MULTI-AGENT CONSTRAINED OPTIMIZATION OF A STRONGLY CONVEX FUNCTION

Erfan Yazdandoost Hamedani *Necdet Serhat Aybat**

Industrial & Manufacturing Engineering Department,
The Pennsylvania State University, PA, USA.
Emails: evy5047@psu.edu, nsa10@psu.edu

ABSTRACT

We consider cooperative multi-agent consensus optimization problems over an undirected network of agents, where only local communications are allowed. The objective is to minimize the sum of agent-specific convex functions over agent-specific private conic constraint sets. We provide convergence rates in sub-optimality, infeasibility and consensus violation when the sum function is strongly convex; examine the effect of underlying network topology on the convergence rates of the proposed decentralized algorithm.

Index Terms— multi-agent distributed optimization, consensus, constrained optimization, convergence rate

1. INTRODUCTION

Decentralized optimization over communication networks has various applications: 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]; v) power control problem in cellular networks [11], to name a few application areas. In many of these applications, the network size is usually 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 the data, and also may violate data privacy in case agents are not willing to share their data even though they are collaborative agents [12, 13].

In this paper, from a broader perspective, we aim to study constrained distributed optimization of a strongly convex function over static communication network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$; in particular, from an application perspective, we are motivated to design an efficient decentralized solution method for *constrained LASSO* (C-LASSO) problems [14] with distributed data. C-LASSO, having the generic form $\min_x\{\lambda \|x\|_1 + \|Cx - d\|_2^2 : \ Ax \leq b\}$, is an important class of statistical problems, which includes fused LASSO, constrained regression, and generalized LASSO problems as its special cases [15, 14, 16] to name a few. In the rest, we provide our results for a more general, constrained decentralized optimization setting. 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 $\mathcal{G}$. In our set up, we also consider the case where each local function φ_i is convex but not necessarily strongly convex for all $i \in \mathcal{N}$. 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 $\mathbf{rank}(C) = n \leq \sum_{i\in\mathcal{N}} m_i$ where $C = [C_i]_{i\in\mathcal{N}}$. Therefore, it is important to note that in the centralized formulation of this problem $\min_x \sum_{i\in\mathcal{N}} \varphi_i(x)$ the objective 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 are special cases of this generic conic constrained decentralized optimization framework. With the motivation of designing an efficient decentralized solution method for the distributed conic constrained problem over a static communication network $\mathcal{G}$ as we briefly discussed above, we propose a *distributed* primal-dual algorithm (DPDA). DPDA is based on the primal-dual algorithm (PDA), recently proposed in [17] 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, $\Phi(\mathbf{x}) \triangleq \rho(\mathbf{x}) + g(\mathbf{x})$ is a *strongly convex* function with modulus $\mu > 0$ such that ρ and h are possibly non-smooth convex functions, g is convex and has a Lipschitz continuous gradient defined on $\mathbf{dom}\,\rho$ with constant L, and T is a linear map. In [17], it is shown for PDA that for any $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, the *ergodic* average sequence $\{\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k\}_{k\geq 0}$ satisfies $\mathcal{L}(\bar{\mathbf{x}}^k, \mathbf{y}) - \mathcal{L}(\mathbf{x}, \bar{\mathbf{y}}^k) = \mathcal{O}(1/k^2)$ for appropriately chosen primal-dual step-sizes.

*Research of N. S. Aybat was partially supported by NSF grants CMMI-1400217 and CMMI-1635106, and ARO grant W911NF-17-1-0298.

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

Problem Definition. Let $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ denote a *connected* undirected graph of N computing nodes, where $\mathcal{N} \triangleq \{1, \ldots, N\}$ and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ denotes the set of edges – without loss of generality assume that $(i,j) \in \mathcal{E}$ implies $i < j$. Suppose nodes i and j can exchange information only if $(i,j) \in \mathcal{E}$, and each node $i \in \mathcal{N}$ has a *private* (local) cost function $\varphi_i : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ such that $\varphi_i(x) \triangleq \rho_i(x) + f_i(x)$, 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 that 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, $\mathbf{prox}_{\rho_i}(x) \triangleq \operatorname{argmin}_{y \in \mathbb{R}^n} \left\{ \rho_i(y) + \frac{1}{2} \|y - x\|_2^2 \right\}$, is *efficiently* computable for $i \in \mathcal{N}$. Let $\mathcal{N}_i \triangleq \{j \in \mathcal{N} : (i,j) \in \mathcal{E} \text{ or } (j,i) \in \mathcal{E}\}$ denote the set of neighboring nodes of $i \in \mathcal{N}$, and $d_i \triangleq |\mathcal{N}_i|$ is the degree of node $i \in \mathcal{N}$. Let $\bar{\varphi}(x) \triangleq \sum_{i \in \mathcal{N}} \varphi_i(x)$ and consider the following problem

$$x^* \in \operatorname*{argmin}_{x \in \mathbb{R}^n} \ \bar{\varphi}(x) \quad \text{s.t.} \quad A_i x - b_i \in \mathcal{K}_i, \quad i \in \mathcal{N}, \qquad (1)$$

where $A_i \in \mathbb{R}^{m_i \times n}$, $b_i \in \mathbb{R}^{m_i}$ and $\mathcal{K}_i \subseteq R^{m_i}$ is a closed, convex cone for $i \in \mathcal{N}$. 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 preimage requires solving an SDP.

