Accelerated Projection Algorithm Based on Smoothing Approximation for Distributed Nonsmooth Optimization

You Zhao, Xiaofeng Liao, *Fellow, IEEE*, and Xing He

***Abstract*—In this article, a distributed smoothing accelerated projection algorithm (DSAPA) is proposed to address constrained nonsmooth convex optimization problems over undirected multiagent networks in a distributed manner, where the objective function is free of the assumption of the Lipschitz gradient or strong convexity. First, based on a distributed exact penalty method, the original optimization problem is translated to a problem of standard assignment without consensus constraints. Then, a novel DSAPA by combining the smoothing approximation with Nesterov's accelerated schemes, is proposed. In addition, we provide a systematic analysis to derive an upper bound on the convergence rate in terms of the objective function based on penalty function and to choose the optimal step size accordingly. Our results demonstrate that the proposed DSAPA can reach $O(\log(k)/k)$ when the optimal step size is chosen. Finally, the effectiveness and correctness of the proposed algorithm are verified by numerical and practical application examples.**

***Index Terms*—Convergence rate, nonsmooth convex, projection operators, smoothing approximation.**

I. Introduction

DISTRIBUTED optimization has always received much attention from a mass of communities, such as sensor networks [1], [2]; machine learning [3]; and sparse signal reconstruction [4], [5], etc. Unlike centralized optimization settings, distributed optimization has solved some engineering problems in a distributed manner, i.e., each agent can only access local information about itself and its neighbors, but not global information. Thus, the aim of the distributed optimization is, in general, to minimize the global objective which is only accessible by the cooperation between individual agent and their neighbors upon multiagent networks. Until now, numerous distributed optimization algorithms have been presented, which can be generally classified into primal consensus algorithms and dual decomposition algorithms. Typical algorithms include the distributed subgradient algorithm [6], distributed primal-dual subgradient algorithm [7], asynchronous distributed gradient algorithm [8], distributed push-sum subgradient algorithm [9], distributed dual proximal gradient (DDPG) algorithm [10], EXTRA [11], D-ADMM [12], primal-dual method of multipliers (PDMM) [13], and so on.

The gradient-based algorithms are widely used to design optimization algorithms due to nonexpensive computational costs. Nedić et al. [6] first designed a distributed (sub)gradient algorithm (DGA) by combining a weighted averaging of local information with gradient descent algorithm. Later, the distributed projected subgradient algorithm (DPGA) was proposed in [14] based on DGA to solve distributed nonsmooth convex optimization problems with set constraints. On the basis of push-sum technology or surplus value, DPGA was further employed for distributed convex optimization problems under directed communication networks equipped with row random weight matrices in [9] and [15]. On the basis of the Nesterov-type accelerated technique, the fast distributed gradient method and its proximal version have been investigated in [16]. However, in the above algorithms, they precisely converge to the optimal solution, but require either a decreasing step size, or a significant increase in the multistep average communication burden in each iteration. These issues can be solved by introducing the gradient tracking scheme [19], [20], [21]. In addition, a fixed step size can be selected to obtain the same convergence rate as the centralized algorithms.

Convergence rate is used as a basic criteria for evaluating the performance of the distributed algorithms for the convex optimization problems. DGA [6] has a convergence rate of $O(\log(k)/\sqrt{k})$ in solving convex optimization problems with nonsmooth convex objective functions, and that of $O(\log(k)/k)$ [22] for strongly convex functions with the Lipschitz gradient assumption. The distributed dual proximal gradient (DDPG) algorithm in [10] has a convergence rate of $O(1/k)$ when solving a "nonsmooth convex" + "nonsmooth strongly

Manuscript received 30 June 2022; revised 23 November 2022 and 24 November 2022; accepted 8 January 2023. Date of publication 17 January 2023; date of current version 5 December 2023. This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0100101, in part by the National Natural Science Foundation of China under Grant 61932006 and Grant 62176027, and in part by the Fundamental Research Funds for the Central Universities, China, under Grant XDJK2020TY003. Recommended by Associate Editor Y. Wang. *(Corresponding author: Xiaofeng Liao.)*

You Zhao and Xiaofeng Liao are with the Key Laboratory of Dependable Services Computing in Cyber Physical Society-Ministry of Education, College of Computer Science, Chongqing University, Chongqing 400044, China (e-mail: zhaoyou1991sdtz@cqu.edu.cn; xfliao@cqu.edu.cn).

Xing He is with the Chongqing Key Laboratory of Nonlinear Circuits and Intelligent Information Processing, School of Electronics and Information Engineering, Southwest University, Chongqing 400715, China (e-mail: hexingdoc@swu.edu.cn).

Digital Object Identifier 10.1109/TCNS.2023.3237496

convex" optimization problem. If the convex objective function has a Lipschitz continuous gradient, a sublinear convergence rate of $O(1/k)$ can be achieved with a fixed step size in [11] and [21]. Moreover, a fast distributed gradient algorithm was proposed and investigated in [16], which has a convergence rate of $O(1/k^{2-\xi})$ for the convex objective function with the Lipschitz gradient assumption, where ξ is an arbitrarily small positive constant. Furthermore, when the objective function is assumed to be strongly convex, the linear convergence rate can be obtained in [11], [21], and [22].

Nevertheless, with respect to nonsmooth convex optimization problems, the decaying step size is inevitable for designing optimization algorithms due to its gradient being non-Lipschitz continuous (i.e., the appropriate step size selection depends heavily on the Lipschitz constant of its gradient). Liu et al. [25] revisited the DPGA under undirected graphs in [14] and obtained its optimal convergence rate $O(1/\sqrt{k})$ when selecting step size $\frac{1}{k}$. Xi et al. [15] proposed a directed-distributed projected subgradient algorithm (D-DPGA) for a constrained nonsmooth convex optimization problem under directed graph and its convergence rate $O(\log(k)/\sqrt{k})$ when selecting a diminishing step size. Li et al. [24] studied a distributed projected subgradient algorithm to tackle a nonsmooth convex optimization problem when the communication topology is time-varying, unbalanced, and directed, and only provided the convergence of their proposed algorithm. Yin et al. [17] presented an asynchronous distributed dual algorithm based on edges that minimize nondifferential convex optimization problems with partially overlapping dependence with a sublinear convergence rate. Shi and Yang [18] proposed surplus-based dual averaging (SDA) for solving a nonsmooth constrained convex optimization over a weight-unbalanced directed graph and obtained a convergence rate $O(1/\sqrt{k})$.

The nonsmooth convex optimization problem, as a crucial class of convex optimization problems, plays an important role in many engineering applications, such as image decomposition and reduction (e.g., L_1-norm minimization problem), visual coding, compressed perception (e.g., Lasso problem), geological exploration (e.g., reconstruction of micro pore structure of shale), wireless communication (e.g., sensor localization problem), artificial intelligence (e.g., L_1-regularized empirical loss models), gene expression analysis, and risk management. Therefore, the nonsmooth convex optimization problems have drawn more and more attention due to the wide range of applications. However, how to design a faster distributed accelerated algorithm to address the constrained nonsmooth convex optimization problems is still challenging. Usually, there are two difficulties as follows.

1) The first dilemma is that the gradient of the nonsmooth objective function is non-Lipschitz continuous, which hinders the selection of the fixed step size of designing distributed algorithms. How to effectively solve this problem is the key point to achieve acceleration and convergence of the proposed algorithms.
2) There exist both optimization variable consensus and set constraints in the considered optimization problems. It is difficult to deal with the above constraints effectively in designing distributed accelerated algorithms.

Inspired by the works in [33], [34], [35], [36], and [37], we focus on designing an accelerated distributed projection algorithm based on the smoothing approximation scheme and Nesterov's accelerated method for the nonsmooth convex optimization problems with local set constraints, which does not need to make the assumptions of the Lipschitz gradient and strong convexity. The contributions are summarized as follows.

1) A novel distributed smoothing accelerated projection algorithm (DSAPA) is developed for solving nonsmooth constrained convex problems in a distributed manner with an accelerated convergence rate, where a distributed exact penalty function method and projection operators are used to deal with the optimization variable consensus and set constraints, and the Nesterov's accelerated strategy (extrapolation or momentum) is utilized to accelerate the convergence rate of DSAPA. We theoretically prove the range of values of the exact penalty parameter, discuss the step size selection condition of the DSAPA, and offer the optimal step size of DSAPA to obtain the optimal accelerated convergence rate.
2) In contrast with the existing state-of-the-art distributed algorithms in [14], [15], [23], [25], and [27], the proposed DSAPA has a faster convergence rate of $O(\log(k)/k)$ when selecting an optimal smoothing approximation parameter, which does not need to make the assumptions of strongly convex and Lipschitz gradient (see Table I). To the best of our knowledge, this is a lower iteration complexity achieved so far for the considered nonsmooth distributed optimization problems without the strong convexity and Lipschitz gradient assumptions.
3) Different from the distributed algorithms in [6] and [9] based on subgradients, our proposed DSAPA avoids the difficulty of derivative selection in the nondifferentiable point. Compared with distributed algorithms in [27] and [28] based on the Fenchel-dual method, distributed algorithms based on proximal operators [29], [30], our proposed DSAPA has no assumptions of closed-form solutions for the conjugate function and the proximal operators of the nonsmooth objective functions. Thus, the DSAPA has wider applicability.
4) Compared with the algorithms in [31] and [32] based on the Nesterov's smoothing method to address nonsmooth optimization problems, the DSAPA allows convergence to the optimal solution for nonsmooth optimization problems when choosing an appropriate parameter, while the algorithms in [31] and [32] only converge to an ϵ-optimal solution (i.e., $\{x|f(x) - f(x^*) \leqslant \epsilon, \epsilon > 0\}$).

The rest of this article is organized as follows. Section II introduces some necessary preliminaries. In Section III, the distributed optimization problem with set constraint is reformulated by exact the nonsmooth penalty method. In Section IV, a novel distributed smoothing accelerated projection algorithm (DSAPA) is proposed and its convergence rate is also analyzed carefully. In Section V, some experimental results are obtained to

TABLE I
COMPARISONS OF DIFFERENT ALGORITHMS

Algorithm	Objective function: convex	Constraint	Stepsize	Gradient	Convergence rate
DGA [6][1]	non-smooth	No	fixed ($\alpha > 0$)	subgradient	$O\left(\frac{1}{\alpha k} + \alpha \mathrm{C}\right)$
Subgradient-push [9]	non-smooth	No	decreasing	subgradient	$O\left(\frac{\log(k)}{\sqrt{k}}\right)$
Gradient-push [23]	smooth strongly	No	decreasing	gradient	$O\left(\frac{\log(k)}{k}\right)$
EXTRA [11]	smooth	No	fixed	gradient	$O\left(\frac{1}{\sqrt{k}}\right)$
DGD [22]	smooth	No	fixed	gradient	$O\left(\frac{1}{k}\right)$
FDGM [17] [2]	smooth	No	fixed	gradient	$O\left(\frac{1}{k^{2-\xi}}\right)$
DDPG [10]	non-smooth strongly+non-smooth	Yes	fixed	proximal gradient	$O\left(\frac{1}{k}\right)$
D-DPS [16]	non-smooth	Yes	decreasing	subgradient	$O\left(\frac{\log(k)}{\sqrt{k}}\right)$
DPGA [26]	non-smooth	Yes	decreasing	subgradient	$O\left(\frac{1}{\sqrt{k}}\right)$
SDA [19]	non-smooth	Yes	decreasing	subgradient	$O\left(\frac{1}{\sqrt{k}}\right)$
DSAPA (12)[3]	non-smooth	Yes	decreasing	smooth approximate gradient	$\begin{cases} O\left(\frac{1}{k^{\theta}}\right), \text{if } \theta \in (0,1) \\ O\left(\frac{\log(k)}{k}\right), \text{if } \theta = 1 \end{cases}$

[1] α and C is a positive constant that depends on Lipscitiz-constant of gradient (From TABLE I, we can see that the DGA in [6] has a convergence rate $O\left(\frac{1}{\alpha k} + \alpha \mathrm{C}\right)$ at first glance. It is worth noting that DGA in [6] equipped with a fixed stepsize $\alpha > 0$ does not converge to the optimal solution of problem (8) without constraint because αC is a positive constant. Although DGA in [6] does not look perfect now, it is a very important result in the development of distributed optimization algorithms, and based on the results in [6], a large number of distributed algorithms have been studied); [2] $\xi > 0$ arbitrarily small; [3] $\theta \in (0, 1]$ is a parameter of decreasing of step size.

demonstrate the effectiveness of the proposed DSAPA. Finally, Section VI concludes this article.

