A Regularized Fenchel Dual Gradient Method for Nonsmooth Optimization over Time-varying Networks

Xuyang Wu, Kin Cheong Sou, and Jie Lu

***Abstract*—In this paper, we develop a regularized Fenchel dual gradient method (RFDGM) for solving nonsmooth convex optimization problems over networks with time-varying topologies, where the nodes are required to find a common decision that minimizes the sum of their local objective functions subject to their local constraints in a fully distributed fashion. Different from most existing distributed optimization algorithms that also cope with time-varying networks, RFDGM is able to handle problems with general convex local objective functions and distinct local constraints, and still has non-asymptotic convergence results. Specifically, under a standard network connectivity condition, we show that RFDGM is guaranteed to reach ϵ-accuracy in both optimality and feasibility within $O(\frac{1}{\epsilon^2}\ln\frac{1}{\epsilon})$ iterations. Such an iteration complexity can be improved to $O(\frac{1}{\epsilon}\ln\frac{1}{\epsilon})$ if the local objective functions are strongly convex but not necessarily differentiable. Finally, simulation results demonstrate the high effectiveness of RFDGM.**

***Index Terms*—distributed optimization, regularization technique, Fenchel duality.**

I. Introduction

There is a recent surge of interest in distributed optimization, motivated by its broad applications such as in-network robust estimation [1], network resource allocation [2], and distributed machine learning [3]. In many of such scenarios, each node in a network is endowed with a local objective function and a local constraint, and all the nodes attempt to jointly find a common decision so that the sum of all the local objective functions is minimized over the intersection of all the local constraint sets. Also, the network is often highly dynamic and of huge size. Accordingly, the nodes can only afford interactions with neighbors and need to cope with time-varying network topologies.

There have been a number of distributed optimization algorithms that are applicable to time-varying networks (e.g., [4]–[17]). Nevertheless, most of these algorithms [4]–[11] only address unconstrained problems or problems with identical local constraints, except [12]–[17] where both *time-varying* networks and *nonidentical* local constraints are allowed.

The consensus-based projected subgradient method [12] and the proximal minimization algorithm [13] both converge for general convex local objective functions. However, [12], [13] only establish asymptotic convergence and do not provide non-asymptotic convergence results such as convergence rates and iteration complexities. DPDA-D [14] and DPDA-TV [15] approximate a centralized primal-dual algorithm by executing a growing number of consensus operations at each iteration k. It is shown that when the constraint sets are bounded, DPDA-D converges with an $O(1/k)$ rate for general convex objective functions, and DPDA-TV converges with an $O(1/k^2)$ rate for strongly convex objective functions. However, the number of the consensus operations carried out by DPDA-D and DPDA-TV per iteration *grows unbounded*, so that DPDA-D and DPDA-TV suffer from high communication costs and, also, their practical performance is not satisfactory [18]. The Dykstra algorithm [16] converges at an $O(1/\sqrt{k})$ rate for strongly convex problems, yet it requires neighboring nodes to intermittently solve an optimization problem that consists of their local objective functions, which can be communication-wise costly and may cause privacy issues.

In our prior work [17], we have developed, based on Fenchel duality [19], a Fenchel dual gradient method (FDGM) that addresses distributed optimization with strongly convex local objectives and distinct local constraints on time-varying networks. FDGM achieves an $O(1/\sqrt{k})$ convergence rate, is highly scalable with respect to the network size and a connectivity parameter, and enjoys fast convergence in practice. Nevertheless, the implementation and analysis of FDGM rely on *strong convexity* of the objective functions, which results in a smooth Fenchel dual problem to be solved by FDGM.

In this paper, we propose a distributed optimization algorithm, referred to as *Regularized Fenchel Dual Gradient Method* (RFDGM), for solving problems with *general convex* local objectives and *nonidentical* local constraints over *time-varying* undirected networks. Since strong convexity is no longer assumed, the Fenchel dual problem of the distributed optimization problem can be non-differentiable. Thus, instead of directly handling the Fenchel dual problem as FDGM does, we introduce a regularization technique to obtain a regularized Fenchel dual problem that is strongly convex and smooth. We show that the gradient of the regularized Fenchel dual function is readily available and can be evaluated by the nodes in parallel. This allows us to construct a weighted gradient method to solve the regularized Fenchel dual problem, which, with properly designed parameters, only requires each node to communicate with its current neighbors. The resulting algorithm, i.e., RFDGM, is characterized by the following:

1) RFDGM *linearly* converges to the optimal value of the regularized Fenchel dual problem on time-varying networks satisfying the standard B-connectivity condition.
2) We provide iteration complexities for RFDGM, which can be explicitly estimated. In particular, for any given $\epsilon > 0$, RFDGM is guaranteed to reach ϵ-accuracy in both optimality and feasibility of the *original primal* problem within $O(\frac{1}{\epsilon^2}\ln\frac{1}{\epsilon})$ iterations. If we further impose strong convexity (but no smoothness) on the local

X. Wu and J. Lu are with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. Email: {wuxy, lujie}@shanghaitech.edu.cn. K. C. Sou is with the Department of Electrical Engineering, National Sun Yat-sen University, Taiwan. Email: sou12@mail.nsysu.edu.tw.

This work has been supported by the National Natural Science Foundation of China under grant 61603254.

objective functions, the iteration complexity is reduced to $O(\frac{1}{\epsilon}\ln\frac{1}{\epsilon})$. This significantly improves the $O(\frac{1}{\epsilon^2})$ iteration complexity of FDGM [17] for solving strongly convex problems, and also eliminates an assumption in [17] which requires the global constraint set to have a nonempty interior.

3) RFDGM can be implemented in a fully distributed way, including parameter selections that ensure convergence.
4) We provide a detailed comparison of RFDGM and the existing distributed optimization methods [12]–[17] that are able to handle time-varying networks and nonidentical local constraints. Briefly, RFDGM does not rely on the strong convexity condition assumed by [15]–[17] and has stronger convergence results than [12], [13], [16], [17] under the same or less restrictive assumptions. Moreover, unlike [14], [15] whose communication costs per iteration are increasing and eventually blow up, each iteration of RFDGM only requires every node to interact with its neighbors once.
5) We demonstrate the superior performance of RFDGM against the alternative methods via simulations on a class of constrained convex optimization problems.

The outline of the paper is as follows: Section II describes the problem formulation. Section III develops RFDGM and Section IV shows the convergence analysis. Section V discusses the parameter selections, and Section VI presents the simulations. Finally, Section VII concludes the paper.

A preliminary conference version of this paper can be found in [20]. Compared to [20], RFDGM in this paper has a more general algorithmic form, which allows local regularization parameters instead of global ones. Accordingly, the convergence analysis of this paper generalizes that in [20]. This paper also contains new convergence results for strongly convex problems and all the proofs including those omitted in [20]. Moreover, this paper includes detailed comparisons between RFDGM and the existing works, as well as an expanded numerical study.

Notation

For any set $X \subseteq \mathbb{R}^d$, $\operatorname{int} X$ is its interior, $\operatorname{rel\,int} X$ is its relative interior, and $|X|$ is its cardinality. For any two sets $X, Y \subseteq \mathbb{R}^d$, we use $X \times Y$ to denote their Cartesian product. We use $\|\cdot\|$ to denote the Euclidean norm and $\|\cdot\|_1$ the ℓ_1 norm. In addition, I_d is the $d \times d$ identity matrix, $\mathbf{O}_d$ is the $d \times d$ all-zero matrix, $\mathbf{1}_d \in \mathbb{R}^d$ is the all-one vector, $\mathbf{0}_d \in \mathbb{R}^d$ is the all-zero vector, and $\otimes$ is the Kronecker product. We also use $\operatorname{diag}(\alpha_1, \ldots, \alpha_n)$ to represent a diagonal matrix with diagonal entries $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. Also, for any matrix $A \in \mathbb{R}^{n\times n}$, $[A]_{ij}$ is its (i,j)-entry and $\operatorname{Null}(A)$ is its null space. For any two matrices $A, B \in \mathbb{R}^{d\times d}$, $A - B \succeq \mathbf{O}_d$ and $B - A \preceq \mathbf{O}_d$ both mean $A - B$ is positive semidefinite. For any symmetric positive semidefinite matrix $A \in \mathbb{R}^{n\times n}$, we use $\lambda_i^{\downarrow}(A)$ to denote the ith largest eigenvalue of A, and let $\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^T A \mathbf{x}}$ for any $\mathbf{x} \in \mathbb{R}^n$. Given $x_1, \ldots, x_n \in \mathbb{R}^d$, $\mathbf{x} = ((x_1)^T, \ldots, (x_n)^T)^T \in \mathbb{R}^{nd}$ represents the vector obtained by stacking $x_1, \ldots, x_n$.

For an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with the vertex set $\mathcal{V}$ and the edge set $\mathcal{E} \subseteq \{\{i,j\} : i, j \in \mathcal{V}, i \neq j\}$, its Laplacian matrix, denoted by $L_{\mathcal{G}}$, is defined as

$$[L_{\mathcal{G}}]_{ij} = \begin{cases} |\mathcal{N}_i|, & \text{if } i = j, \\ -1, & \text{if } \{i,j\} \in \mathcal{E}, \\ 0, & \text{otherwise,} \end{cases} \quad \forall i, j \in \mathcal{V},$$

where $\mathcal{N}_i = \{j \in \mathcal{V} : \{i,j\} \in \mathcal{E}\}$.

For any function $f : \mathbb{R}^d \to \mathbb{R}$, $\partial f(x)$ denotes the subdifferential (i.e., the set of subgradients) of f at $x \in \mathbb{R}^d$. If f is differentiable, $\partial f(x) = \{\nabla f(x)\}$, where $\nabla f(x)$ is the gradient of f at x.

Preliminaries

A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be *convex* on a convex set $X \subseteq \mathbb{R}^d$ if for any $x, y \in X$ and any $s \in \partial f(x)$, $f(y) - f(x) - s^T(y - x) \geq \frac{\theta}{2}\|y - x\|^2$ for some $\theta \geq 0$, where θ is called the *convexity parameter* of f on X.

We say f is *θ-strongly convex* on X if the above inequality holds for some $\theta > 0$, and f is *(globally) strongly convex* if $\theta > 0$ and $X = \mathbb{R}^d$. Note that strong convexity of f does not require f to be differentiable. If f is convex but not strongly convex on X, then $\theta = 0$.

We say f is *M-smooth* if it is differentiable and its gradient ∇f satisfies the Lipschitz condition $\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\|$ $\forall x, y \in \mathbb{R}^d$.

II. Problem Formulation

Consider a multi-hop network consisting of $n \geq 2$ nodes, which can be modeled as a time-varying undirected graph $\mathcal{G}^k = (\mathcal{V}, \mathcal{E}^k)$, where $\mathcal{V} = \{1, 2, \ldots, n\}$ is the set of nodes and $\mathcal{E}^k \subseteq \{\{i,j\} : i, j \in \mathcal{V}, i \neq j\}$ represents the set of links at time $k \geq 0$. We assume $\mathcal{E}^k \neq \emptyset$ $\forall k \geq 0$. Also suppose the following B-connectivity assumption holds, which says that the links occurring in every time interval $[kB, (k+1)B - 1]$, $k \geq 0$ are able to connect all the n nodes.

