Fifty-Fifth Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
October 3-6, 2017

Multi-agent constrained optimization of a strongly convex function over time-varying directed networks

Erfan Yazdandoost Hamedani[1] and Necdet Serhat Aybat[1]

***Abstract*— We consider cooperative multi-agent consensus optimization problems over undirected and directed time-varying communication networks, where only local communications are allowed. The objective is to minimize the sum of agent-specific possibly non-smooth composite convex functions over agent-specific private conic constraint sets; hence, the optimal consensus decision should lie in the intersection of these private sets. Assuming the sum function is strongly convex, we provide convergence rates in sub-optimality, infeasibility and consensus violation; examine the effect of underlying network topology on the convergence rates of the proposed decentralized algorithm.**

I. INTRODUCTION

Decentralized optimization over communication networks is an essential tool for solving various engineering problems: i) distributed parameter estimation in wireless sensor networks [1], [2]; ii) multi-agent cooperative control and coordination in multirobot networks [3], [4]; iii) distributed spectrum sensing in cognitive radio networks [5], [6]; iv) processing distributed big-data in (online) machine learning [7], [8], [9], [10], [11]; v) power control problem in cellular networks [12], 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 agents [13], [14]. Furthermore, in the aforementioned applications, the communication network can be time-varying, i.e., communication links can be on/off over time due to failures or the links may exist among agents depending on their inter-distances, and agents may need to communicate through a directed network, i.e., communication links can be unidirectional.

In this paper, from a broader perspective, we aim to study constrained distributed optimization of a strongly convex function over time-varying (un)directed communication networks $\mathcal{G}^t = (\mathcal{N}, \mathcal{E}^t)$ for $t \geq 0$; in particular, from an application perspective, we are motivated to design an efficient *decentralized* solution method for *constrained LASSO* (C-LASSO) problems [15] with distributed data. C-LASSO, with the generic form

$$\min_x \{\lambda \|x\|_1 + \|Cx - d\|_2^2 : \ Ax \leq b\},$$

is an important class of problems in statistics, which includes fused LASSO, constrained regression, and generalized LASSO problems as its special cases [16], [15], [17] to name a few.

We provide our theoretical results for a more general setting that subsumes C-LASSO as a special case. In the rest, we assume that **i)** each node $i \in \mathcal{N}$ has a *local* conic convex constraint set χ_i, for which projections are *not* easy to compute, and a *local* convex objective function φ_i (possibly non-smooth) such that $\sum_{i\in\mathcal{N}} \varphi_i(x)$ is strongly convex, and **ii)** nodes are willing to collaborate, without sharing their *private* data defining χ_i and φ_i, to compute an optimal consensus decision minimizing the sum of local functions and satisfying all local constraints; moreover, **iii)** nodes are only allowed to communicate with the neighboring nodes over the links in the network.

Although we assume that $\sum_{i\in\mathcal{N}} \varphi_i(x)$ is strongly convex, it is possible that none of the local functions $\{\varphi_i\}_{i\in\mathcal{N}}$ are strongly convex. This kind of structure arises in LASSO problems; in particular, let $\varphi_i : \mathbb{R}^n \to \mathbb{R}$ such that $\varphi_i(x) = \lambda \|x\|_1 + \|C_i x - d_i\|_2^2$ for $C_i \in \mathbb{R}^{m_i \times n}$ and $d_i \in \mathbb{R}^{m_i}$ for $i \in \mathcal{N}$. Note that while φ_i is merely convex for all $i \in \mathcal{N}$, $\sum_{i\in\mathcal{N}} \varphi_i(x)$ is strongly convex when $m_i < n$ for $i \in \mathcal{N}$ and $\mathbf{rank}(C) = n \leq \sum_{i\in\mathcal{N}} m_i \triangleq m$ where $C = [C_i]_{i\in\mathcal{N}} \in \mathbb{R}^{m\times n}$. Therefore, it is important to note that in the centralized formulation of this problem, the objective $\min_x \sum_{i\in\mathcal{N}} \varphi_i(x)$ is strongly convex; however, in the decentralized formulation, this is not the case where we minimize $\sum_{i\in\mathcal{N}} \varphi_i(x_i)$ while imposing consensus among local variables $\{x_i\}_{i\in\mathcal{N}}$. In the numerical section, we considered a distributed C-LASSO problem under a similar strong convexity setting.

Many of the real-life application problems discussed above fit into the conic constrained decentralized optimization framework discussed in this paper. With this motivation, we propose a *distributed* primal-dual algorithm, DPDA-TV, for time-varying communication networks. DPDA-TV is based on a primal-dual algorithm (PDA) recently proposed in [18] for convex-concave saddle-point problems of the form: $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \mathcal{L}(\mathbf{x},\mathbf{y}) \triangleq \Phi(\mathbf{x}) + \langle T\mathbf{x}, \mathbf{y}\rangle - h(\mathbf{y})$, where $\mathcal{X}, \mathcal{Y}$ are vector spaces, $\Phi(\mathbf{x}) \triangleq \rho(\mathbf{x}) + f(\mathbf{x})$ is a *strongly convex* function such that ρ and h are possibly non-smooth convex functions, f is convex and has a Lipschitz continuous gradient defined on $\mathbf{dom}\,\rho$ with constant L, and T is a linear map. In [18], it is shown that for any $(\mathbf{x},\mathbf{y}) \in \mathcal{X}\times\mathcal{Y}$, the *ergodic* average sequence $\{\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k\}_{k\geq 0}$ generated by PDA satisfies $\mathcal{L}(\bar{\mathbf{x}}^k, \mathbf{y}) - \mathcal{L}(\mathbf{x}, \bar{\mathbf{y}}^k) \leq \mathcal{O}(1/k^2)$ for appropriately chosen primal-dual step-size sequences.

E. Yazdandoost Hamedani[1] (evy5047@psu.edu) and N. S. Aybat[1] (nsa10@psu.edu) are with Industrial & Manufacturing Engineering Department, The Pennsylvania State University, University Park, PA, USA. Research of N. S. Aybat was partially supported by NSF grant CMMI-1635106 and ARO grant W911NF-17-1-0298

978-1-5386-3266-6/17/$31.00 ©2017 IEEE

Although, PDA is *not* a distributed algorithm for decentralized consensus optimization, in this paper, we show how to design one based on PDA for solving constrained consensus optimization over (un)directed time-varying networks with $\mathcal{O}(1/k^2)$ rate guarantee – even when all φ_i's are not strongly convex.

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

$$\varphi_i(x) \triangleq \rho_i(x) + f_i(x), \tag{1}$$

where $\rho_i : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ is a possibly *non-smooth* convex function, and $f_i : \mathbb{R}^n \to \mathbb{R}$ is a *smooth* convex function. We assume f_i is differentiable on an open set containing $\mathbf{dom}\,\rho_i$ with a Lipschitz continuous gradient ∇f_i, of which Lipschitz constant is L_i; and the prox map of ρ_i,

$$\mathbf{prox}_{\rho_i}(x) \triangleq \underset{y\in\mathbb{R}^n}{\operatorname{argmin}} \left\{\rho_i(y) + \tfrac{1}{2}\|y - x\|^2\right\}, \tag{2}$$

is *efficiently* computable for $i \in \mathcal{N}$, where $\|.\|$ denotes the Euclidean norm. Consider the following problem:

$$x^* \in \underset{x\in\mathbb{R}^n}{\operatorname{argmin}}\ \bar{\varphi}(x) \triangleq \sum_{i\in\mathcal{N}} \varphi_i(x) \tag{3}$$
$$\text{s.t. } A_i x - b_i \in \mathcal{K}_i, \quad \forall i \in \mathcal{N},$$

where $A_i \in \mathbb{R}^{m_i\times n}$, $b_i \in \mathbb{R}^{m_i}$ and $\mathcal{K}_i \subseteq \mathbb{R}^{m_i}$ is a closed, convex cone. Suppose that projections onto $\mathcal{K}_i$ can be computed efficiently, while the projection onto the *preimage* $\chi_i \triangleq A_i^{-1}(\mathcal{K}_i + b_i)$ is assumed to be *impractical*, e.g., when $\mathcal{K}_i$ is the positive semidefinite cone, projection to the preimage χ_i requires solving an SDP.

Assumption 1.1: The duality gap for (3) is zero, and a primal-dual solution to (3) exists.
A sufficient condition is the existence of a Slater point, i.e., there exists $\bar{x} \in \mathbf{relint}(\mathbf{dom}\,\bar{\varphi})$ such that $A_i\bar{x} - b_i \in \mathbf{int}(\mathcal{K}_i)$ for $i \in \mathcal{N}$, where $\mathbf{dom}\,\bar{\varphi} = \cap_{i\in\mathcal{N}}\,\mathbf{dom}\,\varphi_i$.

