Distributed Constrained Optimization by Consensus-Based Primal-Dual Perturbation Method

Tsung-Hui Chang, *Member, IEEE*, Angelia Nedić, *Member, IEEE*, and Anna Scaglione, *Fellow, IEEE*

***Abstract*—Various distributed optimization methods have been developed for solving problems which have simple local constraint sets and whose objective function is the sum of local cost functions of distributed agents in a network. Motivated by emerging applications in smart grid and distributed sparse regression, this paper studies distributed optimization methods for solving general problems which have a coupled global cost function and have inequality constraints. We consider a network scenario where each agent has no global knowledge and can access only its local mapping and constraint functions. To solve this problem in a distributed manner, we propose a consensus-based distributed primal-dual perturbation (PDP) algorithm. In the algorithm, agents employ the average consensus technique to estimate the global cost and constraint functions via exchanging messages with neighbors, and meanwhile use a local primal-dual perturbed subgradient method to approach a global optimum. The proposed PDP method not only can handle smooth inequality constraints but also non-smooth constraints such as some sparsity promoting constraints arising in sparse optimization. We prove that the proposed PDP algorithm converges to an optimal primal-dual solution of the original problem, under standard problem and network assumptions. Numerical results illustrating the performance of the proposed algorithm for a distributed demand response control problem in smart grid are also presented.**

***Index Terms*—Average consensus, constrained optimization, demand side management control, distributed optimization, primal-dual subgradient method, regression, smart grid.**

I. Introduction

DISTRIBUTED optimization methods are becoming popular options for solving several engineering problems, including parameter estimation, detection and localization problems in sensor networks [1], [2], resource allocation problems in peer-to-peer/multi-cellular communication networks [3], [4], and distributed learning and regression problems in control [5] and machine learning [6]–[8], to name a few. In these applications, rather than pooling together all the relevant parameters that define the optimization problem, distributed agents, which have access to a local subset of such parameters, collaborate with each other to minimize a global cost function, subject to local variable constraints. Specifically, since it is not always efficient for the agents to exchange across the network the local cost and constraint functions, owing to the large size of network, time-varying network topology, energy constraints and/or privacy issues, distributed optimization methods that utilize only local information and messages exchanged between connecting neighbors have been of great interest; see [9]–[16] and references therein.

Contributions: Different from the existing works [9]–[14] where the local variable constraints are usually simple (in the sense that they can be handled via simple projection) and independent among agents, in this paper, we consider a problem formulation that has a general set of convex inequality constraints that couple all the agents' optimization variables. In addition, similar to [17], the considered problem has a global (non-separable) convex cost function that is a function of the sum of local mapping functions of the local optimization variabless. Such a problem formulation appears, for example, in the classical regression problems which have a wide range of applications. In addition, the considered formulation also arises in the demand response control and power flow control problems in the emerging smart grid systems [18]–[20]. More discussions about applications are presented in Section II-B.

In this paper, we assume that each agent knows only the local mapping function and local constraint function. To solve this problem in a distributed fashion, in this paper, we develop a novel *distributed consensus-based primal-dual perturbation (PDP)* algorithm, which combines the ideas of the primal-dual perturbed (sub-)gradient method [21], [22] and the average consensus techniques [10], [23], [24]. In each iteration of the proposed algorithm, agents exchange their local estimates of the global cost and constraint functions with their neighbors, followed by performing one-step of primal-dual variable (sub-)gradient update. Instead of using the primal-dual iterates computed at the preceding iteration as in most of the existing primal-dual subgradient based methods [15], [16], the (sub-)gradients in the proposed distributed PDP algorithm are computed based on some perturbation points which can be efficiently computed using the messages exchanged from neighbors. In particular, we provide two efficient ways to compute the perturbation points that can respectively handle the smooth and non-smooth constraint functions. More importantly, we build convergence analysis results showing that the proposed distributed PDP algorithm ensures a strong convergence of

Manuscript received September 11, 2012; revised April 19, 2013 and November 4, 2013; accepted January 27, 2014. Date of publication February 26, 2014; date of current version May 20, 2014. This work was supported by the National Science Council of Taiwan (R.O.C.) by Grant NSC 102-2221-E-011-005-MY3, by NSF Grants CMMI 07-42538 and CCF 11-11342, and by NSF CCF-1011811. Recommended by Associate Editor F. Paganini.

T.-H. Chang is with the Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei 10607, Taiwan (e-mail: tsunghui.chang@ieee.org).

A. Nedić is with the Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: angelia@illinois.edu).

A. Scaglione is with the Department of Electrical and Computer Engineering, University of California, Davis, CA 95616 USA (e-mail: ascaglione@ucdavis.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2014.2308612

the local primal-dual iterates to a global optimal primal-dual solution of the considered problem. The proposed algorithm is applied to a distributed sparse regression problem and a distributed demand response control problem in smart grid. Numerical results for the two applications are presented to demonstrate the effectiveness of the proposed algorithm.

Related Works: Distributed dual subgradient method (e.g., dual decomposition) [25] is a popular approach to solving a problem with coupled inequality constraints in a distributed manner. However, given the dual variables, this method requires the agents to globally solve the local subproblems, which may require considerable computational efforts if the local cost and constraint functions have some complex structure. Consensus-based distributed primal-dual (PD) subgradient methods have been developed recently in [15], [16] for solving a problem with an objective function which is the sum of local convex cost functions, and with global convex inequality constraints. In addition to having a different cost function from our problem formulation, the works in [15], [16] assumed that all the agents in the network have global knowledge of the inequality constraint function; the two are in sharp contrast to our formulation where a non-separable objective function is considered and each agent can access only its local constraint function. Moreover, these works adopted the conventional PD subgradient updates [26], [27] without perturbation. Numerical results will show that these methods do not perform as well as the proposed algorithm with perturbation. Another recent development is the Bregman-distance based PD subgradient method proposed in [28] for solving an epigraph formulation of a min-max problem. The method in [28], however, assumes that the Lagrangian function has a unique saddle point, in order to guarantee the convergence of the primal-dual iterates. In contrast, our proposed algorithm, which uses the perturbed subgradients, does not require such assumption.

Synopsis: Section II presents the problem formulation, applications, and a brief review of the centralized PD subgradient methods. Section III presents the proposed distributed consensus-based PDP algorithm. The assumptions and convergence analysis results are given in Section IV. Numerical results are presented in Section V. Finally, the conclusions and discussion of future extensions are drawn in Section VI.

II. Problem Formulation, Applications, and Brief Review

A. Problem Formulation

We consider a network with N agents, denoted by $\mathcal{V} = \{1, \ldots, N\}$. We assume that, for all $i = 1, \ldots, N$, each agent i ahas a local decision variable[1] $\boldsymbol{x}_i \in \mathbb{R}^K$, a local constraint set $\mathcal{X}_i \subseteq \mathbb{R}^K$, and a local mapping function $\boldsymbol{f}_i : \mathbb{R}^K \to \mathbb{R}^M$, in which $\boldsymbol{f}_i = (f_{i1}, \ldots, f_{iM})^T$ with each $f_{im} : \mathbb{R}^K \to \mathbb{R}$ being continuous. The network cost function is given by

$$\bar{\mathcal{F}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \triangleq \mathcal{F}\left(\sum_{i=1}^{N} \boldsymbol{f}_i(\boldsymbol{x}_i)\right) \tag{1}$$

[1]Here, without loss of generality, we assume that all the agents have the same variable dimension K. The proposed algorithm and analysis can be easily generalized to the case with different variable dimensions.

where $\mathcal{F} : \mathbb{R}^M \to \mathbb{R}$ and $\bar{\mathcal{F}} : \mathbb{R}^{NK} \to \mathbb{R}$ are continuous. In addition, the agents are subject to a global inequality constraint $\sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{x}_i) \preceq \boldsymbol{0}$, where $\boldsymbol{g}_i : \mathbb{R}^K \to \mathbb{R}^P$ are continuous mappings for all $i = 1, \ldots, N$; specifically, $\boldsymbol{g}_i = (g_{i1}, \ldots, g_{iP})^T$, with each $g_{ip} : \mathbb{R}^K \to \mathbb{R}$ being continuous. The vector inequality $\sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{x}_i) \preceq \boldsymbol{0}$ is understood coordinate-wise.

We assume that each agent i can access $\mathcal{F}(\cdot)$, $\boldsymbol{f}_i(\cdot)$, $\boldsymbol{g}_i(\cdot)$ and $\mathcal{X}_i$ only, for all $i = 1, \ldots, N$. Under this local knowledge constraint, the agents seek to cooperate with each other to minimize the total network cost $\bar{\mathcal{F}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ (or maximize the network utility $-\bar{\mathcal{F}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$). Mathematically, the optimization problem can be formulated as follows:

$$\min_{\substack{\boldsymbol{x}_i \in \mathcal{X}_i, \\ i=1,\ldots,N}} \bar{\mathcal{F}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \quad \text{s.t.} \quad \sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{x}_i) \preceq \boldsymbol{0}. \tag{2}$$

The goal of this paper is to develop a *distributed* algorithm for solving (2) with each agent communicating with their neighbors only.

B. Application to Smart Grid Control

In this subsection, we discuss some smart grid control problems where the problem formulation (2) may arise. Consider a power grid system where a retailer (e.g., the utility company) bids electricity from the power market and serves a residential/industrial neighborhood with N customers. In addition to paying for its market bid, the retailer has to pay additional cost if there is a deviation between the bid purchased in earlier market settlements and the real-time aggregate load of the customers. Any demand excess or shortfall results in a cost for the retailer that mirrors the effort to maintain the power balance. In the smart grid, thanks to the advances in communication and sensory technologies, it is envisioned that the retailer can observe the load of customers and can even control the power usage of some of the appliances (e.g., controlling the charging rate of electrical vehicles and turning ON/OFF air conditioning systems), which is known as the demand side management (DSM); see [29] for a recent review.

We let p_t, $t = 1, \ldots, T$, be the power bids over a time horizon of length T, and let $\psi_{i,t}(\boldsymbol{x}_i)$, $t = 1, \ldots, T$, be the load profile of customer i, where $\boldsymbol{x}_i \in \mathbb{R}^K$ contains some control variables. The structures of $\psi_{i,t}$ and $\boldsymbol{x}_i$ depend on the appliance load model. As mentioned, the retailer aims to minimize the cost caused by power imbalance, e.g., [18], [19], [29]

$$\min_{\boldsymbol{x}_1 \in \mathcal{X}_1, \ldots, \boldsymbol{x}_N \in \mathcal{X}_N} C_{\mathrm{p}}\left[\left(\sum_{i=1}^{N} \boldsymbol{\psi}_i(\boldsymbol{x}_i) - \boldsymbol{p}\right)^{+}\right] + C_{\mathrm{s}}\left[\left(\boldsymbol{p} - \sum_{i=1}^{N} \boldsymbol{\psi}_i(\boldsymbol{x}_i)\right)^{+}\right] \tag{3}$$

where $(x)^+ = \max\{x, 0\}$, $\mathcal{X}_i$ denotes the local control constraint set and $C_{\mathrm{p}}, C_{\mathrm{s}} : \mathbb{R}^T \to \mathbb{R}$ denote the cost functions due to insufficient and excessive power bids, respectively. Moreover, let $\boldsymbol{p} = (p_1, \ldots, p_T)^T$ and $\boldsymbol{\psi}_i = (\psi_{i,1}, \ldots, \psi_{i,T})^T$. By

defining $\boldsymbol{z}=(\sum_{i=1}^N \boldsymbol{\psi}_i(\boldsymbol{x}_i)-\boldsymbol{p})^+$ and assuming that $C_{\rm p}$ is monotonically increasing, one can write (3) as

$$\min_{\boldsymbol{x}_1\in\mathcal{X}_1,\ldots,\boldsymbol{x}_N\in\mathcal{X}_N} C_{\rm p}[\boldsymbol{z}]+C_{\rm s}\left[\boldsymbol{z}-\sum_{i=1}^N\boldsymbol{\psi}_i(\boldsymbol{x}_i)+\boldsymbol{p}\right]$$
$$\text{s.t.}\quad \sum_{i=1}^N\boldsymbol{\psi}_i(\boldsymbol{x}_i)-\boldsymbol{p}-\boldsymbol{z}\preceq\boldsymbol{0} \tag{4}$$

which belongs to the considered formulation in (2). Similar problem formulations also arise in the microgrid control problems [20], [30] where the microgrid controller requires not only to control the loads but also to control the local power generation and local power storage (i.e., power flow control), in order to maintain power balance within the microgrid; see [30] for detailed formulations. Distributed control methods are appealing to the smart grid application since all the agents are identical and failure of one agent would not have significant impact on the performance of the whole system [31]. Besides, it also spares the retailer/microgrid controller from the task of collecting real-time information of customers, which not only infringes on the customer's privacy but also is not easy for a large scale neighborhood. In Section V, the proposed distributed algorithm will be applied to a DSM problem as in (4).

In addition to the smart grid applications, problem (2) incorporates the important regression problems which widely appear in control [5], machine learning [6], [7], data mining [32], [33] and imaging processing [7] applications. Formulation (2) also encompasses the network flow control problems [34]; see [35] for an example which considered maximizing the network lifetime. The proposed distributed algorithm therefore can be applied to these problem as well. For example, in [36], we have shown how the proposed distributed algorithm can be applied to handle a distributed sparse regression problem.

C. Centralized PD Subgradient Method

Let us consider the following Lagrange dual problem of (2):

$$\max_{\boldsymbol{\lambda}\succeq\boldsymbol{0}}\left\{\min_{\boldsymbol{x}\in\mathcal{X}}\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda})\right\} \tag{5}$$

where $\boldsymbol{x}=(\boldsymbol{x}_1^T,\ldots,\boldsymbol{x}_N^T)^T$, $\mathcal{X}=\mathcal{X}_1\times\cdots\times\mathcal{X}_N$, $\boldsymbol{\lambda}\in\mathbb{R}_+^P$ (i.e., the non-negative orthant in $\mathbb{R}^P$) is the dual variable associated with the inequality constraint $\sum_{i=1}^N\boldsymbol{g}_i(\boldsymbol{x}_i)\preceq\boldsymbol{0}$, and $\mathcal{L}:\mathbb{R}^{NK}\times\mathbb{R}_+^P\to\mathbb{R}$ is the Lagrangian function, given by

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda})=\bar{\mathcal{F}}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)+\boldsymbol{\lambda}^T\left(\sum_{i=1}^N\boldsymbol{g}_i(\boldsymbol{x}_i)\right). \tag{6}$$

Throughout the paper, we assume that problem (2) is convex, i.e., $\mathcal{X}$ is closed and convex, $\bar{\mathcal{F}}(\boldsymbol{x})$ is convex in $\boldsymbol{x}$ and each $\boldsymbol{g}_i(\boldsymbol{x}_i)$ is convex in $\boldsymbol{x}_i$. We also assume that the Slater condition holds, i.e., there is an $(\bar{\boldsymbol{x}}_1,\ldots,\bar{\boldsymbol{x}}_N)$ that lies in the relative interior of $\mathcal{X}_1\times\cdots\times\mathcal{X}_N$ such that $\sum_{i=1}^N\boldsymbol{g}_i(\bar{\boldsymbol{x}}_i)\prec\boldsymbol{0}$. Hence, the strong duality holds for problem (2) [37], problem (2) can be handled by solving its dual (5). A classical approach is the dual subgradient method [38]. One limitation of such method is that the inner problem $\min_{\boldsymbol{x}\in\mathcal{X}}\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda})$ needs to be globally solved at each iteration, which, however, is not always easy, especially when $\boldsymbol{f}_i(\boldsymbol{x}_i)$ and $\boldsymbol{g}_i(\boldsymbol{x}_i)$ are complex or when the problem is large scale. Another approach is the primal-dual (PD) subgradient method [26], [39] which handles the inner problem inexactly. More precisely, at iteration k, the PD subgradient method performs

$$\boldsymbol{x}^{(k)}=\mathcal{P}_{\mathcal{X}}\left(\boldsymbol{x}^{(k-1)}-a_k\mathcal{L}_{\boldsymbol{x}}\left(\boldsymbol{x}^{(k-1)},\boldsymbol{\lambda}^{(k-1)}\right)\right) \tag{7a}$$
$$\boldsymbol{\lambda}^{(k)}=\left(\boldsymbol{\lambda}^{(k-1)}+a_k\mathcal{L}_{\boldsymbol{\lambda}}\left(\boldsymbol{x}^{(k-1)},\boldsymbol{\lambda}^{(k-1)}\right)\right)^+ \tag{7b}$$

where $\mathcal{P}_{\mathcal{X}}:\mathbb{R}^{NK}\to\mathcal{X}$ is a projection function, $a_k>0$ is a step size, and

$$\begin{aligned}&\mathcal{L}_{\boldsymbol{x}}\left(\boldsymbol{x}^{(k)},\boldsymbol{\lambda}^{(k)}\right)\\&\triangleq\begin{bmatrix}\mathcal{L}_{\boldsymbol{x}_1}\left(\boldsymbol{x}^{(k)},\boldsymbol{\lambda}^{(k)}\right)\\\vdots\\\mathcal{L}_{\boldsymbol{x}_N}\left(\boldsymbol{x}^{(k)},\boldsymbol{\lambda}^{(k)}\right)\end{bmatrix}\\&=\begin{bmatrix}\nabla\boldsymbol{f}_1^T\left(\boldsymbol{x}_1^{(k)}\right)\nabla\mathcal{F}\left(\sum_{i=1}^N\boldsymbol{f}_i\left(\boldsymbol{x}_i^{(k)}\right)\right)+\nabla\boldsymbol{g}_1^T\left(\boldsymbol{x}_1^{(k)}\right)\boldsymbol{\lambda}^{(k)}\\\vdots\\\nabla\boldsymbol{f}_N^T\left(\boldsymbol{x}_N^{(k)}\right)\nabla\mathcal{F}\left(\sum_{i=1}^N\boldsymbol{f}_i\left(\boldsymbol{x}_i^{(k)}\right)\right)+\nabla\boldsymbol{g}_N^T\left(\boldsymbol{x}_N^{(k)}\right)\boldsymbol{\lambda}^{(k)}\end{bmatrix},\end{aligned} \tag{8a}$$

$$\mathcal{L}_{\boldsymbol{\lambda}}\left(\boldsymbol{x}^{(k)},\boldsymbol{\lambda}^{(k)}\right)\triangleq\sum_{i=1}^N\boldsymbol{g}_i\left(\boldsymbol{x}_i^{(k)}\right) \tag{8b}$$

represent the subgradients of $\mathcal{L}$ at $(\boldsymbol{x}^{(k)},\boldsymbol{\lambda}^{(k)})$ with respect to $\boldsymbol{x}$ and $\boldsymbol{\lambda}$, respectively. Each $\nabla\boldsymbol{g}_i(\boldsymbol{x}_i^{(k)})$ is a $P\times K$ Jacobian matrix with rows equal to the subgradients $\nabla g_{ip}^T(\boldsymbol{x}_i)$, $p=1,\ldots,P$ (gradients if they are continuously differentiable), and each $\nabla\boldsymbol{f}_i(\boldsymbol{x}_i^{(k)})$ is a $M\times K$ Jacobian matrix with rows containing the gradients $\nabla f_{im}^T(\boldsymbol{x}_i)$, $m=1,\ldots,M$.