Assumption 1. *There exists an integer $B > 0$ such that for any $k \geq 0$, the graph $(\mathcal{V}, \bigcup_{t=kB}^{(k+1)B-1} \mathcal{E}^t)$ is connected.*

Assumption 1 does not require every $\mathcal{G}^k$, $k \geq 0$ to be connected. It is a standard connectivity condition on time-varying networks, and is also considered in many existing works such as [5]–[17].

Suppose each node $i \in \mathcal{V}$ in the network has a local objective function $f_i : \mathbb{R}^d \to \mathbb{R}$ and a local constraint set $X_i \subseteq \mathbb{R}^d$. All the nodes attempt to jointly solve the following constrained optimization problem:

$$\begin{array}{ll} \underset{x\in\mathbb{R}^d}{\text{minimize}} & \sum_{i\in\mathcal{V}} f_i(x) \\ \text{subject to} & x \in \bigcap_{i\in\mathcal{V}} X_i. \end{array} \tag{1}$$

Let each node $i \in \mathcal{V}$ hold a local copy $x_i \in \mathbb{R}^d$ of the global decision variable x. Define $\mathbf{x} = ((x_1)^T, \ldots, (x_n)^T)^T \in \mathbb{R}^{nd}$ and $S = \{\mathbf{x} \in \mathbb{R}^{nd} : x_1 = x_2 = \cdots = x_n\}$. Problem (1) can thus be reformulated as

$$\begin{array}{ll} \underset{\mathbf{x}\in\mathbb{R}^{nd}}{\text{minimize}} & F(\mathbf{x}) := \sum_{i\in\mathcal{V}} f_i(x_i) \\ \text{subject to} & \mathbf{x} \in X := X_1 \times \cdots \times X_n, \\ & \mathbf{x} \in S. \end{array} \tag{2}$$

Clearly, problems (1) and (2) share the same optimal value $F^\star$. Moreover, $x^\star \in \mathbb{R}^d$ is an optimal solution of problem (1) if and only if $\mathbf{x}^\star = ((x^\star)^T, \ldots, (x^\star)^T)^T \in \mathbb{R}^{nd}$ is an optimal solution of problem (2).

We impose the following assumption on problem (1).

Assumption 2. *Problem* (1) *satisfies the following:*

(a) Each f_i, $i \in \mathcal{V}$ is continuous and convex on X_i.

(b) Each X_i, $i \in \mathcal{V}$ is closed and convex. In addition, $\operatorname{rel\,int} \bigcap_{i\in\mathcal{V}} X_i \neq \emptyset$.

Assumption 2(a) allows each f_i to be a general convex function that is not necessarily differentiable. The nonemptyness of $\operatorname{rel\,int} \bigcap_{i\in\mathcal{V}} X_i$ in Assumption 2(b) will be used to guarantee zero duality gap later. Besides, unlike the existing algorithms [12], [13], [17] that require each local constraint set X_i to have a nonempty interior, here each X_i may have an empty interior and therefore can be described by linear equality constraints.

Finally, we assume the existence of an optimal solution to problem (1).

Assumption 3. *The optimal set $\mathcal{X}^\star$ of problem* (1) *is nonempty.*

III. Algorithm Development

In this section, we design a distributed algorithm by means of a Fenchel dual approach and a regularization technique to solve problem (1) over time-varying networks.

A. Regularized Fenchel Dual Problem

The Fenchel dual problem [19] of (2) is given by

$$\begin{array}{ll} \underset{\mathbf{w}\in\mathbb{R}^{nd}}{\text{maximize}} & -D(\mathbf{w}) := -\sup_{\mathbf{x}\in X}(\mathbf{w}^T\mathbf{x} - F(\mathbf{x})) \\ \text{subject to} & \mathbf{w} \in S^\perp, \end{array} \tag{3}$$

where the Fenchel dual function D is the convex conjugate function of F and $S^\perp = \{\mathbf{w} = ((w_1)^T, \ldots, (w_n)^T)^T \in \mathbb{R}^{nd} : w_1 + w_2 + \cdots + w_n = \mathbf{0}_d\}$ is the orthogonal complement of S in problem (2). Assumption 2(b) guarantees that there is zero duality gap between problems (2) and (3), i.e., the optimal value $-D^\star$ of problem (3) is equal to $F^\star$, and that the optimal set $W^\star \subseteq S^\perp$ of (3) is nonempty [19].

In what follows we introduce some basic properties of problem (3). For each $i \in \mathcal{V}$, we let

$$G_i(w_i) = \arg\max_{x\in X_i}\langle w_i, x\rangle - f_i(x) \subseteq X_i.$$

Also let $G(\mathbf{w}) = G_1(w_1) \times \cdots \times G_n(w_n)$. For any given $\mathbf{w} \in \mathbb{R}^{nd}$, if $G(\mathbf{w}) \neq \emptyset$, then the subdifferential $\partial D(\mathbf{w})$ of D at $\mathbf{w}$ is exactly $G(\mathbf{w})$ [19]. Observe that D is differentiable at $\mathbf{w}$ if and only if $G(\mathbf{w})$ is a singleton, which can be guaranteed when, for example, each f_i is strictly convex and has bounded level sets.

In our previous work [17], under the assumption that each f_i is strongly convex on X_i, D is guaranteed to be smooth. Owing to this, a weighted gradient method can be applied to solve the smooth Fenchel dual problem, leading to a distributed Fenchel dual gradient method (FDGM). However, FDGM fails here because we no longer assume strong convexity of f_i and thus D may be non-differentiable. Furthermore, directly solving the nonsmooth problem (3) via other algorithms such as the subgradient methods often causes unsatisfactory convergence performance in practice.

To address this issue, we intend to derive an approximation of the Fenchel dual problem (3) which enjoys a more favorable structure. In [21], a double smoothing technique is proposed for regularizing both the primal problem and the *Lagrange* dual problem for linearly-constrained convex optimization problems, so that the resulting regularized Lagrange dual problem is strongly convex and smooth. Inspired by this, below we propose a *primal-Fenchel-dual regularization procedure* for regularizing the primal objective function F and the Fenchel dual function D, leading to a strongly convex and smooth Fenchel dual approximation, which, as will be demonstrated shortly, can be solved in a distributed fashion over time-varying networks.

Primal regularization: We first add a quadratic regularization term $\frac{1}{2}\|\mathbf{x}\|^2_{\Lambda_\gamma}$ to the primal objective function F, where $\Lambda_\gamma = \operatorname{diag}(\gamma_1, \ldots, \gamma_n) \otimes I_d$ and $\gamma_i \geq 0\ \forall i \in \mathcal{V}$. We require the resulting regularized primal objective function

$$F_\gamma(\mathbf{x}) := F(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}\|^2_{\Lambda_\gamma}$$

to be strongly convex on X, or equivalently,

$$\gamma_i + \theta_i > 0, \quad \forall i \in \mathcal{V}, \tag{4}$$

where $\theta_i \geq 0$ is the convexity parameter of f_i on X_i. This can be satisfied by choosing $\gamma_i > 0$ for f_i that is not strongly convex on X_i, i.e., $\theta_i = 0$. If f_i is already strongly convex on X_i, i.e., $\theta_i > 0$, then γ_i can take any nonnegative value including 0.

Subsequently, we consider the convex conjugate function of F_γ, which is given by

$$D_\gamma(\mathbf{w}) := \sup_{\mathbf{x}\in X}\ \langle \mathbf{w}, \mathbf{x}\rangle - F_\gamma(\mathbf{x}).$$

Since F_γ is strongly convex on X, D_γ is well-defined. Moreover, given any $\mathbf{w} \in \mathbb{R}^{nd}$, there uniquely exists

$$\tilde{\mathbf{x}}_\gamma(\mathbf{w}) := \arg\ \max_{\mathbf{x}\in X}\langle \mathbf{w}, \mathbf{x}\rangle - F_\gamma(\mathbf{x}).$$

Accordingly, $D_\gamma(\mathbf{w})$ is differentiable and

$$\nabla D_\gamma(\mathbf{w}) = \tilde{\mathbf{x}}_\gamma(\mathbf{w}).$$

In addition, since F_γ is $\min_{i\in\mathcal{V}}(\gamma_i + \theta_i)$-strongly convex on X, D is $\max_{i\in\mathcal{V}}(\frac{1}{\gamma_i+\theta_i})$-smooth [17, Corollary 1].

Fenchel dual regularization: We then regularize the Fenchel dual objective function D by adding another quadratic regularization term to D_γ, which yields

$$D_{\gamma,\kappa}(\mathbf{w}) := D_\gamma(\mathbf{w}) + \frac{1}{2}\|\mathbf{w}\|^2_{\Lambda_\kappa},$$

where $\Lambda_\kappa = \operatorname{diag}(\kappa_1, \ldots, \kappa_n) \otimes I_d$ and $\kappa_i > 0\ \forall i \in \mathcal{V}$. It follows that $D_{\gamma,\kappa}$ is $\underline{\kappa}$-strongly convex with $\underline{\kappa} := \min_{i\in\mathcal{V}} \kappa_i$ and $L_{\gamma,\kappa}$-smooth with

$$L_{\gamma,\kappa} := \max_{i\in\mathcal{V}}(\frac{1}{\gamma_i + \theta_i} + \kappa_i). \tag{5}$$

Additionally,

$$\nabla D_{\gamma,\kappa}(\mathbf{w}) = \nabla D_{\gamma}(\mathbf{w}) + \Lambda_{\kappa}\mathbf{w} = \tilde{\mathbf{x}}_{\gamma}(\mathbf{w}) + \Lambda_{\kappa}\mathbf{w}. \quad (6)$$

Based on the above primal-Fenchel-dual regularization procedure, we obtain the regularized Fenchel dual problem

$$\begin{array}{ll} \underset{\mathbf{w}\in\mathbb{R}^{nd}}{\text{maximize}} & -D_{\gamma,\kappa}(\mathbf{w}) \\ \text{subject to} & \mathbf{w} \in S^{\perp}, \end{array} \quad (7)$$

which consists of a strongly convex and smooth objective function and a linear equality constraint. Therefore, there exists a unique optimal solution $\mathbf{w}^{\star}_{\gamma,\kappa} \in S^{\perp}$ to problem (7). Moreover, $\mathbf{w} \in S^{\perp}$ satisfies $\nabla D_{\gamma,\kappa}(\mathbf{w}) \in S$ if and only if $\mathbf{w} = \mathbf{w}^{\star}_{\gamma,\kappa}$ [22, Lemma 3.1].

