SIAM J. OPTIM.
Vol. 29, No. 4, pp. 3036–3068

# A DISTRIBUTED ADMM-LIKE METHOD FOR RESOURCE SHARING OVER TIME-VARYING NETWORKS[*]

NECDET SERHAT AYBAT[†] AND ERFAN YAZDANDOOST HAMEDANI[†]

**Abstract.** We consider cooperative multiagent resource sharing problems over time-varying communication networks, where only local communications are allowed. The objective is to minimize the sum of agent-specific composite convex functions subject to a conic constraint that couples agents' decisions. We propose a distributed primal-dual algorithm, DPDA-D, to solve the saddle-point formulation of the sharing problem on time-varying (un)directed communication networks; and we show that the primal-dual iterate sequence converges to a point defined by a primal optimal solution and a consensual dual price for the coupling constraint. Furthermore, we provide convergence rates for suboptimality, infeasibility, and consensus violation of agents' dual price assessments; examine the effect of underlying network topology on the convergence rates of the proposed decentralized algorithm; and compare DPDA-D with centralized methods on the basis pursuit denoising and multichannel power allocation problems.

**Key words.** multiagent distributed optimization, primal-dual method, resource sharing problem, convex optimization, convergence rate

**AMS subject classifications.** 90C25, 90C46, 68W15

**DOI.** 10.1137/17M1151973

## 1. Introduction.

Let $\{\mathcal{G}^t\}_{t\in\mathbb{R}_+}$ denote a time-varying graph of $N$ computing nodes. More precisely, for $t \geq 0$, the graph has the form $\mathcal{G}^t = (\mathcal{N}, \mathcal{E}^t)$, where $\mathcal{N} \triangleq \{1, \ldots, N\}$ and $\mathcal{E}^t \subseteq \mathcal{N} \times \mathcal{N}$ is the set of *directed* edges at time $t$. Suppose each node $i \in \mathcal{N}$ has a *private* constraint function $g_i : \mathbb{R}^{n_i} \to \mathbb{R}^m$ and a *private* cost function $\varphi_i : \mathbb{R}^{n_i} \to \mathbb{R} \cup \{+\infty\}$ such that

$$\varphi_i(\xi_i) \triangleq \rho_i(\xi_i) + f_i(\xi_i), \tag{1.1}$$

where $\rho_i : \mathbb{R}^{n_i} \to \mathbb{R} \cup \{+\infty\}$ is a proper, closed convex function (possibly *nonsmooth*), and $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ is a *smooth* convex function. Assuming each node $i \in \mathcal{N}$ has only access to $\varphi_i$, $g_i$ and a closed convex cone $\mathcal{K} \subseteq R^m$, consider the following problem:

$$\min_{\boldsymbol{\xi}\in\mathbb{R}^n} \ \varphi(\boldsymbol{\xi}) \triangleq \sum_{i\in\mathcal{N}} \varphi_i(\xi_i) \quad \text{s.t.} \quad g(\boldsymbol{\xi}) \triangleq \sum_{i\in\mathcal{N}} g_i(\xi_i) \in -\mathcal{K}, \tag{1.2}$$

where $\xi_i \in \mathbb{R}^{n_i}$ denotes the *local* decision of node $i \in \mathcal{N}$ and $n \triangleq \sum_{i\in\mathcal{N}} n_i$.

ASSUMPTION 1. *For all $i \in \mathcal{N}$, the function $f_i$ is differentiable on an open set containing* $\mathbf{dom}\,\rho_i$, *and $\nabla f_i$ is Lipschitz with constant $L_{f_i}$; the prox map of $\rho_i$,*

$$\mathbf{prox}_{\rho_i}(\xi_i) \triangleq \operatorname*{argmin}_{x_i\in\mathbb{R}^{n_i}} \left\{ \rho_i(x_i) + \tfrac{1}{2}\left\|x_i - \xi_i\right\|^2 \right\}, \tag{1.3}$$

[*] Received by the editors October 16, 2017; accepted for publication (in revised form) July 5, 2019; published electronically December 12, 2019. In our prior work [2], published in the *Proceedings of the* 50*th Asilomar Conference on Signals, Systems and Computer*, we studied the resource sharing problem over static and undirected communication networks; see section 1.2 for details.

https://doi.org/10.1137/17M1151973

**Funding:** This research was partially supported by NSF grants CMMI-1400217 and CMMI-1635106 and by ARO grant W911NF-17-1-0298.

[†] Industrial and Manufacturing Engineering Department, The Pennsylvania State University, University Park, PA 16802 (nsa10@psu.edu, evy5047@psu.edu).

*is* efficiently *computable, where* $\|.\|$ *denotes the Euclidean norm. Moreover,* $g_i$ *is* $\mathcal{K}$*-convex* [7, Chapter 3.6.2] *and Lipschitz continuous with constant* $C_{g_i}$ *and has a Lipschitz continuous Jacobian,* $\mathbf{J}g_i$*, with constant* $L_{g_i}$*.*

In this paper, we design a distributed algorithm for solving (1.2) and provide a unified approach for analyzing the convergence behavior of the proposed method, regardless of whether the communications over the time-varying graph $\{\mathcal{G}^t\}$ are unidirectional or bidirectional. To this aim, we need some definitions and assumptions related to the time-varying graph $\{\mathcal{G}^t\}$. To unify the notation, we assume all edges are directed and consider undirected graphs as a special case of directed graphs.

Definition 1. *For any* $t \geq 0$*,* $\mathcal{G}^t = (\mathcal{N}, \mathcal{E}^t)$ *is a directed graph; let* $\mathcal{N}_i^{t,\text{in}} \triangleq \{j \in \mathcal{N} : \ (j,i) \in \mathcal{E}^t\} \cup \{i\}$ *and* $\mathcal{N}_i^{t,\text{out}} \triangleq \{j \in \mathcal{N} : \ (i,j) \in \mathcal{E}^t\} \cup \{i\}$ *denote the in-neighbors and out-neighbors of node* $i \in \mathcal{N}$ *at time* $t$*, respectively; and let* $d_i^t \triangleq |\mathcal{N}_i^{t,\text{out}}| - 1$ *be the out-degree of node* $i \in \mathcal{N}$*.* $\mathcal{G}^t = (\mathcal{N}, \mathcal{E}^t)$ *is called undirected when* $(i,j) \in \mathcal{E}^t$ *if and only if* $(j,i) \in \mathcal{E}^t$*. For undirected* $\mathcal{G}^t$*, let* $\mathcal{N}_i^t \triangleq \mathcal{N}_i^{t,\text{in}} \setminus \{i\} = \mathcal{N}_i^{t,\text{out}} \setminus \{i\}$ *denote the neighbors of* $i \in \mathcal{N}$*, and* $d_i^t \triangleq |\mathcal{N}_i^t|$ *represents the degree of node* $i \in \mathcal{N}$ *at time* $t$*.*

Assumption 2. *When* $\mathcal{G}^t$ *is a (general) directed graph, node* $i \in \mathcal{N}$ *can receive data from* $j \in \mathcal{N}$ *only if* $j \in \mathcal{N}_i^{t,\text{in}}$*, i.e.,* $(j,i) \in \mathcal{E}^t$*, and can send data to* $j \in \mathcal{N}$ *only if* $j \in \mathcal{N}_i^{t,\text{out}}$*, i.e.,* $(i,j) \in \mathcal{E}^t$*; on the other hand, when* $\mathcal{G}^t$ *is undirected, node* $i \in \mathcal{N}$ *can send and receive data to and from* $j \in \mathcal{N}$ *at time* $t$ *only if* $j \in \mathcal{N}_i^t$*, i.e.,* $(i,j) \in \mathcal{E}^t$*.*

Our objective is to solve (1.2) in a *decentralized* fashion using the computing nodes in $\mathcal{N}$ while the information exchange among the nodes is restricted to edges in $\mathcal{E}^t$ for $t \geq 0$ according to Assumption 2. We are interested in designing algorithms which can distribute the computation over the nodes such that each node's computation is based on the local topology of $\mathcal{G}^t$ and information only available to that node.

Decentralized optimization over communication networks has drawn attention from a wide range of application areas: coordination and control in multirobot networks, parameter estimation in wireless sensor networks, processing distributed big data in machine learning, and distributed power control in cellular networks, to name a few. In these examples, the network size can be prohibitively large for centralized optimization, which requires a fusion center that collects the physically distributed data and runs a centralized optimization method. This process has expensive communication overhead, requires large enough memory to store and process the data, and also may violate data privacy in case agents are not willing to share their data even though they are collaborative [35]. Therefore, a common objective of today's big data networks is to use decentralized optimization techniques to avoid expensive communication overhead required by the centralized setting and to enhance the data privacy. The communication networks in these application areas may be directed, i.e., communication links can be unidirectional, and/or the network may be time-varying, e.g., communication links in a wireless network can be on/off over time due to failures or the links may exist among agents depending on their interdistances.

In the remainder of this section, as a brief preliminary, we discuss the primal-dual algorithm PDA proposed in [9] to solve convex-concave saddle-point problems with a *bilinear* coupling term, explain its connections to ADMM-like algorithms, and briefly discuss some recent work related to ours. It is worth noting that the saddle-point (SP) problem formulation of (1.2) contains a coupling term that is *not* bilinear due to nonlinear $\{g_i\}_{i\in\mathcal{N}}$; therefore, PDA is not applicable. Next, in section 2, we propose DPDA-D, a new distributed algorithm based on PDA and extending it to handle nonlinear constraints, for solving the SP formulation of the multiagent sharing problem in (1.2) when the topology of the connectivity graph is *time-varying* with *(un)directed* communication links. After we state the main theorem establishing the convergence properties of DPDA-D, we provide the proof of the main theorem in section 3. Subsequently, in sections 4, 5, and 6, we discuss certain details related to the applicability of the method in practice. In section 7, we compare our method with Prox-JADMM [15] on the basis pursuit denoising problem, and with Mirror-prox [22] on the multichannel power allocation problem; and finally, in section 8 we state our concluding remarks and briefly discuss potential future work.

**1.1. Preliminary.** In this paper, we study an *inexact* variant of the PDA proposed in [9], extending it to handle nonlinear constraints, to solve the SP formulation of (1.2) in a decentralized manner over a time-varying communication network. There has been active research on efficient algorithms for convex-concave saddle-point problems $\min_{\mathbf{x}}\max_{\mathbf{y}}\mathcal{L}(\mathbf{x},\mathbf{y})$, e.g., [8, 14, 21, 32]. PDA [9] also belongs to this family and is proposed for the convex-concave SP problem:

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}}\mathcal{L}(\mathbf{x},\mathbf{y})\triangleq\Phi(\mathbf{x})+\langle T(\mathbf{x}),\mathbf{y}\rangle-h(\mathbf{y}),\tag{1.4}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are finite-dimensional vector spaces, $\Phi(\mathbf{x})\triangleq\rho(\mathbf{x})+f(\mathbf{x})$, $\rho$ and $h$ are possibly nonsmooth convex functions, $f$ is a convex function and has a Lipschitz continuous gradient defined on $\mathbf{dom}\,\rho$ with Lipschitz constant $L$, and $T:\mathcal{X}\to\mathcal{Y}$ is a *linear* map. Briefly, given $\mathbf{x}^0\in\mathcal{X}$, $\mathbf{y}^0\in\mathcal{Y}$ and algorithm parameters $\nu_x,\nu_y>0$, PDA consists of two proximal-gradient steps that can be written as

(1.5a)
$$\mathbf{x}^{k+1}\leftarrow\operatorname*{argmin}_{\mathbf{x}}\ \rho(\mathbf{x})+f(\mathbf{x}^k)+\left\langle\nabla f(\mathbf{x}^k),\ \mathbf{x}-\mathbf{x}^k\right\rangle+\left\langle T(\mathbf{x}),\mathbf{y}^k\right\rangle+\frac{1}{\nu_x}D_x(\mathbf{x},\mathbf{x}^k),$$

(1.5b)
$$\mathbf{y}^{k+1}\leftarrow\operatorname*{argmin}_{\mathbf{y}}\ h(\mathbf{y})-\left\langle 2T(\mathbf{x}^{k+1})-T(\mathbf{x}^k),\mathbf{y}\right\rangle+\frac{1}{\nu_y}D_y(\mathbf{y},\mathbf{y}^k),$$

where $D_x$ and $D_y$ are Bregman distance functions corresponding to some continuously differentiable strongly convex functions $\psi_x$ and $\psi_y$ such that $\mathbf{dom}\,\psi_x\supset\mathbf{dom}\,\rho$ and $\mathbf{dom}\,\psi_y\supset\mathbf{dom}\,h$. In particular, $D_x(\mathbf{x},\bar{\mathbf{x}})\triangleq\psi_x(\mathbf{x})-\psi_x(\bar{\mathbf{x}})-\langle\nabla\psi_x(\bar{\mathbf{x}}),\ \mathbf{x}-\bar{\mathbf{x}}\rangle$, and $D_y$ is defined similarly. Abusing the notation, below we use $T$ also to denote the corresponding matrix, i.e., $T(\mathbf{x})=T\mathbf{x}$.

In [9], it is shown that, when the convexity modulus for $\psi_x$ and $\psi_y$ is 1, if $\nu_x,\nu_y>0$ are chosen such that $(\frac{1}{\nu_x}-L)\frac{1}{\nu_y}\geq\sigma^2_{\max}(T)$, then for any $\mathbf{x},\mathbf{y}\in\mathcal{X}\times\mathcal{Y}$,

(1.6)
$$\mathcal{L}(\bar{\mathbf{x}}^K,\mathbf{y})-\mathcal{L}(\mathbf{x},\bar{\mathbf{y}}^K)\leq\tfrac{1}{K}(\tfrac{1}{\nu_x}D_x(\mathbf{x},\mathbf{x}^0)+\tfrac{1}{\nu_y}D_y(\mathbf{y},\mathbf{y}^0)-\left\langle T(\mathbf{x}-\mathbf{x}^0),\mathbf{y}-\mathbf{y}^0\right\rangle)$$

holds $\forall K\geq 1$, where $\bar{\mathbf{x}}^K\triangleq\frac{1}{K}\sum_{k=1}^K\mathbf{x}^k$ and $\bar{\mathbf{y}}^K\triangleq\frac{1}{K}\sum_{k=1}^K\mathbf{y}^k$.

It is worth mentioning the connection between PDA and the alternating direction method of multipliers (ADMM). Indeed, when implemented on $\min_{\mathbf{v}\in\mathcal{X}^*,\mathbf{y}\in\mathcal{Y}}\{\Phi^*(\mathbf{v})+h(\mathbf{y}):\ \mathbf{v}+T^\top\mathbf{y}=\mathbf{0}\}$, preconditioned ADMM is equivalent to PDA [8, 9], where $\mathcal{X}^*$ denotes the dual space and $\Phi^*$ is the convex conjugate of $\Phi$. There is also a strong connection between the linearized ADMM algorithm proposed by Aybat et al. in [5] and PDA proposed in [9]—see section 1.4 in the online technical report [1].

**Notation.** $\|\cdot\|$ denotes the Euclidean or the spectral norm depending on its argument, i.e., for a matrix $R$, $\|R\|=\sigma_{\max}(R)$. Given a convex set $\mathcal{S}$, let $\sigma_{\mathcal{S}}(\cdot)$

denote its support function, i.e., $\sigma_{\mathcal{S}}(\theta) \triangleq \sup_{w\in\mathcal{S}} \langle \theta,\ w\rangle$, let $\mathbb{1}_S(\cdot)$ denote the indicator function of $\mathcal{S}$, i.e., $\mathbb{1}_S(w) = 0$ for $w \in \mathcal{S}$ and equal to $+\infty$ otherwise, and let $\mathcal{P}_{\mathcal{S}}(w) \triangleq \operatorname{argmin}\{\|v - w\| :\ v \in \mathcal{S}\}$ denote the Euclidean projection onto $\mathcal{S}$. For a closed convex set $\mathcal{S}$, we define the distance function as $d_{\mathcal{S}}(w) \triangleq \|\mathcal{P}_{\mathcal{S}}(w) - w\|$. Given a convex cone $\mathcal{K} \in \mathbb{R}^m$, let $\mathcal{K}^*$ denote its dual cone, i.e., $\mathcal{K}^* \triangleq \{\theta \in \mathbb{R}^m :\ \langle\theta, w\rangle \geq 0\ \ \forall w \in \mathcal{K}\}$, and $\mathcal{K}^\circ \triangleq -\mathcal{K}^*$ denote the polar cone of $\mathcal{K}$. Note that for any cone $\mathcal{K} \in \mathbb{R}^m$, $\sigma_{\mathcal{K}}(\theta) = 0$ for $\theta \in \mathcal{K}^\circ$ and equal to $+\infty$ if $\theta \notin \mathcal{K}^\circ$, i.e., $\sigma_{\mathcal{K}}(\theta) = \mathbb{1}_{\mathcal{K}^\circ}(\theta)\ \forall\theta \in \mathbb{R}^m$. Given a convex function $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, its convex conjugate is $h^*(w) \triangleq \sup_{\theta\in\mathbb{R}^n} \langle w, \theta\rangle - h(\theta)$, and for differentiable $h : \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{J}h : \mathbb{R}^n \to \mathbb{R}^{m\times n}$ denotes the Jacobian of $h$. Throughout the paper, $\otimes$ denotes the Kronecker product, $\Pi$ denotes the Cartesian product, and $\mathbf{I}_n$ is the $n \times n$ identity matrix. $Q$-norm is defined as $\|z\|_Q \triangleq (z^\top Qz)^{1/2}$ for any positive definite matrix $Q$.

**1.2. Our previous work on resource sharing.** In [2], we considered (1.2) when $g_i(\xi) = r_i - R_i\xi_i$ is affine for $i \in \mathcal{N}$, over a *static* and *undirected* communication network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ as a dual consensus problem. Using Lagrangian duality, we reformulated it as an SP problem, $\min_{\boldsymbol{\xi}} \max_{y\in\mathcal{K}^\circ} \sum_{i\in\mathcal{N}} \varphi_i(\xi_i) + \langle\sum_{i\in\mathcal{N}} R_i\xi_i - r_i,\ y\rangle$, which can be written in a distributed form through creating local copies of dual variable $y \in \mathbb{R}^n$ as $(P) : \min_{\boldsymbol{\xi}} \max_{\mathbf{y}} \{\sum_{i\in\mathcal{N}} \varphi_i(\xi_i) + \langle R_i\xi_i - r_i,\ y_i\rangle :\ \ y_i \in \mathcal{K}^\circ\ \forall i \in \mathcal{N},\ y_i = y_j\ \forall(i, j) \in \mathcal{E}\}$, where $\boldsymbol{\xi} = [\xi_i]_{i\in\mathcal{N}}$ and $\mathbf{y} = [y_i]_{i\in\mathcal{N}}$. Using $M$, the edge-node incidence matrix of $\mathcal{G}$, the consensus constraints $y_i = y_j$ for $(i, j) \in \mathcal{E}$ can be written as $M\mathbf{y} = \mathbf{0}$. Furthermore, by dualizing the consensus constraints, we obtain another SP problem, equivalent to $(P)$, in the form of (1.4):

$$(1.7)\qquad \min_{\boldsymbol{\xi}} \max_{\mathbf{y}\in\Pi_{i\in\mathcal{N}}\mathcal{K}^\circ} \min_{\mathbf{w}} \mathcal{L}(\boldsymbol{\xi}, \mathbf{w}, \mathbf{y}) = \min_{\boldsymbol{\xi},\mathbf{w}} \max_{\mathbf{y}\in\Pi_{i\in\mathcal{N}}\mathcal{K}^\circ} \mathcal{L}(\boldsymbol{\xi}, \mathbf{w}, \mathbf{y}),$$

where $\mathcal{L}(\boldsymbol{\xi}, \mathbf{w}, \mathbf{y}) \triangleq \sum_{i\in\mathcal{N}} \varphi_i(\xi_i) + \langle R_i\xi_i - r_i,\ y_i\rangle - \langle\mathbf{w}, M\mathbf{y}\rangle$. The equality in (1.7) holds as long as $\mathcal{K}$ is a pointed cone—hence $\mathbf{int}\,(\mathcal{K}^\circ) \neq \emptyset$; therefore, for each fixed $\boldsymbol{\xi}$, inner $\max_{\mathbf{y}}$ and $\min_{\mathbf{w}}$ can be interchanged. The saddle-point problem on the right side of (1.7) is special case of (1.4) with a separable structure. Exploiting this special structure, we customized PDA in (1.5) and proposed Algorithm DPDA-S. In [2] we showed that Algorithm DPDA-S can solve the sharing problem (1.2) with an affine conic constraint in a *decentralized* way and established its convergence properties provided that the node-specific primal-dual step-sizes $\{\tau_i, \kappa_i\}_{i\in\mathcal{N}}$ and the algorithm parameter $\gamma > 0$ satisfy $\frac{1}{\tau_i} > L_{f_i}$ and $(\frac{1}{\tau_i} - L_{f_i})(\frac{1}{\kappa_i} - 2\gamma d_i) \geq \|R_i\|^2\ \ \forall i \in \mathcal{N}$, where $d_i$ denotes the degree of $i \in \mathcal{N}$ for the static $\mathcal{G}$. Our result in [2] refines the error bound in (1.6) and establishes the $\mathcal{O}(1/k)$ ergodic rate in terms of suboptimality and infeasibility of the DPDA-S iterate sequence—see Theorem 2 in [2].

The arguments used for proving Theorem 2 in [2] cannot be used for the *time-varying directed* communication network setting considered in this paper since the undirected network is encoded through the use of an $M\mathbf{y} = \mathbf{0}$ constraint. However, when the topology is time-varying or when the edges are directed, it is not immediately clear how one can represent this problem as an SP problem. To extend our previous results to a more general setting of time-varying topology with possibly directed edges, in this paper we develop a new SP formulation that can impose consensus over the dual variables while the formulation is independent of the changing topology. Finally, the new method can also handle nonlinear conic constraints on resource sharing in (1.2).

**1.3. Related work.** Now we briefly review some recent work on the distributed resource sharing problem. From the application perspective, algorithms and their basic convergence analysis have been studied for the economic dispatch problem (EDP), e.g., [38] for power-flow networks and [20, 43] for smart-grids. The variants of EDP considered in [20, 38, 43] are special cases of (1.2). In particular, each node $i \in \mathcal{N}$ has a convex objective function $f_i$, usually a quadratic function; $\rho_i(\xi_i) = \mathbb{1}_{\mathcal{X}_i}(\xi_i)$, where $\mathcal{X}_i$ is a local simple convex set, $g_i(\xi_i) = \xi_i - r_i$, and $\mathcal{K} = \{\mathbf{0}\}$. In [38], the aim is to optimize the total power generation cost in a DC power-flow model; [20, 43] also study a similar problem considering random wind power injection—both papers establish basic convergence results without any rate guarantees. The distributed resource allocation problem can also arise in controlling and coordinating internet services over hybrid edge-cloud networks, for which a distributed ADMM algorithm is proposed in [23] to solve a problem in the form of (1.2) with $\mathcal{K} = \{\mathbf{0}\}$ and $g_i(\xi_i) = \xi_i - r_i$. Yang et al. [42] study EDP considering communication delays in directed time-varying network topology, and an algorithm based on push-sum protocol is proposed.

From the theoretical point of view, there has been active research on the distributed resource allocation problem. In [16], a distributed Lagrangian method (DLM) has been proposed for solving a particular case of (1.2) on a *static* network; more precisely, the objective is to minimize the sum of local convex functions subject to local convex *compact* sets and a coupling constraint of the form $\sum_{i\in\mathcal{N}} \xi_i - r_i = \mathbf{0}$. In [16], the authors establish the convergence rate of $\mathcal{O}(\log(k)/\sqrt{k})$ for the dual function values estimated at the time-weighted average of dual iterates. Reference [17] gives a gradient balancing protocol to solve (1.2) in which $\rho_i(\cdot) = 0$, $g_i(\xi_i) = \xi_i - r_i$ and $\mathcal{K} = \{\mathbf{0}\}$. The authors show that the generated sequence $\boldsymbol{\xi}^k = [\xi_i^k]_{i\in\mathcal{N}}$ satisfies $\sum_{i\in\mathcal{N}} f_i(\xi_i^k) - \varphi^* \leq \mathcal{O}(1/k)$ and is feasible for all $k$ under the assumption that the initial point $\boldsymbol{\xi}^0 = [\xi_i^0]_{i\in\mathcal{N}}$ is feasible—$\varphi^*$ denotes the optimal value; moreover, a linear rate is established when each $f_i$ is strongly convex. For a similar formulation as in [17], an asynchronous gradient-descent method is proposed in [25] for time-varying undirected communication networks; the proposed algorithm produces a feasible iterate sequence such that $\min_{\ell=1,\ldots,k} \max_{i,j\in\mathcal{N}} \left\|\nabla f_i(\xi_i^\ell) - \nabla f_i(\xi_j^\ell)\right\| \leq \mathcal{O}(1/\sqrt{k})$ when each $f_i$ is convex and has a Lipschitz gradient. However, none of these methods can solve (1.2) in its full generality over a time-varying and directed communication network.

In [11], a method based on ADMM is proposed to reduce the computational work of ADMM due to exact minimizations in each iteration. First, a *dual consensus* ADMM is proposed for solving (1.2) over an undirected static network in a distributed fashion when $\mathcal{K} = \{\mathbf{0}\}$, $g_i(\xi_i) = R_i\xi_i - r_i$, and $\varphi_i(\xi_i) = \rho_i(\xi_i) + f_i(A_i\xi_i)$ for $\rho_i$ and $f_i$ as in (1.1). To avoid exact minimizations in ADMM, an inexact variant taking proximal-gradient steps is analyzed. Convergence of the primal-dual sequence is shown when each $f_i$ is strongly convex—without a rate result; and a linear rate is established in the absence of the nonsmooth $\rho_i$, i.e., $\varphi_i(\xi_i) = f_i(A_i\xi_i)$, and assuming each $A_i$ has full column-rank and $f_i$ is strongly convex, i.e., $\varphi_i$ is strongly convex.

