Obviously, (19) is equivalent to (5). The purpose of adding the proximal term $\frac{\kappa}{2}\|\mathbf{x}-\mathbf{z}\|^2$ is to make the objective function in (19a) strongly convex with respect to $\mathbf{x}$ when $\mathbf{z}$ is fixed and $\kappa > 0$ is large enough (however, note that (19a) is not jointly convex with respect to $(x^T, z^T)^T$). Such strong convexity will be exploited for building the algorithm convergence.

Second, let us consider the AL function of (19) as follows

$$L_c(\mathbf{x},\mathbf{z};\boldsymbol{\lambda}) = f(\mathbf{x}) + r(\mathbf{x}) + \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{x}\rangle + \frac{c}{2}\|\mathbf{A}\mathbf{x}\|^2 + \frac{\kappa}{2}\|\mathbf{x}-\mathbf{z}\|^2, \tag{20}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^{|\mathcal{E}|n}$ is the Lagrangian dual variable, and $c > 0$ is a positive penalty parameter. Then, the Lagrange dual problem of (19) can be expressed as

$$\max_{\boldsymbol{\lambda}} \min_{\mathbf{x},\mathbf{z}} L_c(\mathbf{x},\mathbf{z};\boldsymbol{\lambda}). \tag{21}$$

We apply the following inexact stochastic primal-dual updates with momentum for problem (21): for $k = 0, 1, 2, \ldots$,

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \alpha \mathbf{A}\mathbf{x}^k, \tag{22}$$

$$\mathbf{s}^k = \mathbf{x}^k + \eta(\mathbf{x}^k - \mathbf{x}^{k-1}), \tag{23}$$

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} g(\mathbf{x},\mathbf{x}^k,\mathbf{s}^k,\mathbf{z}^k,\boldsymbol{\xi}^k;\boldsymbol{\lambda}^{k+1}), \tag{24}$$

$$\mathbf{z}^{k+1} = \mathbf{z}^k + \beta(\mathbf{x}^{k+1} - \mathbf{z}^k). \tag{25}$$

Specifically, (22) is the dual ascent step with $\alpha > 0$ being the dual step size. In (23), the momentum variable $\mathbf{s}^k$ is introduced for the primal variable $\mathbf{x}$. Let $\mathbf{p}^k = \mathbf{A}^\top \boldsymbol{\lambda}^k$, (22) can be replaced by

$$\mathbf{p}^{k+1} = \mathbf{p}^k + \alpha \mathbf{A}^\top \mathbf{A}\mathbf{x}^k. \tag{26}$$

To update $\mathbf{x}$, we consider the inexact step as in (24) where $g(\mathbf{x},\mathbf{x}^k,\mathbf{s}^k,\mathbf{z}^k,\boldsymbol{\xi}^k;\boldsymbol{\lambda}^{k+1})$ is a surrogate function given by

$$\begin{aligned} &g(\mathbf{x},\mathbf{x}^k,\mathbf{s}^k,\mathbf{z}^k,\boldsymbol{\xi}^k;\boldsymbol{\lambda}^{k+1}) \\ &= \underbrace{f(\mathbf{s}^k) + \langle \bar{G}(\mathbf{s}^k,\boldsymbol{\xi}^k), \mathbf{x}-\mathbf{s}^k\rangle + \frac{\gamma}{2}\|\mathbf{x}-\mathbf{s}^k\|^2}_{\text{(a)}} \\ &\quad + r(\mathbf{x}) + \langle \mathbf{p}^{k+1}, \mathbf{x}\rangle \\ &\quad + \underbrace{\frac{c}{2}\|\mathbf{A}\mathbf{x}\|^2 + \frac{c}{2}\|\mathbf{x}-\mathbf{x}^k\|^2_{\mathbf{B}^\top\mathbf{B}}}_{\text{(b)}} + \frac{\kappa}{2}\|\mathbf{x}-\mathbf{z}^k\|^2. \end{aligned} \tag{27}$$

In (27), the term (a) is a quadratic approximation of f at $\mathbf{s}^k$ using the stochastic gradient $\bar{G}$, where $\gamma > 0$ is a parameter. In term (b) of (27), $\mathbf{B}$ is the signless incidence matrix of the graph $\mathcal{G}$, i.e., $\mathbf{B} = |\mathbf{A}|$, which satisfies $\mathbf{A}^\top\mathbf{A} + \mathbf{B}^\top\mathbf{B} = 2\mathbf{D}$, where $\mathbf{D} = \mathrm{diag}\{d_1, \ldots, d_N\}$ is the degree matrix of $\mathcal{G}$. As shown in [16], the introduction of $\frac{c}{2}\|\mathbf{x}-\mathbf{x}^k\|^2_{\mathbf{B}^\top\mathbf{B}}$ can "diagonalize" $\frac{c}{2}\|\mathbf{A}\mathbf{x}\|^2$ and lead to distributed implementation of (24). In particular, one can show that (24) with (27) can be expressed as

$$\mathbf{x}^{k+1} = \mathrm{prox}_r^{\boldsymbol{\Psi}}\left(\boldsymbol{\Psi}^{-1}(\gamma\mathbf{s}^k + c\mathbf{B}^\top\mathbf{B}\mathbf{x}^k + \kappa\mathbf{z}^k - \bar{G}(\mathbf{s}^k,\boldsymbol{\xi}^k) - \mathbf{p}^{k+1})\right). \tag{28}$$

As seen, due to the graphical structure of $\mathbf{B}^\top\mathbf{B}$, each x_i^{k+1} in (28) can be obtained in a distributed fashion using only x_j^k, $j \in \boldsymbol{\mathcal{N}}_i$ from its neighbors. Lastly, we update $\mathbf{z}$ by applying the gradient descent to $L_c(\mathbf{x}^{k+1},\mathbf{z};\boldsymbol{\lambda}^{k+1})$ with step size β, which then yields (25).

To show how (7)-(10) are obtained, let us define

$$\mathbf{x}^{k+\frac{1}{2}} = \boldsymbol{\Psi}^{-1}\left(\gamma\mathbf{s}^k + c\mathbf{B}^\top\mathbf{B}\mathbf{x}^k + \kappa\mathbf{z}^k - \bar{G}(\mathbf{s}^k,\boldsymbol{\xi}^k) - \mathbf{p}^{k+1}\right). \tag{29}$$

Then, (28) can be written as

$$\mathbf{x}^{k+1} = \mathrm{prox}_r^{\boldsymbol{\Psi}}\left(\mathbf{x}^{k+\frac{1}{2}}\right). \tag{30}$$

Moreover, by subtracting $\mathbf{x}^{k-1+\frac{1}{2}}$ from $\mathbf{x}^{k+\frac{1}{2}}$, one obtains

$$\begin{aligned} \mathbf{x}^{k+\frac{1}{2}} &= \mathbf{x}^{k-1+\frac{1}{2}} + \gamma\boldsymbol{\Psi}^{-1}(\mathbf{s}^k - \mathbf{s}^{k-1}) + \kappa\boldsymbol{\Psi}^{-1}(\mathbf{z}^k - \mathbf{z}^{k-1}) \\ &\quad + c\boldsymbol{\Psi}^{-1}\mathbf{B}^\top\mathbf{B}(\mathbf{x}^k - \mathbf{x}^{k-1}) - \boldsymbol{\Psi}^{-1}(\mathbf{p}^{k+1} - \mathbf{p}^k) \\ &\quad - \boldsymbol{\Psi}^{-1}(\bar{G}(\mathbf{s}^k,\boldsymbol{\xi}^k) - \bar{G}(\mathbf{s}^{k-1},\boldsymbol{\xi}^{k-1})). \end{aligned} \tag{31}$$

After substituting (26) into (31), we obtain

$$\begin{aligned} \mathbf{x}^{k+\frac{1}{2}} &= \mathbf{x}^{k-1+\frac{1}{2}} + \mathbf{U}\mathbf{x}^k - \tilde{\mathbf{U}}\mathbf{x}^{k-1} \\ &\quad + \gamma\boldsymbol{\Psi}^{-1}(\mathbf{s}^k - \mathbf{s}^{k-1}) + \kappa\boldsymbol{\Psi}^{-1}(\mathbf{z}^k - \mathbf{z}^{k-1}) \\ &\quad - \boldsymbol{\Psi}^{-1}(\bar{G}(\mathbf{s}^k,\boldsymbol{\xi}^k) - \bar{G}(\mathbf{s}^{k-1},\boldsymbol{\xi}^{k-1})), \end{aligned} \tag{32}$$

which is exactly (11) since $\mathbf{U} = c\boldsymbol{\Psi}^{-1}\mathbf{B}^\top\mathbf{B} - \alpha\boldsymbol{\Psi}^{-1}\mathbf{A}^\top\mathbf{A}$ and $\tilde{\mathbf{U}} = c\boldsymbol{\Psi}^{-1}\mathbf{B}^\top\mathbf{B}$ by (12) and (13), respectively.

In summary, (22) and (24) can be equivalently written as (32) and (30), and therefore we obtain (23), (32), (30) and (25) as the algorithm updates, which correspond to (7)–(10) in Algorithm 1.

Before ending the section, we remark that it is possible to employ the existing stochastic primal-dual methods such as [33] for solving the non-smooth and non-convex problem (5). However, these methods require strict conditions on $\mathbf{A}$. For example, the stochastic ADMM method in [33] requires $\mathbf{A}$ to have full rank, which cannot hold for the distributed optimization problem (5) since the graph incidence matrix $\mathbf{A}$ for a connected graph must be rank deficient.

III. Convergence Analysis

In this section, we present the main theoretical results of the proposed SPPDM and PPDM algorithms by establishing their convergence conditions and convergence rate.

A. Assumptions

We first make some proper assumptions on problem (5).

Assumption 1:

1) The function $f(\mathbf{x})$ is a continuously differentiable function with Lipschitz continuous gradients, i.e., for constant $L > 0$,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \tag{33}$$

for all $\mathbf{x}, \mathbf{y}$. Moreover, assume that there exists a constant $\mu \geq -L$ (possibly negative) such that

$$f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2, \tag{34}$$

for all $\mathbf{x}, \mathbf{y}$.

2) The objective function $f(\mathbf{x}) + r(\mathbf{x})$ is bounded from below in the feasible set $\{\mathbf{x}|\mathbf{A}\mathbf{x} = 0\}$, i.e.,

$$f(\mathbf{x}) + r(\mathbf{x}) > \underline{f} > -\infty,$$

for some constant $\underline{f}$.

Assumption 2: [34] The epigraph of each $r_i(x_i)$, i.e., $\{(x_i, y_i)|r_i(x_i) \leq y_i\}$, is a polyhedral set and has a compact form as

$$S_{x,i}x_i + S_{y,i}y_i \geq \zeta_i, \tag{35}$$

where $S_{x,i} \in \mathbb{R}^{q_i \times n}$, $S_{y,i} \in \mathbb{R}^{q_i}$ and $\zeta_i \in \mathbb{R}^{q_i}$ are some constant matrix and vectors.

There are many useful functions that satisfy Assumptions 1, such as the sigmoid function $\frac{1}{1+\exp^{-x}}$, which is used as activation functions in neural networks. In addition, the common L_1 regularization term $\|x\|_1$ satisfies Assumption 2, which is widely used for obtaining sparse solutions.

For online/streaming learning, we also make the following standard assumptions that the gradient estimates are unbiased and have a bounded variance.

Assumption 3: [21], [25], [26] The stochastic gradient estimate $G_i(x, \xi)$ satisfies

$$\mathbb{E}[G_i(x, \xi)] = \nabla f_i(x) \tag{36}$$

$$\mathbb{E}[\|G_i(x, \xi) - \nabla f_i(x)\|^2] \leq \sigma^2, \tag{37}$$

for all x, where $\sigma > 0$ is a constant, and the expectation $\mathbb{E}$ is with respect to the random sample $\xi \sim \mathcal{B}_i$.

It is easy to check that the gradient estimate of the mini-batch samples satisfies

$$\mathbb{E}\left[\left\|\frac{1}{|\mathcal{I}|}\sum_{j=1}^{|\mathcal{I}|} G_i(x, \xi_j) - \nabla f_i(x)\right\|^2\right] \leq \sigma^2/|\mathcal{I}|. \tag{38}$$

B. Convergence Analysis of SPPDM

We define the following term

$$Q(\mathbf{x}, \boldsymbol{\lambda}) = \|\mathbf{x} - \operatorname{prox}_r^1(\mathbf{x} - \nabla f(\mathbf{x}) - \mathbf{A}^\top\boldsymbol{\lambda})\|^2 + \|\mathbf{A}\mathbf{x}\|^2 \tag{39}$$

as the optimally gap for a primal-dual solution $(\mathbf{x}, \boldsymbol{\lambda})$ of problem (5). Obviously, one can shown that when $Q(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) = 0$, $(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)$ is a KKT solution of (5). We define that $(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)$ is an ϵ-stationary solution of (5) if $Q(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) < \epsilon$.

The convergence result is stated in the following theorem.