Definition *1:* A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is *strongly convex* with modulus $\mu > 0$ if the following inequality holds

$$f(x) \geq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle + \frac{\mu}{2}\|x - \bar{x}\|^2 \quad \forall x, \bar{x} \in \mathbb{R}^n.$$

Assumption 1.2: Suppose $\bar{f}(x) \triangleq \sum_{i\in\mathcal{N}} f_i(x)$ is strongly convex with modulus $\bar{\mu} > 0$; and each f_i is strongly convex with modulus $\mu_i \geq 0$ for $i \in \mathcal{N}$.

Remark *1.1:* Define $\underline{\mu} \triangleq \min_{i\in\mathcal{N}}\{\mu_i\} \geq 0$. Clearly $\bar{\mu} \geq \sum_{i\in\mathcal{N}} \mu_i$ is always true, and it is possible that $\mu_i = 0$ for all $i \in \mathcal{N}$ but still $\bar{\mu} > 0$; moreover, $\bar{\mu} > 0$ implies that x^* is the unique optimal solution to (3).

Previous Work. Here we briefly review some recent work on distributed consensus optimization for solving $\min_{x\in\mathbb{R}^n}\{\bar{\varphi}(x) : x \in \cap_{i\in\mathcal{N}}\chi_i\}$ over a network of computing agents $\mathcal{N}$, where $\bar{\varphi}(x) = \sum_{i\in\mathcal{N}} \varphi_i(x)$. Although the *unconstrained* consensus optimization, i.e., $\chi_i = \mathbb{R}^n$, is well studied for static or time-varying networks – see [19], [20] and the references therein, the *constrained* case is still an area of active research, e.g., [19], [20], [21], [22], [23], [24], [25], [26], [27], [28]. Our focus in this paper is on the case where $\bar{\varphi}$ is strongly convex such that each $\varphi_i = \rho_i + f_i$ is composite convex, and χ_i has the form $A_i^{-1}(\mathcal{K}_i + b_i)$ for $i \in \mathcal{N}$.

There are many papers investigating unconstrained minimization of a *strongly* convex, smooth objective function $\bar{f}(x) \triangleq \sum_{i\in\mathcal{N}} f_i(x)$ in the multi-agent setting, e.g., [29], [30], [31], [32], [33], [34] considered static communication networks $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ while [35], [36] studied the time-varying networks. In [31], an exact first-order algorithm (EXTRA) is proposed to to minimize $\bar{f}$ over an undirected static network; later in [32], combining EXTRA with the push-sum protocol, its variant EXTRA-push is proposed that can handle communication over *directed* networks that are strongly connected and static – both methods are gradient based with constant step-size. When $\bar{f}$ is smooth and strongly convex with modulus $\bar{\mu} > 0$ while each f_i need not to be strongly convex, it is shown that EXTRA has linear convergence, provided that the step-size $\alpha > 0$, constant among all the nodes, is sufficiently small, i.e., $\alpha = \mathcal{O}(\bar{\mu}/L_{\max}^2)$. Convergence of EXTRA-push, without providing any rate, has been shown under boundedness assumption on the iterate sequence. Similar to EXTRA-push, DEXTRA proposed in [33] also employs the push-sum protocol to minimize strongly convex $\bar{f}$ over static *directed* networks. Assuming that ∇f_i is bounded over $\mathbb{R}^n$ for $i \in \mathcal{N}$, which implies boundedness of the iterate sequence, it is shown in [33] that the iterate sequence converges linearly when the constant step-size α, fixed for all $i \in \mathcal{N}$, is chosen carefully belonging to a *non-trivial* interval. In a follow up paper [34], Xi and Khan proposed Accelerated Distributed Directed Optimization (ADD-OPT) where they improved on the nontrivial step-size condition of DEXTRA and showed that the iterates converge linearly when the constant step-size α is chosen sufficiently small – assuming that the directed network topology is static and each f_i is strongly convex with Lipschitz continuous gradients. Nedić and Olshevsky [35] proposed a stochastic (sub)gradient-push for a more general setting of minimizing (possibly) nonsmooth strongly convex $\bar{f}$ over a *time-varying directed* network when the stochastic error in subgradient samples has zero mean and bounded standard deviation. When $\mu_i > 0$ for all $i \in \mathcal{N}$, choosing a diminishing step-size sequence, they were able to show $\mathcal{O}(\log(k)/k)$ rate result provided that the iterate sequence stays bounded – the boundedness assumption on the iterate sequence can be removed by assuming that functions are smooth, having Lipschitz continuous gradients. In [36], Nedić et al. proposed distributed inexact gradient methods: DIGing and Push-DIGing for *time-varying*, undirected and directed networks, respectively. The iterate sequence is shown to converge linearly provided that the constant step-size α, fixed for all $i \in \mathcal{N}$, is chosen sufficiently small when each f_i is strongly convex with Lipschitz continuous gradient.

For constrained consensus optimization, other than few exceptions, e.g., [23], [24], [25], [26], [27], [28], the existing methods require that each node compute a projection on the local set χ_i in addition to consensus and (sub)gradient steps, e.g., [21], [22]. Moreover, among those few exceptions, only [25], [26], [27], [28] can handle agent-specific constraints without assuming global knowledge of the constraints by all agents. However, *no* rate results in terms of suboptimality, local infeasibility, and consensus violation exist for the primal-dual distributed methods in [25], [26], [27] when implemented for the agent-specific *conic* constraint sets $\chi_i = \{x : A_i x - b_i \in \mathcal{K}_i\}$ studied in this paper. In [25], the authors considered the problem of minimizing 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 inequality constraints on the summation of agent specific constrained functions, i.e., $\sum_{i\in\mathcal{N}} g_i(x) \leq 0$, over a *time-varying directed* network. The authors proposed a consensus-based distributed primal-dual perturbation (PDP) algorithm using a diminishing step-size sequence, and showed that the primal-dual iterate sequence converges to a global optimal primal-dual solution, without providing a rate result. The proposed PDP method can also handle non-smooth constraints with similar convergence guarantees. In a recent work [26], a distributed algorithm on time-varying directed networks for solving saddle-point problems subject to consensus constraints is proposed. The algorithm can also solve consensus optimization problems with *inequality* constraints that can be written as summation of local convex functions of local and global variables. It is shown that using a carefully selected decreasing step-size sequence, the ergodic average of primal-dual sequence converges with $\mathcal{O}(1/\sqrt{k})$ rate in terms of saddle-point evaluation error; however, when applied to constrained optimization problems, *no* rate in terms of either suboptimality or infeasibility is provided. A closely related paper to ours is [27], where a proximal dual consensus ADMM method, PDC-ADMM, is proposed by Chang to solve $\min_x\{\sum_i \varphi_i(x_i) : \sum_{i\in\mathcal{N}} C_i x_i = d,\ x_i \in \chi_i\ i \in \mathcal{N}\}$ over both static and time-varying undirected networks, requiring only proximal-gradient steps, where $\varphi_i = \rho_i + f_i$ is composite convex, $\chi_i = \{x_i : A_i x_i \geq b_i, x_i \in \mathcal{S}_i\}$ and $\mathcal{S}_i$ is a convex compact set. It is shown that both for static and time-varying cases, PCD-ADMM have $\mathcal{O}(1/k)$ ergodic convergence rate in the mean without requiring projection onto χ_i for suboptimality and infeasibility when each f_i is strongly convex and differentiable with a Lipschitz continuous gradient for $i \in \mathcal{N}$. More recently, in [28], Aybat and Yazdandoost Hamedani proposed a distributed primal-dual method to solve (3) over *time-varying* networks when $\varphi_i = \rho_i + f_i$ is composite convex. Assuming f_i is smooth, convergence of the primal-dual iterate sequence is shown, and $\mathcal{O}(1/k)$ ergodic rate is given for suboptimality and infeasibility. In this paper, we aim to improve on this rate by further assuming $\sum_i \varphi_i$ is strongly convex to achieve $\mathcal{O}(1/k^2)$ ergodic rate.

Although our focus is on the convex setting, it is worth emphasizing that distributed constrained non-convex consensus optimization is another area of active research, e.g., [37], [38], [39]. In these papers, the objective is to minimize the sum of agent specific smooth non-convex functions subject to a globally known closed convex set over a time-varying communication network. Under certain assumptions, it is shown that agents' iterates converge to a stationary point.