The idea behind the PD subgradient method lies in the well-known saddle-point relation:

Theorem 1: *(Saddle-Point Theorem) [37] The point* $(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star)\in\mathcal{X}\times\mathbb{R}_+^P$ *is a primal-dual solution pair of problems (2) and (5) if and only if there holds*

$$\mathcal{L}(\boldsymbol{x}^\star,\boldsymbol{\lambda})\le\mathcal{L}(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star)\le\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}^\star)\ \forall\boldsymbol{x}\in\mathcal{X},\quad\boldsymbol{\lambda}\succeq\boldsymbol{0}. \tag{9}$$

According to Theorem 1, if the PD subgradient method converges to a saddle point of the Lagrangian function (6), then it solves the original problem (2). Convergence properties of the PD method in (7) have been studied extensively; see, for example, [26], [27], [39]. In such methods, typically a subsequence of the sequence $(\boldsymbol{x}^{(k)},\boldsymbol{\lambda}^{(k)})$ converges to a saddle point of the Lagrangian function in (6). To ensure the convergence of the whole sequence $(\boldsymbol{x}^{(k)},\boldsymbol{\lambda}^{(k)})$, it is often assumed that the Lagrangian function is strictly convex in $\boldsymbol{x}$ and strictly concave in $\boldsymbol{\lambda}$, which does not hold in general however.

One of the approaches to circumventing this condition is the primal-dual perturbed (PDP) subgradient method in [21], [22]. Specifically, [21] suggests to update $\boldsymbol{x}^{(k-1)}$ and $\boldsymbol{\lambda}^{(k-1)}$

based on some perturbation points, denoted by $\hat{\boldsymbol{\alpha}}^{(k)}$ and $\hat{\boldsymbol{\beta}}^{(k)}$, respectively. The PDP updates are

$$\boldsymbol{x}^{(k)} = \mathcal{P}_{\mathcal{X}}\left(\boldsymbol{x}^{(k-1)} - a_k \mathcal{L}_{\boldsymbol{x}}\left(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}\right)\right) \quad (10a)$$

$$\boldsymbol{\lambda}^{(k)} = \left(\boldsymbol{\lambda}^{(k-1)} + a_k \mathcal{L}_{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \boldsymbol{\lambda}^{(k-1)}\right)\right)^{+}. \quad (10b)$$

Note that, in (10a), we have replaced $\boldsymbol{\lambda}^{(k-1)}$ by $\hat{\boldsymbol{\beta}}^{(k)}$, and, in (10b), replaced $\boldsymbol{x}^{(k-1)}$ by $\hat{\boldsymbol{\alpha}}^{(k)}$, and thus $\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)})$ and $\mathcal{L}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\alpha}}^{(k)}, \boldsymbol{\lambda}^{(k-1)})$ are perturbed subgradients. It was shown in [21] that, with carefully chosen $(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\beta}}^{(k)})$ and the step size a_k, the primal-dual iterates in (10) converge to a saddle point of (5), without any strict convexity and concavity assumptions on $\mathcal{L}$.

There are several ways to generate the perturbation points $\hat{\boldsymbol{\alpha}}^{(k)}$ and $\hat{\boldsymbol{\beta}}^{(k)}$. Our interests lie specifically on those that are computationally as efficient as the PD subgradient updates in (10). Depending on the smoothness of $\{g_{ip}\}_{p=1}^{P}$, we consider the following two methods:

Gradient Perturbation Points: A simple approach to computing the perturbation points is using the conventional gradient updates exactly as in (7), i.e.

$$\hat{\boldsymbol{\alpha}}^{(k)} = \mathcal{P}_{\mathcal{X}}\left(\boldsymbol{x}^{(k-1)} - \rho_1 \mathcal{L}_{\boldsymbol{x}}\left(\boldsymbol{x}^{(k-1)}, \boldsymbol{\lambda}^{(k-1)}\right)\right) \quad (11a)$$

$$\hat{\boldsymbol{\beta}}^{(k)} = \left(\boldsymbol{\lambda}^{(k-1)} + \rho_2 \mathcal{L}_{\boldsymbol{\lambda}}\left(\boldsymbol{x}^{(k-1)}, \boldsymbol{\lambda}^{(k-1)}\right)\right)^{+} \quad (11b)$$

where $\rho_1 > 0$ and $\rho_2 > 0$ are constants. The PDP subgradient method thus combines (10) and (11), which involve two primal and dual subgradient updates. Even though the updates are relatively simple, this method requires smooth constraint functions g_{ip}, $p = 1, \ldots, P$.

In cases when g_{ip}, $p = 1, \ldots, P$, are non-smooth, we propose to use the following proximal perturbation point approach, which is novel and has not appeared in earlier works [21], [22].

Proximal Perturbation Points: When g_{ip}, $p = 1, \ldots, P$, are non-smooth, we compute the perturbation point $\hat{\boldsymbol{\alpha}}^{(k)}$ by the following proximal gradient update[2] [40]:

$$\hat{\boldsymbol{\alpha}}^{(k)} = \arg\min_{\boldsymbol{\alpha}\in\mathcal{X}} \left\{ \sum_{i=1}^{N} \boldsymbol{g}_i^T(\boldsymbol{\alpha}_i)\boldsymbol{\lambda}^{(k-1)} + \frac{1}{2\rho_1}\left\|\boldsymbol{\alpha} - \left(\boldsymbol{x}^{(k-1)} - \rho_1 \nabla \bar{\mathcal{F}}\left(\boldsymbol{x}^{(k-1)}\right)\right)\right\|^2 \right\} \quad (12)$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_N^T)^T$ and

$$\nabla \bar{\mathcal{F}}\left(\boldsymbol{x}^{(k-1)}\right) = \begin{bmatrix} \nabla \boldsymbol{f}_1^T\left(\boldsymbol{x}_1^{(k-1)}\right) \nabla \mathcal{F}\left(\sum_{i=1}^{N} \boldsymbol{f}_i\left(\boldsymbol{x}_i^{(k-1)}\right)\right) \\ \vdots \\ \nabla \boldsymbol{f}_N^T\left(\boldsymbol{x}_N^{(k-1)}\right) \nabla \mathcal{F}\left(\sum_{i=1}^{N} \boldsymbol{f}_i\left(\boldsymbol{x}_i^{(k-1)}\right)\right) \end{bmatrix}. \quad (13)$$

It is worthwhile to note that, when g_{ip}, $p = 1, \ldots, P$, are some *sparsity promoting functions* (e.g., the 1-norm, 2-norm and the nuclear norm) that often arise in sparse optimization problems [7], [41], [42], the proximal perturbation point in (12) can be solved very efficiently and may even have closed-form solutions. For example, if $\boldsymbol{g}_i(\boldsymbol{\alpha}_i) = \|\boldsymbol{\alpha}_i\|_1$ for all i ($P = 1$), and $\mathcal{X} = \mathbb{R}^{KN}$, (12) has a closed-form solution known as the soft thresholding operator [7]

$$\hat{\boldsymbol{\alpha}}^{(k)} = \left(\boldsymbol{b} - \lambda^{(k-1)}\rho_1 \mathbf{1}\right)^{+} + \left(-\boldsymbol{b} - \lambda^{(k-1)}\rho_1 \mathbf{1}\right)^{+} \quad (14)$$

where $\boldsymbol{b} = \boldsymbol{x}^{(k-1)} - \rho_1 \nabla \bar{\mathcal{F}}(\boldsymbol{x}^{(k-1)})$ and $\mathbf{1}$ is an all-one vector.

III. Proposed Consensus-Based Distributed PDP Algorithm

Our goal is to develop a distributed counterpart of the PDP subgradient method in (10). Consider the following saddle-point problem:

$$\max_{\boldsymbol{\lambda}\in\mathcal{D}} \left\{ \min_{\substack{\boldsymbol{x}_i\in\mathcal{X}_i, \\ i=1,\ldots,N}} \mathcal{L}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{\lambda}) \right\} \quad (15)$$

where

$$\mathcal{D} = \left\{ \boldsymbol{\lambda} \succeq \mathbf{0} | \ \|\boldsymbol{\lambda}\| \le D_\lambda \triangleq \frac{\bar{\mathcal{F}}(\bar{\boldsymbol{x}}) - \tilde{q}}{\gamma} + \delta \right\} \quad (16)$$

in which $\bar{\boldsymbol{x}} = (\bar{\boldsymbol{x}}_1^T, \ldots, \bar{\boldsymbol{x}}_N^T)^T$ is a Slater point of (2), $\tilde{q} = \min_{\boldsymbol{x}_i\in\mathcal{X}_i, i=1,\ldots,N} \mathcal{L}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \tilde{\boldsymbol{\lambda}})$ is the dual function value for some arbitrary $\tilde{\boldsymbol{\lambda}} \succeq \mathbf{0}$, $\gamma = \min_{p=1,\ldots,P}\{-\sum_{i=1}^{N} g_{ip}(\bar{\boldsymbol{x}}_i)\}$, and $\delta > 0$ is arbitrary. It has been shown in [43] that the optimal dual solution $\hat{\boldsymbol{\lambda}}^{\star}$ of (5) satisfies

$$\|\hat{\boldsymbol{\lambda}}^{\star}\| \le \frac{\bar{\mathcal{F}}(\bar{\boldsymbol{x}}) - \tilde{q}}{\gamma} \quad (17)$$

and thus $\hat{\boldsymbol{\lambda}}^{\star}$ lies in $\mathcal{D}$. Here we consider the saddle point problem (15), instead of the original Lagrange dual problem (5), because $\mathcal{D}$ bounds the dual variable $\boldsymbol{\lambda}$ and thus also bounds the subgradient $\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{x}^{(k)}, \boldsymbol{\lambda}^{(k)})$ in (8a). This property is important in building the convergence of the distributed algorithm to be discussed shortly. Both (5) and (15) have the same optimal dual solution $\hat{\boldsymbol{\lambda}}^{\star}$ and attain the same optimal objective value. One can further verify that any saddle point of (5) is also a saddle point of (15). However, to relate the saddle points of (15) to solutions of problem (2) some conditions are needed, as given in the following proposition.

Proposition 1 (Primal-Dual Optimality Conditions) [44]: *Let the Slater condition hold and let $(\hat{\boldsymbol{x}}_1^{\star}, \ldots, \hat{\boldsymbol{x}}_N^{\star}, \hat{\boldsymbol{\lambda}}^{\star})$ be a saddle point of (15). Then $(\hat{\boldsymbol{x}}_1^{\star}, \ldots, \hat{\boldsymbol{x}}_N^{\star})$ is an optimal solution for problem (2) if and only if*

$$\sum_{i=1}^{N} \boldsymbol{g}_i(\hat{\boldsymbol{x}}_i^{\star}) \preceq \mathbf{0} \text{ and } (\hat{\boldsymbol{\lambda}}^{\star})^T \left(\sum_{i=1}^{N} \boldsymbol{g}_i(\hat{\boldsymbol{x}}_i^{\star})\right) = \mathbf{0}.$$

To have a distributed optimization algorithm for solving (15), in addition to $\boldsymbol{x}_i^{(k)}$, we let each agent i have a local copy of the dual iterate $\boldsymbol{\lambda}^{(k)}$, denoted by $\boldsymbol{\lambda}_i^{(k)}$. Moreover, each agent i owns two auxiliary variables, denoted by $\boldsymbol{y}_i^{(k)}$ and $\boldsymbol{z}_i^{(k)}$,

[2]If not mentioned specifically, the norm function $\|\cdot\|$ stands for the Euclidean norm.

representing respectively the local estimates of the average values of the argument function $1/N\sum_{i=1}^N \boldsymbol{f}_i(\boldsymbol{x}_i^{(k)})$ and of the inequality constraint function $1/N\sum_{i=1}^N \boldsymbol{g}_i(\boldsymbol{x}_i^{(k)})$, for all $i=1,\ldots,N$. We consider a *time-varying synchronous network* model [11], where the network of agents at time k is represented by a weighted directed graph $\mathcal{G}(k)=(\mathcal{V},\mathcal{E}(k),\boldsymbol{W}(k))$. Here $(i,j)\in\mathcal{E}(k)$ if and only if agent i can receive messages from agent j, and $\boldsymbol{W}(k)\in\mathbb{R}^{N\times N}$ is a weight matrix with each entry $[\boldsymbol{W}(k)]_{ij}$ representing a weight that agent i assigns to the incoming message on link (i,j) at time k. If $(i,j)\in\mathcal{E}(k)$, then $[\boldsymbol{W}(k)]_{ij}>0$ and $[\boldsymbol{W}(k)]_{ij}=0$ otherwise. The agents exchange messages with their neighbors (according to the network graph $\mathcal{G}(k)$) in order to achieve consensus on $\boldsymbol{\lambda}^{(k)}$, $\sum_{i=1}^N \boldsymbol{g}_i(\boldsymbol{x}_i^{(k)})$ and $\sum_{i=1}^N \boldsymbol{f}_i(\boldsymbol{x}_i^{(k)})$; while computing local perturbation points and primal-dual (sub-)gradient updates locally. Specifically, the proposed distributed consensus-based PDP method consists of the following steps at each iteration k:

1) **Averaging consensus:** For $i=1,\ldots,N$, each agent i sends $\boldsymbol{y}_i^{(k-1)}$, $\boldsymbol{z}_i^{(k-1)}$ and $\boldsymbol{\lambda}_i^{(k-1)}$ to all its neighbors j satisfying $(j,i)\in\mathcal{E}(k)$; it also receives $\boldsymbol{y}_j^{(k-1)}$, $\boldsymbol{z}_j^{(k-1)}$ and $\boldsymbol{\lambda}_j^{(k-1)}$ from its neighbors, and combines the received estimates, as follows:

$$\tilde{\boldsymbol{y}}_i^{(k)}=\sum_{j=1}^N [\boldsymbol{W}(k)]_{ij}\,\boldsymbol{y}_j^{(k-1)},\ \tilde{\boldsymbol{z}}_i^{(k)}=\sum_{j=1}^N [\boldsymbol{W}(k)]_{ij}\,\boldsymbol{z}_j^{(k-1)},$$
$$\tilde{\boldsymbol{\lambda}}_i^{(k)}=\sum_{j=1}^N [\boldsymbol{W}(k)]_{ij}\,\boldsymbol{\lambda}_j^{(k-1)}. \tag{18}$$

2) **Perturbation point computation:** For $i=1,\ldots,N$, if functions g_{ip}, $p=1,\ldots,P$, are smooth, then each agent i computes the local perturbation points by

$$\boldsymbol{\alpha}_i^{(k)}=\mathcal{P}_{\mathcal{X}_i}\Big(\boldsymbol{x}_i^{(k-1)}-\rho_1\Big[\nabla\boldsymbol{f}_i^T\Big(\boldsymbol{x}_i^{(k-1)}\Big)\nabla\mathcal{F}\Big(N\tilde{\boldsymbol{y}}_i^{(k)}\Big)+\nabla\boldsymbol{g}_i^T\Big(\boldsymbol{x}_i^{(k-1)}\Big)\tilde{\boldsymbol{\lambda}}_i^{(k)}\Big]\Big), \tag{19a}$$

$$\boldsymbol{\beta}_i^{(k)}=\mathcal{P}_{\mathcal{D}}\Big(\tilde{\boldsymbol{\lambda}}_i^{(k)}+\rho_2 N\tilde{\boldsymbol{z}}_i^{(k)}\Big). \tag{19b}$$

Note that, comparing to (11) and (12), agent i here uses the most up-to-date estimates $N\tilde{\boldsymbol{y}}_i^{(k)}$, $N\tilde{\boldsymbol{z}}_i^{(k)}$ and $\tilde{\boldsymbol{\lambda}}_i^{(k)}$ in place of $\sum_{i=1}^N \boldsymbol{f}_i(\boldsymbol{x}_i^{(k-1)})$, $\sum_{i=1}^N \boldsymbol{g}_i(\boldsymbol{x}_i^{(k-1)})$ and $\boldsymbol{\lambda}^{(k-1)}$. If g_{ip}, $p=1,\ldots,P$, are non-smooth, agent i instead computes $\boldsymbol{\alpha}_i^{(k)}$ by

$$\boldsymbol{\alpha}_i^{(k)}=\arg\min_{\boldsymbol{\alpha}_i\in\mathcal{X}_i}\Big\{\boldsymbol{g}_i^T(\boldsymbol{\alpha}_i)\tilde{\boldsymbol{\lambda}}_i^{(k)}+\frac{1}{2\rho_1}\Big\|\boldsymbol{\alpha}_i-\Big(\boldsymbol{x}_i^{(k-1)}-\rho_1\nabla\boldsymbol{f}_i^T\Big(\boldsymbol{x}_i^{(k-1)}\Big)\nabla\mathcal{F}\Big(N\tilde{\boldsymbol{y}}_i^{(k)}\Big)\Big)\Big\|^2\Big\} \tag{20}$$

for $i=1,\ldots,N$.