Theorem 1: Assume that Assumptions 1-3 hold true, and let parameters satisfy

$$\kappa > -\mu, \gamma > 3L, \tag{40}$$

$$\eta \leq \sqrt{\frac{\kappa + 2c + \gamma - 3L}{2(\gamma - \mu + 3L)}} := \bar{\eta}, \tag{41}$$

moreover, let $0 < \alpha \leq c$ and $\beta > 0$ be both sufficiently small (see (84) and (85)). Then, for a sequence $\{\mathbf{x}^k, \mathbf{z}^k, \boldsymbol{\lambda}^k\}$ generated by Algorithm 1, it holds that

$$\min_{k=0,\ldots,K-1} \mathbb{E}[Q(\mathbf{x}^k, \boldsymbol{\lambda}^{k+1})] \leq C_0\left(\frac{\phi^0 - \underline{f}}{K} + \frac{C_1 N\sigma^2}{|\mathcal{I}|}\right), \tag{42}$$

where C_0 and C_1 are some positive constants depending on the problem parameters (see (101) and (87)). In addition, ϕ^0 is a constant defined in (68).

To prove Theorem 1, the key is to define a novel stochastic potential function $\mathbb{E}[\phi^{k+1}]$ in (68) and analyze the conditions for which $\mathbb{E}[\phi^{k+1}]$ descends monotonically with the iteration number k (Lemma 6). To achieve the goal, several approximation error bounds for the primal variable $\mathbf{x}^k$ (Lemma 2) and the dual variable $\boldsymbol{\lambda}^k$ (Lemma 4) are derived. Interested readers may refer to Appendix B for the details.

By Theorem 1, we immediately obtain the following corollary.

Corollary 1: Let

$$|\mathcal{I}| \geq \frac{2NC_0C_1\sigma^2}{\epsilon} \text{and} K \geq \frac{2C_0(\phi^0 - \underline{f})}{\epsilon}. \tag{43}$$

Then,

$$\min_{k=0,\ldots,K-1} \mathbb{E}[Q(\mathbf{x}^k, \boldsymbol{\lambda}^{k+1})] \leq \epsilon, \tag{44}$$

that is, an ϵ-stationary solution of problem (5) can be obtained in an expected sense.

Remark 3: Given a mini-batch size $|\mathcal{I}| = \mathcal{O}(1/\epsilon)$, Corollary 1 implies that the proposed SPPDM algorithm has the convergence rate of $\mathcal{O}(1/\epsilon)$ to obtain an ϵ-stationary solution. In each communication round k, each agent should receive the information $x_j^k, j \in \mathcal{N}_i$. Assume $x_j^k \in \mathbb{R}^n$ and the neighboring agents in the subset $\mathcal{N}_i := \{j \in V|(i,j) \in \mathcal{E}\}$ have size d_i. As a result, the corresponding total communication complexity of the SPPDM algorithm is proportional to $\frac{2\sum_{i=1}^N d_i}{\epsilon}$ while the computational complexity is $\mathcal{O}(N|\mathcal{I}|/\epsilon) = \mathcal{O}(N/\epsilon^2)$. As shown in Table I, the communication complexity $\mathcal{O}(1/\epsilon)$ of the SPPDM algorithm is smaller than $\mathcal{O}(1/\epsilon^2)$ of D-PSGD [21], D^2 [25], GNSD [26] and R-SGD-M [31].

Remark 4: Our definition of ϵ-stationary point in (39) is a sufficient condition for reaching a $\frac{2}{N}(1 + \frac{L^2}{\sigma_{\min}})\epsilon$-stationary point of D-PSGD, D^2, PR-SGD-M and a $\frac{1}{N}(1 + \frac{1}{\sigma_{\min}})\epsilon$-stationary point of GNSD, where $\sigma_{\min}$ is the smallest non-zero eigenvalue of $\mathbf{A}^T\mathbf{A}$. The proof is shown in the Appendix A.

C. Convergence Analysis of PPDM

When the full gradient $\nabla f(\mathbf{x}^k)$ is available for the PPDM algorithm, one can deduce a similar convergence result.

Theorem 2: Assume Assumptions 1-2 and the same conditions in (40), (41), (85) and (84) hold true.

- Every limit point of the sequence $\{\mathbf{x}^k, \mathbf{z}^k, \boldsymbol{\lambda}^k\}$ generated by the PPDM algorithm is a KKT solution of (5).
- Given $K \geq \frac{C_0(\phi^0 - \underline{f})}{\epsilon}$, we have

$$\min_{k=0,\ldots,K-1} Q(\mathbf{x}^k, \boldsymbol{\lambda}^{k+1}) \leq C_0 \left(\frac{\phi^0 - \underline{f}}{K} \right) \leq \epsilon.$$

The proof is presented in Appendix C.

To our knowledge, Theorem 1 and Theorem 2 are the first results that show the $\mathcal{O}(1/\epsilon)$ communication complexity of the distributed primal-dual method with momentum for non-convex and non-smooth problems. In both PPDM and SPPDM, the number of the total communication round is in the same order as the total number of iterations. Comparing with PPDM, SPPDM randomly draws $\mathcal{I}$ samples to obtain a mini-batch gradient. When the target accuracy ϵ is moderate and a batch size $|\mathcal{I}| = 1/\epsilon$ is proportional to full sample size, Corollary 1 shows that stochastic methods and deterministic methods have the same communication complexity. Numerical results in the next section will demonstrate that the SPPDM and PPDM algorithms can exhibit favorable convergence behavior than the existing methods.

IV. Numerical Results

In this section, we examine the numerical performance of the proposed SPPDM/PPDM algorithm and present comparison results with the existing methods.

A. Distributed Non-Convex Truncated Losses

In this example, we use a real California housing data to train a linear model. Consider the following distributed regression problem with a nonconvex truncated loss [35]

$$\min_x \sum_{i=1}^{N} \left(f_i(x) + \varsigma_i \|x\|_1 \right), \tag{45}$$

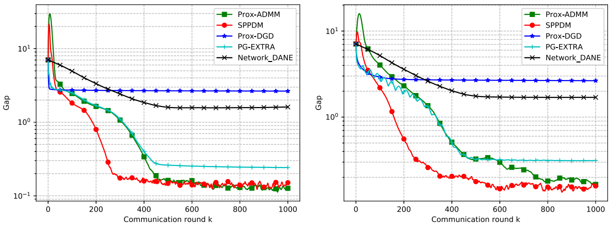
where

$$f_i(x) = \frac{\rho}{m} \sum_{j=1}^{m} \log \left(1 + \frac{\|y_j - h_j^\top x\|^2}{2\rho} \right),$$

and ρ is a parameter to determine the truncation level. (h_j, y_j) is the input data. We set $m = 688$, $n = 8$ and $\rho = 5$. Moreover, we consider a Erdos Renyi graph and a ring graph with $N = 30$ agents. The probability for Erdos Renyi graph is 0.3.

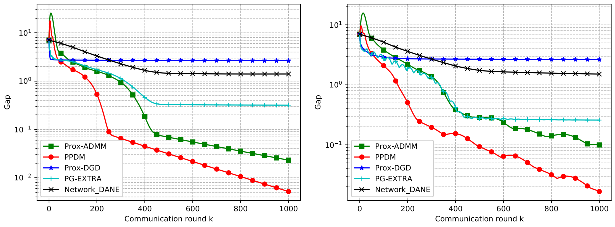
We compare the PPDM algorithm with Prox-DGD [12], PG-EXTRA [8], Prox-ADMM [18] and Network-DANE [36]. Note that theoretically PG-EXTRA and Network-DANE are not guaranteed to converge for the non-convex problem (5). We implement these two methods simply for comparison purpose.

For the PG-EXTRA, we choose the stepsize $\ell = 0.1$ according to the sufficient condition suggested in [8]. Moreover, Network-DANE also uses the stepsize $\ell = 0.1$. According to their convergence conditions, the diminishing step size $\ell = \frac{0.1}{k+1}$ is used for the Prox-DGD.



(a) Rènyi graph with stochastic gradient(b) Ring graph with stochastic gradient

Fig. 1. Comparison of proposed SPPDM with Prox-DGD, PG-EXTRA, Prox-ADMM and Network-DANE in terms of optimal gap; the batch size is $|\mathcal{I}| = 256$.



(a) Rènyi graph with full gradient

(b) Ring graph with full gradient

Fig. 2. Comparison of proposed PPDM with Prox-DGD, PG-EXTRA, Prox-ADMM and Network-DANE in terms of optimal gap; they uses the full gradient.

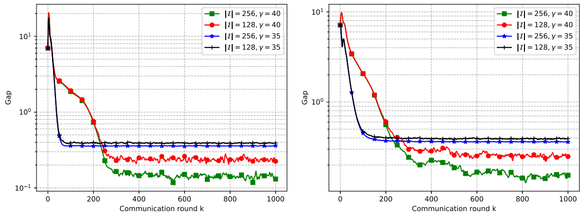
For Prox-DGD and PG-EXTRA, the mixing matrix follows the metropolis weight

$$[\mathbf{W}]_{ij} \triangleq \begin{cases} \frac{1}{\max\{d_i, d_j\}+1}, & \text{for } (i,j) \in \mathcal{E}, \\ 0, & \text{for } (i,j) \notin \mathcal{E} \text{ and } i \neq j. \\ 1 - \sum_{j \neq i} w_{ij}, & \text{for } i = j \end{cases} \tag{46}$$

If not specified, the parameters of the SPPDM/PPDM and the Prox-ADMM are given as $\alpha = .2$, $\kappa = 4$, $c = 1$, $\gamma = 40$, $\beta = 0.1$.[1] For the proposed SPPDM, we consider two cases about η, one is $\eta = 0$ without momentum, and the other is $\eta = 0.98$. When $\eta = 0$, we denote SPPDM as SPPD. We use the optimal gap defined in (39) to measure the performance. We run 5 independents trials for each algorithm with randomly generated data and random initial values. The optimal gap curves obtained by averaging over all 5 trials are plotted in Figs. 1-2.

In Fig. 1, we observe that the SPPDM, PG-EXTRA, Prox-ADMM and Network-DANE all perform better than Prox-DGD in terms of the optimal gap. The reason is that these methods all use constant step sizes rather than the diminishing step size. In addition, the proposed SPPDM has better performance than Prox-ADMM due to add the momentum technique. Compared

[1] By calculation, the Hessian matrix for the function $f_i(x)$ is $\frac{1}{N_i}\sum_{j=1}^{N_i} \frac{2\rho h_j h_j^T(2\rho - \|h_j^T x - y_j\|^2)}{(2\rho + \|h_j^T x - y_j\|^2)^2}$. It shows that the maximum eigenvalue of this Hessian matrix is smaller than 1 ($L < 1$) with the given parameter. Thus, the parameters of SPPDM/PPDM satisfy the conditions stated in Theorem 1. Moreover, in [35] (Proposition 1), the authors have shown that the objective function $f_i(x)$ has Lipschitz continuous gradient.



(a) Rènyi graph (b) Ring graph

Fig. 3. Convergence curves of the optimally gap achieved by our proposed methods with different batch size and step size.

Fig. 1 (a) with Fig. 1 (b), it shows that the proposed method performs better for both Erdos Renyi graph and a ring graph, thus it is robust to the difficult multi-agent network.

The impact of the mini-batch size $|\mathcal{I}|$ and parameters γ are analyzed in Fig. 3. When we consider the computational capability in each node, a smaller mini-batch size may be used. One can see that the smaller mini-batch size we use, the larger error one would suffer, which corroborates Corollary 1. In practice, we may select a moderate mini-batch size to balance the communication complexity and solution accuracy. With the same mini-batch size, the larger values of γ correspond to smaller primal and dual step sizes. Thus, the SPPDM with larger values of γ has slower convergence; whereas as seen from the figures, larger values of γ can lead to a smaller optimal gap.

For the offline setting, the comparison results of the proposed PPDM with the existing methods are shown in Fig. 2. It can be observed that the proposed PPDM enjoys the fastest convergence. Compared with the Prox-ADMM, it is clear to see the advantage of the PPDM with momentum for speeding up the algorithm convergence. It also shows that the proposed method enjoys sub-linear convergence for the offline/batch learning, which is consistent with Theorem 2.

B. Distributed Neural Network

In this simulation, our task is to classify handwritten digits from the MNIST dataset. The local loss function $f_i(\theta_i)$ in each node is the cross-entropy function, which is denoted as

$$f_i(\theta) = -\frac{1}{m}\sum_{j=1}^{m}\langle y_j, \log(h_\theta(x_j))\rangle.$$

In this example, we do not consider nonsmooth term and inequality constraint set. Thus, many existing methods, D-PSGD [21], D^2 [25], Network-SVRG [36] and PR-SGD-M [31] can be applied to train a classification DNN.

We consider two different neural network models, one is the fully connected neural network (FCN) that has at most three hidden layers, the other is the convolutional neural network (CNN) that at most has three convolutional layers. For the fully connected neural network, each hidden layer with 500 neurons. Then we analyze the performance of the compared methods by increasing the depth of the neural network.