In [10], a proximal *dual consensus* ADMM method, PDC-ADMM, is proposed by Chang to minimize $\sum_{i\in\mathcal{N}} \varphi_i$ subject to coupling equality and agent-specific constraints over both static and time-varying *undirected* networks—for the time-varying topology, they assumed that agents are on/off and communication links fail randomly with certain probabilities. The goal in the paper is to solve $\min_{\boldsymbol{\xi}}\{\sum_i \varphi_i(\xi_i) : \sum_{i\in\mathcal{N}} R_i\xi_i = r,\ \xi_i \in \mathcal{X}_i,\ i \in \mathcal{N}\}$, where $\varphi_i$ is closed convex, $\mathcal{X}_i = \{\xi_i \in \mathcal{S}_i :\ C_i\xi_i \leq d_i\}$, and $\mathcal{S}_i$ is a convex compact set for each $i \in \mathcal{N}$. The polyhedral constraints $\xi_i \in \mathcal{X}_i$ are handled using a penalty formulation without requiring projections onto them. It is shown that both for static and time-varying cases, PDC-ADMM have a $\mathcal{O}(1/k)$ ergodic convergence rate in the mean for suboptimality and infeasibility; that said, in each iteration, costly *exact* minimizations involving $\varphi_i$ are needed. To alleviate this burden, Chang also proposed an inexact PDC-ADMM taking prox-gradient steps when $\varphi_i(\xi_i) = \rho_i(\xi_i) + f_i(A_i\xi_i)$ and $A_i$ is a linear map for each $i \in \mathcal{N}$, and showed $\mathcal{O}(1/k)$ ergodic convergence rate when each $f_i$ is *strongly convex* and differentiable with a Lipschitz continuous gradient for $i \in \mathcal{N}$.

In [12], a consensus-based distributed primal-dual perturbation (PDP) algorithm using a diminishing step-size sequence is proposed. The objective is to minimize a composition of a global network function (smooth) with the sum of local objective functions (smooth), i.e., $\mathcal{F}(\sum_{i\in\mathcal{N}} f_i(x))$, subject to local compact sets and an inequality constraint, $\sum_{i\in\mathcal{N}} g_i(x) \leq 0$, over a time-varying directed network. It is shown that the primal-dual iterate sequence converges to an optimal primal-dual solution; however, no rate result is provided. There are fewer papers on resource allocation over time-varying directed networks. Gu et al. [19] consider a special case of (1.2) with $\mathcal{K} = \{\mathbf{0}\}$, $g_i(\xi_i) = \xi_i - r_i$, $f_i$ convex, and $\rho_i(\xi_i) = \mathbb{1}_{\mathcal{X}_i}(\xi_i)$, where $\mathcal{X}_i$ is convex and compact for $i \in \mathcal{N}$. Assuming a Slater point exists which implies boundedness of a dual optimal set, the authors proved the $\mathcal{O}(\log(k)/\sqrt{k})$ rate result. Reference [41] has the same setting in [19] with $\mathcal{X}_i = [\underline{\xi}_i, \bar{\xi}_i]$. Assuming each $f_i$ is smooth and strongly convex, a distributed method is proposed and its convergence is shown without providing a rate result. Finally, while we were preparing this paper, we became aware of recent work [26, 30]. Reference [26] also uses Fenchel conjugation and the *dual consensus* formulation to decompose separable constraints. A distributed algorithm on time-varying *balanced*[1] directed communication networks is proposed for solving saddle-point problems subject to consensus constraints. Assuming each agents' local iterates and subgradient sets are uniformly bounded, it is shown that the ergodic average of the primal-dual sequence converges with a $\mathcal{O}(1/\sqrt{k})$ rate in terms of the saddle-point evaluation error; however, when the method is applied to constrained optimization problems, *no* rate in terms of suboptimality and infeasibility is provided. The other recent work in [30] investigates the connection between the decentralized resource allocation problem and the decentralized consensus optimization problem where the objective is to minimize the sum of convex functions subject to local closed convex sets and $\sum_{i\in\mathcal{N}} \xi_i - r_i = \mathbf{0}$ over *static undirected* networks. Utilizing the mirror relationship between the optimality conditions of these problems, they proposed a method for solving the decentralized resource allocation problem and proved an $o(1/k)$ rate of convergence in terms of *squared* residuals of first-order optimality conditions.

## 2. A distributed algorithm for time-varying network topology.

In this section, we develop a distributed algorithm for solving (1.2) when the communication network topology is *time-varying*, under the following assumption.

ASSUMPTION 3. *A primal-dual solution to* (1.2) *exists and the duality gap is* 0.

Clearly this assumption holds if a Slater point for (1.2) exists, i.e., there exists some $\bar{\boldsymbol{\xi}} \in \mathbf{relint}(\mathbf{dom}\,\varphi \cap \mathbf{dom}\,g)$ such that $g(\bar{\boldsymbol{\xi}}) \in \mathbf{int}(-\mathcal{K})$. The existence of a Slater point is also assumed in many related papers, e.g., [12, 19, 26, 30, 32]. When $\mathcal{K} = \{\mathbf{0}\}$ and $g_i(\xi) = R_i\xi - r_i$ for $i \in \mathcal{N}$, Assumption 3 trivially holds if there exists some $\bar{\boldsymbol{\xi}} \in \mathbf{relint}(\mathbf{dom}\,\varphi)$ that is feasible, i.e., $\sum_{i\in\mathcal{N}} R_i\bar{\xi}_i - r_i = \mathbf{0}$.

Since $\mathbb{1}_{\mathcal{K}}(\cdot) = \sup_{y\in\mathbb{R}^m}\{\langle y,\ \cdot\rangle - \sigma_{\mathcal{K}}(y)\}$, one can reformulate (1.2) as

$$\min_{\boldsymbol{\xi}} \max_{y\in\mathbb{R}^m} \left\{ \sum_{i\in\mathcal{N}} \varphi_i(\xi_i) - \left\langle \sum_{i\in\mathcal{N}} g_i(\xi_i),\ y \right\rangle - \sigma_{\mathcal{K}}(y) \right\}. \tag{2.1}$$

[1] A directed graph $\mathcal{G}$ is balanced when each node has an equal number of in-degree and out-degree.

According to Assumption 3, a dual optimal solution, $y^* \in \mathcal{K}^\circ$ exists and the duality gap is 0 for (1.2). Suppose each node $i \in \mathcal{N}$ has its own estimate $y_i \in \mathbb{R}^m$ of a dual optimal solution, and $\mathbf{y} = [y_i]_{i\in\mathcal{N}}$ denotes these estimates in long-vector form. We define the *consensus set* as

$$\mathcal{C} \triangleq \{\mathbf{y} \in \mathbb{R}^{m|\mathcal{N}|} :\ \exists \bar{y} \in \mathbb{R}^m \ \text{ s.t. } \ y_i = \bar{y} \quad \forall i \in \mathcal{N}\}. \tag{2.2}$$

Suppose we are given a (possibly trivial) bound $B \in (0,\infty]$ such that $\|y^*\| \leq B$. For instance, if a Slater point is available, then a nontrivial bound $B \in (0,\infty)$ on dual solutions can be obtained by solving a convex problem in a distributed way; on the other hand, when the Slater condition holds for (1.2) but a Slater point is not available, then the nodes can collectively compute a Slater point (see Section 6). Let $\mathcal{B}_0 \triangleq \{y \in \mathbb{R}^m :\ \|y\| \leq 2B\}$ and $\mathcal{B} \triangleq \Pi_{i\in\mathcal{N}}\mathcal{B}_0$, i.e., $\mathcal{B} = \{\mathbf{y} :\ \|y_i\| \leq 2B,\ i \in \mathcal{N}\}$. Finally, we also define the *bounded consensus set*,

$$\widetilde{\mathcal{C}} \triangleq \mathcal{C} \cap \mathcal{B} = \{\mathbf{y} \in \mathbb{R}^{m|\mathcal{N}|} :\ \exists \bar{y} \in \mathcal{B}_0 \subset \mathbb{R}^m \ \text{ s.t. } \ y_i = \bar{y} \quad \forall i \in \mathcal{N}\}. \tag{2.3}$$

We can equivalently reformulate (2.1) as the following dual consensus problem:

$$\min_{\boldsymbol{\xi}} \max_{\mathbf{y}\in\widetilde{\mathcal{C}}} \ L(\boldsymbol{\xi},\mathbf{y}) \triangleq \sum_{i\in\mathcal{N}} \Big(\varphi_i(\xi_i) - \langle g_i(\xi_i),\ y_i\rangle - \sigma_{\mathcal{K}}(y_i)\Big), \tag{2.4}$$

i.e., any saddle point of (2.4) is also a saddle point of (2.1), which follows from the definitions of $\sigma_{\mathcal{K}}(\cdot)$ and $\widetilde{\mathcal{C}}$. Define $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^{m|\mathcal{N}|} \times \mathbb{R}^{m|\mathcal{N}|} \to \mathbb{R} \cup \{\pm\infty\}$ such that

$$\mathcal{L}(\boldsymbol{\xi},\mathbf{w},\mathbf{y}) \triangleq \sum_{i\in\mathcal{N}} \Big(\varphi_i(\xi_i) - \langle g_i(\xi_i),\ y_i\rangle - \sigma_{\mathcal{K}}(y_i)\Big) - \langle \mathbf{w},\mathbf{y}\rangle + \sigma_{\widetilde{\mathcal{C}}}(\mathbf{w}) - \mathbb{1}_{\mathcal{B}}(\mathbf{y}). \tag{2.5}$$

Note that for any $\boldsymbol{\xi} \in \mathbf{dom}\,\varphi$, we have $\max_{\mathbf{y}\in\widetilde{\mathcal{C}}} L(\boldsymbol{\xi},\mathbf{y}) = \max_{\mathbf{y}} \min_{\mathbf{w}} \mathcal{L}(\boldsymbol{\xi},\mathbf{w},\mathbf{y})$; hence, (2.4) can be equivalently written as follows:

$$\min_{\boldsymbol{\xi}} \Big\{ \max_{\mathbf{y}} \min_{\mathbf{w}} \ \mathcal{L}(\boldsymbol{\xi},\mathbf{w},\mathbf{y}) \Big\} \ = \ \min_{\boldsymbol{\xi},\mathbf{w}} \max_{\mathbf{y}} \mathcal{L}(\boldsymbol{\xi},\mathbf{w},\mathbf{y}), \tag{2.6}$$

where interchanging $\max_{\mathbf{y}}$ and $\min_{\mathbf{w}}$ is trivially justified when $\mathcal{B}$ is bounded; in case $B = +\infty$, i.e., $\mathcal{B}_0 = \mathbb{R}^m$, one can directly verify that $\min_{\mathbf{w}} \max_{\mathbf{y}} \mathcal{L}(\boldsymbol{\xi},\mathbf{w},\mathbf{y}) = \min_{\mathbf{w}} \max_{\mathbf{y}} \mathcal{L}(\boldsymbol{\xi},\mathbf{w},\mathbf{y})$ and is equal to $\varphi(\boldsymbol{\xi})$ if $g(\boldsymbol{\xi}) \in -\mathcal{K}$, and $+\infty$ otherwise.

Since we can equivalently solve $\min_{\boldsymbol{\xi},\mathbf{w}} \max_{\mathbf{y}} \mathcal{L}(\boldsymbol{\xi},\mathbf{w},\mathbf{y})$ in (2.6) to solve (1.2), we next generalize PDA iterations in (1.5a)–(1.5b) to solve this saddle-point problem.

Definition 2. *Let* $\mathcal{X} \triangleq \Pi_{i\in\mathcal{N}}\mathbb{R}^{n_i} \times \Pi_{i\in\mathcal{N}}\mathbb{R}^m$ *and* $\mathcal{X} \ni \mathbf{x} = [\boldsymbol{\xi}^\top \mathbf{w}^\top]^\top$ *for* $\boldsymbol{\xi} = [\xi_i]_{i\in\mathcal{N}} \in \mathbb{R}^n$ *and* $\mathbf{w} \in \mathbb{R}^{n_0}$, *where* $n \triangleq \sum_{i\in\mathcal{N}} n_i$ *and* $n_0 \triangleq m|\mathcal{N}|$; *let* $\mathcal{Y} \triangleq \Pi_{i\in\mathcal{N}}\mathbb{R}^m$ *and* $\mathcal{Y} \ni \mathbf{y} = [y_i]_{i\in\mathcal{N}} \in \mathbb{R}^{n_0}$. *Given parameters* $\gamma > 0$, *and* $\tau_i, \kappa_i > 0$ *for* $i \in \mathcal{N}$, *let* $\mathbf{D}_\gamma \triangleq \frac{1}{\gamma}\mathbf{I}_{n_0}$, $\mathbf{D}_\tau \triangleq \operatorname{diag}([\frac{1}{\tau_i}\mathbf{I}_{n_i}]_{i\in\mathcal{N}})$, *and* $\mathbf{D}_\kappa \triangleq \operatorname{diag}([\frac{1}{\kappa_i}\mathbf{I}_m]_{i\in\mathcal{N}})$. *Defining* $\psi_x(\mathbf{x}) \triangleq \frac{1}{2}\boldsymbol{\xi}^\top \mathbf{D}_\tau \boldsymbol{\xi} + \frac{1}{2}\mathbf{w}^\top \mathbf{D}_\gamma \mathbf{w}$ *and* $\psi_y(\mathbf{y}) \triangleq \frac{1}{2}\mathbf{y}^\top \mathbf{D}_\kappa \mathbf{y}$ *leads to the following Bregman distance functions:* $D_x(\mathbf{x},\bar{\mathbf{x}}) = \frac{1}{2}\left\|\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}\right\|^2_{\mathbf{D}_\tau} + \frac{1}{2}\left\|\mathbf{w} - \bar{\mathbf{w}}\right\|^2_{\mathbf{D}_\gamma}$, *and* $D_y(\mathbf{y},\bar{\mathbf{y}}) = \frac{1}{2}\left\|\mathbf{y} - \bar{\mathbf{y}}\right\|^2_{\mathbf{D}_\kappa}$. *To simplify notation, also define* $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$ *and* $\mathcal{Z} \ni \mathbf{z} = [\mathbf{x}^\top \mathbf{y}^\top]^\top$.

Definition 3. *Suppose* $\varphi_i = \rho_i + f_i$ *is a composite convex function defined as in* (1.1) *for* $i \in \mathcal{N}$. *Let* $\Phi(\mathbf{x}) \triangleq \rho(\mathbf{x}) + f(\mathbf{x})$ *and* $h(\mathbf{y}) \triangleq \sum_{i\in\mathcal{N}} h_i(y_i)$, *where* $\rho(\mathbf{x}) \triangleq \sigma_{\widetilde{\mathcal{C}}}(\mathbf{w}) + \sum_{i\in\mathcal{N}} \rho_i(\xi_i)$, $f(\mathbf{x}) \triangleq \sum_{i\in\mathcal{N}} f_i(\xi_i)$ *and* $h_i(y_i) \triangleq \sigma_{\mathcal{K}}(y_i) + \mathbb{1}_{\mathcal{B}_0}(y_i)$ *for* $i \in \mathcal{N}$. *Let* $G : \mathbb{R}^n \to \mathbb{R}^{n_0}$ *such that* $G(\boldsymbol{\xi}) \triangleq [g_i(\xi_i)]_{i\in\mathcal{N}}$ $\forall \mathbf{x} \in \mathcal{X}$ *and define* $T : \mathbb{R}^n \times \mathbb{R}^{n_0} \to \mathbb{R}^{n_0}$ *such that* $T(\mathbf{x}) \triangleq -G(\boldsymbol{\xi}) - \mathbf{w}$; *hence,* $\mathbf{J}T(\mathbf{x}) = [-\mathbf{J}G(\boldsymbol{\xi}) \ \ -\mathbf{I}_{n_0}]$.

With the aim of solving (1.2) as an SP problem, let $\Phi$, $h$, and $T$ be as given in Definition 3, and consider $\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}}\Phi(\mathbf{x})+\langle T(\mathbf{x}),\mathbf{y}\rangle-h(\mathbf{y})$. Hence, given the initial iterates $\boldsymbol{\xi}^0,\mathbf{w}^0,\mathbf{y}^0$ and parameters $\gamma>0$, $\tau_i,\kappa_i>0$ for $i\in\mathcal{N}$, choosing Bregman functions $D_x$ and $D_y$ as in Definition 2, and setting $\nu_x=\nu_y=1$, we propose a modified version of PDA iterations to handle nonlinear $T(\cdot)$; indeed, after linearizing $T(\mathbf{x})$ around $\mathbf{x}^k$ in (1.5a), the iterations in (1.5) can be written as follows for $k\geq 0$:

(2.7a)
$$\mathbf{w}^{k+1}\leftarrow\operatorname*{argmin}_{\mathbf{w}}\sigma_{\widetilde{\mathcal{C}}}(\mathbf{w})-\langle\mathbf{y}^k,\ \mathbf{w}\rangle+\frac{1}{2\gamma}\|\mathbf{w}-\mathbf{v}^k\|^2,$$

(2.7b)
$$\mathbf{v}^{k+1}\leftarrow\mathbf{w}^{k+1},$$

(2.7c)
$$\xi_i^{k+1}\leftarrow\operatorname*{argmin}_{\xi_i}\rho_i(\xi_i)+\langle\nabla f_i(\xi_i^k)-\mathbf{J}g_i(\xi_i^k)^\top y_i^k,\ \xi_i\rangle+\frac{1}{2\tau_i}\|\xi_i-\xi_i^k\|^2,\ i\in\mathcal{N},$$

(2.7d)
$$y_i^{k+1}\leftarrow\operatorname*{argmin}_{y_i\in\mathcal{K}^\circ\cap\mathcal{B}_0}\ \langle 2g_i(\xi_i^{k+1})-g_i(\xi_i^k)+2v_i^{k+1}-v_i^k,\ y_i\rangle+\frac{1}{2\kappa_i}\|y_i-y_i^k\|^2,\ i\in\mathcal{N},$$

where we initialize $\mathbf{v}^0=\mathbf{w}^0$. The reason we introduced an auxiliary sequence $\{\mathbf{v}^k\}_{k\geq0}$ such that $\mathbf{v}^k=[v_i^k]_{i\in\mathcal{N}}$ will be explained shortly. Briefly, in its currently stated form, the computation in (2.7) can be considered as linearized PDA iterations—$T(\cdot)$ in (1.5a)–(1.5b) is linearized around $\mathbf{x}^k$; however, this naive scheme is not suitable for our purposes, i.e., the $\mathbf{w}^{k+1}$ update in (2.7a) is not practical to be computed in a *distributed* manner. Therefore, instead of setting $\mathbf{v}^{k+1}$ to $\mathbf{w}^{k+1}$, we will replace (2.7b) and assign $\mathbf{v}^{k+1}$ to an approximation of $\mathbf{w}^{k+1}$ such that this approximation can be efficiently computed in a distributed way—this modified version of (2.7) will be analyzed as an *inexact* variant of *linearized* PDA.

Using the extended Moreau decomposition for proximal operators, for $k\geq0$,

$$\begin{aligned}\mathbf{w}^{k+1}&=\operatorname*{argmin}_{\mathbf{w}}\sigma_{\widetilde{\mathcal{C}}}(\mathbf{w})+\frac{1}{2\gamma}\left\|\mathbf{w}-(\mathbf{v}^k+\gamma\mathbf{y}^k)\right\|^2=\mathbf{prox}_{\gamma\sigma_{\widetilde{\mathcal{C}}}}(\mathbf{v}^k+\gamma\mathbf{y}^k),\\ &=\gamma\Big[\tfrac{1}{\gamma}\mathbf{v}^k+\mathbf{y}^k-\mathcal{P}_{\widetilde{\mathcal{C}}}(\tfrac{1}{\gamma}\mathbf{v}^k+\mathbf{y}^k)\Big].\end{aligned}\tag{2.8}$$

For an arbitrary $\mathbf{y}=[y_i]_{i\in\mathcal{N}}\in\mathbb{R}^{n_0}$, $\mathcal{P}_{\widetilde{\mathcal{C}}}(\mathbf{y})$ can be computed as $\mathcal{P}_{\widetilde{\mathcal{C}}}(\mathbf{y})=\mathbf{1}\otimes\operatorname{argmin}_{x\in\mathcal{B}_0}\sum_{i\in\mathcal{N}}\|x-y_i\|^2=\ \mathbf{1}\otimes\operatorname{argmin}_{x\in\mathcal{B}_0}\|x-\frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}}y_i\|^2$, where $\mathbf{1}\in\mathbb{R}^{|\mathcal{N}|}$ denotes the vector of all ones. Hence, we can write $\mathcal{P}_{\widetilde{\mathcal{C}}}(\mathbf{y})=\mathcal{P}_{\mathcal{B}}\left((W\otimes\mathbf{I}_m)\mathbf{y}\right)$, where $W\triangleq\frac{1}{|\mathcal{N}|}\mathbf{1}\mathbf{1}^\top\in\mathbb{R}^{|\mathcal{N}|\times|\mathcal{N}|}$. Equivalently,

$$\mathcal{P}_{\widetilde{\mathcal{C}}}(\mathbf{y})=\mathcal{P}_{\mathcal{B}}\left(\mathbf{1}\otimes p(\mathbf{y})\right),\quad\text{where}\quad p(\mathbf{y})\triangleq\frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}}y_i.\tag{2.9}$$

Note that $\mathcal{P}_{\mathcal{B}}(\mathbf{y})=\mathbf{y}\ \forall\mathbf{y}\in\mathcal{Y}$ when $B=\infty$; and for $B<\infty$, $\mathcal{P}_{\mathcal{B}}(\cdot)$ is easy to compute locally since $\mathcal{B}=\Pi_{i\in\mathcal{N}}\mathcal{B}_0$ and $\mathcal{P}_{\mathcal{B}_0}(y)=y\min\{1,\ 2B/\|y\|\}$ for $y\in\mathbb{R}^m$. Furthermore, the $\boldsymbol{\xi}$-step and $\mathbf{y}$-step of the PDA implementation in (2.7) can also be computed locally at each node; however, computing $\mathbf{w}^{k+1}$ requires communication among the nodes. Indeed, evaluating the average operator $p(\cdot)$ is *not* a simple operation in a decentralized computational setting which only allows for communication among neighboring nodes (see Assumption 2). To overcome the issue with decentralized computation of the averaging operator $p(\cdot)$, we will use *multicommunication rounds* to approximate $p(\cdot)$ and analyze the resulting primal-dual iterations as an *inexact* primal-dual algorithm. In [13], the idea of using *multicommunication rounds* has also been exploited within a distributed primal algorithm for *unconstrained* consensus optimization problems over *undirected* communication networks.

We define a *communication round* as an operation over $\mathcal{G}^t$ such that every node simultaneously sends and receives data to and from its neighboring nodes according to Assumption 2—the details of this operation will be discussed shortly. We assume that communication among neighbors occurs *instantaneously* and nodes operate *synchronously*; and we further assume that for each iteration $k \geq 0$, there exists an approximate averaging operator $\mathcal{R}^k(\cdot)$ which can be computed in a decentralized fashion and approximates $\mathcal{P}_{\widetilde{\mathcal{C}}}(\cdot)$ with a decreasing approximation error in $k$.

Assumption 4. *Given a time-varying network $\{\mathcal{G}^t\}_{t\in\mathbb{R}_+}$ such that $\mathcal{G}^t = (\mathcal{N}, \mathcal{E}^t)$ for $t \geq 0$, suppose that there is a global clock known to all $i \in \mathcal{N}$. Assume that the local operations in* (2.7c) *and* (2.7d) *can be completed between two tics of the clock for all $i \in \mathcal{N}$ and $k \geq 1$, and every time the clock ticks a communication round with instantaneous messaging between neighboring nodes takes place subject to Assumption* 2*. Suppose that for each $k \geq 0$ there exists $\mathcal{R}^k(\cdot) = [\mathcal{R}_i^k(\cdot)]_{i\in\mathcal{N}}$ such that $\mathcal{R}_i^k(\cdot)$ can be computed with local information available to node $i \in \mathcal{N}$ and decentralized computation of $\mathcal{R}^k$ requires $q_k$ communication rounds. Furthermore, we assume that there exist $\Gamma > 0$ and $\alpha \in (0,1)$ such that $N\Gamma \geq 1$ and for all $k \geq 0$, $\mathcal{R}^k$ satisfies*

$$\mathcal{R}^k(\mathbf{w}) \in \mathcal{B}, \qquad \|\mathcal{R}^k(\mathbf{w}) - \mathcal{P}_{\widetilde{\mathcal{C}}}(\mathbf{w})\| \leq N\ \Gamma\alpha^{q_k}\ \|\mathbf{w}\| \quad \forall\ \mathbf{w} \in \mathbb{R}^{n_0}. \tag{2.10}$$

The "unit time" is defined to be the length of the interval between two tics of the clock. The assumption that every node $i \in \mathcal{N}$ can finish its $\xi_i$ and $y_i$ updates in one unit time is mainly for the sake of notational simplicity throughout the analysis. All of our results still hold as long as there exists a uniform bound $\Delta \in \mathbb{Z}_+$ such that the local operations in (2.7c) and (2.7d) can be completed in $\Delta$ unit time for all $i \in \mathcal{N}$ and $k \geq 1$. In the rest, we assume that $\Delta = 1$ as in Assumption 4.