Notations: Let R be the set of real numbers and R^n be a set of n-dimensional column vectors. The superscript T indicates the transpose operation. $\|x\|_1 = \sum_{i=1}^n |x_i|$ denotes the 1-norm of x. $\|x\| = (\sum_{i=1}^n x_i^2)^{\frac{1}{2}}$ denotes the Euclidean norm. $\|x\|_0$ is the 0-norm. Let $\mathbf{1}, \mathbf{0}$ be a vector with all entries being 1 and 0, respectively. For $\boldsymbol{\mu} = \mathrm{diag}\{\mu_1, \ldots, \mu_m\} \in R^{mn \times mn}$ with $\mu_i \in R^n, i = 1, \ldots, m$, one has $\frac{1}{\boldsymbol{\mu}} = \mathrm{diag}\{\frac{1}{\mu_1}, \ldots, \frac{1}{\mu_m}\}$ and $\boldsymbol{L} \in R^{m \times m}, \boldsymbol{L} = \mathrm{diag}\{L_1, \ldots, L_m\}$, $\frac{\boldsymbol{\mu}}{\boldsymbol{L}} = \mathrm{diag}\{\frac{\mu_1}{L_1}, \ldots, \frac{\mu_m}{L_m}\}$ and $\|x\|_{\boldsymbol{L}}^2 = x^T \boldsymbol{L} x$. $x \preccurlyeq y$ means $x_i \leqslant y_i, i = 1, \ldots, m$. Let $\Omega_1 \times \ldots \times \Omega_m$ be the Cartesian product of sets $\Omega_1, \ldots, \Omega_m$. For $x \in R$, $\mathrm{sgn}(x) = -1$, if $x < 0$, $\mathrm{sgn}(x) = 1$, if $x > 0$, and $\mathrm{sgn}(x) = [-1, 1]$, if $x = 0$.

II. PRELIMINARY RESULTS

A. Projection Operator

For a nonempty, closed, and convex set $\Omega \in R^n$, the projection operator of Ω is defined by $P_\Omega(x) = \arg\min_{u \in \Omega} \|u - x\|$. The normal cone of set Ω is given by $N_\Omega(x) = \{v \in R^n | v^T(y - x) \leqslant 0 \forall y \in \Omega\}$, where "cl$(\cdot)$" is the closures of set $\cdot$, and $d_\Omega(x)$ is $\min_{u \in \Omega} \|x - u\|$.

Lemma 1: When the constrained set Ω is a box or affine set, there exists a closed-form solution of its projection operator as follows:1) If Ω is a box set, i.e., $\Omega = \{x \in R^n \mid x_{i,\min} \leq x_i \leq x_{i,\max}, i = 1, \ldots, n\}$, then

$$P_\Omega(x_i) = \max\{\min\{x_i, x_{i,\max}\}, x_{i,\min}\}. \tag{1}$$

2) If Ω is an affine set, i.e., $\Omega = \{x \in R^n \mid Ax = b\}$ and A satisfies $A \in R^{m \times n}$, and $\mathrm{Rank}(A) = m$, such that

$$P_\Omega(x) = x + A^T\left(AA^T\right)^{-1}(b - Ax). \tag{2}$$

B. Convex Analysis

Let $g(x): R^n \to R$ be a locally Lipschitz function and D_g be a set at which g is differentiable, the Clarke generalized gradient of $g(x)$ is given by

$$\partial g(x) = \mathrm{co}\left\{\lim_{x_k \to x: x_k \in D_g} \nabla \partial g(x_k)\right\}$$

where co(S) is the convex hull of set S.

A function $g: \Omega \to R$ is generally convex (possible nonsmooth) if it satisfies

$$g(u) - g(w) \geqslant (u - w)^T g_f(w) \forall u, w \in \Omega \tag{3}$$

where $g_f(w) \in \partial g$ and $\Omega \subset R^n$ is a convex set.

C. Smoothing Approximation

Definition 1: [35] Let $\hat{g}: \Omega \subset R^n \times (0, +\infty) \to R$ be a smoothing function of g, where $g: \Omega \subset R^n \to R$ is locally Lipschitz and $\hat{g}$ enjoys the following properties.

i) (Continuous differentiable property) $\hat{g}(\cdot, \mu)$ is continuous differentiable in R^m with any fixed $\mu > 0$, and $\hat{g}(x, \cdot)$ is differentiable in $(0, +\infty]$ with any fixed $x \in \Omega \subset R^n$.

ii) (Approximation property) $\lim_{\mu \to 0^+} \hat{g}(x, \mu) = g(x)$ for any fixed $x \in \Omega \subset R^n$.

iii) (Gradient boundedness with respect to μ) There exists a positive $\kappa_{\hat{g}} > 0$, such that

$$|\nabla_\mu \hat{g}(x, \mu)| \leq \kappa_{\hat{g}} \quad \forall \mu \in (0, +\infty), \ x \in \Omega \subset R^n.$$

iv) (Gradient consistency) $\{\lim_{z \to x, \mu \to 0} \nabla_z \hat{g}(z, \mu)\} \subseteq \partial g(x)$.

In addition, for any fixed $x \in R^n$, smoothing function $\hat{g}$ satisfies the following properties.

v) (General approximation property) $\lim_{z \to x, \mu \to 0} \hat{g}(z, \mu) = g(x)$.

vi) (Lipschitz continuous with respect of μ) $|\hat{g}(x, \mu_1) - \hat{g}(x, \mu_2)| \leq \kappa_{\hat{g}} |\mu_1 - \mu_2| \ \forall \mu_1, \mu_2 \in (0, \infty], x \in \Omega \subset R^n$.

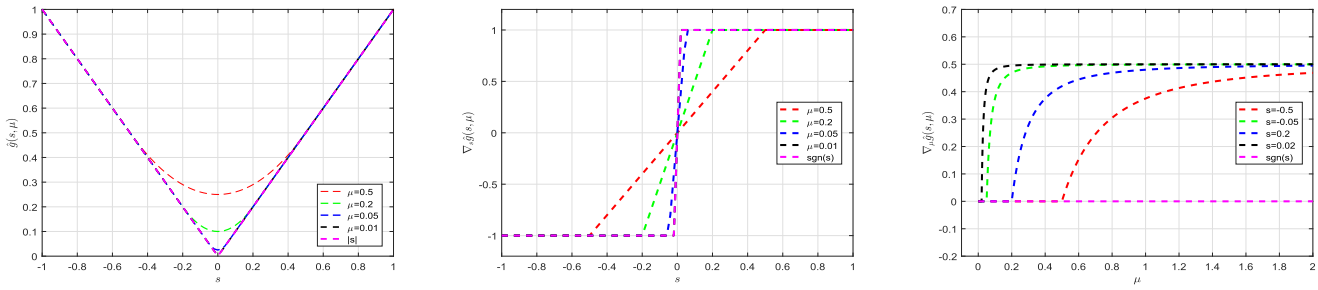

Fig. 1. Smoothing function $\hat{g}(s,\mu)$ with different μ (left). The gradient of $\nabla_s \hat{g}(s,\mu)$ with different μ (middle). The gradient of $\nabla_\mu \hat{g}(s,\mu)$ with different s (right).

vii) (Lipschitz continuous of gradient) $\hat{g}(x,\mu)$ is a convex function, and there exists $\mathfrak{l} > 0$, such that $\|\nabla \hat{g}(x,\mu) - \nabla \hat{g}(y,\mu)\| \leq \frac{\mathfrak{l}}{\mu}\|x-y\| \forall x, y \in \Omega \subset R^n$.

Next, an important property of the combination of smoothing functions is provided by the following.

Lemma 2 ([36]): Let $\hat{g}_1, \ldots, \hat{g}_m$ be smoothing functions of $g_1, \ldots, g_m$, $a_i \geq 0$ and g_i be regular for any $i = 1, 2, \ldots, m$, then $\sum_{i=1}^m a_i \hat{g}_i$ is a smoothing function of $\sum_{i=1}^m a_i g_i$ with $\kappa_{\sum_{i=1}^n a_i \hat{g}_i} = \sum_{i=1}^m a_i \kappa_{\hat{g}_i}$.

A smoothing approximation function of $|s|, s \in R$, presented in [35] is used in this article as follows:

$$\hat{g}(s,\mu) = \begin{cases} |s|, & \text{if } |s| > \mu \\ \frac{s^2}{2\mu} + \frac{\mu}{2}, & \text{if } |s| \leq \mu. \end{cases} \tag{4}$$

Note that $\lim_{\mu \to 0^+} \hat{g}(s,\mu) = |s|$ [see Fig. 1 (left)] from ii) in Definition 1.

Moreover, we can obtain the derivative of $\hat{g}(s,\mu)$ with respect to s [see Fig. 1 (middle)] as follows:

$$\nabla_s \hat{g}(s,\mu) = \begin{cases} \text{sgn}(s), & \text{if } |s| > \mu \\ \frac{s}{\mu}, & \text{if } |s| \leq \mu. \end{cases} \tag{5}$$

The derivative of function $\hat{g}(s,\mu)$ with respect to μ [see Fig. 1 (right)] is represented as

$$\nabla_\mu \hat{g}(s,\mu) = \begin{cases} 0, & \text{if } |s| > \mu \\ \frac{1}{2} - \frac{s^2}{2\mu^2}, & \text{if } |s| \leq \mu. \end{cases}$$

D. Graph Theory

An undirected communication topology graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ of order m consists of node set $\mathcal{V} = \{v_1, v_2, \ldots, v_m\}$, edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, and $\mathcal{A} = \{a_{ij}\}_{m \times m}$ with nonnegative elements $a_{ij} = a_{ji} > 0$ if $(i,j) \in \mathcal{E}$ and $a_{ij} = a_{ji} = 0$ otherwise. The couples among the agents are unordered in the undirected graph, which implies that the information is exchanged between agent i and agent j. The undirected graph path between agent i and agent j is a sequence of edges from (i, i_1), (i_1, i_2), $\ldots$, (i_s, j), where $i_1, \cdots, i_s$, j denotes different agents. Denote $\mathcal{N}_i = \{j | (i,j) \in \mathcal{E}\}$ as the set of the neighbors of agent i. If any pair of distinct nodes i and j $(i, j = 1, 2, \ldots, m)$ exists as a path between them, then an undirected graph $\mathcal{G}$ is connected.

III. Nonsmooth Convex Optimization Problem and Its Reformulation

A. Problem Formulation

Consider a scenario that involves in m agents under an undirected network. For every agent, there exists a local nonsmooth convex objective function $f_i : R^n \to R$ and a local nonempty, closed, and convex feasible constraint $\Omega_i \in R^n$. With all agents working together with their neighbors to achieve a consensus solution to optimize the global objective function $\sum_{i=1}^m f_i(x)$ in the constraint set $\bigcap_{i=1}^m \Omega_i$. Thus, the optimization problem can be modeled as follows:

$$\min \bar{F}(x) = \sum_{i=1}^m f_i(x), \text{ s.t. } x \in \bigcap_{i=1}^m \Omega_i. \tag{6}$$

Note that the problem (6) encompasses many practical application problems, such as signal processing and sensor network localization problems.

Moreover, the local constraints are necessary or unavoidable in light of limitations of the agent's performance in computational and communication capabilities.

Assumption 1: The function f_i is said to be Lipschitz continuous on the set $\Omega_i \forall i \in 1, \ldots, m$, if it satisfies

$$\|f_i(u) - f_i(v)\| \leqslant l_i \|u - v\|, u, v \in \Omega_i, i = 1, \ldots, m \tag{7}$$

with a Lipschitz constant $l_i > 0$.

Note that the Assumption 1 is easy to be satisfied in practice due to the set $\Omega_i \forall i \in 1, \ldots, m$ is a nonempty, closed, and convex set.

Assumption 2: $f_i, i = 1, \ldots, m$ is nonsmooth convex, i.e., it satisfies (3).

Assumption 3: The undirected communication topology graph is connected.

B. Reformulation

Note that the problem (6) is not a standard distributed problem. Under Assumption 3, the problem (6) can be equivalent to the following distributed form:

$$\begin{aligned} \min \ & F(x) = \sum_{i=1}^m f_i(x_i) \\ \text{s.t. } & x_i \in \Omega_i \subset R^n, x_i = x_j, i \in \mathcal{V}_i, j \in \mathcal{N}_i \end{aligned} \tag{8}$$

where $\mathcal{N}_i$ is the neighbor set of agent i. Further, the penalty method is used to transform the problem (8) into

$$\begin{aligned} \min\ \ & \Gamma(x)=\sum_{i=1}^{m} f_i(x_i)+\frac{\Upsilon}{2}\sum_{i=1}^{m}\sum_{j\in\mathcal{N}_i}\|x_i-x_j\|_1 \\ \text{s.t.}\ \ & x_i\in\Omega_i, i\in\mathcal{V}_i \end{aligned} \tag{9}$$

where $\|\cdot\|_1$ is 1-norm, $\Upsilon>0$ is a penalty parameter.