3) **Primal-dual perturbed subgradient update:** For $i=1,\ldots,N$, each agent i updates its primal and dual variables $(\boldsymbol{x}_i^{(k)},\boldsymbol{\lambda}_i^{(k)})$ based on the local perturbation point $(\boldsymbol{\alpha}_i^{(k)},\boldsymbol{\beta}_i^{(k)})$

$$\boldsymbol{x}_i^{(k)}=\mathcal{P}_{\mathcal{X}_i}\Big(\boldsymbol{x}_i^{(k-1)}-a_k\Big[\nabla\boldsymbol{f}_i^T\Big(\boldsymbol{x}_i^{(k-1)}\Big)\nabla\mathcal{F}\Big(N\tilde{\boldsymbol{y}}_i^{(k)}\Big)+\nabla\boldsymbol{g}_i^T\Big(\boldsymbol{x}_i^{(k-1)}\Big)\boldsymbol{\beta}_i^{(k)}\Big]\Big) \tag{21}$$

$$\boldsymbol{\lambda}_i^{(k)}=\mathcal{P}_{\mathcal{D}}\Big(\tilde{\boldsymbol{\lambda}}_i^{(k)}+a_k\boldsymbol{g}_i\Big(\boldsymbol{\alpha}_i^{(k)}\Big)\Big). \tag{22}$$

4) **Auxiliary variable update:** For $i=1,\ldots,N$, each agent i updates variable $\boldsymbol{y}_i^{(k)}$, $\boldsymbol{z}_i^{(k)}$ with the changes of the local argument function $\boldsymbol{f}_i(\boldsymbol{x}_i^{(k)})$ and the constraint function $\boldsymbol{g}_i(\boldsymbol{x}_i^{(k)})$

$$\boldsymbol{y}_i^{(k)}=\tilde{\boldsymbol{y}}_i^{(k)}+\boldsymbol{f}_i\Big(\boldsymbol{x}_i^{(k)}\Big)-\boldsymbol{f}_i\Big(\boldsymbol{x}_i^{(k-1)}\Big) \tag{23}$$

$$\boldsymbol{z}_i^{(k)}=\tilde{\boldsymbol{z}}_i^{(k)}+\boldsymbol{g}_i\Big(\boldsymbol{x}_i^{(k)}\Big)-\boldsymbol{g}_i\Big(\boldsymbol{x}_i^{(k-1)}\Big). \tag{24}$$

Algorithm 1 summarizes the above steps. We prove that Algorithm 1 converges under proper problem and network assumptions in the next section. Readers who are interested more in numerical performance of Algorithm 1 may go directly to Section V.

Algorithm 1 Distributed Consensus-Based PDP Algorithm

1: **Given** initial variables $\boldsymbol{x}_i^{(0)}\in\mathcal{X}_i$, $\boldsymbol{\lambda}_i^{(0)}\succeq\boldsymbol{0}$, $\boldsymbol{y}_i^{(0)}=\boldsymbol{f}_i(\boldsymbol{x}_i^{(0)})$ and $\boldsymbol{z}_i^{(0)}=\boldsymbol{g}_i(\boldsymbol{x}_i^{(0)})$ for each agent i, $i=1,\ldots,N$; Set $k=1$.
2: **repeat**
3: **Averaging consensus**: For $i=1,\ldots,N$, each agent i computes (18).
4: **Perturbation point computation**: For $i=1,\ldots,N$, if $\{g_{ip}\}_{p=1}^P$ are smooth, then each agent i computes the local perturbation points by (19); otherwise, each agent i instead computes $\boldsymbol{\alpha}_i^{(k)}$ by (20).
5: **Local variable updates**: For $i=1,\ldots,N$, each agent i updates (21), (22), (23) and (24) sequentially.
6: **Set** $k=k+1$.
7: **until** a predefined stopping criterion (e.g., a maximum iteration number) is satisfied.

IV. Convergence Analysis

Next, in Section IV-A, we present additional assumptions on problem (2) and the network model. The main convergence results are presented in Section IV-B. The proofs are presented in Sections IV-C and D.

A. Assumptions

Our results will make use of the following assumption.

Assumption 1*: (a) The sets $\mathcal{X}_i$, $i=1,\ldots,N$, are compact. In particular, for $i=1,\ldots,N$, there is a constant $D_x>0$ such that*

$$\|\boldsymbol{x}_i\|\le D_x\quad\forall\boldsymbol{x}_i\in\mathcal{X}_i; \tag{25}$$

(b) The functions $f_{i1},\ldots,f_{iM}$, $i=1,\ldots,N$, are continuously differentiable.

Note that Assumptions 1(a) and (b) imply that $f_{i1},\ldots,f_{iM}$ have uniformly bounded gradients (denoted by ∇f_{im}, $m=1,\ldots,M$) and are Lipschitz continuous, i.e., for some $L_f>0$

$$\max_{1\le m\le M}\|\nabla f_{im}(\boldsymbol{x}_i)\|\le L_f,\quad\forall\boldsymbol{x}_i\in\mathcal{X}_i \tag{26}$$

$$\max_{1\le m\le M}|f_{im}(\boldsymbol{x}_i)-f_{im}(\boldsymbol{y}_i)|\le L_f\|\boldsymbol{x}_i-\boldsymbol{y}_i\|\ \ \forall\boldsymbol{x}_i,\boldsymbol{y}_i\in\mathcal{X}_i. \tag{27}$$

Similarly, Assumption 1(a) and the convexity of functions $g_{i1}, \ldots, g_{iP}$ imply that all g_{ip} have uniformly bounded subgradients, which is equivalent to all g_{ip} being Lipschitz continuous. Thus, for some $L_g > 0$, we have

$$\max_{1\le p\le P} \|\nabla g_{ip}(\boldsymbol{x}_i)\| \le L_g \quad \forall \boldsymbol{x}_i \in \mathcal{X}_i \tag{28}$$

$$\max_{1\le p\le P} |g_{ip}(\boldsymbol{x}_i) - g_{ip}(\boldsymbol{y}_i)| \le L_g \|\boldsymbol{x}_i - \boldsymbol{y}_i\| \quad \forall \boldsymbol{x}_i, \boldsymbol{y}_i \in \mathcal{X}_i. \tag{29}$$

In addition, by Assumption 1 and the continuity of each g_{ip} (which is implied by the convexity of g_{ip}) each $\boldsymbol{f}_i$ and $\boldsymbol{g}_i$ are also bounded on $\mathcal{X}$, i.e., there exist constants $C_f > 0$ and $C_g > 0$ such that for all $i = 1, \ldots, N$

$$\|\boldsymbol{f}_i(\boldsymbol{x}_i)\| \le C_f, \quad \|\boldsymbol{g}_i(\boldsymbol{x}_i)\| \le C_g, \quad \forall \boldsymbol{x}_i \in \mathcal{X}_i \tag{30}$$

where $\|\boldsymbol{f}_i(\boldsymbol{x}_i)\| = \sqrt{\sum_{m=1}^{M} f_{im}^2(\boldsymbol{x}_i)}$ and $\|\boldsymbol{g}_i(\boldsymbol{x}_i)\| = \sqrt{\sum_{p=1}^{P} g_{ip}^2(\boldsymbol{x}_i)}$.

We also make use of the following assumption on the network utility costs $\mathcal{F}$ and $\bar{\mathcal{F}}$:

Assumption 2*: (a) The function $\mathcal{F}$ is continuously differentiable and has bounded and Lipschitz continuous gradients, i.e., for some $G_{\mathcal{F}} > 0$ and $L_{\mathcal{F}} > 0$, we have*

$$\|\nabla \mathcal{F}(\boldsymbol{x}) - \nabla \mathcal{F}(\boldsymbol{y})\| \le G_{\mathcal{F}} \|\boldsymbol{x} - \boldsymbol{y}\| \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^M \tag{31}$$

$$\|\nabla \mathcal{F}(\boldsymbol{y})\| \le L_{\mathcal{F}} \quad \forall \boldsymbol{y} \in \mathbb{R}^M \tag{32}$$

(b) The function $\bar{\mathcal{F}}$ has Lipschitz continuous gradients, i.e., for some $G_{\bar{\mathcal{F}}} > 0$

$$\left\|\nabla \bar{\mathcal{F}}(\boldsymbol{x}) - \nabla \bar{\mathcal{F}}(\boldsymbol{y})\right\| \le G_{\bar{\mathcal{F}}} \|\boldsymbol{x} - \boldsymbol{y}\| \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}. \tag{33}$$

Note that the convexity of $\bar{\mathcal{F}}$ Assumption 1(a) indicate that $\bar{\mathcal{F}}$ is Lipschitz continuous, i.e., for some $L_{\bar{\mathcal{F}}} > 0$

$$\left|\bar{\mathcal{F}}(\boldsymbol{x}) - \bar{\mathcal{F}}(\boldsymbol{y})\right| \le L_{\bar{\mathcal{F}}} \|\boldsymbol{x} - \boldsymbol{y}\| \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}. \tag{34}$$

Assumptions 1 and 2 are used to ensure that the (sub-)gradients of the Lagrangian function $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ with respect to $\boldsymbol{x}$ are well behaved for applying (sub-)gradient-based methods. In cases that g_{ip}, $p = 1, \ldots, P$, are smooth, we make use of the following additional assumption:

Assumption 3*: The functions g_{ip}, $p = 1, \ldots, P$, are continuously differentiable and have Lipschitz continuous gradients, i.e., there exists a constant $G_g > 0$ such that*

$$\max_{1\le p\le P} \|\nabla g_{ip}(\boldsymbol{x}_i) - \nabla g_{ip}(\boldsymbol{y}_i)\| \le G_g \|\boldsymbol{x}_i - \boldsymbol{y}_i\| \forall \boldsymbol{x}_i,\ \boldsymbol{y}_i \in \mathcal{X}_i. \tag{35}$$

We also have the following assumption on the network model [11], [17]:

Assumption 4*: The weighted graphs $\mathcal{G}(k) = (\mathcal{V}, \mathcal{E}(k), \boldsymbol{W}(k))$ satisfy:*

(a) *There exists a scalar $0 < \eta < 1$ such that $[\boldsymbol{W}(k)]_{ii} > \eta$ for all i, k and $[\boldsymbol{W}(k)]_{ij} > \eta$ if $[\boldsymbol{W}(k)]_{ij} > 0$.*

(b) *$\boldsymbol{W}(k)$ is doubly stochastic: $\sum_{j=1}^{N} [\boldsymbol{W}(k)]_{ij} = 1$ for all i, k and $\sum_{i=1}^{N} [\boldsymbol{W}(k)]_{ij} = 1$ $\forall j, k$.*

(c) *There is an integer Q such that $(\mathcal{V}, \cup_{\ell=1,\ldots,Q} \mathcal{E}(k+\ell))$ is strongly connected for all k.*

Assumption 4 ensures that all the agents can sufficiently and equally influence each other in a long run.

B. Main Convergence Results

Let $A_k = \sum_{\ell=1}^{k} a_\ell$, and let

$$\hat{\boldsymbol{x}}_i^{(k-1)} = \frac{1}{A_k} \sum_{\ell=1}^{k} a_\ell \boldsymbol{x}_i^{(\ell-1)}, \quad i = 1, \ldots, N \tag{36}$$

be the running weighted-averages of the primal iterates $\boldsymbol{x}_i^{(0)}, \ldots, \boldsymbol{x}_i^{(k-1)}$ generated by agent i until time $k-1$. Our main convergence result for Algorithm 1 is given in the following theorem.

Theorem 2*: Let Assumptions 1–4 hold, and let $\rho_1 \le 1/(G_{\bar{\mathcal{F}}} + D_\lambda \sqrt{P} G_g)$. Assume that the step size sequence $\{a_k\}$ is non-increasing and such that $a_k > 0$ for all $k \ge 1$, $\sum_{k=1}^{\infty} a_k = \infty$ and $\sum_{k=1}^{\infty} a_k^2 < \infty$. Let the sequences $\{\hat{\boldsymbol{x}}^{(k)}\}$ and $\{\boldsymbol{\lambda}_i^{(k)}\}$, $i = 1, \ldots, N$, be generated by Algorithm 1 using the gradient perturbation points in (19). Then, $\{\hat{\boldsymbol{x}}^{(k)}\}$ and $\{\boldsymbol{\lambda}_i^{(k)}\}$, $i = 1, \ldots, N$, converge to an optimal primal solution $\boldsymbol{x}^\star \in \mathcal{X}$ and an optimal dual solution $\boldsymbol{\lambda}^\star$ of problem (2), respectively.*

Theorem 2 indicates that the proposed distributed primal-dual algorithm asymptotically yields an optimal primal and dual solution pair for the original problem (2). The same convergence result holds if the constraint functions g_{ip}, $p = 1, \ldots, P$, are non-smooth and the perturbation points $\boldsymbol{\alpha}_i^{(k)}$ are computed according to (20).

Theorem 3*: Let Assumptions 1, 2, and 4 hold, and let $\rho_1 \le 1/G_{\bar{\mathcal{F}}}$. Assume that the step size sequence $\{a_k\}$ is non-increasing and such that $a_k > 0$ for all $k \ge 1$, $\sum_{k=1}^{\infty} a_k = \infty$ and $\sum_{k=1}^{\infty} a_k^2 < \infty$. Let the sequences $\{\hat{\boldsymbol{x}}^{(k)}\}$ and $\{\boldsymbol{\lambda}_i^{(k)}\}$, $i = 1, \ldots, N$, be generated by Algorithm 1 using the perturbation points in (20) and (19b). Then, $\{\hat{\boldsymbol{x}}^{(k)}\}$ and $\{\boldsymbol{\lambda}_i^{(k)}\}$, $i = 1, \ldots, N$, converge to an optimal primal solution $\boldsymbol{x}^\star \in \mathcal{X}$ and an optimal}dual solution $\boldsymbol{\lambda}^\star$ of problem (2), respectively.*

The proofs of Theorems 2 and 3 are presented in the next two subsections, respectively.

C. Proof of Theorem 2

In this subsection, we present the major steps for proving Theorem 2. Three key lemmas that will be used in the proof are presented first. The first is a deterministic version of the lemma in [45, Lemma 11, Chapter 2.2].

Lemma 1*: Let $\{b_k\}$, $\{d_k\}$ and $\{c_k\}$ be non-negative sequences. Suppose that $\sum_{k=1}^{\infty} c_k < \infty$ and*

$$b_k \le b_{k-1} - d_{k-1} + c_{k-1} \quad \forall k \ge 1$$

then the sequence $\{b_k\}$ converges and $\sum_{k=1}^{\infty} d_k < \infty$.

Moreover, by extending the results in [17, Theorem 4.2] and [11, Lemma 8(a)], we establish the following result on the consensus of $\{\boldsymbol{\lambda}_i^{(k)}\}$, $\{\boldsymbol{y}_i^{(k)}\}$, and $\{\boldsymbol{z}_i^{(k)}\}$ among agents.

Lemma 2: *Suppose that Assumptions 1 and 4 hold. If $\{a_k\}$ is a positive, non-increasing sequence satisfying $\sum_{k=1}^{\infty} a_k^2 < \infty$, then*

$$\sum_{k=1}^{\infty} a_k \left\|\boldsymbol{\lambda}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k)}\right\| < \infty,\ \lim_{k\to\infty}\left\|\boldsymbol{\lambda}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k)}\right\| = 0 \quad (37)$$

$$\sum_{k=1}^{\infty} a_k \left\|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\right\| < \infty,\ \lim_{k\to\infty}\left\|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\right\| = 0, \quad (38)$$

$$\sum_{k=1}^{\infty} a_k \left\|\tilde{\boldsymbol{y}}_i^{(k)} - \hat{\boldsymbol{y}}^{(k-1)}\right\| < \infty,\ \lim_{k\to\infty}\left\|\tilde{\boldsymbol{y}}_i^{(k)} - \hat{\boldsymbol{y}}^{(k-1)}\right\| = 0, \quad (39)$$

$$\sum_{k=1}^{\infty} a_k \left\|\tilde{\boldsymbol{z}}_i^{(k)} - \hat{\boldsymbol{z}}^{(k-1)}\right\| < \infty,\ \lim_{k\to\infty}\left\|\tilde{\boldsymbol{z}}_i^{(k)} - \hat{\boldsymbol{z}}^{(k-1)}\right\| = 0 \quad (40)$$

for all $i = 1, \ldots, N$, where

$$\hat{\boldsymbol{y}}^{(k)} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{f}_i\left(\boldsymbol{x}_i^{(k)}\right), \quad \hat{\boldsymbol{z}}^{(k)} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{g}_i\left(\boldsymbol{x}_i^{(k)}\right),$$
$$\hat{\boldsymbol{\lambda}}^{(k)} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\lambda}_i^{(k)}. \quad (41)$$

The proof is omitted here due to the space limitation; interested readers may refer to the electronic companion [46]. Lemma 2 implies that the local variables $\boldsymbol{\lambda}_i^{(k)}$, $\boldsymbol{y}_i^{(k)}$ and $\boldsymbol{z}_i^{(k)}$ at distributed agents will eventually achieve consensus on the values of $\hat{\boldsymbol{\lambda}}^{(k)}$ $\hat{\boldsymbol{y}}^{(k)}$ and $\hat{\boldsymbol{z}}^{(k)}$, respectively.