Next, the 6×10^4 training samples are divided into 10 subsets and assigned to the $N = 10$ agents in two ways. The first is the *IID* case, where the samples are sufficiently shuffled, and then partitioned into 10 subsets with equal size ($m = 6000$). The second is the *Non-IID* case, where we first sort the samples according to their labels, divide them into 20 shards of size 3000, and assign each of 10 agents 2 shards. Thus most agents have samples of two digits only, which is popular appeared in the Federated learning.

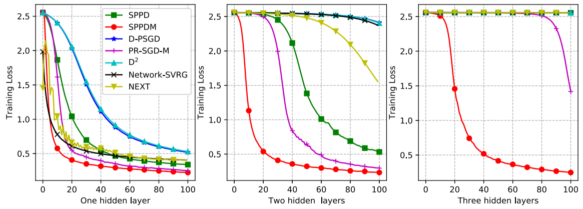


Fig. 4. Comparison of proposed SPPDM/SPPD algorithms with different methods under the IID case by using FCN model.

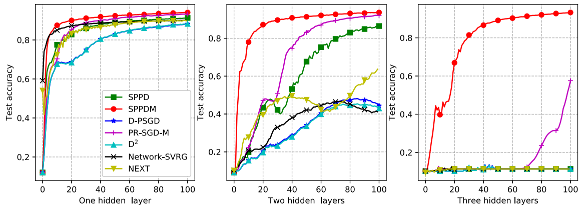


Fig. 5. Comparison of proposed SPPDM/SPPD algorithms with different methods under the IID case by using FCN model.

The communication graph is also a ring. We compare the SPPDM with the D-PSGD [21], D^2 [25], Network-SVRG [36], NEXT [13] and PR-SGD-M [31]. The same mixing matrix in (46) is used for the three methods. Moreover, a fixed step size of $\ell = 0.05$ is used to ensure the convergence of these three methods in the simulation. For the proposed SPPDM, we set parameter $c = 1$, $\gamma = 3$, $\alpha = 0.001$, $\kappa = 0.1$, $\beta = 0.9$, and $\eta = 0.9$. The batch size is $|\mathcal{I}| = 128$.

We calculate the loss value and the classification accuracy based on the average model $\bar{\theta} = 1/N\sum_{i=1}^{N}\theta_i$. Fig. 4 and Fig. 5 show the training loss and the classification accuracy for the IID case for FCN model by averaging over all 3 trials, respectively. From Fig. 4, we see that D^2 and D-PSGD have a similar performance; meanwhile, the proposed SPPDM performs better than the other methods. Besides, compared the left figure, middle figure and right figure in Figs. 4- 5, one can see that SPPDM, PPDM, PR-SGD-M, NEXT enjoy fast decreasing of the loss function and increasing of the classification accuracy, especially for the deep networks [37], respectively. The reason is that both SPPDM and PR-SGD-M use the momentum technique. We should point out that the communication overhead of PR-SGD-M is twice of the SPPDM since the PR-SGD-M requires the agents to exchange not only the variable x_i but also the momentum variables. Lastly, comparing the SPPDM with SPPD, it shows again that the momentum techniques can accelerate the algorithm convergence.

From Figs. 6- 7 for the IID case by using the CNN model, we can also see the similar performance that SPPDM and PR-SGD-M perform better with the increasing of the convolutional layer. Figs. 8- 9 present the result for the non-IID case by using the

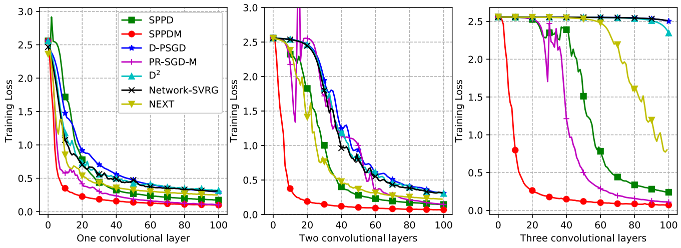


Fig. 6. Comparison of proposed SPPDM/SPPD algorithms with different methods under the IID case by using CNN model.

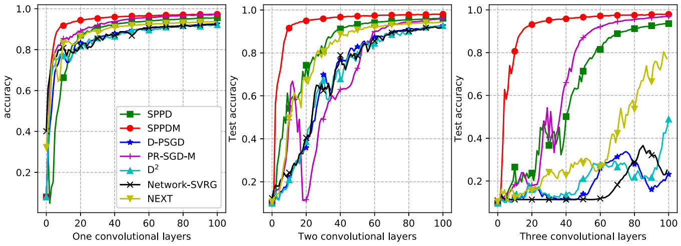


Fig. 7. Comparison of proposed SPPDM/SPPD algorithms with different methods under the IID case by using CNN model.

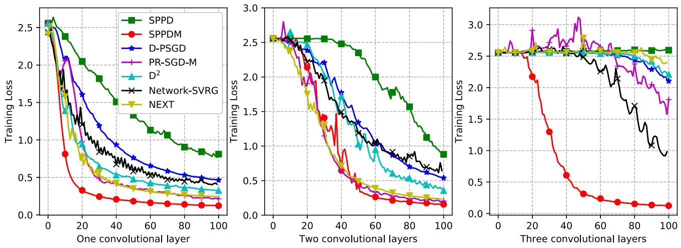


Fig. 8. Comparison of proposed SPPDM/SPPD algorithms with different methods under the non-IID case by using CNN model.

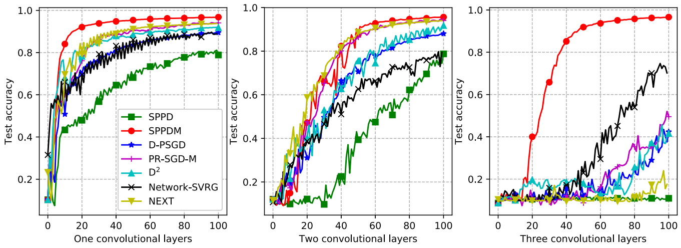


Fig. 9. Comparison of proposed SPPDM/SPPD algorithms with different methods under the non-IID case by using CNN model.

CNN model, we can observe that the D^2 and Network-SVRG perform better than D-PSGD and SPPD, the reason may be they use the variance reduction technique. In fact, by comparing the curves in Fig. 6 with those in Fig. 8, one can see that the convergence of SPPDM is also slowed under the Non-IID case, but it still performs best among the methods under test.

V. Conclusion

In this paper, we have proposed a distributed stochastic proximal primal-dual algorithm with momentum for minimizing a non-convex and non-smooth function (5) over a connected multi-agent network. We have shown (in Remark 1 and Remark 2) that the proposed algorithm has a close connection with some of the existing algorithms that are for convex and smooth problems, and therefore can be regarded as an enhanced counterpart of these existing algorithms. Theoretically, under Assumptions 1-3, we have built the convergence conditions of the proposed algorithms in Theorem 1 and Theorem 2. In particular, we have shown that the proposed SPPDM can achieve an ϵ-stationary solution with $\mathcal{O}(1/\epsilon^2)$ computational complexity and $\mathcal{O}(1/\epsilon)$ communication complexity, where the latter is better than many of the existing methods which have $\mathcal{O}(1/\epsilon^2)$ communication complexity (see Table 1). Experimental results have demonstrated that the proposed algorithms with momentum can effectively speed up the convergence. For distributed learning under non-IID data distribution (Figs. 8-9), we have also shown the proposed SPPDM performs better than the existing methods.

Appendix A
Proof of Remark 4

When the objective function only has smooth term that considered by the algorithms D-PSGD, D^2, PR-SGD-M and GNSD, the proposed measure $Q(\mathbf{x},\boldsymbol{\lambda})$ can be rewritten as

$$Q(\mathbf{x},\boldsymbol{\lambda}) = \|\nabla f(\mathbf{x}) + \mathbf{A}^T\boldsymbol{\lambda}\|^2 + \|\mathbf{A}\mathbf{x}\|^2.$$

Firstly, PSGD, D^2, PR-SGD-M use the measure $\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\bar{x})\|^2$ to define an ϵ-stationary point, where $\bar{x} = \frac{1}{N}\sum_{i=1}^N x_i$ and N is the number of the agents. Then we show that our definition of ϵ-stationary point is the sufficient condition for $\frac{2}{N}(1+\frac{L^2}{\sigma_{\min}})\epsilon$-stationary point in D-PSGD, D^2, PR-SGD-M, which means $\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\bar{x})\|^2 \le \frac{2}{N}(1+\frac{L^2}{\sigma_{\min}})\epsilon$, where $\sigma_{\min}$ is the smallest non-zero eigenvalue of $\mathbf{A}^T\mathbf{A}$.

Let $\mathbf{p} = [p_1, p_2, \ldots, p_N]^T = \mathbf{A}^T\boldsymbol{\lambda}$ and recall that the definition of $\mathbf{A}$, we have $\mathbf{A}\mathbf{1} = 0$, where $\mathbf{1}\in\mathbb{R}^{Nn}$ is an all-one vector. Thus $\sum_{i=1}^N p_i = (\mathbf{A}\mathbf{1})^T\boldsymbol{\lambda} = 0$. Then it implies

$$\begin{aligned}\left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(x_i)\right\|^2 &= \left\|\frac{1}{N}\sum_{i=1}^N (\nabla f_i(x_i) + p_i)\right\|^2\\ &\le \frac{1}{N}\sum_{i=1}^N \|\nabla f_i(x_i) + p_i\|^2\\ &= \frac{1}{N}\|\nabla f(\mathbf{x}) + \mathbf{A}^T\boldsymbol{\lambda}\|^2, \qquad (47)\end{aligned}$$

where the inequality comes from the convexity of $\|\cdot\|^2$. On the other hand, we have

$$\begin{aligned}\sum_{i=1}^N \|x_i - \bar{x}\|^2 &= \|\mathbf{x} - \bar{\mathbf{x}}\|^2 \le \frac{1}{\sigma_{\min}}\|\mathbf{x} - \bar{\mathbf{x}}\|^2_{\mathbf{A}^T\mathbf{A}}\\ &= \frac{1}{\sigma_{\min}}\|\mathbf{A}\mathbf{x} - \mathbf{A}\bar{\mathbf{x}}\|^2 = \frac{1}{\sigma_{\min}}\|\mathbf{A}\mathbf{x}\|^2, \qquad (48)\end{aligned}$$

Then we have

$$\begin{aligned}
&\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\bar{x})\right\|^2\\
&\le 2\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i)\right\|^2+2\left\|\frac{1}{N}\sum_{i=1}^{N}(\nabla f_i(\bar{x})-\nabla f_i(x_i))\right\|^2\\
&\le \frac{2}{N}\left\|\nabla f(\mathbf{x})+\mathbf{A}^T\boldsymbol{\lambda}\right\|^2+\frac{2}{N}\sum_{i=1}^{N}\|\nabla f_i(\bar{x})-\nabla f_i(x_i)\|^2\\
&\le \frac{2}{N}\left\|\nabla f(\mathbf{x})+\mathbf{A}^T\boldsymbol{\lambda}\right\|^2+\frac{2L^2}{N}\sum_{i=1}^{N}\|\bar{x}-x_i\|^2\\
&\le \frac{2}{N}\left\|\nabla f(\mathbf{x})+\mathbf{A}^T\boldsymbol{\lambda}\right\|^2+\frac{2L^2}{N\sigma_{\min}}\|\mathbf{A}\mathbf{x}\|^2\\
&\le \frac{2}{N}\left(1+\frac{L^2}{\sigma_{\min}}\right)\left(\left\|\nabla f(\mathbf{x})+\mathbf{A}^T\boldsymbol{\lambda}\right\|^2+\|\mathbf{A}\mathbf{x}\|^2\right)\\
&= \frac{2}{N}\left(1+\frac{L^2}{\sigma_{\min}}\right)Q(\mathbf{x},\boldsymbol{\lambda}), \qquad (49)
\end{aligned}$$

where the first inequality uses Cauchy-Schwartz inequality; the second inequality comes from (47); the third inequality dues to the assumption that f has Lipschitz continuous gradients and L is the Lipschitz constant; the fourth inequality dues to (48). Thus $Q(\mathbf{x},\boldsymbol{\lambda})\le\epsilon$, we have $\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\bar{x})\|^2\le\frac{2}{N}(1+\frac{L^2}{\sigma_{\min}})\epsilon$.