Assumption 1.1. *The duality gap for* (1) *is zero, and a primal-dual solution to* (1) *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 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. *Each f_i is strongly convex with modulus $\mu_i \geq 0$ for $i \in \mathcal{N}$, $\bar{f}(x) \triangleq \sum_{i \in \mathcal{N}} f_i(x)$ is strongly convex with modulus $\bar{\mu} > 0$, and define $\underline{\mu} \triangleq \min_{i \in \mathcal{N}} \{\mu_i\}$.*

Remark. Note that while $\bar{\mu} \geq \sum_{i \in \mathcal{N}} \mu_i$, 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 (1).

Contribution. To the best of our knowledge, only a handful of methods, e.g., [18, 19, 20] can handle constrained consensus problems similar to (1) 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 primal-dual distributed methods in [19, 20] when implemented for the agent-specific conic constraint sets χ_i studied in this paper; moreover, none of these three methods exploits the strong convexity of the sum $\bar{\varphi} = \sum_{i \in \mathcal{N}} \varphi_i$. We believe that DPDA proposed in this paper is one of the first decentralized algorithms to solve (1) with $\mathcal{O}(1/k^2)$ ergodic rate guarantee on both sub-optimality and infeasibility.

Notation. Throughout $\|.\|$ denotes either the Euclidean norm or the spectral norm. Given a convex set $\mathcal{S}$, let $\sigma_{\mathcal{S}}(.)$ denote its support function, i.e., $\sigma_{\mathcal{S}}(\theta) \triangleq \sup_{w \in \mathcal{S}} \langle \theta,\, w \rangle$. 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 $\mathcal{K}^\circ \triangleq -\mathcal{K}^*$ denotes 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 $\theta \notin \mathcal{K}^\circ$. $\otimes$ denotes the Kronecker product, and $\mathbf{I}_n$ is the $n \times n$ identity matrix. $\mathbb{S}_+^n$ denotes the cone of symmetric positive semidefinite matrices. For $Q \succeq 0$, i.e., $Q \in \mathbb{S}_+^n$, we define $\|z\|_Q \triangleq \sqrt{z^\top Q z}$ and $\lambda_{\min}^+(Q)$ denotes smallest positive eigenvalue of Q.

2. METHODOLOGY

Let $x_i \in \mathbb{R}^n$ denote the *local* decision vector of node $i \in \mathcal{N}$. By taking advantage of the fact that $\mathcal{G}$ is *connected*, we can reformulate (1) as the following *distributed consensus* optimization problem:

$$\min_{x_i \in \mathbb{R}^n,\, i \in \mathcal{N}} \left\{ \sum_{i \in \mathcal{N}} \varphi_i(x_i) \mid \begin{array}{c} x_i = x_j : \lambda_{ij},\ \forall (i,j) \in \mathcal{E}, \\ A_i x_i - b_i \in \mathcal{K}_i : \theta_i,\ \forall i \in \mathcal{N} \end{array} \right\}, \qquad (2)$$

where $\lambda_{ij} \in \mathbb{R}^n$ and $\theta_i \in \mathbb{R}^{m_i}$ are the corresponding dual variables. Let $\mathbf{x} = [x_i]_{i \in \mathcal{N}} \in \mathbb{R}^{n|\mathcal{N}|}$. The consensus constraints $x_i = x_j$ for $(i,j) \in \mathcal{E}$ can be formulated as $M\mathbf{x} = 0$, where $M \in \mathbb{R}^{n|\mathcal{E}| \times n|\mathcal{N}|}$ is a block matrix such that $M = H \otimes \mathbf{I}_n$ where H is the oriented edge-node incidence matrix, i.e., the entry $H_{(i,j),l}$, corresponding to edge $(i,j) \in \mathcal{E}$ and $l \in \mathcal{N}$, is equal to 1 if $l = i$, -1 if $l = j$, and 0 otherwise. Note that $M^\mathsf{T} M = H^\mathsf{T} H \otimes \mathbf{I}_n = \Omega \otimes \mathbf{I}_n$, where $\Omega \in \mathbb{R}^{|\mathcal{N}| \times |\mathcal{N}|}$ denotes the graph Laplacian of $\mathcal{G}$, i.e., $\Omega_{ii} = d_i$, $\Omega_{ij} = -1$ if $(i,j) \in \mathcal{E}$ or $(j,i) \in \mathcal{E}$, and equal to 0 otherwise.