The local perturbation points $\boldsymbol{\alpha}_i^{(k)}$ and $\boldsymbol{\beta}_i^{(k)}$ in (19) and (20) will also achieve consensus asymptotically. In particular, following (11), we define:

$$\hat{\boldsymbol{\alpha}}_i^{(k)} = \mathcal{P}_{\mathcal{X}_i}\Big(\boldsymbol{x}_i^{(k-1)} - \rho_1\Big[\nabla \boldsymbol{f}_i^T\left(\boldsymbol{x}_i^{(k-1)}\right)\nabla\mathcal{F}\left(N\hat{\boldsymbol{y}}^{(k)}\right) + \nabla \boldsymbol{g}_i^T\left(\boldsymbol{x}_i^{(k-1)}\right)\hat{\boldsymbol{\lambda}}^{(k-1)}\Big]\Big), \quad (42a)$$

$$\hat{\boldsymbol{\beta}}^{(k)} = \mathcal{P}_{\mathcal{D}}\left(\hat{\boldsymbol{\lambda}}^{(k-1)} + \rho_2 N\hat{\boldsymbol{z}}^{(k)}\right). \quad (42b)$$

for $i = 1, \ldots, N$, as the 'centralized' counterparts of (19); similarly, following (12), we define

$$\hat{\boldsymbol{\alpha}}_i^{(k)} = \arg\min_{\boldsymbol{\alpha}_i\in\mathcal{X}_i} \boldsymbol{g}_i^T(\boldsymbol{\alpha}_i)\hat{\boldsymbol{\lambda}}^{(k-1)} + \frac{1}{2\rho_1}\Big\|\boldsymbol{\alpha}_i - \left(\boldsymbol{x}_i^{(k-1)} - \rho_1\nabla \boldsymbol{f}_i^T\left(\boldsymbol{x}_i^{(k-1)}\right)\nabla\mathcal{F}\left(N\hat{\boldsymbol{y}}^{(k-1)}\right)\right)\Big\|^2 \quad (43)$$

for $i = 1, \ldots, N$, as the centralized counterparts of the proximal perturbation point in (20). We show in Appendix A the following lemma.

Lemma 3: *Let Assumptions 1 and 2 hold. For $\{\boldsymbol{\alpha}_i^{(k)}, \boldsymbol{\beta}_i^{(k)}\}_{i=1}^N$ in (19) and $(\hat{\boldsymbol{\alpha}}_1^{(k)}, \ldots, \hat{\boldsymbol{\alpha}}_N^{(k)}, \hat{\boldsymbol{\beta}}^{(k)})$ in (42), it holds that*

$$\left\|\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{\alpha}_i^{(k)}\right\| \le \rho_1 L_g\sqrt{P}\left\|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\right\| + \rho_1 G_{\mathcal{F}} L_f\sqrt{M}N\left\|\tilde{\boldsymbol{y}}_i^{(k)} - \hat{\boldsymbol{y}}^{(k-1)}\right\| \quad (44)$$

$$\left\|\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}_i^{(k)}\right\| \le \left\|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\right\| + \rho_2 N\left\|\tilde{\boldsymbol{z}}_i^{(k)} - \hat{\boldsymbol{z}}^{(k-1)}\right\| \quad (45)$$

$i = 1, \ldots, N$. Equation (44) also holds for the proximal perturbation point $\boldsymbol{\alpha}_i^{(k)}$ in (20) and $\hat{\boldsymbol{\alpha}}_i^{(k)}$ in (43).

Lemma 3 says that, when $\tilde{\boldsymbol{\lambda}}_i^{(k)}$, $\tilde{\boldsymbol{y}}_i^{(k)}$ and $\tilde{\boldsymbol{z}}_i^{(k)}$ at distributed agents achieve consensus, each $\boldsymbol{\alpha}_i^{(k)}$ converges to $\hat{\boldsymbol{\alpha}}_i^{(k)}$, and all the $\boldsymbol{\beta}_i^{(k)}$ converge to the common point $\hat{\boldsymbol{\beta}}^{(k)}$.

Now we are ready to prove Theorem 2. The proof primarily consists of showing two facts: (a) the primal-dual iterate pairs $(\hat{\boldsymbol{x}}_1^{(k)}, \ldots, \hat{\boldsymbol{x}}_N^{(k)}, \hat{\boldsymbol{\lambda}}^{(k)})$ will converge to a saddle point of (15), and (b) $(\hat{\boldsymbol{x}}_1^{(k)}, \ldots, \hat{\boldsymbol{x}}_N^{(k)}, \hat{\boldsymbol{\lambda}}^{(k)})$ asymptotically satisfies the primal-dual optimality conditions in Proposition 1. Thus, $(\hat{\boldsymbol{x}}_1^{(k)}, \ldots, \hat{\boldsymbol{x}}_N^{(k)}, \hat{\boldsymbol{\lambda}}^{(k)})$ is asymptotically primal-dual optimal to problem (2). To show the first fact, we use (21), (22), and Lemma 3 to characterize the basic relations of the primal and dual iterates.

Lemma 4: *Let Assumptions 1 and 2 hold. Then, for any $\boldsymbol{x} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_N^T)^T \in \mathcal{X}$ and $\boldsymbol{\lambda} \in \mathcal{D}$, the following two inequalities are true*:

$$\begin{aligned}\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\right\|^2 &\le \left\|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}\right\|^2 - 2a_k\left(\mathcal{L}\left(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}\right) - \mathcal{L}\left(\boldsymbol{x}, \hat{\boldsymbol{\beta}}^{(k)}\right)\right) \\ &\quad + a_k^2 N\left(\sqrt{M}L_f L_{\mathcal{F}} + D_\lambda\sqrt{P}L_g\right)^2 + 2a_k N D_x \\ &\quad \times \sqrt{M}L_f G_{\mathcal{F}}\sum_{i=1}^{N}\left\|\tilde{\boldsymbol{y}}_i^{(k)} - \hat{\boldsymbol{y}}^{(k-1)}\right\| + 2a_k D_x\sqrt{P} \\ &\quad \times L_g\sum_{i=1}^{N}\left(\left\|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\right\| + \rho_2 N\left\|\tilde{\boldsymbol{z}}_i^{(k)} - \hat{\boldsymbol{z}}_i^{(k-1)}\right\|\right), \end{aligned} \quad (46)$$

$$\begin{aligned}\sum_{i=1}^{N}\left\|\boldsymbol{\lambda}_i^{(k)} - \boldsymbol{\lambda}\right\|^2 &\le \sum_{i=1}^{N}\left\|\boldsymbol{\lambda}_i^{(k-1)} - \boldsymbol{\lambda}\right\|^2 \\ &\quad + 2\alpha_k\left(\mathcal{L}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right) - \mathcal{L}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \boldsymbol{\lambda}\right)\right) + a_k^2 N C_g^2 \\ &\quad + 2a_k\left(2\rho_1 D_\lambda P L_g^2 + C_g\right)\left\|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\right\| \\ &\quad + 4\rho_1 N D_\lambda G_{\mathcal{F}}\sqrt{PM}L_g L_f a_k\left\|\tilde{\boldsymbol{y}}_i^{(k)} - \hat{\boldsymbol{y}}_i^{(k-1)}\right\|. \end{aligned} \quad (47)$$

The detailed proof is given in the electronic companion [46]. The second ingredient is a relation between the primal-dual iterates $(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)})$ and the perturbation points $(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\beta}}^{(k)})$, as given below.}

Lemma 5*: Let Assumptions 1, 2, and 3 hold. For the gradient perturbation points $(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\beta}}^{(k)})$ in (42), it holds true that*

$$\begin{aligned}&\mathcal{L}\left(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}\right) - \mathcal{L}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right)\\ &\quad\geq \left(\frac{1}{\rho_1} - (G_{\bar{\mathcal{F}}} + D_\lambda\sqrt{P}G_g)\right)\left\|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\right\|^2\\ &\qquad+ \frac{1}{\rho_2}\left\|\hat{\boldsymbol{\lambda}}^{(k-1)} - \hat{\boldsymbol{\beta}}^{(k)}\right\|^2. \end{aligned} \tag{48}$$

Moreover, let $\rho_1 \leq 1/(G_{\bar{\mathcal{F}}} + D_\lambda\sqrt{P}G_g)$, and suppose that $\mathcal{L}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) \to 0$ and $(\boldsymbol{x}^{(k-1)}, \boldsymbol{\lambda}^{(k-1)})$ converges to some limit point $(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star) \in \mathcal{X} \times \mathcal{D}$ as $k \to \infty$. Then $(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star)$ is a saddle point of (15).

The proof is presented in Appendix B. Using the preceding lemmas, we show the first key fact, namely, that $(\hat{\boldsymbol{x}}_1^{(k)}, \ldots, \hat{\boldsymbol{x}}_N^{(k)}, \hat{\boldsymbol{\lambda}}^{(k)})$ converges to a saddle point of (15).

Lemma 6*: Let Assumptions 1–4 hold, and let $\rho_1 \leq 1/(G_{\bar{\mathcal{F}}} + D_\lambda\sqrt{P}G_g)$. Assume that the step size $a_k > 0$ is a non-increasing sequence satisfying $\sum_{k=1}^\infty a_k = \infty$ and $\sum_{k=1}^\infty a_k^2 < \infty$. Then*

$$\lim_{k\to\infty}\left\|\boldsymbol{x}_i^{(k)} - \hat{\boldsymbol{x}}_i^\star\right\| = 0, \quad i = 1, \ldots, N,$$

$$\lim_{k\to\infty}\left\|\hat{\boldsymbol{\lambda}}^{(k)} - \hat{\boldsymbol{\lambda}}^\star\right\| = 0 \tag{49}$$

$$\lim_{k\to\infty}\left\|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\right\| = 0, \quad \lim_{k\to\infty}\left\|\hat{\boldsymbol{\lambda}}^{(k-1)} - \hat{\boldsymbol{\beta}}^{(k)}\right\| = 0 \tag{50}$$

where $\hat{\boldsymbol{x}}^\star = ((\hat{\boldsymbol{x}}_1^\star)^T, \ldots, (\hat{\boldsymbol{x}}_N^\star)^T)^T \in \mathcal{X}$ and $\hat{\boldsymbol{\lambda}}^\star \in \mathcal{D}$ form a saddle point of problem (15).

Proof: By the compactness of the set $\mathcal{X}$ and the continuity of the functions $\bar{\mathcal{F}}$ and $\boldsymbol{g}_i$, problem (2) has a solution. Due to the Slater condition, the dual problem also has a solution. By construction of the set $\mathcal{D}$ in (16), all dual optimal solutions are contained in the set $\mathcal{D}$. We let $\boldsymbol{x}^\star = ((\boldsymbol{x}_1^\star)^T, \ldots, (\boldsymbol{x}_N^\star)^T)^T \in \mathcal{X}$ and $\boldsymbol{\lambda}^\star \in \mathcal{D}$ be an arbitrary saddle point of (15), and we apply Lemma 4 with $\boldsymbol{x} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_N^T)^T = \boldsymbol{x}^\star$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}^\star$. By summing (46) and (47), we obtain the following inequality:

$$\begin{aligned}&\left(\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star\right\|^2 + \sum_{i=1}^N\left\|\boldsymbol{\lambda}_i^{(k)} - \boldsymbol{\lambda}^\star\right\|^2\right)\\ &\quad\leq \left(\left\|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^\star\right\|^2 + \sum_{i=1}^N\left\|\boldsymbol{\lambda}_i^{(k-1)} - \boldsymbol{\lambda}^\star\right\|^2\right)\\ &\qquad+ \tilde{c}_k - 2a_k\left(\mathcal{L}\left(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}\right) - \mathcal{L}\left(\boldsymbol{x}^\star, \boldsymbol{\beta}^{(k)}\right)\right.\\ &\qquad\qquad\left.-\mathcal{L}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right) + \mathcal{L}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \boldsymbol{\lambda}^\star\right)\right)\end{aligned} \tag{51}$$

where

$$\begin{aligned}\tilde{c}_k \triangleq\ & a_k^2 N\left[(\sqrt{M}L_f L_{\mathcal{F}} + D_\lambda\sqrt{P}L_g)^2 + C_g^2\right]\\ &+ 2\left[D_x\sqrt{P}L_g + C_g + 2\rho_1 P D_\lambda L_g^2\right]\\ &\times\sum_{i=1}^N\left(a_k\left\|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\right\|\right)\\ &+ 2N\sqrt{M}L_f G_{\mathcal{F}}(D_x + 2\rho_1 D_\lambda\sqrt{P}L_g)\\ &\times\sum_{i=1}^N\left(a_k\left\|\tilde{\boldsymbol{y}}_i^{(k)} - \hat{\boldsymbol{y}}^{(k-1)}\right\|\right)\\ &+ 2N\rho_2 D_x\sqrt{P}L_g\sum_{i=1}^N\left(a_k\left\|\tilde{\boldsymbol{z}}_i^{(k)} - \hat{\boldsymbol{z}}^{(k-1)}\right\|\right).\end{aligned} \tag{52}$$

First of all, by Theorem 1, we have

$$\mathcal{L}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \boldsymbol{\lambda}^\star\right) - \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star) \geq 0,\ \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star) - \mathcal{L}\left(\boldsymbol{x}^\star, \hat{\boldsymbol{\beta}}^{(k)}\right) \geq 0$$

implying that $\mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \boldsymbol{\lambda}^\star) - \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\beta}^{(k)}) \geq 0$. Hence we deduce from (52) that

$$\begin{aligned}&\left(\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star\right\|^2 + \sum_{i=1}^N\left\|\boldsymbol{\lambda}_i^{(k)} - \boldsymbol{\lambda}^\star\right\|^2\right)\\ &\quad\leq \left(\left\|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^\star\right\|^2 + \sum_{i=1}^N\left\|\boldsymbol{\lambda}_i^{(k-1)} - \boldsymbol{\lambda}^\star\right\|^2\right) + \tilde{c}_k\\ &\qquad- 2a_k\left(\mathcal{L}\left(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}\right) - \mathcal{L}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right)\right).\end{aligned} \tag{53}$$

Secondly, by $\sum_{k=1}^\infty a_k^2 < \infty$ and by Lemma 2, we see that all the four terms in $\tilde{c}_k$ are summable over k, and thus $\sum_{k=1}^\infty \tilde{c}_k < \infty$. Thirdly, by Lemma 5 and under the premise of $\rho_1 \leq 1/(G_{\bar{\mathcal{F}}} + D_\lambda\sqrt{P}G_g)$, we have $\mathcal{L}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) \geq 0$. Therefore, by applying Lemma 1 to (53), we conclude that the sequence $\{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star\|^2 + \sum_{i=1}^N\|\boldsymbol{\lambda}_i^{(k)} - \boldsymbol{\lambda}^\star\|^2\}$ converges for any saddle point $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$, and it holds that $\sum_{k=1}^\infty a_k(\mathcal{L}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)})) < \infty$. Because $\sum_{k=1}^\infty a_k = \infty$, the preceding relation implies that

$$\liminf_{k\to\infty}\mathcal{L}\left(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}\right) - \mathcal{L}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right) = 0. \tag{54}$$

Equation (54) implies that there exists a subsequence $\ell_1, \ell_2, \ldots$ such that

$$\mathcal{L}\left(\boldsymbol{x}^{(\ell_k-1)}, \hat{\boldsymbol{\beta}}^{(\ell_k)}\right) - \mathcal{L}\left(\hat{\boldsymbol{\alpha}}^{(\ell_k)}, \hat{\boldsymbol{\lambda}}^{(\ell_k-1)}\right) \to 0 \text{ as } k \to \infty. \tag{55}$$

According to Lemma 5, the above equation indicates that

$$\lim_{k\to\infty}\left\|\boldsymbol{x}^{(\ell_k-1)} - \hat{\boldsymbol{\alpha}}^{(\ell_k)}\right\| = 0,\ \lim_{k\to\infty}\left\|\hat{\boldsymbol{\lambda}}^{(\ell_k-1)} - \hat{\boldsymbol{\beta}}^{(\ell_k)}\right\| = 0. \tag{56}$$

Moreover, because $\{(\boldsymbol{x}^{(\ell_k-1)}, \hat{\boldsymbol{\lambda}}^{(\ell_k-1)})\} \subset \mathcal{X} \times \mathcal{D}$ is a bounded sequence, there must exist a limit point, say $(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star) \in \mathcal{X} \times \mathcal{D}$, such that

$$\boldsymbol{x}^{(\ell_k-1)} \to \hat{\boldsymbol{x}}^\star, \quad \hat{\boldsymbol{\lambda}}^{(\ell_k-1)} \to \hat{\boldsymbol{\lambda}}^\star, \text{ as } k \to \infty. \tag{57}$$

Under the premise of $\rho_1 \leq 1/(G_{\bar{\mathcal{F}}} + D_\lambda\sqrt{P}G_g)$, and by (55) and (57), we obtain from Lemma 5 that $(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star) \in \mathcal{X} \times \mathcal{D}$ is a saddle point of (15). Moreover, because

$$\left\|\boldsymbol{x}^{(\ell_k)} - \hat{\boldsymbol{x}}^\star\right\|^2 + \sum_{i=1}^N \left\|\boldsymbol{\lambda}_i^{(\ell_k)} - \hat{\boldsymbol{\lambda}}^\star\right\|^2 \leq \left\|\boldsymbol{x}^{(\ell_k)} - \hat{\boldsymbol{x}}^\star\right\|^2 + \sum_{i=1}^N \left(\left\|\boldsymbol{\lambda}_i^{(\ell_k)} - \hat{\boldsymbol{\lambda}}^{(\ell_k)}\right\| + \left\|\hat{\boldsymbol{\lambda}}^{(\ell_k)} - \hat{\boldsymbol{\lambda}}^\star\right\|\right)^2$$

we obtain from Lemma 2 and (57) that the sequence $\{\|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}^\star\|^2 + \sum_{i=1}^N \|\boldsymbol{\lambda}_i^{(k)} - \hat{\boldsymbol{\lambda}}^\star\|^2\}$ has a limit value equal to zero. Since the sequence $\{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star\|^2 + \sum_{i=1}^N \|\boldsymbol{\lambda}_i^{(k)} - \boldsymbol{\lambda}^\star\|^2\}$ converges for any saddle point of (15), we conclude that $\{\|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}^\star\|^2 + \sum_{i=1}^N \|\boldsymbol{\lambda}_i^{(k)} - \hat{\boldsymbol{\lambda}}^\star\|^2\}$ in fact converges to zero, and therefore (49) is proved. Finally, relation (50) can also be obtained by (49), (53) and (48), provided that $\rho_1 \leq 1/(G_{\bar{\mathcal{F}}} + D_\lambda\sqrt{P}G_g)$. ■

According to [44, Lemma 3], if $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^\star$ as $k \to \infty$, then its weighted running average $\bar{\boldsymbol{x}}^{(k)}$ defined in (36) also converges to $\boldsymbol{x}^\star$ as $k \to \infty$. What remains is to show the second fact that $(\hat{\boldsymbol{x}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k)})$ asymptotically satisfies the optimality conditions given by Proposition 1. We prove in Appendix C that the following lemma holds.