B. Regularized Fenchel Dual Gradient Method

To address the regularized Fenchel dual problem (7), we propose the following algorithm referred to as *Regularized Fenchel Dual Gradient Method* (RFDGM):

$$\begin{aligned} &\mathbf{w}^0 \in S^{\perp}, \\ &\mathbf{x}^k = \tilde{\mathbf{x}}_{\gamma}(\mathbf{w}^k), \quad \forall k \geq 0, \\ &\mathbf{w}^{k+1} = \mathbf{w}^k - \alpha^k (H_{\mathcal{G}^k} \otimes I_d)(\mathbf{x}^k + \Lambda_{\kappa}\mathbf{w}^k), \quad \forall k \geq 0, \end{aligned} \quad (8)$$

where $H_{\mathcal{G}^k} \in \mathbb{R}^{n\times n}$ is a weight matrix associated with the interaction graph $\mathcal{G}^k$ and $\alpha^k > 0$ is the step-size at time k. In the proposed RFDGM, i.e., (8), the first equation forces the initial dual iterate $\mathbf{w}^0$ to stay in the dual feasible set $S^{\perp}$, which, along with a properly designed $H_{\mathcal{G}^k}$, can ensure dual feasibility of all the subsequent dual iterates. The second equation describes the primal update and the third equation presents the dual update. From (6), the term $\mathbf{x}^k + \Lambda_{\kappa}\mathbf{w}^k$ in the dual update is exactly the gradient of the regularized Fenchel dual function $D_{\gamma,\kappa}$ at $\mathbf{w}^k$. Therefore, RFDGM (8) can be viewed as a weighted gradient method aimed at solving the regularized Fenchel dual problem (7).

To adapt (8) to time-varying networks, we impose the following conditions on the time-varying weight matrix $H_{\mathcal{G}^k} \in \mathbb{R}^{n\times n}$. For each $k \geq 0$, let $H_{\mathcal{G}^k}$ take the form of

$$[H_{\mathcal{G}^k}]_{ij} = \begin{cases} \sum_{s\in\mathcal{N}_i^k} h_{is}^k, & \text{if } i = j, \\ -h_{ij}^k, & \text{if } \{i,j\} \in \mathcal{E}^k, \\ 0, & \text{otherwise}, \end{cases} \quad \forall i,j \in \mathcal{V}, \quad (9)$$

where $\mathcal{N}_i^k = \{j \in \mathcal{V} : \{i,j\} \in \mathcal{E}^k\}$ is the set of neighbors of node i at time k, and let $H_{\mathcal{G}^k}$ satisfy the assumption below.

Assumption 4. *For each $k \geq 0$, $h_{ij}^k = h_{ji}^k > 0$ $\forall \{i,j\} \in \mathcal{E}^k$. In addition,*

$$\underline{h} := \inf_{k\geq 0} \min_{\{i,j\}\in\mathcal{E}^k} h_{ij}^k > 0,$$
$$\bar{h} := \sup_{k\geq 0} \max_{\{i,j\}\in\mathcal{E}^k} h_{ij}^k < \infty.$$

The above conditions on $H_{\mathcal{G}^k}$ suggest that $H_{\mathcal{G}^k}$ is uniformly bounded and symmetric with $\mathbf{1}_n$ in the null space. Also, since $H_{\mathcal{G}^k}$ is diagonally dominant, we have $H_{\mathcal{G}^k} \succeq \mathbf{O}_n$. Moreover, it can be shown that as long as $\mathbf{w}^0 \in S^{\perp}$, every subsequent dual iterate $\mathbf{w}^k$, $k \geq 1$ remains in $S^{\perp}$ and thus is dual feasible [17, Proposition 3].

Finally, we elaborate how RFDGM (8) can be implemented over the time-varying graph $\mathcal{G}^k$ in a distributed fashion. To do so, we evenly partition $\mathbf{x}^k$ and $\mathbf{w}^k$ into n blocks, i.e., $\mathbf{x}^k = ((x_1^k)^T, \ldots, (x_n^k)^T)^T$ and $\mathbf{w}^k = ((w_1^k)^T, \ldots, (w_n^k)^T)^T$. This allows us to rewrite the primal and dual updates in (8) as

$$\begin{aligned} x_i^k &= \arg\ \max\nolimits_{x\in X_i} \langle w_i^k, x\rangle - f_i(x) - \frac{\gamma_i}{2}\|x\|^2, \\ w_i^{k+1} &= w_i^k - \alpha^k \sum_{j\in\mathcal{N}_i^k} h_{ij}^k (x_i^k - x_j^k + \kappa_i w_i^k - \kappa_j w_j^k), \end{aligned}$$

for all $k \geq 0$ and $i \in \mathcal{V}$. Then, we associate each $x_i^k \in X_i$ and $w_i^k \in \mathbb{R}^d$ with node i. To satisfy $\mathbf{w}^0 \in S^{\perp}$, each node i may simply set $w_i^0 = \mathbf{0}_d$. At each $k \geq 0$, every node i with at least one neighbor updates its primal variable x_i^k and dual variable w_i^k by requesting $x_j^k + \kappa_j w_j^k$ from each neighbor $j \in \mathcal{N}_i^k$. Algorithm 1 below details all these actions taken by the nodes.

Algorithm 1 RFDGM

1: **Initialization:** Each node $i \in \mathcal{V}$ selects $w_i^0 \in \mathbb{R}^d$ so that $\mathbf{w}^0 \in S^{\perp}$ (or simply let $w_i^0 = \mathbf{0}_d$) and sets $x_i^0 = \arg\max_{x\in X_i} \langle w_i^0, x\rangle - f_i(x) - \frac{\gamma_i}{2}\|x\|^2$.
2: **for** $k = 0, 1, \ldots$ **do**
3: Each node $i \in \mathcal{V}$ with $\mathcal{N}_i^k \neq \emptyset$ sends its $x_i^k + \kappa_i w_i^k$ to all $j \in \mathcal{N}_i^k$.
4: Upon receiving $x_j^k + \kappa_j w_j^k$ $\forall j \in \mathcal{N}_i^k$, each node $i \in \mathcal{V}$ with $\mathcal{N}_i^k \neq \emptyset$ updates $w_i^{k+1} = w_i^k - \alpha^k \sum_{j\in\mathcal{N}_i^k} h_{ij}^k (x_i^k - x_j^k + \kappa_i w_i^k - \kappa_j w_j^k)$.
5: Each node $i \in \mathcal{V}$ with $\mathcal{N}_i^k \neq \emptyset$ computes $x_i^{k+1} = \arg\max_{x\in X_i} (w_i^{k+1})^T x - f_i(x) - \frac{\gamma_i}{2}\|x\|^2$.
6: Each node $i \in \mathcal{V}$ with $\mathcal{N}_i^k = \emptyset$ takes no action, i.e., $w_i^{k+1} = w_i^k$ and $x_i^{k+1} = x_i^k$.
7: **end for**

Before executing Algorithm 1, each node $i \in \mathcal{V}$ needs to determine the values of the regularization parameters γ_i and κ_i. Subsequently at each iteration k, every pair of neighboring nodes i and j need to agree on the weight $h_{ij}^k = h_{ji}^k$ associated with the link $\{i,j\} \in \mathcal{E}^k$. Moreover, all the nodes with a nonempty neighborhood should use the same step-size α^k. Later in Section V, we will discuss how to determine, in a distributed way, the values of these parameters that guarantee RFDGM to converge.

IV. Convergence Analysis

In this section, we investigate the convergence behavior of RFDGM and provide its iteration complexity in achieving any given accuracy in primal optimality, dual optimality, and primal feasibility. Recall from Section III-B that dual feasibility is constantly guaranteed by the initial condition of RFDGM and the structure of $H_{\mathcal{G}^k}$.

A. Linear Convergence Rate of RFDGM

We first show that RFDGM reaches the optimal value of the regularized problem (7) at a linear rate.

To present the result, we let $\mathcal{T}^k$, $k \geq 0$ be an arbitrary *connected* spanning subgraph of $(\mathcal{V}, \bigcup_{t=kB}^{(k+1)B-1} \mathcal{E}^t)$, which always exists due to Assumption 1. As a result, the algebraic connectivity $\lambda^k := \lambda_{n-1}^{\downarrow}(L_{\mathcal{T}^k})$ of $\mathcal{T}^k$, i.e., the second smallest eigenvalue of the Laplacian matrix $L_{\mathcal{T}^k}$ of $\mathcal{T}^k$, is positive [23], and hence $0 < \underline{\lambda}_T := \inf_{k\geq 0} \lambda^k \leq n$. Also, let ϖ^k be the maximum degree of $\mathcal{T}^k$ and let $\bar{\varpi} = \sup_{k\geq 0} \varpi^k \in [1, n-1]$. In addition, let $\delta > 0$ be such that for any $k \geq 0$,

$$H_{\mathcal{G}^k} \preceq \delta \left(\operatorname{diag}\left(\frac{1}{\gamma_1+\theta_1} + \kappa_1, \ldots, \frac{1}{\gamma_n+\theta_n} + \kappa_n \right) \right)^{-1}. \tag{10}$$

As $\frac{1}{\gamma_i+\theta_i} + \kappa_i > 0 \ \forall i \in \mathcal{V}$, the existence of δ is guaranteed.

Based on the above, we establish the linear convergence rate of RFDGM in the following theorem.

Theorem 1. *Suppose Assumptions 1, 2, 4 and* (4) *hold. Let* $(\mathbf{w}^k)_{k=0}^{\infty}$ *be the dual iterates generated by RFDGM. Let* $[\underline{c}, \bar{c}] \subsetneq (0, 2)$ *and suppose*

$$\alpha^k \in [\underline{c}/\delta, \bar{c}/\delta], \ \forall k \geq 0, \tag{11}$$

with $\delta > 0$ *satisfying* (10)*. Then, for each* $k \geq 0$*,*

$$\begin{aligned} &D_{\gamma,\kappa}(\mathbf{w}^k) - D_{\gamma,\kappa}^{\star} \\ &\leq \left(1 - \frac{2\underline{\kappa}\rho\underline{\lambda}_T}{\eta}\right)^{\lfloor k/B \rfloor} (D_{\gamma,\kappa}(\mathbf{w}^0) - D_{\gamma,\kappa}^{\star}), \end{aligned} \tag{12}$$

where $\rho = \min\{\frac{\underline{c}}{\delta} - \frac{\underline{c}^2}{2\delta}, \frac{\bar{c}}{\delta} - \frac{\bar{c}^2}{2\delta}\} \in (0, \infty)$*,* $\eta = 3B\bar{\varpi}L_{\gamma,\kappa}\bar{c}^2/\delta + 3/\underline{h} \in (0, \infty)$*,* $L_{\gamma,\kappa}$ *is given in* (5)*, and* $-D_{\gamma,\kappa}^{\star}$ *is the optimal value of problem* (7)*.*

Proof. See Appendix A. □

In [24], a distributed weighted gradient method for addressing linearly-constrained convex optimization problems in the form of (7) is proposed, which has a *time-invariant* weight matrix and thus can only handle *fixed* networks. Also, [24] derives a linear convergence rate under strong convexity and smoothness of the objective functions. Recall from Section III-B that RFDGM can be viewed as a weighted gradient method with a *time-varying* weight matrix applied to (7). Thus, Theorem 1 extends the linear convergence result in [24] for static graphs to time-varying graphs.