Lemma 3 (Sufficient condition): If Assumption 1 and Assumption 3 hold and $\Upsilon\geqslant\sqrt{m}\max\limits_{1\leqslant i\leqslant m}\{l_i\}$, then, the optimal solution x^* of the problem (9) is equivalent to that in problem (6).

Proof: Let $\bar{x}=1/m\sum_{i=1}^{m}x_i$ and $D^2(x)!=\sum_{i=1}^{m}\|x_i-\bar{x}\|^2\leqslant\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}\|x_i-x_j\|^2\leqslant\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}\|x_i-x_j\|_1^2$. Moreover, for any $p,q\in\mathcal{V}$, there exists a path $\mathcal{P}_{pq}\subset\mathcal{E}$ owing to Assumption 3 holds, such that

$$\begin{aligned} h(x) &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j\in\mathcal{N}_i}\|x_i-x_j\|_1=\frac{1}{2}\sum_{(p,q)\in\mathcal{E}}\|x_p-x_q\|_1 \\ &\geqslant \frac{1}{2}\sum_{(p,q)\in\mathcal{P}_{ij}}\|x_p-x_q\|_1\geqslant\|x_i-x_j\|_1. \end{aligned} \tag{10}$$

Furthermore, we have $D(x)^2\leqslant mh^2(x)\Rightarrow D(x)\leqslant\sqrt{m}h(x)$.

Let $\Upsilon\geqslant\sqrt{m}\max\limits_{1\leqslant i\leqslant m}\{l_i\}$, thus we have

$$\begin{aligned} F(x)+\Upsilon h(x) &\geqslant F(x)+\max_{1\leqslant i\leqslant m}\{l_i\}D(x) \\ &= F(\bar{x})+F(x)-F(\bar{x})+\max_{1\leqslant i\leqslant m}\{l_i\}D(x)\geqslant F(\bar{x}). \end{aligned} \tag{11}$$

The first inequality holds which comes from the condition (10) and the second inequality is satisfied due to the Lipschitz continuous property in (7). From (11), one has that $\min F(x)+\Upsilon h(x)\geqslant\min\limits_{x_i=x_j}F(\bar{x})$. This means that the equation holds if and only if $x_i=\bar{x}, i\in\mathcal{V}$. ■

IV. Main Results

For problem (9) with nonsmooth generalized convex functions, we propose the following distributed smoothing accelerated projection algorithm (DSAPA).

A. Distributed Smoothing Accelerated Projection Algorithm (DSAPA)

For agent $i=1,\ldots,m$, the DSAPA is described as follows:

$$\begin{cases} x_i^{k+1}=P_{\Omega_i}\left(y_i^k-\frac{\mu_i^k}{L_i}\left(\nabla\hat{f}_i\left(y_i^k,\mu_i^k\right)\right.\right. \\ \qquad\left.\left.+\Upsilon\sum_{j\in\mathcal{N}_i}\nabla\hat{h}_i\left(y_i^k-y_j^k,\mu_i^k\right)\right)\right) \\ y_i^{k+1}=x_i^{k+1}+\alpha_i^{k+1}\left(x_i^{k+1}-x_i^k\right) \\ \alpha_i^{k+1}=\frac{t_i^k-1}{t_i^{k+1}},\mu_i^k=\frac{\mu_i^0}{(k+1)^\theta},\ \theta\in(0,1] \\ t_i^{k+1}=\frac{a}{2}+\sqrt{\left(\frac{\mu_i^k}{\mu_i^{k+1}}\right)t_i^k},\ a\in(0,1] \\ x_i^0=y_i^0\in\Omega_i, t_i^0=1,\mu_i^0\in(0,+\infty) \end{cases} \tag{12}$$

where $L_i=\frac{(\mathfrak{l}_i+\Upsilon|\mathcal{N}_i|)}{\mu_i}$ with $\mathfrak{l}_i$ is a Lipschitz constant of $\hat{f}_i$ and $|\mathcal{N}_i|$ means the number of neighbors of the agent i

$$\nabla_y\hat{h}_i\left(y_i^k-y_j^k,\mu_i^k\right)=\sum_{\tau=1}^{n}\nabla_y\hat{h}_{i,\tau}\left(y_{i,\tau}^k-y_{j,\tau}^k,\mu_i^k\right)$$

$$\nabla_y\hat{h}_{i,\tau}\left(y_{i,\tau}^k-y_{j,\tau}^k,\mu_i^k\right)=\begin{cases}\operatorname{sgn}\left(y_{i,\tau}^k-y_{j,\tau}^k\right), \text{if } \left|y_{i,\tau}^k-y_{j,\tau}^k\right|>\mu_i^k \\ \frac{y_{i,\tau}^k-y_{j,\tau}^k}{\mu_i^k}, \qquad \text{if } \left|y_{i,\tau}^k-y_{j,\tau}^k\right|\leqslant\mu_i^k.\end{cases}$$

The compact form of DSAPA (12) is given by

$$\begin{cases} y^{k+1}=x^{k+1}+\boldsymbol{\alpha}^{k+1}\left(x^{k+1}-x^k\right) \\ x^{k+1}=P_\Omega\left(y^k-\frac{\boldsymbol{\mu}^k}{\boldsymbol{L}}\nabla\hat{\Gamma}\left(y^k,\boldsymbol{\mu}^k\right)\right)\end{cases} \tag{13}$$

where $\boldsymbol{\alpha}^{k+1}=\operatorname{diag}(\alpha_1^{k+1},\ldots,\alpha_m^{k+1})$, $\boldsymbol{\mu}^k=\operatorname{diag}(\mu_1^k,\ldots,\mu_m^k)$, and $\frac{\boldsymbol{\mu}^k}{\boldsymbol{L}}=\operatorname{diag}\left(\frac{\mu_1^k}{L_1},\ldots,\frac{\mu_m^k}{L_m}\right)$.

Remark 1: It should be noted that the parameter Υ of DSAPA (12) is related to the Lipschitz constant of objective functions since $\Upsilon\geqslant\sqrt{m}\max\limits_{1\leqslant i\leqslant m}\{l_i\}$. For $\max\limits_{1\leqslant i\leqslant m}\{l_i\}$, it can be obtained in the following distributed manner, i.e., the agent $i, i=1,\ldots,n$ only communicate with their neighbors (the agent j, $j\in\mathcal{N}_i$) once after running the $\max\{\cdot\}$ operation of l_i, l_j. In addition, the parameters θ, a, and m of DSAPA (12) can also be obtained by using some existing distributed algorithms in advance, for example, decentralized minimum-time consensus [26] can be used to estimate θ, a, and m in finite-time under a distributed manner.

Proposition 1: The sequences α^k and t^k satisfy the following equation.

(i) For $k\geqslant1$, $i=1,\ldots,m$, $\frac{\sqrt{\mu_i^0}}{2(2-\theta)}(k+2)^{1-\frac{\theta}{2}}\leqslant t_i^{k+1}\sqrt{\mu_i^{k+1}}\leqslant\frac{(4-\theta)\sqrt{\mu_i^0}}{2-\theta}(k+1)^{1-\frac{\theta}{2}}$.

(ii) $t_i^{k+1}(\mu_i^{k+1})^2\leqslant\frac{(4-\theta)(\mu_i^0)^2}{2-\theta}(k+1)^{1-2\theta}$, $i=1,\ldots,m$.

(iii) $(\alpha_i^k)^2\leqslant\frac{\mu_i^{k+1}}{\mu_i^k}$, $i=1,\ldots,m$.

(iv) $\mu_i^{k+1}(t_i^{k+1})(t_i^{k+1}-1)\leqslant\mu_i^k(t_i^k)^2$, $i=1,\ldots,m$.

(v) $(\alpha_i^{k+1})^2\leqslant(\frac{\mu_i^{k+1}}{\mu_i^k})-\sqrt{(\frac{\mu_i^{k+1}}{\mu_i^k})}\frac{1}{t_i^{k+1}}$ for $k\geqslant\frac{2^{1+\frac{\theta}{2}}}{a^2}$, $i=1,\ldots,m$.

Proof: The proof is shown in Appendix A. ■

Proposition 2: Let $q(y,\boldsymbol{\mu})=P_\Omega(y-\frac{\boldsymbol{\mu}}{\boldsymbol{L}}\nabla\Gamma(y,\mu))$, then for any $x,y\in R^{mn}$, and $\mu_{i,j}\in(0,\infty), i=1,\ldots,m, j=1,\ldots,n$, the following condition holds:

$$\begin{aligned} &\hat{\Gamma}(x,\boldsymbol{\mu})-\hat{\Gamma}(q(y,\boldsymbol{\mu}),\boldsymbol{\mu})+\mathbf{1}^T\frac{\boldsymbol{\mu}}{\boldsymbol{L}}\mathrm{I}_\Omega(x) \\ &\quad\geqslant(x-q(y,\boldsymbol{\mu}))^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}}(x-q(y,\boldsymbol{\mu})) \\ &\quad\quad-(x-y)^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}}(x-y) \end{aligned} \tag{14}$$

where $\hat{\Gamma}(q(y,\boldsymbol{\mu}),\boldsymbol{\mu})=\hat{f}(q(y,\boldsymbol{\mu}),\boldsymbol{\mu})+\Upsilon\hat{h}(q(y,\boldsymbol{\mu}),\boldsymbol{\mu})$, $\hat{\Gamma}(x,\boldsymbol{\mu})=\hat{f}(x,\boldsymbol{\mu})+\Upsilon\hat{h}(x,\boldsymbol{\mu})$, $\mathrm{I}_\Omega(x)=(\mathrm{I}_{\Omega_1}(x_1),\ldots,\mathrm{I}_{\Omega_m}(x_m))^T\in R^{mn}$, $\mathrm{I}_{\Omega_i}(x_i)=\begin{cases}\mathbf{0},\mathbf{x_i}\in\Omega_\mathbf{i}, \\ +\infty, x_i\notin\Omega_i,\end{cases}$ $i=1,\ldots,m$.

Proof: For proof, see Appendix B. ■

Proposition 3: Let $W(k)=\hat{\Gamma}(x^k,\boldsymbol{\mu}^k)+\kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^k\mathbf{1}+(x^k-x^{k-1})^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^{k-1}}(x^k-x^{k-1})$, then it is nonincreasing and $\lim_{k\to+\infty}W(k)$ exists. In addition, the sequence $\{x^k\}$ is bounded.