Lemma 7: *Under the assumptions of Lemma 6, it holds*

$$\lim_{k\to\infty} \left\| \left(\sum_{i=1}^N \boldsymbol{g}_i\left(\hat{\boldsymbol{x}}_i^{(k)}\right)\right)^+ \right\| = 0,$$

$$\lim_{k\to\infty} \left(\hat{\boldsymbol{\lambda}}^{(k)}\right)^T \left(\sum_{i=1}^N \boldsymbol{g}_i\left(\hat{\boldsymbol{x}}_i^{(k)}\right)\right) = 0. \tag{58}$$

By Lemma 6, Lemma 7, and Proposition 1, we conclude that Theorem 2 is true. Finally, we remark that when the step size a_k has the form of $a/(b+k)$ where $a > 0, b \geq 0$, one can simply consider the running average below [44]

$$\bar{\boldsymbol{x}}^{(k)} = \frac{1}{k}\sum_{\ell=0}^{k-1} \boldsymbol{x}^{(\ell)} = \left(1 - \frac{1}{k}\right)\bar{\boldsymbol{x}}^{(k-1)} + \frac{1}{k}\boldsymbol{x}^{(k-1)} \tag{59}$$

instead of the running weighted-average in (36) while Lemma 7 still holds true.

A) Proof of Theorem 3: Theorem 3 essentially can be obtained in the same line as the proof of Theorem 2, except for Lemma 5. What we need to show here is that the centralized proximal perturbation point $\hat{\boldsymbol{\alpha}}^{(k)}$ in (43) and $\hat{\boldsymbol{\beta}}^{(k)}$ in (42b) and the primal-dual iterates $(\boldsymbol{x}^{(k-1)}, \boldsymbol{\lambda}^{(k-1)})$ satisfy a result similar to Lemma 5. The lemma below is proved in Appendix D.

Lemma 8: *Let Assumptions 1 and 2 hold. For the centralized perturbation points $\hat{\boldsymbol{\alpha}}^{(k)}$ in (43) and $\hat{\boldsymbol{\beta}}^{(k)}$ in (42b), it holds true that*

$$\mathcal{L}\left(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}\right) - \mathcal{L}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right) \geq \left(\frac{1}{2\rho_1} - \frac{G_{\bar{\mathcal{F}}}}{2}\right) \times \left\|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\right\|^2 + \frac{1}{\rho_2}\left\|\hat{\boldsymbol{\lambda}}^{(k-1)} - \hat{\boldsymbol{\beta}}^{(k)}\right\|^2. \tag{60}$$

Moreover, let $\rho_1 \leq 1/G_{\bar{\mathcal{F}}}$, and let $\mathcal{L}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) \to 0$ and $(\boldsymbol{x}^{(k-1)}, \boldsymbol{\lambda}^{(k-1)}) \to (\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star)$ as $k \to \infty$, where $(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star) \in \mathcal{X} \times \mathcal{D}$. Then $(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star)$ is a saddle point of (15).

V. Simulation Results

In this section, we examine the efficacy of the proposed distributed PDP method (Algorithm 1) by considering the DSM problem discussed in Section II-B. We consider the DSM problem presented in (3) and (4). The cost functions were set to $C_{\rm p}(\cdot) = \pi_{\rm p}\|\cdot\|^2$ and $C_{\rm s}(\cdot) = \pi_{\rm s}\|\cdot\|^2$, respectively, where $\pi_{\rm p}$ and $\pi_{\rm s}$ are some price parameters. The load profile function $\boldsymbol{\psi}_i(\boldsymbol{x}_i)$ is based on the load model in [18], which were proposed to model deferrable, non-interruptible loads such as electrical vehicle, washing machine and tumble dryer et. al. According to [18], $\boldsymbol{\psi}_i(\boldsymbol{x}_i)$ can be modeled as a linear function, i.e., $\boldsymbol{\psi}_i(\boldsymbol{x}_i) = \boldsymbol{\Psi}_i\boldsymbol{x}_i$, where $\boldsymbol{\Psi}_i \in \mathbb{R}^{T\times T}$ is a coefficient matrix composed of load profiles of appliances of customer i. The control variable $\boldsymbol{x}_i \in \mathbb{R}^T$ determines the operation scheduling of appliances of customer i. Due to some physical conditions and quality of service constraints, each $\boldsymbol{x}_i$ is subject to a local constraint set $\mathcal{X}_i = \{\boldsymbol{x}_i \in \mathbb{R}^T \mid \boldsymbol{A}_i\boldsymbol{d}_i \preceq \boldsymbol{b}_i, \boldsymbol{l}_i \leq \boldsymbol{d}_i \leq \boldsymbol{u}_i\}$ where $\boldsymbol{A}_i \in \mathbb{R}^{T\times T}$ and $\boldsymbol{l}_i, \boldsymbol{u}_i \in \mathbb{R}^T$ [18]. The problem formulation corresponding to (3) is thus given by

$$\min_{\substack{\boldsymbol{x}_i\in\mathcal{X}_i,\\ i=1,\ldots,N}} \pi_{\rm p}\left\|\left(\sum_{i=1}^N \boldsymbol{\Psi}_i\boldsymbol{x}_i - \boldsymbol{p}\right)^+\right\|^2 + \pi_{\rm s}\left\|\left(\boldsymbol{p} - \sum_{i=1}^N \boldsymbol{\Psi}_i\boldsymbol{x}_i\right)^+\right\|^2. \tag{61}$$

Analogous to (4), problem (61) can be reformulated as

$$\min_{\substack{\boldsymbol{x}_i\in\mathcal{X}_i, i=1,\ldots,N,\\ \boldsymbol{z}\succeq\boldsymbol{0}}} \pi_{\rm p}\|\boldsymbol{z}\|^2 + \pi_{\rm s}\left\|\boldsymbol{z} - \sum_{i=1}^N \boldsymbol{\Psi}_i\boldsymbol{x}_i + \boldsymbol{p}\right\|^2 \tag{62a}$$

$$\text{s.t.} \sum_{i=1}^N \boldsymbol{\Psi}_i\boldsymbol{x}_i - \boldsymbol{p} - \boldsymbol{z} \preceq \boldsymbol{0} \tag{62b}$$

to which the proposed distributed PDP method can be applied. We consider a scenario with 400 customers $(N = 400)$, and follow the same methods as in [47] to generate the power bidding $\boldsymbol{p}$ and coefficients $\boldsymbol{\Psi}_i$, $\boldsymbol{A}_i$, $\boldsymbol{b}_i$, $\boldsymbol{l}_i$, $\boldsymbol{u}_i$, $i = 1, \ldots, N$. The network graph $\mathcal{G}$ was randomly generated. The price parameters $\pi_{\rm p}$ and $\pi_{\rm s}$ were simply set to $1/N$ and $0.8/N$, respectively. In addition to the distributed PD method in [15], we also compare the proposed distributed PDP method with the distributed dual subgradient (DDS) method[3] [18], [25]. This method is based on the same idea as the dual decomposition technique [25], where, given the dual variables, each customer globally solves the corresponding inner minimization problem. The average consensus subgradient technique [10] is applied to the dual domain for distributed dual optimization.

[3]One can utilize the linear structure to show that (61) is equivalent to the following saddle point problem (by Lagrange dual)

$$\max_{\substack{\boldsymbol{\lambda}\succeq\boldsymbol{0},\\ \boldsymbol{\eta}\succeq\boldsymbol{0}}} \left\{ \min_{\substack{\boldsymbol{x}_i\in\mathcal{X}_i\\ i=1,\ldots,N}} -\frac{1}{4\pi_{\rm p}}\|\boldsymbol{\lambda}\|^2 - \frac{1}{4\pi_{\rm s}}\|\boldsymbol{\eta}\|^2 + (\boldsymbol{\lambda}-\boldsymbol{\eta})^T\left(\sum_{i=1}^N \boldsymbol{\Psi}_i\boldsymbol{x}_i - \boldsymbol{p}\right)\right\} \tag{63}$$

to which the method in [15] and the DDS method [25] can be applied.

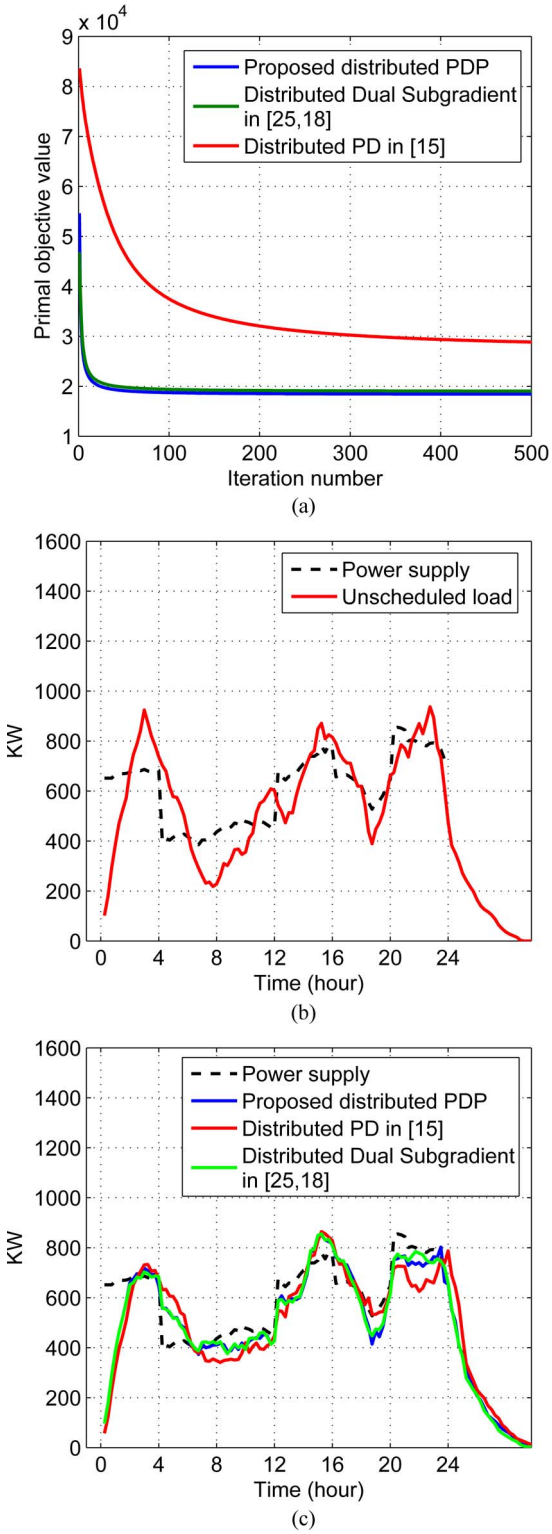


Fig. 1. Numerical results for the smart grid DSM problem (61) with 400 customers. (a) Convergence curve; (b) Unscheduled load profile; (c) Scheduled load profiles.

Fig. 1(a) shows the convergence curves of the three methods under test. The curves shown in this figure are the corresponding objective values in (61) of the running average iterates of the three methods. The step size of the distributed PD method in [15] was set to $a_k = 15/10 + k$ and that of the DDS method was set to $a_k = 0.05/10 + k$. For the proposed distributed PDP method, a_k, ρ_1 and ρ_2 were respectively set to $a_k = 0.1/10 + k$ and $\rho_1 = \rho_2 = 0.001$. From this figure, we observe that the proposed distributed PDP method and the DDS method exhibit comparable convergence behavior; both methods converge within 100 iterations and outperform the distributed PD method in [15]. One should note that the DDS method is computational more expensive than the proposed distributed PDP method since, in each iteration, the former requires to globally solve the inner minimization problem while the latter takes two primal gradient updates only. For the proposed PDP Algorithm 1, the complexity order per iteration per customer is given by $\mathcal{O}(4T)$ [see (19), (21) and (22)]. For the DDS method, each customer has to solve the inner linear programming (LP) in (63) $\min_{\boldsymbol{x}_i \in \mathcal{X}_i} (\boldsymbol{\lambda} - \boldsymbol{\eta})^T \boldsymbol{\Psi}_i \boldsymbol{x}_i$ per iteration. According to [48], the worst-case complexity of interior point methods for solving an LP is given by $\mathcal{O}(T^{0.5}(3T^2 + T^3)) \approx \mathcal{O}(T^{3.5})$.

In Fig. 1(b), we display the load profiles of the power supply and the unscheduled load (without DSM), while, in Fig. 1(c), we show the load profiles scheduled by the three optimization methods under consideration. The results were obtained by respectively combining each of the optimization method with the certainty equivalent control (CEC) approach in [18, Algorithm 1] to handle a stochastic counterpart of problem (61). The stopping criterion was set to the maximum iteration number of 500. We can observe from this figure that, for all the three methods, the power balancing can be much improved compared to that without DSM control. However, we still can observe from Fig. 1(c) that the proposed PDP method and the DDS method exhibit better results than the distributed PD method in [15]. Specifically, the cost in (61) is 4.49×10^4 KW for the unscheduled load whereas that of the load scheduled by the proposed distributed PDP method is 2.44×10^4 KW (45.65% reduction). The cost for the load scheduled by the distributed DDS method is slightly lower which is 2.38×10^4 KW; whereas that scheduled by the distributed PD method in [15] has a higher cost of 3.81×10^4 KW.

As discussed in Section II-B, problem (2) also incorporates the important regression problems. In [36], we have applied the proposed PDP method to solving a distributed sparse regression problem (with a non-smooth constraint function). The simulation results can be found in [36].

VI. Conclusion

We have presented a distributed consensus-based PDP algorithm for solving problem of the form (2), which has a globally coupled cost function and inequality constraints. The algorithm employs the average consensus technique and the primal-dual perturbed (sub-) gradient method. We have provided a convergence analysis showing that the proposed algorithm enables the agents across the network to achieve a global optimal primal-dual solution of the considered problem in a distributed manner. The effectiveness of the proposed algorithm has been demonstrated by applying it to a smart grid demand response control problem and a sparse linear regression problem [36]. In particular, the proposed algorithm is shown to have better convergence property than the distributed PD method in [15] which does not have perturbation. In addition, the proposed

algorithm performs comparably with the distributed dual subgradient method [25] for the demand response control problem, even though the former is computationally cheaper.