Contribution. To the best of our knowledge, only a handful of methods, e.g., [25], [26], [27], [28] can handle consensus problems, similar to (3), with agent-specific *local* constraint sets $\{\chi_i\}_{i\in\mathcal{N}}$ without requiring each agent $i \in \mathcal{N}$ to project onto χ_i. However, no rate results in terms of suboptimality, local infeasibility, and consensus violation exist for the distributed methods in [25], [26], [27] when implemented for *conic* sets $\{\chi_i\}_{i\in\mathcal{N}}$ studied in this paper; moreover, none of these four methods exploits the strong convexity of the sum function $\bar{\varphi} = \sum_{i\in\mathcal{N}} \varphi_i$. We believe DPDA-TV proposed in this paper is one of the first decentralized algorithms to solve (3) with $\mathcal{O}(1/k^2)$ ergodic rate on both sub-optimality and infeasibility. More precisely, we show that when $\bar{\varphi}$ is strongly convex and each φ_i is composite convex with smooth f_i for $i \in \mathcal{N}$, our proposed method reduces the suboptimality and infeasibility with $\mathcal{O}(1/k^2)$ rate as k, the number primal-dual iterations, increases, and it requires $\mathcal{O}(k\log(k))$ local communications for all k iterations in total. To the best of our knowledge, this is the best rate result for our setting. It is worth noting that, our results imply that DPDA-TV can compute a point in the intersection of closed convex sets with $\mathcal{O}(1/k)$ rate for the solution error $\sum_{i\in\mathcal{N}} \left\|x_i^k - x^*\right\|$ – in a decentralized way over time-varying directed communication networks, which is faster than $\mathcal{O}(1/\sqrt{k})$ rate of Dykstra's algorithm – see [40].

Notation. Throughout $\|.\|$ denotes either the Euclidean norm or the spectral norm, and $\langle\theta, w\rangle \triangleq \theta^\top w$ for $\theta, w \in \mathbb{R}^n$. Given a convex set $\mathcal{S}$, let $\sigma_{\mathcal{S}}(.)$ denote its support function, i.e., $\sigma_{\mathcal{S}}(\theta) \triangleq \sup_{w\in\mathcal{S}} \langle\theta,\ w\rangle$, and let $\mathcal{P}_{\mathcal{S}}(w) \triangleq \operatorname{argmin}\{\|v - w\| :\ v \in \mathcal{S}\}$ denote the 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 a given cone $\mathcal{K} \in \mathbb{R}^m$, $\sigma_{\mathcal{K}}(\theta) = 0$ for $\theta \in \mathcal{K}^\circ$ and equal to $+\infty$ if $\theta \notin \mathcal{K}^\circ$. Given a convex function $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, its convex conjugate is defined as $g^*(w) \triangleq \sup_{\theta\in\mathbb{R}^n} \langle w, \theta\rangle - g(\theta)$. $\otimes$ denotes the Kronecker product, $\mathbf{1}_n \in \mathbb{R}^n$ is the vector all ones, and $\mathbf{I}_n$ is the $n \times n$ identity matrix. $\mathbb{S}^n_{++}$ ($\mathbb{S}^n_+$) denotes the cone of symmetric positive (semi)definite matrices. For $Q \succ 0$, i.e., $Q \in \mathbb{S}^n_{++}$, Q-norm is defined as $\|z\|_Q \triangleq \sqrt{z^\top Q z}$. Finally, Π denotes the Cartesian product.

II. METHODOLOGY

In this section we develop a distributed primal-dual algorithm for solving (3) when the communication network topology is *time-varying*. We will adopt the following definition and assumption for the *time-varying* network model.

Definition *2:* Given $t \geq 0$, for an undirected graph $\mathcal{G}^t = (\mathcal{N}, \mathcal{E}^t)$, let $\mathcal{N}_i^t \triangleq \{j \in \mathcal{N} :\ (i,j) \in \mathcal{E}^t\ \text{ or }\ (j,i) \in \mathcal{E}^t\}$

denote the set of neighboring nodes of $i \in \mathcal{N}$, and $d_i^t \triangleq |\mathcal{N}_i^t|$ represent the degree of node $i \in \mathcal{N}$ at time t; for a directed graph $\mathcal{G}^t = (\mathcal{N}, \mathcal{E}^t)$, 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 at time t, respectively; and $d_i^t \triangleq |\mathcal{N}_i^{t,\text{out}}|$ be the out-degree of node i.

Assumption 2.1: Suppose that $\{\mathcal{G}^t\}_{t\in\mathbb{R}_+}$ is a collection of either all directed or all undirected graphs. When $\mathcal{G}^t$ is an undirected graph, 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., $(i,j) \in \mathcal{E}^t$ or $(j,i) \in \mathcal{E}^t$; on the other hand, when $\mathcal{G}^t$ is a 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., $(i,j) \in \mathcal{E}^t$.

We assume a *compact* domain, i.e., let $\Delta_i \triangleq \max_{x_i,x_i'\in\mathbf{dom}\,\varphi_i} \|x - x'\|$ and $\Delta \triangleq \max_{i\in\mathcal{N}} \Delta_i < \infty$. Let $\mathcal{B}_0 \triangleq \{x \in \mathbb{R}^n : \ \|x\| \leq 2\Delta\}$ and $\mathcal{B} \triangleq \Pi_{i\in\mathcal{N}}\mathcal{B}_0 \subset \mathcal{X} \triangleq \Pi_{i\in\mathcal{N}}\mathbb{R}^n$; and let $\widetilde{\mathcal{C}} \triangleq \mathcal{C} \cap \mathcal{B}$ be a set of bounded consensus decisions, where $\mathcal{C}$ is the consensus cone defined as:

$$\mathcal{C} \triangleq \{\mathbf{x} \in \mathcal{X} : \ \exists \bar{x} \in \mathbb{R}^n \ \mathbf{s.t.} \ x_i = \bar{x} \quad \forall \ i \in \mathcal{N}\}. \tag{4}$$

Definition *3:* Suppose $\mathcal{X} \triangleq \Pi_{i\in\mathcal{N}}\mathbb{R}^n$ and $\mathcal{X} \ni \mathbf{x} \triangleq [x_i]_{i\in\mathcal{N}}$; let $\varphi : \mathcal{X} \to \mathbb{R}\cup\{\infty\}$ such that $\varphi(\mathbf{x}) = \rho(\mathbf{x}) + f(\mathbf{x})$ where $\rho(\mathbf{x}) \triangleq \sum_{i\in\mathcal{N}} \rho_i(x_i)$ and $f(\mathbf{x}) \triangleq \sum_{i\in\mathcal{N}} f_i(x_i)$. Given $\alpha \geq 0$, define $\varphi_\alpha(\mathbf{x}) = \rho(\mathbf{x}) + f_\alpha(\mathbf{x})$ where $f_\alpha(\mathbf{x}) \triangleq f(\mathbf{x}) + \frac{\alpha}{2} d_{\mathcal{C}}^2(\mathbf{x})$.

Recall that when $\underline{\mu} > 0$, according to Assumption 1.2, $f(\mathbf{x}) = \sum_{i\in\mathcal{N}} f_i(x_i)$ is strongly convex on $\mathcal{X}$ with modulus $\underline{\mu}$. On the other hand, as emphasized in the introduction, although $\bar{f}(x) = \sum_{i\in\mathcal{N}} f_i(x)$ is strongly convex with modulus $\bar{\mu} > 0$, it is possible that f may be merely convex with $\underline{\mu} = 0$, which implies that f is strongly convex only on $\mathcal{C}$; in the next lemma, we show that by suitably regularizing f, one can obtain a strongly convex function on $\mathcal{X}$ even when $\underline{\mu} = 0$.

Lemma 2.1: Consider $f(\mathbf{x}) = \sum_{i\in\mathcal{N}} f_i(x_i)$ under Assumption 1.2 and suppose $\underline{\mu} = 0$. Given $\alpha > 0$, let $f_\alpha(\mathbf{x}) \triangleq f(\mathbf{x}) + \frac{\alpha}{2} d_{\mathcal{C}}^2(\mathbf{x})$. Then f_α is strongly convex with modulus $\mu_\alpha \triangleq \frac{\bar{\mu}/|\mathcal{N}| + \alpha}{2} - \sqrt{\left(\frac{\bar{\mu}/|\mathcal{N}| - \alpha}{2}\right)^2 + 4\bar{L}^2} > 0$ for any $\alpha > \frac{4}{\bar{\mu}}|\mathcal{N}|\bar{L}^2$, where $\bar{L} = \sqrt{\frac{\sum_{i\in\mathcal{N}} L_i^2}{|\mathcal{N}|}}$.

Proof: For proof, see the online technical report [41]. ■

Remark *2.1:* Using Lemma 2.1, one can design accelerated algorithms for time-varying network topologies – hence, in a way, it generalizes Proposition 3.6. of [31] where a similar result is obtained using a regularizer defined by some mixing matrices dependent on the static topology of the network.