Consider the $k$th iteration of PDA as shown in (2.7). Instead of setting $\mathbf{v}^{k+1}$ to $\mathbf{w}^{k+1}$ as in (2.7b), we propose approximating $\mathbf{w}^{k+1}$ using the inexact averaging operator $\mathcal{R}^k(\cdot) = [\mathcal{R}_i^k(\cdot)]_{i\in\mathcal{N}}$ of Assumption 4 and set $\mathbf{v}^{k+1}$ to this approximation. This way, we can skip the (2.7a) step and avoid explicitly computing $\mathbf{w}^{k+1}$ as in (2.8), which requires using the exact averaging to compute $\mathcal{P}_{\widetilde{\mathcal{C}}}(\cdot)$. More precisely, to obtain an *inexact* variant of (2.7), we replace (2.7b) with the following:

$$\mathbf{v}^{k+1} \leftarrow \gamma\left[\tfrac{1}{\gamma}\mathbf{v}^k + \mathbf{y}^k - \mathcal{R}^k\left(\tfrac{1}{\gamma}\mathbf{v}^k + \mathbf{y}^k\right)\right]. \tag{2.11}$$

Thus, PDA iterations in (2.7), for solving the saddle-point formulation, $\min_{\boldsymbol{\xi},\mathbf{w}} \max_{\mathbf{y}} \mathcal{L}(\boldsymbol{\xi}, \mathbf{w}, \mathbf{y})$, of the distributed resource allocation problem in (1.2), can be computed inexactly, but in a *decentralized* way for a time-varying connectivity network $\{\mathcal{G}^t\}_{t\geq 0}$ provided that $\mathcal{R}^k$ satisfying Assumption 4 exists for $\{\mathcal{G}^t\}_{t\geq 0}$. We call this inexact version of the linearized PDA Algorithm DPDA-D, and the node-specific computations of DPDA-D are displayed in Figure 1 below. Indeed, the iterate sequence $\{\boldsymbol{\xi}^k, \mathbf{v}^k, \mathbf{y}^k\}_{k\geq 0}$ generated by Algorithm DPDA-D is the same sequence generated by the recursion in (2.11), (2.7c), and (2.7d). As emphasized previously, the sequence $\{\mathbf{w}^k\}_{k\geq 0}$ will not be explicitly computed; instead we will use it in the analysis of the inexact algorithm. Next, we discuss the existence of inexact average operators $\mathcal{R}^k$ satisfying Assumption 4 under various assumptions on time-varying network $\{\mathcal{G}^t\}_{t\geq 0}$.

```
**Algorithm DPDA-D** ( $\boldsymbol{\xi}^0, \gamma, \{\tau_i, \kappa_i\}_{i\in\mathcal{N}}$ )
Initialization: $v_i^0 \leftarrow \mathbf{0}, \quad y_i^0 \leftarrow \mathbf{0} \quad i \in \mathcal{N}$
Iteration $k$: $(k \geq 0)$
  1. $v_i^c \leftarrow \frac{1}{\gamma} v_i^k + y_i^k, \qquad v_i^{k+1} \leftarrow \gamma v_i^c - \gamma \mathcal{R}_i^k(\mathbf{v}^c), \quad i \in \mathcal{N}, \quad$ where $\quad \mathbf{v}^c = [v_i^c]_{i\in\mathcal{N}}$
  2. $\xi_i^{k+1} \leftarrow \mathbf{prox}_{\tau_i \rho_i}\Big(\xi_i^k - \tau_i\Big(\nabla f_i(\xi_i^k) - \mathbf{J} g_i(\xi_i^k)^\top y_i^k\Big)\Big), \quad i \in \mathcal{N}$
  3. $y_i^{k+1} \leftarrow \mathcal{P}_{\mathcal{K}^\circ \cap \mathcal{B}_0}\Big(y_i^k - \kappa_i\Big(2 g_i(\xi_i^{k+1}) - g_i(\xi_i^k) + (2 v_i^{k+1} - v_i^k)\Big)\Big), \quad i \in \mathcal{N}$
```

FIG. 1. *Distributed PDA for time-varying $\{\mathcal{G}^t\}$ (DPDA-D).*

### 2.1. Inexact averaging operators.

Let $t_k \in \mathbb{Z}_+$ be the total number of *communication rounds* done before the $k$th iteration of DPDA-D, in Figure 1, and let $q_k \in \mathbb{Z}_+$ be the number of communication rounds to be performed within the $k$th iteration while evaluating $\mathcal{R}^k$. According to Assumption 4, each node $i \in \mathcal{N}$ can finish $\xi_i^{k+1}$ and $y_i^{k+1}$ computations within *one* unit of time, i.e., between two consecutive tics of the clock, $\forall k \geq 0$, and communication rounds occur every time the global clock tics; hence, $\mathcal{G}^t$ represents the connectivity network at the time of $t$th communication round $\forall t \in \mathbb{Z}_+$. Thus, only $\{\mathcal{G}^t\}_{t\in\mathbb{Z}_+}$ among $\{\mathcal{G}^t\}_{t\in\mathbb{R}_+}$ is relevant since the topology of the time-varying network is only pertinent at those times when communication happens among neighboring nodes. For implementation in practice, it is sufficient for each node to count the number of global clock tics since the last update.

DEFINITION 4. *Let $V^t \in \mathbb{R}^{|\mathcal{N}|\times|\mathcal{N}|}$ be a matrix encoding the topology of $\mathcal{G}^t = (\mathcal{N}, \mathcal{E}^t)$ in some way for $t \in \mathbb{Z}_+$. We define $W^{t,s} \triangleq V^t V^{t-1}, \ldots, V^{s+1}$ for any $t, s \in \mathbb{Z}_+$ such that $t \geq s+1$.*

Let $\{\mathcal{G}^t\}$ be a time-varying directed graph; we adopt the information exchange model in [29] satisfying the assumptions stated in Assumption 5.

ASSUMPTION 5. *For all $t \in \mathbb{Z}_+$,* (i) *every $i \in \mathcal{N}$ knows $\mathcal{N}_i^{t,\text{out}}$ and there exists $\zeta \in (0,1)$ such that for $i \in \mathcal{N}$, $V_{ij}^t \geq \zeta$ if $j \in \mathcal{N}_i^{t,\text{in}}$, and $V_{ij}^t = 0$ otherwise.* (ii) *$\mathcal{G}^t$ is $M$-strongly-connected, i.e., there exist an $\mathbb{Z} \ni M \geq 1$ (possibly unknown to nodes) such that the graph with edge set $\mathcal{E}_M^k = \bigcup_{t=kM}^{(k+1)M-1} \mathcal{E}^t$ is strongly connected for $k \in \mathbb{Z}_+$.*

#### 2.1.1. Undirected $\{\mathcal{G}^t\}_{t\in\mathbb{Z}_+}$.

Let $\{\mathcal{G}^t\}$ be a time-varying undirected graph; $\mathcal{N}_i^t$ is defined as in Definition 1, and $d_i^t = |\mathcal{N}_i^t|$ for $i \in \mathcal{N}$. For the undirected case, we assume $\{V^t\}_{t\in\mathbb{Z}_+}$ is *doubly stochastic* and satisfies Assumption 5. For instance, $V^t$ can be set as the Metropolis edge weight matrix [6] corresponding to $\mathcal{G}^t$, i.e., for each $i \in \mathcal{N}$ set $V_{ij}^t = (\max\{d_i^t,\ d_j^t\} + 1)^{-1}$ for $j \in \mathcal{N}_i^t$, $V_{ij}^t = 0$ for $j \notin \mathcal{N}_i^t \cup \{i\}$ and $V_{ii}^t = 1 - \sum_{j\in\mathcal{N}_i^t} V_{ij}^t$. Suppose that there exists $d_{\max}$ such that $d_i^t \leq d_{\max}$ $\forall i \in \mathcal{N}$ and $t \in \mathbb{Z}_+$. Under this assumption, it is trivial to check $\zeta = (d_{\max} + 1)^{-1}$.

For $V^t$ satisfying (i) in Assumption 5, given any $w \in \mathbb{R}^{|\mathcal{N}|}$, the matrix-vector multiplication $V^t w \in \mathbb{R}^{|\mathcal{N}|}$ can be computed in a distributed way, i.e., the $i$th component $(V^t w)_i = \sum_{j\in\mathcal{N}_i\cup\{i\}} V_{ij}^t w_j$ can be computed at node $i \in \mathcal{N}$ requiring only local communication of $i$ with nodes in $\mathcal{N}_i^t$. The next result shows how this distributed operation can be used to approximate the average (also see [31]).

LEMMA 2.1. *Let $\{V^t\}_{t\in\mathbb{Z}_+}$ be a sequence of* doubly stochastic *matrices satisfying Assumption 5. For any $s, t \in \mathbb{Z}_+$ such that $t \geq s$, $\|(W^{t,s} \otimes \mathbf{I}_m)\mathbf{w} - \mathbf{1} \otimes p(\mathbf{w})\| \leq \frac{8}{7}\alpha^{t-s}\|\mathbf{w}\|$ for any $\mathbf{w} = [w_i]_{i\in\mathcal{N}} \in \mathbb{R}^{n_0}$, where $\alpha = (1 - \frac{\zeta}{2N^2})^{1/2M}$.*

*Proof.* The proof immediately follows from [29, Lemma 5]. □

For $\mathbf{w} = [w_i]_{i\in\mathcal{N}} \in \mathbb{R}^{n_0}$ such that $w_i \in \mathbb{R}^m$ for $i \in \mathcal{N}$, define

$$\mathcal{R}^k(\mathbf{w}) \triangleq \mathcal{P}_{\mathcal{B}}\left((W^{t_k+q_k,t_k} \otimes \mathbf{I}_m)\ \mathbf{w}\right) \tag{2.12}$$

to approximate $\mathcal{P}_{\widetilde{\mathcal{C}}}(\cdot)$ in (2.8). Note that $\mathcal{R}^k(\cdot)$ can be computed in a *distributed fashion* requiring $q_k$ communications with the neighbors for each node. In particular, components of $\mathcal{R}^k(\mathbf{w})$ can be computed at each node as $\mathcal{R}^k(\mathbf{w}) = [\mathcal{R}^k_i(\mathbf{w})]_{i\in\mathcal{N}}$ such that $\mathcal{R}^k_i(\mathbf{w}) \triangleq \mathcal{P}_{\mathcal{B}_0}(\sum_{j\in\mathcal{N}_i\cup\{i\}} W^{t_k+q_k,t_k}_{ij} w_j)$. Moreover, the approximation error, $\mathcal{R}^k(\mathbf{w}) - \mathcal{P}_{\widetilde{\mathcal{C}}}(\mathbf{w})$, for any $\mathbf{w}$ can be bounded as in (2.10) using the nonexpansivity of $\mathcal{P}_{\mathcal{B}}$ and Lemma 2.1. More precisely, $\mathcal{R}^k$ defined in (2.12) satisfies Assumption 4.

**2.1.2. Directed $\{\mathcal{G}^t\}_{t\in\mathbb{Z}_+}$.** Let $\{\mathcal{G}^t\}$ be a time-varying directed graph, and $\mathcal{N}_i^{t,\text{in}}$, $\mathcal{N}_i^{t,\text{out}}$ be defined as in Definition 1 for $i \in \mathcal{N}$. Recall $d_i^t = |\mathcal{N}_i^{t,\text{out}}|-1$. Since the definition of $\widetilde{\mathcal{C}}$ in (2.3) does not depend on the topology of the network, using the push-sum protocol [24] within DPDA-D, one can also handle time-varying *directed* communication networks. Indeed, given any $\mathbf{w} = [w_i]_{i\in\mathcal{N}}$, nodes can inexactly compute $\mathcal{P}_{\widetilde{\mathcal{C}}}(\mathbf{w})$ in a distributed fashion with increasing approximation quality; consider the weight-matrix sequence $\{V^t\}_{t\in\mathbb{Z}_+}$: for any $t \geq 0$,

$$V^t_{ij} = \frac{1}{d^t_j + 1} \ \text{ if } \ j \in \mathcal{N}_i^{t,\text{in}}; \quad V^t_{ij} = 0 \ \text{ if } \ j \notin \mathcal{N}_i^{t,\text{in}}, \quad i \in \mathcal{N}. \tag{2.13}$$

For $\mathbf{w} = [w_i]_{i\in\mathcal{N}} \in \mathbb{R}^{n_0}$ such that $w_i \in \mathbb{R}^m$ for $i \in \mathcal{N}$, define

$$\mathcal{R}^k(\mathbf{w}) \triangleq \mathcal{P}_{\mathcal{B}}\left(\operatorname{diag}(W^{t_k+q_k,t_k}\mathbf{1}_N \otimes \mathbf{I}_m)^{-1}\ (W^{t_k+q_k,t_k} \otimes \mathbf{I}_m)\ \mathbf{w}\right) \tag{2.14}$$

to approximate $\mathcal{P}_{\widetilde{\mathcal{C}}}(\cdot)$ in (2.8). $\mathcal{R}^k(\cdot)$ can be computed in a *distributed fashion* requiring $q_k$ communication rounds and is a compact representation of push-sum operation.

LEMMA 2.2. *Consider $\mathcal{R}^k$ defined in* (2.14) *for $k \geq 0$. Assuming $\{\mathcal{G}^t\}_{t\in\mathbb{Z}_+}$ is uniformly strongly connected ($M$-strongly connected),* (2.10) *holds for some $\Gamma > 0$ and $\alpha \in (0,1)$ such that $\Gamma \leq 8N^{NM}$ and $\alpha \leq \left(1 - \frac{1}{N^{NM}}\right)^{\frac{1}{M}}$.*

*Proof.* The result follows from the proof Lemma 1 in [27]. □

**3. Convergence of algorithm DPDA-D.** Define $\bar{C}_g \triangleq \sum_{i\in\mathcal{N}} C_{g_i}/N$ and $\bar{R}_x \triangleq \max\{\left\|\boldsymbol{\xi}^* - \boldsymbol{\xi}^0\right\|_{\mathbf{L}_g}, \left\|\boldsymbol{\xi}^* - \boldsymbol{\xi}^0\right\|_{\mathbf{L}'}\}/\sqrt{N}$, where $\mathbf{L}' \triangleq \operatorname{diag}([(1 + L_{f_i} + C_{g_i})\mathbf{I}_{n_i}]_{i\in\mathcal{N}})$ and $\mathbf{L}_g \triangleq \operatorname{diag}([(L_{g_i})\mathbf{I}_{n_i}]_{i\in\mathcal{N}})$.

THEOREM 3.1. *Suppose Assumptions* 1, 2, 3, *and* 4 *hold. For any $\gamma > 0$, let the primal-dual step-sizes $\{\tau_i, \kappa_i\}_{i\in\mathcal{N}}$ be chosen such that for some $\beta > 0$,*

$$\tau_i = (\max\{1, L_{f_i} + \beta L_{g_i}\} + C_{g_i})^{-1}, \quad \kappa_i = \left(C_{g_i} + \frac{5\gamma}{2}\right)^{-1} \quad \forall\ i \in \mathcal{N}. \tag{3.1}$$

*Given $B \in (0,\infty]$, starting from $\mathbf{v}^0 = \mathbf{y}^0 = \mathbf{0}$ and an arbitrary $\boldsymbol{\xi}^0$, let $\{(\boldsymbol{\xi}^k, \mathbf{v}^k)\}_{k\geq 0}$ be the primal, and $\{\mathbf{y}^k\}_{k\geq 0}$ be the dual iterate sequence generated by Algorithm DPDA-D, displayed in Figure* 1, *using $q_k \in \mathbb{Z}_+$ communication rounds for the $k$th iteration such that $C_0 \triangleq \sum_{k=0}^{\infty} \alpha^{q_k}(k+1) < \infty$. For any $\gamma > 0$, if $\beta > 0$ is chosen as discussed below, then $\{(\boldsymbol{\xi}^k, \mathbf{y}^k)\}_{k\geq 0}$ converges to $(\boldsymbol{\xi}^*, \mathbf{y}^*)$ such that $\mathbf{y}^* = \mathbf{1} \otimes y^*$ and $(\boldsymbol{\xi}^*, y^*)$*

*is an optimal primal-dual solution to* (1.2). *Moreover, both infeasibility,* $F(\bar{\boldsymbol{\xi}}^K, \bar{\mathbf{y}}^K)$, *and suboptimality,* $|\varphi(\bar{\boldsymbol{\xi}}^K) - \varphi(\boldsymbol{\xi}^*)|$, *are* $\mathcal{O}(1/K)$, *i.e.,* $\forall K \geq 1$,

$$F(\bar{\boldsymbol{\xi}}^K, \bar{\mathbf{y}}^K) \triangleq d_{\mathcal{C}}(\bar{\mathbf{y}}^K) + \|y^*\| \, d_{\mathcal{K}}\left(-g(\bar{\boldsymbol{\xi}}^K)\right) \leq \frac{\Lambda(\gamma,\beta)}{K}, \tag{3.2}$$

$$0 \leq \varphi(\bar{\boldsymbol{\xi}}^K) - \varphi(\boldsymbol{\xi}^*) + \|y^*\| \, d_{\mathcal{K}}\left(-g(\bar{\boldsymbol{\xi}}^K)\right) \leq \frac{\Lambda(\gamma,\beta)}{K} - F(\bar{\boldsymbol{\xi}}^K, \bar{\mathbf{y}}^K) \tag{3.3}$$

*for some* $\Lambda(\gamma,\beta) \in \mathbb{R}_+$, *where* $\bar{\boldsymbol{\xi}}^K = \frac{1}{K}\sum_{k=1}^K \boldsymbol{\xi}^k$ *and* $\bar{\mathbf{y}}^K = \frac{1}{K}\sum_{k=1}^K \mathbf{y}^k$ *for* $K \geq 1$.

Case 1. *If a dual bound is known, i.e.,* $B < \infty$, *then* (3.2) *and* (3.3) *hold for* $\beta = 2B$; *moreover, setting the free parameter* $\gamma = (N^{3/2}\Gamma C_0 B)^{-1}$ *gives*

$$\Lambda(\gamma,\beta) = \mathcal{O}\Big(NB(\bar{R}_x^2 + \bar{C}_g B) + N^{3/2}\Gamma C_0 B\Big). \tag{3.4}$$

Case 2. *If the dual bound does not exist, set* $B = \infty$ *within DPDA-D. Assuming* $q_k \geq \log_{1/\alpha}(24N\Gamma(k+1))$ *for* $k \geq 0$, *there exists* $\bar{\beta} > 0$ *such that* (3.2) *and* (3.3) *hold* $\forall \beta \geq \bar{\beta}$; *moreover, selecting* $\gamma = N^{\frac{3}{2}}\Gamma C_0 \bar{R}_x^2$ *gives* $\Lambda(\gamma,\beta) = \mathcal{O}(N^{\frac{9}{2}}\Gamma^3 C_0^3 \bar{R}_x^2 \max\{1, \|y^*\|^2\})$. *Finally, when* $g_i$ *is affine for* $i \in \mathcal{N}$ *and* $\{\tau_i\}$ *are independent of* $\beta$, $\gamma = (N^{\frac{3}{2}}\Gamma C_0)^{-1}$ *leads to* $\Lambda(\gamma,\beta) = \mathcal{O}(N^3\Gamma^2 C_0^2(\bar{R}_x^2 + \bar{C}_g \max\{1, \|y^*\|^2\}))$.

*Remark* 3.1. We assume agents know $q_k$ as a function of $k$ at the initialization; hence, synchronicity can be achieved among nodes if simply each node counts the number of times the global clock tics, where at each tic one communication round occurs according to Assumption 4.

*Remark* 3.2. Suppose we are given $(0,1) \ni \bar{\alpha} \geq \alpha$. For any $c > 0$, choosing $q_k = \lceil (2+c)\log_{1/\bar{\alpha}}(k+1) \rceil$ for $k \geq 0$ satisfies the condition in Theorem 3.1, i.e., $C_0 = \sum_{k=0}^\infty \alpha^{q_k}(k+1) \leq \frac{1}{c} + 1$. Moreover, this choice of $\{q_k\}_{k\in\mathbb{Z}_+}$ implies that the total number of communication rounds right before the $K$th iteration is equal to $t_K = \sum_{k=0}^{K-1} q_k = (2+c)[(K-1)\log_{1/\bar{\alpha}}(K) + \log_{1/\bar{\alpha}}(e)]$, where $e$ is Euler's number.

Corollary 3.2. *Under the premise of Theorem* 3.1, *let* $\{\mathcal{G}^t\}$ *be an undirected time-varying graph and* $\{q_k\}$ *be as in Remark* 3.2 *with* $(0,1) \ni \bar{\alpha} = \iota\alpha$ *for some* $\iota > 1$. *Let* $Q(\epsilon)$ *be the total number of communications needed to compute an* $\epsilon$*-optimal and* $\epsilon$*-feasible solution* $(\boldsymbol{\xi}^\epsilon, \mathbf{y}^\epsilon)$ *for* $\gamma = 1/\mathcal{O}(\sqrt{N})$, *i.e.,* $F(\boldsymbol{\xi}^\epsilon, \mathbf{y}^\epsilon) < \epsilon$ *and* $|\varphi(\boldsymbol{\xi}^\epsilon) - \varphi(\boldsymbol{\xi}^*)| < \epsilon$. *If a dual bound* $B < \infty$ *is known, then* $Q(\epsilon) = \mathcal{O}(\frac{N^4}{\epsilon}\log(\frac{N}{\epsilon}))$. *If a Slater point does not exist, i.e.,* $B = \infty$, *then* $Q(\epsilon) = \mathcal{O}(\frac{N^{4.5}}{\epsilon}\log(\frac{N^{1.5}}{\epsilon}))$; *moreover,* $Q(\epsilon) = \mathcal{O}(\frac{N^4}{\epsilon}\log(\frac{N}{\epsilon}))$ *is achieved when* $g_i$ *is an affine function for* $i \in \mathcal{N}$.

*Proof.* Theorem 3.1 implies that $(\boldsymbol{\xi}^\epsilon, \mathbf{y}^\epsilon)$ can be computed in $K^\epsilon = \Lambda(\gamma,\beta)/\epsilon$ DPDA-D iterations, which requires $t_{K^\epsilon} = \mathcal{O}(K^\epsilon \log(K^\epsilon)/\log(\frac{1}{\alpha}))$ communications in total (see Remark 3.2). Lemma 2.1 implies that $\Gamma = 1/N$; hence, setting $\gamma$ as described in Theorem 3.1, we bound $\Lambda(\gamma,\beta)$ with $\mathcal{O}(N)$ for Case 1, $\mathcal{O}(N^{1.5})$ for Case 2 in general, and with $\mathcal{O}(N)$ when $g_i$'s are linear. Thus, the result follows from $\log(\frac{1}{\alpha}) \geq \zeta/N^2$, where $\zeta$ can be as small as $\mathcal{O}(1/N)$. ∎

Note that when $\{\mathcal{G}^t\}$ is a general time-varying directed graph, we employ the push-sum protocol with $\Gamma = N^{NM}$ (see Lemma 2.2), which leads to exponential $\mathcal{O}(1)$ bounds, e.g., $\Lambda(\gamma,\beta) = \mathcal{O}(N^{NM+\frac{3}{2}}B)$ for Case 1. To the best of our knowledge, polynomial bounds for directed graphs in $N$ is still an open question [28]. That said, setting $\{q_k\}$ as in Remark 3.2, DPDA-D can compute an $\epsilon$-solution in $\mathcal{O}(\frac{1}{\epsilon}\log(\frac{1}{\epsilon}))$ communications even for general directed graphs and choosing $q_k = (2+c)\log_{1/\bar{\alpha}}(k+1)$ in

Case 1 leads to $\mathcal{O}(1)$ constant $N^{2NM+1.5}$, which is better than the $\mathcal{O}(\frac{1}{\epsilon^2}\log^2(\frac{1}{\epsilon}))$ result in [19, 27] with the $\mathcal{O}(1)$ constant of $N^{2NM+2}$. The method in [19] has $\log(k)/\sqrt{k}$ rate and requires exact minimization of convex $f_i$ over compact $\mathcal{X}_i$ at each iteration. The method in [27] can be used to solve the dual of (1.2) when $\rho_i$ is the indicator function of some compact convex set $\mathcal{X}_i$ and $f_i$ is convex for $i \in \mathcal{N}$ but the subproblem that needs to be solved at each iteration is fairly complicated as in [19].

*Remark* 3.3. Since $\sum_{k=1}^{\infty} \alpha^{\sqrt[p]{k}} k < \infty$ for any $p \geq 1$, if one chooses $q_k = \sqrt[p]{k}$ for $k \geq 1$, then $t_K = \sum_{k=0}^{K-1} q_k = \mathcal{O}(K^{1+1/p})$. This choice of $\{q_k\}_{k\in\mathbb{Z}_+}$, unlike the one in Remark 3.2, is independent of the parameter $\alpha \in (0,1)$ but leads to a larger $C_0 = \sum_{k=0}^{\infty} \alpha^{q_k}(k+1) = \mathcal{O}(\alpha/\log^{2p}(\alpha))$ for $\alpha \in (1/e, 1)$. On the other hand, a priori running DPDA-D, a practical way to estimate $\alpha \in (0,1)$ is to run an average of consensus iterations with a random initialization until iterates stagnate around the average; this leads to a rate coefficient $\alpha_i$ for $i \in \mathcal{N}$. Next, nodes can do a max consensus to compute $\bar{\alpha} = \max_{i\in\mathcal{N}} \alpha_i$ and use it to set $q_k = (2+c)\log_{1/\bar{\alpha}}(k+1)$.

*Remark* 3.4. Suppose the dual bound is not available. If $q_k = (2+c)\log_{1/\bar{\alpha}}(k+1)$ for some $c > 0$ and $(0,1) \ni \bar{\alpha} \geq \alpha$, then $q_k \geq \log_{1/\alpha}(24N\Gamma(k+1))$ $\forall k \geq \tilde{K} \triangleq \lceil (24N\Gamma)^{1/(1+c)} \rceil$. If $q_k = \sqrt[p]{k}$ for some $p \geq 1$, then $q_k \geq \log_{1/\alpha}(24N\Gamma(k+1))$ $\forall k \geq \tilde{K} = \lceil (\log_{1/\alpha}(24N\Gamma) + p\log_{1/\alpha} p)^p \rceil$. Hence, the rate results of Theorem 3.1 will hold after the transient period of $\tilde{K}$ iterations.

**3.1. Auxiliary results to prove Theorem 3.1.** Let $\{\boldsymbol{\xi}^k, \mathbf{v}^k, \mathbf{y}^k\}_{k\geq 0}$ be the iterate sequence generated by DPDA-D as shown in Figure 1 and $\{\mathbf{w}^k\}_{k\geq 0}$ be the auxiliary sequence where $\mathbf{w}^k$ is given in (2.8) for $k \geq 1$ and we set $\mathbf{w}^0 \triangleq \mathbf{v}^0 = \mathbf{0}$. We first define the error sequence $\{\mathbf{e}^k\}_{k\geq 0}$: let $\mathbf{e}^k \triangleq (\mathbf{v}^k - \mathbf{w}^k)/\gamma$ $\forall k \geq 0$; hence, $\mathbf{e}^0 = \mathbf{e}^1 = \mathbf{0}$ and for $k \geq 0$, we have

$$\mathbf{e}^{k+1} = \mathcal{P}_{\widetilde{\mathcal{C}}}\left(\tfrac{1}{\gamma}\mathbf{v}^k + \mathbf{y}^k\right) - \mathcal{R}^k\left(\tfrac{1}{\gamma}\mathbf{v}^k + \mathbf{y}^k\right). \tag{3.5}$$

In order to prove Theorem 3.1, we first prove Lemma 3.3, which helps us to bound $\mathcal{L}(\boldsymbol{\xi}^k, \mathbf{v}^k, \mathbf{y}) - \mathcal{L}(\boldsymbol{\xi}, \mathbf{v}, \mathbf{y}^k)$ for any given $(\boldsymbol{\xi}, \mathbf{v}, \mathbf{y}) \in \mathcal{Z}$ and $k \geq 1$, where $\mathcal{L}$ is defined in (2.5); and then we provide a few other technical results which will be used together with Lemma 3.3 to show the asymptotic convergence of $\{\boldsymbol{\xi}^k, \mathbf{v}^k, \mathbf{y}^k\}$ in Theorem 3.1.