B. Regularization Error Analysis

Although RFDGM has been proved to solve the regularized Fenchel dual problem (7) at a linear rate, it is unknown how well it addresses the original Fenchel dual problem (3) and the primal problem (2). In this subsection, we express errors with respect to the original Fenchel dual optimality, primal feasibility, and primal optimality in terms of an error with respect to the optimality of the regularized Fenchel dual problem (7).

To this end, we select the regularization parameters as follows: Given an arbitrary $\epsilon > 0$, let

$$\gamma_i = \frac{\epsilon}{4nD_i^X}, \quad \kappa_i = \frac{\epsilon}{2(D_i^W)^2}, \tag{13}$$

$\bar{D}^X$	$\max_{i\in\mathcal{V}} D_i^X \in (0,\infty) \cup \{\infty\}$ where D_i^X satisfies (14) and is finite if X_i is bounded
$\bar{D}^W$	$\max_{i\in\mathcal{V}} D_i^W \in (0,\infty)$
$\underline{D}^W$	$\min_{i\in\mathcal{V}} D_i^W \in (0,\infty)$, where D_i^W satisfies (14)
$L_{\gamma,\kappa}$	$\max_{i\in\mathcal{V}}(\frac{1}{\gamma_i+\theta_i} + \kappa_i) \in (0,\infty)$ due to (4)
$\bar{h}$	$\sup_{k\geq 0} \max_{\{i,j\}\in\mathcal{E}^k} h_{ij}^k \in (0,\infty)$
$\underline{h}$	$\inf_{k\geq 0} \min_{\{i,j\}\in\mathcal{E}^k} h_{ij}^k \in (0,\infty)$
$\bar{\varpi}$	$\sup_{k\geq 0} \varpi^k \in [1, n-1]$ where ϖ^k is the maximum degree of $\mathcal{T}^k$
$\bar{\lambda}$	any upper bound of $\sup_{k\geq 0} \lambda_1^{\downarrow}(L_{\mathcal{G}^k})$ one option of $\bar{\lambda}$ is n
$\underline{\lambda}_T$	$\inf_{k\geq 0} \lambda_{n-1}^{\downarrow}(L_{\mathcal{T}^k}) \in (0, n]$
$\underline{\theta}$	$\min_{i\in\mathcal{V}} \theta_i$

TABLE I
Constants in convergence results

where $D_i^X \in (0,\infty) \cup \{\infty\}$ and $D_i^W \in (0,\infty)$ satisfy

$$D_i^X \geq \sup_{x\in X_i} \frac{\|x\|^2}{2}, \ D_i^W \geq \min_{\mathbf{w}^\star \in W^\star} \|\mathbf{w}^\star\|. \tag{14}$$

Note that if X_i is unbounded, D_i^X has to be ∞ which forces $\gamma_i = 0$. Otherwise, we choose D_i^X to be finite which yields a positive and finite γ_i. Consequently, the condition (4) required by RFDGM can be guaranteed when X_i is bounded for non-strongly convex f_i (i.e., $\theta_i = 0$). In addition, since the optimal set $W^\star$ of the Fenchel dual problem (3) is nonempty [19], we can always find a positive and finite D_i^W, so that κ_i is also positive and finite. We will discuss the ways of determining D_i^W without knowing $W^\star$ later in Section V.

Below, we express the optimality error $D(\mathbf{w}) - D^\star$ with respect to the original Fenchel dual problem (3) as well as the feasibility error $\|P_{S^\perp}(\tilde{\mathbf{x}}_\gamma(\mathbf{w}))\|$ and the optimality error $|F(\tilde{\mathbf{x}}_\gamma(\mathbf{w})) - F^\star|$ with respect to the primal problem (2) in terms of the optimality error $D_{\gamma,\kappa}(\mathbf{w}) - D_{\gamma,\kappa}^\star$ with respect to the regularized Fenchel dual problem (7).

Lemma 1. *Suppose Assumptions 2–3 and* (4)*,* (13)*,* (14) *hold. For any* $\mathbf{w} \in S^\perp$*,*

$$D(\mathbf{w}) - D^\star \leq D_{\gamma,\kappa}(\mathbf{w}) - D_{\gamma,\kappa}^\star + \frac{\epsilon}{2}, \tag{15}$$

$$\begin{aligned} \|P_{S^\perp}(\tilde{\mathbf{x}}_\gamma(\mathbf{w}))\| &\leq \frac{\epsilon}{\sqrt{2}\underline{D}^W} \\ &+ \left(\sqrt{2L_{\gamma,\kappa}} + \frac{\sqrt{\epsilon}}{\underline{D}^W}\right) \sqrt{D_{\gamma,\kappa}(\mathbf{w}) - D_{\gamma,\kappa}^\star}, \end{aligned} \tag{16}$$

$$\begin{aligned} F(\tilde{\mathbf{x}}_\gamma(\mathbf{w})) - F^\star &\leq \left(\frac{\bar{D}^W}{\underline{D}^W} + \frac{1}{4}\right)\epsilon \\ &+ 2\left(\bar{D}^W\sqrt{\frac{2L_{\gamma,\kappa}}{\epsilon}} + \frac{\bar{D}^W}{\underline{D}^W}\right)(D_{\gamma,\kappa}(\mathbf{w}) - D_{\gamma,\kappa}^\star) \\ &+ 2\left(\bar{D}^W\sqrt{L_{\gamma,\kappa}} + \sqrt{2\epsilon}\frac{\bar{D}^W}{\underline{D}^W}\right)\sqrt{D_{\gamma,\kappa}(\mathbf{w}) - D_{\gamma,\kappa}^\star}, \end{aligned} \tag{17}$$

$$\begin{aligned} F(\tilde{\mathbf{x}}_\gamma(\mathbf{w})) - F^\star &\geq -\left(\frac{\bar{D}^W}{\underline{D}^W} + \frac{3}{4}\right)\epsilon \\ &- 2\left(\bar{D}^W\sqrt{\frac{2L_{\gamma,\kappa}}{\epsilon}} + \frac{\bar{D}^W}{\underline{D}^W} + \frac{1}{2}\right)(D_{\gamma,\kappa}(\mathbf{w}) - D_{\gamma,\kappa}^\star) \end{aligned}$$

$$-2\left(\bar{D}^W\sqrt{L_{\gamma,\kappa}}+\sqrt{2\epsilon}\frac{\bar{D}^W}{\underline{D}^W}\right)\sqrt{D_{\gamma,\kappa}(\mathbf{w})-D^\star_{\gamma,\kappa}}, \quad (18)$$

where the constants are defined in Table I.

Proof. See Appendix B. □

Combining Theorem 1 and Lemma 1 results in the following corollary, which says that the dual optimality error, primal optimality error, and primal feasibility error generated by RFDGM all converge to $O(\epsilon)$.

Corollary 1. *Suppose all the conditions in Theorem 1 and Lemma 1 hold. Let $(\mathbf{w}^k)_{k=0}^\infty$ and $(\mathbf{x}^k)_{k=0}^\infty$ be the dual and primal iterates generated by RFDGM. Then,*

$$\lim_{k\to\infty} D(\mathbf{w}^k)-D^\star \le \frac{\epsilon}{2},$$
$$\lim_{k\to\infty}\|P_{S^\perp}(\mathbf{x}^k)\| \le \frac{\epsilon}{\sqrt{2}\underline{D}^W},$$
$$-\left(\frac{\bar{D}^W}{\underline{D}^W}+\frac{3}{4}\right)\epsilon \le \lim_{k\to\infty} F(\mathbf{x}^k)-F^\star \le \left(\frac{\bar{D}^W}{\underline{D}^W}+\frac{1}{4}\right)\epsilon,$$

where the constants are defined in Table I.

C. Iteration Complexity: General Convex Objective Functions

This subsection analyzes the iteration complexity of RFDGM, i.e., the number of iterations needed to achieve $O(\epsilon)$-accuracy, in solving problem (1) with general convex f_i's whose convexity parameters θ_i's are considered as 0.

To make the regularization parameter selections (13)–(14) satisfy the condition (4) required by RFDGM, we assume that every local constraint set X_i is bounded, so that $D_i^X < \infty$ and thus $\gamma_i > 0$ for all $i \in \mathcal{V}$.

Assumption 5. *Each X_i, $i \in \mathcal{V}$ is bounded.*

Assumption 5, together with Assumption 2, indicates that each X_i, $i \in \mathcal{V}$ is convex and compact. Note that it is not rare to assume compactness of X_i $\forall i \in \mathcal{V}$ in the literature of distributed optimization with nonidentical local constraints (e.g., [12]–[15]). In addition, Assumption 5 suffices to guarantee Assumption 3.

Remark 1. *Assumption 5 can be relaxed by allowing some X_i to be unbounded if the corresponding f_i is strongly convex on X_i with $\theta_i > 0$. In that case, $\gamma_i+\theta_i > 0$ by default and the iteration complexity below still holds (up to some constants).*

Theorem 2. *Suppose Assumptions 1, 2, 4, 5 and* (13)–(14) *hold. Let $[\underline{c},\bar{c}] \subsetneq (0,2)$, $\delta \in (0, L_{\gamma,\kappa}\bar{h}\bar{\lambda}]$ satisfy* (10) *(which always exists), and* (11) *hold. Let $(\mathbf{w}^k)_{k=0}^\infty$ and $(\mathbf{x}^k)_{k=0}^\infty$ be the dual and primal iterates generated by RFDGM. For any $\sigma > 0$, define*

$$K_{\epsilon,\sigma} = \frac{3B(\bar{D}^W)^2(B\bar{\varpi}\bar{c}^2+\bar{\lambda}\bar{h}/\underline{h})}{2\underline{\lambda}_T\max\{\underline{c}-\underline{c}^2/2,\bar{c}-\bar{c}^2/2\}}\left(\frac{1}{(\underline{D}^W)^2}+\frac{8n\bar{D}^X}{\epsilon^2}\right) \times \ln\frac{D(\mathbf{w}^0)-D^\star+\frac{\epsilon}{4}(1+\frac{\|\mathbf{w}^0\|^2}{(\underline{D}^W)^2})}{\sigma}. \quad (19)$$