Proof: The proof can be found in Appendix C. ■

Theorem 1: Under Assumptions 1–3 and let $\Upsilon\geqslant\sqrt{m}\max_{1\leqslant i\leqslant m}\{l_i\}$, any clustering point $\{x^k\}$ generated by DSAPA (13) is an optimal solution to problem (9), i.e., problem (8), and the DSAPA (13) has an arithmetical convergence rate

$$f\left(x^k\right)+\Upsilon h\left(x^k\right)-\left(f\left(x^*\right)+\Upsilon h\left(x^*\right)\right)=\begin{cases}O\left(\frac{1}{k^\theta}\right),\text{if }\theta\in(0,1)\\ O\left(\frac{\log(k)}{k}\right),\text{if }\theta=1.\end{cases}$$

Proof: Since $\mathrm{I}_\Omega(x^k)=\mathbf{0}$ and let $q(y^k,\boldsymbol{\mu}^k)=x^{k+1}$ in (45), one can get

$$\begin{aligned}&\hat{\Gamma}\left(x^k,\boldsymbol{\mu}^k\right)-\hat{\Gamma}\left(x^{k+1},\boldsymbol{\mu}^k\right)\\&=\hat{f}\left(x^k,\boldsymbol{\mu}^k\right)+\Upsilon\hat{h}\left(x^k,\boldsymbol{\mu}^k\right)\\&\quad-\left(\hat{f}\left(x^{k+1},\boldsymbol{\mu}^k\right)+\Upsilon\hat{h}\left(x^{k+1},\boldsymbol{\mu}^k\right)\right)\\&\geqslant\left(y^k-x^{k+1}\right)^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^k}\left(y^k-x^{k+1}\right)\\&\quad+\left(x^k-y^k\right)^T\frac{\boldsymbol{L}}{\boldsymbol{\mu}^k}\left(y^k-x^{k+1}\right).\end{aligned}\tag{15}$$

Using $\kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^k\mathbf{1}\geqslant\kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^{k+1}\mathbf{1}$ and letting $z^k=\hat{\Gamma}(x^k,\boldsymbol{\mu}^k)+\kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^k\mathbf{1}-\Gamma(x^*)$ with $x^*=\min_{x\in\Omega}\Gamma(x)$, we deduce that

$$\begin{aligned}z^k-z^{k+1}&\geqslant\left(y^k-x^{k+1}\right)^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^k}\left(y^k-x^{k+1}\right)\\&\quad+\left(x^k-y^k\right)^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^k}\left(y^k-x^{k+1}\right).\end{aligned}\tag{16}$$

Next, let $x=x^*$, $y=y^k$ and $\boldsymbol{\mu}=\boldsymbol{\mu}^k$ in (14), one has

$$\begin{aligned}&\hat{\Gamma}\left(x^*,\boldsymbol{\mu}^k\right)-\hat{\Gamma}\left(x^{k+1},\boldsymbol{\mu}^k\right)\\&=\hat{f}\left(x^*,\boldsymbol{\mu}^k\right)+\Upsilon\hat{h}\left(x^*,\boldsymbol{\mu}^k\right)\\&\quad-\left(\hat{f}\left(x^{k+1},\boldsymbol{\mu}^k\right)+\Upsilon\hat{h}\left(x^{k+1},\boldsymbol{\mu}^k\right)\right)\\&\geqslant\left(y^k-x^{k+1}\right)^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^k}\left(y^k-x^{k+1}\right)\\&\quad+\left(x^*-y^k\right)^T\frac{\boldsymbol{L}}{\boldsymbol{\mu}^k}\left(y^k-x^{k+1}\right)\end{aligned}\tag{17}$$

due to $\mathrm{I}_\Omega(x^*)=\mathbf{0}$.

In addition, from (vi) in Definition 1, i.e., $|\hat{\Gamma}(x^*,\boldsymbol{\mu}^k)-\Gamma(x^*)|\leqslant\kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^k\mathbf{1}$, $|\hat{\Gamma}(x^{k+1},\boldsymbol{\mu}^{k+1})-\hat{\Gamma}(x^{k+1},\boldsymbol{\mu}^k)|\leqslant\kappa_{\hat{\Gamma}}\mathbf{1}^T(\boldsymbol{\mu}^k-\boldsymbol{\mu}^{k+1})\mathbf{1}$, the condition $\hat{\Gamma}(x^*,\boldsymbol{\mu}^k)-\hat{\Gamma}(x^{k+1},\boldsymbol{\mu}^k)$ becomes

$$\hat{\Gamma}\left(x^*,\boldsymbol{\mu}^k\right)-\hat{\Gamma}\left(x^{k+1},\boldsymbol{\mu}^k\right)\leqslant-z^{k+1}+2\kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^k\mathbf{1}.\tag{18}$$

Replacing (18) into (17) yields

$$\begin{aligned}&-z^{k+1}+2\kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^k\mathbf{1}\\&\geqslant\left(y^k-x^{k+1}\right)^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^k}\left(y^k-x^{k+1}\right)\\&\quad+\left(x^*-y^k\right)^T\frac{\boldsymbol{L}}{\boldsymbol{\mu}^k}\left(y^k-x^{k+1}\right).\end{aligned}\tag{19}$$

Multiplying (16) by $\mathbf{1}^T\boldsymbol{t}^k(\boldsymbol{t}^k-I)$ and (19) by $\mathbf{1}^T\boldsymbol{t}^k$, then adding them together to get

$$\begin{aligned}&\mathbf{1}^T\boldsymbol{t}^k\left(\boldsymbol{t}^k-I\right)z^k-\mathbf{1}^T\left(\boldsymbol{t}^k\right)^2z^{k+1}+2\kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{t}^k\boldsymbol{\mu}^k\mathbf{1}\\&\geqslant\left(y^k-x^{k+1}\right)^T\frac{\boldsymbol{L}\left(\boldsymbol{t}^k\right)^2}{2\boldsymbol{\mu}^k}\left(y^k-x^{k+1}\right)\\&\quad+\left(\left(\boldsymbol{t}^k-I\right)x^k-\boldsymbol{t}^ky^k+x^*\right)^T\frac{\boldsymbol{L}\boldsymbol{t}^k}{\boldsymbol{\mu}^k}\left(y^k-x^{k+1}\right).\end{aligned}\tag{20}$$

Since

$$\begin{aligned}&\|\mathfrak{b}-\mathfrak{a}\|^2_{\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^k}}+2\left(\mathfrak{a}-\mathfrak{b}\right)^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^k}\left(\mathfrak{c}-\mathfrak{a}\right)\\&=\|\mathfrak{b}-\mathfrak{c}\|^2_{\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^k}}-\|\mathfrak{a}-\mathfrak{c}\|^2_{\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^k}}\end{aligned}$$

with $\|\mathfrak{b}-\mathfrak{a}\|^2_{\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^k}}=(\mathfrak{b}-\mathfrak{a})^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^k}(\mathfrak{b}-\mathfrak{a})$.

Setting $w^k=\boldsymbol{t}^{k-1}x^k-(\boldsymbol{t}^{k-1}-I)x^{k-1}-x^*$, $w^{k+1}=\boldsymbol{t}^kx^{k+1}-(\boldsymbol{t}^k-I)x^k-x^*$ in the right hand of (20), we get

$$\begin{aligned}&\left(y^k-x^{k+1}\right)\frac{\boldsymbol{L}\left(\boldsymbol{t}^k\right)^2}{2\boldsymbol{\mu}^k}\left(y^k-x^{k+1}\right)\\&\quad+\left(\left(\boldsymbol{t}^k-I\right)x^k-\boldsymbol{t}^ky^k+x^*\right)^T\frac{\boldsymbol{L}\boldsymbol{t}^k}{\boldsymbol{\mu}^k}\left(y^k-x^{k+1}\right)\\&=\left(w^{k+1}\right)^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^k}w^{k+1}-\left(w^k\right)^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^k}w^k.\end{aligned}\tag{21}$$

By combining (21) and (20) with (iv) in Proposition 1, i.e., $\mu_i^kt_i^k(t_i^k-1)\leqslant\mu_i^{k-1}(t_i^{k-1})^2, i=1,\ldots,m$ and $z^{k+1}\succcurlyeq\mathbf{0}$, we have

$$\begin{aligned}&\left(w^{k+1}\right)^T\frac{\boldsymbol{L}}{2}w^{k+1}+\mathbf{1}^T\boldsymbol{\mu}^k\left(\boldsymbol{t}^k\right)^2z^{k+1}\\&\leqslant\left(w^k\right)^T\frac{\boldsymbol{L}}{2}w^k+\mathbf{1}^T\boldsymbol{\mu}^{k-1}\left(\boldsymbol{t}^{k-1}\right)^2z^k\\&\quad+2\kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{t}^k\left(\boldsymbol{\mu}^k\right)^2\mathbf{1}.\end{aligned}\tag{22}$$

By performing recursion on (22), one has

$$\begin{aligned}&\left(w^{k+1}\right)^T\frac{\boldsymbol{L}}{2}w^{k+1}+\mathbf{1}^T\boldsymbol{\mu}^k\left(\boldsymbol{t}^k\right)^2z^{k+1}\\&\leqslant\left(w^1\right)^T\frac{\boldsymbol{L}}{2}w^1+\mathbf{1}^T\boldsymbol{\mu}^0\left(\boldsymbol{t}^0\right)^2z^1\\&\quad+2\kappa_{\hat{\Gamma}}\sum_{j=1}^k\mathbf{1}^T\boldsymbol{t}^i\left(\boldsymbol{\mu}^j\right)^2\mathbf{1}\end{aligned}\tag{23}$$

where the first and second items in the right-hand of (23) becomes

$$\begin{aligned}
&\left(x^*-x^1\right)^T\frac{\boldsymbol{L}}{2}\left(x^*-x^1\right)\\
&\quad+\mathbf{1}^T\boldsymbol{\mu}^0\mathbf{1}\left(\hat{f}\left(x^*,\boldsymbol{\mu}^0\right)+\Upsilon\hat{h}\left(x^*,\boldsymbol{\mu}^0\right)\right.\\
&\quad\left.-f\left(x^*\right)-\Upsilon h\left(x^*\right)+\kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^1\mathbf{1}\right)\\
&\leqslant\left(x^*-y^0\right)^T\frac{\boldsymbol{L}}{2}\left(x^*-y^0\right)\\
&\quad+\mathbf{1}^T\boldsymbol{\mu}^0\mathbf{1}\left(\hat{f}\left(x^*,\boldsymbol{\mu}^0\right)+\Upsilon\hat{h}\left(x^*,\boldsymbol{\mu}^0\right)\right.\\
&\quad\left.-f\left(x^*\right)-\Upsilon h\left(x^*\right)+\kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^1\mathbf{1}\right)\\
&\leqslant\left(x^*-y^0\right)^T\frac{\boldsymbol{L}}{2}\left(x^*-y^0\right)+2\kappa_{\hat{\Gamma}}\mathbf{1}^T\left(\boldsymbol{\mu}^0\right)^2\mathbf{1}
\end{aligned}\tag{24}$$

where the first inequality holds by using (14) and $\boldsymbol{\mu}^1\preccurlyeq\boldsymbol{\mu}^0$, and the second inequality is satisfied since $\hat{\Gamma}(x^*,\boldsymbol{\mu}^0)-\Gamma(x^*)\leqslant\kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^0\mathbf{1}$.

In addition, based on (ii) in Proposition 1, the third item in the right-hand side of (23) becomes

$$\begin{aligned}
&2\sum_{j=1}^{k}t^j\kappa_{\hat{\Gamma}}\mathbf{1}^T\left(\boldsymbol{\mu}^i\right)^2\mathbf{1}\\
&\leqslant\frac{2\kappa_{\hat{\Gamma}}\left(4-\theta\right)\mathbf{1}^T\left(\boldsymbol{\mu}^0\right)^2\mathbf{1}}{\left(2-\theta\right)}\int_1^{k+1}s^{1-2\theta}ds\\
&\leqslant\frac{2\kappa_{\hat{\Gamma}}\left(4-\theta\right)\mathbf{1}^T\left(\boldsymbol{\mu}^0\right)^2\mathbf{1}}{\left(2-\theta\right)\left(2-2\theta\right)}\left(k+1\right)^{2-2\theta},\text{if }\theta\in(0,1)
\end{aligned}$$

or

$$\begin{aligned}
&2\sum_{j=1}^{k}t^j\kappa_{\hat{\Gamma}}\mathbf{1}^T\left(\boldsymbol{\mu}^j\right)^2\mathbf{1}\\
&\leqslant\frac{6\kappa_{\hat{\Gamma}}\mathbf{1}^T\left(\boldsymbol{\mu}^0\right)^2\mathbf{1}}{2-\theta}\sum_{i=1}^{k}\left(k+1\right)^{-1}\\
&\leqslant\frac{6\kappa_{\hat{\Gamma}}\mathbf{1}^T\left(\boldsymbol{\mu}^0\right)^2\mathbf{1}}{2-\theta}\int_1^{k+1}s^{-1}ds\\
&\leqslant\frac{6\kappa_{\hat{\Gamma}}\mathbf{1}^T\left(\boldsymbol{\mu}^0\right)^2\mathbf{1}}{2-\theta}\log\left(k+1\right),\text{if }\theta=1
\end{aligned}$$

that is

$$\begin{aligned}
&2\sum_{j=1}^{k}t^j\kappa_{\hat{\Gamma}}\mathbf{1}^T\left(\boldsymbol{\mu}^j\right)^2\mathbf{1}\\
&\leqslant\begin{cases}\frac{2\kappa_{\hat{\Gamma}}(4-\theta)\mathbf{1}^T\left(\boldsymbol{\mu}^0\right)^2\mathbf{1}}{(2-\theta)(2-2\theta)}\left(k+1\right)^{2-2\theta},&\text{if }\theta\in(0,1)\\6\kappa_{\hat{\Gamma}}\mathbf{1}^T\left(\boldsymbol{\mu}^0\right)^2\mathbf{1}\log\left(k+1\right),&\text{if }\theta=1.\end{cases}
\end{aligned}\tag{25}$$

According to (24) and (25), the inequality (22) can be rewritten as

$$\begin{aligned}
&\left(w^{k+1}\right)^T\frac{\boldsymbol{L}}{2}w^{k+1}+\mathbf{1}^T\boldsymbol{\mu}^k\left(\boldsymbol{t}^k\right)^2z^{k+1}\\
&\leqslant C+\begin{cases}\frac{2\kappa_{\hat{\Gamma}}(4-\theta)\mathbf{1}^T\left(\boldsymbol{\mu}^0\right)^2\mathbf{1}}{(2-\theta)(2-2\theta)}\left(k+1\right)^{2-2\theta},&\text{if }\theta\in(0,1)\\6\kappa_{\hat{\Gamma}}\mathbf{1}^T\left(\boldsymbol{\mu}^0\right)^2\mathbf{1}\log\left(k+1\right),&\text{if }\theta=1\end{cases}
\end{aligned}\tag{26}$$

where $C=(x^*-y^0)^T\frac{\boldsymbol{L}}{2}(x^*-y^0)+2\kappa_{\hat{\Gamma}}\mathbf{1}^T(\boldsymbol{\mu}^0)^2\mathbf{1}$.