Definition 5. *Let $\mathbf{D}_\gamma$ and $\mathbf{D}_\kappa$ be the diagonal matrices given in Definition 2. Define a diagonal matrix $\mathbf{C} \triangleq \operatorname{diag}([C_{g_i}]_{i\in\mathcal{N}})$, and $H \triangleq [\mathbf{C} \;\; \mathbf{I}_N]$. Given some $\beta > 0$, define the symmetric matrix*

$$\bar{\mathbf{Q}}(\beta) \triangleq \begin{bmatrix} \bar{\mathbf{D}}(\beta) & -H^\top \\ -H & \bar{\mathbf{D}}_\kappa \end{bmatrix}, \qquad \textit{where} \qquad \bar{\mathbf{D}}(\beta) \triangleq \begin{bmatrix} \bar{\mathbf{D}}_\tau(\beta) & \mathbf{0} \\ \mathbf{0} & \frac{1}{\gamma}\mathbf{I}_N \end{bmatrix},$$

$\bar{\mathbf{D}}_\tau(\beta) \triangleq \operatorname{diag}([\frac{1}{\tau_i} - \max\{1, L_{f_i} + \beta L_{g_i}\}]_{i\in\mathcal{N}})$ *and* $\bar{\mathbf{D}}_\kappa \triangleq \operatorname{diag}([\frac{1}{\kappa_i}]_{i\in\mathcal{N}})$. *Let* $\mathbf{u} : \mathcal{Z}\times\mathcal{Z} \to \mathbb{R}^{3N}$ *such that* $\mathbf{u}(\mathbf{z}, \bar{\mathbf{z}}) \triangleq \left[ [\|\xi_i - \bar{\xi}_i\|]_{i\in\mathcal{N}}^\top \;\; [\|w_i - \bar{w}_i\|]_{i\in\mathcal{N}}^\top \;\; [\|y_i - \bar{y}_i\|]_{i\in\mathcal{N}}^\top \right]^\top \in \mathbb{R}^{3N}$.

Lemma 3.3. *Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be the spaces defined in Definition 2. Let $\{\tilde{\mathbf{x}}^k\}_{k\geq 0} \subset \mathcal{X}$ be the primal and $\{\mathbf{y}^k\}_{k\geq 0} \subset \mathcal{Y}$ be the dual iterate sequences generated by Algorithm DPDA-D in Figure 1, using some positive step-sizes, $\{\tau_i, \kappa_i\}_{i\in\mathcal{N}}$ and $\gamma$, and initializing from an arbitrary $\boldsymbol{\xi}^0$ and $\mathbf{v}^0 = \mathbf{y}^0 = \mathbf{0}$, where $\tilde{\mathbf{x}}^k = [\boldsymbol{\xi}^{k\top} \;\, \mathbf{v}^{k\top}]^\top$ for $k \geq 0$. Define $\{\mathbf{x}^k\}$ and $\{\mathbf{z}^k\}$ such that $\mathbf{x}^k = [\boldsymbol{\xi}^{k\top} \;\, \mathbf{w}^{k\top}]^\top \in \mathcal{X}$ and $\mathbf{z}^k = [\mathbf{x}^{k\top}\mathbf{y}^{k\top}]^\top \in \mathcal{Z}$*

*for $k \geq 0$. Let $\{\beta_k\}_{k\geq 0}$ be such that $\beta_k \geq \max_{i\in\mathcal{N}} \left\|y_i^k\right\|$ for $k \geq 0$, and then for any $\mathbf{x} = [\boldsymbol{\xi}^\top \ \mathbf{w}^\top]^\top \in \mathcal{X}$, and $\mathbf{y} \in \mathcal{Y}$, $\{\mathbf{z}^k\}_{k\geq 0} \subset \mathcal{Z}$ satisfies*

$$\begin{aligned} &\mathcal{L}(\mathbf{x}^{k+1}, \mathbf{y}) - \mathcal{L}(\mathbf{x}, \mathbf{y}^{k+1}) \\ &\quad\leq \left[D_x(\mathbf{x}, \mathbf{x}^k) + D_y(\mathbf{y}, \mathbf{y}^k) - \left\langle T(\mathbf{x}) - T(\mathbf{x}^k),\ \mathbf{y} - \mathbf{y}^k \right\rangle\right] \\ &\quad\quad - \left[D_x(\mathbf{x}, \mathbf{x}^{k+1}) + D_y(\mathbf{y}, \mathbf{y}^{k+1}) - \left\langle T(\mathbf{x}) - T(\mathbf{x}^{k+1}),\ \mathbf{y} - \mathbf{y}^{k+1} \right\rangle\right] \\ &\quad\quad + E^{k+1}(\mathbf{z}) - \tfrac{1}{2}\mathbf{u}(\mathbf{z}^{k+1}, \mathbf{z}^k)^\top \bar{\mathbf{Q}}(\beta_k)\ \mathbf{u}(\mathbf{z}^{k+1}, \mathbf{z}^k) \quad \forall\ k \geq 0, \end{aligned} \tag{3.6}$$

*where $\mathbf{u}(\cdot,\cdot)$ is given in Definition 5, $\mathbf{z}^k = [\mathbf{x}^{k\top}\ \mathbf{y}^{k\top}]^\top$, $D_x$ and $D_y$ are Bregman functions in Definition 2, $T(\cdot)$ is given in Definition 3, $E^{k+1}(\mathbf{z}) \triangleq \left\|\mathbf{e}^k\right\| \left\|\mathbf{w} - \mathbf{w}^{k+1}\right\| + \gamma \left\|2\mathbf{e}^{k+1} - \mathbf{e}^k\right\| \left\|\mathbf{y} - \mathbf{y}^{k+1}\right\|$, and $\mathbf{e}^k \triangleq (\mathbf{v}^k - \mathbf{w}^k)/\gamma$ for $k \geq 0$,*

*Proof.* Given $\{\mathbf{v}^k\}_{k\geq 0}$ generated as in Figure 1, let the $\{\mathbf{w}^k\}_{k\geq 0}$ sequence be defined according to (2.7a)—recall that the $\{\mathbf{w}^k\}_{k\geq 0}$ sequence is never actually computed in practice; this sequence will help us in our analysis of DPDA-D.

Let $\Phi$, $h$, and possibly *nonlinear* map $T(\cdot)$ be as given in Definition 3; hence, our objective is to compute a saddle point for $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \Phi(\mathbf{x}) + \langle T(\mathbf{x}), \mathbf{y}\rangle - h(\mathbf{y})$ to solve (1.2). Using this notation and the fact that $\mathbf{v}^k = \mathbf{w}^k + \gamma\mathbf{e}^k$ for $k \geq 0$, we can represent $\{\boldsymbol{\xi}^k\}$, $\{\mathbf{w}^k\}$, and $\{\mathbf{y}^k\}$ sequences in a more compact form as follows:

$$\mathbf{x}^{k+1} = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmin}}\ \rho(\mathbf{x}) + f(\mathbf{x}^k) + \left\langle \nabla f(\mathbf{x}^k) + \mathbf{J}T(\mathbf{x}^k)^\top \mathbf{y}^k + U\mathbf{e}^k,\ \mathbf{x} - \mathbf{x}^k \right\rangle + D_x(\mathbf{x}, \mathbf{x}^k), \tag{3.7a}$$

$$\mathbf{y}^{k+1} = \underset{\mathbf{y}\in\mathcal{Y}}{\operatorname{argmin}}\ h(\mathbf{y}) - \left\langle 2T(\mathbf{x}^{k+1}) - T(\mathbf{x}^k) - \gamma(2\mathbf{e}^{k+1} - \mathbf{e}^k), \mathbf{y} \right\rangle + D_y(\mathbf{y}, \mathbf{y}^k), \tag{3.7b}$$

where $U = [\mathbf{0}\ \mathbf{I}_{n_0}]^\top \in \mathbb{R}^{(n+n_0)\times n_0}$ and $\{\mathbf{v}^k\}$ is updated according to (2.11). Let $S_1^k(\mathbf{w}) \triangleq \left\langle \mathbf{e}^k,\ \mathbf{w} - \mathbf{w}^{k+1} \right\rangle$. Since $\rho$ is a proper, closed, convex function and $D_x$ is a Bregman function, Property 1 in [39] applied to (3.7a) implies that for any $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned} &\rho(\mathbf{x}) - \rho(\mathbf{x}^{k+1}) + \left\langle \nabla f(\mathbf{x}^k) + \mathbf{J}T(\mathbf{x}^k)^\top \mathbf{y}^k,\ \mathbf{x} - \mathbf{x}^{k+1} \right\rangle \\ &\qquad\qquad \geq D_x(\mathbf{x}, \mathbf{x}^{k+1}) - D_x(\mathbf{x}, \mathbf{x}^k) + D_x(\mathbf{x}^{k+1}, \mathbf{x}^k) - S_1^k(\mathbf{w}). \end{aligned} \tag{3.8}$$

Moreover, the convexity of $f_i$ and Lipschitz continuity of $\nabla f_i$ implies that for any $\xi_i \in \mathbb{R}^{n_i}$,

$$\begin{aligned} f_i(\xi_i) \geq f_i(\xi_i^k) + \left\langle \nabla f_i(\xi_i^k), \xi_i - \xi_i^k \right\rangle &\geq f_i(\xi_i^{k+1}) \\ &+ \left\langle \nabla f_i(\xi_i^k), \xi_i - \xi_i^{k+1} \right\rangle - \frac{L_{f_i}}{2}\|\xi_i^{k+1} - \xi_i^k\|^2. \end{aligned}$$

Similarly, since $-y_i^k \in \mathcal{K}^*$, the $\mathcal{K}$-convexity of $g_i$ and Lipschitz continuity of $\mathbf{J}g_i$ imply

$$\begin{aligned} -\left\langle g_i(\xi_i),\ y_i^k \right\rangle &\geq -\left\langle g_i(\xi_i^k),\ y_i^k \right\rangle - \left\langle \mathbf{J}g_i(\xi_i^k)^\top y_i^k,\ \xi_i - \xi_i^k \right\rangle \\ &\geq -\left\langle g_i(\xi_i^{k+1}),\ y_i^k \right\rangle - \left\langle \mathbf{J}g_i(\xi_i^k)^\top y_i^k,\ \xi_i - \xi_i^{k+1} \right\rangle - \frac{\beta_k L_{g_i}}{2}\left\|\xi_i^{k+1} - \xi_i^k\right\|^2. \end{aligned}$$

Summing the last two inequalities first for each $i$, then summing over $i \in \mathcal{N}$, and combining the sum with (3.8), we get

$$\begin{aligned}\Phi(\mathbf{x}) - \Phi(\mathbf{x}^{k+1}) + \left\langle T(\mathbf{x}) - T(\mathbf{x}^{k+1}),\ \mathbf{y}^k\right\rangle \\ \geq D_x(\mathbf{x}, \mathbf{x}^{k+1}) - D_x(\mathbf{x}, \mathbf{x}^k) + \tfrac{1}{2}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2_{\tilde{\mathbf{D}}^k} - S_1^k(\mathbf{w}),\end{aligned} \tag{3.9}$$

where

$$\tilde{\mathbf{D}}^k \triangleq \begin{bmatrix}\tilde{\mathbf{D}}^k_\tau & \mathbf{0}\\ \mathbf{0} & \mathbf{D}_\gamma\end{bmatrix}$$

and $\tilde{\mathbf{D}}^k_\tau \triangleq \mathrm{diag}([(\frac{1}{\tau_i} - (L_{f_i} + \beta_k L_{g_i}))\mathbf{I}_{n_i}]_{i\in\mathcal{N}})$.

Finally, since $h$ is a proper, closed, convex function and $D_y$ is a Bregman function, Property 1 in [39] applied to (3.7b) implies

$$\begin{aligned}h(\mathbf{y}) - h(\mathbf{y}^{k+1}) - \left\langle 2T(\mathbf{x}^{k+1}) - T(\mathbf{x}^k),\ \mathbf{y} - \mathbf{y}^{k+1}\right\rangle \\ \geq D_y(\mathbf{y}, \mathbf{y}^{k+1}) - D_y(\mathbf{y}, \mathbf{y}^k) + \tfrac{1}{2}\left\|\mathbf{y}^{k+1} - \mathbf{y}^k\right\|^2_{\mathbf{D}_\kappa} - S_2^k(\mathbf{y}),\end{aligned} \tag{3.10}$$

where $S_2^k(\mathbf{y}) = \gamma\left\langle(2\mathbf{e}^{k+1} - \mathbf{e}^k),\ \mathbf{y} - \mathbf{y}^{k+1}\right\rangle$. Summing (3.9) and (3.10) and rearranging the terms yields

$$\begin{aligned}\mathcal{L}(\mathbf{x}^{k+1}, \mathbf{y}) - \mathcal{L}(\mathbf{x}, \mathbf{y}^{k+1}) \leq & S^k(\mathbf{z}) + \Big[D_x(\mathbf{x}, \mathbf{x}^k) + D_y(\mathbf{y}, \mathbf{y}^k) - \Big\langle T(\mathbf{x}) - T(\mathbf{x}^k),\ \mathbf{y} - \mathbf{y}^k\Big\rangle\Big] \\ & - \Big[D_x(\mathbf{x}, \mathbf{x}^{k+1}) + D_y(\mathbf{y}, \mathbf{y}^{k+1}) - \Big\langle T(\mathbf{x}) - T(\mathbf{x}^{k+1}),\ \mathbf{y} - \mathbf{y}^{k+1}\Big\rangle\Big],\end{aligned}$$

$$S^k(\mathbf{z}) \triangleq S_1^k(\mathbf{w}) + S_2^k(\mathbf{y}) + \Big\langle T(\mathbf{x}^{k+1}) - T(\mathbf{x}^k),\ \mathbf{y}^{k+1} - \mathbf{y}^k\Big\rangle - \frac{1}{2}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2_{\tilde{\mathbf{D}}^k} - \frac{1}{2}\left\|\mathbf{y}^{k+1} - \mathbf{y}^k\right\|^2_{\mathbf{D}_\kappa}.$$

Using the Cauchy–Schwartz inequality and Lipschitz continuity of $g_i$ for all $i \in \mathcal{N}$, one can bound $S^k(\mathbf{z})$ as follows:

$$\begin{aligned}S^k(\mathbf{z}) \leq & \|\mathbf{e}^k\|\|\mathbf{w} - \mathbf{w}^{k+1}\| \\ & + \gamma\|2\mathbf{e}^{k+1} - \mathbf{e}^k\|\|\mathbf{y} - \mathbf{y}^{k+1}\| - \tfrac{1}{2}\mathbf{u}(\mathbf{z}^{k+1}, \mathbf{z}^k)^\top \bar{\mathbf{Q}}(\beta_k)\ \mathbf{u}(\mathbf{z}^{k+1}, \mathbf{z}^k)\end{aligned}$$

$\forall \mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{Y}$, and $k \geq 0$. □

Given some $\beta > 0$, the next lemma gives a sufficient condition on the local stepsizes for $\bar{\mathbf{Q}}(\beta)$ to be positive (semi)-definite.

LEMMA 3.4. *Consider $\bar{\mathbf{Q}}(\beta)$ given in Definition 5 for some $\beta > 0$. If positive $\{\tau_i, \kappa_i\}_{i\in\mathcal{N}}$ and $\gamma$ satisfy $\tau_i \leq \frac{1}{\max\{1, L_{f_i} + \beta L_{g_i}\}}$, $\kappa_i \leq \frac{1}{\gamma}$ and $(\frac{1}{\tau_i} - \max\{1, L_{f_i} + \beta L_{g_i}\})(\frac{1}{\kappa_i} - \gamma) > C_{g_i}^2$ $\forall i \in \mathcal{N}$, then $\bar{\mathbf{Q}}(\beta) \succ \mathbf{0}$. Moreover, $\bar{\mathbf{Q}}(\beta) \succeq \mathbf{0}$ if the strict inequalities in the last condition are relaxed to $\geq$-relation for some $i \in \mathcal{N}$.*

*Proof.* Given a permutation matrix

$$\mathbf{P} \triangleq \begin{bmatrix}\mathbf{I}_N & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}_N\\ \mathbf{0} & \mathbf{I}_N & \mathbf{0}\end{bmatrix},$$

$\bar{\mathbf{Q}}(\beta) \succ \mathbf{0}$ is equivalent to $\mathbf{P}\bar{\mathbf{Q}}(\beta)\mathbf{P}^{-1} \succ \mathbf{0}$. Since $\gamma > 0$, the Schur complement condition implies

$$\mathbf{P}\bar{\mathbf{Q}}(\beta)\mathbf{P}^{-1} = \begin{bmatrix}\bar{\mathbf{D}}_\tau(\beta) & -\mathbf{C} & \mathbf{0}\\ -\mathbf{C} & \bar{\mathbf{D}}_\kappa & -\mathbf{I}_N\\ \mathbf{0} & -\mathbf{I}_N & \frac{1}{\gamma}\mathbf{I}_N\end{bmatrix} \succ \mathbf{0} \ \Leftrightarrow\ \begin{bmatrix}\bar{\mathbf{D}}_\tau(\beta) & -\mathbf{C}\\ -\mathbf{C} & \bar{\mathbf{D}}_\kappa\end{bmatrix} - \gamma\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_N\end{bmatrix} \succ \mathbf{0}. \tag{3.11}$$

Note $\bar{\mathbf{D}}_\tau(\beta) \succ 0$; hence, using the Schur complement again, one can conclude that the condition on the right-hand side of (3.11) holds if and only if $\bar{\mathbf{D}}_\kappa - \gamma \mathbf{I}_N - \mathbf{C}\bar{\mathbf{D}}_\tau(\beta)^{-1}\mathbf{C} \succ \mathbf{0}$, and equivalently $(\frac{1}{\kappa_i} - \gamma) - (\frac{1}{\tau_i} - \max\{1, L_{f_i} + \beta L_{g_i}\})^{-1} C_{g_i}^2 > 0$ $\forall i \in \mathcal{N}$. Hence, the conditions in Lemma 3.4 are both necessary and sufficient for $\bar{\mathbf{Q}}(\beta) \succ \mathbf{0}$. If the strict inequalities in the last condition are relaxed to include equality for some $i \in \mathcal{N}$, then it is sufficient for $\bar{\mathbf{Q}}(\beta) \succeq \mathbf{0}$. □

Note that if $\{\mathbf{y}^k\} \subseteq \mathcal{B}$, then we can set $\beta^k = 2B$ $\forall k \geq 0$; hence, Lemma 3.4 implies that if the local step-size condition in (3.1) holds (possibly with equality for some $i \in \mathcal{N}$), then $\bar{\mathbf{Q}}(\beta^k)$ in (3.6) is positive (semi)-definite $\forall k \geq 0$, which helps to simplify the analysis of Theorem 3.1.

**3.2. Proof of Theorem 3.1.** Using the two technical lemmas in the appendix, we are ready to prove Theorem 3.1. The proof is divided into three subsections, where we first show that the dual iterate sequence $\{\mathbf{y}^k\}$ stays bounded even if a dual bound is not provided, i.e., $B = \infty$; second, we prove the convergence of the iterate sequences; finally, we provide rate statements for the infeasibility and suboptimality.

Under Assumption 3, a saddle point $(\boldsymbol{\xi}^*, \mathbf{w}^*, \mathbf{y}^*)$ for $\min_{\boldsymbol{\xi},\mathbf{w}} \max_{\mathbf{y}} \mathcal{L}(\boldsymbol{\xi}, \mathbf{w}, \mathbf{y})$ exists, where $\mathcal{L}$ is given in (2.5); moreover, any saddle point $(\boldsymbol{\xi}^*, \mathbf{w}^*, \mathbf{y}^*)$ satisfies that $\mathbf{y}^* = \mathbf{1} \otimes y^*$ for some $y^* \in \mathcal{B}_0$ such that $(\boldsymbol{\xi}^*, y^*)$ is a primal-dual solution to (1.2). Thus, $y^* \in \mathcal{K}^\circ$ and $\mathcal{L}(\boldsymbol{\xi}^*, \mathbf{w}^*, \mathbf{y}^*) = \varphi(\boldsymbol{\xi}^*)$. Indeed, this implies $\langle \mathbf{y}^*, \mathbf{w}^* \rangle - \sigma_{\widetilde{\mathcal{C}}}(\mathbf{w}^*) = 0$, which leads to $\sum_{i\in\mathcal{N}} w_i^* = \mathbf{0}$, i.e., $\mathbf{w}^* \in \mathcal{C}^\circ$. Hence, we have $0 = \langle \mathbf{y}^*, \mathbf{w}^* \rangle = \sigma_{\widetilde{\mathcal{C}}}(\mathbf{w}^*)$, and it trivially follows that if $(\boldsymbol{\xi}^*, \mathbf{w}^*, \mathbf{y}^*)$ is a saddle point of $\mathcal{L}$ with $\mathbf{w}^* \neq \mathbf{0}$, then $(\boldsymbol{\xi}^*, \mathbf{0}, \mathbf{y}^*)$ is another saddle point of $\mathcal{L}$. Therefore, under Assumption 3, there is always a saddle point of the form $(\boldsymbol{\xi}^*, \mathbf{0}, \mathbf{y}^*)$, i.e., with $\mathbf{w}^* = \mathbf{0}$. In the rest, let $\mathbf{z}^*$ be a saddle point with components $(\boldsymbol{\xi}^*, \mathbf{0}, \mathbf{y}^*)$.

Next, we state few useful observations later used in the proof. Given some $\beta > 0$, when primal-dual step-sizes are chosen as stated in (3.1), Lemma 3.4 implies that $\bar{\mathbf{Q}}(\beta) \succ 0$ and it follows from the definitions of $D_x, D_y$, and $T$ that $\forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}$,

$$
\begin{aligned}
(3.12) \quad & D_x(\mathbf{x}, \mathbf{x}') + D_y(\mathbf{y}, \mathbf{y}') - \langle T(\mathbf{x}) - T(\mathbf{x}'),\ \mathbf{y} - \mathbf{y}' \rangle \\
& \geq \sum_{i\in\mathcal{N}} \tfrac{1}{2} \max\{1, L_{f_i} + \beta L_{g_i}\} \left\| \xi_i - \xi_i' \right\|^2 + \frac{1}{4\gamma} \left\| \mathbf{w} - \mathbf{w}' \right\|^2 + \frac{\gamma}{4} \left\| \mathbf{y} - \mathbf{y}' \right\|^2 .
\end{aligned}
$$

Moreover, the error term $E^{k+1}(\mathbf{z})$, defined in Lemma 3.3, trivially satisfies

$$
(3.13) \qquad E^{k+1}(\mathbf{z}) \leq \gamma(2\|\mathbf{e}^{k+1}\| + \|\mathbf{e}^k\|)(\tfrac{1}{\gamma}\|\mathbf{w}^{k+1} - \mathbf{w}\| + \|\mathbf{y}^{k+1} - \mathbf{y}\|) \quad \forall\ k \geq 0.
$$

**3.2.1. Boundedness of dual iterate sequence.** Next, we show that $\{\mathbf{y}^k\}_{k\geq 0}$ and $\{\mathbf{w}^k\}_{k\geq 0}$ are bounded. More specifically, our aim is to show that there exist $\bar{\beta}, \varsigma, \nu \in \mathbb{R}_+$ such that if we choose the step-sizes as in (3.1) for any $\gamma > 0$ and $\beta \geq \bar{\beta}$, then

$$
(3.14) \qquad \max_{i\in\mathcal{N}}\{\|y_i^k\|\} \leq \beta, \quad \|\mathbf{w}^k\| \leq \varsigma, \quad \|\mathbf{e}^k\| \leq \nu \alpha^{q_{k-1}} k
$$

$\forall k \geq 0$, where $q_{-1} \triangleq 0$ and $q_0 \triangleq 0$. Below we provide the analysis for two sperate cases. We first define two quantities that are repeatedly used in the proof. Define $C_0 \triangleq \sum_{k=1}^{\infty} \alpha^{q_{k-1}} k < +\infty$, which implies $C_0 > 1$. Let $A_0 \triangleq D_x(\mathbf{x}^*, \mathbf{x}^0) + D_y(\mathbf{y}^*, \mathbf{y}^0) - \langle T(\mathbf{x}^*) - T(\mathbf{x}^0),\ \mathbf{y}^* - \mathbf{y}^0 \rangle$. Since we initialize $\mathbf{w}^0 = \mathbf{y}^0 = \mathbf{0}$, the proof of Lemma 3.4 implies that $A_0 \leq \left\| \boldsymbol{\xi}^* - \boldsymbol{\xi}^0 \right\|_{\mathbf{D}_\tau}^2 + \|\mathbf{y}^*\|_{\mathbf{D}_\kappa}^2$. Recall the definitions of $\bar{C}_g$ and $\bar{R}_x$ given in section 3. Using (3.1), we get

$$
(3.15) \qquad A_0 \leq \left\| \boldsymbol{\xi}^* - \boldsymbol{\xi}^0 \right\|_{\mathbf{D}_\tau}^2 + \|\mathbf{y}^*\|_{\mathbf{D}_\kappa}^2 \leq (\beta + 1) N \bar{R}_x^2 + (\bar{C}_g + \tfrac{5}{2}\gamma) N \left\| y^* \right\|^2 \triangleq \bar{A}_0.
$$

In the rest we assume $\bar{C}_g \geq 1$.