Then, for any $k \ge K_{\epsilon,\sigma}$,

$$D(\mathbf{w}^k)-D^\star \le \sigma+\frac{\epsilon}{2}, \quad (20)$$

$$\|P_{S^\perp}(\mathbf{x}^k)\| \le \left(\sqrt{\frac{8n\bar{D}^X}{\epsilon}+\frac{\epsilon}{(\underline{D}^W)^2}}+\frac{\sqrt{\epsilon}}{\underline{D}^W}\right)\sqrt{\sigma} + \frac{\epsilon}{\sqrt{2}\underline{D}^W}, \quad (21)$$

$$F(\mathbf{x}^k)-F^\star \le \left(\frac{\bar{D}^W}{\underline{D}^W}+\frac{1}{4}\right)\epsilon + 2\bar{D}^W\left(\sqrt{\frac{8n\bar{D}^X}{\epsilon^2}+\frac{1}{(\underline{D}^W)^2}}+\frac{1}{\underline{D}^W}\right)\sigma + 2\bar{D}^W\left(\sqrt{\frac{4n\bar{D}^X}{\epsilon}+\frac{\epsilon}{2(\underline{D}^W)^2}}+\frac{\sqrt{2\epsilon}}{\underline{D}^W}\right)\sqrt{\sigma}, \quad (22)$$

$$F(\mathbf{x}^k)-F^\star \ge -\left(\frac{\bar{D}^W}{\underline{D}^W}+\frac{3}{4}\right)\epsilon - 2\bar{D}^W\left(\sqrt{\frac{8n\bar{D}^X}{\epsilon^2}+\frac{1}{(\underline{D}^W)^2}}+\frac{1}{\underline{D}^W}+\frac{1}{2\bar{D}^W}\right)\sigma - 2\bar{D}^W\left(\sqrt{\frac{4n\bar{D}^X}{\epsilon}+\frac{\epsilon}{2(\underline{D}^W)^2}}+\frac{\sqrt{2\epsilon}}{\underline{D}^W}\right)\sqrt{\sigma}, \quad (23)$$

where the constants are defined in Table I. In particular, by choosing $\sigma = O(\epsilon^3)$, it is guaranteed that $D(\mathbf{w}^k)-D^\star \le O(\epsilon)$, $\|P_{S^\perp}(\mathbf{x}^k)\| \le O(\epsilon)$, and $|F(\mathbf{x}^k)-F^\star| \le O(\epsilon)$ for any $k \ge K_{\epsilon,\sigma} = O(\frac{1}{\epsilon^2}\ln\frac{1}{\epsilon})$.

Proof. See Appendix C. □

Theorem 2 states that for arbitrarily small $\epsilon > 0$, RFDGM yields errors of order $O(\epsilon)$ in dual optimality, primal optimality, and primal feasibility within $O(\frac{1}{\epsilon^2}\ln\frac{1}{\epsilon})$ iterations.

Next, we provide an upper bound on $K_{\epsilon,\sigma}$, which gives a computable estimate on the number of iterations required to achieve $O(\epsilon)$-accuracy in optimality and feasibility as in (20)-(23). Assume that B, n, $\bar{h}$, $\underline{h}$, and $\bar{\lambda}$ are known *a priori*[1]. Then, since the algebraic connectivity of any connected graphs is bounded below by $4/(n(n-1))$ [25], we have $\underline{\lambda}_T \ge 4/(n(n-1))$. Also, we have $\bar{\varpi} \le n$ and $-D^\star = F^\star \le F(\hat{\mathbf{x}})$ for any $\hat{\mathbf{x}} \in X \cap S$. Therefore,

$$K_{\epsilon,\sigma} \le \frac{3Bn^2(\bar{D}^W)^2(Bn\bar{c}^2+\bar{\lambda}\bar{h}/\underline{h})}{8\max\{\underline{c}-\underline{c}^2/2,\bar{c}-\bar{c}^2/2\}}\left(\frac{1}{(\underline{D}^W)^2}+\frac{8n\bar{D}^X}{\epsilon^2}\right) \times \ln\frac{D(\mathbf{w}^0)+F(\hat{\mathbf{x}})+\frac{\epsilon}{4}(1+\frac{\|\mathbf{w}^0\|^2}{(\underline{D}^W)^2})}{\sigma}.$$

Given ϵ and σ, the above upper bound on $K_{\epsilon,\sigma}$ can be explicitly evaluated provided that each D_i^X and D_i^W are known. We will illustrate ways of determining D_i^X and D_i^W in Section V.

D. Iteration Complexity: Strongly Convex Objective Functions

Here, we establish the iteration complexity of RFDGM in the case where the f_i's are strongly convex (but can be non-differentiable) and the X_i's may be unbounded.

Assumption 6. *Each f_i, $i \in \mathcal{V}$ is strongly convex on X_i, i.e., $\theta_i > 0$ $\forall i \in \mathcal{V}$.*

[1]We may simply choose $\bar{\lambda}$ to be n.

Assumption 6 is common for distributed optimization methods to establish non-asymptotic convergence results (e.g., [6], [8], [9], [15]–[17]). Also, Assumption 6 ensures (4) as well as the unique existence of an optimal solution to problem (1), so that Assumption 3 holds.

Compared to Section IV-C, although we additionally impose Assumption 6 in this subsection, we eliminate the compactness of the X_i's required by Theorem 2 and provide an improved iteration complexity of $O(\frac{1}{\epsilon}\ln\frac{1}{\epsilon})$ in the theorem below.

Theorem 3. *Suppose Assumptions 1, 2, 4, 6 and* (13)–(14) *hold. Let* $[\underline{c},\bar{c}] \subsetneq (0,2)$, $\delta \in (0, L_{\gamma,\kappa}\bar{h}\bar{\lambda}]$ *satisfy* (10)*, and* (11) *hold. Let* $(\mathbf{w}^k)_{k=0}^\infty$ *and* $(\mathbf{x}^k)_{k=0}^\infty$ *be the dual and primal iterates generated by RFDGM. For any* $\sigma > 0$*, define*

$$\tilde{K}_{\epsilon,\sigma} = \frac{3B(\bar{D}^W)^2(\frac{1}{\underline{\theta}} + \frac{\epsilon}{2(\underline{D}^W)^2})(B\bar{\varpi}\bar{c}^2 + \bar{\lambda}\bar{h}/\underline{h})}{\epsilon\underline{\lambda}_T \max\{\underline{c} - \underline{c}^2/2, \bar{c} - \bar{c}^2/2\}} \times \ln\frac{D(\mathbf{w}^0) - D^\star + \frac{\epsilon}{4}(1 + \frac{\|\mathbf{w}^0\|^2}{(\underline{D}^W)^2})}{\sigma}. \tag{24}$$

Then, for any $k \geq \tilde{K}_{\epsilon,\sigma}$,

$$D(\mathbf{w}^k) - D^\star \leq \sigma + \frac{\epsilon}{2}, \tag{25}$$

$$\|P_{S^\perp}(\mathbf{x}^k)\| \leq \frac{\epsilon}{\sqrt{2}\underline{D}^W} + \left(\sqrt{2\left(\frac{1}{\underline{\theta}} + \frac{\epsilon}{2(\underline{D}^W)^2}\right)} + \frac{\sqrt{\epsilon}}{\underline{D}^W}\right)\sqrt{\sigma}, \tag{26}$$

$$F(\mathbf{x}^k) - F^\star \leq \left(\frac{\bar{D}^W}{\underline{D}^W} + \frac{1}{4}\right)\epsilon + 2\left(\bar{D}^W\sqrt{\frac{2}{\epsilon}\left(\frac{1}{\underline{\theta}} + \frac{\epsilon}{2(\underline{D}^W)^2}\right)} + \frac{\bar{D}^W}{\underline{D}^W}\right)\sigma + 2\left(\bar{D}^W\sqrt{\frac{1}{\underline{\theta}} + \frac{\epsilon}{2(\underline{D}^W)^2}} + \sqrt{2\epsilon}\frac{\bar{D}^W}{\underline{D}^W}\right)\sqrt{\sigma}, \tag{27}$$

$$F(\mathbf{x}^k) - F^\star \geq -\left(\frac{\bar{D}^W}{\underline{D}^W} + \frac{3}{4}\right)\epsilon - 2\left(\bar{D}^W\sqrt{\frac{2}{\epsilon}\left(\frac{1}{\underline{\theta}} + \frac{\epsilon}{2(\underline{D}^W)^2}\right)} + \frac{\bar{D}^W}{\underline{D}^W} + \frac{1}{2}\right)\sigma - 2\left(\bar{D}^W\sqrt{\frac{1}{\underline{\theta}} + \frac{\epsilon}{2(\underline{D}^W)^2}} + \sqrt{2\epsilon}\frac{\bar{D}^W}{\underline{D}^W}\right)\sqrt{\sigma}, \tag{28}$$

where the constants are defined in Table I. In particular, by choosing $\sigma = O(\epsilon^2)$*, it is guaranteed that* $D(\mathbf{w}^k) - D^\star \leq O(\epsilon)$, $\|P_{S^\perp}(\mathbf{x}^k)\| \leq O(\epsilon)$*, and* $|F(\mathbf{x}^k) - F^\star| \leq O(\epsilon)$ *for any* $k \geq \tilde{K}_{\epsilon,\sigma} = O(\frac{1}{\epsilon}\ln\frac{1}{\epsilon})$.

Proof. See Appendix D. □

Note that we may also apply FDGM [17] to solve problem (1) under Assumptions 2 and 6 on time-varying networks under Assumption 1. The difference between FDGM and the proposed RFDGM is that no regularization is carried out in FDGM. In [17], we show that the iteration complexity of FDGM is $O(1/\epsilon^2)$, which is worse than the iteration complexity $O(\frac{1}{\epsilon}\ln\frac{1}{\epsilon})$ of RFDGM provided in Theorem 3. This indicates that *the regularization technique introduced in Section III-A not only enables RFDGM to handle more general problems than FDGM but also significantly reduces the iteration complexity of FDGM*. Moreover, the convergence analysis of FDGM additionally requires the global constraint $\cap_{i\in\mathcal{V}}X_i$ to have a nonempty interior, while all the convergence results throughout this section do not.

Similar to Section IV-C, we also provide a computable upper bound on $\tilde{K}_{\epsilon,\sigma}$. For any $\hat{\mathbf{x}} \in X \cap S$, we have

$$\tilde{K}_{\epsilon,\sigma} \leq \frac{3Bn^2(\bar{D}^W)^2(Bn\bar{c}^2 + \bar{\lambda}\bar{h}/\underline{h})}{8\max\{\underline{c} - \underline{c}^2/2, \bar{c} - \bar{c}^2/2\}}\left(\frac{1}{(\underline{D}^W)^2} + \frac{2}{\epsilon\underline{\theta}}\right) \times \ln\frac{D(\mathbf{w}^0) + F(\hat{\mathbf{x}}) + \frac{\epsilon}{4}(1 + \frac{\|\mathbf{w}^0\|^2}{(\underline{D}^W)^2})}{\sigma},$$

which can also be evaluated explicitly with the knowledge of the D_i^W's. We will discuss how to acquire D_i^W in Section V.

E. Comparison with Related Algorithms

Finally, we compare the assumptions and the convergence results of RFDGM and the existing distributed optimization methods [12]–[17] that are also capable of solving problem (1) over time-varying networks.