Let x^* be the unique solution to (3), $\mathbf{x}^* \triangleq \mathbf{1} \otimes x^*$ satisfies $d_{\mathcal{C}}(\mathbf{x}^*) = 0$; hence, $d_{\mathcal{C}}(\mathbf{x}) \geq 0$ for $\mathbf{x} \in \mathcal{X}$ implies that

$$\min_{\mathbf{x}\in\widetilde{\mathcal{C}}} \{\varphi_\alpha(\mathbf{x}) : \ A_i x_i - b_i \in \mathcal{K}_i, \ i \in \mathcal{N}\} \tag{5}$$

is equivalent to (3) for any $\alpha \geq 0$. Next, consider the following reformulation of (5) as a saddle point problem:

$$\min_{\mathbf{x}} \max_{\boldsymbol{\lambda},\boldsymbol{\theta}} \ \mathcal{L}(\mathbf{x},\mathbf{y}) \triangleq \varphi_\alpha(\mathbf{x}) + \langle \boldsymbol{\lambda}, \ \mathbf{x}\rangle - \sigma_{\widetilde{\mathcal{C}}}(\boldsymbol{\lambda}) + \sum_{i\in\mathcal{N}} \langle \theta_i, A_i x_i - b_i\rangle - \sigma_{\mathcal{K}_i}(\theta_i), \tag{6}$$

where $\boldsymbol{\theta} = [\theta_i]_{i\in\mathcal{N}}$ and $\boldsymbol{\lambda} \in \mathbb{R}^{n|\mathcal{N}|}$. Therefore, for any given $\alpha \geq 0$, one can compute a primal-dual optimal solution to (3) through computing a saddle-point to (6). In the rest, we first consider a naive implementation of a PDA in [18] to solve (6), which does not result in a decentralized method; thus, we subsequently discuss how to fix it to design a distributed algorithm that works over time-varying communication networks.

Let $\mathcal{Y} \triangleq \Pi_{i\in\mathcal{N}}\mathbb{R}^{m_i} \times \mathbb{R}^{m_0}$, $\mathcal{Y} \ni \mathbf{y} = [\boldsymbol{\theta}^\top \boldsymbol{\lambda}^\top]^\top$ such that $\boldsymbol{\theta} = [\theta_i]_{i\in\mathcal{N}} \in \mathbb{R}^m$ and $\boldsymbol{\lambda} \in \mathbb{R}^{m_0}$, where $m \triangleq \sum_{i\in\mathcal{N}} m_i$ and $m_0 \triangleq n|\mathcal{N}|$. Let $h : \mathcal{Y} \to \mathbb{R}\cup\{\infty\}$ such that $h(\mathbf{y}) \triangleq \sigma_{\widetilde{\mathcal{C}}}(\boldsymbol{\lambda}) + \sum_{i\in\mathcal{N}} \sigma_{\mathcal{K}_i}(\theta_i) + \langle b_i, \theta_i\rangle$. Define the block-diagonal matrix $A \triangleq \mathbf{diag}([A_i]_{i\in\mathcal{N}}) \in \mathbb{R}^{m\times m_0}$ and $T = [A^\top \ \mathbf{I}_{m_0}]^\top$. Given parameters $\gamma^k > 0$, $\kappa_i^k > 0$ for $i \in \mathcal{N}$, let $\mathbf{D}_{\gamma^k} \triangleq \frac{1}{\gamma^k}\mathbf{I}_{m_0}$, $\mathbf{D}_{\boldsymbol{\kappa}^k} \triangleq \mathbf{diag}([\frac{1}{\kappa_i^k}\mathbf{I}_{m_i}]_{i\in\mathcal{N}})$, and $\mathbf{D}_{\boldsymbol{\kappa}^k,\gamma^k} \triangleq \begin{bmatrix} \mathbf{D}_{\boldsymbol{\kappa}^k} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{\gamma^k} \end{bmatrix}$. Finally, define the Bregman function $D_k(\mathbf{y},\bar{\mathbf{y}}) = \frac{1}{2}\|\mathbf{y} - \bar{\mathbf{y}}\|^2_{\mathbf{D}_{\boldsymbol{\kappa}^k,\gamma^k}}$ for each $k \geq 0$. Hence, with φ_α, h and T, defined as above, given the initial iterates $\mathbf{x}^0$ and $\mathbf{y}^0 = [\boldsymbol{\theta}^{0^\top} \boldsymbol{\lambda}^{0^\top}]^\top$, the PDA iterations of Algorithm 4 in [18], applied to $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \varphi_\alpha(\mathbf{x}) + \langle T\mathbf{x}, \mathbf{y}\rangle - h(\mathbf{y})$, take the following form for $k \geq 0$:

$$\theta_i^{k+1} \leftarrow \operatorname*{argmin}_{\theta_i} \sigma_{\mathcal{K}_i}(\theta_i) - \langle A_i(x_i^k + \eta^k(x_i^k - x_i^{k-1})) - b_i, \ \theta_i\rangle + \frac{1}{2\kappa_i^k}\|\theta_i - \theta_i^k\|^2, \quad i \in \mathcal{N}, \tag{7a}$$

$$\boldsymbol{\lambda}^{k+1} \leftarrow \operatorname*{argmin}_{\boldsymbol{\lambda}} \sigma_{\widetilde{\mathcal{C}}}(\boldsymbol{\lambda}) - \langle \mathbf{x}^k + \eta^k(\mathbf{x}^k - \mathbf{x}^{k-1}), \ \boldsymbol{\lambda}\rangle + \frac{1}{2\gamma^k}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^k\|^2, \tag{7b}$$

$$\mathbf{x}^{k+1} \leftarrow \operatorname*{argmin}_{\mathbf{x}} \rho(\mathbf{x}) + \langle \nabla f_\alpha(\mathbf{x}^k), \ \mathbf{x}\rangle + \langle A\mathbf{x} - b, \ \boldsymbol{\theta}^{k+1}\rangle + \langle \mathbf{x}, \ \boldsymbol{\lambda}^{k+1}\rangle + \frac{1}{2\tau^k}\|\mathbf{x} - \mathbf{x}^k\|^2, \tag{7c}$$

where $\mathbf{x}^{-1} = \mathbf{x}^0$ – a possible choice for $\alpha \geq 0$ and the positive parameter sequences $\{\gamma^k\}_k$, $\{\tau^k\}_k$, $\{\kappa_i^k\}_k$ for $i \in \mathcal{N}$ is given in Remark 2.2 and Figure 1.

For $k \geq 0$, using extended Moreau decomposition for proximal operators, $\boldsymbol{\lambda}^{k+1}$ in (7b) can be computed as $\boldsymbol{\lambda}^{k+1} = \mathbf{prox}_{\gamma^k\sigma_{\widetilde{\mathcal{C}}}}(\boldsymbol{\lambda}^k + \gamma^k(\mathbf{x}^k + \eta^k(\mathbf{x}^k - \mathbf{x}^{k-1})) = \gamma^k\left(\boldsymbol{\omega}^k - \mathcal{P}_{\widetilde{\mathcal{C}}}(\boldsymbol{\omega}^k)\right)$, where $\boldsymbol{\omega}^k \triangleq \frac{1}{\gamma^k}\boldsymbol{\lambda}^k + \mathbf{x}^k + \eta^k(\mathbf{x}^k - \mathbf{x}^{k-1})$ for $k \geq 0$. Moreover, ∇f_α for the $\mathbf{x}$-step in (7c) can be computed as

$$\nabla f_\alpha(\mathbf{x}^k) = \nabla f(\mathbf{x}^k) + \alpha\mathcal{P}_{\mathcal{C}^\circ}(\mathbf{x}^k) = \nabla f(\mathbf{x}^k) + \alpha\big(\mathbf{x}^k - \mathcal{P}_{\mathcal{C}}(\mathbf{x}^k)\big). \tag{8}$$

For any $\mathbf{x} = [x_i]_{i\in\mathcal{N}} \in \mathcal{X}$, $\mathcal{P}_{\widetilde{\mathcal{C}}}(\mathbf{x})$ and $\mathcal{P}_{\mathcal{C}}(\mathbf{x})$ can be computed as $\mathcal{P}_{\widetilde{\mathcal{C}}}(\mathbf{x}) = \mathcal{P}_{\mathcal{B}}(\mathcal{P}_{\mathcal{C}}(\mathbf{x}))$, and $\mathcal{P}_{\mathcal{C}}(\mathbf{x}) = \mathbf{1} \otimes p(\mathbf{x})$, where $p(\mathbf{x}) \triangleq \frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}} x_i$, $\mathcal{P}_{\mathcal{B}}(\mathbf{x}) = [\mathcal{P}_{\mathcal{B}_0}(x_i)]_{i\in\mathcal{N}}$ and $\mathcal{P}_{\mathcal{B}_0}(x_i) = x_i \min\{1, \frac{2\Delta}{\|x_i\|}\}$ for $i \in \mathcal{N}$.