*Case* 1. *Bound $B$ on $\|y^*\|$ is available, i.e., $B\in(0,\infty)$.* In this part, we assume that a nontrivial dual bound $B\in(0,\infty)$ is available. Suppose we set $\bar{\beta}=2B$ and we choose the step-sizes as in (3.1) for some $\gamma>0$ and $\beta\geq\bar{\beta}$. Trivially, from (2.7d), we have $\max_{i\in\mathcal{N}}\left\|y_i^k\right\|\leq 2B\leq\beta$ for $k\geq 0$. Hence, Lemma 3.3 shows that $\forall k\geq 0$, (3.6) holds for $\beta_k=\beta$. Moreover, the step-size conditions in (3.1) and Lemma 3.4 imply that $\bar{\mathbf{Q}}(\beta)\succ\mathbf{0}$. Therefore, for any $\ell\geq 0$, dropping the last term in (3.6), summing over $k\in\{0,\ldots,\ell\}$, and using Jensen's inequality, we get $\forall\mathbf{z}\in\mathcal{Z}$,

$$
\begin{aligned}
(3.16)\quad &(\ell+1)(\mathcal{L}(\bar{\mathbf{x}}^{\ell+1},\mathbf{y})-\mathcal{L}(\mathbf{x},\bar{\mathbf{y}}^{\ell+1}))\\
&\leq\Big[D_x(\mathbf{x},\mathbf{x}^0)+D_y(\mathbf{y},\mathbf{y}^0)-\big\langle T(\mathbf{x})-T(\mathbf{x}^0),\ \mathbf{y}-\mathbf{y}^0\big\rangle\Big]\\
&-\Big[D_x(\mathbf{x},\mathbf{x}^{\ell+1})+D_y(\mathbf{y},\mathbf{y}^{\ell+1})-\big\langle T(\mathbf{x})-T(\mathbf{x}^{\ell+1}),\ \mathbf{y}-\mathbf{y}^{\ell+1}\big\rangle\Big]+\sum_{k=0}^{\ell}E^{k+1}(\mathbf{z}),
\end{aligned}
$$

where $\bar{\mathbf{x}}^{\ell+1}\triangleq\frac{1}{(\ell+1)}\sum_{k=1}^{\ell+1}\mathbf{x}^k$ and $\bar{\mathbf{y}}^{\ell+1}\triangleq\frac{1}{(\ell+1)}\sum_{k=1}^{\ell+1}\mathbf{y}^k$. For any $\ell\geq 0$, setting $\mathbf{z}=\mathbf{z}^*$ in (3.16) and using $\mathcal{L}(\bar{\mathbf{x}}^{\ell+1},\mathbf{y}^*)-\mathcal{L}(\mathbf{x}^*,\bar{\mathbf{y}}^{\ell+1})\geq 0$ and (3.12) we obtain

$$
(3.17)\qquad \tfrac{1}{4\gamma}\|\mathbf{w}^{\ell+1}\|^2+\tfrac{\gamma}{4}\|\mathbf{y}^*-\mathbf{y}^{\ell+1}\|^2\leq A_0+\sum_{k=0}^{\ell}E^{k+1}(\mathbf{z}^*).
$$

Hence, using (3.13), (3.17), and the fact that $\mathbf{w}^*=0\ \forall\ell\geq 0$, we have

$$
\begin{aligned}
&\tfrac{\gamma}{8}(\tfrac{1}{\gamma}\|\mathbf{w}^{\ell+1}\|+\|\mathbf{y}^*-\mathbf{y}^{\ell+1}\|)^2\leq\tfrac{1}{4\gamma}\|\mathbf{w}^{\ell+1}\|^2+\tfrac{\gamma}{4}\|\mathbf{y}^*-\mathbf{y}^{\ell+1}\|^2\\
(3.18)\qquad &\leq A_0+\sum_{k=1}^{\ell+1}\gamma(2\|\mathbf{e}^k\|+\|\mathbf{e}^{k-1}\|)(\tfrac{1}{\gamma}\|\mathbf{w}^k\|+\|\mathbf{y}^*-\mathbf{y}^k\|).
\end{aligned}
$$

Next, we use Lemma 9.2 with $u_k=\frac{1}{\gamma}\|\mathbf{w}^k\|+\|\mathbf{y}^*-\mathbf{y}^k\|$, $S_k=\frac{8}{\gamma}A_0$ for $k\geq 0$, and $\lambda_k=8(2\left\|\mathbf{e}^k\right\|+\left\|\mathbf{e}^{k-1}\right\|)$ for $k\geq 1$. Note (3.12) and $\mathbf{w}^0=\mathbf{y}^0=\mathbf{0}$ imply that $A_0\geq\frac{\gamma}{4}\left\|\mathbf{y}^*\right\|^2$; hence, we have $u_0^2\leq S_0$. Thus, Lemma 9.2 implies that $\forall\ell\geq 0$,

$$
(3.19)\qquad \tfrac{1}{\gamma}\|\mathbf{w}^{\ell+1}\|+\|\mathbf{y}^*-\mathbf{y}^{\ell+1}\|\leq\tfrac{1}{2}\sum_{k=1}^{\ell+1}\lambda_k+\sqrt{\frac{8A_0}{\gamma}+\left(\tfrac{1}{2}\sum_{k=1}^{\ell+1}\lambda_k\right)^2}\leq 24\sum_{k=1}^{\ell+1}\left\|\mathbf{e}^k\right\|+\sqrt{\frac{8A_0}{\gamma}}.
$$

For each $i\in\mathcal{N}$ and $k\geq 0$, the definition of $\mathcal{R}^k$ in (2.10) implies $\mathcal{R}_i^k(\mathbf{y})\in\mathcal{B}_0\ \forall\mathbf{y}$; hence, from (2.11), $\|v_i^{k+1}\|\leq\|v_i^k+\gamma y_i^k\|+\gamma\|\mathcal{R}_i^k(\frac{1}{\gamma}\mathbf{v}^k+\mathbf{y}^k)\|\leq\|v_i^k\|+4\gamma B$; thus, $\max_{i\in\mathcal{N}}\left\|v_i^k\right\|\leq 4\gamma Bk$ for $k\geq 0$, and we trivially get the following bound:

$$
(3.20)\qquad \|\mathbf{v}^k\|\leq 4\gamma\sqrt{N}\ B\ k\quad\forall\ k\geq 0.
$$

Hence, for $k\geq 0$, since $\|\mathbf{y}^k\|\leq 2\sqrt{N}\ B$, it follows from (2.10), (3.5), and (3.20) that

$$
(3.21)\qquad \|\mathbf{e}^{k+1}\|\leq N\ \Gamma\alpha^{q_k}\|\tfrac{1}{\gamma}\mathbf{v}^k+\mathbf{y}^k\|\leq 2N^{3/2}B\Gamma\alpha^{q_k}(2k+1)\implies\nu=4N^{3/2}B\Gamma.
$$

Therefore, $\left\|\mathbf{e}^k\right\|$ satisfies (3.14) for $\nu=4N^{3/2}B\Gamma$. Using this result within (3.19), we obtain

$$
(3.22)\qquad \|\mathbf{w}^{\ell+1}\|\leq 24\gamma\nu\sum_{k=1}^{\ell+1}\alpha^{q_{k-1}}k+\sqrt{8A_0\gamma}\leq\varsigma\triangleq 24\gamma\nu C_0+\sqrt{8A_0\gamma}\quad\forall\ \ell\geq 0.
$$

*Case* 2. *Bound $B$ on $\|y^*\|$ is not available, i.e., $B=\infty$.* We set $B=+\infty$ in Algorithm 1. We prove the claim in (3.14) using induction; indeed, we construct $\bar\beta, \varsigma, \nu \in \mathbb{R}_+$ and show for any $\gamma>0$, $\beta \geq \bar\beta$ and $K \geq 1$ that if (3.14) holds $\forall k \in \mathcal{I} \triangleq \{0,\ldots,K-1\}$, then $\left\|\mathbf{e}^K\right\| \leq \nu\alpha^{q_{K-1}}K$ also holds and this implies $\left\|\mathbf{w}^K\right\| \leq \varsigma$ and $\max_{i\in\mathcal{N}}\{\left\|y_i^K\right\|\} \leq \beta$, which would complete the induction.

Since $\mathbf{v}^0=\mathbf{w}^0=\mathbf{0}$ and $\mathbf{y}^0=\mathbf{0}$, (3.14) trivially holds for $k=0$. Suppose for some $\beta>0$, (3.14) holds for $k\in\mathcal{I}$; hence, $\max_{i\in\mathcal{N}}\{\|y_i^k\|\} \leq \beta$ for $k\in\mathcal{I}$, and using the same arguments as in Case 1, it can be shown that (3.16), (3.17), (3.18), and (3.19) hold $\forall \ell \in \mathcal{I}$. Next, using (3.5), (3.19) implies that

$$\begin{aligned}
\|\mathbf{e}^K\| &= \left\|\mathcal{P}_{\widetilde{\mathcal{C}}}\left(\tfrac{1}{\gamma}\mathbf{w}^{K-1}+\mathbf{e}^{K-1}+\mathbf{y}^{K-1}\right)-\mathcal{R}^{K-1}\left(\tfrac{1}{\gamma}\mathbf{w}^{K-1}+\mathbf{e}^{K-1}+\mathbf{y}^{K-1}\right)\right\| \\
&\leq N\ \Gamma\alpha^{q_{K-1}}\left(\|\mathbf{e}^{K-1}\|+24\sum_{k=1}^{K-1}\|\mathbf{e}^k\|+\sqrt{\frac{8A_0}{\gamma}}+\|\mathbf{y}^*\|\right) \\
&\leq N\ \Gamma\nu\alpha^{q_{K-1}}\left(\alpha^{q_{K-2}}(K-1)+24\sum_{k=1}^{K-1}\alpha^{q_{k-1}}k+\left(\sqrt{\frac{8A_0}{\gamma}}+\|\mathbf{y}^*\|\right)/\nu\right). \qquad (3.23)
\end{aligned}$$

The assumption, $q_k \geq \log_{1/\alpha}(24N\Gamma(k+1))$ for $k\geq 0$, and $q_{-1}=0$ imply that $\alpha^{q_{k-1}}k \leq \frac{1}{24N\Gamma}$ for $k \geq 0$. Thus, for $\nu \triangleq \frac{24}{23}N\Gamma(\|\mathbf{y}^*\|+\sqrt{\frac{8A_0}{\gamma}})$, (3.23) is indeed bounded above by $\nu\alpha^{q_{K-1}}K$, which proves the induction on $\left\|\mathbf{e}^K\right\|$. Hence, using this result within (3.19) for $\ell=K-1$, we obtain

$$\tfrac{1}{\gamma}\|\mathbf{w}^K\|+\|\mathbf{y}^K\|\leq 24\nu\sum_{k=1}^{K}\alpha^{q_{k-1}}k+\sqrt{\frac{8A_0}{\gamma}}+\|\mathbf{y}^*\| \leq 24\nu C_0+\sqrt{\frac{8A_0}{\gamma}}+\|\mathbf{y}^*\|, \qquad (3.24)$$

where $C_0 \triangleq \sum_{k=1}^{\infty}\alpha^{q_{k-1}}k < +\infty$ and is independent of $\beta$. Thus, $\left\|\mathbf{w}^K\right\| \leq \gamma\beta$ and $\max_{i\in\mathcal{N}}\{\left\|y_i^K\right\|\} \leq \beta$ for all $\beta \geq (\frac{576}{23}N\Gamma C_0+1)(\sqrt{\frac{8A_0}{\gamma}}+\sqrt{N}\left\|y^*\right\|)$. Hence, using the bound on $A_0$ in (3.15), we derive a sufficient condition on $\beta$:

$$\beta \geq (\tfrac{576}{23}N\Gamma C_0+1)\sqrt{N}\left(\left\|y^*\right\|+\sqrt{\tfrac{8}{\gamma}\Big((\beta+1)\bar{R}_x^2+(\bar{C}_g+\tfrac{5}{2}\gamma)\left\|y^*\right\|^2\Big)}\right). \qquad (3.25)$$

Note that (3.25) implies that there exists $\bar\beta\in\mathbb{R}$ such that $\bar\beta \geq \|\mathbf{y}^*\|$ and $\forall\beta\geq\bar\beta$ and $\varsigma=\gamma\beta$, (3.14) holds when the step-sizes are chosen as in (3.1) using $\beta$. Thus, when primal step-sizes $[\tau_i]_{i\in\mathcal{N}}$ are chosen sufficiently small and $\{q_k\}$ are chosen such that $q_k \geq \log_{1/\alpha}(24N\Gamma(k+1))$ and $\sum_{k=1}^{\infty}\alpha^{q_{k-1}}k<\infty$, both $\{\mathbf{y}^k\}$ and $\{\mathbf{w}^k\}$ are *bounded*. Moreover, solving the quadratic inequality in (3.25), we get

$$\beta=\mathcal{O}\Big(\tfrac{1}{\gamma}N^3\Gamma^2C_0^2\bar{R}_x^2+N^{3/2}\Gamma C_0\Big(\left\|y^*\right\|+\sqrt{\tfrac{1}{\gamma}(\bar{R}_x^2+\bar{C}_g\left\|y^*\right\|^2)}\Big)\Big). \qquad (3.26)$$

If $g_i$ is an affine function ($L_{g_i}=0$) $\forall i\in\mathcal{N}$, then choosing $q_k$ as before and setting $\tau_i=(\max\{1,L_{f_i}\}+C_{g_i})^{-1}$ for $i\in\mathcal{N}$ guarantees that $\{\mathbf{y}^k\}_k$ and $\{\mathbf{w}^k\}_k$ are bounded. Moreover, since $\mathbf{D}_\tau$ does not depend on $\beta$, the term $(\beta+1)\bar{R}_x^2$ on the right-hand side of (3.25) becomes $\bar{R}_x^2$; thus,

$$\beta=\mathcal{O}\Big(N^{3/2}\Gamma C_0\Big(\left\|y^*\right\|+\sqrt{\tfrac{1}{\gamma}(\bar{R}_x^2+\bar{C}_g\left\|y^*\right\|^2)}\Big)\Big).$$

**3.2.2. Convergence of iterates.** In section 3.2.1, we showed that there exist $\bar{\beta}, \varsigma, \nu \in \mathbb{R}_+$ such that if we choose the step-sizes as in (3.1) for any $\gamma > 0$ and $\beta \geq \bar{\beta}$, then (3.14) holds $\forall k \geq 0$. Consider a saddle point $\mathbf{z}^* = [\mathbf{x}^{*\top} \mathbf{y}^{*\top}]^\top$ of $\mathcal{L}$ in (2.5), where $\mathbf{x}^* = [\boldsymbol{\xi}^{*\top} \mathbf{w}^{*\top}]^\top$. Trivially, (3.13) and (3.14) imply that

$$\sum_{k=0}^{\infty} E^{k+1}(\mathbf{z}^*) \leq 3\gamma \max_{k \geq 0}\{\tfrac{1}{\gamma}\|\mathbf{w}^{k+1} - \mathbf{w}^*\| + \|\mathbf{y}^{k+1} - \mathbf{y}^*\|\} \sum_{k=0}^{\infty}\|\mathbf{e}^{k+1}\| < \infty. \tag{3.27}$$

Evaluating (3.6) at $\mathbf{z} = \mathbf{z}^*$, we get

$$0 \leq \mathcal{L}(\mathbf{x}^{k+1}, \mathbf{y}^*) - \mathcal{L}(\mathbf{x}^*, \mathbf{y}^{k+1}) \leq \mathbf{a}^k - \mathbf{a}^{k+1} - \mathbf{b}^k + \mathbf{c}^k \tag{3.28}$$

for $k \geq 0$, where $\mathbf{a}^k \triangleq D_x(\mathbf{x}^*, \mathbf{x}^k) + D_y(\mathbf{y}^*, \mathbf{y}^k) - \langle T(\mathbf{x}^*) - T(\mathbf{x}^k),\ \mathbf{y}^* - \mathbf{y}^k \rangle$, $\mathbf{b}^k \triangleq \frac{1}{2}\left\|\mathbf{u}(\mathbf{z}^{k+1}, \mathbf{z}^k)\right\|^2_{\bar{\mathbf{Q}}(\beta)}$, and $\mathbf{c}^k \triangleq E^{k+1}(\mathbf{z}^*)$ for $k \geq 0$. Clearly, $\mathbf{b}^k \geq 0$ and $\mathbf{c}^k \geq 0$ for $k \geq 0$. Moreover, from (3.12), we get $\mathbf{a}^k \geq 0$ for $k \geq 0$. Since $\sum_{k=0}^{\infty} E^{k+1}(\mathbf{z}^*) < \infty$, Lemma 9.1 implies that $\lim_{k\to\infty} \mathbf{a}^k$ exists. Thus, $\{\mathbf{a}^k\}$ is a bounded sequence; and due to (3.12), $\{\mathbf{z}^k\}$ is bounded as well. Consequently, there exists a subsequence $\{\mathbf{z}^{k_n}\}_n$ such that $\mathbf{z}^{k_n} \to \mathbf{z}^{\#}$ as $n \to \infty$. Thus, there exists $N_1$ such that $\forall n \geq N_1$, we have $\left\|\mathbf{z}^{k_n} - \mathbf{z}^{\#}\right\| < \frac{\epsilon}{2}$. Moreover, Lemma 9.1 also implies $\sum_{k=0}^{\infty}\left\|\mathbf{u}(\mathbf{z}^{k+1}, \mathbf{z}^k)\right\|^2_{\bar{\mathbf{Q}}(\beta)} < \infty$. Since $\bar{\mathbf{Q}}(\beta) \succ \mathbf{0}$, for any $\epsilon > 0$, there exists $N_2$ such that $\forall n \geq N_2$, we have $\left\|\mathbf{z}^{k_n+1} - \mathbf{z}^{k_n}\right\| < \frac{\epsilon}{2}$. Therefore, by letting $N = \max\{N_1, N_2\}$ we get $\left\|\mathbf{z}^{k_n+1} - \mathbf{z}^{\#}\right\| < \epsilon$, i.e., $\mathbf{z}^{k_n+1} \to \mathbf{z}^{\#}$ as $n \to \infty$.

Note that (3.14) implies $\left\|\mathbf{e}^k\right\| \to 0$ as $k \to \infty$ for any $\{q_k\}$ such that $\sum_{k=1}^{\infty} \alpha^{q_k} k < +\infty$. Recall that $\psi_x(\mathbf{x}) = \frac{1}{2}\left\|\boldsymbol{\xi}\right\|^2_{\mathbf{D}_\tau} + \frac{1}{2}\left\|\mathbf{w}\right\|^2_{\mathbf{D}_\gamma}$, and $\psi_y(\mathbf{y}) = \frac{1}{2}\left\|\mathbf{y}\right\|^2_{\mathbf{D}_\kappa}$ are the strongly convex functions corresponding to Bregman distance functions $D_x$ and $D_y$, respectively. In particular, $D_x(\mathbf{x}, \bar{\mathbf{x}}) = \psi_x(\mathbf{x}) - \psi_x(\bar{\mathbf{x}}) - \langle \nabla\psi_x(\bar{\mathbf{x}}),\ \mathbf{x} - \bar{\mathbf{x}} \rangle$, and $D_y$ is defined similarly. The optimality conditions for (3.7) imply that $\forall n \in \mathbb{Z}_+$, $\mathbf{q}^n \in \partial\rho(\mathbf{x}^{k_n+1})$ and $\mathbf{p}^n \in \partial h(\mathbf{y}^{k_n+1})$, where $\mathbf{q}^n \triangleq \nabla\psi_x(\mathbf{x}^{k_n}) - \nabla\psi_x(\mathbf{x}^{k_n+1}) - (\nabla f(\mathbf{x}^{k_n}) + \mathbf{J}T(\mathbf{x}^{k_n})^\top \mathbf{y}^{k_n} + U\mathbf{e}^{k_n})$, and $\mathbf{p}^n \triangleq \nabla\psi_y(\mathbf{y}^{k_n}) - \nabla\psi_y(\mathbf{y}^{k_n+1}) + 2T(\mathbf{x}^{k_n+1}) - T(\mathbf{x}^{k_n}) + \gamma(2\mathbf{e}^{k_n+1} - \mathbf{e}^{k_n})$. Since $\nabla\psi_x$ and $\nabla\psi_y$ are continuously differentiable on $\mathbf{dom}\,\rho$ and $\mathbf{dom}\,h$, respectively, and since $\rho$ and $h$ are proper, closed convex functions, it follows from Theorem 24.4 in [37] that $\partial\rho(\mathbf{x}^{\#}) \ni \lim_n \mathbf{q}^n = -\nabla f(\mathbf{x}^{\#}) - \mathbf{J}T(\mathbf{x}^{\#})^\top \mathbf{y}^{\#}$, and $\partial h(\mathbf{y}^{\#}) \ni \lim_n \mathbf{p}^n = T(\mathbf{x}^{\#})$, which also implies that $\mathbf{z}^{\#}$ is a saddle point of (1.4).

Since (3.28) is true for any saddle point $\mathbf{z}^*$, by setting $\mathbf{z}^* = \mathbf{z}^{\#}$ in (3.28), one can conclude that $\mathbf{s}^{\#} \triangleq \lim_k \mathbf{s}^k \geq 0$ exists, where $\mathbf{s}^k \triangleq D_x(\mathbf{x}^{\#}, \mathbf{x}^k) + D_y(\mathbf{y}^{\#}, \mathbf{y}^k) - \langle T(\mathbf{x}^{\#}) - T(\mathbf{x}^k),\ \mathbf{y}^{\#} - \mathbf{y}^k \rangle$ for $k \geq 0$. Since $\lim_n \langle T(\mathbf{x}^{\#}) - T(\mathbf{x}^{k_n}), \mathbf{y}^{\#} - \mathbf{y}^{k_n} \rangle = 0$ (from $\mathbf{z}^{k_n} \to \mathbf{z}^{\#}$), clearly $\mathbf{s}^{\#} = \lim_{n\to\infty}\ \mathbf{s}^{k_n} = 0$, which together with (3.12) implies that $\mathbf{z}^k \to \mathbf{z}^{\#}$.

**3.2.3. Convergence rate.** Recall that we initialize $\mathbf{v}^0 = \mathbf{w}^0 = \mathbf{0}$ and $\mathbf{y}^0 = \mathbf{0}$; hence, the inequality in (3.16) can be written more explicitly as follows: let $\bar{\boldsymbol{\xi}}^K \triangleq \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\xi}^k$, and $\bar{\mathbf{w}}^K \triangleq \frac{1}{K}\sum_{k=1}^{K} \mathbf{w}^k$, and then for any $\boldsymbol{\xi}$, $\mathbf{w}$, and $\mathbf{y}$, and $\forall K \geq 1$,

$$\mathcal{L}(\bar{\boldsymbol{\xi}}^K, \bar{\mathbf{w}}^K, \mathbf{y}) - \mathcal{L}(\boldsymbol{\xi}, \mathbf{w}, \bar{\mathbf{y}}^K) \leq \Theta(\mathbf{z})/K, \tag{3.29}$$

where $\Theta(\mathbf{z}) \triangleq \frac{1}{2\gamma}\|\mathbf{w}\|^2 + \langle \mathbf{y},\ \mathbf{w} \rangle + \sum_{i\in\mathcal{N}}\left[\frac{1}{2\tau_i}\|\xi_i - \xi_i^0\|^2 + \frac{1}{2\kappa_i}\|y_i\|^2 + \langle g_i(\xi_i) - g_i(\xi_i^0), y_i \rangle\right]$ $+ \sum_{k=0}^{K-1} E^{k+1}(\mathbf{z})$. Given the step-size condition in (3.1), the Schur complement condition guarantees that $\left[\begin{smallmatrix} \frac{1}{\tau_i} & C_{g_i} \\ C_{g_i} & \frac{1}{\kappa_i} \end{smallmatrix}\right] \preceq \left[\begin{smallmatrix} \frac{2}{\tau_i} & 0 \\ 0 & \frac{2}{\kappa_i} \end{smallmatrix}\right]$ for any $i \in \mathcal{N}$; therefore,

$$\Theta(\mathbf{z}) \leq \sum_{i\in\mathcal{N}} \Big[\frac{1}{\tau_i}\|\xi_i - \xi_i^0\|^2 + \frac{1}{\kappa_i}\|y_i\|^2\Big] + \frac{1}{2\gamma}\|\mathbf{w}\|^2 + \langle \mathbf{y}, \mathbf{w}\rangle + \sum_{k=0}^{K-1} E^{k+1}(\mathbf{z}). \tag{3.30}$$