Moreover, from $\mathbf{1}^T\boldsymbol{\mu}^k(\boldsymbol{t}^k)^2\mathbf{1}\geqslant\frac{a^2\mathbf{1}^T\boldsymbol{\mu}^0\mathbf{1}}{4(2-\theta)^2}(k+1)^{2-\theta}$ in (i) of Proposition 1 and $\hat{\Gamma}(x^{k+1},\boldsymbol{\mu}^{k+1})-\Gamma(x^*)+\kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^{k+1}\mathbf{1}\geqslant\Gamma(x^{k+1})-\Gamma(x^*)$, we obtain

$$\begin{aligned}
&\Gamma\left(x^{k+1}\right)-\Gamma\left(x^*\right)\\
&=f\left(x^{k+1}\right)+\Upsilon h\left(x^{k+1}\right)-f\left(x^*\right)-\Upsilon h\left(x^*\right)\\
&\leqslant\frac{2\left(2-\theta\right)^2}{a^2\mathbf{1}^T\boldsymbol{\mu}^0\mathbf{1}}\left(k+1\right)^{\theta-2}\left(x^*-y^0\right)^T\boldsymbol{L}\left(x^*-y^0\right)\\
&\quad+\frac{8\left(2-\theta\right)^2\kappa_{\hat{\Gamma}}}{a^2}\mathbf{1}^T\boldsymbol{\mu}^0\mathbf{1}\left(k+1\right)^{\theta-2}\\
&\quad+\begin{cases}\frac{8\kappa_{\hat{\Gamma}}(2-\theta)(4-\theta)\mathbf{1}^T\left(\boldsymbol{\mu}^0\right)\mathbf{1}}{(2-\theta)a^2}\left(k+1\right)^{-\theta},&\text{if }\theta\in(0,1)\\\frac{24\kappa_{\hat{\Gamma}}\mathbf{1}^T\left(\boldsymbol{\mu}^0\right)\mathbf{1}\log(k+1)}{(k+1)a^2},&\text{if }\theta=1.\end{cases}
\end{aligned}\tag{27}$$

In the end, we obtain

$$\begin{aligned}
&\Gamma\left(x^k\right)-\Gamma\left(x^*\right)\\
&=f\left(x^k\right)+\Upsilon h\left(x^k\right)-f\left(x^*\right)-\Upsilon h\left(x^*\right)\\
&=\begin{cases}O\left(\frac{1}{k^\theta}\right),&\text{if }\theta\in(0,1)\\O\left(\frac{\log(k)}{k}\right),&\text{if }\theta=1.\end{cases}
\end{aligned}\tag{28}$$

Thus, the proof is completed. ■

Theorem 2: Suppose Assumptions 1–3 hold and let $\Upsilon\geqslant\sqrt{m}\max_{1\leqslant i\leqslant m}\{l_i\}$, the sequence $\{x^k\}$ generated by DSAPA (13) satisfies

$$(\mathrm{I}):\sum_{\mathrm{k}=\mathcal{K}}^{+\infty}\left(\mathrm{x}^{\mathrm{k}}-\mathrm{x}^{\mathrm{k}-1}\right)^{\mathrm{T}}\frac{2^{\theta-1}\boldsymbol{L}}{\left(\boldsymbol{\mu}^{\mathrm{k}}\boldsymbol{t}^{\mathrm{k}}\right)}\left(\mathrm{x}^{\mathrm{k}}-\mathrm{x}^{\mathrm{k}-1}\right)<+\infty$$

$$\text{if }k\geqslant\mathcal{K}=\frac{2^{1+\frac{\theta}{2}}}{a^2}.$$

$$(\mathrm{II}):\left\|\mathrm{x}^{\mathrm{k}+1}-\mathrm{x}^{\mathrm{k}}\right\|^2=\begin{cases}O\left(\frac{1}{(k+1)^{2\theta}}\right),&\text{if }\theta\in(0,1)\\O\left(\frac{\log(k+1)}{(k+1)^2}\right),&\text{if }\theta=1.\end{cases}$$

Proof: (I): With the use of (48), (v) in Proposition 1 and $\frac{-1}{\boldsymbol{\mu}^k\boldsymbol{\mu}^{k-1}}\preccurlyeq\frac{-1}{(\boldsymbol{\mu}^{k-1})^2}$. Then for $k\geqslant\mathcal{K}=\frac{2^{1+\frac{\theta}{2}}}{a^2}$, it follows that

$$\begin{aligned}
&\hat{f}\left(x^k,\boldsymbol{\mu}^k\right)+\Upsilon\hat{h}\left(x^k,\boldsymbol{\mu}^k\right)+\kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^k\mathbf{1}\\
&\quad+\left(x^k-x^{k-1}\right)^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^{k-1}}\left(x^k-x^{k-1}\right)\\
&\quad-\left(\hat{f}\left(x^{k+1},\boldsymbol{\mu}^{k+1}\right)+\Upsilon\hat{h}\left(x^{k+1},\boldsymbol{\mu}^{k+1}\right)\right.\\
&\quad\left.+\kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^{k+1}\mathbf{1}+\left(x^k-x^{k+1}\right)^T\right.
\end{aligned}$$

$$\times \ \frac{\boldsymbol{L}}{2\boldsymbol{\mu}^k}\left(x^k - x^{k+1}\right)\Big) \geqslant \left(x^k - x^{k-1}\right)^T \frac{\boldsymbol{L}}{2\left(\boldsymbol{\mu}^{k-1}\boldsymbol{t}^k\right)}\left(x^k - x^{k-1}\right) \tag{29}$$

i.e.,

$$\left(x^k - x^{k-1}\right)^T \frac{\boldsymbol{L}}{2\left(\boldsymbol{\mu}^{k-1}\boldsymbol{t}^k\right)}\left(x^k - x^{k-1}\right) + W(k+1) \leqslant W(k), \text{if } k \geqslant \mathcal{K}. \tag{30}$$

Since $\mu_i^{k-1} = \frac{\mu_i^0}{(k+1)^\theta} \leqslant \frac{2^\theta \mu_i^0}{(k+2)^\theta} = 2^\theta \mu_i^k$, $i = 1, \ldots, m$, summing the above inequalities from $\mathcal{K}$ to $+\infty$, we have

$$\sum_{k=\mathcal{K}}^{+\infty}\left(x^k - x^{k-1}\right)^T \frac{2^{\theta-1}\boldsymbol{L}}{\left(\boldsymbol{\mu}^k\boldsymbol{t}^k\right)}\left(x^k - x^{k-1}\right) < +\infty. \tag{31}$$

(II): The inequality (26) implies

$$\left(w^{k+1}\right)^T \frac{\boldsymbol{L}}{2} w^{k+1} \leqslant C + \begin{cases} \frac{2\kappa_{\hat{f}}(4-\theta)\mathbf{1}^T\left(\boldsymbol{\mu}^0\right)^2\mathbf{1}}{(2-\theta)(2-2\theta)}(k+1)^{2-2\theta}, & \text{if } \theta \in (0,1) \\ 6\kappa_{\hat{f}}\mathbf{1}^T\left(\boldsymbol{\mu}^0\right)^2\mathbf{1}\log(k+1), & \text{if } \theta = 1 \end{cases} \tag{32}$$

due to $\mathbf{1}^T\boldsymbol{\mu}^k(\boldsymbol{t}^k)^2 z^{k+1}\mathbf{1} \geqslant 0$.

Furthermore, (32) becomes

$$\left(\boldsymbol{t}^k\left(x^{k+1} - x^k\right)\right)^T \frac{\boldsymbol{L}}{2}\left(\boldsymbol{t}^k\left(x^{k+1} - x^k\right)\right) \leqslant B + C + \begin{cases} \frac{2\kappa_{\hat{f}}(4-\theta)\mathbf{1}^T\left(\boldsymbol{\mu}^0\right)^2\mathbf{1}}{(2-\theta)(2-2\theta)}(k+1)^{2-2\theta}, & \text{if } \theta \in (0,1) \\ 6\kappa_{\hat{f}}\mathbf{1}^T\left(\boldsymbol{\mu}^0\right)^2\mathbf{1}\log(k+1), & \text{if } \theta = 1 \end{cases} \tag{33}$$

due to $(x^k - x^*)^T\frac{\boldsymbol{L}}{2}(x^k - x^*) \leqslant \frac{1}{2}\max_{1\leqslant i\leqslant m}\{L_i\}\max_{x\in\Omega}\|x - x^*\| = B < +\infty$ which is satisfied from (50) (the sequence $\{x^k\}$ is bounded).

Since $t_i^k \geqslant \frac{(k+1)a}{2}$, $i = 1, \ldots, m$, we finally get

$$\|x^{k+1} - x^k\|^2 \leqslant \frac{8(B+C)}{\min\limits_{1\leqslant i\leqslant n}\{L_i\}(k+1)^2 a^2} + \begin{cases} \frac{16\kappa_{\hat{f}}(4-\theta)\mathbf{1}^T\left(\boldsymbol{\mu}^0\right)^2\mathbf{1}(k+1)^{-2\theta}}{(2-\theta)(2-2\theta)a^2\min\limits_{1\leqslant i\leqslant n}\{L_i\}}, & \text{if } \theta \in (0,1) \\ \frac{48\kappa_{\hat{f}}\mathbf{1}^T\left(\boldsymbol{\mu}^0\right)^2\mathbf{1}\log(k+1)}{\min\limits_{1\leqslant i\leqslant n}\{L_i\}(k+1)^2 a^2}, & \text{if } \theta = 1 \end{cases} \tag{34}$$

i.e.,

$$\|x^{k+1} - x^k\|^2 = \begin{cases} O\left(\frac{1}{(k+1)^{2\theta}}\right), & \text{if } \theta \in (0,1) \\ O\left(\frac{\log(k+1)}{(k+1)^2}\right), & \text{if } \theta = 1. \end{cases} \tag{35}$$

Thus, the proof is completed. ■

V. Numerical Simulations

In this section, we use DSAPA (13) to deal with distributed constrained optimization problems to demonstrate its effectiveness and superiority.

Example 1: Consider a nonsmooth constrained convex optimization problem over an undirected annular connected graph as follows:

$$\min\ F(x) = \sum_{i=1}^{m}|x - i|, \text{s.t. } x \in \Omega. \tag{36}$$

Let x_i be an estimate variable which is only known to the agent i, the problem (36) can be illustrated as follows: $F(x) = \sum_{i=1}^m f_i(x_i)$, $f_i(x_i) = |x_i - i|$ and the set $\Omega_i = \{x \in R | x_i \geqslant i\}, i = 1, \ldots, 5$, $\Omega_i = \{x \in R | i+1 \geqslant x_i \geqslant i-1\}, i = 6, \ldots, 10$, $\Omega_i = \{x \in R | x_i \leqslant i+2\}$, $i = 11, \ldots, 15$, $\Omega_i = \{x \in R | \|x_i\|_\infty \leqslant 10\}, i = 16, \ldots, 19$. For the simulation, we set $m = 19$ and three step sizes, i.e, $\theta = 0.5, 0.8$, and 1 in DSAPA and make a comparison test with DPGA [25]. Fig. 2 (left) displays the convergence trajectories of x in DSAPA (13) with various parameters ($\theta = 0.5, 0.8, 1$) and DPGA [25] with its optimal step size $\frac{C}{k}$, $C > m$. As can be seen from the results, DSAPA and DPGA [25] can converge to the same optimal solutions. Fig. 2 (middle) shows that the DSAPA (13) a faster rate of convergence when the parameter $\theta \in (0, 1]$ keeps increasing. Moreover, DPGA [25] provides a faster convergence rate than DSAPA (13) with $\theta = 0.5$ and is less than DSAPA (13) with $\theta = 0.8, 1$. Furthermore, Fig. 2 (right) presents convergence rates of $\|x^{k+1} - x^k\|^2$ of DSAPA (13) and DPGA [25].