Although $\boldsymbol{\theta}$-step of the PDA implementation in (7) can be computed locally at each node, computing $\mathbf{x}$-step and

$\boldsymbol{\lambda}$-step require communication among the nodes to evaluate $\mathcal{P}_{\widetilde{\mathcal{C}}}(\boldsymbol{\omega}^k)$ and $\mathcal{P}_{\mathcal{C}}(\mathbf{x}^k)$. Indeed, evaluating the average operator $p(.)$ is *not* a simple operation in a decentralized computational setting which only allows for communication among the neighbors. In order to overcome this issue, we will approximate the average operator $p(.)$ using multi-communication rounds, and analyze the resulting iterations as an *inexact* primal-dual algorithm. We define a *communication round* at time t 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 2.1 – 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 PDA iteration $k \geq 0$, there exists an approximate averaging operator $\mathcal{R}^k(\cdot)$ which can be computed in a decentralized fashion and approximate $\mathcal{P}_{\mathcal{C}}(\cdot)$ with decreasing *approximation error* as k, the number of PDA iterations, increases. This *inexact* version of PDA using approximate averaging operator $\mathcal{R}^k(\cdot)$ and running on time-varying communication network $\{\mathcal{G}^t\}_{t\in\mathbb{R}_+}$ will be called DPDA-TV, of which details is discussed next.

Assumption 2.2: 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 requiring to compute $\Pi_{\mathcal{K}_i}$ as in (7a), and $\mathbf{prox}_{\rho_i}$ and ∇f_i as in (7c) can be completed between two ticks of the clock for all $i \in \mathcal{N}$ and $k \geq 0$; and every time the clock ticks a communication round with instantaneous messaging between neighboring nodes takes place subject to Assumption 2.1. 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 $\beta \in (0,1)$ such that for all $k \geq 0$,

$$\|\mathcal{R}^k(\mathbf{w}) - \mathcal{P}_{\mathcal{C}}(\mathbf{w})\| \leq N\ \Gamma\beta^{q_k}\ \|\mathbf{w}\|, \quad \forall\ \mathbf{w} \in \mathbb{R}^{m_0}. \quad (9)$$

Now we briefly talk about such operators. 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 $W^{t,s} \triangleq V^t V^{t-1} ... V^{s+1}$ for any $t, s \in \mathbb{Z}_+$ such that $t \geq s+1$. For directed time-varying graph $\mathcal{G}^t$, set $V^t \in \mathbb{R}^{|\mathcal{N}|\times|\mathcal{N}|}$ as follows: for each $i \in \mathcal{N}$,

$$V_{ij}^t = \frac{1}{d_j^t} \ \text{ if } \ j \in \mathcal{N}_i^{t,\text{in}}; \ \ V_{ij}^t = 0 \ \text{ if } \ j \notin \mathcal{N}_i^{t,\text{in}}. \quad (10)$$

Let $t_k \in \mathbb{Z}_+$ be the total number of *communication rounds* done before the k-th iteration of DPDA-TV, and let $q_k \in \mathbb{Z}_+$ be the number of communication rounds to be performed within the k-th iteration while evaluating $\mathcal{R}^k$. For $\mathbf{x} \in \mathcal{X}$, define

$$\mathcal{R}^k(\mathbf{x}) \triangleq \mathbf{diag}(W^{t_k+q_k,t_k}\mathbf{1}_{|\mathcal{N}|})^{-1}(W^{t_k+q_k,t_k} \otimes \mathbf{I}_n)\ \mathbf{x} \quad (11)$$

to approximate $\mathcal{P}_{\mathcal{C}}(\cdot)$. Note that $\mathcal{R}^k(\cdot)$ can be computed in a *distributed fashion* requiring q_k communication rounds – $\mathcal{R}^k$ is nothing but the push-sum protocol [42]. Assuming that the digraph sequence $\{\mathcal{G}^t\}_{t\in\mathbb{Z}_+}$ is uniformly strongly connected (M-strongly connected), it follows from [42], [43] that $\mathcal{R}^k$ satisfies Assumption 2.2. When $\{\mathcal{G}^t\}_{t\in\mathbb{Z}_+}$ is undirected, then choosing $\{V^t\}_{t\in\mathbb{Z}_+}$ according to Metropolis weights, one can show that under certain conditions,

$$\mathcal{R}^k(\mathbf{x}) \triangleq (W^{t_k+q_k,t_k} \otimes \mathbf{I}_m)\mathbf{x} \quad (12)$$

satisfies Assumption 2.2, e.g., see [44].

Note that for $\widetilde{\mathcal{R}}^k(\cdot) \triangleq \mathcal{P}_{\mathcal{B}}(\mathcal{R}^k(\cdot))$, we have $\widetilde{\mathcal{R}}^k(\mathbf{w}) \in \mathcal{B}$, and $\|\widetilde{\mathcal{R}}^k(\mathbf{w}) - \mathcal{P}_{\widetilde{\mathcal{C}}}(\mathbf{w})\| \leq N\ \Gamma\beta^{q_k}\ \|\mathbf{w}\|$ for $\mathbf{w} \in \mathbb{R}^{m_0}$ due to non-expansivity of $\mathcal{P}_{\mathcal{B}}$. Consider the k-th iteration of PDA as shown in (7). Instead of computing $\boldsymbol{\lambda}^{k+1}$ and $\mathbf{x}^{k+1}$ as shown in (7b) and (7c), which require computing $\mathcal{P}_{\mathcal{C}}$, we propose replacing (7b) and (7c) with similar update rules that use the inexact averaging operator $\mathcal{R}^k$ to approximate $\mathcal{P}_{\mathcal{C}}$. Hence, we obtain an *inexact* variant of (7) replacing (7b) and (7c) with

$$\boldsymbol{\lambda}^{k+1} \leftarrow \gamma^k \left(\boldsymbol{\omega}^k - \mathcal{P}_{\mathcal{B}}\big(\mathcal{R}^k(\boldsymbol{\omega}^k)\big)\right), \quad (13a)$$

where $\boldsymbol{\omega}^k = \frac{1}{\gamma^k}\boldsymbol{\lambda}^k + \mathbf{x}^k + \eta^k(\mathbf{x}^k - \mathbf{x}^{k-1})$, and

$$\mathbf{x}^{k+1} \leftarrow \mathbf{prox}_{\tau^k\rho}\Big(\mathbf{x}^k - \tau^k\mathbf{s}^k\Big), \quad (13b)$$

where $\mathbf{s}^k = \nabla f(\mathbf{x}^k) + A^\top\boldsymbol{\theta}^{k+1} + \boldsymbol{\lambda}^{k+1} + \alpha\big(\mathbf{x}^k - \mathcal{R}^k(\mathbf{x}^k)\big)$.

Thus, PDA given in (7) can be computed inexactly, and in a *decentralized* way for any time-varying connectivity network $\{\mathcal{G}^t\}_{t\in\mathbb{Z}_+}$, via the node-specific computations as in the distributed primal-dual algorithm displayed in Fig. 1 below. Indeed, the iterate sequence $\{\mathbf{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\theta}^k\}_{k\geq 0}$ generated by DPDA-TV displayed in Fig. 1 is the same sequence generated by the recursion in (7a), (13a), and (13b).

Algorithm DPDA-TV ($\mathbf{x}^0, \boldsymbol{\theta}^0, \alpha, \mu, \delta_1, \delta_2, \{q_k\}$)

Initialization: $\mathbf{x}^{-1} \leftarrow \mathbf{x}^0$, $\boldsymbol{\lambda}^0 \leftarrow \mathbf{0}$, $\delta_1, \delta_2 > 0$,
$\tau^0 \leftarrow \frac{1}{L_{\max}+\delta_2+\alpha}$, $\tilde{\tau}^0 \leftarrow (\frac{1}{\tau^0} - \mu)^{-1}$, $\eta^0 \leftarrow 0$,
$\gamma^0 \leftarrow \frac{\delta_2}{1+\delta_1}$, $\kappa_i^0 \leftarrow \gamma^0\frac{\delta_1}{\|A_i\|^2}$ $i \in \mathcal{N}$

Step k: ($k \geq 0$)