In the rest, fix $K \geq 1$ and a saddle point $(\boldsymbol{\xi}^*, \mathbf{w}^*, \mathbf{y}^*)$ of $\mathcal{L}$ in (2.5) such that $\mathbf{w}^* = \mathbf{0}$. Let $\hat{y}^K \triangleq 2\|y^*\|\,\mathcal{P}_{\mathcal{K}^\circ}(-g(\bar{\boldsymbol{\xi}}^K))/\|\mathcal{P}_{\mathcal{K}^\circ}(-g(\bar{\boldsymbol{\xi}}^K))\| \in \mathcal{K}^\circ$, and define $\hat{\mathbf{y}}^K = [\hat{y}_i^K]_{i\in\mathcal{N}}$ such that $\hat{y}_i^K = \hat{y}^K \ \ \forall i \in \mathcal{N}$, i.e., $\hat{\mathbf{y}}^K = \mathbf{1} \otimes \hat{y}^K \in \widetilde{\mathcal{C}}$, and also define $\hat{\mathbf{w}}^K \triangleq \left\|\mathcal{P}_{\mathcal{C}^\circ}(\bar{\mathbf{y}}^K)\right\|^{-1} \mathcal{P}_{\mathcal{C}^\circ}(\bar{\mathbf{y}}^K)$, where $\mathcal{C} \supset \widetilde{\mathcal{C}}$ defined in (2.2) is a closed convex cone and $\mathcal{C}^\circ$ denotes its polar cone. Note that $\hat{\mathbf{y}}^K \in \mathcal{C}$ and $\hat{\mathbf{w}}^K \in \mathcal{C}^\circ$ imply $\left\langle \hat{\mathbf{y}}^K, \hat{\mathbf{w}}^K \right\rangle \leq 0$. Recall that every closed convex cone $\mathcal{Q} \subset \mathbb{R}^m$ induces an orthogonal decomposition on $\mathbb{R}^m$, i.e., according to Moreau decomposition, for any $y \in \mathbb{R}^m$, there exist $y^1 \in \mathcal{Q}$ and $y^2 \in \mathcal{Q}^\circ$ such that $y = y^1 + y^2$ and $y^1 \perp y^2$; in particular, $y^1 = \mathcal{P}_{\mathcal{Q}}(y)$ and $y^2 = \mathcal{P}_{\mathcal{Q}^\circ}(y)$. Thus, $\left\langle \hat{\mathbf{w}}^K,\ \bar{\mathbf{y}}^K \right\rangle = \left\langle \hat{\mathbf{w}}^K,\ \mathcal{P}_{\mathcal{C}}(\bar{\mathbf{y}}^K) + \mathcal{P}_{\mathcal{C}^\circ}(\bar{\mathbf{y}}^K) \right\rangle = \left\|\mathcal{P}_{\mathcal{C}^\circ}(\bar{\mathbf{y}}^K)\right\| = d_{\mathcal{C}}(\bar{\mathbf{y}}^K)$. Note that for each $i \in \mathcal{N}$ we have $\bar{y}_i^K \in \mathcal{K}^\circ$ since $y_i^k \in \mathcal{K}^\circ \ \forall k = 1, \ldots, K$ and $\mathcal{K}$ is convex; hence, $\sigma_{\mathcal{K}}(\bar{y}_i^K) = 0$ for $i \in \mathcal{N}$. Moreover, $\hat{\mathbf{w}}^K \in \mathcal{C}^\circ$ implies $\sigma_{\mathcal{C}}(\hat{\mathbf{w}}^K) = \mathbb{1}_{\mathcal{C}^\circ}(\hat{\mathbf{w}}^K) = 0$; and since $\widetilde{\mathcal{C}} \subset \mathcal{C}$, we also have $\sigma_{\widetilde{\mathcal{C}}}(\hat{\mathbf{w}}^K) \leq \sigma_{\mathcal{C}}(\hat{\mathbf{w}}^K) = 0$. Therefore, we can conclude that $\sigma_{\widetilde{\mathcal{C}}}(\hat{\mathbf{w}}^K) = 0$ since $\mathbf{0} \in \widetilde{\mathcal{C}}$. These observations imply that

$$\mathcal{L}(\boldsymbol{\xi}^*, \hat{\mathbf{w}}^K, \bar{\mathbf{y}}^K) = \varphi(\boldsymbol{\xi}^*) - \sum_{i\in\mathcal{N}} \left\langle g_i(\xi_i^*),\ \bar{y}_i^K \right\rangle - d_{\mathcal{C}}(\bar{\mathbf{y}}^K). \tag{3.31}$$

Similarly, from the definition of $\hat{y}^K \in \mathcal{K}^\circ$, $-\sum_{i\in\mathcal{N}} \left\langle g_i(\bar{\xi}_i^K),\ \hat{y}^K \right\rangle = 2\|y^*\|\, d_{\mathcal{K}}(-g(\bar{\boldsymbol{\xi}}^K))$, and since $\hat{\mathbf{y}}^K \in \widetilde{\mathcal{C}}$, we also have $\left\langle \bar{\mathbf{w}}^K,\ \hat{\mathbf{y}}^K \right\rangle - \sigma_{\widetilde{\mathcal{C}}}(\bar{\mathbf{w}}^K) \leq \sup_{\mathbf{w}} \left\langle \mathbf{w},\ \hat{\mathbf{y}}^K \right\rangle - \sigma_{\widetilde{\mathcal{C}}}(\mathbf{w}) = \mathbb{1}_{\widetilde{\mathcal{C}}}(\hat{\mathbf{y}}^K) = 0$. Note $\sigma_{\mathcal{K}}(\hat{y}^K) = 0$ since $\hat{y}^K \in \mathcal{K}^\circ$. Thus, we conclude that $\mathcal{L}$ satisfies

$$\mathcal{L}(\bar{\boldsymbol{\xi}}^K, \bar{\mathbf{w}}^K, \hat{\mathbf{y}}^K) \geq \varphi(\bar{\boldsymbol{\xi}}^K) + 2\|y^*\|\, d_{\mathcal{K}}\Big(-g(\bar{\boldsymbol{\xi}}^K)\Big). \tag{3.32}$$

Combining (3.31) and (3.32), we get

$$\begin{aligned} &\mathcal{L}(\bar{\boldsymbol{\xi}}^K, \bar{\mathbf{w}}^K, \hat{\mathbf{y}}^K) - \mathcal{L}(\boldsymbol{\xi}^*, \hat{\mathbf{w}}^K, \bar{\mathbf{y}}^K) \\ &\quad \geq \varphi(\bar{\boldsymbol{\xi}}^K) - \varphi(\boldsymbol{\xi}^*) + 2\|y^*\|\, d_{\mathcal{K}}\Big(-g(\bar{\boldsymbol{\xi}}^K)\Big) + d_{\mathcal{C}}(\bar{\mathbf{y}}^K) + \sum_{i\in\mathcal{N}} \left\langle g_i(\xi_i^*),\ \bar{y}_i^K \right\rangle. \end{aligned} \tag{3.33}$$

Moreover, $\left\langle \hat{\mathbf{y}}^K,\ \hat{\mathbf{w}}^K \right\rangle \leq 0$, (3.29), and (3.30) imply that

$$\mathcal{L}(\bar{\boldsymbol{\xi}}^K, \bar{\mathbf{w}}^K, \hat{\mathbf{y}}^K) - \mathcal{L}(\boldsymbol{\xi}^*, \hat{\mathbf{w}}^K, \bar{\mathbf{y}}^K) \leq \Theta(\hat{\mathbf{z}}^K)/K \leq \frac{\Lambda_1 + \sum_{k=0}^{K-1} E^{k+1}(\hat{\mathbf{z}}^K)}{K} \triangleq \frac{\Lambda(\gamma, \beta)}{K}, \tag{3.34}$$

where $\hat{\mathbf{z}}^K = [{\boldsymbol{\xi}^*}^\top\ (\hat{\mathbf{w}}^K)^\top\ (\hat{\mathbf{y}}^K)^\top]^\top$ and $\Lambda_1 \triangleq \frac{1}{2\gamma} + \sum_{i\in\mathcal{N}} [\frac{1}{\tau_i}\|\xi_i^* - \xi_i^0\|^2 + \frac{4}{\kappa_i}\|y^*\|^2]$. Recall that we fixed a saddle point $(\boldsymbol{\xi}^*, \mathbf{w}^*, \mathbf{y}^*)$ such that $\mathbf{w}^* = \mathbf{0}$; hence, we have $\mathcal{L}(\boldsymbol{\xi}^*, \mathbf{w}^*, \mathbf{y}^*) = \varphi(\boldsymbol{\xi}^*)$ and $\sigma_{\widetilde{\mathcal{C}}}(\mathbf{w}^*) = 0$. Moreover, since $(\boldsymbol{\xi}^*, \mathbf{w}^*, \mathbf{y}^*)$ is a saddle point, we have $\mathcal{L}(\bar{\boldsymbol{\xi}}^K, \mathbf{w}^*, \mathbf{y}^*) - \mathcal{L}(\boldsymbol{\xi}^*, \mathbf{w}^*, \mathbf{y}^*) \geq 0$ and $\mathcal{L}(\boldsymbol{\xi}^*, \mathbf{w}^*, \mathbf{y}^*) - \mathcal{L}(\boldsymbol{\xi}^*, \mathbf{w}^*, \bar{\mathbf{y}}^K) \geq 0$; therefore, these facts imply that

$$\sum_{i\in\mathcal{N}} \langle g_i(\xi_i^*),\ \bar{y}_i^K \rangle \geq 0, \qquad \varphi(\bar{\boldsymbol{\xi}}^K) - \varphi(\boldsymbol{\xi}^*) + \|y^*\|\, d_{\mathcal{K}}\Big(-g(\bar{\boldsymbol{\xi}}^K)\Big) \geq 0, \tag{3.35}$$

where we used $y^* \in \mathcal{K}^\circ$, i.e., $\langle y^*, y\rangle \leq \langle y^*, \mathcal{P}_{\mathcal{K}^\circ}(y)\rangle \leq \|y^*\|\, d_{\mathcal{K}}(y)\ \forall y \in \mathbb{R}^m$. Therefore, combining (3.33), (3.34), and (3.35) gives us the *infeasibility* and *consensus* results in (3.2) and also the upper bound in (3.3); while the inequality on the left in (3.35) gives us the lower bound for the *suboptimality*.

To show that $\Lambda(\gamma, \beta)$ is finite and independent of $K$, we bound $\sum_{k=0}^{K-1} E^{k+1}(\hat{\mathbf{z}}^K)$. As in (3.27), using (3.13), (3.14), and (3.19), we get

$$\begin{aligned} \sum_{k=0}^{K-1} E^{k+1}(\hat{\mathbf{z}}^K) &\leq 3\gamma \max_{k=0,\ldots,K-1}\{\tfrac{1}{\gamma}\left\|\mathbf{w}^{k+1} - \hat{\mathbf{w}}^K\right\| + \left\|\mathbf{y}^{k+1} - \hat{\mathbf{y}}^K\right\|\} \sum_{k=0}^{K-1}\left\|\mathbf{e}^{k+1}\right\| \\ &\triangleq \Lambda_2 \leq 3\nu C_0\Big(1 + \sqrt{8A_0\gamma} + \gamma(24\nu C_0 + 3\sqrt{N}\,\|y^*\|)\Big) \end{aligned} \tag{3.36}$$

(recall $\sum_{k=0}^{\infty}\left\|\mathbf{e}^{k+1}\right\| = \nu C_0$). Below we specify the bound in (3.36) for both cases.

For Case 1 where $B$ is known, $\nu = 4N^{\frac{3}{2}}\Gamma B$ (see (3.21)) and $\|y^*\| \leq B$; hence, using these facts and the bound on $A_0$ given in (3.15) within (3.36), we get

$$\Lambda_2 \leq N^{\frac{3}{2}}\Gamma C_0 B\ \mathcal{O}\Big(1 + \sqrt{\gamma N(B\bar{R}_x^2 + B^2\bar{C}_g)} + \gamma N^{\frac{3}{2}}\Gamma C_0 B\Big).$$

Moreover, the second inequality in (3.15) implies $\Lambda_1 = \mathcal{O}(\frac{1}{\gamma} + N(B\bar{R}_x^2 + B^2\bar{C}_g) + \gamma N B^2)$. Our aim is to optimize the $\mathcal{O}(1)$ constant of $\Lambda_1 + \Lambda_2$ via carefully selecting the free parameter $\gamma$. Setting $\gamma = (N^{3/2}\Gamma C_0 B)^{-1}$ gives $\Lambda(\gamma, \beta) = \mathcal{O}(NB(\bar{R}_x^2 + \bar{C}_g B) + N^{3/2}\Gamma C_0 B(1 + \sqrt{\frac{\bar{R}_x^2 + \bar{C}_g B}{N^{1/2}\Gamma C_0}}))$ which implies the $N$ dependency in (3.4).

For Case 2, where $B$ is not known, $\nu = \frac{24}{23}N\Gamma(\|\mathbf{y}^*\| + \sqrt{\frac{8A_0}{\gamma}})$—see the discussion below (3.23); hence, from (3.36), we get $\Lambda_2 \leq \mathcal{O}(N^2\Gamma^2 C_0^2(A_0 + \gamma N \max\{1, \|y^*\|^2\}))$. For the sake of simplicity, suppose $\|y^*\| \geq 1$. Moreover, $\Lambda_1 = \mathcal{O}(\frac{1}{\gamma} + \bar{A}_0)$, and since $A_0 \leq \bar{A}_0$ (see (3.15)), $\Lambda_1 + \Lambda_2 = \mathcal{O}(\frac{1}{\gamma} + N^2\Gamma^2 C_0^2 \bar{A}_0)$. Selecting $\gamma = N^{\frac{3}{2}}\Gamma C_0 \bar{R}_x^2$, (3.15) and the bound on $\beta$ in (3.26) together imply that $\Lambda_1 + \Lambda_2 = \mathcal{O}(N^{\frac{9}{2}}\Gamma^3 C_0^3 \bar{R}_x^2 \|y^*\|^2)$, assuming $N^{\frac{3}{2}}\Gamma > \bar{C}_g/\bar{R}_x^2$ and $N > 1/\bar{R}_x^2$, which are reasonable since we are interested in the bounds when $N$ is large. Moreover, when $g_i$'s are linear functions ($L_{g_i} = 0$) the bound $\bar{A}_0$ can be simplified, i.e., $\bar{A}_0 = N(\bar{R}_x^2 + (\bar{C}_g + \frac{5\gamma}{2})\,\|y^*\|^2)$. Therefore, choosing $\gamma = (N^{\frac{3}{2}}\Gamma C_0)^{-1}$, we get $\Lambda_1 + \Lambda_2 = \mathcal{O}(N^3\Gamma^2 C_0^2(\bar{R}_x^2 + \bar{C}_g\,\|y^*\|^2))$.

*Remark* 3.5. For Case 1, assuming $\sum_{k=0}^{\infty}\alpha^{q_k}(k+1)^2 < +\infty$ in addition to $C_0 < +\infty$, one can observe that using (3.20) and (3.21), the $\mathcal{O}(1)$ term takes a simpler form: $\Lambda(\gamma, \beta) = \Lambda_1 + \sum_{k=0}^{K-1} E^{k+1}(\hat{\mathbf{z}}^K) \leq \frac{1}{\gamma} + N(B\bar{R}_x^2 + B^2\bar{C}_g) + \gamma N B^2 + 12N^{\frac{3}{2}}\Gamma B[\sum_{k=1}^{K}\alpha^{q_{k-1}}k + 4\sqrt{N}B\gamma\sum_{k=1}^{K}\alpha^{q_{k-1}}k(k+1)]$.

**4. Fully distributed step-size rule.** Recall that the step-size selection rule in (3.1) of Theorem 3.3 requires some sort of coordination among the nodes in $\mathcal{N}$ because there is a fixed $\gamma > 0$ coupling and affecting all nodes' step-size choice. To overcome this issue, we will define $\gamma_i > 0$ for each node, which are node-specific and can be chosen independently. Let $\mathbf{D}_\gamma \triangleq \mathrm{diag}([\frac{1}{\gamma_i}\mathbf{I}_m]_{i\in\mathcal{N}}) \succ 0$ and define $\boldsymbol{\gamma} \triangleq [\gamma_i]_{i\in\mathcal{N}}$ and $\widehat{\mathcal{C}} \triangleq \{\mathbf{p} \in \mathcal{Y} :\ \exists \bar{y} \in \mathbb{R}^m \text{ s.t. } \frac{1}{\sqrt{\gamma_i}}\ p_i = \bar{y}\quad \forall i \in \mathcal{N},\quad \|\bar{y}\| \leq 2B\}$—here, $\mathbf{p} = [p_i]_{i\in\mathcal{N}}$. Recall the definition of the Bregman distance function given in Definition 2: $D_x(\mathbf{x}, \bar{\mathbf{x}}) = \frac{1}{2}\left\|\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}\right\|_{\mathbf{D}_\tau}^2 + \frac{1}{2}\left\|\mathbf{w} - \bar{\mathbf{w}}\right\|_{\mathbf{D}_\gamma}^2$. Switching to $\mathbf{D}_\gamma$ as defined above, (2.7a) should be replaced with $\mathbf{w}^{k+1} \leftarrow \mathrm{argmin}_{\mathbf{w}}\, \sigma_{\widetilde{\mathcal{C}}}(\mathbf{w}) - \langle \mathbf{y}^k,\ \mathbf{w}\rangle + \frac{1}{2}\|\mathbf{w} - \mathbf{v}^k\|_{\mathbf{D}_\gamma}^2$. Using the change of variables $\hat{\mathbf{w}} \triangleq \mathbf{D}_\gamma^{\frac{1}{2}}\mathbf{w}$, it can be rewritten as

$$\mathbf{w}^{k+1} \leftarrow \mathbf{D}_\gamma^{-\frac{1}{2}} \operatorname*{argmin}_{\hat{\mathbf{w}}} \sigma_{\widehat{\mathcal{C}}}\,(\hat{\mathbf{w}}) + \frac{1}{2}\|\hat{\mathbf{w}} - (\mathbf{D}_\gamma^{\frac{1}{2}}\mathbf{v}^k + \mathbf{D}_\gamma^{-\frac{1}{2}}\mathbf{y}^k)\|^2, \tag{4.1}$$

where we use the fact that $\sigma_{\widehat{\mathcal{C}}}(\mathbf{D}_\gamma^{-\frac{1}{2}}\hat{\mathbf{w}}) = \sigma_{\widehat{\mathcal{C}}}\,(\hat{\mathbf{w}})$. Now, we can write (4.1) in a proximal form and using Moreau's decomposition, we get

$$\begin{aligned}\mathbf{w}^{k+1} &= \mathbf{D}_\gamma^{-\frac{1}{2}}\mathbf{prox}_{\sigma_{\widehat{\mathcal{C}}}}(\mathbf{D}_\gamma^{\frac{1}{2}}\mathbf{v}^k + \mathbf{D}_\gamma^{-\frac{1}{2}}\mathbf{y}^k)\\ &= \mathbf{D}_\gamma^{-\frac{1}{2}}(\mathbf{D}_\gamma^{\frac{1}{2}}\mathbf{v}^k + \mathbf{D}_\gamma^{-\frac{1}{2}}\mathbf{y}^k - \mathcal{P}_{\widehat{\mathcal{C}}}(\mathbf{D}_\gamma^{\frac{1}{2}}\mathbf{v}^k + \mathbf{D}_\gamma^{-\frac{1}{2}}\mathbf{y}^k)).\end{aligned}$$

Note that $\mathbf{p} \in \widehat{\mathcal{C}}$ implies that $\mathbf{D}_\gamma^{\frac{1}{2}}\mathbf{p} = \mathbf{1}_N \otimes \bar{y}$ for some $\bar{y} \in \mathbb{R}^m$ such that $\|\bar{y}\| \leq 2B$. Therefore, for $\mathbf{y} = [y_i]_{i\in\mathcal{N}} \in \mathbb{R}^{n_0}$, the projection of $\mathbf{D}_\gamma^{-\frac{1}{2}}\mathbf{y}$ onto $\widehat{\mathcal{C}}$ can be computed as

$$\begin{aligned}\mathcal{P}_{\widehat{\mathcal{C}}}(\mathbf{D}_\gamma^{-\frac{1}{2}}\mathbf{y}) &= \operatorname*{argmin}_{\mathbf{p}\in\widehat{\mathcal{C}}} \tfrac{1}{2}\left\|\mathbf{D}_\gamma^{-\frac{1}{2}}\mathbf{y} - \mathbf{p}\right\|^2 = \mathbf{D}_\gamma^{-\frac{1}{2}}\left(\mathbf{1} \otimes \operatorname*{argmin}_{\|\bar{y}\|\leq 2B} \frac{1}{2}\left\|\mathbf{D}_\gamma^{-\frac{1}{2}}\mathbf{y} - \mathbf{D}_\gamma^{-\frac{1}{2}}(\mathbf{1}\otimes\bar{y})\right\|^2\right)\\ &= \mathbf{D}_\gamma^{-\frac{1}{2}}\mathcal{P}_{\mathcal{B}}\left(\frac{1}{\sum_{i\in\mathcal{N}}\gamma_i}(\mathbf{1}\mathbf{1}^\top \otimes \mathbf{I}_m)\mathbf{D}_\gamma^{-1}\mathbf{y}\right). \qquad (4.2)\end{aligned}$$

Let $\mathcal{P}_\gamma(\mathbf{y}) \triangleq \mathbf{1}_N \otimes \mathcal{P}_{\mathcal{B}_0}\left(\frac{1}{\sum_{i\in\mathcal{N}}\gamma_i}\sum_{i\in\mathcal{N}}\gamma_i y_i\right)$; hence, we get that

$$\mathbf{w}^{k+1} = \mathbf{D}_\gamma^{-1}\Big(\mathbf{D}_\gamma\mathbf{v}^k + \mathbf{y}^k - \mathcal{P}_\gamma(\mathbf{D}_\gamma\mathbf{v}^k + \mathbf{y}^k)\Big). \tag{4.3}$$

Thus, we propose approximating $\mathcal{P}_\gamma(\cdot)$ using an approximate convex combination operator $\mathcal{R}_\gamma^k(\cdot) = [\mathcal{R}_i^k(\cdot)]_{i\in\mathcal{N}}$ such that it can be computed in a distributed way, i.e., $\mathcal{R}_i^k(\cdot)$ can be computed at $i \in \mathcal{N}$ using local communication. More precisely, suppose $\mathcal{R}_\gamma^k$ satisfies a slightly modified version of Assumption 4, where (2.10) is replaced with

$$\mathcal{R}_\gamma^k(\mathbf{w}) \in \mathcal{B} \qquad \|\mathcal{R}_\gamma^k(\mathbf{w}) - \mathcal{P}_\gamma(\mathbf{w})\| \leq N\ \Gamma\alpha^{q_k}\,\|\mathbf{w}\| \quad \forall\ \mathbf{w} \in \mathbb{R}^{n_0}. \tag{4.4}$$

Provided that such an operator exists, instead of (2.11), we set $\mathbf{v}^{k+1}$ as follows:

$$\mathbf{v}^{k+1} \leftarrow \mathbf{D}_\gamma^{-1}\Big(\mathbf{D}_\gamma\mathbf{v}^k + \mathbf{y}^k - \mathcal{R}_\gamma^k(\mathbf{D}_\gamma\mathbf{v}^k + \mathbf{y}^k)\Big). \tag{4.5}$$

With this modification, we can still show that the iterate sequence converges to a primal-dual optimal solution with $\mathcal{O}(1/K)$ ergodic rate provided that primal-dual step-sizes $\{\tau_i, \kappa_i\}_{i\in\mathcal{N}}$ and $\{\gamma_i\}_{i\in\mathcal{N}}$ are chosen such that $\tau_i = (\max\{1, L_{f_i} + \beta L_{g_i}\} + C_{g_i})^{-1}$, $\kappa_i = (C_{g_i} + \frac{5\gamma_i}{2})^{-1}\ \forall i \in \mathcal{N}$.

In the rest of this section, for both undirected and directed time-varying communication networks, we provide an operator $\mathcal{R}_\gamma^k$ satisfying (4.4). For $\mathbf{y} = [y_i]_{i\in\mathcal{N}} \in \mathcal{Y}$, define $p_\gamma(\mathbf{y}) \triangleq \frac{1}{\sum_{i\in\mathcal{N}}\gamma_i}\sum_{i\in\mathcal{N}}\gamma_i y_i$; hence, we have $\mathcal{P}_\gamma(\mathbf{y}) = \mathbf{1}_N \otimes \mathcal{P}_{\mathcal{B}_0}(p_\gamma(\mathbf{y}))$. Therefore, we should consider distributed approximation of $p_\gamma(\mathbf{y})$. Given $y_i \in \mathbb{R}^m$ and $\gamma_i > 0$, which are only known at node $i \in \mathcal{N}$, we next discuss extensions of techniques discussed in Section 2.1 to compute the convex combination $\sum_{i\in\mathcal{N}}\gamma_i y_i / \sum_{i\in\mathcal{N}}\gamma_i$.

First, suppose that $\{\mathcal{G}^t\}$ is a time-varying undirected graph and $\{V^t\}_{t\in\mathbb{Z}_+}$ a corresponding sequence of weight matrices satisfying Assumption 5. For $\mathbf{w} = [w_i]_{i\in\mathcal{N}} \in \mathcal{Y}$ such that $w_i \in \mathbb{R}^m$ for $i \in \mathcal{N}$, define

$$\mathcal{R}_\gamma^k(\mathbf{w}) \triangleq \mathcal{P}_{\mathcal{B}}\left(\left(\operatorname{diag}(W^{t_k+q_k,t_k}\boldsymbol{\gamma})^{-1}W^{t_k+q_k,t_k} \otimes \mathbf{I}_m\right)\ \mathbf{D}_\gamma^{-1}\mathbf{w}\right) \tag{4.6}$$

to approximate $\mathcal{P}_\gamma(\cdot)$ in (4.3). Note that $\mathcal{R}^k_\gamma(\cdot)$ can be computed in a *distributed fashion* requiring $q_k$ communications with the neighbors for each node. Using Lemma 2.1, it is easy to show that $\mathcal{R}^k_\gamma$ given in (4.6) satisfies the condition in (4.4).

Second, suppose that $\{\mathcal{G}^t\}$ is a time-varying $M$-strongly-connected directed graph, and $\{V^t\}_{t\in\mathbb{Z}_+}$ the corresponding weight-matrix sequence as defined in (2.13) within Section 2.1.2—so that (4.6) can be computed over a time-varying *directed* network. Given any $\mathbf{w} = [w_i]_{i\in\mathcal{N}}$ and $\{\gamma_i\}_{i\in\mathcal{N}}$, the results in [27] immediately imply that for any $s \in \mathbb{Z}_+$, the vector $(\,\mathrm{diag}(W^{t,s}\boldsymbol{\gamma})^{-1}W^{t,s}\otimes\mathbf{I}_m)\mathbf{D}_\gamma^{-1}\mathbf{w}$ converges to the consensus convex combination vector $\mathbf{1}_N \otimes p_\gamma(\mathbf{w})$ with a geometric rate as $t$ increases. Indeed, this can be trivially achieved by using a different initialization for the push-sum method. Next, we state a slightly modified version of the convergence result in Lemma 2.2.