Secondly, for the GNSD, the performance measure is defined as $\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i)\|^2+\frac{1}{N}\sum_{i=1}^{N}\|x_i-\bar{x}\|^2$. Then we show that our definition of ϵ-stationary point is the sufficient condition for $\frac{1}{N}(1+\frac{1}{\sigma_{\min}})\epsilon$-stationary point in GNSD. Using (47) and (48), we have

$$\begin{aligned}
&\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i)\right\|^2+\frac{1}{N}\sum_{i=1}^{N}\|x_i-\bar{x}\|^2\\
&\le \frac{1}{N}\left\|\nabla f(\mathbf{x})+\mathbf{A}^T\boldsymbol{\lambda}\right\|^2+\frac{1}{N\sigma_{\min}}\|\mathbf{A}\mathbf{x}\|^2\\
&\le \frac{1}{N}\left(1+\frac{1}{\sigma_{\min}}\right)Q(\mathbf{x},\boldsymbol{\lambda}).
\end{aligned}$$

Appendix B
Proof of Theorem 1

Let us recapitulate the augmented Lagrange function in (20) below

$$\begin{aligned}
L_c(\mathbf{x},\mathbf{z};\boldsymbol{\lambda})&=f(\mathbf{x})+r(\mathbf{x})+\langle\boldsymbol{\lambda},\mathbf{A}\mathbf{x}\rangle\\
&\quad+\frac{c}{2}\|\mathbf{A}\mathbf{x}\|^2+\frac{\kappa}{2}\|\mathbf{x}-\mathbf{z}\|^2. \qquad (50)
\end{aligned}$$

We introduce some auxiliary functions as follows

$$d(\mathbf{z};\boldsymbol{\lambda})=\min_{\mathbf{x}}L_c(\mathbf{x},\mathbf{z};\boldsymbol{\lambda}) \qquad (51)$$

$$\mathbf{x}(\mathbf{z};\boldsymbol{\lambda})=\arg\min_{\mathbf{x}}L_c(\mathbf{x},\mathbf{z};\boldsymbol{\lambda}) \qquad (52)$$

$$P(\mathbf{z})=\min_{\mathbf{A}\mathbf{x}=0}f(\mathbf{x})+r(\mathbf{x})+\frac{\kappa}{2}\|\mathbf{x}-\mathbf{z}\|^2 \qquad (53)$$

$$\mathbf{x}(\mathbf{z})=\arg\min_{\mathbf{A}\mathbf{x}=0}f(\mathbf{x})+r(\mathbf{x})+\frac{\kappa}{2}\|\mathbf{x}-\mathbf{z}\|^2. \qquad (54)$$

Besides, we define the full gradient iterate $\hat{\mathbf{x}}^{k+1}$ and $\hat{\mathbf{z}}^{k+1}$,

$$\hat{\mathbf{x}}^{k+1}:=\arg\min_{\mathbf{x}}g(\mathbf{x},\mathbf{w}^k;\boldsymbol{\lambda}^{k+1}) \qquad (55)$$

$$\hat{\mathbf{z}}^{k+1}:=\mathbf{z}^k+\beta(\hat{\mathbf{x}}^{k+1}-\mathbf{z}^k), \qquad (56)$$

where $\mathbf{w}^k=[\mathbf{x}^k,\mathbf{s}^k,\mathbf{z}^k]$ and

$$\begin{aligned}
&g(\mathbf{x},\mathbf{w}^k;\boldsymbol{\lambda}^{k+1})\\
&=f(\mathbf{s}^k)+\langle\nabla f(\mathbf{s}^k),\mathbf{x}-\mathbf{s}^k\rangle+\frac{\gamma}{2}\|\mathbf{x}-\mathbf{s}^k\|^2+r(\mathbf{x})\\
&+\langle\boldsymbol{\lambda}^{k+1},\mathbf{A}\mathbf{x}\rangle+\frac{c}{2}\|\mathbf{A}\mathbf{x}\|^2+\frac{c}{2}\|\mathbf{x}-\mathbf{x}^k\|^2_{\mathbf{B}^\top\mathbf{B}}+\frac{\kappa}{2}\|\mathbf{x}-\mathbf{z}^k\|^2. \qquad (57)
\end{aligned}$$

We also define

$$g(\mathbf{x},\mathbf{w}^k,\boldsymbol{\xi}^k;\boldsymbol{\lambda}^{k+1}):=g(\mathbf{x},\mathbf{x}^k,\mathbf{s}^k,\mathbf{z}^k,\boldsymbol{\xi}^k;\boldsymbol{\lambda}^{k+1})$$

for (27) at our disposal.

A. Some Error Bounds

Firstly, we show the upper bound between $\mathbf{x}^{k+1}$ and $\hat{\mathbf{x}}^{k+1}$.

Lemma 1: Suppose Assumption 3 holds, we have

$$\mathbb{E}[\|\mathbf{x}^{k+1}-\hat{\mathbf{x}}^{k+1}\|^2]\le\frac{N\sigma^2}{(\gamma+2c+\kappa)^2|\mathcal{I}|}. \qquad (58)$$

Proof: See Reference [32]. ■

Lemma 2: Suppose $\kappa>-\mu$. There exists some positive constants σ_1,σ_2 such that the following primal error bound holds

$$\|\mathbf{x}^k-\mathbf{x}(\mathbf{z}^k;\boldsymbol{\lambda}^{k+1})\|\le\sigma_1\|\mathbf{x}^k-\hat{\mathbf{x}}^{k+1}\|+\sigma_2\|\mathbf{x}^k-\mathbf{s}^k\|. \qquad (59)$$

Proof: Based on $\kappa>-\mu$, we know that L_c in (50) is strongly convex in $\mathbf{x}$ with modulus $\kappa+\mu$ and Lipschitz constant $\kappa+L+c\sigma_A^2$, where σ_A is the spectral norm of the matrix. Thus, we can apply [38, Theorem 3.1] to upper bound the distance between $\mathbf{x}^k$ and the optimal solution $\mathbf{x}(\mathbf{z}^k;\boldsymbol{\lambda}^{k+1})$

$$\|\mathbf{x}^k-\mathbf{x}(\mathbf{z}^k;\boldsymbol{\lambda}^{k+1})\|\le\varrho\|\tilde{\nabla}_{\mathbf{x}}L_c(\mathbf{x}^k,\mathbf{z}^k;\boldsymbol{\lambda}^{k+1})\|, \qquad (60)$$

where $\varrho=\frac{\kappa+L+c\sigma_A^2+1}{\kappa+\mu}$ and

$$\tilde{\nabla}_{\mathbf{x}}L_c(\mathbf{x},\mathbf{z};\boldsymbol{\lambda})=\mathbf{x}-\mathrm{prox}_r^{\boldsymbol{\Psi}}(\mathbf{x}-\nabla_{\mathbf{x}}(L_c(\mathbf{x},\mathbf{z};\boldsymbol{\lambda})-r(\mathbf{x})))$$

is known as the proximal gradient.

We can bound $\|\tilde{\nabla}_{\mathbf{x}}L_c(\mathbf{x}^k,\mathbf{z}^k;\boldsymbol{\lambda}^{k+1})\|$ as follows

$$\begin{aligned}
&\|\tilde{\nabla}_{\mathbf{x}}L_c(\mathbf{x}^k,\mathbf{z}^k;\boldsymbol{\lambda}^{k+1})\|\\
&=\|\mathbf{x}^k-\mathrm{prox}_r^{\boldsymbol{\Psi}}(\mathbf{x}^k-\nabla_{\mathbf{x}}(L_c(\mathbf{x}^k,\mathbf{z}^k;\boldsymbol{\lambda}^{k+1})-r(\mathbf{x}^k)))\|\\
&\le\|\mathbf{x}^k-\hat{\mathbf{x}}^{k+1}\|\\
&+\|\hat{\mathbf{x}}^{k+1}-\mathrm{prox}_r^{\boldsymbol{\Psi}}(\mathbf{x}^k-\nabla_{\mathbf{x}}(L_c(\mathbf{x}^k,\mathbf{z}^k;\boldsymbol{\lambda}^{k+1})-r(\mathbf{x}^k)))\|\\
&=\|\mathbf{x}^k-\hat{\mathbf{x}}^{k+1}\|\\
&+\Big\|\mathrm{prox}_r^{\boldsymbol{\Psi}}(\hat{\mathbf{x}}^{k+1}-\nabla_{\mathbf{x}}(g(\hat{\mathbf{x}}^{k+1},\mathbf{w}^k;\boldsymbol{\lambda}^{k+1})-r(\hat{\mathbf{x}}^{k+1})))
\end{aligned}$$

$$
\begin{aligned}
&- \mathrm{prox}_r^{\Psi}(\mathbf{x}^k - \nabla_{\mathbf{x}}(L_c(\mathbf{x}^k, \mathbf{z}^k; \boldsymbol{\lambda}^{k+1}) - r(\mathbf{x}^k)))\Big\| \\
&\leq (2 + 2cd_{\max} + \gamma + \kappa)\|\mathbf{x}^k - \hat{\mathbf{x}}^{k+1}\| + (\gamma + L)\|\mathbf{x}^k - \mathbf{s}^k\|,
\end{aligned}
$$

where $d_{\max} = \max\{d_1, \ldots, d_N\}$; the second equality is obtained by using the optimality condition of $\hat{\mathbf{x}}^{k+1}$ in (55), and the second inequality is based on the nonexpansive property of the proximal operator. Denote

$$\sigma_1 = \varrho(2 + 2cd_{\max} + \gamma + \kappa), \tag{61}$$

$$\sigma_2 = \gamma + L. \tag{62}$$

The proof is complete. ■

Lemma 3: (Lemma 3.2 in [18]) Suppose $\kappa > -\mu$, and Assumption 1 holds. There exists some positive constants σ_3, σ_4 such that the following error bounds hold

$$\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\| \geq \sigma_3\|\mathbf{x}(\mathbf{z}; \boldsymbol{\lambda}_1) - \mathbf{x}(\mathbf{z}; \boldsymbol{\lambda}_2)\| \tag{63}$$

$$\|\mathbf{z}_1 - \mathbf{z}_2\| \geq \sigma_4\|\mathbf{x}(\mathbf{z}_1; \boldsymbol{\lambda}) - \mathbf{x}(\mathbf{z}_2; \boldsymbol{\lambda})\|, \tag{64}$$

where

$$\sigma_3 = (\kappa + \mu)/\sigma_A \tag{65}$$

$$\sigma_4 = (\kappa + \mu)/\kappa. \tag{66}$$

Lemma 4: Suppose that Assumptions 1-2 hold and $\kappa > \mu$. Then, there exist some positive scalars σ_5, Δ such that the following dual error bound holds

$$\|\mathbf{x}(\mathbf{z}, \boldsymbol{\lambda}) - \mathbf{x}(\mathbf{z})\| \leq \sigma_5\|\mathbf{A}\mathbf{x}(\mathbf{z}; \boldsymbol{\lambda})\|, forany \mathbf{z}, \boldsymbol{\lambda}. \tag{67}$$

where σ_5 depends only on the constants L, κ, σ_A, μ and the matrices $\mathbf{A}, \mathbf{S}_x, \mathbf{S}_y$.

Proof: The lemma is an extension of [19, Lemma 3.2], where the non-smooth term $r(\mathbf{x})$ of (5) is limited to an indicator function of a polyhedral set. Due to limited space, the detailed proof are relegated to the supplementary document [39]. ■

B. Descent Lemmas

In order to show the convergence of Algorithm 1, we consider a new potential function,

$$
\begin{aligned}
\mathbb{E}[\phi^{k+1}] \triangleq\ & \mathbb{E}[L_c(\mathbf{x}^{k+1}, \mathbf{z}^{k+1}; \boldsymbol{\lambda}^{k+1}) + \tau\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2] \\
& + \mathbb{E}[2P(\mathbf{z}^{k+1}) - 2d(\mathbf{z}^{k+1}; \boldsymbol{\lambda}^{k+1})],
\end{aligned} \tag{68}
$$

for some $\tau > 0$. By the weak duality, we have

$$L_c(\mathbf{x}, \mathbf{z}; \boldsymbol{\lambda}) \geq d(\mathbf{z}; \boldsymbol{\lambda}), P(\mathbf{z}) \geq d(\mathbf{z}; \boldsymbol{\lambda}). \tag{69}$$

Thus, we have $\mathbb{E}[\phi^k] \geq \mathbb{E}[P(\mathbf{z}^k)]$. According to the definition of $P(\mathbf{z}^k)$ in (53) and Assumption 1 (ii), we obtain $P(\mathbf{z}^k) \geq \underline{f}$. As a result, $\mathbb{E}[\phi^k]$ is bounded below by $\underline{f}$.