Since x^* is the unique solution to (1) and since $\mathbf{x}^* \triangleq \mathbf{1} \otimes x^*$ satisfies $(\Omega \otimes \mathbf{I}_n)\mathbf{x}^* = 0$, one can reformulate (1) as a saddle point problem. Indeed, for any $\alpha \geq 0$, one can solve (1) through solving

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \mathcal{L}(\mathbf{x}, \mathbf{y}) \triangleq \frac{\alpha}{2} \|\mathbf{x}\|_{\Omega \otimes \mathbf{I}_n}^2 + \langle \boldsymbol{\lambda}, M\mathbf{x} \rangle + \sum_{i \in \mathcal{N}} \Big(\varphi_i(x_i) + \langle \theta_i, A_i x_i - b_i \rangle - \sigma_{\mathcal{K}_i}(\theta_i) \Big). \qquad (3)$$

Next, given $\alpha \geq 0$, we consider the direct implementation of PDA [17] to solve (3) for appropriately chosen algorithm parameters such as primal-dual step sizes and a componentwise separable Bregman distance function on $\mathcal{X}$.

Definition 2. *Let $\mathcal{X} \triangleq \Pi_{i \in \mathcal{N}} \mathbb{R}^n$ and $\mathcal{X} \ni \mathbf{x} = [x_i]_{i \in \mathcal{N}}$; $\mathcal{Y} \triangleq \Pi_{i \in \mathcal{N}} \mathbb{R}^{m_i} \times \mathbb{R}^{m_0}$, $\mathcal{Y} \ni \mathbf{y} = [\boldsymbol{\theta}^\top \boldsymbol{\lambda}^\top]^\top$ and $\boldsymbol{\theta} =$*

$[\theta_i]_{i\in\mathcal{N}} \in \mathbb{R}^m$, *where* $m \triangleq \sum_{i\in\mathcal{N}} m_i$. *Let* $\Phi : \mathcal{X} \to \mathbb{R}\cup\{\infty\}$ *such that* $\Phi(\mathbf{x}) = \rho(\mathbf{x}) + g(\mathbf{x})$ *where* $\rho(\mathbf{x}) \triangleq \sum_{i\in\mathcal{N}} \rho_i(x_i)$, $g(\mathbf{x}) \triangleq f(\mathbf{x}) + \frac{\alpha}{2}\|\mathbf{x}\|^2_{\Omega\otimes\mathbf{I}_n}$ *and* $f(\mathbf{x}) \triangleq \sum_{i\in\mathcal{N}} f_i(x_i)$, *and let* $h : \mathcal{Y} \to \mathbb{R}\cup\{\infty\}$ *such that* $h(\mathbf{y}) \triangleq \sum_{i\in\mathcal{N}} \sigma_{\mathcal{K}_i}(\theta_i) + \langle b_i, \theta_i\rangle$. *Define the block-diagonal matrix* $A \triangleq \mathbf{diag}([A_i]_{i\in\mathcal{N}}) \in \mathbb{R}^{m\times n|\mathcal{N}|}$ *and* $T = [A^\top\ M^\top]^\top$.

Given some positive parameters $\gamma^k, \tau^k > 0$, $\kappa_i^k > 0$ for $i \in \mathcal{N}$ – we shortly discuss how to select them, and given Φ, h and T as in Definition 2, and the initial iterates $\mathbf{x}^0$ and $\mathbf{y}^0 = [{\boldsymbol{\theta}^0}^\top {\boldsymbol{\lambda}^0}^\top]^\top$, the PDA iterations take the following form:

$$\theta_i^{k+1} \leftarrow \operatorname*{argmin}_{\theta_i} \sigma_{\mathcal{K}_i}(\theta_i) - \langle A_i(x_i^k + \eta^k(x_i^k - x_i^{k-1})) - b_i, \theta_i\rangle + \frac{1}{2\kappa_i^k}\|\theta_i - \theta_i^k\|^2, \quad i \in \mathcal{N}, \tag{4a}$$

$$\boldsymbol{\lambda}^{k+1} \leftarrow \operatorname*{argmin}_{\boldsymbol{\lambda}} -\langle M(\mathbf{x}^k + \eta^k(\mathbf{x}^k - \mathbf{x}^{k-1})), \boldsymbol{\lambda}\rangle + \frac{1}{2\gamma^k}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^k\|^2 = \boldsymbol{\lambda}^k + \gamma^k M(\mathbf{x}^k + \eta^k(\mathbf{x}^k - \mathbf{x}^{k-1})), \tag{4b}$$