APPENDIX A
PROOF OF LEMMA 3

We first show (45). By definitions in (42b) and (19b), and by the non-expansiveness of projection, we readily obtain

$$\begin{aligned}&\left\|\hat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}_i^{(k)}\right\|\\&\quad\le\left\|\mathcal{P}_{\mathcal{D}}\left(\tilde{\boldsymbol{\lambda}}_i^{(k)}+\rho_2 N\tilde{\boldsymbol{z}}_i^{(k)}\right)-\mathcal{P}_{\mathcal{D}}\left(\hat{\boldsymbol{\lambda}}^{(k-1)}+\rho_2 N\hat{\boldsymbol{z}}^{(k-1)}\right)\right\|\\&\quad\le\left\|\tilde{\boldsymbol{\lambda}}_i^{(k)}-\hat{\boldsymbol{\lambda}}^{(k-1)}\right\|+\rho_2 N\left\|\tilde{\boldsymbol{z}}_i^{(k)}-\hat{\boldsymbol{z}}^{(k-1)}\right\|.\end{aligned}$$

Equation (44) for the $\boldsymbol{\alpha}_i^{(k)}$ in (19a) and $\hat{\boldsymbol{\alpha}}_i^{(k)}$ in (42a) can be shown in a similar line, as shown in the following:

$$\begin{aligned}&\left\|\boldsymbol{\alpha}_i^{(k)}-\hat{\boldsymbol{\alpha}}_i^{(k)}\right\|\\&\quad=\Big\|\mathcal{P}_{\mathcal{X}_i}\Big(\boldsymbol{x}_i^{(k-1)}-\rho_1\Big[\nabla\boldsymbol{f}_i^T\left(\boldsymbol{x}_i^{(k-1)}\right)\nabla\mathcal{F}\left(N\tilde{\boldsymbol{y}}_i^{(k)}\right)\\&\qquad\qquad+\nabla\boldsymbol{g}_i^T\left(\boldsymbol{x}_i^{(k-1)}\right)\tilde{\boldsymbol{\lambda}}_i^{(k)}\Big]\Big)\\&\qquad-\mathcal{P}_{\mathcal{X}_i}\Big(\boldsymbol{x}_i^{(k-1)}-\rho_1\Big[\nabla\boldsymbol{f}_i^T\left(\boldsymbol{x}_i^{(k-1)}\right)\nabla\mathcal{F}\left(N\hat{\boldsymbol{y}}^{(k-1)}\right)\\&\qquad\qquad+\nabla\boldsymbol{g}_i^T\left(\boldsymbol{x}_i^{(k-1)}\right)\hat{\boldsymbol{\lambda}}^{(k-1)}\Big]\Big)\Big\|\\&\quad\le\rho_1\left\|\nabla\boldsymbol{f}_i^T\Big(\boldsymbol{x}_i^{(k-1)}\Big)\right\|\left\|\nabla\mathcal{F}\Big(N\tilde{\boldsymbol{y}}_i^{(k)}\Big)-\nabla\mathcal{F}\left(N\hat{\boldsymbol{y}}^{(k-1)}\right)\right\|\\&\qquad+\rho_1\left\|\nabla\boldsymbol{g}_i\left(\boldsymbol{x}_i^{(k-1)}\right)\right\|\left\|\tilde{\boldsymbol{\lambda}}_i^{(k)}-\hat{\boldsymbol{\lambda}}^{(k-1)}\right\|\\&\quad\le\rho_1 L_g\sqrt{P}\left\|\tilde{\boldsymbol{\lambda}}_i^{(k)}-\hat{\boldsymbol{\lambda}}^{(k-1)}\right\|\\&\qquad+\rho_1 G_{\mathcal{F}}L_f\sqrt{M}N\left\|\tilde{\boldsymbol{y}}_i^{(k)}-\hat{\boldsymbol{y}}^{(k-1)}\right\|\end{aligned}\tag{A1}$$

where, in the second inequality, we have used the boundedness of gradients (cf. (26), (28)) and the Lipschitz continuity of $\nabla\mathcal{F}$ (Assumption 2).

To show that (44) holds for $\boldsymbol{\alpha}_i^{(k)}$ in (20) and $\hat{\boldsymbol{\alpha}}_i^{(k)}$ in (43), we use the following lemma:

Lemma 9 [49, Lemma 4.1]: *If $\boldsymbol{y}^\star=\arg\min_{\boldsymbol{y}\in\mathcal{Y}}J_1(\boldsymbol{y})+J_2(\boldsymbol{y})$, where $J_1:\mathbb{R}^n\to\mathbb{R}$ and $J_2:\mathbb{R}^n\to\mathbb{R}$ are convex functions and $\mathcal{Y}$ is a closed convex set. Moreover, J_2 is continuously differentiable. Then $\boldsymbol{y}^\star=\arg\min_{\boldsymbol{y}\in\mathcal{Y}}\{J_1(\boldsymbol{y})+\nabla J_2^T(\boldsymbol{y}^\star)\boldsymbol{y}\}$.*

By applying the above lemma to (20) using $J_1(\boldsymbol{\alpha}_1)=\boldsymbol{g}_i^T(\boldsymbol{\alpha}_i)\tilde{\boldsymbol{\lambda}}_i^{(k)}$ and

$$J_2(\boldsymbol{\alpha}_i)=\frac{1}{2\rho_1}\left\|\boldsymbol{\alpha}_i-\Big(\boldsymbol{x}_i^{(k-1)}-\rho_1\nabla\boldsymbol{f}_i^T\Big(\boldsymbol{x}_i^{(k-1)}\Big)\nabla\mathcal{F}\Big(N\tilde{\boldsymbol{y}}_i^{(k)}\Big)\Big)\right\|^2$$

we obtain

$$\begin{aligned}\boldsymbol{\alpha}_i^{(k)}=\arg\min_{\boldsymbol{\alpha}_i\in\mathcal{X}_i}\boldsymbol{g}_i^T(\boldsymbol{\alpha}_i)\tilde{\boldsymbol{\lambda}}_i^{(k)}+\Big(\nabla\boldsymbol{f}_i^T\left(\boldsymbol{x}_i^{(k-1)}\right)\\\times\nabla\mathcal{F}\left(N\tilde{\boldsymbol{y}}_i^{(k)}\right)+\frac{1}{\rho_1}\left(\boldsymbol{\alpha}_i^{(k)}-\boldsymbol{x}_i^{(k-1)}\right)\Big)^T\boldsymbol{\alpha}_i.\end{aligned}\tag{A2}$$

Similarly, applying Lemma 9 to (43), we obtain

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_i^{(k)}=\arg\min_{\boldsymbol{\alpha}_i\in\mathcal{X}_i}\boldsymbol{g}_i^T(\boldsymbol{\alpha}_i)\hat{\boldsymbol{\lambda}}^{(k-1)}+\Big(\nabla\boldsymbol{f}_i^T\left(\boldsymbol{x}_i^{(k-1)}\right)\\\times\nabla\mathcal{F}\left(N\hat{\boldsymbol{y}}^{(k-1)}\right)+\frac{1}{\rho_1}\left(\hat{\boldsymbol{\alpha}}_i^{(k)}-\boldsymbol{x}_i^{(k-1)}\right)\Big)^T\boldsymbol{\alpha}_i.\end{aligned}\tag{A3}$$

From (A2) it follows that:

$$\begin{aligned}&\boldsymbol{g}_i^T\left(\boldsymbol{\alpha}_i^{(k)}\right)\tilde{\boldsymbol{\lambda}}_i^{(k)}\\&\quad+\left(\nabla\boldsymbol{f}_i^T\Big(\boldsymbol{x}_i^{(k-1)}\Big)\nabla\mathcal{F}\Big(N\tilde{\boldsymbol{y}}_i^{(k)}\Big)+\frac{1}{\rho_1}\Big(\boldsymbol{\alpha}_i^{(k)}-\boldsymbol{x}_i^{(k-1)}\Big)\right)^T\boldsymbol{\alpha}_i^{(k)}\\&\quad\le\boldsymbol{g}_i^T\left(\hat{\boldsymbol{\alpha}}_i^{(k)}\right)\tilde{\boldsymbol{\lambda}}_i^{(k)}+\Big(\nabla\boldsymbol{f}_i^T\left(\boldsymbol{x}_i^{(k-1)}\right)\nabla\mathcal{F}\left(N\tilde{\boldsymbol{y}}_i^{(k)}\right)\\&\qquad+\frac{1}{\rho_1}\left(\boldsymbol{\alpha}_i^{(k)}-\boldsymbol{x}_i^{(k-1)}\right)\Big)^T\hat{\boldsymbol{\alpha}}_i^{(k)}\end{aligned}$$

which is equivalent to

$$\begin{aligned}0\le&\Big(\boldsymbol{g}_i^T\Big(\hat{\boldsymbol{\alpha}}_i^{(k)}\Big)-\boldsymbol{g}_i^T\Big(\boldsymbol{\alpha}_i^{(k)}\Big)\Big)\tilde{\boldsymbol{\lambda}}_i^{(k)}+\nabla\boldsymbol{f}_i^T\Big(\boldsymbol{x}_i^{(k-1)}\Big)\nabla\mathcal{F}\Big(N\tilde{\boldsymbol{y}}_i^{(k)}\Big)\\&\times\Big(\hat{\boldsymbol{\alpha}}_i^{(k)}-\boldsymbol{\alpha}_i^{(k)}\Big)+\frac{1}{\rho_1}\left(\boldsymbol{\alpha}_i^{(k)}-\boldsymbol{x}_i^{(k-1)}\right)\left(\hat{\boldsymbol{\alpha}}_i^{(k)}-\boldsymbol{\alpha}_i^{(k)}\right).\end{aligned}\tag{A4}$$

Similarly, (A3) implies that

$$\begin{aligned}0\le&\left(\boldsymbol{g}_i^T\left(\boldsymbol{\alpha}_i^{(k)}\right)-\boldsymbol{g}_i^T\left(\hat{\boldsymbol{\alpha}}_i^{(k)}\right)\right)\hat{\boldsymbol{\lambda}}^{(k-1)}\\&+\nabla\boldsymbol{f}_i^T\left(\boldsymbol{x}_i^{(k-1)}\right)\nabla\mathcal{F}\left(N\hat{\boldsymbol{y}}^{(k-1)}\right)\left(\boldsymbol{\alpha}_i^{(k)}-\hat{\boldsymbol{\alpha}}_i^{(k)}\right)\\&+\frac{1}{\rho_1}\left(\hat{\boldsymbol{\alpha}}_i^{(k)}-\boldsymbol{x}_i^{(k-1)}\right)\left(\boldsymbol{\alpha}_i^{(k)}-\hat{\boldsymbol{\alpha}}_i^{(k)}\right).\end{aligned}\tag{A5}$$

By combining (A4) and (A5), we obtain

$$\begin{aligned}&\frac{1}{\rho_1}\left\|\hat{\boldsymbol{\alpha}}_i^{(k)}-\boldsymbol{\alpha}_i^{(k)}\right\|^2\\&\quad\le\left(\boldsymbol{g}_i^T\left(\hat{\boldsymbol{\alpha}}_i^{(k)}\right)-\boldsymbol{g}_i^T\left(\boldsymbol{\alpha}_i^{(k)}\right)\right)\left(\tilde{\boldsymbol{\lambda}}_i^{(k)}-\hat{\boldsymbol{\lambda}}^{(k-1)}\right)\\&\qquad+\nabla\boldsymbol{f}_i^T\left(\boldsymbol{x}_i^{(k-1)}\right)\Big(\nabla\mathcal{F}\left(N\tilde{\boldsymbol{y}}_i^{(k)}\right)\\&\qquad\qquad-\nabla\mathcal{F}\left(N\hat{\boldsymbol{y}}^{(k-1)}\right)\Big)\left(\hat{\boldsymbol{\alpha}}_i^{(k)}-\boldsymbol{\alpha}_i^{(k)}\right)\\&\quad\le\Big(\sqrt{P}L_g\left\|\tilde{\boldsymbol{\lambda}}_i^{(k)}-\hat{\boldsymbol{\lambda}}^{(k-1)}\right\|\\&\qquad+G_{\mathcal{F}}L_f\sqrt{M}N\left\|\tilde{\boldsymbol{y}}_i^{(k)}-\hat{\boldsymbol{y}}^{(k-1)}\right\|\Big)\left\|\hat{\boldsymbol{\alpha}}_i^{(k)}-\boldsymbol{\alpha}_i^{(k)}\right\|\end{aligned}$$

where we have used the boundedness of gradients [cf. (26), (28)], the Lipschitz continuity of $\nabla\mathcal{F}$ (Assumption 2) as well as the Lipschitz continuity of $\boldsymbol{g}_i$ [in (29)]. The desired result in (44) follows from the preceding relation. ■

APPENDIX B
PROOF OF LEMMA 5

We first prove that relation (48) holds for the perturbation points $\hat{\boldsymbol{\alpha}}_i^{(k)}$ and $\hat{\boldsymbol{\beta}}^{(k)}$ in (42) assuming that Assumption 3 is satisfied. Note that (42a) is equivalent to

$$\hat{\boldsymbol{\alpha}}_i^{(k)}=\arg\min_{\boldsymbol{\alpha}_i\in\mathcal{X}_i}\left\|\boldsymbol{\alpha}_i-\boldsymbol{x}_i^{(k-1)}+\rho_1\mathcal{L}_{\boldsymbol{x}_i}\left(\boldsymbol{x}^{(k-1)},\hat{\boldsymbol{\lambda}}^{(k-1)}\right)\right\|^2$$

where $\mathcal{L}_{\boldsymbol{x}_i}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) = \nabla \boldsymbol{f}_i^T(\boldsymbol{x}_i^{(k-1)})\nabla\mathcal{F}(N\hat{\boldsymbol{y}}^{(k)}) + \nabla \boldsymbol{g}_i^T(\boldsymbol{x}_i^{(k-1)})\hat{\boldsymbol{\lambda}}^{(k-1)}$. By the optimality condition, we have that, for all $\boldsymbol{x}_i \in \mathcal{X}_i$

$$\left(\boldsymbol{x}_i - \hat{\boldsymbol{\alpha}}_i^{(k)}\right)^T \left(\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{x}_i^{(k-1)} + \rho_1 \mathcal{L}_{\boldsymbol{x}_i}\left(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right)\right) \geq 0.$$

By choosing $\boldsymbol{x}_i = \boldsymbol{x}_i^{(k-1)}$, one obtains

$$\left(\boldsymbol{x}_i^{(k-1)} - \hat{\boldsymbol{\alpha}}_i^{(k)}\right)^T \mathcal{L}_{\boldsymbol{x}_i}\left(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right) \geq \frac{1}{\rho_1}\left\|\boldsymbol{x}_i^{(k-1)} - \hat{\boldsymbol{\alpha}}_i^{(k)}\right\|^2$$

which, by summing over $i = 1, \ldots, N$, gives rise to

$$\left(\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\right)^T \mathcal{L}_{\boldsymbol{x}}\left(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right) \geq \frac{1}{\rho_1}\left\|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\right\|^2.$$

Further write the above equation as follows:

$$\begin{aligned}
&\left(\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\right)^T \mathcal{L}_{\boldsymbol{x}}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right) \\
&\quad \geq \frac{1}{\rho_1}\left\|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\right\|^2 - \left(\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\right)^T \\
&\qquad \times \left(\mathcal{L}_{\boldsymbol{x}}\left(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right) - \mathcal{L}_{\boldsymbol{x}}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right)\right) \\
&\quad \geq \frac{1}{\rho_1}\left\|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\right\|^2 - \|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\| \\
&\qquad \times \left\|\mathcal{L}_{\boldsymbol{x}}\left(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right) - \mathcal{L}_{\boldsymbol{x}}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right)\right\|. \quad \text{(A6)}
\end{aligned}$$

By (8), Assumption 2, Assumption 3 and the boundedness of $\hat{\boldsymbol{\lambda}}^{(k-1)} \in \mathcal{D}$, we can bound the second term in (A6) as

$$\begin{aligned}
&\left\|\mathcal{L}_{\boldsymbol{x}}\left(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right) - \mathcal{L}_{\boldsymbol{x}}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right)\right\| \\
&\quad \leq \left\|\nabla\bar{\mathcal{F}}\left(\boldsymbol{x}^{(k-1)}\right) - \nabla\bar{\mathcal{F}}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\right\| \\
&\qquad + \left\|\hat{\boldsymbol{\lambda}}^{(k-1)}\right\| \left\|\begin{bmatrix} \nabla \boldsymbol{g}_1^T\left(\boldsymbol{x}_1^{(k-1)}\right) - \nabla \boldsymbol{g}_1^T\left(\hat{\boldsymbol{\alpha}}_1^{(k)}\right) \\ \vdots \\ \nabla \boldsymbol{g}_N^T\left(\boldsymbol{x}_N^{(k-1)}\right) - \nabla \boldsymbol{g}_N^T\left(\hat{\boldsymbol{\alpha}}_N^{(k)}\right) \end{bmatrix}\right\|_F \\
&\quad \leq (G_{\bar{\mathcal{F}}} + D_\lambda\sqrt{P}G_g)\left\|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\right\| \quad \text{(A7)}
\end{aligned}$$

where $\|\cdot\|_f$ denotes the Frobenious norm. By combining (A6) and (A7), we obtain

$$\begin{aligned}
&\left(\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\right)^T \mathcal{L}_{\boldsymbol{x}}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right) \\
&\quad \geq \left(\frac{1}{\rho_1} - (G_{\bar{\mathcal{F}}} + D_\lambda\sqrt{P}G_g)\right)\left\|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\right\|^2. \quad \text{(A8)}
\end{aligned}$$

Since $\mathcal{L}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) \geq (\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)})^T \mathcal{L}_{\boldsymbol{x}}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)})$ by the convexity of $\mathcal{L}$ in $\boldsymbol{x}$, we further obtain

$$\begin{aligned}
&\mathcal{L}\left(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right) - \mathcal{L}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right) \\
&\quad \geq \left(\frac{1}{\rho_1} - (G_{\bar{\mathcal{F}}} + D_\lambda\sqrt{P}G_g)\right)\left\|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\right\|^2. \quad \text{(A9)}
\end{aligned}$$

On the other hand, by (42b), we know that $\hat{\boldsymbol{\beta}}^{(k)} = \arg\min_{\beta\in\mathcal{D}} \|\boldsymbol{\beta} - \hat{\boldsymbol{\lambda}}^{(k-1)} - \rho_2 \sum_{i=1}^N \boldsymbol{g}_i(\boldsymbol{x}_i^{(k-1)})\|^2$. By the optimality condition and the linearity of $\mathcal{L}$ in $\boldsymbol{\lambda}$, we have

$$\begin{aligned}
&\mathcal{L}\left(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}\right) - \mathcal{L}\left(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}\right) \\
&\quad = -\left(\hat{\boldsymbol{\lambda}}^{(k-1)} - \hat{\boldsymbol{\beta}}^{(k)}\right)^T \times \left(\sum_{i=1}^N \boldsymbol{g}_i\left(\boldsymbol{x}_i^{(k-1)}\right)\right) \\
&\quad \geq \frac{1}{\rho_2}\left\|\hat{\boldsymbol{\lambda}}^{(k-1)} - \hat{\boldsymbol{\beta}}^{(k)}\right\|^2. \quad \text{(A10)}
\end{aligned}$$

Combining (A9) and (A10) yields (48).