Example 2: For this example, an classical sparse signal reconstruction problem in compressed sensing is considered as follows:

$$\min_{x\in R^n}\ \|x\|_1, \text{s.t. } Ax = b \tag{37}$$

where $A \in R^{m\times n}$, and $b \in R^m$.

Note that the problem (37) is not a standard distributed optimization problem. According to Assumption 3, distributed consensus theory, and the restricted isometry property in compressed sensing, solving problem (37) is equivalent in solving a distributed optimization problem as follows:

$$\begin{aligned} \min_{X\in R^{Mn}}\ & \sum_{i=1}^{M}\|X_i\|_1 \\ \text{s.t. } & A_iX_i = b_i \in R^{m_i}, i = 1, \ldots, M \\ & X_i = X_j \in R^n, i, j = 1, \ldots, M \\ & A_i \in R^{m_i\times n}, \sum_{i=1}^{M} m_i = m \end{aligned} \tag{38}$$

where the decomposition matrix A by row is shown in Fig. 3.

Applying DSAPA (13) to dispose of the problem (38) with $n = 128$, $m = 50$, and sparsity $s = 10$ under five agents. The (top, left) and (top, middle) in Fig. 4 display the trajectories of x with $\theta = 0.8$ and $\theta = 1$ are globally asymptotically stable. Furthermore, the (bottom, left) and (bottom, right) in Fig. 4 demonstrate that it is feasible to reconstruct the sparse signals in a distributed manner by the stable solution of DSAPA (13). The (top, right) and (bottom, right) in Fig. 4 show that the approach DSAPA (13) with $\theta = 1$ has a faster convergence rate than that

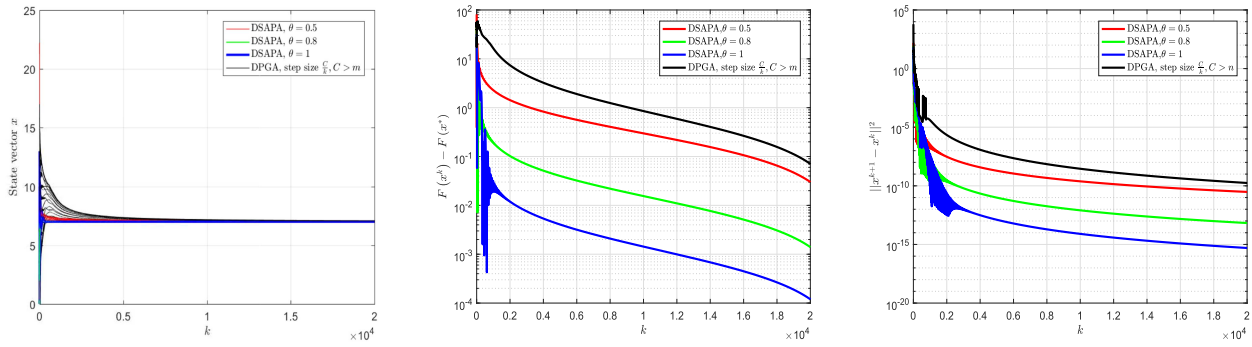

Fig. 2. Transient behaviors of x (left). Convergence rates of $F(x^k) - F(x^*)$ (middle). Convergence rates of $\|x^{k+1} - x^k\|^2$ (right).

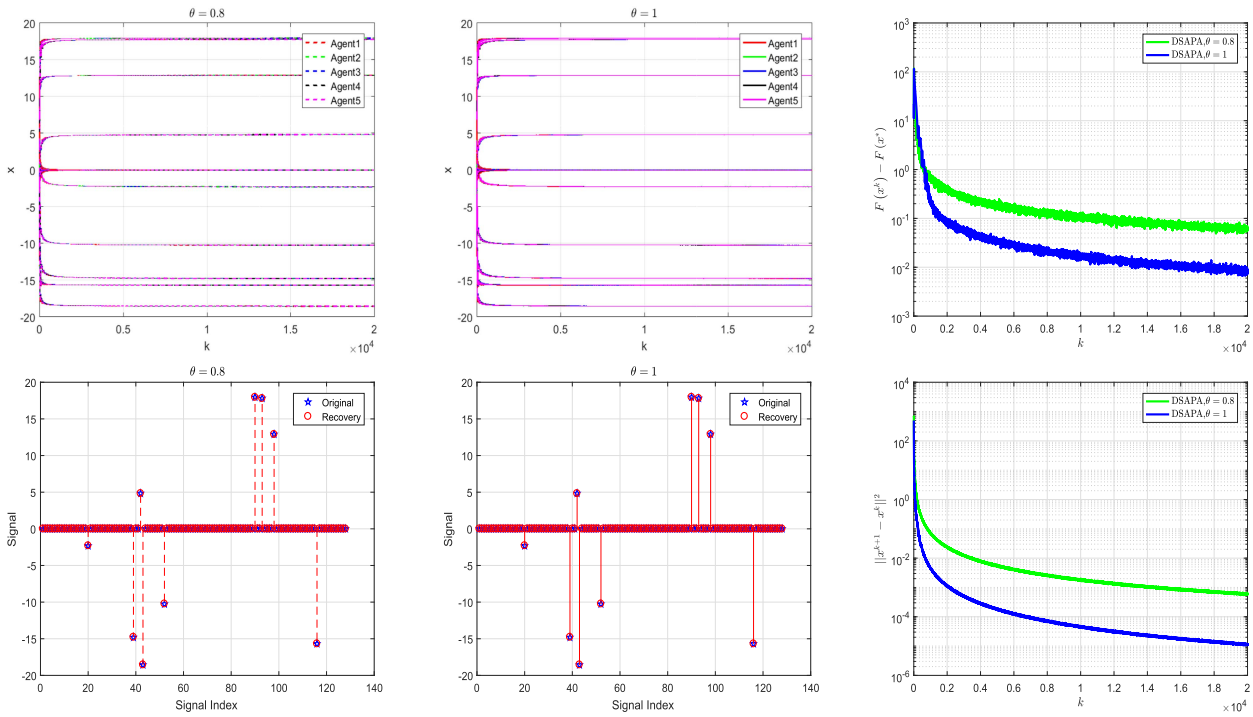

Fig. 3. Transient behaviors of x with $\theta = 0.8$ (top, left). Transient behaviors of x with $\theta = 1$ (top, middle). Convergence rates of $F(x^k) - F(x^*)$ (top, right). Reconstructed signals x with $\theta = 0.8$ (bottom, left). Reconstructed signals x with $\theta = 1$ (bottom, middle). Convergence rates of $\|x^{k+1} - x^k\|^2$ (bottom, right).

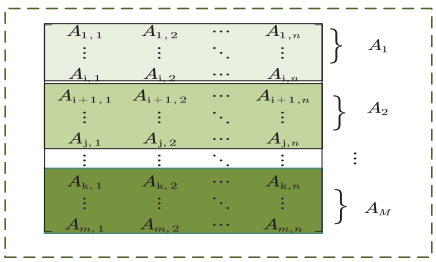

Fig. 4. Decomposition matrix A by row.

with $\theta = 0.8$ which matches the conclusions in Theorems 1 and 2.

Example 3: In this example, the proposed DSAPA (13) is used to optimize a sensor network localization problem in a distributed manner. We consider a scenario in which there are five sensors and ten anchors in the plane R^2. We label the locations of the ten anchors with $b_\iota \in R^2(\iota\{1, 2, \ldots, 10\})$ and mark the positions of the five sensors by $x_i = (x_i^1, x_i^2) \in R^2(i \in \{1, 2, \ldots, 5\})$. Fig. 5 (left) shows the linkages among all sensors and anchors. As can be seen from Fig. 5 (left) that we employ yellow solid squares to indicate the anchors' locations $b_l(\iota \in \{1, 2, \ldots, 10\})$ and red solid pentagrams to represent the optimal anchors' locations $x_i^*(\ i \in \{1, 2, \ldots, 5\})$. Moreover, we use the blue dotted lines to indicate links between sensors, the green dotted lines to indicate the links between sensors and anchors, and the orange rectangle to represent feasible region. It is not difficult to see that the constraint $(\|x_i\|_\infty \leq 2,\ i \in \{1, 2, \ldots, 5\})$ makes the optimal values $x_i^*(i = 1, 2, 3, 4, 5)$ within the feasible domain. There exists an infinite set of specifications for each sensor to limit the sensor locations. The associated goal is to

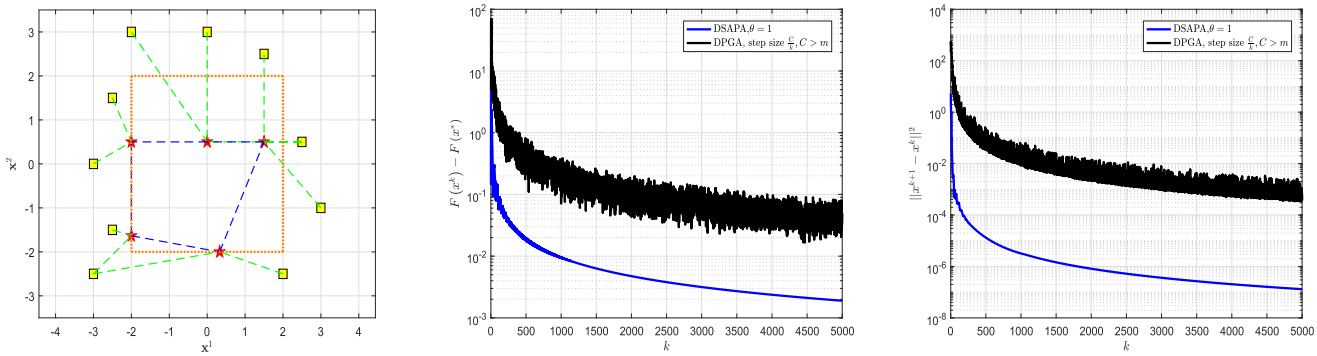

Fig. 5. Topology and optimal location of sensors and anchors (left). Convergence rates of $F(x^k) - F(x^*)$ (middle). Convergence rates of $\|x^{k+1} - x^k\|^2$ (right).

minimize the lengths of the connection between the sensors and the anchors. Thus, the optimization problem is formulated as

$$\min \sum_{i=1}^{5} \frac{1}{2} \left(\sum_{\iota \in \mathcal{N}_i} \|x_i - b_\iota\|_1 + \sum_{j \in \mathcal{N}_i} \|x_i - x_j\|_1 \right)$$
$$\text{s. t. } \|x_i\|_\infty \leq 2,\ i \in \{1, 2, \ldots, 5\}, \iota \in \mathcal{N}_i.$$

Applying DSAPA (13) to deal with this problem. For agent i, the local objective function f_i is

$$f_i(\mathrm{x}) = \frac{1}{2} \left(\sum_{j \in \mathcal{N}_i} \|x_i - x_j\|_1 + \sum_{\iota \in \mathcal{N}_i} \|x_i - b_\iota\|_1 \right)$$

where $\mathrm{x} = (x_1^T, x_2^T, x_3^T, x_4^T, x_5^T)^T$.

The middle and right in Fig. 5 display the convergence rates of $F(x^k) - F(x^*)$ and $\|x^{k+1} - x^k\|^2$, respectively. It is easy to see that the proposed DSAPA (13) has a faster convergence rates of $F(x^k) - F(x^*)$ and $\|x^{k+1} - x^k\|^2$ than DPGA in [25].

VI. Conclusion

In this article, we have proposed a DSAPA for solving constrained nonsmooth convex optimization problems without the Lipschitz gradient and strong convexity assumptions, where the optimization variable consensus and set constraints are handled by exploiting methods based on exact smoothing penalty method and projection operators. Smoothing approximation technique is also utilized to deal with the nonsmooth objective function and the Nesterov's accelerated technique is applied to accelerate the proposed DSAPA. Furthermore, we have carefully proven the convergence rate of the DSAPA when taking different smoothing parameters (step sizes), and given the optimal convergence rate, that is, $O(\log/(k)k)$. Finally, the effectiveness of the proposed DSAPA is illustrated by simulations and applications to specific examples. Considering that the communication topology in practical engineering applications is directed graph as well as time-varying, we will study nonsmooth acceleration algorithms for solving nonsmooth constrained optimization problems over time-varying directed graphs in our future work. Moreover, differentially private distributed online learning over time-varying digraphs will also be worth considering, and the smoothing approximation of the 1-norm penalty problem in general case is to be further investigated in our future work.