$$\mathbf{x}^{k+1} \leftarrow \operatorname*{argmin}_{\mathbf{x}} \left\langle M^\top\boldsymbol{\lambda}^{k+1} + \alpha(\Omega\otimes\mathbf{I}_n)\mathbf{x}^k, \mathbf{x}\right\rangle + \sum_{i\in\mathcal{N}} \rho_i(x_i) + \langle \nabla f_i(x_i^k) + A_i^\top\theta_i^{k+1}, x_i\rangle + \frac{1}{2\tau^k}\|x_i - x_i^k\|^2. \tag{4c}$$

Since $\mathcal{K}_i$ is a cone, $\mathbf{prox}_{\kappa_i^k\sigma_{\mathcal{K}_i}}(\cdot) = \mathcal{P}_{\mathcal{K}_i^\circ}(\cdot)$; hence, $\theta_i^{k+1} = \mathcal{P}_{\mathcal{K}_i^\circ}\Big(\theta_i^k + \kappa_i^k\Big(A_i(x_i^k + \eta^k(x_i^k - x_i^{k-1})) - b_i\Big)\Big)$ for $i \in \mathcal{N}$. Using recursion in (4b), we can write $\boldsymbol{\lambda}^k$ as a partial summation of primal iterates $\{\mathbf{x}^\ell\}_{\ell=0}^{k-1}$, i.e., $\boldsymbol{\lambda}^k = \boldsymbol{\lambda}^0 + \sum_{\ell=0}^{k-1}\gamma^\ell M(\mathbf{x}^\ell + \eta^\ell(\mathbf{x}^\ell - \mathbf{x}^{\ell-1}))$. Let $\boldsymbol{\lambda}^0 \leftarrow \mathbf{0}$, and define $\{\mathbf{s}^k\}_{k\geq 0}$ such that $\mathbf{s}^0 = \mathbf{0}$ and $\mathbf{s}^{k+1} = \mathbf{s}^k + \gamma^k(\mathbf{x}^k + \eta^k(\mathbf{x}^k - \mathbf{x}^{k-1}))$ for $k \geq 0$; hence, $\boldsymbol{\lambda}^k = M\mathbf{s}^k$ for $k \geq 0$. Using $M^\top M = \Omega\otimes\mathbf{I}_n$, we obtain $\langle \mathbf{x}, M^\top\boldsymbol{\lambda}^{k+1}\rangle = \langle \mathbf{x},\ (\Omega\otimes\mathbf{I}_n)\mathbf{s}^{k+1}\rangle = \sum_{i\in\mathcal{N}}\langle x_i, \sum_{j\in\mathcal{N}_i}(s_i^{k+1} - s_j^{k+1})\rangle$. Thus, PDA iterations given in (4) for the static graph $\mathcal{G}$ can be computed in a *decentralized* way, via the node-specific computations as in distributed primal dual algorithm (DPDA) displayed in Fig. 1 below.

Definition 3. *A weighted Laplacian matrix* $W \in \mathbb{S}_+^{|\mathcal{N}|}$ *is such that* $W_{ij} = W_{ji} < 0$ *for* $(i,j) \in \mathcal{E}$, $W_{ij} = W_{ji} = 0$ *for* $(i,j) \notin \mathcal{E}$, *and* $W_{ii} = -\sum_{j\in\mathcal{N}} W_{ij}$ *for* $i \in \mathcal{N}$.

Remark. When $\underline{\mu} > 0$, according to Assumption 1.2, $f(\mathbf{x}) = \sum_{i\in\mathcal{N}} f_i(x_i)$ is strongly convex with modulus $\underline{\mu}$. That said, 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 not when $\underline{\mu} = 0$.

Inspired from Proposition 3.6. in [21], we prove the following Lemma with a slight difference in choosing parameter, showing that by suitably regularizing f, one can obtain a strongly convex function 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* $f_\alpha(\mathbf{x}) \triangleq f(\mathbf{x}) + \alpha\, r(\mathbf{x})$, *where* $r(\mathbf{x}) \triangleq \frac{1}{2}\|\mathbf{x}\|^2_{W\otimes\mathbf{I}_n}$. *Then*