Lemma 5: For a sequence $\{\mathbf{x}^k, \mathbf{z}^k, \boldsymbol{\lambda}^k\}$ generated by Algorithm 1, if $\kappa > -\mu, \gamma > 3L, 0 < \beta < 1$ and

$$0 \leq \eta \leq \sqrt{\frac{\kappa + 2c + \gamma - 3L}{2(\gamma - \mu + 3L)}} := \bar{\eta}, \tag{70}$$

there exit some positive constants τ, $\hat{\sigma}_1$ and $\hat{\sigma}_2$ such that

$$\hat{\sigma}_1 \triangleq \frac{\kappa + 2c + \gamma - 3L}{2} - 2\tau \geq 0 \tag{71}$$

$$\hat{\sigma}_2 \triangleq \frac{\mu - \gamma - 3L}{2}\bar{\eta}^2 + \tau \geq 0, \tag{72}$$

then

$$
\begin{aligned}
&\mathbb{E}[L_c(\mathbf{x}^k, \mathbf{z}^k; \boldsymbol{\lambda}^k) + \tau\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2] \\
&- \mathbb{E}[L_c(\mathbf{x}^{k+1}, \mathbf{z}^{k+1}; \boldsymbol{\lambda}^{k+1}) - \tau\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2] \\
&\geq -\alpha\mathbb{E}[\|\mathbf{A}\mathbf{x}^k\|^2] + \frac{\kappa}{2\beta}\mathbb{E}[\|\mathbf{z}^k - \mathbf{z}^{k+1}\|^2] \\
&+ \hat{\sigma}_1\mathbb{E}[\|\mathbf{x}^k - \hat{\mathbf{x}}^{k+1}\|^2] + \hat{\sigma}_2\mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2] \\
&- \left(\frac{1}{2\gamma} + \frac{3L + 4\tau}{2(\gamma + 2c + \kappa)^2}\right)\frac{N\sigma^2}{|\mathcal{I}|}.
\end{aligned} \tag{73}
$$

Proof: Firstly, according to (20) and (22), we have

$$\mathbb{E}[L_c(\mathbf{x}^k, \mathbf{z}^k; \boldsymbol{\lambda}^k) - L_c(\mathbf{x}^k, \mathbf{z}^k; \boldsymbol{\lambda}^{k+1})] = -\alpha\mathbb{E}[\|\mathbf{A}\mathbf{x}^k\|^2]. \tag{74}$$

Secondly, we have

$$
\begin{aligned}
&\mathbb{E}[L_c(\mathbf{x}^k, \mathbf{z}^k; \boldsymbol{\lambda}^{k+1}) - L_c(\mathbf{x}^{k+1}, \mathbf{z}^k; \boldsymbol{\lambda}^{k+1})] \\
&= \mathbb{E}[L_c(\mathbf{x}^k, \mathbf{z}^k; \boldsymbol{\lambda}^{k+1}) - g(\mathbf{x}^k, \mathbf{w}^k; \boldsymbol{\lambda}^{k+1})] \\
&+ \mathbb{E}[g(\mathbf{x}^k, \mathbf{w}^k \boldsymbol{\lambda}^{k+1}) - g(\hat{\mathbf{x}}^{k+1}, \mathbf{w}^k; \boldsymbol{\lambda}^{k+1})] \\
&+ \mathbb{E}[g(\hat{\mathbf{x}}^{k+1}, \mathbf{w}^k; \boldsymbol{\lambda}^{k+1}) - g(\mathbf{x}^{k+1}, \mathbf{w}^k, \boldsymbol{\xi}^k; \boldsymbol{\lambda}^{k+1})] \\
&+ \mathbb{E}[g(\mathbf{x}^{k+1}, \mathbf{w}^k, \boldsymbol{\xi}^k; \boldsymbol{\lambda}^{k+1}) - L_c(\mathbf{x}^{k+1}, \mathbf{z}^k; \boldsymbol{\lambda}^{k+1})].
\end{aligned} \tag{75}
$$

Next, we bound each of the terms in the right hand side of (75). Based on the definition of function g in (57), we have

$$
\begin{aligned}
&\mathbb{E}[L_c(\mathbf{x}^k, \mathbf{z}^k; \boldsymbol{\lambda}^{k+1}) - g(\mathbf{x}^k, \mathbf{w}^k; \boldsymbol{\lambda}^{k+1})] \\
&= \mathbb{E}[f(\mathbf{x}^k) - f(\mathbf{s}^k) - \langle\nabla f(\mathbf{s}^k), \mathbf{x}^k - \mathbf{s}^k\rangle - \frac{\gamma}{2}\|\mathbf{x}^k - \mathbf{s}^k\|^2] \\
&\geq \frac{\mu - \gamma}{2}\mathbb{E}[\|\mathbf{x}^k - \mathbf{s}^k\|^2],
\end{aligned} \tag{76}
$$

where the inequality comes from (34) in Assumption 1. Using the strongly convexity of the objective function g (with modulus $\kappa + 2c + \gamma$) and the definition of $\hat{\mathbf{x}}^{k+1}$ in (55), we can obtain

$$
\begin{aligned}
&\mathbb{E}[g(\mathbf{x}^k, \mathbf{w}^k; \boldsymbol{\lambda}^{k+1}) - g(\hat{\mathbf{x}}^{k+1}, \mathbf{w}^k; \boldsymbol{\lambda}^{k+1})] \\
&\geq \frac{\kappa + 2c + \gamma}{2}\mathbb{E}[\|\mathbf{x}^k - \hat{\mathbf{x}}^{k+1}\|^2].
\end{aligned} \tag{77}
$$

In addition, we have

$$
\begin{aligned}
&\mathbb{E}[g(\hat{\mathbf{x}}^{k+1}, \mathbf{w}^k; \boldsymbol{\lambda}^{k+1}) - g(\mathbf{x}^{k+1}, \mathbf{w}^k, \boldsymbol{\xi}^k; \boldsymbol{\lambda}^{k+1})] \\
&= \mathbb{E}[g(\hat{\mathbf{x}}^{k+1}, \mathbf{w}^k, \boldsymbol{\xi}^k; \boldsymbol{\lambda}^{k+1}) - g(\mathbf{x}^{k+1}, \mathbf{w}^k, \boldsymbol{\xi}^k; \boldsymbol{\lambda}^{k+1})] \\
&\geq 0,
\end{aligned} \tag{78}
$$

where the first equality dues to (36) in Assumption 3, and the above inequality dues to $\mathbf{x}^{k+1}$ is the optimal solution in (24).

Lastly, we can bound

$$\begin{aligned}
&\mathbb{E}[g(\mathbf{x}^{k+1},\mathbf{w}^k,\boldsymbol{\xi}^k;\boldsymbol{\lambda}^{k+1})-L_c(\mathbf{x}^{k+1},\mathbf{z}^k;\boldsymbol{\lambda}^{k+1})]\\
&=\mathbb{E}[f(\mathbf{s}^k)]+\frac{1}{|\mathcal{I}|}\sum_{j=1}^{|\mathcal{I}|}\mathbb{E}[\langle G(\mathbf{s}^k,\boldsymbol{\xi}_j^k),\mathbf{x}^{k+1}-\mathbf{s}^k\rangle]\\
&+\mathbb{E}\left[\frac{\gamma}{2}\|\mathbf{x}^{k+1}-\mathbf{s}^k\|^2]+\frac{c}{2}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2_{\mathbf{B}^\top\mathbf{B}}-f(\mathbf{x}^{k+1})\right]\\
&\geq\frac{1}{|\mathcal{I}|}\sum_{j=1}^{|\mathcal{I}|}\mathbb{E}[\langle G(\mathbf{s}^k,\boldsymbol{\xi}_j^k)-\nabla f(\mathbf{s}^k),\mathbf{x}^{k+1}-\mathbf{s}^k\rangle]\\
&+\frac{\gamma-L}{2}\mathbb{E}[\|\mathbf{x}^{k+1}-\mathbf{s}^k\|^2]+\frac{c}{2}\mathbb{E}[\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2_{\mathbf{B}^\top\mathbf{B}}]\\
&\geq-\frac{N\sigma^2}{2\gamma|\mathcal{I}|}-\frac{L}{2}\mathbb{E}[\|\mathbf{x}^{k+1}-\mathbf{s}^k\|^2], \qquad (79)
\end{aligned}$$

where the first inequality is obtained by applying the descent lemma [40, Lemma1.2.3]

$$f(\mathbf{x}^{k+1})\leq f(\mathbf{s}^k)+\langle\nabla f(\mathbf{s}^k),\mathbf{x}^{k+1}-\mathbf{s}^k\rangle+\frac{L}{2}\|\mathbf{x}^{k+1}-\mathbf{s}^k\|^2$$

owing to gradient Lipschitz continuity in (33); the second inequality holds by using the Young's inequality $a^\top b\geq-\frac{\|a\|^2}{2\gamma}-\frac{\gamma\|b\|^2}{2}$ and (37) in Assumption 3. Using the convexity of the operator $\|\cdot\|^2$, we have

$$\begin{aligned}
\|\mathbf{x}^{k+1}-\mathbf{s}^k\|^2\leq 3\|\mathbf{x}^{k+1}-\hat{\mathbf{x}}^{k+1}\|^2+3\|\hat{\mathbf{x}}^{k+1}-\mathbf{x}^k\|^2&\\
+3\|\mathbf{x}^k-\mathbf{s}^k\|^2.& \qquad (80)
\end{aligned}$$

Substituting (58) and (80) into (79) gives rise to

$$\begin{aligned}
&\mathbb{E}[g(\mathbf{x}^{k+1},\mathbf{w}^k,\boldsymbol{\xi}^k;\boldsymbol{\lambda}^{k+1})-L_c(\mathbf{x}^{k+1},\mathbf{z}^k;\boldsymbol{\lambda}^{k+1})]\\
&\geq-\left(\frac{1}{2\gamma}+\frac{3L}{2(\gamma+2c+\kappa)^2}\right)\frac{N\sigma^2}{|\mathcal{I}|}\\
&-\frac{3L}{2}\mathbb{E}[\|\hat{\mathbf{x}}^{k+1}-\mathbf{x}^k\|^2]-\frac{3L}{2}\mathbb{E}[\|\mathbf{s}^k-\mathbf{x}^k\|^2]. \qquad (81)
\end{aligned}$$

By further substituting (76)-(78) and (81) into (75), we obtain

$$\begin{aligned}
&\mathbb{E}[L_c(\mathbf{x}^k,\mathbf{z}^k;\boldsymbol{\lambda}^{k+1})+\tau\|\mathbf{x}^k-\mathbf{x}^{k-1}\|^2\\
&-L_c(\mathbf{x}^{k+1},\mathbf{z}^k;\boldsymbol{\lambda}^{k+1})-\tau\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2]\\
&\geq\frac{\mu-\gamma}{2}\mathbb{E}[\|\mathbf{x}^k-\mathbf{s}^k\|^2]+\frac{\kappa+2c+\gamma}{2}\mathbb{E}[\|\mathbf{x}^k-\hat{\mathbf{x}}^{k+1}\|^2]\\
&-\frac{3L}{2}\mathbb{E}[\|\hat{\mathbf{x}}^{k+1}-\mathbf{x}^k\|^2]-\frac{3L}{2}\mathbb{E}[\|\mathbf{s}^k-\mathbf{x}^k\|^2]\\
&-\left(\frac{1}{2\gamma}+\frac{3L}{2(\gamma+2c+\kappa)^2}\right)\frac{\sigma^2}{|\mathcal{I}|}+\tau\mathbb{E}[\|\mathbf{x}^k-\mathbf{x}^{k-1}\|^2]\\
&-2\tau\mathbb{E}[\|\mathbf{x}^{k+1}-\hat{\mathbf{x}}^{k+1}\|^2]-2\tau\mathbb{E}[\|\hat{\mathbf{x}}^{k+1}-\mathbf{x}^k\|^2]\\
&=\hat{\sigma}_1\mathbb{E}[\|\mathbf{x}^k-\hat{\mathbf{x}}^{k+1}\|^2]+\hat{\sigma}_2\mathbb{E}[\|\mathbf{x}^k-\mathbf{x}^{k-1}\|^2]\\
&-\left(\frac{1}{2\gamma}+\frac{3L+4\tau}{2(\gamma+2c+\kappa)^2}\right)\frac{N\sigma^2}{|\mathcal{I}|}, \qquad (82)
\end{aligned}$$

where $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are defined in (71) and (72), respectively, and the equality is obtained by applying (23).