Table II lists the assumptions imposed in [12]–[17] and this paper. It can be seen that strong convexity of the objective functions is required by DPDA-TV [15], the Dykstra algorithm [16], and FDGM [17], among which DPDA-TV additionally assumes the constraint sets X_i's to be compact. The projected subgradient method [12], the proximal minimization method [13], DPDA-D [14], and our proposed RFDGM are able to handle general convex objective functions, which are all accompanied by the compactness condition on the X_i's. However, note that RFDGM does not require the X_i's to be bounded when solving strongly convex problems. Furthermore, RFDGM eliminates the nonemptyness of the interior of the global constraint set $\cap_{i\in\mathcal{V}}X_i$ assumed in [12], [13], [17], so that RFDGM can tackle, for instance, constraints described by linear equations.

The projected subgradient method [12] and the proximal minimization method [13] are guaranteed to be asymptotically convergent, but no non-asymptotic results are provided. The Dykstra algorithm [16] and FDGM [17] both establish iteration complexities of $O(\frac{1}{\epsilon^2})$ to reach ϵ-accuracy in solving strongly convex problems, which are much higher than that of $O(\frac{1}{\epsilon}\ln\frac{1}{\epsilon})$ for RFDGM in Theorem 3.

When the constraint sets are compact, DPDA-D [14] derives an iteration complexity of $O(\frac{1}{\epsilon})$ for problems with general convex objective functions and DPDA-TV [15] derives $O(\frac{1}{\sqrt{\epsilon}})$ for problems with strongly convex objective functions. These results look better than those in Theorems 2 and 3 (though Theorem 3 allows unbounded X_i's). Nevertheless, the iteration complexities only indicate the number of iterations needed to reach ϵ-accuracy but do not reflect the complexities of each iteration. Indeed, at every iteration of DPDA-D and DPDA-TV, there is an inner loop that consists of an increasing number of consensus operations. Specifically, the number of consensus operations at each iteration k is $O(k^{1/p})$, $p \geq 1$ or $O((\ln k)^2)$ for DPDA-D, and is at least $(20+4c)(\log_{1/\beta} k)$, $c > 0$, $\beta \in (0,1)$ for DPDA-TV. Observe that such numbers

of consensus operations grow unbounded, which may cause implementation issues in practice. In contrast, RFDGM, along with the existing algorithms in [12], [13], [17], only requires each node to transmit a d-dimensional vector to its neighbors at every iteration.

V. Parameter Selection

In this section, we investigate the selections of the parameters of RFDGM, including the regularization parameters, the weight matrices, and the step-sizes, which suffice to guarantee the convergence results in Section IV.

A. Regularization Parameters

Given the prespecified accuracy $\epsilon > 0$, the regularization parameters γ_i and κ_i are selected as (13), which depend on the constants D_i^X and D_i^W satisfying (14). We may simply set $D_i^X = \max_{x\in X_i} \|x\|^2/2$ for each $i \in \mathcal{V}$, which can be evaluated by node i on its own and is equal to ∞ for unbounded X_i. To determine D_i^W, we only need to obtain an upper bound on $\|\mathbf{w}^\star\|$ for any optimum $\mathbf{w}^\star \in W^\star$ of the Fenchel dual problem (3). Although we can empirically take a sufficiently large value to estimate such an upper bound, below we provide two theoretical approaches to deriving upper bounds on $\|\mathbf{w}^\star\|$ under different conditions of the X_i's.

1) Nonempty interior of the global constraint set: Assume $\operatorname{int}\cap_{i\in\mathcal{V}} X_i \neq \emptyset$ and arbitrarily pick $x' \in \operatorname{int}\cap_{i\in\mathcal{V}} X_i$. According to [17, Proposition 3],

$$\|\mathbf{w}^\star\| \leq \frac{(\sum_{i\in\mathcal{V}} \max_{x_i\in B(x',r_c)} f_i(x_i)) - F^\star}{r_c},\ \forall \mathbf{w}^\star \in W^\star,$$

where $r_c \in (0,\infty)$ is such that $B(x',r_c) := \{x \in \mathbb{R}^d : \|x - x'\| \leq r_c\} \subseteq \bigcap_{i\in\mathcal{V}} X_i$. Note that r_c can be set to $\min_{i\in\mathcal{V}} r_i$, where $r_i \in (0,\infty)$ satisfies $B(x',r_i) \subseteq X_i$. Also note that

$$\begin{aligned} -F^\star = D^\star &\leq D(\mathbf{0}_{nd}) \\ &= \max_{\mathbf{x}\in X} -F(\mathbf{x}) = -\sum_{i\in\mathcal{V}} \min_{x_i\in X_i} f_i(x_i), \end{aligned}$$

and $\min_{x_i\in X_i} f_i(x_i)\ \forall i \in \mathcal{V}$ are finite if either Assumption 5 or Assumption 6 holds. Therefore, by letting

$$\bar{u}_i = \max_{x_i\in B(x',r_i)} f_i(x_i) - \min_{x_i\in X_i} f_i(x_i) \in (0,\infty),$$

we have

$$\|\mathbf{w}^\star\| \leq \sum_{i\in\mathcal{V}} \bar{u}_i / \min_{i\in\mathcal{V}} r_i.$$

Suppose x' is known to all the nodes. Then, $\bar{u}_i$ and r_i can be assessed by node i itself, so that the upper bound $\sum_{i\in\mathcal{V}} \bar{u}_i/\min_{i\in\mathcal{V}} r_i$ can be collaboratively determined by the nodes using distributed consensus methods (e.g., [26]).

2) Identical local constraints: Suppose Assumption 5 holds and $X_1 = \cdots = X_n$. Then, for any $\mathbf{x} = ((x_1)^T,\ldots,(x_n)^T)^T \in X$, $P_S(\mathbf{x}) = \mathbf{1}_d \otimes (\sum_{i=1}^n x_i/n) \in X$. Let $\mathbf{x}^\star = ((x^\star)^T,\ldots,(x^\star)^T)^T \in X\cap S$ be an optimal solution to problem (2). According to the first-order optimality condition of problem (2), there exists a subgradient $g(\mathbf{x}^\star) \in \partial F(\mathbf{x}^\star)$ such that

$$\langle g(\mathbf{x}^\star), \mathbf{x} - \mathbf{x}^\star\rangle \geq 0, \quad \forall \mathbf{x} \in X \cap S. \tag{29}$$

Then, consider the following proposition.

Proposition 1. *Suppose Assumptions 2, 5 hold and $X_i = X_j$ $\forall i,j \in \mathcal{V}$. Let $g(\mathbf{x}^\star) \in \partial F(\mathbf{x}^\star)$ be such that* (29) *holds. Then, $\tilde{\mathbf{w}}^\star := P_{S^\perp}(g(\mathbf{x}^\star))$ is an optimal solution to problem* (3).

Proof. See Appendix E. □

Since $\|P_{S^\perp}(g(\mathbf{x}^\star))\| \leq \|g(\mathbf{x}^\star)\|$, we have

$$\|\tilde{\mathbf{w}}^\star\| \leq \|g(\mathbf{x}^\star)\| \leq \sum_{i\in\mathcal{V}} \bar{g}_i,$$

where $\bar{g}_i = \max_{x\in X_i} \max_{y\in\partial f_i(x)} \|y\|$. Note that $\bar{g}_i$ is finite because X_i and $\partial f_i(x)\ \forall x \in X_i$ are compact, and can be evaluated by each node i using its local objective function and local constraint set only. Again, $\sum_{i\in\mathcal{V}} \bar{g}_i$ can be determined in a distributed fashion at relatively low cost compared to addressing the distributed optimization problem.

With D_i^X and D_i^W at hand, we are also able to obtain upper bounds on $K_{\epsilon,\sigma}$ and $\tilde{K}_{\epsilon,\sigma}$ as discussed in Section IV-C and Section IV-D.

B. Weight Matrices and Step-sizes

To execute RFDGM, every pair of neighboring nodes i and j at time k need to agree with each other on the value of $h_{ij}^k = h_{ji}^k$ in the predetermined interval $[\underline{h},\bar{h}]$. If the weight matrix $H_{\mathcal{G}^k}$ happens to be, say, the Laplacian matrix $L_{\mathcal{G}^k}$ of $\mathcal{G}^k$, then the weights $h_{ij}^k\ \forall i \in \mathcal{V}\ \forall j \in \mathcal{N}_i^k$ are 1 by default and there is no communication cost for selecting the weights. Another typical weight matrix is the Metropolis weight matrix given by

$$[H_{\mathcal{G}^k}]_{ij} = \begin{cases} \sum_{s\in\mathcal{N}_i^k} \frac{1}{\max\{|\mathcal{N}_i^k|c_i, |\mathcal{N}_s^k|c_s\}}, & \text{if } i = j, \\ -\frac{1}{\max\{|\mathcal{N}_i^k|c_i, |\mathcal{N}_j^k|c_j\}}, & \text{if } \{i,j\} \in \mathcal{E}^k, \\ 0, & \text{otherwise}, \end{cases} \quad \forall i,j \in \mathcal{V}, \tag{30}$$

where $c_i = \frac{1}{\gamma_i+\theta_i} + \kappa_i\ \forall i \in \mathcal{V}$. For the Metropolis weight matrix, each pair of neighboring nodes i and j need to share their $|\mathcal{N}_i^k|c_i$ and $|\mathcal{N}_j^k|c_j$.

To guarantee the convergence of RFDGM, the step-sizes $\alpha^k\ \forall k \geq 0$ should be selected according to (11), which in turn rely on the value of δ given by (10). Since $\lambda_1^\downarrow(H_{\mathcal{G}^k}) \leq \bar{h}\lambda_1^\downarrow(L_{\mathcal{G}^k}) \leq \bar{h}n$, we can always set $\delta = L_{\gamma,\kappa}\bar{h}n$. Likewise, it can be calculated via distributed consensus methods (e.g., [26]). In particular, if $H_{\mathcal{G}^k}$ is given by the Metropolis weight matrix (30), we can simply set $\delta = 2$ [17].

VI. Numerical Examples

In this section, we simulate the convergence performance of RFDGM in solving a class of distributed optimization problems with general convex objective functions, and compare it with that of the alternative methods in the literature.

Consider the following constrained l_1-regularized problem:

$$\begin{aligned} \underset{x\in\mathbb{R}^d}{\text{minimize}} \quad & \textstyle\sum_{i\in\mathcal{V}} (x^T A_i x + b_i^T x + \frac{1}{n}\|x\|_1) \\ \text{subject to} \quad & p_i \leq x \leq q_i, \quad \forall i \in \mathcal{V}, \end{aligned} \tag{31}$$

Algorithm	strongly convex f_i's	compact X_i's	nonempty interior of $\cap_{i\in\mathcal{V}} X_i$
projected subgradient [12]		$\surd$	$\surd$
proximal minimization [13]		$\surd$	$\surd$
DPDA-D [14]		$\surd$	
DPDA-TV [15]	$\surd$	$\surd$	
Dykstra [16]	$\surd$		
FDGM [17]	$\surd$		$\surd$
RFDGM (Theorem 2)		$\surd$	
RFDGM (Theorem 3)	$\surd$		

TABLE II
Assumptions of RFDGM and related methods. Here, $\surd$ means the assumption is required.

where each $A_i \in \mathbb{R}^{d\times d}$ is positive semidefinite, $b_i \in \mathbb{R}^d$, and $p_i \le x \le q_i$ is an element-wise inequality with $p_i, q_i \in \mathbb{R}^d$. We let $n = 50$ and $d = 5$. We also carefully choose p_i and q_i such that the interior of $\cap_{i\in\mathcal{V}}[p_i, q_i]$ is nonempty, which is required in [12], [13].