1. $p_i^k \leftarrow x_i^k + \eta^k(x_i^k - x_i^{k-1}), \quad i \in \mathcal{N}$
2. $\theta_i^{k+1} \leftarrow \mathcal{P}_{\mathcal{K}_i^\circ}\Big(\theta_i^k + \kappa_i^k\big(A_i p_i^k - b_i\big)\Big), \quad i \in \mathcal{N}$
3. $\omega_i^k \leftarrow \frac{1}{\gamma^k}\nu_i^k + p_i^k, \quad i \in \mathcal{N}$
4. $\lambda_i^{k+1} \leftarrow \gamma^k\Big(\omega_i^k - \mathcal{P}_{\mathcal{B}_0}\Big(\mathcal{R}_i^k\big(\boldsymbol{\omega}^k\big)\Big)\Big), \quad i \in \mathcal{N}$
5. $s_i^k \leftarrow \nabla f_i(x_i^k) + A_i^\top\theta_i^{k+1} + \lambda_i^{k+1} + \alpha\big(x_i^k - \mathcal{R}_i^k(\mathbf{x}^k)\big), \ i \in \mathcal{N}$
6. $x_i^{k+1} \leftarrow \mathbf{prox}_{\tau^k\rho_i}\big(x_i^k - \tau^k s_i^k\big), \quad i \in \mathcal{N}$
7. $\eta^{k+1} \leftarrow \frac{1}{\sqrt{1+\mu\tilde{\tau}^k}}$, $\tilde{\tau}^{k+1} \leftarrow \eta^{k+1}\tilde{\tau}^k$, $\tau^{k+1} \leftarrow (\frac{1}{\tilde{\tau}^{k+1}} + \mu)^{-1}$
8. $\gamma^{k+1} \leftarrow \gamma^k/\eta^{k+1}$, $\kappa_i^{k+1} \leftarrow \gamma^{k+1}\frac{\delta_1}{\|A_i\|^2}$ $i \in \mathcal{N}$

Fig. 1: Distributed Primal Dual Algorithm for Time-Varying $\{\mathcal{G}^t\}_{t\geq 0}$ (DPDA-TV)

Remark *2.2:* If $\underline{\mu} > 0$, then we set $\alpha = 0$ and choose $\mu = \underline{\mu}$; otherwise, when $\underline{\mu} = 0$, it follows from Lemma 2.1 that for any $\alpha > \frac{4}{\underline{\mu}}|\mathcal{N}|\bar{L}^2$, f_α is strongly convex with modulus $\mu_\alpha > 0$; hence, we set $\mu = \mu_\alpha$ for some $\alpha > \frac{4}{\underline{\mu}}|\mathcal{N}|\bar{L}^2$.

Next, we quantify the suboptimality and infeasibility of the DPDA-TV iterate sequence.

Theorem 2.2: Suppose Assumptions 1.1, 1.2, 2.1 and 2.2 hold. Starting from $\boldsymbol{\lambda}^0 = \mathbf{0}$, $\boldsymbol{\theta}^0 = \mathbf{0}$, and an arbitrary $\mathbf{x}^0$, let $\{\mathbf{x}^k, \boldsymbol{\theta}^k, \boldsymbol{\lambda}^k\}_{k\geq 0}$ be the iterate sequence generated by Algorithm DPDA-TV, displayed in Fig. 1, using $q_k \geq (5+c)\log_{1/\beta}(k+1)$ communication rounds for the k-th iteration for $k \geq 0$. Then $\{\mathbf{x}^k\}_{k\geq 0}$ converges to $\mathbf{x}^* = \mathbf{1}\otimes x^*$ such that x^* is the optimal solution to (3). Moreover, the following bounds hold for all $K \geq 1$:

$$\max\left\{|\varphi(\bar{\mathbf{x}}^K) - \varphi(\mathbf{x}^*)|,\ d_{\mathcal{C}}(\bar{\mathbf{x}}^K) + \sum_{i\in\mathcal{N}} \|\theta_i^*\|\, d_{\mathcal{K}_i}(A_i\bar{\mathbf{x}}_i^K - b_i)\right\}$$
$$\leq \frac{\Lambda(K)}{N_K} = \mathcal{O}\Big(\frac{1}{K^2}\Big),$$
$$\left\|\mathbf{x}^K - \mathbf{x}^*\right\|^2 \leq \frac{\tilde{\tau}^K}{\gamma^K}\, 2\gamma^0\, \Lambda(K) = \mathcal{O}\Big(\frac{1}{K^2}\Big),$$

and the parameters satisfy $\max\{N_K,\ \gamma^K/\tilde{\tau}^K\} = \mathcal{O}(K^2)$, where $N_K = \sum_{k=1}^K \gamma^{k-1}/\gamma^0$, $\bar{\mathbf{x}}^K = N_K^{-1}\sum_{k=1}^K \frac{\gamma^{k-1}}{\gamma^0}\mathbf{x}^k$, and $\Lambda(K) = \mathcal{O}\big(\sum_{k=1}^K \beta^{q_{k-1}}k^4\big)$; hence, $\sup_{K\in\mathbb{Z}_+}\Lambda(K) < \infty$.

Proof: For proof see the online technical report [41]. ■

Remark *2.3:* Note that, at the K-th iteration, the suboptimality, infeasibility and consensus violation are $\mathcal{O}\left(\frac{1}{N_K}\,\Lambda(K)\right)$ in the ergodic sense, and the distance of iterates to $\mathbf{x}^*$ is $\mathcal{O}\left(\frac{\tilde{\tau}^K}{\gamma^K}\,\Lambda(K)\right)$ where $\Lambda(K)$ denotes the error accumulations due to average approximation. Moreover, $\Lambda(K)$ can be bounded above for all $K \geq 1$ as $\Lambda(K) \leq C_1\sum_{k=1}^K \beta^{q_{k-1}}k^4$ for some $C_1 > 0$; therefore, for any $c > 0$, choosing $\{q_k\}_{k\in\mathbb{Z}_+}$ as stated in Theorem 2.2 ensures that $\sum_{k=1}^\infty \beta^{q_{k-1}}k^4 < 1 + \frac{1}{c}$. Moreover, for any $c > 0$, setting $q_k = (5+c)\log_{\frac{1}{\beta}}(k+1)$ for $k \geq 0$ implies that the total number of communication rounds right before the K-th iteration is equal to $t_K = \sum_{k=0}^{K-1} q_k \leq (5+c)K\log_{\frac{1}{\beta}}(K)$.

III. NUMERICAL EXPERIMENT

In this section, we illustrate the practical performance of DPDA-TV when the network is either undirected or directed for solving synthetic C-LASSO problems. We first test the effect of network topology on the performance of DPDA-TV, and then we compare DPDA-TV with another distributed primal-dual algorithm, DPDA-D, proposed in [28] for solving (3) – it is shown in [28] that DPDA-D converge with $\mathcal{O}(1/K)$ ergodic rate when $\bar{\varphi}$ is merely convex. That said, when $\bar{\varphi}$ is strongly convex with modulus $\mu > 0$ as assumed in this paper, using the fact that $\varphi(\bar{\mathbf{x}}^K) - \varphi(\mathbf{x}^*) \geq \frac{\mu}{2}\left\|\bar{\mathbf{x}}^K - \mathbf{x}^*\right\|^2$, it immediately follows that $\left\|\bar{\mathbf{x}}^K - \mathbf{x}^*\right\|^2 \leq \mathcal{O}(1/K)$.

We consider an isotonic C-LASSO problem over time-varying network $\{\mathcal{G}^t\}_{t\geq 0}$. Given some $\lambda > 0$, this problem can be formulated in a centralized form as $x^* \triangleq \operatorname{argmin}_{x\in\mathbb{R}^n}\ \Big\{\frac{1}{2}\|Cx-d\|^2 + \lambda\|x\|_1 :\ Ax \leq \mathbf{0}\Big\}$, where the matrix $C = [C_i]_{i\in\mathcal{N}} \in \mathbb{R}^{m|\mathcal{N}|\times n}$, $d = [d_i]_{i\in\mathcal{N}} \in \mathbb{R}^{m|\mathcal{N}|}$, and $A \in \mathbb{R}^{n-1\times n}$. In fact, the matrix A captures the isotonic feature of the unknown target vector $x^\#$, and can be written explicitly as, $A(\ell,\ell) = 1$ and $A(\ell,\ell+1) = -1$, for $1 \leq \ell \leq n-1$, otherwise zero. Each agent i has access to C_i, d_i, and A; hence, by making local copies of x, the decentralized formulation can be expressed as

$$\min_{\substack{\mathbf{x}=[x_i]_{i\in\mathcal{N}}\in\mathcal{C},\\ Ax_i\leq\mathbf{0}\ i\in\mathcal{N}}} \frac{1}{2}\sum_{i\in\mathcal{N}}\|C_ix_i - d_i\|^2 + \frac{\lambda}{|\mathcal{N}|}\sum_{i\in\mathcal{N}}\|x_i\|_1, \qquad (14)$$

where $\mathcal{C}$ is the consensus set - see (4).