Suppose that $\mathcal{L}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) \to 0$ and $(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)})$ converges to some limit point $(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star)$ as $k \to \infty$. Since $\rho_1 \leq 1/(G_{\bar{\mathcal{F}}} + D_\lambda\sqrt{P}G_g)$, we infer from (48) that $\|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\| \to 0$ and $\|\hat{\boldsymbol{\lambda}}^{(k-1)} - \hat{\boldsymbol{\beta}}^{(k)}\| \to 0$, as $k \to \infty$. It then follows from (42) and the fact that projection is a continuous mapping [49] that $(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star) \in \mathcal{X} \times \mathcal{D}$ satisfies

$$\begin{aligned}
\hat{\boldsymbol{x}}_i^\star &= \mathcal{P}_{\mathcal{X}_i}\left(\hat{\boldsymbol{x}}_i^\star - \rho_1\left[\nabla \boldsymbol{f}_i^T(\hat{\boldsymbol{x}}_i^\star)\nabla\mathcal{F}\left(\sum_{i=1}^M \boldsymbol{f}_i(\hat{\boldsymbol{x}}_i^\star)\right)\right.\right. \\
&\qquad \left.\left. + \nabla \boldsymbol{g}_i^T(\hat{\boldsymbol{x}}_i^\star)\hat{\boldsymbol{\lambda}}^\star\right]\right), \ i = 1, \ldots, N, \\
\hat{\boldsymbol{\lambda}}^\star &= \mathcal{P}_{\mathcal{D}}\left(\hat{\boldsymbol{\lambda}}^\star + \rho_2 \sum_{i=1}^N \boldsymbol{g}_i(\hat{\boldsymbol{x}}_i^\star)\right)
\end{aligned}$$

which, respectively, imply that $\hat{\boldsymbol{x}}^\star = \arg\min_{\boldsymbol{x}\in\mathcal{X}} \mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{\lambda}}^\star)$ and $\hat{\boldsymbol{\lambda}}^\star = \arg\max_{\boldsymbol{\lambda}\in\mathcal{D}} \mathcal{L}(\hat{\boldsymbol{x}}^\star, \boldsymbol{\lambda})$ i.e., $(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star)$ is a saddle point of problem (15). ■

APPENDIX C
PROOF OF LEMMA 7

By (47) in Lemma 4 and the fact of $\mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \boldsymbol{\lambda}) = \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) + (\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^{(k-1)})^T \mathcal{L}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)})$, we have

$$\begin{aligned}
&\left(\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^{(k-1)}\right)^T \boldsymbol{g}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right) \leq \frac{\bar{c}_k}{2a_k} \\
&\quad + \frac{1}{2a_k}\left(\sum_{j=1}^N \left\|\boldsymbol{\lambda}_j^{(k-1)} - \boldsymbol{\lambda}\right\|^2 - \sum_{i=1}^N \left\|\boldsymbol{\lambda}_i^{(k)} - \boldsymbol{\lambda}\right\|^2\right) \quad \text{(A11)}
\end{aligned}$$

where $\boldsymbol{g}(\hat{\boldsymbol{\alpha}}^{(k)}) = \sum_{i=1}^N \boldsymbol{g}_i(\hat{\boldsymbol{\alpha}}_i^{(k)})$ and

$$\begin{aligned}
\bar{c}_k \triangleq\ & a_k^2 N C_g^2 + 2a_k\left(2\rho_1 D_\lambda P L_g^2 + C_g\right)\left\|\tilde{\boldsymbol{\lambda}}_i^{(k)} - \hat{\boldsymbol{\lambda}}^{(k-1)}\right\| \\
& + 4\rho_1 N D_\lambda G_{\mathcal{F}}\sqrt{PM} L_g L_f a_k \left\|\tilde{\boldsymbol{y}}_i^{(k)} - \hat{\boldsymbol{y}}_i^{(k-1)}\right\|.
\end{aligned}$$

By following a similar argument as in [27, Proposition 5.1] and by (A11), (16), (29), and (30), one can show that

$$\begin{aligned}
&\left(\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^\star\right)^T \boldsymbol{g}\left(\boldsymbol{x}^{(k-1)}\right) \\
&\leq \frac{\bar{c}_k}{2a_k} + \frac{1}{2a_k}\left(\sum_{j=1}^N \left\|\boldsymbol{\lambda}_j^{(k-1)} - \boldsymbol{\lambda}\right\|^2 - \sum_{i=1}^N \left\|\boldsymbol{\lambda}_i^{(k)} - \boldsymbol{\lambda}\right\|^2\right) \\
&\quad + 2N\sqrt{P}D_\lambda L_g \left\|\boldsymbol{x}^{(k-1)} - \hat{\boldsymbol{\alpha}}^{(k)}\right\| + NC_g \left\|\hat{\boldsymbol{\lambda}}^{(k-1)} - \hat{\boldsymbol{\lambda}}^\star\right\|. \quad \text{(A12)}
\end{aligned}$$

By taking the weighted running average of (A12), we obtain

$$\begin{aligned}
&(\boldsymbol{\lambda}-\hat{\boldsymbol{\lambda}}^{\star})^{T}\boldsymbol{g}\left(\hat{\boldsymbol{x}}^{(k-1)}\right)\\
&\leq \frac{1}{A_k}\sum_{\ell=1}^{k}a_\ell(\boldsymbol{\lambda}-\hat{\boldsymbol{\lambda}}^{\star})^{T}\boldsymbol{g}\left(\boldsymbol{x}^{(\ell-1)}\right)\\
&\leq \frac{1}{2A_k}\sum_{\ell=1}^{k}\bar{c}_\ell+\frac{1}{2A_k}\\
&\quad\times\left(\sum_{j=1}^{N}\left\|\boldsymbol{\lambda}_j^{(0)}-\boldsymbol{\lambda}\right\|^2-\sum_{i=1}^{N}\left\|\boldsymbol{\lambda}_i^{(k)}-\boldsymbol{\lambda}\right\|^2\right)\\
&\quad+\frac{2N\sqrt{P}D_\lambda L_g}{A_k}\sum_{\ell=1}^{k}a_\ell\left\|\boldsymbol{x}^{(\ell-1)}-\hat{\boldsymbol{\alpha}}^{(\ell)}\right\|\\
&\quad+\frac{NC_g}{A_k}\sum_{\ell=1}^{k}a_\ell\left\|\hat{\boldsymbol{\lambda}}^{(\ell-1)}-\hat{\boldsymbol{\lambda}}^{\star}\right\|\\
&\leq \frac{1}{2A_k}\sum_{\ell=1}^{k}\bar{c}_\ell+\frac{2ND_\lambda^2}{A_k}\\
&\quad+\frac{2N\sqrt{P}D_\lambda L_g}{A_k}\sum_{\ell=1}^{k}a_\ell\left\|\boldsymbol{x}^{(\ell-1)}-\hat{\boldsymbol{\alpha}}^{(\ell)}\right\|\\
&\quad+\frac{NC_g}{A_k}\sum_{\ell=1}^{k}a_\ell\left\|\hat{\boldsymbol{\lambda}}^{(\ell-1)}-\hat{\boldsymbol{\lambda}}^{\star}\right\| \qquad \text{(A13)}
\end{aligned}$$

where the first inequality is owing to the fact that $\boldsymbol{g}(\boldsymbol{x})$ is convex, and the last inequality is obtained by dropping $-\sum_{i=1}^{N}\|\boldsymbol{\lambda}_i^{(k)}-\boldsymbol{\lambda}\|^2$ followed by applying (16). We claim that

$$\lim_{k\to\infty}\xi^{(k-1)}=0. \qquad \text{(A14)}$$

To see this, note that the first and second terms in $\xi^{(k-1)}$ converge to zero as $k\to\infty$ since $\lim_{k\to\infty}A_k=\infty$ and $\sum_{\ell=1}^{\infty}\bar{c}_\ell<\infty$. The term $(1/A_k)\sum_{\ell=1}^{k}a_\ell\|\hat{\boldsymbol{\lambda}}^{(\ell-1)}-\hat{\boldsymbol{\lambda}}^{\star}\|$ also converges to zero since, by Lemma 6, $\lim_{k\to\infty}\|\hat{\boldsymbol{\lambda}}^{(k)}-\hat{\boldsymbol{\lambda}}^{\star}\|=0$ and so does its weighted running average by [44, Lemma 3]. Similarly, the term $(1/A_k)\sum_{\ell=1}^{k}a_\ell\|\boldsymbol{x}^{(\ell-1)}-\hat{\boldsymbol{\alpha}}^{(\ell)}\|$ also converges to zero since $\lim_{k\to\infty}\|\boldsymbol{x}^{(k-1)}-\hat{\boldsymbol{\alpha}}^{(k)}\|=0$ due to (50).

Now let $\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}^{\star}+\delta((\boldsymbol{g}(\hat{\boldsymbol{x}}^{(k-1)}))^{+}/\|(\boldsymbol{g}(\hat{\boldsymbol{x}}^{(k-1)}))^{+}\|)$ which lies in $\mathcal{D}$, since $\|\boldsymbol{\lambda}\|\leq\|\hat{\boldsymbol{\lambda}}^{\star}\|+\delta\leq D_\lambda$ by (17). Substituting $\boldsymbol{\lambda}$ into (A13) gives rise to

$$\delta\left\|\left(\boldsymbol{g}\left(\hat{\boldsymbol{x}}^{(k-1)}\right)\right)^{+}\right\|\leq\xi^{(k-1)}. \qquad \text{(A15)}$$

As a result, the first term in (58) is obtained by taking $k\to\infty$ in (A15) and by (A14).

To show that the second limit in (58) holds true, we first let $\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}^{\star}+\delta(\hat{\boldsymbol{\lambda}}^{(k-1)}/\|\hat{\boldsymbol{\lambda}}^{(k-1)}\|)\in\mathcal{D}$. By substituting it into (A13) and by (16), we obtain $(\hat{\boldsymbol{\lambda}}^{(k-1)})^{T}\boldsymbol{g}(\hat{\boldsymbol{x}}^{(k-1)})\leq(D_\lambda/\delta)\xi^{(k-1)}$ which, by taking $k\to\infty$, leads to

$$\limsup_{k\to\infty}\left(\hat{\boldsymbol{\lambda}}^{(k-1)}\right)^{T}\boldsymbol{g}\left(\hat{\boldsymbol{x}}^{(k-1)}\right)\leq 0. \qquad \text{(A16)}$$

On the other hand, by letting $\boldsymbol{\lambda}=\mathbf{0}\in\mathcal{D}$, from (A13) we have $-(\hat{\boldsymbol{\lambda}}^{(k-1)})^{T}\boldsymbol{g}(\hat{\boldsymbol{x}}^{(k-1)})\leq\xi^{(k-1)}+(\hat{\boldsymbol{\lambda}}^{\star}-\hat{\boldsymbol{\lambda}}^{(k-1)})^{T}\boldsymbol{g}(\hat{\boldsymbol{x}}^{(k-1)})\leq\xi^{(k-1)}+NC_g\|\hat{\boldsymbol{\lambda}}^{(k-1)}-\hat{\boldsymbol{\lambda}}^{\star}\|$. Since $\lim_{k\to\infty}\xi^{(k-1)}=0$ and $\lim_{k\to\infty}\|\hat{\boldsymbol{\lambda}}^{(k)}-\hat{\boldsymbol{\lambda}}^{\star}\|=0$ by Lemma 6, it follows that $\liminf_{k\to\infty}(\hat{\boldsymbol{\lambda}}^{(k-1)})^{T}\boldsymbol{g}(\hat{\boldsymbol{x}}^{(k-1)})\geq 0$, which along with (A16) yields the second term in (58). ■

APPENDIX D
PROOF OF LEMMA 8

The definition of $\hat{\boldsymbol{\alpha}}^{(k)}$ in (43) implies that

$$\begin{aligned}
&\boldsymbol{g}_i^{T}\left(\hat{\boldsymbol{\alpha}}_i^{(k)}\right)\hat{\boldsymbol{\lambda}}^{(k-1)}\\
&\quad+\left(\hat{\boldsymbol{\alpha}}_i^{(k)}-\boldsymbol{x}_i^{(k-1)}\right)^{T}\nabla\boldsymbol{f}_i^{T}\left(\boldsymbol{x}_i^{(k-1)}\right)\nabla\mathcal{F}\left(N\hat{\boldsymbol{y}}^{(k-1)}\right)\\
&\quad+\frac{1}{2\rho_1}\left\|\hat{\boldsymbol{\alpha}}_i^{(k)}-\boldsymbol{x}_i^{(k-1)}\right\|^2\leq\boldsymbol{g}_i^{T}\left(\boldsymbol{x}_i^{(k-1)}\right)\hat{\boldsymbol{\lambda}}^{(k-1)}
\end{aligned}$$

which, by summing over $i=1,\ldots,N$, yields

$$\begin{aligned}
&\boldsymbol{g}^{T}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\hat{\boldsymbol{\lambda}}^{(k-1)}+\left(\hat{\boldsymbol{\alpha}}^{(k)}-\boldsymbol{x}^{(k-1)}\right)^{T}\nabla\bar{\mathcal{F}}\left(\boldsymbol{x}^{(k-1)}\right)\\
&\qquad+\frac{1}{2\rho_1}\left\|\hat{\boldsymbol{\alpha}}^{(k)}-\boldsymbol{x}^{(k-1)}\right\|^2\leq\boldsymbol{g}^{T}\left(\boldsymbol{x}^{(k)}\right)\hat{\boldsymbol{\lambda}}^{(k-1)} \qquad \text{(A17)}
\end{aligned}$$

where $\boldsymbol{g}(\hat{\boldsymbol{\alpha}}^{(k)})=\sum_{i=1}^{N}\boldsymbol{g}_i^{T}(\hat{\boldsymbol{\alpha}}_i^{(k)})$. By substituting the decent lemma in [49, Lemma 2.1]

$$\begin{aligned}
\bar{\mathcal{F}}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)&\leq\bar{\mathcal{F}}\left(\boldsymbol{x}^{(k-1)}\right)+\left(\hat{\boldsymbol{\alpha}}^{(k)}-\boldsymbol{x}^{(k-1)}\right)^{T}\nabla\bar{\mathcal{F}}\left(\boldsymbol{x}^{(k-1)}\right)\\
&\quad+\frac{G_{\bar{\mathcal{F}}}}{2}\left\|\hat{\boldsymbol{\alpha}}^{(k)}-\boldsymbol{x}^{(k-1)}\right\|^2 \qquad \text{(A18)}
\end{aligned}$$

into (A17), we then obtain

$$\begin{aligned}
&\left(\frac{1}{2\rho_1}-\frac{G_{\bar{\mathcal{F}}}}{2}\right)\left\|\hat{\boldsymbol{\alpha}}^{(k)}-\boldsymbol{x}^{(k-1)}\right\|^2\\
&\qquad\leq\mathcal{L}\left(\boldsymbol{x}^{(k-1)},\hat{\boldsymbol{\lambda}}^{(k-1)}\right)-\mathcal{L}\left(\hat{\boldsymbol{\alpha}}^{(k)},\hat{\boldsymbol{\lambda}}^{(k-1)}\right) \qquad \text{(A19)}
\end{aligned}$$

which, after combining with (A10), yields (60).