To create a B-connected time-varying graph $\mathcal{G}^k = (\mathcal{V}, \mathcal{E}^k)$, we first randomly generate a connected graph $\mathcal{G}' = (\mathcal{V}, \mathcal{E}')$. Then, we divide $\mathcal{E}'$ into B subsets, and cyclically choose each subset as $\mathcal{E}^k$, $k \ge 0$. Clearly, the resulting graph satisfies B-connectivity in Assumption 1. We set $B = 5$ and $B = 20$ to test the effect of B.

In the simulation, we choose to run RFDGM, the consensus-based projected subgradient method [12], the proximal minimization algorithm [13], and DPDA-D [14], which are theoretically guaranteed to solve problem (31). For RFDGM, we set $\gamma_i = 10^{-1}$ and $\kappa_i = 10^{-4}$ for each $i \in \mathcal{V}$, and choose the Metropolis weight matrix (30) as $H_{\mathcal{G}^k}$. For the other three algorithms, we let the average matrix be as follows:

$$[W_{\mathcal{G}^k}]_{ij} = \begin{cases} \frac{1}{\max\{|\mathcal{N}_i^k|,|\mathcal{N}_j^k|\}+1}, & \{i,j\} \in \mathcal{E}^k, \\ 1 - \sum_{s\in\mathcal{N}_i^k} \frac{1}{\max\{|\mathcal{N}_i^k|,|\mathcal{N}_j^k|\}+1}, & i = j, \\ 0, & \text{otherwise}, \end{cases} \quad \forall i,j \in \mathcal{V}. \tag{32}$$

The consensus-based projected subgradient method and the proximal minimization algorithm both require diminishing step-sizes to guarantee convergence. Thus, we set their step-sizes to be c/k for some fine-tuned constant c. We also fine-tune the step-sizes of RFDGM and DPDA-D, and let DPDA-D execute $\lceil (\ln k)^2 + 1 \rceil$ consensus operations at each iteration k as is suggested in [14].

Fig. 1 displays the primal optimality error $|F(\mathbf{x}^k) - F^\star|$ and the consensus error (i.e., primal infeasibility) $\frac{1}{n}\sum_{i=1}^n \|x_i^k - \bar{x}^k\|$ generated by the aforementioned distributed algorithms, where $\bar{x}^k = \frac{1}{n}\sum_{i\in\mathcal{V}} x_i^k$ is the average of all the x_i^k's. It can be observed that RFDGM converges faster than the consensus-based projected subgradient method and DPDA-D in both optimality error and consensus error for both $B = 5$ and $B = 20$. Also recall that the communication cost of RFDGM per iteration is much lower than that of DPDA-D and is the same as those of the other two algorithms. Although the consensus error of the proximal minimization algorithm vanishes faster than that of RFDGM, its optimality error decreases much slower. In addition, through comparison between Fig. 1 (a)–(b) and Fig. 1 (c)–(d), we conclude that larger B leads to slower convergence of RFDGM. This is natural because larger B means slower information fusion and is consistent with our convergence analysis in Section IV.

VII. Conclusion

We have developed a regularized Fenchel dual gradient method, referred to as RFDGM, for constrained, nonsmooth, distributed convex optimization on time-varying networks. The proposed RFDGM is constructed based on a primal-Fenchel-dual regularization procedure and a weighted gradient method for solving a regularized Fenchel dual problem. We have shown that RFDGM is linearly convergent and capable of converging to suboptimality of any given accuracy. We have also provided the corresponding iteration complexities for problems with general convex objective functions and compact constraint sets as well as problems with strongly convex objective functions and possibly unbounded constraint sets. Moreover, we have described several approaches to selecting the algorithm parameters in a decentralized manner. Future works include extending RFDGM to time-varying directed networks and investigating more general, non-quadratic regularization techniques for accelerating the convergence.

Appendix

A. Proof of Theorem 1

This proof will utilize the following two lemmas.

Lemma 2. *[17, Lemma 1] For each $k \ge 0$,*

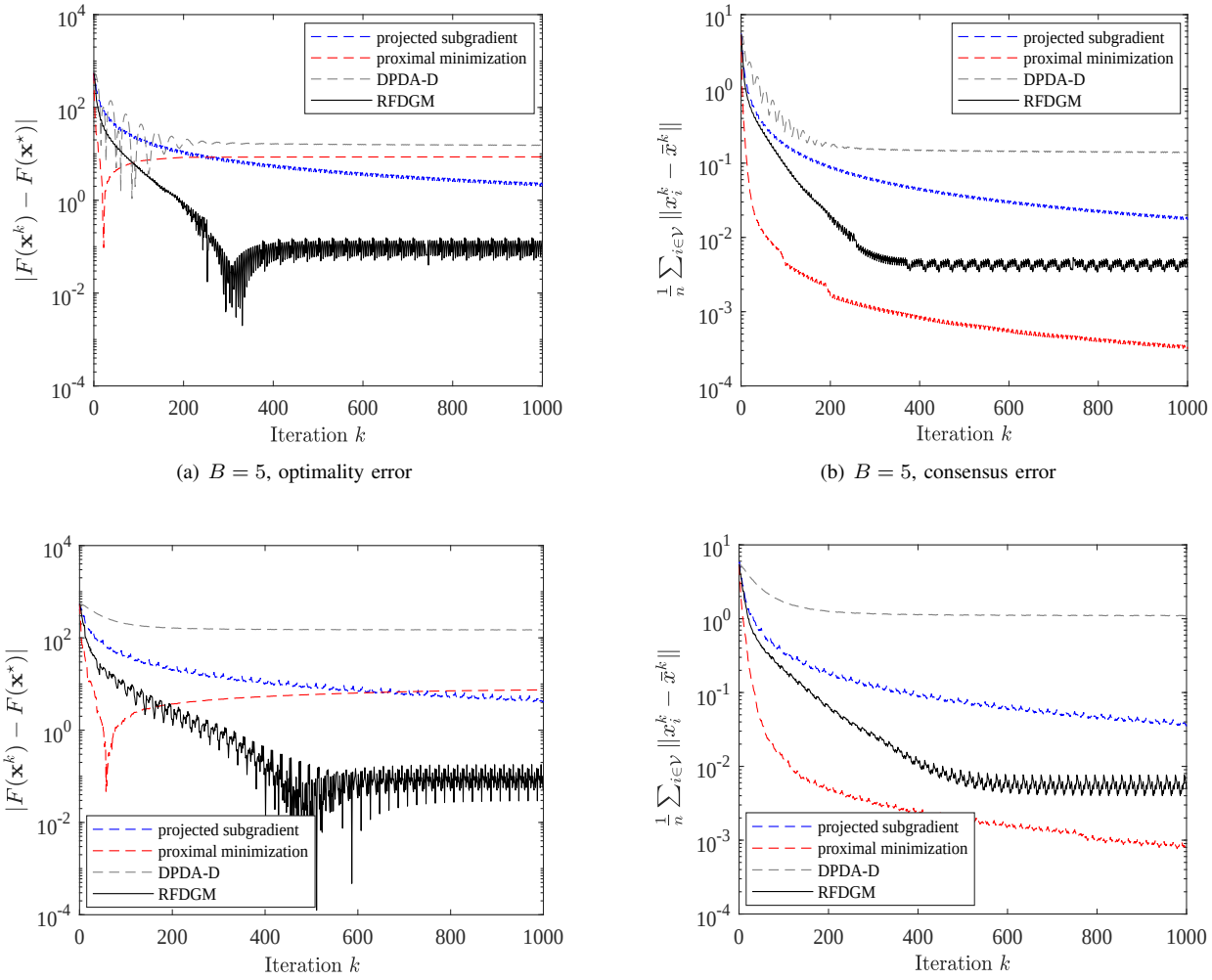
$$D_{\gamma,\kappa}(\mathbf{w}^{k+1}) - D_{\gamma,\kappa}(\mathbf{w}^k) \le -\rho(\nabla D_{\gamma,\kappa}(\mathbf{w}^k))^T (H_{\mathcal{G}^k} \otimes I_d) \nabla D_{\gamma,\kappa}(\mathbf{w}^k).$$

Lemma 3. *[17, Lemma 2] For each $k \ge 0$,*

$$\sum_{t=kB}^{(k+1)B-1} (\nabla D_{\gamma,\kappa}(\mathbf{w}^t))^T (H_{\mathcal{G}^t} \otimes I_d) \nabla D_{\gamma,\kappa}(\mathbf{w}^t) \ge (\nabla D_{\gamma,\kappa}(\mathbf{w}^{kB}))^T (L_{\mathcal{T}^k} \otimes I_d) \nabla D_{\gamma,\kappa}(\mathbf{w}^{kB})/\eta.$$

By combining the above two lemmas, we obtain

$$\begin{aligned} D_{\gamma,\kappa}(\mathbf{w}^{(k+1)B}) - D_{\gamma,\kappa}(\mathbf{w}^{kB}) &\le -\frac{\rho}{\eta}(\nabla D_{\gamma,\kappa}(\mathbf{w}^{kB}))^T (L_{\mathcal{T}^k} \otimes I_d) \nabla D_{\gamma,\kappa}(\mathbf{w}^{kB}) \\ &\le -\frac{\rho \underline{\lambda}_T}{\eta} \|P_{S^\perp}(\nabla D_{\gamma,\kappa}(\mathbf{w}^{kB}))\|^2, \quad \forall k \ge 0, \end{aligned} \tag{33}$$



(a) $B = 5$, optimality error

(b) $B = 5$, consensus error

(c) $B = 20$, optimality error

(d) $B = 20$, consensus error

Fig. 1. Convergence of the consensus-based projected subgradient method, the proximal minimization algorithm, DPDA-D, and RFDGM.

where the last inequality is due to $\mathrm{Null}(L_{\mathcal{T}^k}) = S$. Because $D_{\gamma,\kappa}$ is $\underline{\kappa}$-strongly convex and because of [24, Eq. (19)],

$$D_{\gamma,\kappa}(\mathbf{w}^{kB}) - D^\star_{\gamma,\kappa} \le \frac{1}{2\underline{\kappa}}\|P_{S^\perp}(\nabla D_{\gamma,\kappa}(\mathbf{w}^{kB}))\|^2, \quad \forall k \ge 0.$$

This, along with (33), implies that for any $k \ge 0$,

$$\begin{aligned} &D_{\gamma,\kappa}(\mathbf{w}^{(k+1)B}) - D_{\gamma,\kappa}(\mathbf{w}^{kB}) \\ &\le -\frac{2\underline{\kappa}\rho\underline{\lambda}_T}{\eta}(D_{\gamma,\kappa}(\mathbf{w}^{kB}) - D^\star_{\gamma,\kappa}). \end{aligned}$$

Due to Lemma 2, $D_{\gamma,\kappa}(\mathbf{w}^k)$ is non-increasing. Therefore, the above inequality leads to (12).