In the rest, we set $n = 20$, $m = n+2$, $\lambda = 0.05$ and $\mathcal{K}_i = -\mathbb{R}_+^{n-1}$ for $i \in \mathcal{N}$. Moreover, for each $i \in \mathcal{N}$, we generate $C_i \in \mathbb{R}^{m\times n}$ as follows: after mn entries, i.i.d. with standard Gaussian distribution, are sampled, the condition number of C_i is normalized by sampling the singular values from $[1,3]$ uniformly at random. We generate the first 5 and the last 5 components of $x^\#$ by sampling from $[-10,0]$ and $[0,10]$ uniformly at random in ascending order, respectively, and the other middle 10 components are set to zero; hence, $[x^\#]_j \leq [x^\#]_{j+1}$ for $j = 1,\ldots,n-1$. Finally, we set $d_i = C_i(x^\# + \epsilon_i)$, where $\epsilon_i \in \mathbb{R}^n$ is a random vector with i.i.d. components following Gaussian distribution with zero mean and standard deviation of 10^{-3}.

Generating initial undirected network: $\mathcal{G}_0 = (\mathcal{N}, \mathcal{E}_0)$ is generated as a random small-world network. Given $|\mathcal{N}|$ and the desired number of edges $|\mathcal{E}_0|$, we choose $|\mathcal{N}|$ edges creating a random cycle over nodes, and then the remaining $|\mathcal{E}_0| - |\mathcal{N}|$ edges are selected uniformly at random.

Generating time-varying undirected network: Given $|\mathcal{N}|$ and the desired number of edges $|\mathcal{E}_0|$ for the initial graph, we generate a random small-world $\mathcal{G}_0 = (\mathcal{N}, \mathcal{E}_0)$ as described above. 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 $\lceil p|\mathcal{E}_0|\rceil$ edges of $\mathcal{G}_0$ uniformly at random and we set $\mathcal{E}^{kM-1} = \mathcal{E}_0 \setminus \bigcup_{t=(k-1)M}^{kM-2}\mathcal{E}^t$. In all experiments, we set $M = 5$, $p = 0.8$ and the number of communications per iteration is set to $q_k = 10\ln(k+1)$.

A. Effect of Network Topology on DPDA-TV

In this section, we test the performance of DPDA-TV on *undirected* communication networks. To illustrate the effect of network topology, we consider four scenarios in which the number of nodes $|\mathcal{N}| \in \{10,\ 40\}$ and the average number of edges per node ($|\mathcal{E}|/|\mathcal{N}|$) is either ≈ 1.5 or ≈ 4.5. For each scenario, we plot both the relative error, i.e., $\max_{i\in\mathcal{N}}\left\|x_i^k - x^*\right\| / \|x^*\|$ and the infeasibility, i.e., $\max_{i\in\mathcal{N}} d_{\mathcal{K}_i}(A\bar{x}_i^k) = \max_{i\in\mathcal{N}}\left\|(A\bar{x}_i^k)_+\right\|$ versus iteration number k. All the plots show the average statistics over all 25 randomly generated replications.

DPDA-TV on time-varying undirected networks: We first generated initial undirected small-world networks $\mathcal{G}_0 = (\mathcal{N}, \mathcal{E}_0)$ as described for $(|\mathcal{N}|, |\mathcal{E}_0|) \in \{(10,15),\ (10,45),\ (40,60),\ (40,180)\}$. Next, we generated $\{\mathcal{G}^t\}_{t\geq 1}$ as described above by setting $M = 5$ and $p = 0.8$. For each consensus round $t \geq 1$, V^t is formed according to Metropolis weights, i.e., for each $i \in \mathcal{N}$, $V_{ij}^t = 1/(\max\{d_i, d_j\}+1)$ if $j \in \mathcal{N}_i^t$, $V_{ii}^t = 1 - \sum_{i\in\mathcal{N}_i} V_{ij}^t$, and $V_{ij}^t = 0$ otherwise – see (12) for our choice of $\mathcal{R}^k$.

For DPDA-TV, displayed in Fig. 1, we chose $\delta_1 = \delta_2 = 1$, which lead to the initial step-sizes as $\gamma^0 = \frac{1}{2}$,

$\tau^0 = \frac{1}{L_{\max}+1}$, and $\kappa^0 = \frac{1}{2\|A\|^2}$. In Fig. 2, we plot $\max_{i\in\mathcal{N}} \left\|x_i^k - x^*\right\| / \|x^*\|$ and $\max_{i\in\mathcal{N}} \left\|(A\bar{x}_i^k)_+\right\|$ statistics for DPDA-TV versus iteration number k. Note that compared to average edge density, the network size has more influence on the convergence rate, i.e., the smaller the network faster the convergence is. On the other hand, for fixed size network, as expected, higher the density faster the convergence is.

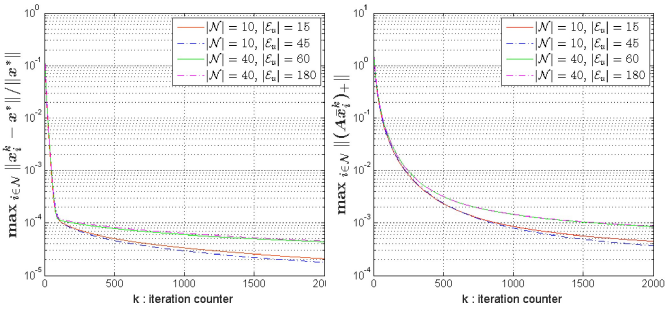

Fig. 2: Effect of network topology on the convergence rate of DPDA-TV

B. Comparison against DPDA-D [28]

We also compared DPDA-TV against DPDA-D, in terms of the relative error and infeasibility of the ergodic iterate sequence, i.e., $\max_{i\in\mathcal{N}} \left\|\bar{x}_i^k - x^*\right\| / \|x^*\|$ and $\max_{i\in\mathcal{N}} \left\|(A\bar{x}_i^k)_+\right\|$. We further report the the relative error of the actual iterate sequence, i.e., $\max_{i\in\mathcal{N}} \left\|x_i^k - x^*\right\| / \|x^*\|$. In this section we fix the number of nodes to $|\mathcal{N}| = 10$ and the average edge density to $|\mathcal{E}|/|\mathcal{N}| = 4.5$ – we observed the same convergence behavior for other networks with different density and size.

Time-varying undirected network: We generated the network sequence $\{\mathcal{G}^t\}_{t\in\mathbb{Z}_+}$ and chose the parameters as in the previous section. Moreover, the constant step-sizes of DPDA-D are set to the initial steps-sizes of DPDA-TV. As it can be seen in Fig. 3, DPDA-TV has faster convergence compared to DPDA-D, confirming the theoretical guarantees: $\mathcal{O}(1/k^2)$ DPDA-TV versus $\mathcal{O}(1/k)$ for DPDA-D.

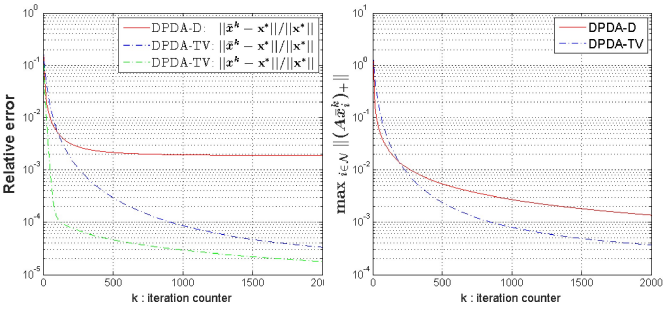

Fig. 3: DPDA-TV vs DPDA-D over undirected time-varying network

Time-varying directed network: In this scenario, we generated time-varying communication networks similar to [36]. Let $\mathcal{G}_d = (\mathcal{N}, \mathcal{E}_d)$ be the directed graph shown in Fig. 5 where it has $|\mathcal{N}| = 12$ nodes and $|\mathcal{E}_d| = 12$ directed edges. We set $\mathcal{G}_0 = \mathcal{G}_d$, and we generate $\{\mathcal{G}^t\}_{t\geq 1}$ as in the undirected case above using the parameters $M = 5$ and $p = 0.8$; hence, $\{\mathcal{G}^t\}_{t\in\mathbb{Z}_+}$ is M-strongly-connected. Moreover, communication weight matrices V^t are formed according to rule (10). We chose the initial step-sizes for DPDA-TV as in the time-varying undirected case, and the constant step-sizes of DPDA-D is set to the initial steps-sizes of DPDA-TV. In Fig. 4 we compare DPDA-TV against DPDA-D. We observe that over time-varying directed networks DPDA-TV again outperforms DPDA-D for both statistics.

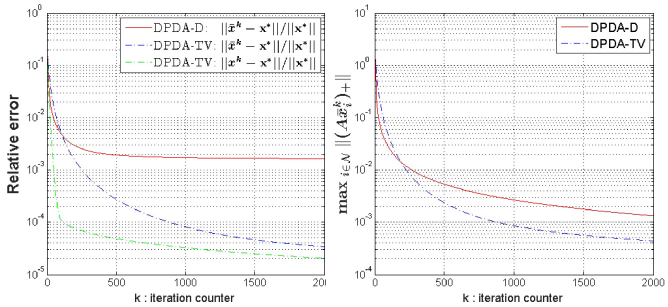

Fig. 4: DPDA-TV vs DPDA-D over directed time-varying network.