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

Initialization: $\mathbf{x}^{-1} \leftarrow \mathbf{x}^0$, $\mathbf{s}^0 \leftarrow \mathbf{0}$,
$\delta_1, \delta_2 > 0$, $\mu \in (0,\ \max\{\underline{\mu},\ \mu_\alpha\})$
$\tau^0 \leftarrow \min_{i\in\mathcal{N}} \frac{1}{L_i + \delta_2 + 2d_i\alpha}$, $\tilde{\tau}^0 \leftarrow (\frac{1}{\tau^0} - \mu)^{-1}$,
$\eta^0 \leftarrow 0$, $\gamma^0 \leftarrow \min_{i\in\mathcal{N}} \frac{\delta_2}{2d_i + \delta_1}$, $\kappa_i^0 \leftarrow \gamma^0\frac{\delta_1}{\|A_i\|^2}$ $i \in \mathcal{N}$
Step k: $(k \geq 0), \forall i \in \mathcal{N}$
1. $q_i^k \leftarrow x_i^k + \eta^k(x_i^k - x_i^{k-1})$,
2. $\theta_i^{k+1} \leftarrow \mathcal{P}_{\mathcal{K}_i^\circ}\Big(\theta_i^k + \kappa_i^k\big(A_i q_i^k - b_i\big)\Big)$,
3. $s_i^{k+1} \leftarrow s_i^k + \gamma^k q_i^k$,
4. $x_i^{k+1} \leftarrow \mathbf{prox}_{\tau^k\rho_i}\Big(x_i^k - \tau^k\Big(\nabla f_i(x_i^k) + A_i^\top\theta_i^{k+1} + \sum_{j\in\mathcal{N}_i}(s_i^{k+1} - s_j^{k+1}) + \alpha\sum_{j\in\mathcal{N}_i}(x_i^k - x_j^k)\Big)\Big)$,
5. $\eta^{k+1} \leftarrow \frac{1}{\sqrt{1+\mu\tilde{\tau}^k}}$, $\tilde{\tau}^{k+1} \leftarrow \eta^{k+1}\tilde{\tau}^k$,
6. $\tau^{k+1} \leftarrow (\frac{1}{\tilde{\tau}^{k+1}} + \mu)^{-1}$
7. $\gamma^{k+1} \leftarrow \gamma^k/\eta^{k+1}$, $\kappa_i^{k+1} \leftarrow \gamma^{k+1}\frac{\delta_1}{\|A_i\|^2}$

Fig. 1: Distributed Primal Dual Algorithm (DPDA)

f_α *is strongly convex with modulus* $\mu_\alpha \triangleq \frac{\bar{\mu}/|\mathcal{N}| + \alpha\lambda_2}{2} - \sqrt{\Big(\frac{\bar{\mu}/|\mathcal{N}| - \alpha\lambda_2}{2}\Big)^2 + 4\bar{L}^2} > 0$ *for any for* $\alpha > \frac{4|\mathcal{N}|}{\lambda_2\bar{\mu}}\bar{L}^2$, *where* $\bar{L} = \sqrt{\frac{\sum_{i\in\mathcal{N}} L_i^2}{|\mathcal{N}|}}$ *and* $\lambda_2 = \lambda_{\min}^+(W)$.

Proof. Let $\mathbf{x}^* = \mathbf{1}_{|\mathcal{N}|}\otimes x^*$, where x^* is the unique optimal solution to (1), and according to Assumption 1.2, $\bar{f}$ is strongly convex with modulus $\bar{\mu} > 0$. Fix some arbitrary $\alpha > \frac{4}{\lambda_2\bar{\mu}}\sum_{i\in\mathcal{N}} L_i^2$ and $\mathbf{x} \in \mathbb{R}^{n|\mathcal{N}|}$. Then, using $\mathbf{Null}(W) = \mathbf{Span}\{\mathbf{1}\}$, any $\mathbf{x} \in \mathbb{R}^{n|\mathcal{N}|}$ can be decomposed into $\mathbf{u} \in \mathbf{Span}\{\mathbf{1}\}$ and $\mathbf{v} \in \mathbf{Span}\{\mathbf{1}\}^\perp$ where $\mathbf{x} = \mathbf{u} + \mathbf{v}$ and $\|\mathbf{x}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$. From definition of f_α we have that,

$$\langle \nabla f_\alpha(\mathbf{x}) - \nabla f_\alpha(\mathbf{x}^*),\ \mathbf{x} - \mathbf{x}^*\rangle = \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*),\ \mathbf{x} - \mathbf{x}^*\rangle + \alpha\|\mathbf{x} - \mathbf{x}^*\|^2_{W\otimes\mathbf{I}_n}. \tag{5}$$

Let $\bar{L} \triangleq \sqrt{\frac{\sum_{i\in\mathcal{N}} L_i^2}{N}}$, where $N \triangleq |\mathcal{N}|$. The inner product on the rhs of (5) can be bounded by using convexity, Lipschitz differentiability, and strong convexity of f as follows:

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*),\ \mathbf{x} - \mathbf{x}^*\rangle \geq \frac{\bar{\mu}}{N}\|\mathbf{x}^* - \mathbf{x}\|^2 - 2\bar{L}\|\mathbf{x}^* - \mathbf{u}\|\|\mathbf{v}\|. \tag{6}$$

From (5), (6) and the fact that $\|\mathbf{x} - \mathbf{x}^*\|^2_{W\otimes\mathbf{I}_n} = \|\mathbf{v}\|^2_{W\otimes\mathbf{I}_n} \geq \lambda_2\|\mathbf{v}\|^2$, we get

$$\langle \nabla f_\alpha(\mathbf{x}) - \nabla f_\alpha(\mathbf{x}^*),\ \mathbf{x} - \mathbf{x}^*\rangle \geq \frac{\bar{\mu}}{N}\|\mathbf{x}^* - \mathbf{x}\|^2 - 2\bar{L}\|\mathbf{x}^* - \mathbf{u}\|\|\mathbf{v}\| + \alpha\lambda_2\|\mathbf{v}\|^2. \tag{7}$$

Next, fix $\omega > 0$. We consider two cases:
(i) $\|\mathbf{v}\| \leq \omega\|\mathbf{u} - \mathbf{x}^*\|$; hence, from (7),

$$\begin{aligned}\langle \nabla f_\alpha(\mathbf{x}) - \nabla f_\alpha(\mathbf{x}^*),\ \mathbf{x} - \mathbf{x}^*\rangle &\geq \big(\frac{\bar{\mu}}{N} - 2\omega\bar{L}\big)\|\mathbf{u} - \mathbf{x}^*\|^2 + \alpha\lambda_2\|\mathbf{v}\|^2 \\ &\geq \min\Big\{\frac{\bar{\mu}}{N} - 2\omega\bar{L},\ \alpha\lambda_2\Big\}\|\mathbf{x} - \mathbf{x}^*\|^2,\end{aligned} \tag{8}$$

and (ii): $\|\mathbf{v}\| \geq \omega \|\mathbf{u} - \mathbf{x}^*\|$; hence,

$$\begin{aligned}\langle \nabla f_\alpha(\mathbf{x}) - \nabla f_\alpha(\mathbf{x}^*),\, \mathbf{x} - \mathbf{x}^* \rangle \\ \geq \frac{\bar{\mu}}{N} \|\mathbf{u} - \mathbf{x}^*\|^2 + \left(\alpha\lambda_2 - \frac{2\bar{L}}{\omega}\right) \|\mathbf{v}\|^2 \\ \geq \min\left\{\frac{\bar{\mu}}{N},\, \alpha\lambda_2 - \frac{2\bar{L}}{\omega}\right\} \|\mathbf{x} - \mathbf{x}^*\|^2 . \qquad (9)\end{aligned}$$

Combining (8) and (9) we conclude that,

$$\begin{aligned}\langle \nabla f_\alpha(\mathbf{x}) - \nabla f_\alpha(\mathbf{x}^*),\, \mathbf{x} - \mathbf{x}^* \rangle \\ \geq \min\left\{\frac{\bar{\mu}}{N} - 2\bar{L}\omega,\, \alpha\lambda_2 - \frac{2\bar{L}}{\omega}\right\} \|\mathbf{x} - \mathbf{x}^*\|^2 . \qquad (10)\end{aligned}$$

Since $\omega \geq 0$ is arbitrary, f_α is strongly convex with modulus $\mu_\alpha = \max_{\omega \geq 0} \min\left\{\frac{\bar{\mu}}{N} - 2\bar{L}\omega,\, \alpha\lambda_2 - \frac{2\bar{L}}{\omega}\right\}$. Note μ_α is attained for $\omega_\alpha \geq 0$ such that $\frac{\bar{\mu}}{N} - 2\bar{L}\omega_\alpha = \alpha\lambda_2 - \frac{2\bar{L}}{\omega_\alpha}$, which implies that $\omega_\alpha = \frac{1}{2}\left(\frac{\bar{\mu}/N - \alpha\lambda_2}{2\bar{L}} + \sqrt{\frac{\bar{\mu}/N - \alpha\lambda_2}{2\bar{L}} + 4}\right)$. Moreover, $\mu_\alpha = \frac{\bar{\mu}}{N} - 2\bar{L}\omega_\alpha$ is the value given in the statement of the lemma, and we have $\frac{\bar{\mu}}{N} > \mu_\alpha > 0$ for any $\alpha > \frac{4N}{\lambda_2 \bar{\mu}} \bar{L}^2$. It is worth mentioning that μ_α is a concave increasing function of α over $\mathbb{R}_{++}$, and $\sup_{\alpha>0} \mu_\alpha = \lim_{\alpha \nearrow \infty} \mu_\alpha = \frac{\bar{\mu}}{N}$. □