LEMMA 4.1. *Suppose that the digraph sequence $\{\mathcal{G}^t\}_{t\geq 1}$ is uniformly strongly connected ($M$-strongly connected), where $\mathcal{G}^t = (\mathcal{N}, \mathcal{E}^t)$. Given node-specific data $\{w_i\}_{i\in\mathcal{N}} \subset \mathbb{R}^m$ and $\{\gamma_i\}_{i\in\mathcal{N}} \subset \mathbb{R}_{++}$, for any fixed integer $s \geq 0$, the following bound holds for all integers $t > s$:*

$$\left\|\mathrm{diag}(W^{t,s}\boldsymbol{\gamma}\otimes\mathbf{I}_m)^{-1}(W^{t,s}\otimes\mathbf{I}_m)\ \mathbf{D}_\gamma^{-1}\mathbf{w} - \mathbf{1}_N\otimes p_\gamma(\mathbf{w})\right\| \leq \frac{8\sqrt{N}}{\gamma_{\min}\delta}\sum_{i\in\mathcal{N}}\gamma_i\,\|w_i\|\ \ \alpha^{t-s-1}$$

*for some $\delta \geq \frac{1}{N^{NM}}$ and $0 < \alpha \leq \left(1 - \frac{1}{N^{NM}}\right)^{\frac{1}{M}}$, where $N = |\mathcal{N}|$ and $\gamma_{\min} = \min_{i\in\mathcal{N}}\gamma_i$.*

*Proof.* The proof follows from Corollary 2 and the proof of Lemma 1 in [27]. □

Thus, $\mathcal{R}^k_\gamma(\cdot)$ defined in (4.6) satisfies the requirement $\|\mathcal{R}^k_\gamma(\mathbf{w}) - \mathcal{P}_\gamma(\mathbf{w})\| \leq N\Gamma\ \alpha^{q_k}\,\|\mathbf{w}\|$ in (4.4) for

$$\Gamma = \frac{\|\boldsymbol{\gamma}\|}{\gamma_{\min}\sqrt{N}}\frac{8}{\delta\alpha}$$

and for some $\alpha \in (0, 1)$ and $\delta > 0$ as stated in Lemma 4.1.

## 5. A distributed algorithm for static network topology.

We extend the results in [2] to nonlinear constraint functions $\{g_i\}_{i\in\mathcal{N}}$. Given an *undirected*, *static* communication network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, following the discussion in section 1.2, the corresponding SP problem for the static network is given as $\min_{\boldsymbol{\xi},\mathbf{w}}\max_{\mathbf{y}}\{\sum_{i\in\mathcal{N}}\varphi_i(\xi_i) - \langle g_i(\xi_i),\ y_i\rangle - \langle\mathbf{w},\ M\mathbf{y}\rangle :\ y_i \in \mathcal{K}^\circ\ \forall i \in \mathcal{N}\}$. In Figure 2, we propose a modified version of the DPDA-S algorithm [2] (see Section 1.2) to solve (1.2) over $\mathcal{G}$.

```
Algorithm DPDA-S ( $\boldsymbol{\xi}^0, \gamma, \{\tau_i, \kappa_i\}_{i\in\mathcal{N}}$ )
Initialization: $y_i^0 \leftarrow 0,\ s_i^0 \leftarrow 0,\quad i \in \mathcal{N}$
Step $k$: ($k \geq 0$)
  1. $\xi_i^{k+1} \leftarrow \mathbf{prox}_{\tau_i\rho_i}\Big(\xi_i^k - \tau_i\Big(\nabla f_i(\xi_i^k) - \mathbf{J}g_i^\top y_i^k\Big)\Big),\quad p_i^{k+1} \leftarrow \sum_{j\in\mathcal{N}_i}(s_i^k - s_j^k),\quad i \in \mathcal{N}$
  2. $y_i^{k+1} \leftarrow \mathcal{P}_{\mathcal{K}^\circ\cap\mathcal{B}_0}\Big[y_i^k - \kappa_i\Big(2g_i(\xi_i^{k+1}) - g_i(\xi_i^k) + \gamma p_i^{k+1}\Big)\Big],\quad i \in \mathcal{N}$
  3. $s_i^{k+1} \leftarrow y_i^{k+1} + \sum_{\ell=0}^{k+1} y_i^\ell,\quad i \in \mathcal{N}$
```

FIG. 2. *Distributed PDA for static $\mathcal{G}$ (DPDA-S).*

DPDA-S needs only *one* communication round per iteration; moreover, since DPDA-S does not require inexact averaging (hence no error accumulation), its analysis is much simpler than and directly follows from the analysis of DPDA-D. The following theorem states the convergence rate for the iterates of DPDA-S.

THEOREM 5.1. *Suppose Assumptions* 1, 2 *with* $\mathcal{G}^t = \mathcal{G}$ *for* $t \geq 0$ *and* 3 *hold. For any* $\gamma > 0$, *let the primal-dual step-sizes* $\{\tau_i, \kappa_i\}_{i\in\mathcal{N}}$ *be chosen such that*

$$\tau_i = (\max\{1, L_{f_i} + \beta L_{g_i}\} + C_{g_i})^{-1}, \quad \kappa_i = (C_{g_i} + \gamma(4d_{\max} + \tfrac{1}{2}))^{-1} \quad \forall\ i \in \mathcal{N}. \tag{5.1}$$

*for some* $\beta > 0$. *Given* $B \in (0, \infty]$, *let* $\mathcal{B}_0 \triangleq \{y \in \mathbb{R}^m : \ \|y\| \leq 2B\}$. *Starting from* $\mathbf{s}^0 = \mathbf{y}^0 = \mathbf{0}$ *and an arbitrary* $\boldsymbol{\xi}^0$, *let* $\{(\boldsymbol{\xi}^k)\}_{k\geq 0}$ *be the primal, and* $\{\mathbf{y}^k\}_{k\geq 0}$ *be the dual, iterate sequence generated by DPDA-S, displayed in Figure* 2. *For any* $\gamma > 0$, *if* $\beta > 0$ *is chosen as discussed below, then* $\{(\boldsymbol{\xi}^k, \mathbf{y}^k)\}_{k\geq 0}$ *converges to* $(\boldsymbol{\xi}^*, \mathbf{y}^*)$ *such that* $\mathbf{y}^* = \mathbf{1} \otimes y^*$ *and* $(\boldsymbol{\xi}^*, y^*)$ *is an optimal primal-dual solution to* (1.2). *Moreover, both infeasibility,* $F(\bar{\boldsymbol{\xi}}^K, \bar{\mathbf{y}}^K)$, *and suboptimality,* $|\varphi(\bar{\boldsymbol{\xi}}^K) - \varphi(\boldsymbol{\xi}^*)|$, *are* $\mathcal{O}(1/K)$, *i.e.,*

$$F(\bar{\boldsymbol{\xi}}^K, \bar{\mathbf{y}}^K) \triangleq \|M\bar{\mathbf{y}}\| + \|y^*\|\, d_{\mathcal{K}}\left(-g(\bar{\boldsymbol{\xi}}^K)\right) \leq \frac{\Lambda(\gamma,\beta)}{K}, \tag{5.2}$$

$$0 \leq \varphi(\bar{\boldsymbol{\xi}}^K) - \varphi(\boldsymbol{\xi}^*) + \|y^*\|\, d_{\mathcal{K}}\left(-g(\bar{\boldsymbol{\xi}}^K)\right) \leq \frac{\Lambda(\gamma,\beta)}{K} - F(\bar{\boldsymbol{\xi}}^K, \bar{\mathbf{y}}^K) \tag{5.3}$$

$\forall K \geq 1$, *where* $\Lambda(\gamma,\beta) = \frac{1}{2\gamma} + \sum_{i\in\mathcal{N}} \frac{1}{\tau_i}\|\xi_i^* - \xi_i^0\|^2 + \frac{4}{\kappa_i}\|y^*\|^2$.

Case 1. *If a dual bound is known, i.e.,* $B < \infty$, *then* (5.2) *and* (5.3) *hold for* $\beta = 2B$; *moreover, setting* $\gamma = (NB)^{-1}$ *gives* $\Lambda(\gamma,\beta) = \mathcal{O}(NB(\bar{R}_x^2 + \bar{C}_g B + 1))$.

Case 2. *If the dual bound does not exist, then set* $B = \infty$ *within DPDA-S. There exists* $\bar{\beta} > 0$ *such that* (5.2) *and* (5.3) *hold* $\forall \beta \geq \bar{\beta}$; *moreover, selecting* $\gamma = \bar{R}_x^2\sqrt{N/d_{\max}}$ *leads to* $\Lambda(\gamma,\beta) = \mathcal{O}(N^{\frac{3}{2}}\sqrt{d_{\max}}\bar{R}_x^2 \max\{1, \|y^*\|^2\})$ *and* $\bar{\beta} = \mathcal{O}(\sqrt{Nd_{\max}}\ \max\{1, \|y^*\|\})$ *for* $N$ *sufficiently large.*[2]

*Proof.* The results follow from the analysis of DPDA-D in section 3 and [2]. □

*Remark* 5.1. In [2, Theorem 2], the rate result is provided for the case that $g_i$ is affine for $i \in \mathcal{N}$. For this case, a dual bound is not needed; hence, the suboptimality and infeasibility rate is $\mathcal{O}(\Lambda/K)$ for some $\Lambda = \mathcal{O}(N(\bar{R}_x^2 + \bar{C}_g\,\|y^*\|^2))$ when $\gamma = 1/N$.

## 6. Computing a dual bound.

Recall that the definition of $\widetilde{\mathcal{C}}$ in (2.3) involves a bound $B$ such that $\|y^*\| \leq B$ for some dual optimal solution $y^*$. In this section, we show that given a Slater point we can find a ball containing the optimal dual set for problem (1.2). To this end, we first derive some results without assuming convexity.

Let $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ be arbitrary functions of $\boldsymbol{\xi}$ and $\mathcal{K} \subset \mathbb{R}^m$ be a cone. For now, we do not assume convexity for $\varphi$, $g$, and $\mathcal{K}$, which are the components of the following generic problem:

$$\varphi^* \triangleq \min_{\boldsymbol{\xi}} \varphi(\boldsymbol{\xi}) \quad \text{s.t.} \quad g(\boldsymbol{\xi}) \in -\mathcal{K} \ : \ y \in \mathcal{K}^\circ, \tag{6.1}$$

where $y \in \mathbb{R}^m$ denotes the dual vector. Let $q$ denote the dual function, i.e.,

$$q(y) \triangleq \begin{cases} \inf_{\boldsymbol{\xi}} \varphi(\boldsymbol{\xi}) - y^\top g(\boldsymbol{\xi}) & \text{if } y \in \mathcal{K}^\circ; \\ -\infty & \text{otherwise} \end{cases} \tag{6.2}$$

We assume that there exists $\hat{y} \in \mathcal{K}^\circ$ such that $q(\hat{y}) > -\infty$. Since $q$ is a closed concave function, this assumption implies that $-q$ is a *proper* closed convex function. Next we show that for any $\bar{y} \in \mathbf{dom}\, q = \{y \in \mathbb{R}^m : \ q(y) > -\infty\}$, the superlevel set $Q_{\bar{y}} \triangleq \{y \in \mathbf{dom}\, q : \ q(y) \geq q(\bar{y})\} \subset \mathcal{K}^\circ$ is contained in a Euclidean ball centered at the origin, of which the radius can be computed efficiently. A special case of this dual boundedness result is well known when $\mathcal{K} = \mathbb{R}^m_+$ [40]—see Lemma 1.1 in [33]; however, it is not trivial to extend this result to an *arbitrary* cone $\mathcal{K}$ with $\mathbf{int}(\mathcal{K}) \neq \emptyset$.

[2] For simple bounds, we assume $N > 1/\bar{R}_x^2$ and $\sqrt{Nd_{\max}} \geq \bar{C}_g/\bar{R}_x^2$.

LEMMA 6.1. *Let $\bar{\boldsymbol{\xi}}$ be a Slater point for* (6.1), *i.e.,* $\bar{\boldsymbol{\xi}} \in \mathbf{relint}(\mathbf{dom}\,\varphi)$ *such that* $-g(\bar{\boldsymbol{\xi}}) \in \mathbf{int}(\mathcal{K})$. *Then* $\forall \bar{y} \in \mathbf{dom}\, q$, *the superlevel set* $Q_{\bar{y}}$ *is bounded as follows:*

$$\|y\| \leq (\varphi(\bar{\boldsymbol{\xi}}) - q(\bar{y}))/r^* \quad \forall y \in Q_{\bar{y}}, \tag{6.3}$$

*where* $0 < r^* \triangleq \min_w \{-w^\top g(\bar{\boldsymbol{\xi}}) : \ \|w\|= 1, \ w \in \mathcal{K}^*\}$. *Note that this is not a convex problem due to the equality constraint; instead, one can upper bound* (6.3) *using* $0 < \tilde{r} \leq r^*$, *which can be efficiently computed by solving a convex problem*

$$\tilde{r} \triangleq \min_w \{-w^\top g(\bar{\boldsymbol{\xi}}) : \ \|w\|_1= 1, \ w \in \mathcal{K}^*\}. \tag{6.4}$$

*Proof.* For any $y \in Q_{\bar{y}} \subset \mathcal{K}^\circ$, we have that

$$q(\bar{y}) \leq q(y) = \inf_{\boldsymbol{\xi}} \{\varphi(\boldsymbol{\xi}) - y^\top g(\boldsymbol{\xi})\} \leq \varphi(\bar{\boldsymbol{\xi}}) - y^\top g(\bar{\boldsymbol{\xi}}), \tag{6.5}$$

which implies that $y^\top g(\bar{\boldsymbol{\xi}}) \leq \varphi(\bar{\boldsymbol{\xi}}) - q(\bar{y})$. Since $-g(\bar{\boldsymbol{\xi}}) \in \mathbf{int}(\mathcal{K})$ and $y \in \mathcal{K}^\circ$, we clearly have $y^\top g(\bar{\boldsymbol{\xi}}) > 0$ whenever $y \neq \mathbf{0}$. Indeed, since $-g(\bar{\boldsymbol{\xi}}) \in \mathbf{int}(\mathcal{K})$, there exist $r > 0$ such that $-g(\bar{\boldsymbol{\xi}}) + ru \in \mathcal{K}$ $\forall \|u\|\leq 1$. Hence, for $y \neq \mathbf{0}$, by choosing $u = y/\|y\|$ and using the fact that $y \in \mathcal{K}^\circ$, we get that $0 \geq (-g(\bar{\boldsymbol{\xi}}) + ry/\|y\|)^\top y$. Therefore, (6.5) implies that $\forall y \in Q_{\bar{y}}$, $r\|y\|\leq y^\top g(\bar{\boldsymbol{\xi}}) \leq \varphi(\bar{\boldsymbol{\xi}}) - q(\bar{y})$; hence, $\|y\|\leq \frac{\varphi(\bar{\boldsymbol{\xi}})-q(\bar{y})}{r}$. Now, we will characterize the largest radius $r^* > 0$ such that $\mathcal{B}(-g(\bar{\boldsymbol{\xi}}), r^*) \subset \mathcal{K}$, where $\mathcal{B}(-g(\bar{\boldsymbol{\xi}}), r) \triangleq \{-g(\bar{\boldsymbol{\xi}}) + ru : \ \|u\| \leq 1\}$. Note that $r^* > 0$ can be written explicitly as follows: $r^* = \max\{r : \ d_{\mathcal{K}}( - g(\bar{\boldsymbol{\xi}}) + ru) \leq 0 \ \forall u \text{ s.t. } \|u\|\leq 1\}$. Let $\gamma(r) \triangleq \sup\{d_{\mathcal{K}}( - g(\bar{\boldsymbol{\xi}}) + ru) : \ \|u\| \leq 1\}$; hence, $r^* = \max\{r : \ \gamma(r) \leq 0\}$. Note that for any fixed $u \in \mathbb{R}^m$, $d_{\mathcal{K}}( - g(\bar{\boldsymbol{\xi}}) + ru)$ as a function of $r$ is a composition of a convex function $d_{\mathcal{K}}(\cdot)$ with an affine function in $r$; hence, it is convex in $r \in \mathbb{R}$ $\forall u \in \mathbb{R}^m$. Moreover, since the supremum of convex functions is also convex, $\gamma(r)$ is convex in $r$. From the definition of $d_{\mathcal{K}}(\cdot)$, we have

$$\gamma(r) = \sup_{\|u\|\leq 1} \inf_{\boldsymbol{\xi}\in\mathcal{K}} \|\boldsymbol{\xi} + g(\bar{\boldsymbol{\xi}}) - ru\| = \sup_{\|u\|\leq 1} \inf_{\boldsymbol{\xi}\in\mathcal{K}} \sup_{\|w\|\leq 1} w^\top (\boldsymbol{\xi} + g(\bar{\boldsymbol{\xi}}) - ru). \tag{6.6}$$

Since $\{w \in \mathbb{R}^m : \ \|w\|\leq 1\}$ is a compact set, and the function in (6.6) is a bilinear function of $w$ and $\boldsymbol{\xi}$ for each $u$, the inner $\inf_{\boldsymbol{\xi}}$ and $\sup_w$ can be interchanged to obtain

$$\begin{aligned}\gamma(r) &= \sup_{\|u\|\leq 1} \sup_{\|w\|\leq 1} \inf_{\boldsymbol{\xi}\in\mathcal{K}} w^\top \Big(\boldsymbol{\xi} + g(\bar{\boldsymbol{\xi}}) - ru\Big) \\ &= \sup_{\substack{\|u\|\leq 1 \\ \|w\|\leq 1 \\ w\in\mathcal{K}^*}} w^\top (g(\bar{\boldsymbol{\xi}}) - ru) = \sup_{\substack{\|w\|\leq 1 \\ w\in\mathcal{K}^*}} w^\top g(\bar{\boldsymbol{\xi}}) + r\|w\|.\end{aligned}$$

Let $w^*(r)$ be one of the maximizers. It is easy to see that $\|w^*(r)\| = 1$, since the supremum of a convex function over a convex set is attained on the boundary of the set. Therefore,

$$\gamma(r) = \sup_{\substack{\|w\|=1 \\ w\in\mathcal{K}^*}} w^\top g(\bar{\boldsymbol{\xi}}) + r.$$

Since $r^* = \max\{r : \ \gamma(r) \leq 0\}$,

$$(P_1): \qquad r^* = \max\Big\{r : \ r \leq -\sup\{w^\top g(\bar{\boldsymbol{\xi}}) : \ \|w\|= 1, \ w \in \mathcal{K}^*\}\Big\} = \min_{\substack{\|w\|=1 \\ w\in\mathcal{K}^*}} -w^\top g(\bar{\boldsymbol{\xi}}).$$

Note that $(P_1)$ is not a convex problem due to the boundary constraint, $\|w\|= 1$. Next, we define a related convex problem:

$$\min_{\substack{\|w\|_1=1\\ w\in\mathcal{K}^*}} -w^\top g(\bar{\boldsymbol{\xi}}) \le r^* = \min_{\substack{\|w\|=1\\ w\in\mathcal{K}^*}} -w^\top g(\bar{\boldsymbol{\xi}})$$

to lowerbound $r^*$ so that we can upper bound the right-hand side of (6.3). Let $w^*$ be an optimal solution to $(P_1)$ and define $\bar{w} = w^*/\|w^*\|_1$. Clearly, $\|\bar{w}\|_1= 1$ and $\bar{w} \in \mathcal{K}^*$. Moreover, since $\|w^*\|_1\ge \|w^*\|= 1$ we have that

$$0 < \tilde{r} = \min_{\substack{\|w\|_1=1\\ w\in\mathcal{K}^*}} -w^\top g(\bar{\boldsymbol{\xi}}) \le -\bar{w}^\top g(\bar{\boldsymbol{\xi}}) = -\frac{1}{\|w^*\|_1} {w^*}^\top g(\bar{\boldsymbol{\xi}}) \le -{w^*}^\top g(\bar{\boldsymbol{\xi}}) = r^*. \qquad \square$$

*Remark* 6.1. Consider the problem in (1.2). Given a Slater point $\bar{\boldsymbol{\xi}}$, one needs to solve the minimization problem (6.4) in a distributed fashion, e.g., using the method in [3], to obtain a dual bound $B \in (0, +\infty)$. Suppose $\varphi_i(\cdot) \ge \underline{\varphi}$ $\forall i \in \mathcal{N}$ and $N$ is known by all agents. Once $\tilde{r}$, the optimal value to (6.4), is computed, one can set $B = (\varphi(\bar{\boldsymbol{\xi}}) - N\underline{\varphi})/\tilde{r}$, i.e., $\bar{y} = \mathbf{0}$. Moreover, if a Slater point exists but is not available, one can solve the problem of $\bar{\boldsymbol{\xi}} = \operatorname{argmin}_{\boldsymbol{\xi}}\ \mathcal{F}(\sum_{i\in\mathcal{N}} g_i(\xi_i))$ in a distributed fashion using methods proposed in [12] to obtain a Slater point where $\mathcal{F} : \mathbb{R}^m \to \mathbb{R}$ is a generalized logarithm function for the proper cone $\mathcal{K}$ (see [7, section 11.6.1] for the definition). Next, $B$ can be computed as discussed previously.

*Remark* 6.2. Let $g_j : \mathbb{R}^n \to \mathbb{R}$ be the components of $g : \mathbb{R}^n \to \mathbb{R}^m$ for $j = 1, \ldots, m$, i.e., $g(\boldsymbol{\xi}) = [g_j(\boldsymbol{\xi})]_{j=1}^m$. When $\mathcal{K} = \mathbb{R}^m_+$, [33, Lemma 1.1] implies that for any $\bar{y} \in \mathbf{dom}\, q$ and $\bar{\boldsymbol{\xi}}$ such that $g_j(\bar{\boldsymbol{\xi}}) < 0$ $\forall j = 1, \ldots, m$, every $y \in Q_{\bar{y}}$ satisfies $\|y\| \le (\varphi(\bar{\boldsymbol{\xi}}) - q(\bar{y}))/\bar{r}$, where $\bar{r} \triangleq \min\{-g_j(\bar{\boldsymbol{\xi}}) :\ j = 1, \ldots, m\}$. Note that our result in Lemma 6.1 gives the same bound since $r^* = \min_w\{-w^\top g(\bar{\boldsymbol{\xi}}) :\ \|w\|= 1,\ w \in \mathbb{R}^m_+\} = \bar{r}$.

## 7. Numerical experiments.

We implemented the DPDA-D algorithm and tested its performance on two different sets of problems.

### 7.1. Basis pursuit denoising (BPD) problem.

Let $\boldsymbol{\xi}^* \in \mathbb{R}^n$ be an unknown sparse vector, i.e., most of its elements are zero. Suppose $r \in \mathbb{R}^m$ denotes a vector of $m \ll n$ noisy linear measurements of $\boldsymbol{\xi}^*$ using the measurement matrix $R \in \mathbb{R}^{m\times n}$, i.e., $\|R\boldsymbol{\xi}^* - r\| \le \epsilon$ for some $\epsilon \ge 0$. The BPD problem can be formulated as

$$\min_{\boldsymbol{\xi}}\ \|\boldsymbol{\xi}\|_1 \quad \text{s.t.} \quad \|R\boldsymbol{\xi} - r\| \le \epsilon. \tag{7.1}$$

BPD appears in the context of compressed sensing [18] and the objective is to recover the unknown sparse $\boldsymbol{\xi}^*$ from a small set of measurement or transform values in $r$.

Given a set of computing nodes $\mathcal{N}$, suppose each node $i \in \mathcal{N}$ knows $r \in \mathbb{R}^m$ and stores only $n_i$ columns of $R$ corresponding to a submatrix $R_i \in \mathbb{R}^{m\times n_i}$ such that $n = \sum_{i\in\mathcal{N}} n_i$ and $R = [R_i]_{i\in\mathcal{N}}$. Partitioning the decision vector $\boldsymbol{\xi} = [\xi_i]_{i\in\mathcal{N}}$ accordingly, the BPD problem in (7.1) can be rewritten as follows:

$$\min_{\xi_i\in\mathbb{R}^{n_i},\ i\in\mathcal{N}}\ \sum_{i\in\mathcal{N}} \|\xi_i\|_1 \quad \text{s.t.} \quad \|\sum_{i\in\mathcal{N}} R_i\xi_i - r\| \le \epsilon. \tag{7.2}$$

Note that (7.2) can be cast into a form similar to (1.2). Indeed, let $\chi : \mathbb{R} \to \mathbb{R}\cup\{+\infty\}$ such that $\chi(t) = 0$ if $t = \epsilon$, and $+\infty$ otherwise; and let $\mathcal{K}$ denote the second-order cone, i.e., $\mathcal{K} = \{(y,t) \in \mathbb{R}^m \times \mathbb{R} : \|y\| \leq t\}$. Hence, (7.2) can be written as

$$\min_{t\in\mathbb{R},\xi_i\in\mathbb{R}^{n_i},\ i\in\mathcal{N}} \sum_{i\in\mathcal{N}} \|\xi_i\|_1 + \chi(t) \quad \text{s.t.} \quad \Big(\sum_{i\in\mathcal{N}} R_i\xi_i - r, t\Big) \in \mathcal{K}.$$

First, we test the effect of network topology on the performance of the proposed algorithm, and then to benchmark this distributed algorithm, we also solve the same problem in a centralized way using the Prox-JADMM algorithm proposed in [15]. Note that Prox-JADMM solves the problem in a centralized fashion, which naturally has a faster convergence than a decentralized algorithm. The aim of this comparison is to show that the convergence of the proposed decentralized algorithm would be competitive with a *centralized* method when the nature of the problem requires one to store and access the data in a decentralized manner. In the online technical report [1], we also examined the performance of the DPDA-S algorithm [2] and benchmarked it against Prox-JADMM as well.

**7.1.1. Problem generation.** In what follows, we consider two different forms of the problem in (7.1): noisy, i.e., $\epsilon > 0$, and noise free, i.e., $\epsilon = 0$. In our experiments, we set $n = 120$ and $m = 20$. For the noisy case, as suggested in [4], the target signal $\boldsymbol{\xi}^*$ is generated by choosing $\kappa = 20$ of its elements, uniformly at random, drawn from the standard Gaussian distribution and the rest of the elements are set to 0. Moreover, each element of $R = [R_{ij}]$ is independent and identically distributed (i.i.d.) with standard normal distribution, and the measurement $r = R\boldsymbol{\xi}^* + \eta$, where $\eta \in \mathbb{R}^m$ such that each of its elements is i.i.d. according to Gaussian distribution with mean 0 and variance $\sigma^2 = \kappa\ 10^{-S/10}$—this would generate a measurement vector $r$ with the signal-to-noise ratio (SNR) equal to $S$, where $\mathrm{SNR}(r) \triangleq 10\log_{10}(\mathbb{E}[\|R\boldsymbol{\xi}^*\|^2]/\mathbb{E}[\|\eta\|^2])$. In our experiments, we consider $S = 30$dB or 40dB. Finally, $\epsilon > 0$ is chosen such that $\Pr(\|\eta\|^2 \leq \epsilon^2) = 1 - \alpha$, and we let $\alpha = 0.05$. For the noise-free case, the noise parameters, i.e., $\sigma^2$ and $\epsilon$, are set to 0; hence, the constraint for the noise-free case is a linear one, i.e., $\sum_{i\in\mathcal{N}} R_i\xi_i = r$—the rest of the problem components are generated as in the noisy case.