Thirdly, according to the definition of the $\mathbf{z}$ update in (25), we have

$$\begin{aligned}
&\mathbb{E}[L_c(\mathbf{x}^{k+1},\mathbf{z}^k;\boldsymbol{\lambda}^{k+1})-L_c(\mathbf{x}^{k+1},\mathbf{z}^{k+1};\boldsymbol{\lambda}^{k+1})]\\
&\geq\frac{\kappa}{2\beta}(2-\beta)\mathbb{E}[\|\mathbf{z}^k-\mathbf{z}^{k+1}\|^2]\\
&\geq\frac{\kappa}{2\beta}\mathbb{E}[\|\mathbf{z}^k-\mathbf{z}^{k+1}\|^2], \qquad (83)
\end{aligned}$$

where the last inequality is due to $0<\beta<1$. By combining (74), (82) and (83), we obtain (73). Besides, (71) and (72) implies (70). ■

Lemma 6: Under Assumptions 1-3, if $\kappa>-\mu$, $\gamma>3L$, η is a constant satisfies the condition (41), and

$$0<\alpha\leq\min\left\{\frac{\hat{\sigma}_1}{4\sigma_A\sigma_1^2},\frac{\hat{\sigma}_2}{4\sigma_A^2\sigma_2^2\eta^2},c\right\}, \qquad (84)$$

$$0<\beta<\min\left\{\frac{\alpha}{12\kappa\sigma_5^2},\frac{\sigma_4}{36},1\right\}, \qquad (85)$$

where σ_1, σ_2, σ_4 and σ_5 are constants denoted in (61) and (62), (66) and (67), respectively. Then we have

$$\begin{aligned}
&\mathbb{E}[\phi^k-\phi^{k+1}]\\
&\geq\frac{\kappa(1-\beta)\beta}{4}\mathbb{E}[\|\hat{\mathbf{x}}^{k+1}-\mathbf{z}^k\|^2]+\frac{\alpha}{2}\mathbb{E}[\|\mathbf{A}\mathbf{x}(\mathbf{z}^k,\boldsymbol{\lambda}^{k+1})\|^2]\\
&+\frac{\hat{\sigma}_1}{2}\mathbb{E}[\|\mathbf{x}^k-\hat{\mathbf{x}}^{k+1}\|^2]+\frac{\hat{\sigma}_2}{2}\mathbb{E}[\|\mathbf{x}^k-\mathbf{x}^{k-1}\|^2]-\frac{C_1N\sigma^2}{|\mathcal{I}|},\\
&\qquad (86)
\end{aligned}$$

where $\hat{\mathbf{x}}^{k+1}$ and $\hat{\mathbf{z}}^{k+1}$ are defined in (55) and (56), and

$$C_1=\left(\frac{1}{2\gamma}+\frac{6L+8\tau+\kappa(1-\beta)}{4(\gamma+2c+\kappa)^2}\right). \qquad (87)$$

Proof: From the definition of $d(\mathbf{z};\boldsymbol{\lambda})$ in (51), we have

$$\begin{aligned}
&\mathbb{E}[d(\mathbf{z}^k;\boldsymbol{\lambda}^{k+1})-d(\mathbf{z}^k;\boldsymbol{\lambda}^k)]\\
&=\mathbb{E}[L_c(\mathbf{x}(\mathbf{z}^k;\boldsymbol{\lambda}^{k+1}),\mathbf{z}^k;\boldsymbol{\lambda}^{k+1})-L_c(\mathbf{x}(\mathbf{z}^k;\boldsymbol{\lambda}^k),\mathbf{z}^k;\boldsymbol{\lambda}^k)]\\
&\geq\mathbb{E}[L_c(\mathbf{x}(\mathbf{z}^k;\boldsymbol{\lambda}^{k+1}),\mathbf{z}^k;\boldsymbol{\lambda}^{k+1})-L_c(\mathbf{x}(\mathbf{z}^k;\boldsymbol{\lambda}^{k+1}),\mathbf{z}^k;\boldsymbol{\lambda}^k)]\\
&=\alpha\mathbb{E}[\langle\mathbf{A}\mathbf{x}^k,\mathbf{A}\mathbf{x}(\mathbf{z}^k,\boldsymbol{\lambda}^{k+1})\rangle],
\end{aligned}$$

where the inequality is due to $\mathbf{x}(\mathbf{z}^k;\boldsymbol{\lambda}^k)=\arg\min_{\mathbf{x}}L_c(\mathbf{x},\mathbf{z}^k;\boldsymbol{\lambda}^k)$ and the second equality comes from the iterates in (22). Using a similar technique, we have

$$\begin{aligned}
&\mathbb{E}[d(\mathbf{z}^{k+1};\boldsymbol{\lambda}^{k+1})-d(\mathbf{z}^k;\boldsymbol{\lambda}^{k+1})]\\
&\geq\frac{\kappa}{2}\mathbb{E}[(\mathbf{z}^{k+1}-\mathbf{z}^k)^\top(\mathbf{z}^{k+1}+\mathbf{z}^k-2\mathbf{x}(\mathbf{z}^{k+1},\boldsymbol{\lambda}^{k+1}))].
\end{aligned}$$

Combing the above two inequalities, we know

$$\begin{aligned}
&\mathbb{E}[d(\mathbf{z}^{k+1};\boldsymbol{\lambda}^{k+1})-d(\mathbf{z}^k;\boldsymbol{\lambda}^k)]\\
&\geq\alpha\mathbb{E}[\langle\mathbf{A}\mathbf{x}^k,\mathbf{A}\mathbf{x}(\mathbf{z}^k,\boldsymbol{\lambda}^{k+1})\rangle]\\
&+\frac{\kappa}{2}\mathbb{E}[(\mathbf{z}^{k+1}-\mathbf{z}^k)^\top(\mathbf{z}^{k+1}+\mathbf{z}^k-2\mathbf{x}(\mathbf{z}^{k+1},\boldsymbol{\lambda}^{k+1}))]. \qquad (88)
\end{aligned}$$

Based on Danskin's theorem [41, Proposition B.25] in convex analysis and $P(\mathbf{z})$ defined in (53) with $\kappa > -\mu$, we can have

$$\nabla P(\mathbf{z}^k) = \kappa(\mathbf{z}^k - \mathbf{x}(\mathbf{z}^k)).$$

Thus, it shows

$$\begin{aligned}
&\|\nabla P(\mathbf{z}^k) - \nabla P(\mathbf{z}^{k+1})\| \\
&\le \kappa\|\mathbf{z}^k - \mathbf{z}^{k+1}\| + \kappa\|\mathbf{x}(\mathbf{z}^{k+1}) - \mathbf{x}(\mathbf{z}^k)\| \\
&\le \kappa\tilde{\sigma}_4\|\mathbf{z}^{k+1} - \mathbf{z}^k\|,
\end{aligned}$$

where $\tilde{\sigma}_4 = 1 + \sigma_4^{-1}$ and the final inequality is due to Lemma 3. The above inequality shows the gradient of $P(\mathbf{z}^k)$ is Lipschitz continuous, which therefore it satisfies the descent lemma

$$\begin{aligned}
&\mathbb{E}[P(\mathbf{z}^{k+1}) - P(\mathbf{z}^k)] \\
&\le \mathbb{E}[\kappa(\mathbf{z}^{k+1} - \mathbf{z}^k)^\top(\mathbf{z}^k - \mathbf{x}(\mathbf{z}^k))] + \frac{\kappa\tilde{\sigma}_4}{2}\mathbb{E}[\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2].
\end{aligned} \tag{89}$$

By combining (88), (89) and (73), we obtain

$$\begin{aligned}
&\mathbb{E}[\phi^k - \phi^{k+1}] \\
&\ge -\alpha\mathbb{E}[\|\mathbf{A}\mathbf{x}^k\|^2] + \frac{\kappa}{2\beta}\mathbb{E}[\|\mathbf{z}^k - \mathbf{z}^{k+1}\|^2] \\
&+ \hat{\sigma}_1\mathbb{E}[\|\mathbf{x}^k - \hat{\mathbf{x}}^{k+1}\|^2] - 2\mathbb{E}[\kappa(\mathbf{z}^{k+1} - \mathbf{z}^k)^\top(\mathbf{z}^k - \mathbf{x}(\mathbf{z}^k))] \\
&+ \hat{\sigma}_2\mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2] - \left(\frac{1}{2\gamma} + \frac{3L + 4\tau}{2(\gamma + 2c + \kappa)^2}\right)\frac{N\sigma^2}{|\mathcal{I}|} \\
&- \kappa\tilde{\sigma}_4\mathbb{E}[\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2] + 2\alpha\mathbb{E}[\langle\mathbf{A}\mathbf{x}^k, \mathbf{A}\mathbf{x}(\mathbf{z}^k, \boldsymbol{\lambda}^{k+1})\rangle] \\
&+ \kappa\mathbb{E}[(\mathbf{z}^{k+1} - \mathbf{z}^k)^\top(\mathbf{z}^{k+1} + \mathbf{z}^k - 2\mathbf{x}(\mathbf{z}^{k+1}, \boldsymbol{\lambda}^{k+1}))] \\
&= \alpha\mathbb{E}[\|\mathbf{A}\mathbf{x}(\mathbf{z}^k, \boldsymbol{\lambda}^{k+1})\|^2] - \alpha\mathbb{E}[\|\mathbf{A}(\mathbf{x}^k - \mathbf{x}(\mathbf{z}^k, \boldsymbol{\lambda}^{k+1}))\|^2] \\
&+ \hat{\sigma}_1\mathbb{E}[\|\mathbf{x}^k - \hat{\mathbf{x}}^{k+1}\|^2] + \left(\frac{\kappa}{2\beta} + \kappa - \kappa\tilde{\sigma}_4\right)\mathbb{E}[\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2] \\
&+ 2\kappa\mathbb{E}[(\mathbf{z}^{k+1} - \mathbf{z}^k)^\top(\mathbf{x}(\mathbf{z}^k) - \mathbf{x}(\mathbf{z}^{k+1}; \boldsymbol{\lambda}^{k+1}))] \\
&+ \hat{\sigma}_2\mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2] - \left(\frac{1}{2\gamma} + \frac{3L + 4\tau}{2(\gamma + 2c + \kappa)^2}\right)\frac{N\sigma^2}{|\mathcal{I}|},
\end{aligned} \tag{90}$$

where the equality comes from completing the square

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{A}(\mathbf{x}^k - \mathbf{x}(\mathbf{z}^k, \boldsymbol{\lambda}^{k+1}))\|^2] \\
&= \mathbb{E}[\|\mathbf{A}\mathbf{x}^k\|^2 - 2\langle\mathbf{A}\mathbf{x}^k, \mathbf{A}vx(\mathbf{z}^k, \boldsymbol{\lambda}^{k+1}) + \|\mathbf{A}\mathbf{x}(\mathbf{z}^k, \boldsymbol{\lambda}^{k+1})\|^2\rangle].
\end{aligned}$$

We further bound the right-hand-side terms of (90). By using the Young's inequality, we have

$$\begin{aligned}
&2(\mathbf{z}^{k+1} - \mathbf{z}^k)^\top(\mathbf{x}(\mathbf{z}^k) - \mathbf{x}(\mathbf{z}^k; \boldsymbol{\lambda}^{k+1})) \\
&\ge -\frac{\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2}{6\beta} - 6\beta\|\mathbf{x}(\mathbf{z}^k) - \mathbf{x}(\mathbf{z}^k; \boldsymbol{\lambda}^{k+1})\|^2 \\
&\ge -\frac{\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2}{6\beta} - 6\beta\sigma_5^2\|\mathbf{A}\mathbf{x}(\mathbf{z}; \boldsymbol{\lambda})\|^2,
\end{aligned} \tag{91}$$

where the lase inequality dues to (67) in Lemma 4. Besides, using the error bound (64) in Lemma 3, we have

$$\begin{aligned}
&(\mathbf{z}^{k+1} - \mathbf{z}^k)^\top(\mathbf{x}(\mathbf{z}^k; \boldsymbol{\lambda}^{k+1}) - \mathbf{x}(\mathbf{z}^{k+1}; \boldsymbol{\lambda}^{k+1})) \\
&\ge -\|\mathbf{z}^{k+1} - \mathbf{z}^k\|\|\mathbf{x}(\mathbf{z}^k; \boldsymbol{\lambda}^{k+1}) - \mathbf{x}(\mathbf{z}^{k+1}; \boldsymbol{\lambda}^{k+1})\| \\
&\ge -\frac{1}{\sigma_4}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2.
\end{aligned} \tag{92}$$

Also, based on the error bound (59) in Lemma 2, we obtain

$$\begin{aligned}
&\|\mathbf{A}(\mathbf{x}^k - \mathbf{x}(\mathbf{z}^k, \boldsymbol{\lambda}^{k+1}))\|^2 \\
&\le 2\sigma_A^2\sigma_1^2\|\mathbf{x}^k - \hat{\mathbf{x}}^{k+1}\|^2 + 2\sigma_A^2\sigma_2^2\|\mathbf{x}^k - \mathbf{s}^k\|^2.
\end{aligned} \tag{93}$$

By substituting (91), (92) and (93) into (90), we therefore obtain

$$\begin{aligned}
&\mathbb{E}[\phi^k - \phi^{k+1}] \\
&\ge (\alpha - 6\kappa\beta\sigma_5^2)\mathbb{E}[\|\mathbf{A}\mathbf{x}(\mathbf{z}^k, \boldsymbol{\lambda}^{k+1})\|^2] \\
&+ \left(\frac{\kappa}{2\beta} + \kappa - \kappa\tilde{\sigma}_5 - \frac{\kappa}{6\beta} - \frac{2\kappa}{\sigma_4}\right)\mathbb{E}[\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2] \\
&+ \left(\hat{\sigma}_1 - 2\alpha\sigma_A^2\sigma_1^2\right)\mathbb{E}[\|\mathbf{x}^k - \hat{\mathbf{x}}^{k+1}\|^2] \\
&+ (\hat{\sigma}_2 - 2\alpha\sigma_A^2\sigma_2^2\eta^2)\mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2] \\
&- \left(\frac{1}{2\gamma} + \frac{3L + 4\tau}{2(\gamma + 2c + \kappa)^2}\right)\frac{N\sigma^2}{|\mathcal{I}|}.
\end{aligned} \tag{94}$$

From (85), we know $\beta < \frac{\sigma_4}{36}$. By recalling $\tilde{\sigma}_4 = 1 + \sigma_4^{-1}$, we have

$$\frac{\kappa}{2\beta} + \kappa - \kappa\tilde{\sigma}_4 - \frac{\kappa}{6\beta} - \frac{2\kappa}{\sigma_4} \ge \frac{\kappa}{4\beta}.$$