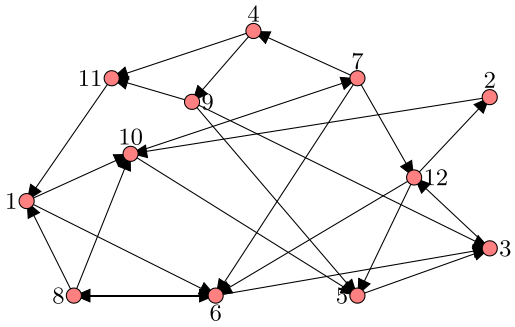

Fig. 5: $\mathcal{G}_d = (\mathcal{N}, \mathcal{E}_d)$ directed strongly connected graph

IV. Conclusions

We proposed a primal-dual algorithm DPDA-TV for solving cooperative multi-agent consensus optimization 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 local conic constraints. We proved that when the sum of local objective functions is strongly convex, while each function can be merely convex, DPDA-TV iterate sequence converges with $\mathcal{O}(1/k^2)$ ergodic rate in terms of suboptimality, infeasibility and consensus violation, requiring a total of $\mathcal{O}(k\log(k))$ local communications in all k DPDA-TV iterations. To the best of our knowledge, this is the best rate result for our setting.

References

[1] Joel B Predd, SB Kulkarni, and H Vincent Poor. Distributed learning in wireless sensor networks. *IEEE Signal Processing Magazine*, 23(4):56–69, 2006.

[2] Ioannis D Schizas, Alejandro Ribeiro, and Georgios B Giannakis. Consensus in ad hoc WSNs with noisy links - Part I: Distributed estimation of deterministic signals. *Signal Processing, IEEE Transactions on*, 56(1):350–364, 2008.

[3] Ke Zhou and Stergios I Roumeliotis. Multirobot active target tracking with combinations of relative observations. *IEEE Transactions on Robotics*, 27(4):678–695, 2011.

[4] Yongcan Cao, Wenwu Yu, Wei Ren, and Guanrong Chen. An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial informatics*, 9(1):427–438, 2013.

[5] Juan Andrés Bazerque and Georgios B Giannakis. Distributed spectrum sensing for cognitive radio networks by exploiting sparsity. *IEEE Transactions on Signal Processing*, 58(3):1847–1862, 2010.

[6] Juan Andrés Bazerque, Gonzalo Mateos, and Georgios B Giannakis. Group-lasso on splines for spectrum cartography. *IEEE Transactions on Signal Processing*, 59(10):4648–4663, 2011.

[7] Konstantinos I Tsianos, Sean Lawlor, and Michael G Rabbat. Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 1543–1550. IEEE, 2012.

[8] John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2012.

[9] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[10] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 713–720, 2011.

[11] Zaid J Towfic, Jianshu Chen, and Ali H Sayed. Collaborative learning of mixture models using diffusion adaptation. In *Machine Learning for Signal Processing (MLSP), 2011 IEEE International Workshop on*, pages 1–6. IEEE, 2011.

[12] Sundhar Srinivasan Ram, Venugopal V Veeravalli, and Angelia Nedic. Distributed non-autonomous power control through distributed convex optimization. In *INFOCOM 2009, IEEE*, pages 3001–3005. IEEE, 2009.

[13] Jianshu Chen and Ali H Sayed. Diffusion adaptation strategies for distributed optimization and learning over networks. *IEEE Transactions on Signal Processing*, 60(8):4289–4305, 2012.

[14] Reza Olfati-Saber, J Alex Fax, and Richard M Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007.

[15] Brian R Gaines and Hua Zhou. Algorithms for fitting the constrained lasso. *arXiv preprint arXiv:1611.01511*, 2016.

[16] Benjamin Hofner, Thomas Kneib, and Torsten Hothorn. A unified framework of constrained regression. *Statistics and Computing*, 26(1-2):1–14, 2016.

[17] Gareth M James, Courtney Paulson, and Paat Rusmevichientong. Penalized and constrained regression. Technical report, Marshall School of Business, University of Southern California, 2013.

[18] Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1):253–287, 2016.

[19] A. Nedic and A. Ozdaglar. *Convex Optimization in Signal Processing and Communications*, chapter Cooperative Distributed Multi-agent Optimization, pages 340–385. Cambridge University Press, 2010.

[20] A. Nedić. Distributed optimization. In *Encyclopedia of Systems and Control*, pages 1–12. Springer, 2014.

[21] Angelia Nedić, Asuman Ozdaglar, and Pablo A Parrilo. Constrained consensus and optimization in multi-agent networks. *Automatic Control, IEEE Transactions on*, 55(4):922–938, 2010.

[22] Kunal Srivastava, Angelia Nedić, and Dušan M Stipanović. Distributed constrained optimization over noisy networks. In *Decision and Control (CDC), 2010 49th IEEE Conference on*, pages 1945–1950. IEEE, 2010.

[23] Minghui Zhu and Sonia Martínez. On distributed convex optimization under inequality and equality constraints. *Automatic Control, IEEE Transactions on*, 57(1):151–164, 2012.

[24] Deming Yuan, Shengyuan Xu, and Huanyu Zhao. Distributed primal–dual subgradient method for multiagent optimization via consensus algorithms. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 41(6):1715–1724, 2011.

[25] Tsung-Hui Chang, Angelia Nedic, and Anna Scaglione. Distributed constrained optimization by consensus-based primal-dual perturbation method. *Automatic Control, IEEE Transactions on*, 59(6):1524–1538, 2014.

[26] David Mateos-Núñez and Jorge Cortés. Distributed subgradient methods for saddle-point problems. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 5462–5467, Dec 2015.

[27] T. H. Chang. A proximal dual consensus ADMM method for multi-agent constrained optimization. *IEEE Transactions on Signal Processing*, 64(14):3719–3734, July 2016.

[28] Necdet Serhat Aybat and Erfan Yazdandoost Hamedani. A primal-dual method for conic constrained distributed optimization problems. In *Advances in Neural Information Processing Systems*, pages 5050–5058, 2016.

[29] A. Makhdoumi and A. Ozdaglar. Convergence rate of distributed admm over networks. *IEEE Transactions on Automatic Control*, 62(10):5082–5095, Oct 2017.

[30] T. H. Chang, M. Hong, and X. Wang. Multi-agent distributed optimization via inexact consensus ADMM. *IEEE Transactions on Signal Processing*, 63(2):482–497, Jan 2015.

[31] Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

[32] Jinshan Zeng and Wotao Yin. ExtraPush for convex smooth decentralized optimization over directed networks. *Journal of Computational Mathematics*, 35(4):381–394, 2017.

[33] C. Xi, Q. Wu, and U. A. Khany. Fast distributed optimization over directed graphs. In *2016 American Control Conference (ACC)*, pages 6507–6512, July 2016.

[34] C. Xi, R. Xin, and U. A. Khan. Add-opt: Accelerated distributed directed optimization. *IEEE Transactions on Automatic Control*, PP(99):1–1, 2017. arXiv preprint arXiv:1607.04757.

[35] Angelia Nedić and Alex Olshevsky. Stochastic gradient-push for strongly convex functions on time-varying directed graphs. *IEEE Transactions on Automatic Control*, 61(12):3936–3947, 2016.

[36] Angelia Nedich, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *arXiv preprint arXiv:1607.03218*, 2016.

[37] Paolo Di Lorenzo and Gesualdo Scutari. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.

[38] Ying Sun, Gesualdo Scutari, and Daniel Palomar. Distributed nonconvex multiagent optimization over time-varying networks. In *Signals, Systems and Computers, 2016 50th Asilomar Conference on*, pages 788–794. IEEE, 2016.

[39] Ying Sun and Gesualdo Scutari. Distributed nonconvex optimization for sparse representation. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 4044–4048. IEEE, 2017.

[40] Antonin Chambolle and Thomas Pock. A remark on accelerated block coordinate descent for computing the proximity operators of a sum of convex functions. *SMAI Journal of Computational Mathematics*, 1:29–54, 2015.

[41] Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. Multi-agent constrained optimization of a strongly convex function over time-varying directed networks. *arXiv preprint arXiv:1706.07907*, 2017.

[42] David Kempe, Alin Dobra, and Johannes Gehrke. Gossip-based computation of aggregate information. In *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on*, pages 482–491. IEEE, 2003.

[43] Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2015.

[44] Angelia Nedić and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *Automatic Control, IEEE Transactions on*, 54(1):48–61, 2009.