**Generating an undirected small-world network.** Let $\mathcal{G}_u = (\mathcal{N}, \mathcal{E}_u)$ be generated as a random small-world network. Given $|\mathcal{N}|$ and the desired number of edges $|\mathcal{E}_u|$, we choose $|\mathcal{N}|$ edges creating a random cycle over nodes, and then the remaining $|\mathcal{E}_u| - |\mathcal{N}|$ edges are selected uniformly at random.

**Generating a time-varying undirected network.** We first generate a random small-world $\mathcal{G}_u = (\mathcal{N}, \mathcal{E}_u)$ as described above. Next, given $M \in \mathbb{Z}_+$, and $p \in (0,1)$, for each $k \in \mathbb{Z}_+$, we generate $\mathcal{G}^t = (\mathcal{N}, \mathcal{E}^t)$, the communication network at time $t \in \{(k-1)M, \ldots, kM-2\}$ by sampling $\lceil p\ |\mathcal{E}_u|\rceil$ edges of $\mathcal{G}_u$ uniformly at random and we set $\mathcal{E}^{kM-1} = \mathcal{E}_u \setminus \bigcup_{t=(k-1)M}^{kM-2} \mathcal{E}^t$. In all experiments, we set $M = 5$, $p = 0.8$ and the number of communications per iteration is set to $q_k = 10\log(k+1)$.

**7.1.2. Effect of network topology.** In this section, we test the effect of network topology on the performance of DPDA-D on *undirected* communication networks. We consider four scenarios in which the number of nodes $N \in \{10,\ 40\}$ and the average number of edges per node, $|\mathcal{E}^t|/N$, is either 1.2 or $\approx 3.6$. For each scenario, we plot relative suboptimality, i.e., $|\varphi(\boldsymbol{\xi}^k) - \varphi(\boldsymbol{\xi}^*)|/|\varphi(\boldsymbol{\xi}^*)|$, infeasibility, i.e., $(\|\sum_{i\in\mathcal{N}} R_i\xi_i^k - r\| - \epsilon)_+$, and consensus violation, i.e., $\max_{i\in\mathcal{N}} \|y_i^k - \frac{1}{|\mathcal{N}|}\sum_{j\in\mathcal{N}} y_j^k\|$ versus iteration number $k$. All the plots show the average statistics over 50 randomly generated replications. In each of these independent replications, both $R$ and $\boldsymbol{\xi}^*$ are also randomly generated in addition to random communication networks.

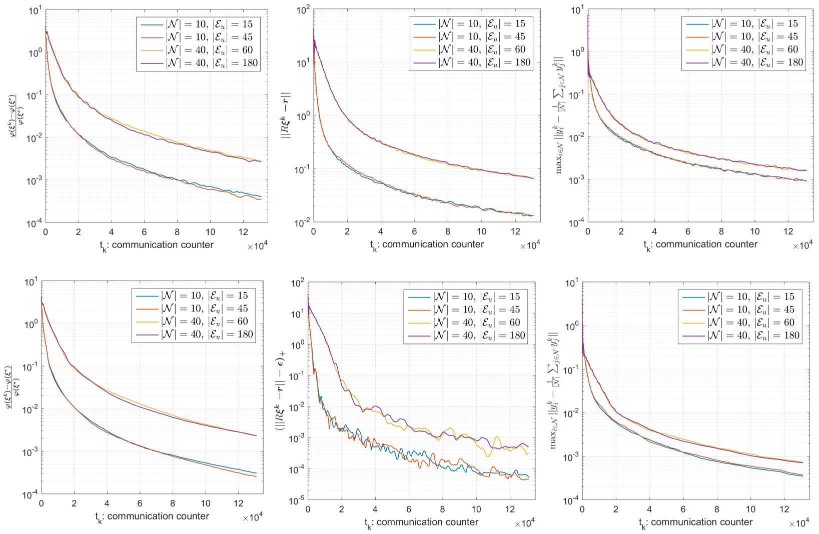

FIG. 3. *Effect of network topology on the convergence of DPDA-D: the top row corresponds to noise free and the bottom row corresponds to noisy experiments with* $S = 30$dB.

**Testing DPDA-D on time-varying undirected communication networks.** We first generated an undirected small-world network $\mathcal{G}_u = (\mathcal{N}, \mathcal{E}_u)$ as described earlier. Next, we generated $\{\mathcal{G}^t\}_{t\geq 0}$ as described in section 7.1.1. We chose the initial point $\boldsymbol{\xi}^0$ of DPDA-D such that the components are i.i.d with the standard uniform distribution and set the step-sizes as follows: $\gamma = 1$, $\tau_i = \frac{1}{|\mathcal{N}|\|R_i\|}$, and $\kappa_i = \frac{1}{\gamma + \|R_i\|/|\mathcal{N}|}$ for $i \in \mathcal{N}$. The performance of DPDA-D in terms of suboptimality, infeasibility, and consensus violation is displayed in Figure 3. It is clear that when compared to the effect of average edge density, the network size $|\mathcal{N}|$ has more influence on the convergence rate, i.e., the smaller the network the faster the convergence is; however, the average edge density does not seem to have a significant impact on the convergence.

**7.1.3. Benchmarking DPDA-D against a centralized algorithm.** In this section we benchmark DPDA-D on both undirected and directed networks against the Prox-JADMM algorithm on BPD problems under three different noise levels: $S = 30$ dB, $S = 40$ dB, and noise free, i.e., $S = +\infty$ dB. Prox-JADMM is a multiblock ADMM using Jacobian type updates and the block-$i$ update has an additional proximal term $\frac{1}{2}\left\|\xi_i - \xi_i^k\right\|_{P_i}^2$ for each $i \in \mathcal{N}$, where $\{P_i\}_{i\in\mathcal{N}}$ are positive-definite matrices satisfying certain conditions. We choose the parameters for the Prox-JADMM algorithm as suggested in section 3.2 of [15], i.e., by setting the matrix $P_i$ in the proximal term to be $P_i = (N\mathbf{I} - 10\ R_i^\top R_i)/\|r\|_1$ for $i \in \mathcal{N}$, and $\{P_i\}_{i\in\mathcal{N}}$ are adaptively updated by the strategy discussed in section 2.3 of [15].

**Time-varying undirected network.** For undirected time-varying networks we fix $N = 10$ and $|\mathcal{E}^t|/N = 1.2$, i.e., $|\mathcal{E}_u|/N = 1.5$—we observe the same convergence behavior for the other network scenarios discussed in section 7.1.2. In each replication, we generate the network sequence $\{\mathcal{G}^t\}_{t\geq 0}$ and choose the parameters as in time-varying network experiments of section 7.1.2. Figure 4 shows the comparison between the two methods in terms of suboptimality, infeasibility and consensus violation. We observe that different noise-levels lead to similar convergence patterns; however, the lower SNR ratio leads to faster convergence, and the noise-free case has the slowest

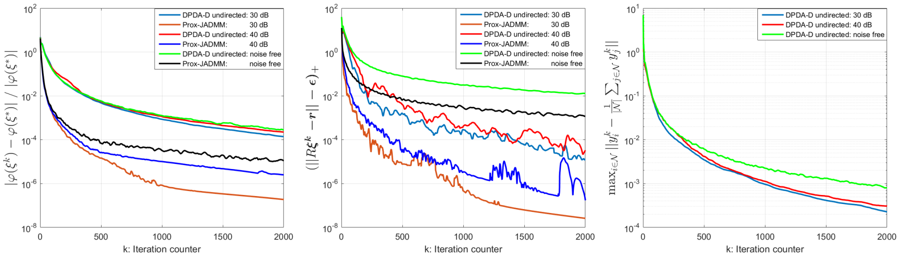

FIG. 4. *Comparison of DPDA-D and Prox-JADMM over an undirected time-varying network for three different noise levels.*

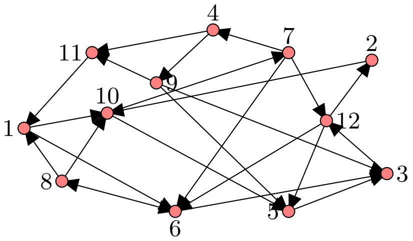

FIG. 5. $\mathcal{G}_d = (\mathcal{N}, \mathcal{E}_d)$ *directed strongly connected graph.*

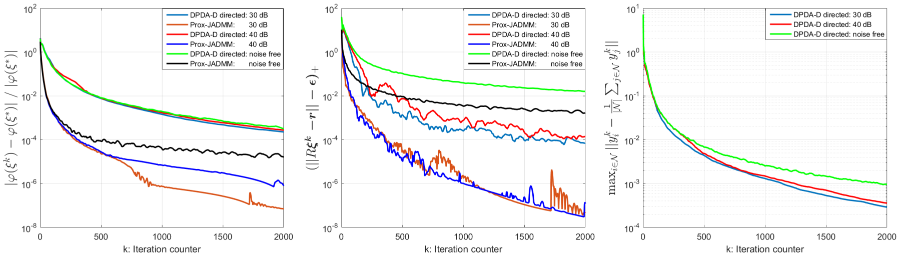

FIG. 6. *Comparison of DPDA-D and Prox-JADMM over a directed time-varying network with three noise levels.*

convergence. For all noise levels DPDA-D is competitive against Prox-JADMM—a slightly slower rate of DPDA-D is the price we pay for the decentralized setting to reach consensus on the dual price over the time-varying network.

**Time-varying directed network.** In this scenario, similar to [34] we consider the *strongly connected* directed graph $\mathcal{G}_d = (\mathcal{N}, \mathcal{E}_d)$ in Figure 5 with $N = 12$ nodes and $|\mathcal{E}_d| = 24$ directed edges. We generated $\{\mathcal{G}^t\}_{t\geq 0}$ as in the undirected case, but using $\mathcal{G}_d$ instead of $\mathcal{G}_u$, with parameters $M = 5$, $p = 0.8$, and $q_k = 10\log(k+1)$; hence, $\{\mathcal{G}^t\}_{t\geq 0}$ is $M$-strongly-connected. Moreover, communication weight matrices $V^t$ are formed according to rule (2.13), and we used the approximate averaging operator $\mathcal{R}^k$ given in (2.14). We set the step-sizes as in the time-varying undirected case. Figure 6 illustrates the comparison between DPDA-D and Prox-JADMM in terms of suboptimality, infeasibility, and consensus violation when the network is both time-varying and directed. The results of this experiment are similar to those for the time-varying undirected case; hence, using unidirectional communications instead of bidirectional did not adversely affect the convergence of DPDA-D.

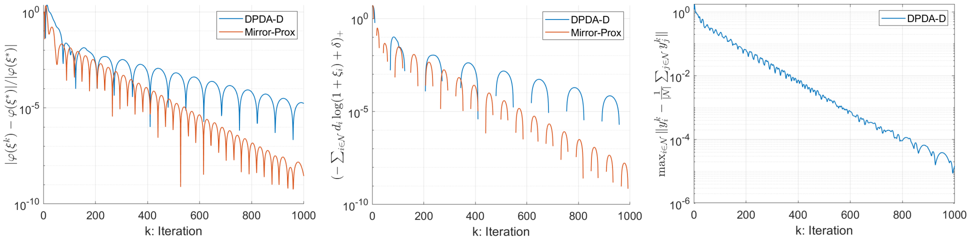

FIG. 7. *Comparison between DPDA-D and mirror-prox.*

**7.2. Multichannel power allocation problem.** Multichannel power allocation is a classic problem in information theory. Suppose there are a set of nodes connected to each other over a time-varying wireless communication network and all transmitting information to a receiver. Let the communication graph be $\mathcal{G}^t = (\mathcal{N}, \mathcal{E}^t)$ at time $t > 0$. Each node $i \in \mathcal{N}$ transmits information over a different channel with bandwidth $b_i$ (given) with a signal transmission power $s_i \in [0, u_i]$ watts and the signal is exposed to Gaussian white (uncorrelated) noise of additive nature, with power $w_i$ watts (given). According to the Shannon–Hartley equation the maximum capacity of the channel associated with node $i \in \mathcal{N}$ is $b_i \log_2(1 + s_i/w_i)$. Suppose we want to minimize the total power of the system subject to certain capacity requirement $\delta > 0$, i.e., $\min\{\sum_{i\in\mathcal{N}} s_i : \sum_{i\in\mathcal{N}} b_i \log_2(1 + s_i/w_i) \geq \delta,\ 0 \leq s_i \leq u_i\}$.

For numerical experiments, we consider the particular setup described in [26]:

$$\min_{\boldsymbol{\xi}=[\xi_i]_{i\in\mathcal{N}}} \sum_{i\in\mathcal{N}} c_i\xi_i \quad \text{s.t.} \quad \sum_{i\in\mathcal{N}} b_i \log(1 + \xi_i) \geq \delta, \quad \boldsymbol{\xi} \in [0,1]^{|\mathcal{N}|}, \tag{7.3}$$

where $\mathbf{c} = [c_i]_{i\in\mathcal{N}} \in \mathbb{R}^{|\mathcal{N}|}$ and $\mathbf{b} = [b_i]_{i\in\mathcal{N}} \in \mathbb{R}^{|\mathcal{N}|}$ are chosen uniformly at random between 0 and 1. We consider both static and dynamic networks; dynamic ones are generated as in section 7.1.1 with $|\mathcal{N}| = 50$ nodes and $|\mathcal{E}_u| = 150$ edges, and the static one is set to $\mathcal{G} = (\mathcal{N}, \mathcal{E}_u)$. In the experiments we set $\delta = 5$. For benchmarking, we compared our algorithm against Consensus-Based Saddle-Point Subgradient (CoBa-SPS)[3] [26] and Mirror-prox [22]—the former one is a decentralized algorithm while the latter one is a centralized algorithm. The Mirror-prox algorithm requires the global Lipschitz constant of $\nabla\mathcal{L}$, where $\mathcal{L}(\boldsymbol{\xi}, \mathbf{y}) = \mathbf{c}^\top\boldsymbol{\xi} + \left\langle \sum_{i\in\mathcal{N}} b_i \log(1 + \xi_i) - \delta,\ \mathbf{y} \right\rangle$ for $\boldsymbol{\xi} \in [0,1]^{|\mathcal{N}|}$, which is $\sqrt{2}\,\|\mathbf{b}\|$. Mirror-prox and CoBa-SPS also require a bound on the dual solutions. Similar to [26], for the Slater point $\bar{\boldsymbol{\xi}} = \mathbf{1}_{|\mathcal{N}|}$, we have that $\|y^*\| \leq \frac{N \max_{i\in\mathcal{N}} c_i}{\log(2)(\sum_{i\in\mathcal{N}} b_i) - \delta}$. We compare DPDA-S against CoBa-SPS and Mirror-prox. Since CoBa-SPS can only handle a static network, when the network topology is time-varying, we compare DPDA-D only against Mirror-prox, where we set $q_k = 10\log(k + 1)$ within DPDA-D. We choose our step-sizes according to (3.1), where $L_{f_i} = 0$, $L_{g_i} = C_{g_i} = b_i$, and we set $\gamma = 1/|\mathcal{N}|$. Figure 7 shows the performance of DPDA-D in terms of suboptimality and infeasibility as well as consensus violation. The performance of our method is comparable with the centralized Mirror-prox and a slightly slower rate of DPDA-D is the price we pay for the decentralized setting. Figure 8 compares the performance of DPDA-S against CoBa-SPS and Mirro-prox. Although CoBa-SPS finds a feasible solution and remains feasible, the iterates are far

[3] The code is available online and it is used to implement problem (7.3).

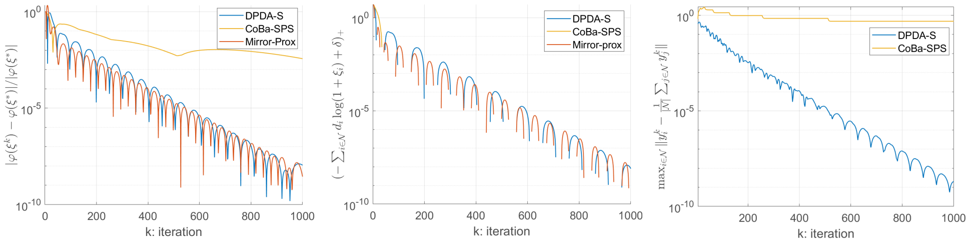

FIG. 8. *Comparison among DPDA-S, Mirror-prox, and CoBa-SPS.*

from optimality; and both suboptimality and consensus violations decrease with a slow rate. DPDA-S has a superior performance compared to CoBa-SPS. The discontinuity within infeasibility plots (middle figures) is due to achieving occasional feasibility when both primal and dual iterates are approaching their optimal solutions.

## 8. Conclusions.

We propose a distributed primal-dual algorithm, DPDA-D, for solving cooperative multiagent convex resource sharing problems over time-varying (un)directed communication networks, where only local communications are allowed. The objective is to minimize the sum of agent-specific composite convex functions subject to a conic constraint that couples agents' decisions. We show that the DPDA-D iterate sequence converges to $\epsilon$-suboptimality/infeasibility within $\mathcal{O}(1/\epsilon)$ number of iterations. To the best of our knowledge, this is the best rate result for our setting. Moreover, DPDA-D employs agent-specific constant step-sizes using local information. As a potential future work, we plan to analyze convergence rates of similar primal-dual algorithms under certain strong convexity assumptions.

## 9. Appendix.

LEMMA 9.1 (see [36]). *Let $\{a^k\}$, $\{b^k\}$, $\{c^k\}$, and $\{d^k\}$ be nonnegative real sequences such that $a^{k+1} \leq (1+d^k)a^k - b^k + c^k$ $\forall k \geq 0$, $\sum_{k=0}^{\infty} c^k < \infty$, and $\sum_{k=0}^{\infty} d^k < \infty$. Then $a = \lim_{k\to\infty} a^k$ exists, and $\sum_{k=0}^{\infty} b^k < \infty$.*

LEMMA 9.2. *Assume that $\{u_k\}_{k=0}^K \subset \mathbb{R}_+$ satisfies $u_0^2 \leq S_0$ and $u_k^2 \leq S_k + \sum_{i=1}^k \lambda_i u_i$ $\forall k \in \{1,\ldots,K\}$ for some $\{S_k\}_{k=0}^K$ nondecreasing in $k$ and $\{\lambda_k\}_{k=1}^K \subset \mathbb{R}_+$. Then, the following inequality holds $\forall k \in \{1,\ldots,K\}$:*

$$u_k \leq \frac{1}{2}\sum_{i=1}^k \lambda_i + \left(S_k + \Big(\frac{1}{2}\sum_{i=1}^k \lambda_i\Big)^2\right)^{1/2}.$$

## REFERENCES

[1] N. S. AYBAT AND E. Y. HAMEDANI, *A Distributed ADMM-Like Method for Resource Sharing Under Conic Constraints over Time-Varying Networks*, https://arxiv.org/abs/1611.07393, 2016.

[2] N. S. AYBAT AND E. Y. HAMEDANI, *Distributed primal-dual method for multi-agent sharing problem with conic constraints*, in Proceedings of the 50th Asilomar Conference on Signals, Systems and Computers, 2016, pp. 777–782.

[3] N. S. AYBAT AND E. Y. HAMEDANI, *A primal-dual method for conic constrained distributed optimization problems*, in Proceedings of the Advances in Neural Information Processing Systems, 2016, pp. 5049–5057.

[4] N. S. AYBAT AND G. IYENGAR, *A first-order augmented Lagrangian method for compressed sensing*, SIAM J. Optim., 22 (2012), pp. 429–459.

[5] N. S. Aybat, Z. Wang, T. Lin, and S. Ma, *Distributed linearized alternating direction method of multipliers for composite convex consensus optimization*, IEEE Trans. Automat. Control, 63 (2017), pp. 5–20.

[6] S. Boyd, P. Diaconis, and L. Xiao, *Fastest mixing Markov chain on a graph*, SIAM Rev., 46 (2004), pp. 667–689.

[7] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

[8] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vision, 40 (2011), pp. 120–145.

[9] A. Chambolle and T. Pock, *On the ergodic convergence rates of a first-order primal-dual algorithm*, Math. Program., 159 (2016), pp. 253–287.

[10] T.-H. Chang, *A proximal dual consensus ADMM method for multi-agent constrained optimization*, IEEE Trans. Signal Process., 64 (2016), pp. 3719–3734.

[11] T.-H. Chang, M. Hong, and X. Wang, *Multi-agent distributed optimization via inexact consensus admm*, IEEE Trans. Signal Process., 63 (2015), pp. 482–497.

[12] T.-H. Chang, A. Nedic, and A. Scaglione, *Distributed constrained optimization by consensus-based primal-dual perturbation method*, IEEE Trans. Automat. Control, 59 (2014), pp. 1524–1538.

[13] A. I. Chen and A. Ozdaglar, *A fast distributed proximal-gradient method*, Proceedings of the IEEE 50th Annual Allerton Conference on Communication, Control, and Computing, 2012, pp. 601–608.

[14] Y. Chen, G. Lan, and Y. Ouyang, *Optimal primal-dual methods for a class of saddle point problems*, SIAM J. Optim., 24 (2014), pp. 1779–1814.

[15] W. Deng, M.-J. Lai, Z. Peng, and W. Yin, *Parallel multi-block ADMM with $o(1/k)$ convergence*, J. Sci. Comput., 71 (2017), pp. 712–736.

[16] T. T. Doan and C. L. Beck, *Distributed Lagrangian Methods for Network Resource Allocation*, https://arxiv.org/abs/1609.06287, 2017.

[17] T. T. Doan and A. Olshevsky, *Distributed resource allocation on dynamic networks in quadratic time*, Systems Control Lett., 99 (2017), pp. 57–63.

[18] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.

[19] C. Gu, Z. Wu, J. Li, and Y. Guo, *Distributed Convex Optimization with Coupling Constraints Over Time-Varying Directed Graphs*, https://arxiv.org/abs/1805.07916, 2018.

[20] F. Guo, C. Wen, J. Mao, and Y.-D. Song, *Distributed economic dispatch for smart grids with random wind power*, IEEE Trans. Smart Grid, 7 (2016), pp. 1572–1583.

[21] B. He and X. Yuan, *Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective*, SIAM J. Imaging Sci., 5 (2012), pp. 119–149.

[22] N. He, A. Juditsky, and A. Nemirovski, *Mirror prox algorithm for multi-term composite minimization and semi-separable problems*, Comput. Optim. Appl., 61 (2015), pp. 275–319.

[23] H. Huang, Q. Ling, W. Shi, and J. Wang, *Collaborative resource allocation over a hybrid cloud center and edge server network.*, J. Comput. Math., 35 (2017).

[24] D. Kempe, A. Dobra, and J. Gehrke, *Gossip-based computation of aggregate information*, in Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science, IEEE, 2003, pp. 482–491.

[25] H. Lakshmanan and D. P. De Farias, *Decentralized resource allocation in dynamic networks of agents*, SIAM J. Optim., 19 (2008), pp. 911–940.

[26] D. Mateos-Núnez and J. Cortés, *Distributed saddle-point subgradient algorithms with Laplacian averaging*, IEEE Trans. Automat. Control, 62 (2017), pp. 2720–2735.

[27] A. Nedić and A. Olshevsky, *Distributed optimization over time-varying directed graphs*, IEEE Trans. Automat. Control, 60 (2015), pp. 601–615.

[28] A. Nedić and A. Olshevsky, *Stochastic gradient-push for strongly convex functions on time-varying directed graphs*, IEEE Trans. Automat. Control, 61 (2016), pp. 3936–3947.

[29] A. Nedic, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, *On distributed averaging algorithms and quantization effects*, IEEE Trans. Automat. Control, 54 (2009), pp. 2506–2517.

[30] A. Nedić, A. Olshevsky, and W. Shi, *Improved Convergence Rates for Distributed Resource Allocation*, https://arxiv.org/abs/1706.05441, 2017.

[31] A. Nedić and A. Ozdaglar, *Distributed subgradient methods for multi-agent optimization*, IEEE Trans. Automat. Control, 54 (2009), pp. 48–61.

[32] A. Nedić and A. Ozdaglar, *Subgradient methods for saddle-point problems*, J. Optim. Theory Appl., 142 (2009), pp. 205–228.

[33] A. Nedic and A. Ozdaglar, *Cooperative distributed multi-agent optimization*, Convex Optimization in Signal Processing and Communications, Cambridge University Press, Cambridge, 2010, pp. 340–385.

[34] A. Nedich, A. Olshevsky, and W. Shi, *Achieving Geometric Convergence for Distributed Optimization Over Time-Varying Graphs*, preprint, https://arxiv.org/abs/1607.03218, 2016.

[35] R. Olfati-Saber, J. A. Fax, and R. M. Murray, *Consensus and cooperation in networked multi-agent systems*, Proc. IEEE, 95 (2007), pp. 215–233.

[36] H. Robbins and D. Siegmund, *A convergence theorem for non negative almost supermartingales and some applications*, Optimizing Methods in Statistics, Academic Press, New York, 1971, pp. 233–257.

[37] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1997.

[38] H. Seifi and M. S. Sepasian, *Electric Power System Planning: Issues, Algorithms and Solutions*, Springer, New York, 2011.

[39] P. Tseng, *On Accelerated Proximal Gradient Methods for Convex-Concave Optimization*, http://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf, 2008.

[40] H. Uzawa, *Iterative methods in concave programming*, in Studies in Linear and Nonlinear Programming, Stanford University Press, Palo Alto, CA, 1958, pp. 154–165.

[41] Y. Xu, T. Han, K. Cai, Z. Lin, G. Yan, and M. Fu, *A distributed algorithm for resource allocation over dynamic digraphs.*, IEEE Trans. Signal Process., 65 (2017), pp. 2600–2612.

[42] T. Yang, J. Lu, D. Wu, J. Wu, G. Shi, Z. Meng, and K. H. Johansson, *A distributed algorithm for economic dispatch over time-varying directed networks with delays*, IEEE Trans. Ind. Electron., 64 (2017), pp. 5095–5106.

[43] Y. Zhang and G. B. Giannakis, *Efficient decentralized economic dispatch for microgrids with wind power integration*, in Proceedings of the Sixth Annual Green Technologies Conference (GreenTech), IEEE, Piscataway, NJ, 2014, pp. 7–12.