As $\beta < \frac{\alpha}{12\kappa\sigma_5^2}$ by (85), we have $\alpha - 6\kappa\beta\sigma_5^2 \ge \frac{\alpha}{2}$. Similarly, based on (84), we have

$$\hat{\sigma}_1 - 2\alpha\sigma_A^2\sigma_1^2 \ge \frac{\hat{\sigma}_1}{2}, \hat{\sigma}_2 - 2\alpha\sigma_A^2\sigma_2^2\eta^2 \ge \frac{\hat{\sigma}_2}{2}.$$

Thus, it follows from (94) that

$$\begin{aligned}
&\mathbb{E}[\phi^k - \phi^{k+1}] \\
&\ge \frac{\kappa}{4\beta}\mathbb{E}[\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2] + \frac{\alpha}{2}\mathbb{E}[\|\mathbf{A}\mathbf{x}(\mathbf{z}^k, \boldsymbol{\lambda}^{k+1})\|^2] \\
&+ \frac{\hat{\sigma}_1}{2}\mathbb{E}[\|\mathbf{x}^k - \hat{\mathbf{x}}^{k+1}\|^2] + \frac{\hat{\sigma}_2}{2}\mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2] \\
&- \left(\frac{1}{2\gamma} + \frac{3L + 4\tau}{2(\gamma + 2c + \kappa)^2}\right)\frac{N\sigma^2}{|\mathcal{I}|}.
\end{aligned} \tag{95}$$

Note that by using the definition of $\hat{\mathbf{z}}^{k+1}$ in (56) and by (25), we have

$$\hat{\mathbf{z}}^{k+1} = \mathbf{z}^{k+1} + \beta(\hat{\mathbf{x}}^{k+1} - \mathbf{x}^{k+1}). \tag{96}$$

Thus, we can bound $\mathbb{E}[\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2]$ as

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2] \\
&\ge \left(1 - \frac{1}{\beta}\right)\mathbb{E}[\|\mathbf{z}^{k+1} - \hat{\mathbf{z}}^{k+1}\|^2] + (1 - \beta)\mathbb{E}[\|\hat{\mathbf{z}}^{k+1} - \mathbf{z}^k\|^2]
\end{aligned}$$

$$
\begin{aligned}
&= \beta(\beta-1)\mathbb{E}[\|\mathbf{x}^{k+1} - \hat{\mathbf{x}}^{k+1}\|^2] + (1-\beta)\mathbb{E}[\|\hat{\mathbf{z}}^{k+1} - \mathbf{z}^k\|^2] \\
&\geq \frac{\beta(\beta-1)}{(\gamma+2c+\kappa)^2}\frac{N\sigma^2}{|\mathcal{I}|} + (1-\beta)\beta^2\mathbb{E}[\|\hat{\mathbf{x}}^{k+1} - \mathbf{z}^k\|^2],
\end{aligned}
$$

where the last inequality comes from (58) and (56). By substituting the above inequality (95), we obtain (86). ■

C. Proof of Theorem 1

We are ready to prove Theorem 1. By summing (86) for $k = 0, 1, \ldots, K-1$, we obtain

$$
\begin{aligned}
&\mathbb{E}[\phi^0 - \phi^K] \\
&\geq \frac{\kappa(1-\beta)\beta}{4}\sum_{k=0}^{K-1}\mathbb{E}[\|\hat{\mathbf{x}}^{k+1} - \mathbf{z}^k\|^2] - K\frac{C_1 N\sigma^2}{|\mathcal{I}|} \\
&+ \frac{\alpha}{2}\sum_{k=0}^{K-1}\mathbb{E}[\|\mathbf{A}\mathbf{x}(\mathbf{z}^k, \boldsymbol{\lambda}^{k+1})\|^2] + \frac{\hat{\sigma}_1}{2}\sum_{k=0}^{K-1}\mathbb{E}[\|\mathbf{x}^k - \hat{\mathbf{x}}^{k+1}\|^2] \\
&+ \frac{\hat{\sigma}_2}{2}\sum_{k=0}^{K-1}\mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2]. \qquad (97)
\end{aligned}
$$

Recall the definition of $Q(\mathbf{x}, \boldsymbol{\lambda})$ in (39)

$$
Q(\mathbf{x}, \boldsymbol{\lambda}) = \|\mathbf{x} - \mathrm{prox}_r^1(\mathbf{x} - \nabla f(\mathbf{x}) - \mathbf{A}^\top\boldsymbol{\lambda})\|^2 + \|\mathbf{A}\mathbf{x}\|^2. \qquad (98)
$$

To obtain the desired result, we first consider

$$
\begin{aligned}
&\mathbb{E}[\|\mathbf{x}^k - \mathrm{prox}_r^1(\mathbf{x}^k - \nabla_{\mathbf{x}} f(\mathbf{x}^k) - \mathbf{A}^\top\boldsymbol{\lambda}^{k+1})\|^2] \\
&\leq 2\mathbb{E}[\|\mathbf{x}^k - \hat{\mathbf{x}}^{k+1}\|^2] \\
&+ 2\mathbb{E}[\|\hat{\mathbf{x}}^{k+1} - \mathrm{prox}_r^1(\mathbf{x}^k - \nabla_{\mathbf{x}} f(\mathbf{x}^k) - \mathbf{A}^\top\boldsymbol{\lambda}^{k+1})\|^2] \quad (99)
\end{aligned}
$$

where the inequality dues to $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. Notice

$$
\begin{aligned}
&\mathbb{E}[\|\hat{\mathbf{x}}^{k+1} - \mathrm{prox}_r^1(\mathbf{x}^k - \nabla_{\mathbf{x}} f(\mathbf{x}^k) - \mathbf{A}^\top\boldsymbol{\lambda}^{k+1})\|^2] \\
&= \mathbb{E}[\|\mathrm{prox}_r^1(\hat{\mathbf{x}}^{k+1} - \nabla_{\mathbf{x}} g(\hat{\mathbf{x}}^{k+1}, \mathbf{w}^k; \boldsymbol{\lambda}^{k+1})) \\
&- \mathrm{prox}_r^1(\mathbf{x}^k - \nabla_{\mathbf{x}} f(\mathbf{x}^k) - \mathbf{A}^\top\boldsymbol{\lambda}^{k+1})\|^2] \\
&\leq \mathbb{E}[\|\hat{\mathbf{x}}^{k+1} - \mathbf{x}^k - \nabla_{\mathbf{x}} g(\hat{\mathbf{x}}^{k+1}, \mathbf{w}^k; \boldsymbol{\lambda}^{k+1}) \\
&+ \nabla_{\mathbf{x}} f(\mathbf{x}^k) + \mathbf{A}^\top\boldsymbol{\lambda}^{k+1}\|^2] \\
&\leq 2\mathbb{E}[\|\hat{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2] \\
&+ 2\mathbb{E}[\|\nabla_{\mathbf{x}} g(\hat{\mathbf{x}}^{k+1}, \mathbf{w}^k; \boldsymbol{\lambda}^{k+1}) - \nabla_{\mathbf{x}} f(\mathbf{x}^k) - \mathbf{A}^\top\boldsymbol{\lambda}^{k+1}\|^2], \\
&= 2\mathbb{E}[\|\hat{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2] + 2\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{s}^k) - \nabla_{\mathbf{x}} f(\mathbf{x}^k) \\
&+ \gamma(\hat{\mathbf{x}}^{k+1} - \mathbf{s}^k) + c\mathbf{D}(\hat{\mathbf{x}}^{k+1} - \mathbf{x}^k) + c\mathbf{A}^T\mathbf{A}\mathbf{x}^k \\
&+ \kappa(\hat{\mathbf{x}}^{k+1} - \mathbf{z}^k)\|^2] \\
&\leq 2\mathbb{E}[\|\hat{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2] + 10\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{s}^k) - \nabla_{\mathbf{x}} f(\mathbf{x}^k)\|^2 \\
&+ \|\gamma(\hat{\mathbf{x}}^{k+1} - \mathbf{s}^k)\|^2 + \|c\mathbf{D}(\hat{\mathbf{x}}^{k+1} - \mathbf{x}^k)\|^2 + \|c\mathbf{A}^T\mathbf{A}\mathbf{x}^k\|^2 \\
&+ \|\kappa(\hat{\mathbf{x}}^{k+1} - \mathbf{z}^k)\|^2] \\
&\leq (2 + 10c^2 d_{max}^2 + 20\gamma^2)\mathbb{E}[\|\hat{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2] \\
&+ 10c^2\sigma_A^2\mathbb{E}[\|\mathbf{A}\mathbf{x}^k\|^2] \\
&+ (10L^2 + 20\gamma^2)\mathbb{E}[\|\mathbf{x}^k - \mathbf{s}^k\|^2] + 10\kappa^2\mathbb{E}[\|\hat{\mathbf{x}}^{k+1} - \mathbf{z}^k\|^2],
\end{aligned}
$$

where $d_{\max}$ is the largest value of matrix $\mathbf{D}$, the first equality is due to the optimal condition for (55), i.e., $\hat{\mathbf{x}}^{k+1} = \mathrm{prox}_r^1(\hat{\mathbf{x}}^{k+1} - \nabla_{\mathbf{x}} g(\hat{\mathbf{x}}^{k+1}, \mathbf{w}^k; \boldsymbol{\lambda}^{k+1})$; the first inequality is owing to the nonexpansive property of the proximal operator; the second inequality dues to $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$; the second equality is obtained by the definition of function g in (57); the last inequality dues to the L-smooth in (33).

Next, we show the upper bound for $\|\mathbf{A}\mathbf{x}^k\|$ as

$$
\begin{aligned}
&\mathbb{E}[\|\mathbf{A}\mathbf{x}^k\|^2] \\
&\leq 2\mathbb{E}[\|\mathbf{A}\mathbf{x}(\mathbf{z}^k, \boldsymbol{\lambda}^{k+1})\|^2] + 2\sigma_A^2\mathbb{E}[\|\mathbf{x}^k - \mathbf{x}(\mathbf{z}^k, \boldsymbol{\lambda}^{k+1})\|^2] \\
&\leq 2\mathbb{E}[\|\mathbf{A}\mathbf{x}(\mathbf{z}^k, \boldsymbol{\lambda}^{k+1})\|^2] + 4\sigma_A^2\sigma_1^2\mathbb{E}[\|\mathbf{x}^k - \hat{\mathbf{x}}^{k+1}\|^2] \\
&+ 4\sigma_A^2\sigma_2^2\mathbb{E}[\|\mathbf{x}^k - \mathbf{s}^k\|^2], \qquad (100)
\end{aligned}
$$

where the last inequality comes from Lemma 2. Now, we consider the upper bound of (98). Using the above inequalities (99)-(100), we can obtain

$$
\begin{aligned}
&\min_{k=0,\ldots,K-1}\mathbb{E}[Q(\mathbf{x}^k, \boldsymbol{\lambda}^{k+1})] \\
&\leq \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\mathbf{x}^k - \mathrm{prox}_r^1(\mathbf{x}^k - \nabla_{\mathbf{x}} f(\mathbf{x}^k) - \mathbf{A}^\top\boldsymbol{\lambda}^{k+1})\|^2] \\
&+ \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\mathbf{A}\mathbf{x}^k\|^2] \\
&\leq \frac{K_1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\mathbf{x}^k - \hat{\mathbf{x}}^{k+1}\|^2 + \frac{K_2}{K}\sum_{k=0}^{K-1}\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2] \\
&+ \frac{K_3}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\hat{\mathbf{x}}^{k+1} - \mathbf{z}^k\|^2] + \frac{K_4}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\mathbf{A}\mathbf{x}(\mathbf{z}^k, \boldsymbol{\lambda}^{k+1})\|^2].
\end{aligned}
$$

where

$$
\begin{aligned}
&K_1 = 6 + 40\gamma + 20c^2 d_{\max}^2 + 4(20c^2\sigma_A^2 + 1)\sigma_A^2\sigma_1^2, \\
&K_2 = (20L^2 + 20\gamma^2)\bar{\eta}^2 + 4(20c^2\sigma_A^2 + 1)\sigma_A^2\sigma_2^2\bar{\eta}^2, \\
&K_3 = 20\kappa^2, K_4 = 2(20c^2\sigma_A^2 + 1).
\end{aligned}
$$

Further applying (97), we have

$$
\begin{aligned}
\min_{k=0,\ldots,K-1}\mathbb{E}[Q(\mathbf{x}^k, \boldsymbol{\lambda}^{k+1})] &\leq C_0\left(\frac{\mathbb{E}[\phi^0 - \phi^K]}{K} + \frac{C_1 N\sigma^2}{|\mathcal{I}|}\right) \\
&\leq C_0\left(\frac{\phi^0 - \underline{f}}{K} + \frac{C_1 N\sigma^2}{|\mathcal{I}|}\right),
\end{aligned}
$$

where $\underline{f}$ is the lower bound of ϕ and C_0 is defined as follows,

$$
C_0 \triangleq \frac{2K_1}{\hat{\sigma}_1} + \frac{K_2}{\hat{\sigma}_2} + \frac{4K_3\beta}{\kappa(1-\beta)} + \frac{2K_4}{\alpha}, \qquad (101)
$$