Appendix A
Proof of Proposition 1

Proof: i). We will prove i) in two steps.

Since $t_i^{k+1} = \frac{a}{2} + \sqrt{(\frac{\mu_i^k}{\mu_i^{k+1}})t_i^k}$, one has

$$\frac{a}{2} + \sqrt{\left(\frac{\mu_i^k}{\mu_i^{k+1}}\right)t_i^k} = t_i^{k+1} \leqslant \sqrt{\frac{\mu_i^k}{\mu_i^{k+1}}} + \sqrt{\frac{\mu_i^k}{\mu_i^{k+1}}}t_i^k. \quad (39)$$

Step 1: From the right-hand inequality of (39), we have $\sqrt{\mu_i^{k+1}}t_i^{k+1} \leqslant \sqrt{\mu_i^k} + \sqrt{\mu_i^k}t_i^k$. Furthermore, one has

$$\sqrt{\mu_i^{k+1}}t_i^{k+1} \leqslant \sqrt{\mu_i^k} + \sqrt{\mu_i^k}t_i^k \leqslant \sum_{j=0}^{k} \sqrt{\mu_i^j} + \sqrt{\mu_i^0}t_i^0$$
$$= \sqrt{\mu_i^0} + \sqrt{\mu_i^0} \sum_{j=1}^{k+1} j^{-\frac{\theta}{2}} \leqslant \sqrt{\mu_i^0} \left(1 + \int_0^{k+1} t^{-\frac{\theta}{2}} dt\right)$$
$$= \frac{(4-\theta)\sqrt{\mu_i^0}}{2-\theta} (k+1)^{1-\frac{\theta}{2}}$$
$$\Rightarrow t_i^{k+1}\sqrt{\mu_i^{k+1}} \leqslant \frac{(4-\theta)\sqrt{\mu_i^0}}{2-\theta} (k+1)^{1-\frac{\theta}{2}}. \quad (40)$$

Step 2: The left-hand equality of (39) implies

$$t_i^{k+1}\sqrt{\mu_i^{k+1}} = \frac{a\sqrt{\mu_i^{k+1}}}{2} + \sqrt{\mu_i^k}t_i^k$$
$$= \frac{2-a}{2}\sqrt{\mu_i^0} + \frac{a}{2}\left(\frac{\sqrt{\mu_i^0}}{1^{\frac{\theta}{2}}} + \frac{\sqrt{\mu_i^0}}{2^{\frac{\theta}{2}}} + \cdots + \frac{\sqrt{\mu_i^0}}{(k+2)^{\frac{\theta}{2}}}\right)$$
$$\geqslant \frac{a}{2}\sqrt{\mu_i^0}\left(\left(\frac{2}{a} - 1\right) + \int_1^{k+3} t^{-\frac{\theta}{2}} dt\right)$$
$$\geqslant \frac{a\sqrt{\mu_i^0}}{2-\theta}\left((k+3)^{1-\frac{\theta}{2}} - 1\right) \geqslant \frac{a\sqrt{\mu_i^0}}{2(2-\theta)} (k+2)^{1-\frac{\theta}{2}}$$

$$\Rightarrow \sqrt{\mu_i^{k+1}} t_i^{k+2} \geqslant \frac{a\sqrt{\mu_i^0}}{2\,(2-\theta)}\,(k+2)^{1-\frac{\theta}{2}} \tag{41}$$

where the first inequality holds due to $a \in (0,1]$ and the second inequality is satisfied by using $\frac{4-2a-2\theta+a\theta-2}{a(2-\theta)} \geqslant \frac{\theta-2}{2-\theta} = -1$, $(k+3)^{1-\frac{\theta}{2}} \geqslant 2, k \geqslant 1$.

Thus, (40) and (41) imply the condition i) holds.

ii) Since $t_i^{k+1}(\mu_i^{k+1})^2 = t_i^{k+1}\sqrt{\mu_i^{k+1}}\sqrt{(\mu_i^{k+1})^3}$ and $(\frac{\mu_i^0}{(k+2)^\theta})^{\frac{3}{2}} \leqslant (\frac{\mu_i^0}{(k+1)^\theta})^{\frac{3}{2}}$, we have $t_i^{k+1}(\mu_i^{k+1})^2 \leqslant \frac{(4-\theta)(\mu_i^0)^2}{2-\theta}(k+1)^{1-2\theta}$.

iii) From (39) we have $\sqrt{(\frac{\mu_i^k}{\mu_i^{k+1}})}t_i^k \leqslant t_i^{k+1}$ and $(\frac{t_i^k}{t_i^{k+1}})^2 = (\alpha_i^k)^2 \leqslant \frac{\mu_i^{k+1}}{\mu_i^k}$.

iv) The condition $t_i^{k+1} = \frac{a}{2} + \sqrt{(\frac{\mu_i^k}{\mu_i^{k+1}})}t_i^k$, implies $\mu_i^{k+1}t_i^{k+1}(t_i^{k+1}-1) = (t_i^k)^2\mu_i^k - \mu_i^{k+1}t_i^{k+1}(1-a) - \frac{\mu_i^{k+1}a^2}{4}$, thus, we have $\mu_i^{k+1}(t_i^{k+1})(t_i^{k+1}-1) \leqslant \mu_i^k(t_i^k)^2$.

v) According to $\frac{a}{2} + \sqrt{(\frac{\mu_i^k}{\mu_i^{k+1}})}t_i^k = t_i^{k+1}$ and $\mu_i^k \geqslant \mu_i^{k+1}$, one has $t_i^{k+1} \geqslant \frac{a}{2} + t_i^k = \frac{ak+1}{2} \geqslant \frac{a(k+1)}{2}$ and $t_i^k - 1 = \sqrt{(\frac{\mu_i^{k+1}}{\mu_i^k})}(t_i^{k+1} - \frac{a}{2}) - 1 \leqslant \sqrt{(\frac{\mu_i^{k+1}}{\mu_i^k})}t_i^{k+1} - 1$.

Further, $\frac{(t_i^k-1)^2}{(t_i^{k+1})^2} \leqslant (\frac{\mu_i^{k+1}}{\mu_i^k}) + \frac{1}{(t_i^{k+1})^2} - 2\sqrt{(\frac{\mu_i^{k+1}}{\mu_i^k})}\frac{1}{(t_i^{k+1})}$. In order to obtain the condition **(v)**, we only need to prove $\frac{1}{(t_i^{k+1})^2} \leqslant \sqrt{(\frac{\mu_i^{k+1}}{\mu_i^k})}\frac{1}{(t_i^{k+1})}$, i.e., $(k+1)^2(\frac{k+1}{k+2})^\theta \geqslant \frac{4}{a^2}$. Since $(\frac{k+1}{k+2})^\theta \geqslant (\frac{1}{2})^\theta$, we just have to prove $(k+1)^2 \geqslant \frac{2^{2+\theta}}{a^2}$, i.e., $k \geqslant \frac{2^{1+\frac{\theta}{2}}}{a^2}$. ■

Appendix B
Proof of Proposition 2

Proof: Define an auxiliary function of variables x, y, and matrix $\boldsymbol{\mu}$ as follows:

$$\begin{aligned} H(x,y,\boldsymbol{\mu}) \quad &= \hat{\Gamma}(y,\boldsymbol{\mu}) + \nabla\hat{\Gamma}(y,\boldsymbol{\mu})^T(x-y) \\ &\quad + \frac{1}{2}(x-y)^T\frac{\boldsymbol{L}}{\boldsymbol{\mu}}(x-y) + \mathbf{1}^T\frac{\boldsymbol{L}}{\boldsymbol{\mu}}\mathrm{I}_\Omega(x) \end{aligned} \tag{42}$$

then, the optimal condition of x in (42) satisfies $\mathbf{0} \in \frac{\boldsymbol{\mu}}{\boldsymbol{L}}\nabla\hat{\Gamma}(y,\boldsymbol{\mu}) + N_\Omega(x) + x - y$. From $(I+N_\Omega)^{-1} = P_\Omega$, one has $x = P_\Omega(y - \frac{\boldsymbol{\mu}}{\boldsymbol{L}}\nabla\hat{\Gamma}(y,\boldsymbol{\mu}))$, i.e., $q(y,\boldsymbol{\mu}) = \min_x H(x,y,\boldsymbol{\mu})$.

Note that

$$\begin{aligned} &H(x,y,\boldsymbol{\mu}) - H(q(y,\boldsymbol{\mu}),y,\mu) \\ &\geqslant (x-q(y,\boldsymbol{\mu}))^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}}(x-q(y,\boldsymbol{\mu})) \end{aligned} \tag{43}$$

holds from the definitions of $q(y,\boldsymbol{\mu})$ and $H(x,y,\boldsymbol{\mu})$ in (42).

In addition

$$\begin{aligned} &\hat{\Gamma}(q(y,\boldsymbol{\mu}),\boldsymbol{\mu}) + \mathbf{1}^T\frac{\boldsymbol{L}}{\boldsymbol{\mu}}\mathrm{I}_\Omega(q(y,\boldsymbol{\mu})) \\ &\leqslant H(x,y,\boldsymbol{\mu}) - (x-q(y,\boldsymbol{\mu}))^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}}(x-q(y,\boldsymbol{\mu})) \\ &\leqslant \hat{\Gamma}(x,\boldsymbol{\mu}) + (x-y)^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}}(x-y) + \mathbf{1}^T\frac{\boldsymbol{L}}{\boldsymbol{\mu}}\mathrm{I}_\Omega(x) \\ &\quad - (x-q(y,\boldsymbol{\mu}))^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}}(x-q(y,\boldsymbol{\mu})) \\ &= \hat{\Gamma}(x,\boldsymbol{\mu}) - (q(y,\boldsymbol{\mu})-y)^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}}(q(y,\boldsymbol{\mu})-y) \\ &\quad - (x-y)^T\frac{\boldsymbol{L}}{\boldsymbol{\mu}}(q(y,\boldsymbol{\mu})-y) + \mathbf{1}^T\frac{\boldsymbol{L}}{\boldsymbol{\mu}}\mathrm{I}_\Omega(x) \end{aligned} \tag{44}$$

where the second inequality holds according to the convexity of $\hat{\Gamma}(x,\boldsymbol{\mu})$ with respect to x, the condition (14) is satisfied by using $q(y,\boldsymbol{\mu}) \in \Omega$. Therefore, the proof is completed. ■

Appendix C
Proof of Proposition 3

Proof: Letting $x = x^k$, $y = y^k$ and $\boldsymbol{\mu} = \boldsymbol{\mu}^k$ in (14), we have

$$\begin{aligned} &\hat{\Gamma}(x^k,\boldsymbol{\mu}^k) - \hat{\Gamma}(q(y^k,\boldsymbol{\mu}^k),\boldsymbol{\mu}^k) + \mathbf{1}^T\frac{\boldsymbol{\mu}}{\boldsymbol{L}}\mathrm{I}_\Omega(x^k) \\ &\geqslant -(x^k-y^k)^T\frac{L}{2\boldsymbol{\mu}^k}(x^k-y^k) \\ &\quad + (x^k-q(y^k,\boldsymbol{\mu}^k))^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^k}(x^k-q(y^k,\boldsymbol{\mu}^k)). \end{aligned} \tag{45}$$

Combining $x^{k+1} = q(y^k,\boldsymbol{\mu}^k)$, $y^k = x^k + \boldsymbol{\alpha}^k(x^k-x^{k-1})$ with $y^k = x^k + \boldsymbol{\alpha}^k(x^k-x^{k-1})$, (45) yields that

$$\begin{aligned} &\hat{\Gamma}(x^k,\boldsymbol{\mu}^k) - \hat{\Gamma}(x^{k+1},\boldsymbol{\mu}^k) + I_\Omega(x^k) \\ &\geqslant (x^k-x^{k+1})^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^k}(x^k-x^{k+1}) \\ &\quad - (x^k-x^{k-1})^T\frac{\boldsymbol{L}(\boldsymbol{\alpha}^k)^2}{2\boldsymbol{\mu}^k}(x^k-x^{k-1}). \end{aligned} \tag{46}$$