Remark 2.1. *When $\underline{\mu} > 0$, i.e., all f_i's are strongly convex, the parameter α can be set to zero; hence, $g(\mathbf{x}) = f(\mathbf{x})$ is strongly convex with modulus $\mu_g = \underline{\mu}$. Otherwise, when $\underline{\mu} = 0$, α should be chosen according to Lemma 2.1; hence, $g(\mathbf{x}) = f_\alpha(\mathbf{x})$ is strongly convex with modulus $\mu_g = \mu_\alpha$. The condition $\alpha > \frac{4}{\bar{\mu}\lambda^+_{\min}(W)} \sum_{i\in\mathcal{N}} L_i^2$ is similar to the one in [21], where α should be greater than $\frac{|\mathcal{N}| L^2_{\max}}{2\bar{\mu}\lambda^+_{\min}(W)}$ for some $W \in \mathbb{S}_+^{|\mathcal{N}|}$ which is a parameter for their algorithm satisfying certain conditions and $L_{\max} = \max_{i\in\mathcal{N}} L_i$.*

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

Theorem 2.2. *Suppose Assumption 1.1 holds. Let $\{\mathbf{x}^k, \boldsymbol{\theta}^k\}_{k\geq 0}$ be the sequence generated by Algorithm DPDA, displayed in Fig. 1, initialized from an arbitrary $\mathbf{x}^0$ and $\boldsymbol{\theta}^0 = \mathbf{0}$. Then $\{\mathbf{x}^k\}_{k\geq 0}$ converges to $\mathbf{x}^* = \mathbf{1}\otimes x^*$ such that x^* is the optimal solution to* (1)*; moreover, the following error bounds,*

$$\left\|M\bar{\mathbf{x}}^K\right\| + \sum_{i\in\mathcal{N}} \|\theta_i^*\|\, d_{\mathcal{K}_i}(A_i\bar{x}_i^K - b_i) \leq \Theta_0/N_K,$$

$$|\Phi(\bar{\mathbf{x}}^K) - \varphi(\mathbf{x}^*)| \leq \Theta_0/N_K, \quad \left\|\mathbf{x}^K - \mathbf{x}^*\right\|^2 \leq \frac{\tilde{\tau}^K}{\gamma^K} 2\gamma^0\Theta_0,$$

hold for all $K \geq 1$, where $\bar{\mathbf{x}}^K = N_K^{-1}\sum_{k=1}^K \gamma^{k-1}\mathbf{x}^k$, $N_K = \sum_{k=1}^K \gamma^{k-1} = \mathcal{O}(K^2)$, and $\Theta_0 \triangleq \sum_{i\in\mathcal{N}}\left[\frac{1}{2\tau^0}\|x_i^0 - x^*\|^2 + \frac{2}{\kappa_i^0}\|\theta_i^*\|^2\right] + \frac{1}{2\gamma^0}$. Moreover, $\tilde{\tau}^K/\gamma^K = \mathcal{O}(1/K^2)$.

Proof. Due to its technical nature and lack of space, the proof is included in the online technical report [22]. □

Remark 2.2. *Note that the result in Theorem 2.2 can be extended to weighted graphs by replacing the Laplacian matrix Ω in g, with the weighted Laplacian W, and also replacing consensus constraint $M\mathbf{x} = 0$ in* (2) *with $(W \otimes \mathbf{I}_n)\mathbf{x} = 0$.*

3. NUMERICAL SECTION

In this section we illustrate the performance of DPDA by implementing on constrained LASSO problem and compare it with distributed primal-dual algorithm DPDA-S in [18] which is proposed for solving (1) with $\mathcal{O}(1/K)$ ergodic convergence rate when $\bar{\varphi}$ is merely convex. We consider an isotonic constrained LASSO problem over network $\mathcal{G}(\mathcal{N}, \mathcal{E})$, which can be formulated in a centralized form as $\min_{x\in\mathbb{R}^n}\left\{\frac{1}{2}\|Cx - d\|^2 + \lambda\|x\|_1 \;:\; Ax \leq \mathbf{0}\right\}$ 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, matrix A captures the isotonic feature of 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 it is zero. By making local copies of x, the decentralized formulation can be expressed as

$$\min_{\substack{M\mathbf{x}=0,\\ Ax_i\leq\mathbf{0}\; i\in\mathcal{N}}} \frac{1}{2}\sum_{i\in\mathcal{N}} \|C_i x_i - d_i\|^2 + \frac{\lambda}{|\mathcal{N}|}\sum_{i\in\mathcal{N}} \|x_i\|_1 . \qquad (11)$$

Graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ is generated as a random small-world network. Given $|\mathcal{N}|$ and the desired number of edges $|\mathcal{E}|$, we choose $|\mathcal{N}|$ edges creating a random cycle over nodes, and then the remaining $|\mathcal{E}| - |\mathcal{N}|$ edges are selected uniformly at random. We set $n = 20$, $m = n + 2$, and $\lambda = 0.05$. For any $i \in \mathcal{N}$, we set $\mathcal{K}_i = -\mathbb{R}_+^{n-1}$, and entries of C_i are sampled from standard Gaussian distribution and the condition number of C_i is normalized by sampling the singular values from $U[1,3]$. We let $d_i = C_i(x^* + \epsilon_i)$, where the first 5 and the last 5 components of x^* are generated by choosing from $U[-10,0]$ and $U[0,10]$ in ascending order, respectively, and other 10 components are set to zero, and components of $\epsilon_i \in \mathbb{R}^n$ are i.i.d with Gaussian distribution having zero mean and standard deviation of 10^{-3}. We tested our method on problem (11), by setting $\delta_1 = d_{\max}$ and $\delta_2 = 2L_{\max}$ which lead to initial step-sizes $\gamma^0 = \frac{2}{3}\frac{L_{\max}}{d_{\max}}$, $\tau^0 = \frac{1}{3L_{\max}}$, and $\kappa^0 = \frac{2}{3}\frac{L_{\max}}{\|A\|^2}$. Moreover, we compared our method with DPDA-S by setting its constant step-sizes to DPDA's initial steps-sizes, in terms of relative error $\left(\max_{i\in\mathcal{N}} \left\|\bar{x}_i^k - x^*\right\| / \|x^*\|\right)$, and infeasibilty $\left(\max_{i\in\mathcal{N}} d_{\mathcal{K}_i}(A\bar{x}_i^k)\right)$. As it can be seen in Fig. 2, our method converges faster than DPDA-S in both statistics.

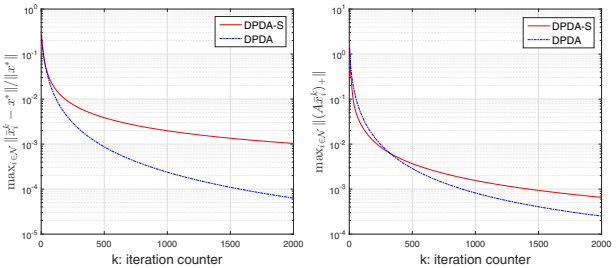

Fig. 2: Comparison of DPDA and DPDA-S

4. REFERENCES

[1] Joel B Predd, SB Kulkarni, and H Vincent Poor, "Distributed learning in wireless sensor networks," *IEEE Signal Processing Magazine*, vol. 23, no. 4, pp. 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*, vol. 56, no. 1, pp. 350–364, 2008.

[3] Ke Zhou and Stergios I Roumeliotis, "Multirobot active target tracking with combinations of relative observations," *IEEE Transactions on Robotics*, vol. 27, no. 4, pp. 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*, vol. 9, no. 1, pp. 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*, vol. 58, no. 3, pp. 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*, vol. 59, no. 10, pp. 4648–4663, 2011.

[7] John C Duchi, Alekh Agarwal, and Martin J Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Transactions on Automatic control*, vol. 57, no. 3, pp. 592–606, 2012.

[8] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[9] 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)*, 2011, pp. 713–720.

[10] 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*. IEEE, 2011, pp. 1–6.

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

[12] Jianshu Chen and Ali H Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4289–4305, 2012.

[13] Reza Olfati-Saber, J Alex Fax, and Richard M Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.

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

[15] Benjamin Hofner, Thomas Kneib, and Torsten Hothorn, "A unified framework of constrained regression," *Statistics and Computing*, vol. 26, no. 1-2, pp. 1–14, 2016.

[16] Gareth M James, Courtney Paulson, and Paat Rusmevichientong, "Penalized and constrained regression," Tech. Rep., Technical report, 2013. 15, 2013.

[17] Antonin Chambolle and Thomas Pock, "On the ergodic convergence rates of a first-order primal–dual algorithm," *Mathematical Programming*, vol. 159, no. 1, pp. 253–287, 2016.

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

[19] Tsung-Hui Chang, Angelia Nedic, and Anna Scaglione, "Distributed constrained optimization by consensus-based primal-dual perturbation method," *Automatic Control, IEEE Transactions on*, vol. 59, no. 6, pp. 1524–1538, 2014.

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

[21] Wei Shi, Qing Ling, Gang Wu, and Wotao Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[22] Erfan Yazdandoost Hamedani and Necdet Serhat Aybat, "Multi-agent constrained optimization of a strongly convex function over time-varying directed networks," *Technical Report [Online]. Available at arXiv:1706.07907*, 2017.