APPENDIX C
PROOF OF THEOREM 2

Proof: If we know the full gradient $\nabla f(\mathbf{x}^k)$, i.e., $G(\mathbf{x}^k, \boldsymbol{\xi}^k) = \nabla f(\mathbf{x}^k)$, then $\sigma^2 = 0$. Substituting it into (86), we have

$$\begin{aligned}&\phi^k - \phi^{k+1}\\ &\geq \frac{\kappa(1-\beta)\beta}{4}\left\|\hat{\mathbf{x}}^{k+1} - \mathbf{z}^k\right\|^2 + \frac{\alpha}{2}\left\|\mathbf{A}\mathbf{x}(\mathbf{z}^k, \boldsymbol{\lambda}^{k+1})\right\|^2\\ &+ \frac{\hat{\sigma}_1}{2}\left\|\mathbf{x}^k - \hat{\mathbf{x}}^{k+1}\right\|^2 + \frac{\hat{\sigma}_2}{2}\left\|\mathbf{x}^k - \mathbf{x}^{k-1}\right\|^2 \geq 0.\end{aligned}$$

Thus ϕ^k is monotonically decreasing and it has lower bound $\underline{f}$. This implies that

$$\max\{\|\mathbf{x}^k - \hat{\mathbf{x}}^{k+1}\|, \|\mathbf{z}^k - \hat{\mathbf{x}}^{k+1}\|, \|\mathbf{A}\mathbf{x}(\mathbf{z}^k; \boldsymbol{\lambda}^{k+1})\|\} \to 0.$$

Thus, according to [19, Theorem 2.4], every limit point generated by PPDM algorithm is a KKT point of problem (5). In addition, substituting $\sigma^2 = 0$ into (42) and picking $K \geq \frac{C_0(\phi^0 - \underline{f})}{\epsilon}$, we have

$$\min_{k=0,\ldots,K-1} Q(\mathbf{x}^k, \boldsymbol{\lambda}^{k+1}) \leq C_0\left(\frac{\phi^0 - \underline{f}}{K}\right) \leq \epsilon.$$

Therefore, the proof is completed. ■

REFERENCES

[1] S. Boyd *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[2] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surv. Tut.*, vol. 19, no. 3, pp. 1628–1656, Jul.-Sep. 2017.

[3] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments-part I: Agreement at a linear rate," *IEEE Trans. Signal Process.*, vol. 69, pp. 1242–1256, 2021.

[4] T.-H. Chang, M. Hong, and X. Wang, "Multi-agent distributed optimization via inexact consensus ADMM," *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 482–497, Jan. 2015.

[5] T.-H. Chang, M. Hong, H.-T. Wai, X. Zhang, and S. Lu, "Distributed learning in the non-convex world: From batch to streaming data, and beyond," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 26–38, May 2020.

[6] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.

[7] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An exact first-order algorithm for decentralized consensus optimization," *SIAM J. Optim.*, vol. 25, no. 2, pp. 944–966, 2015.

[8] W. Shi, Q. Ling, G. Wu, and W. Yin, "A proximal gradient algorithm for decentralized composite optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 22, pp. 6013–6023, Nov. 2015.

[9] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the ADMM in decentralized consensus optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1750–1761, Apr. 2014.

[10] S. A. Alghunaim, E. K. Ryu, K. Yuan, and A. H. Sayed, "Decentralized proximal gradient algorithms with linear convergence rates," in *IEEE Transactions on Automatic Control*, vol. 66, no. 6, pp. 2787–2794, 2020.

[11] J. Xu, Y. Tian, Y. Sun, and G. Scutari, "Distributed algorithms for composite optimization: Unified and tight convergence analysis," 2020, *arXiv:2002.11534*.

[12] J. Zeng and W. Yin, "On nonconvex decentralized gradient descent," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2834–2848, Jun. 2018.

[13] P.D. Lorenzo and G. Scutari, "NEXT: In-network nonconvex optimization," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 2, pp. 120–136, Jun. 2016.

[14] G. Scutari and Y. Sun, "Parallel and distributed successive convex approximation methods for big-data optimization," *Multi-agent Optimization*, pp. 141–308, 2018.

[15] G. Scutari and Y. Sun, "Distributed nonconvex constrained optimization over time-varying digraphs," *Math. Program.*, vol. 176, no. 1-2, pp. 497–544, 2019.

[16] M. Hong, D. Hajinezhad, and M.-M. Zhao, "Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1529–1538.

[17] H. Sun and M. Hong, "Distributed non-convex first-order optimization and information processing: Lower complexity bounds and rate optimal algorithms," *IEEE Trans. Signal Process.*, vol. 67, no. 22, pp. 5912–5928, Nov. 2019.

[18] J. Zhang and Z.-Q. Luo, "A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization," *SIAM J. Optim.*, vol. 30, no. 3, pp. 2272–2302, 2020.

[19] J. Zhang and Z. Luo, "A global dual error bound and its application to the analysis of linearly constrained nonconvex optimization," 2020, *arXiv:2006.16440*.

[20] M. Hong and T.-H. Chang, "Stochastic proximal gradient consensus over random networks," *IEEE Trans. Signal Process.*, vol. 65, no. 11, pp. 2933–2948, Jun. 2017.

[21] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5330–5340.

[22] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[23] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," 2018, *arXiv:1812.06127*.

[24] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5132–5143.

[25] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, "D^2: Decentralized training over decentralized data," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4848–4856.

[26] S. Lu, X. Zhang, H. Sun, and M. Hong, "GNSD: A gradient-tracking based nonconvex stochastic algorithm for decentralized optimization," in *Proc. IEEE Data Sci. Workshop*, 2019, pp. 315–321..

[27] P. Bianchi and J. Jakubowicz, "Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization," *IEEE Trans. Autom. Control*, vol. 58, no. 2, pp. 391–405, Feb. 2013.

[28] D. Jakovetić, J. Xavier, and J. M. Moura, "Fast distributed gradient methods," *IEEE Trans. Autom. Control*, vol. 59, no. 5, pp. 1131–1146, May 2014.

[29] H. Li, C. Fang, W. Yin, and Z. Lin, "A sharp convergence rate analysis for distributed accelerated gradient methods," 2018, *arXiv:1810.01053*.

[30] Y. Yan, T. Yang, Z. Li, Q. Lin, and Y. Yang, "A unified analysis of stochastic momentum methods for deep learning," 2018, *arXiv:1808.10396*.

[31] H. Yu, R. Jin, and S. Yang, "On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 7184–7193.

[32] Z. Wang, J. Zhang, T.-H. Chang, J. Li, and Z.-Q. Luo, "Distributed stochastic consensus optimization with momentum for nonconvex nonsmooth problems," 2020, *arXiv:2011.05082*.

[33] F. Huang and S. Chen, "Mini-batch stochastic ADMMs for nonconvex nonsmooth optimization," 2018, *arXiv:1802.03284*.

[34] M. Hong and Z.-Q. Luo, "On the linear convergence of the alternating direction method of multipliers," *Math. Program.*, vol. 162, no. 1-2, pp. 165–199, 2017.

[35] Y. Xu, S. Zhu, S. Yang, C. Zhang, R. Jin, and T. Yang, "Learning with non-convex truncated losses by SGD," in *Proc. Uncertainty Artif. Intell.*, 2020, pp. 701–711.

[36] B. Li, S. Cen, Y. Chen, and Y. Chi, "Communication-efficient distributed optimization in networks with gradient tracking and variance reduction," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 1662–1672.

[37] B. Wang, T. M. Nguyen, A. L. Bertozzi, R. G. Baraniuk, and S. J. Osher, "Scheduled restart momentum for accelerated stochastic gradient descent," 2020, *arXiv:2002.10583*.

[38] J.-S. Pang, "A posteriori error bounds for the linearly-constrained variational inequality problem," *Math. Operations Res.*, vol. 12, no. 3, pp. 474–484, 1987.

[39] Z. Wang, J. Zhang, T.-H. Chang, J. Li, and Z.-Q. Luo, "Supplementary material for distributed consensus optimization with momentum

for nonconvex nonsmooth problems," 2020. [Online]. Available: https://www.researchgate.net/publication/343418255

[40] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course.* Dordrecht, The Netherlands: Kluwer Academic, 2004.

[41] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA:Athena Scientific, 1999.

Zhiguo Wang received the Ph.D. degree from Sichuan University, Chengdu, China, in 2018. He is currently a Postdoctor with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China. His current research interests include information fusion, target tracking, machine learning, nonconvex optimization, and analysis and control of uncertain systems.

Jiawei Zhang is currently working toward the Ph.D. degree with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China. His research interests include optimization theory and designing algorithms for machine learning, data science with theoretical guarantees.

Tsung-Hui Chang (Senior Member, IEEE) received the B.S. degree in electrical engineering and the Ph.D. degree in communications engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2003 and 2008, respectively. He is currently an Associate Professor with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China. Previously, he was a Postdoctoral Researcher with NTHU during 2008–2011, the University of California, Davis, CA, USA, during 2011–2012, and was a Faculty Member of the National Taiwan University of Science and Technology (NTUST), Taipei, Taiwan, during 2012–2015. His research interests include data communications, machine learning, and large-scale optimization. He was the recipient of the Young Scholar Research Award of NTUST in 2014, IEEE Communication Society Asian-Pacific Outstanding Young Researcher Award in 2015, and the IEEE Signal Processing Society Best Paper Award in 2018. He was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING during August 2014–December 2018 and IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS during January 2015– December 2018. He is currently an Associate Editor for the IEEE OPEN JOURNAL OF SIGNAL PROCESSING from Janutary 2020 to present and a Senior Area Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING from February 2021 to present. He is also an Elected Member of the IEEE SPS Signal Processing for Communications and Networking Technical Committee (SPCOM TC) (January 2020–).

Jian Li (Fellow, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering from The Ohio State University, Columbus, OH, USA, in 1987 and 1991, respectively. She is currently a Professor with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA. His publications include the *Robust Adaptive Beamforming (2005, Wiley)*, *Spectral Analysis: the Missing Data Case (2005, Morgan & Claypool)*, *MIMO Radar Signal Processing (2009, Wiley)*, and *Waveform Design for Active Sensing Systems – A Computational Approach (2011, Cambridge University Press)*. Her current research interests include spectral estimation, statistical and array signal processing, and their applications to radar, sonar, and biomedical engineering.

She is a Fellow of the European Academy of Sciences (Brussels) and a Fellow of IET. She was the recipient of the 1994 National Science Foundation Young Investigator Award and the 1996 Office of Naval Research Young Investigator Award. She was an Executive Committee Member of the 2002 International Conference on Acoustics, Speech, and Signal Processing, Orlando, Florida, May 2002. She was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 1999 to 2005, an Associate Editor for the IEEE SIGNAL PROCESSING MAGAZINE from 2003 to 2005, and a Member of the Editorial Board of Signal Processing, a publication of the European Association for Signal Processing (EURASIP), from 2005 to 2007. From 2010 to 2012, she was a Member of the Editorial Board of the IEEE SIGNAL PROCESSING MAGAZINE. She was a Member of the Sensor Array and Multichannel Technical Committee of the IEEE Signal Processing Society for 12 years. She is the coauthor of the paper that was the recipient of the M. Barry Carlton Award for the best paper published in IEEE Transactions on Aerospace and Electronic Systems in 2005. She is also the coauthor of a paper published in IEEE Transactions on Signal processing that was the recipient of the Best Paper Award in 2013 from the IEEE Signal Processing Society.

Zhi-Quan Luo (Fellow, IEEE) received the B.S. degree in applied mathematics from Peking University, Beijing, China, and the Ph.D. degree in operations research from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1989. From 1989 to 2003, he held a Faculty position with the Department of Electronics and Communication Engineering, McMaster University, Hamilton, ON, Canada. From 2001 to 2003, he held a tier-1 Canada Research Chair in information processing. After that, he has been a Full Professor with the Department of Electronics and Communication Engineering, University of Minnesota, Minneapolis, MN, USA, and held an endowed ADC Chair in digital technology. He is currently the Vice President (Academic) of the The Chinese University of Hong Kong, Shenzhen, China, and the Director of Shenzhen Research Institute of Big Data (SRIBD). He has authored or coauthored more than 300 refereed papers, books, and special issues. His research interests mainly include mathematical issues in information sciences, with particular focus on the design, analysis and applications of large-scale optimization algorithms. He is a Fellow of SIAM and is selected to the Royal Society of Canada. He was the recipient of the four best paper awards from the IEEE Signal Processing Society, One Best Paper Award from EUSIPCO, the Farkas Prize from INFORMS and the prize of Paul Y. Tseng Memorial Lectureship in Continuous Optimization, and some best paper awards from international conferences. He has served as an Associate Editor for many internationally recognized journals and the Editor-in-Chief of the IEEE TRANSACTIONS ON SIGNAL PROCESSING.