To show the second part of this lemma, let us recall (A3) that $\hat{\boldsymbol{\alpha}}_i^{(k)}$ in (43) can be alternatively written as

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}_i^{(k)}=\arg\min_{\boldsymbol{\alpha}_i\in\mathcal{X}_i}\boldsymbol{g}_i^{T}(\boldsymbol{\alpha}_i)\hat{\boldsymbol{\lambda}}^{(k-1)}+\Big(\nabla\boldsymbol{f}_i^{T}\left(\boldsymbol{x}_i^{(k-1)}\right)\\
\times\nabla\mathcal{F}\left(N\hat{\boldsymbol{y}}^{(k-1)}\right)\frac{1}{\rho_1}\left(\hat{\boldsymbol{\alpha}}_i^{(k)}-\boldsymbol{x}_i^{(k-1)}\right)\Big)^{T}\boldsymbol{\alpha}_i
\end{aligned}$$

which implies that, for all $\boldsymbol{x}_i\in\mathcal{X}_i$, we have

$$\begin{aligned}
&\boldsymbol{g}_i^{T}\left(\hat{\boldsymbol{\alpha}}_i^{(k)}\right)\hat{\boldsymbol{\lambda}}^{(k-1)}\\
&\quad+\Big(\nabla\boldsymbol{f}_i^{T}\Big(\boldsymbol{x}_i^{(k-1)}\Big)\nabla\mathcal{F}\Big(N\hat{\boldsymbol{y}}^{(k-1)}\Big)+\frac{1}{\rho_1}\Big(\hat{\boldsymbol{\alpha}}_i^{(k)}-\boldsymbol{x}_i^{(k-1)}\Big)\Big)^{T}\\
&\quad\times\left(\hat{\boldsymbol{\alpha}}_i^{(k)}-\boldsymbol{x}_i^{(k-1)}\right)\\
&\qquad\leq\boldsymbol{g}_i^{T}(\boldsymbol{x}_i)\hat{\boldsymbol{\lambda}}^{(k-1)}\\
&\qquad\quad+\left(\nabla\boldsymbol{f}_i^{T}\left(\boldsymbol{x}_i^{(k-1)}\right)\nabla\mathcal{F}\left(N\hat{\boldsymbol{y}}^{(k-1)}\right)\right.\\
&\qquad\qquad\left.+\frac{1}{\rho_1}\left(\hat{\boldsymbol{\alpha}}_i^{(k)}-\boldsymbol{x}_i^{(k-1)}\right)\right)^{T}\left(\boldsymbol{x}_i-\boldsymbol{x}_i^{(k-1)}\right).
\end{aligned}$$

By summing the above inequality over $i = 1, \ldots, N$, one obtains, for all $\boldsymbol{x} \in \mathcal{X}$

$$\begin{aligned}
&\boldsymbol{g}^T\left(\hat{\boldsymbol{\alpha}}^{(k)}\right)\hat{\boldsymbol{\lambda}}^\star + \nabla\bar{\mathcal{F}}^T\left(\boldsymbol{x}^{(k-1)}\right)\left(\hat{\boldsymbol{\alpha}}^{(k)} - \boldsymbol{x}^{(k-1)}\right) \\
&\quad + \frac{1}{\rho_1}\left\|\hat{\boldsymbol{\alpha}}^{(k)} - \boldsymbol{x}^{(k-1)}\right\|^2 \\
&\quad \le \boldsymbol{g}^T(\boldsymbol{x})\hat{\boldsymbol{\lambda}}^\star + \nabla\bar{\mathcal{F}}^T\left(\boldsymbol{x}^{(k-1)}\right)\left(\boldsymbol{x} - \boldsymbol{x}^{(k-1)}\right) \\
&\qquad + \frac{1}{\rho_1}\sum_{i=1}^{N}\left(\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{x}_i^{(k-1)}\right)\left(\boldsymbol{x}_i - \boldsymbol{x}_i^{(k-1)}\right) \\
&\qquad + \left(\hat{\boldsymbol{\lambda}}^\star - \hat{\boldsymbol{\lambda}}^{(k-1)}\right)\left(\boldsymbol{g}\left(\hat{\boldsymbol{\alpha}}^{(k)}\right) - \boldsymbol{g}(\boldsymbol{x})\right) \\
&\quad \le \boldsymbol{g}^T(\boldsymbol{x})\hat{\boldsymbol{\lambda}}^\star + \bar{\mathcal{F}}(\boldsymbol{x}) - \bar{\mathcal{F}}\left(\boldsymbol{x}^{(k-1)}\right) \\
&\qquad + \frac{2D_x}{\rho_1}\sum_{i=1}^{N}\left\|\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{x}_i^{(k-1)}\right\| + 2C_g\left\|\hat{\boldsymbol{\lambda}}^\star - \hat{\boldsymbol{\lambda}}^{(k-1)}\right\|
\end{aligned}$$

where we have utilized the convexity of $\bar{\mathcal{F}}$, boundedness of $\mathcal{X}_i$ and the constraint functions [cf. Assumption 1 and (30)] in obtaining the last inequality. By applying (A18) to the above inequality and by the premise of $1/\rho_1 \ge G_{\bar{\mathcal{F}}} > G_{\bar{\mathcal{F}}}/2$, we further obtain, for all $\boldsymbol{x} \in \mathcal{X}$

$$\begin{aligned}
\mathcal{L}(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star) &\le \mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{\lambda}}^\star) + \frac{2D_x}{\rho_1}\sum_{i=1}^{N}\left\|\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{x}_i^{(k-1)}\right\| \\
&\quad + 2C_g\left\|\hat{\boldsymbol{\lambda}}^\star - \hat{\boldsymbol{\lambda}}^{(k-1)}\right\| + \left|\mathcal{L}(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star) - \mathcal{L}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^\star\right)\right| \qquad \text{(A20)}
\end{aligned}$$

in which one can bound the last term, using (34), (27), (29) and (16), by

$$\left|\mathcal{L}(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star) - \mathcal{L}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^\star\right)\right| \le (L_{\bar{\mathcal{F}}} + ND_\lambda\sqrt{P}L_g)\left\|\hat{\boldsymbol{x}}^\star - \hat{\boldsymbol{\alpha}}^{(k)}\right\|. \qquad \text{(A21)}$$

Suppose that $\mathcal{L}(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\beta}}^{(k)}) - \mathcal{L}(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) \to 0$ and $(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)})$ converges to some limit point $(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star)$ as $k \to \infty$. Then, by (60) and since $1/\rho_1 \ge G_{\bar{\mathcal{F}}}$, we have $\|(\boldsymbol{x}^{(k-1)}, \hat{\boldsymbol{\lambda}}^{(k-1)}) - (\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\beta}}^{(k)})\| \to 0$, as $k \to \infty$. Therefore

$$\begin{aligned}
\lim_{k\to\infty}\Bigg(&\frac{2D_x}{\rho_1}\sum_{i=1}^{N}\left\|\hat{\boldsymbol{\alpha}}_i^{(k)} - \boldsymbol{x}_i^{(k-1)}\right\| + 2C_g\left\|\hat{\boldsymbol{\lambda}}^\star - \hat{\boldsymbol{\lambda}}^{(k-1)}\right\| \\
&+ \left|\mathcal{L}\left(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\lambda}}^\star\right) - \mathcal{L}(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star)\right|\Bigg) = 0.
\end{aligned}$$

Thus, it follows from (A20), (A21) and the above equation that $\mathcal{L}(\hat{\boldsymbol{x}}^\star, \hat{\boldsymbol{\lambda}}^\star) \le \mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{\lambda}}^\star)$ for all $\boldsymbol{x} \in \mathcal{X}$. The rest of the proof is similar to that of Lemma 5. ■

REFERENCES

[1] V. Lesser, C. Ortiz, and M. Tambe, *Distributed Sensor Networks: A Multiagent Perspective*. Norwell, MA: Kluwer Academic Publishers, 2003.

[2] M. Rabbat and R. Nowak, "Distributed optimization in sensor networks," in *Proc. ACM IPSN*, Berkeley, CA, USA, Apr. 26–27, 2004, pp. 20–27.

[3] M. Chiang, P. Hande, T. Lan, and W. C. Tan, "Power control in wireless cellular networks," *Found. Trends Netw.*, vol. 2, no. 4, pp. 381–533, 2008.

[4] C. Shen, T.-H. Chang, K.-Y. Wang, Z. Qiu, and C.-Y. Chi, "Distributed robust multicell coordianted beamforming with imperfect csi: An ADMM approach," *IEEE Trans. Signal Processing*, vol. 60, no. 6, pp. 2988–3003, 2012.

[5] D. Belomestny, A. Kolodko, and J. Schoenmakers, "Regression methods for stochastic control problems and their convergence analysis," *SIAM J. Control Optim.*, vol. 48, no. 5, pp. 3562–3588, 2010.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, Prediction*. New York: Springer-Verlag, 2001.

[7] M. Elad, *Sparse and Redundant Rerpesentations*. New York: Springer Science+Business Media, 2010.

[8] R. Cavalcante, I. Yamada, and B. Mulgrew, "An adaptive projected subgradient approach to learning in diffusion networks," *IEEE Trans. Signal Processing*, vol. 57, no. 7, pp. 2762–2774, Aug. 2009.

[9] B. Johansson, T. Keviczky, M. Johansson, and K. Johansson, "Subgradient methods and consensus algorithms for solving convex optimization problems," in *Proc. IEEE CDC*, Cancun, Mexica, Dec. 9–11, 2008, pp. 4185–4190.

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

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

[12] I. Lobel and A. Ozdaglar, "Distributed subgradient methods for convex optimization over random networks," *IEEE Trans. Autom. Control*, vol. 56, no. 6, pp. 1291–1306, Jun. 2011.

[13] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Distributed stochastic subgradeint projection algorithms for convex optimization," *J. Optim. Theory Appl.*, vol. 147, pp. 516–545, 2010.

[14] J. Chen and A. H. Sayed, "Diffusion adaption strategies for distributed optimization and learning networks," *IEEE. Trans. Signal Process.*, vol. 60, no. 8, pp. 4289–4305, Aug. 2012.

[15] 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.

[16] D. Yuan, S. Xu, and H. Zhao, "Distributed primal-dual subgradient method for multiagent optimization via consensus algorithms," *IEEE Trans. Systems, Man, Cybern. B*, vol. 41, no. 6, pp. 1715–1724, Dec. 2011.

[17] S. S. Ram, A. Nedić, and V. V. Veeravalli, "A new class of distributed optimization algorithm: Application of regression of distributed data," *Optim. Methods Software*, vol. 27, no. 1, pp. 71–88, 2012.

[18] T.-H. Chang, M. Alizadeh, and A. Scaglione, "Coordinated home energy management for real-time power balancing," in *Proc. IEEE PES General Meeting*, San Diego, CA, Jul. 22–26, 2012, pp. 1–8.

[19] N. Li, L. Chen, and S. H. Low, "Optimal demand response based on utility maximization in power networks," in *Proc. IEEE PES General Meeting*, Detroit, MI, USA, Jul. 24–29, 2011, pp. 1–8.

[20] J. C. V. Quintero, "Decentralized Control Techniques Applied to Electric Power Distributed Generation in Microgrids," MS thesis, Departament d'Enginyeria de Sistemes, Automática i Informńtica Industrial, Universitat Politécnica de Catalunya, Barcelona, Spain, 2009.

[21] M. Kallio and A. Ruszczyński, "'Perturbation Methods for Saddle Point Computation,'," International Institute for Applied Systems Analysis, Rep. WP-94-38, 1994.

[22] M. Kallio and C. H. Rosa, "Large-scale convex optimization via saddle-point computation," *Oper. Res.*, pp. 93–101, 1999.

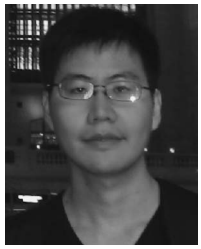
[23] A. Olshevsky and J. N. Tsitsiklis, "Convergence rates in distributed consensus averaging," in *Proc. IEEE CDC*, San Diego, CA, USA, Dec. 13–15, 2006, pp. 3387–3392.

[24] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Syst. Control Lett.*, vol. 53, pp. 65–78, 2004.

[25] B. Yang and M. Johansson, "Distributed optimization and games: A tutorial overview," in *Networked Control Systems*, A. Bemporad, M. Heemels, and M. Johansson, Eds. New York: Springer-Verlag, 2010, ch. 4, LNCIS 406.

[26] H. Uzawa, "Iterative methods in concave programming," in *Studies in Linear and Nonlinear Programming*, K. Arrow, L. Hurwicz, and H. Uzawa, Eds. Stanford, CA: Stanford Univ. Press, 1958, pp. 154–165.

[27] A. Nedić and A. Ozdaglar, "Subgradeint methods for saddle-point problems," *J. Optim. Theory Appl.*, vol. 142, pp. 205–228, 2009.
[28] K. Srivastava, A. Nedić, and D. Stipanović, "Distributed Bregman-distance algorithms for min-max optimization," in *Agent-Based Optimization*, I. Czarnowski, P. Jedrzejowicz, and J. Kacprzyk, Eds. New York: Springer Studies in Computational Intelligence (SCI), 2012, in book.
[29] M. Alizadeh, X. Li, Z. Wang, A. Scaglione, and R. Melton, "Demand side management in the smart grid: Information processing for the power switch," *IEEE Signal Process. Mag.*, vol. 59, no. 5, pp. 55–67, Sep. 2012.
[30] X. Guan, Z. Xu, and Q.-S. Jia, "Energy-efficient buildings facilitated by microgrid," *IEEE Trans. Smart Grid*, vol. 1, no. 3, pp. 243–252, Dec. 2010.
[31] N. Cai and J. Mitra, "A decentralized control architecture for a microgrid with power electronic interfaces," in *Proc. North Amer. Power Symp. (NAPS)*, Sep. 26–28, 2010, pp. 1–8.
[32] D. Hershberger and H. Kargupta, "Distributed multivariate regression using wavelet based collective data mining," *J. Parallel Distrib. Comput.*, vol. 61, no. 3, pp. 372–400, March 2001.
[33] H. Kargupta, B.-H. Park, D. Hershberger, and E. Johnson, "Collective data mining: A new perspective toward distributed data mining," in *Proc. Advances in Distributed Data Mining*, H. Kargupta and P. Chan, Eds., 1999.
[34] D. P. Bertsekas, *Network Optimization: Continuous and Discrete Models*. Cambridge, MA: Athena Scientific, 1998.
[35] R. Madan and S. Lall, "Distributed algorithms for maximum lifetime routing in wireless sensor networks," *IEEE. Trans. Wireless Commun.*, vol. 5, no. 8, pp. 2185–2193, Aug. 2006.
[36] T.-H. Chang, A. Nedich, and A. Scaglione, "Distributed sparse regression by consensus-based primal-dual perturbation optimization," in *Proc. IEEE Global Conf. Signal Info. Process. (GlobalSIP)*, Austin, TX, Dec. 3–5, 2013.
[37] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge Univ. Press, 2004.
[38] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex Analysis and Optimization*. Cambridge, MA: Athena Scientific, 2003.
[39] I. Y. Zabotin, "A subgradient method for finding a saddle point of a convex-concave function," *Issled. Prikl. Mat.*, vol. 15, pp. 6–12, 1988.
[40] Y. Nesterov, "Smooth minimization of nonsmooth functions," *Math. Program.*, vol. 103, no. 1, pp. 127–152, 2005.
[41] G. Mateos, J. A. Bazerque, and G. B. Giannakis, "Distributed sparse linear regression," *IEEE Trans. Signal Processing*, vol. 58, no. 10, pp. 5262–5276, Dec. 2010.
[42] 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, April 2012.
[43] A. Nedić and A. Ozdaglar, "Approximate primal solutions and rate analysis for dual subgradient methods," *SIAM J. OPtim.*, vol. 19, no. 4, pp. 1757–1780, 2009.
[44] T. Larsson, Patriksson, and A.-B. Strömberg, "Ergodic, primal convergence in dual subgradient schemes for convex programming," *Math. Program.*, vol. 86, pp. 238–312, 1999.
[45] B. T. Polyak, *Introduction to Optimization*. New York: Optimization Software Inc., 1987.
[46] T.-H. Chang, A. Nedich, and A. Scaglione, Electronic Companion for Distributed Constrained Optimization by Consensus-Based Primal-Dual Perturbation Method 2013. [Online]. Available: http://arxiv.org
[47] T.-H. Chang, M. Alizadeh, and A. Scaglione, "Real-time power balancing via decentralized coordinated home energy scheduling," *IEEE Trans. Smart Grid*, vol. 4, no. 3, pp. 1490–1504, Sep. 2013.
[48] I. J. Lustig and A. D. F. S. R. E. Marsten, "Interior point methods for linear programming: Computational state of the art," *ORSA J. Comput.*, vol. 6, no. 1, pp. 1–14, 1994.
[49] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Upper Saddle River, NJ, USA: Prentice-Hall, 1989.



Tsung-Hui Chang (S'07–M'08) received the B.S. degree in electrical engineering and the Ph.D. degree in communications engineering from the National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2003 and 2008, respectively.

Since September 2012, he has been with the Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology (NTUST), Taipei, Taiwan, as an Assistant Professor. Before joining NTUST, he held research positions with NTHU (2008–2011), and University of California at Davis, CA (2011–2012). He was also a Visiting Scholar of the University of Minnesota, Twin Cities, MN, the Chinese University of Hong Kong and Xidian University, Xian, China. His research interests are widely in signal processing problems in wireless communications and smart grid, and convex optimization methods and its applications.

Angelia Nedić (M'02) received the B.S. degree in mathematics from the University of Montenegro, Podgorica, Montenegro, in 1987, the M.S. degree in mathematics from the University of Belgrade, Belgrade, Serbia, in 1990, the Ph.D. degree in mathematics and mathematical physics from Moscow State University, Moscow, Russia, in 1994, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, in 2002.

She has been at the BAE Systems Advanced Information Technology from 2002 to 2006. Since 2006, she has been with the Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign (UIUC), where she is holding an Associate Professor position. Her current research interest is focused on large-scale convex optimization, distributed multi-agent optimization, equilibrium problems, and duality theory.

Dr. Nedić received an NSF Faculty Early Career Development Award in 2007 in Operations Research, and the Donald Biggar Willett Scholar title in 2013 from the College of Engineering at UIUC.

Anna Scaglione (F'11) received the M.Sc. and Ph.D. degrees from the Universita di Roma "La Sapienza," Rome, Italy, in 1995 and 1999, respectively.

She is currently Professor in Electrical and Computer Engineering, University of California at Davis. She was previously a member of the faculty at Cornell University, Ithaca, NY, from 2001 to 2006 and at the University of New Mexico, Albuquerque, from 2000 to 2001. Her expertise is in the broad area of signal processing for communication systems and networks. Her current research focuses on studying and enabling decentralized learning and signal processing in networks of sensors. She also focuses on sensor systems and networking models for the demand side management and reliable energy delivery.

Dr. Scaglione received the 2000 IEEE Signal Processing Transactions Best Paper Award, NSF Career Award in 2002, the Best Paper Award (MILCOM 2005), the 2013 IEEE Donald G. Fink Prize Paper Award, and the 2013 IEEE Signal Processing Society Young Author Best Paper Award.