Considering that $\boldsymbol{\mu}^{k+1} \preccurlyeq \boldsymbol{\mu}^k$, which implies $\kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^{k+1}\mathbf{1} \leqslant \kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^k\mathbf{1}$. Moreover, since $x^0 \in \Omega$, $x^{k+1} = P_\Omega(y^k - \frac{\boldsymbol{\mu}^k}{\boldsymbol{L}}\nabla\hat{\Gamma}(y^k,\boldsymbol{\mu}^k)) \in \Omega$, i.e., $\mathrm{I}_\Omega(x^k) = 0$. Thus, we obtain

$$\begin{aligned} &\hat{\Gamma}(x^k,\boldsymbol{\mu}^k) + \kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^k\mathbf{1} \\ &\quad + (x^k-x^{k-1})^T\frac{\boldsymbol{L}(\boldsymbol{\alpha}^k)^2}{2\boldsymbol{\mu}^k}(x^k-x^{k-1}) \\ &\geqslant \hat{\Gamma}(x^{k+1},\boldsymbol{\mu}^{k+1}) + \kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^{k+1}\mathbf{1} \\ &\quad + (x^k-x^{k+1})^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^k}(x^k-x^{k+1}). \end{aligned} \tag{47}$$

From iii) in Proposition 1, i.e., $(\boldsymbol{\alpha}^k)^2 \preccurlyeq \frac{\boldsymbol{\mu}^k}{\boldsymbol{\mu}^{k-1}}$, we have

$$\begin{aligned} W(k) &= \hat{\Gamma}(x^k,\boldsymbol{\mu}^k) + \kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^k\mathbf{1} \\ &\quad + (x^k-x^{k-1})^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^{k-1}}(x^k-x^{k-1}) \\ &\geqslant \hat{\Gamma}(x^{k+1},\boldsymbol{\mu}^{k+1}) + \kappa_{\hat{\Gamma}}\mathbf{1}^T\boldsymbol{\mu}^{k+1}\mathbf{1} \\ &\quad + (x^k-x^{k+1})^T\frac{\boldsymbol{L}}{2\boldsymbol{\mu}^k}(x^k-x^{k+1}) = W(k+1). \end{aligned} \tag{48}$$

Since $W(k)$ is bounded from below, i.e., $W(k) \geqslant \hat{\Gamma}(x^k, \boldsymbol{\mu}^k) + \kappa_\Gamma \mathbf{1}^T \boldsymbol{\mu}^k \mathbf{1} \geqslant \hat{\Gamma}(x^k, \boldsymbol{\mu}^k) \geqslant \hat{\Gamma}(x^*, \boldsymbol{\mu}^*)$, thus

$$\lim_{k \to +\infty} W(k) \text{ exists.} \tag{49}$$

Moreover, according to (48), we obtain

$$\{x^k\} \in \left\{ x | \hat{\Gamma}(x, \boldsymbol{\mu}) \leqslant \hat{\Gamma}\left(x^1, \boldsymbol{\mu}^1\right) + \kappa_{\hat{\Gamma}} \mathbf{1}^T \boldsymbol{\mu}^1 \mathbf{1} + \left(x^1 - x^0\right)^T \frac{\boldsymbol{L}}{2\boldsymbol{\mu}^0} \left(x^1 - x^0\right) \forall x \in \Omega \right\}.$$

By using inequality $\hat{\Gamma}(x^1, \boldsymbol{\mu}^1) + \kappa_{\hat{\Gamma}} \mathbf{1}^T \boldsymbol{\mu}^1 \mathbf{1} + (x^1 - x^0)^T \frac{\boldsymbol{L}}{2\boldsymbol{\mu}^0} (x^1 - x^0) \leqslant \hat{\Gamma}(x^0, \boldsymbol{\mu}^0) + \kappa_{\hat{\Gamma}} \mathbf{1}^T \boldsymbol{\mu}^0 \mathbf{1}, x^0 = y^0 \in \Omega$, we have

$$\{x^k\} \in \left\{ x \in R^{mn} | \hat{\Gamma}(x, \boldsymbol{\mu}) \leqslant \hat{\Gamma}\left(x^0, \boldsymbol{\mu}^0\right) + \kappa_{\hat{\Gamma}} \mathbf{1}^T \boldsymbol{\mu}^0 \mathbf{1} \forall x^0 = y^0 \in \Omega \right\} \tag{50}$$

i.e, the sequence $\{x^k\}$ is bounded. ■

REFERENCES

[1] M. Rabba and R. Nowak, "Distributed optimization in sensor networks," in *Proc. Int. Symp. Inf. Process. Sens. Netw. (Part CPS Week)*, 2004, pp. 20–27.

[2] H. Wang and C. Li, "Distributed quantile regression over sensor networks," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 4, no. 2, pp. 338–348, Jun. 2018.

[3] M. Li et al., "Scaling distributed machine learning with the parameter server," in *Proc. USENIX Symp. Oper. Syst. Des. Implement.*, 2014, pp. 583–598.

[4] J. F. C. Mota, J. M. F. Xavier, P. M. Q. Aguiar, and M. Puschel, "Distributed basis pursuit," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1942–1956, Apr. 2012.

[5] Y. Zhao, X. Liao, X. He, and R. Tang, "Centralized and collective neurodynamic optimization approaches for sparse signal reconstruction via L_1-minimization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7488–7501, Dec. 2022.

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

[7] M. Zhu and S. Martínez, "On distributed convex optimization under inequality and equality constraints," *IEEE Trans. Autom. Control*, vol. 57, no. 1, pp. 151–164, Jan. 2012.

[8] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Convergence of asynchronous distributed gradient methods over stochastic networks," *IEEE Trans. Autom. Control.*, vol. 63, no. 2, pp. 434–448, Feb. 2018.

[9] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Trans. Autom. Control.*, vol. 60, no. 3, pp. 601–615, Mar. 2015.

[10] I. Notarnicola and G. Notarstefano, "Asynchronous distributed optimization via randomized dual proximal gradient," *IEEE Trans. Autom. Control.*, vol. 62, no. 5, pp. 2095–2106, May 2017.

[11] 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.

[12] J. F. C. Mota, J. M. F. Xavier, P. M. Q. Aguiar, and M. Püschel, "D-ADMM: A communication-efficient distributed algorithm for separable optimization," *IEEE Trans. Signal Proces.*, vol. 61, no. 10, pp. 2718–2723, May 2013.

[13] G. Zhang and R. Heusdens, "Distributed optimization using the primal-dual method of multipliers," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 4, no. 1, pp. 173–187, Mar. 2018.

[14] A. Nedić, A. Ozdaglar, and P. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Trans. Autom. Control.*, vol. 55, no. 4, pp. 922–938, Apr. 2010.

[15] C. Xi and U. A. Khan, "Distributed subgradient projection algorithm over directed graphs," *IEEE Trans. Autom. Control.*, vol. 62, no. 8, pp. 3986–3992, Aug. 2017.

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

[17] Y. Lin, I. Shames, and D. Nesic, "Asynchronous distributed optimization via dual decomposition and block coordinate subgradient methods," *IEEE Trans. Control Netw. Syst*, vol. 8, no. 3, pp. 1348–1359, Sep. 2021.

[18] C.-X. Shi and G. -H. Yang, "Distributed optimization under unbalanced digraphs with node errors: Robustness of surplus-based dual averaging algorithm," *IEEE Trans. Control Netw. Syst*,, vol. 8, no. 1, pp. 331–341, Mar. 2021.

[19] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant step-sizes," in *Proc IEEE 54th Conf Decis. Control*, 2015, pp. 2055–2060.

[20] A. Nedić, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM J. Optim.*, vol. 27, no. 4, pp. 2597–2633, 2017.

[21] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Trans. Control. Netw Syst.*, vol. 5, no. 3, pp. 1245–1260, Sep. 2018.

[22] A. Nedić and A. Olshevsky, "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," *IEEE Trans. Autom. Control.*, vol. 61, no. 12, pp. 3936–3947, Dec. 2016.

[23] P. Wang, P. Lin, W. Ren, and Y. Song, "Distributed subgradient-based multiagent optimization with more general step sizes," *IEEE Trans. Autom. Control.*, vol. 63, no. 7, pp. 2295–2302, Jul. 2018.

[24] H. Li, G. Q. Lü, and T. Huang, "Distributed projection subgradient algorithm over time-varying general unbalanced directed graphs," *IEEE Trans. Autom. Control.*, vol. 64, no. 3, pp. 1309–1316, Mar. 2018.

[25] S. Liu, Z. Qiu, and L. Xie, "Convergence rate analysis of distributed optimization with projected subgradient algorithm," *Automatica*, vol. 83, pp. 162–169, 2017.

[26] Y. Yuan, G. B. Stan, L. Shi, M. Barahona, and J. Goncalves, "Decentralised minimum-time consensus," *Automatica*, vol. 49, no. 5, pp. 1227–1235, 2013.

[27] X. Wu and J. Lu, "Fenchel dual gradient methods for distributed convex optimization over time-varying networks," *IEEE Trans. Autom. Control.*, vol. 64, no. 11, pp. 4629–4636, Nov. 2019.

[28] C. A. Uribe, S. Lee, A. Gasnikov, and A. Nedić, "A dual approach for optimal algorithms in distributed optimization over networks," in *Proc. Inf. Theory Appl. Workshop*, 2020, pp. 1–37.

[29] K. Margellos, A. Falsone, S. Garatti, and M. Prandini, "Distributed constrained optimization and consensus in uncertain networks via proximal minimization," *IEEE Trans. Autom. Control.*, vol. 63, no. 5, pp. 1372–1387, May 2018.

[30] X. Li, G. Feng, and L. Xie, "Distributed proximal algorithms for multiagent optimization with coupled inequality constraints," *IEEE Trans. Autom. Control.*, vol. 66, no. 3, pp. 1223–1230, Mar. 2021.

[31] A. Sidford and K. Tian, "Coordinate methods for accelerating l_∞ regression and faster approximate maximum flow," in *Proc. IEEE 59th Annu. Symp. Found. Comput. Sci*, 2018, pp. 922–933.

[32] Y. Nesterov, "Smooth minimization of non-smooth functions," *Math. Prog.*, vol. 103, no. 1, pp. 127–152, 2005.

[33] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[34] D. Kim and J. A. Fessler, "Another look at the fast iterative shrinkage/thresholding algorithm (FISTA)," *SIAM J. Optim.*, vol. 28, no. 1. pp. 223–250, 2018.

[35] X. Chen, "Smoothing methods for nonsmooth, nonconvex minimization," *Math. Prog.*, vol. 134, no. 1, pp. 71–99, 2012.

[36] W. Bian and X. Chen, "Neural network for nonsmooth, nonconvex constrained minimization via smooth approximation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 3, pp. 545–556, Mar. 2014.

[37] W. Bian, "Smoothing accelerated algorithm for constrained nonsmooth convex optimization problems," *Sci. China Math.*, vol. 50, no. 12, pp. 1651–1666, 2020.

You Zhao received the M.S. degree in signal and information processing from the School of Electronics and Information Engineering, Southwest University, Chongqing, China, in 2018. He is currently working toward the Ph.D. degree computer science and technology with the School of Computer Science, Chongqing University, Chongqing, China.

His research interests include neurodynamic optimization, distributed optimization, and compressed sensing.

Xiaofeng Liao (Fellow, IEEE) received the B.S. and M.S. degrees in mathematics from Sichuan University, Chengdu, China, in 1986 and 1992, respectively, and the Ph.D. degree in circuits and systems from the University of Electronic Science and Technology of China, Chengdu, in 1997.

He was a Research Associate with the Chinese University of Hong Kong, Hong Kong, from November 1997 to April 1998, and from October 1999 to October 2000. He was a Professor and the Dean of the College of Electronic and Information Engineering, Southwest University, Chongqing, China, from July 2012 to July 2018. He is currently a Professor and the Dean of the College of Computer Science, Chongqing University, Chongqing, China. He is also a Yangtze River Scholar of the Ministry of Education of China, Beijing, China. His research interests include optimization and control, machine learning, neural networks, bifurcation and chaos, and cryptography.

Prof. Liao currently serves as an Associate Editor for IEEE TRANSACTIONS ON CYBERNETICS and IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Xing He received the Ph.D. degree in computer science and technology from Chongqing University, Chongqing, China, in 2013.

From 2012 to 2013, he was a Research Assistant with the Texas A&M University at Qatar, Doha, Qatar. He is currently a Professor with the School of Electronics and Information Engineering, Southwest University, Chongqing. His research interests include neural networks, optimization methods, smart grids, and nonlinear dynamical systems.