B. Proof of Lemma 1

For any $\mathbf{w} \in \mathbb{R}^{nd}$, by the definition of D and D_γ, we have

$$\begin{aligned} D_\gamma(\mathbf{w}) &= \sup_{\mathbf{y}\in X} \langle \mathbf{w}, \mathbf{y}\rangle - F(\mathbf{y}) - \frac{\|\mathbf{y}\|^2_{\Lambda_\gamma}}{2} \\ &\ge \langle \mathbf{w}, \mathbf{x}\rangle - F(\mathbf{x}) - \frac{\|\mathbf{x}\|^2_{\Lambda_\gamma}}{2} \\ &= D(\mathbf{w}) - \frac{\|\mathbf{x}\|^2_{\Lambda_\gamma}}{2}, \quad \forall \mathbf{x} \in G(\mathbf{w}), \end{aligned} \tag{34}$$

$$\begin{aligned} D_\gamma(\mathbf{w}) &= \sup_{\mathbf{y}\in X} \langle \mathbf{w}, \mathbf{y}\rangle - F(\mathbf{y}) - \frac{\|\mathbf{y}\|^2_{\Lambda_\gamma}}{2} \\ &\le \sup_{\mathbf{y}\in X} \langle \mathbf{w}, \mathbf{y}\rangle - F(\mathbf{y}) = D(\mathbf{w}), \end{aligned} \tag{35}$$

where $G(\mathbf{w})$ is defined in Section III-A. Let $\mathbf{w} \in S^\perp$ and $\mathbf{w}^\star = \arg\min_{\mathbf{w}\in W^\star} \|\mathbf{w}\|$. For any $\mathbf{x} \in X$,

$$\|\mathbf{x}\|^2_{\Lambda_\gamma} = \sum_{i\in\mathcal{V}} \frac{\epsilon}{4nD_i^X}\|x_i\|^2 \le \frac{\epsilon}{2}, \tag{36}$$

which, together with (34), gives

$$D_\gamma(\mathbf{w}) \ge D(\mathbf{w}) - \frac{\epsilon}{4}. \tag{37}$$

Moreover, since $D_{\gamma,\kappa}(\mathbf{w}) \ge D_\gamma(\mathbf{w})$, we have

$$D_{\gamma,\kappa}(\mathbf{w}) \ge D(\mathbf{w}) - \frac{\epsilon}{4}. \tag{38}$$

By (35) and $\|\mathbf{w}^\star\|^2_{\Lambda_\kappa} \le \frac{\epsilon\|\mathbf{w}^\star\|^2}{2(\underline{D}^W)^2} \le \frac{\epsilon}{2}$,

$$D_{\gamma,\kappa}(\mathbf{w}^\star_{\gamma,\kappa}) \le D_{\gamma,\kappa}(\mathbf{w}^\star) = D_\gamma(\mathbf{w}^\star) + \frac{1}{2}\|\mathbf{w}^\star\|^2_{\Lambda_\kappa}$$

$$\leq D(\mathbf{w}^\star)+\frac{\epsilon}{4}. \tag{39}$$

Combining (38) and (39) yields (15).

From (6), we have

$$\begin{aligned}&\|P_{S^\perp}(\tilde{\mathbf{x}}_\gamma(\mathbf{w}))\| = \|P_{S^\perp}(\nabla D_{\gamma,\kappa}(\mathbf{w})-\Lambda_\kappa\mathbf{w})\|\\ \leq&\|P_{S^\perp}(\nabla D_{\gamma,\kappa}(\mathbf{w}))\| + \|\Lambda_\kappa(\mathbf{w}-\mathbf{w}^\star_{\gamma,\kappa})\| + \|\Lambda_\kappa\mathbf{w}^\star_{\gamma,\kappa}\|.\end{aligned} \tag{40}$$

Because $\nabla D_{\gamma,\kappa}(\mathbf{w}^\star_{\gamma,\kappa})\in S$ and $\mathbf{w},\mathbf{w}^\star_{\gamma,\kappa}\in S^\perp$,

$$\langle \nabla D_{\gamma,\kappa}(\mathbf{w}^\star_{\gamma,\kappa}), \mathbf{w}-\mathbf{w}^\star_{\gamma,\kappa}\rangle = 0. \tag{41}$$

This, along with the $L_{\gamma,\kappa}$-smoothness of $D_{\gamma,\kappa}$ and [27, Eq.(2.1.7)], gives

$$\|\nabla D_{\gamma,\kappa}(\mathbf{w})-\nabla D_{\gamma,\kappa}(\mathbf{w}^\star_{\gamma,\kappa})\| \leq \sqrt{2L_{\gamma,\kappa}(D_{\gamma,\kappa}(\mathbf{w})-D^\star_{\gamma,\kappa})}.$$

Due again to $\nabla D_{\gamma,\kappa}(\mathbf{w}^\star_{\gamma,\kappa})\in S$,

$$\begin{aligned}\|P_{S^\perp}(\nabla D_{\gamma,\kappa}(\mathbf{w}))\| &= \|\nabla D_{\gamma,\kappa}(\mathbf{w}) - P_S(\nabla D_{\gamma,\kappa}(\mathbf{w}))\|\\ &\leq \|\nabla D_{\gamma,\kappa}(\mathbf{w}) - \nabla D_{\gamma,\kappa}(\mathbf{w}^\star_{\gamma,\kappa})\|.\end{aligned}$$

Combining the above two inequalities,

$$\|P_{S^\perp}(\nabla D_{\gamma,\kappa}(\mathbf{w}))\| \leq \sqrt{2L_{\gamma,\kappa}\big(D_{\gamma,\kappa}(\mathbf{w})-D^\star_{\gamma,\kappa}\big)}. \tag{42}$$

By (41) and the convexity of D_γ, we have

$$\begin{aligned}&D_{\gamma,\kappa}(\mathbf{w})-D^\star_{\gamma,\kappa}\\ =&D_\gamma(\mathbf{w})-D_\gamma(\mathbf{w}^\star_{\gamma,\kappa})+\frac{\|\mathbf{w}\|^2_{\Lambda_\kappa}}{2}-\frac{\|\mathbf{w}^\star_{\gamma,\kappa}\|^2_{\Lambda_\kappa}}{2}\\ \geq&\langle\nabla D_\gamma(\mathbf{w}^\star_{\gamma,\kappa}),\mathbf{w}-\mathbf{w}^\star_{\gamma,\kappa}\rangle+\frac{\|\mathbf{w}\|^2_{\Lambda_\kappa}}{2}-\frac{\|\mathbf{w}^\star_{\gamma,\kappa}\|^2_{\Lambda_\kappa}}{2}\\ =&\langle-\Lambda_\kappa\mathbf{w}^\star_{\gamma,\kappa},\mathbf{w}-\mathbf{w}^\star_{\gamma,\kappa}\rangle+\frac{\|\mathbf{w}\|^2_{\Lambda_\kappa}}{2}-\frac{\|\mathbf{w}^\star_{\gamma,\kappa}\|^2_{\Lambda_\kappa}}{2}\\ =&\frac{\|\mathbf{w}-\mathbf{w}^\star_{\gamma,\kappa}\|^2_{\Lambda_\kappa}}{2},\end{aligned} \tag{43}$$

and therefore

$$\begin{aligned}&\|\Lambda_\kappa(\mathbf{w}-\mathbf{w}^\star_{\gamma,\kappa})\| \leq \sqrt{\max_{i\in\mathcal{V}}\kappa_i}\|\mathbf{w}-\mathbf{w}^\star_{\gamma,\kappa}\|_{\Lambda_\kappa}\\ \leq&\sqrt{\max_{i\in\mathcal{V}}\kappa_i}\sqrt{2(D_{\gamma,\kappa}(\mathbf{w})-D^\star_{\gamma,\kappa})}\\ =&\frac{\sqrt{\epsilon}}{\underline{D}^W}\sqrt{D_{\gamma,\kappa}(\mathbf{w})-D^\star_{\gamma,\kappa}}.\end{aligned} \tag{44}$$

From (37) and (35),

$$\begin{aligned}&D(\mathbf{w}^\star)+\frac{\|\mathbf{w}^\star\|^2_{\Lambda_\kappa}}{2}\geq D_\gamma(\mathbf{w}^\star)+\frac{\|\mathbf{w}^\star\|^2_{\Lambda_\kappa}}{2}=D_{\gamma.\kappa}(\mathbf{w}^\star)\\ \geq&D_{\gamma,\kappa}(\mathbf{w}^\star_{\gamma,\kappa})=D_\gamma(\mathbf{w}^\star_{\gamma,\kappa})+\frac{\|\mathbf{w}^\star_{\gamma,\kappa}\|^2_{\Lambda_\kappa}}{2}\\ \geq&D(\mathbf{w}^\star_{\gamma,\kappa})-\frac{\epsilon}{4}+\frac{\|\mathbf{w}^\star_{\gamma,\kappa}\|^2_{\Lambda_\kappa}}{2}\geq D(\mathbf{w}^\star)-\frac{\epsilon}{4}+\frac{\|\mathbf{w}^\star_{\gamma,\kappa}\|^2_{\Lambda_\kappa}}{2}.\end{aligned} \tag{45}$$

Thus,

$$\begin{aligned}&\|\Lambda_\kappa\mathbf{w}^\star_{\gamma,\kappa}\| \leq \sqrt{\max_{i\in\mathcal{V}}\kappa_i}\|\mathbf{w}^\star_{\gamma,\kappa}\|_{\Lambda_\kappa}\\ \leq&\sqrt{(\max_{i\in\mathcal{V}}\kappa_i)((\max_{i\in\mathcal{V}}\kappa_i)\|\mathbf{w}^\star\|^2+\frac{\epsilon}{2})}\leq\frac{\epsilon}{\sqrt{2}\underline{D}^W}.\end{aligned} \tag{46}$$

Substituting (42), (44), and (46) into (40) results in (16).

Since $\mathbf{w}\in S^\perp$, we have

$$\begin{aligned}D_\gamma(\mathbf{w}) &= \langle\mathbf{w},\tilde{\mathbf{x}}_\gamma(\mathbf{w})\rangle - F(\tilde{\mathbf{x}}_\gamma(\mathbf{w}))-\frac{1}{2}\|\tilde{\mathbf{x}}_\gamma(\mathbf{w})\|^2_{\Lambda_\gamma}\\ &\leq \langle\mathbf{w},P_{S^\perp}(\tilde{\mathbf{x}}_\gamma(\mathbf{w}))\rangle - F(\tilde{\mathbf{x}}_\gamma(\mathbf{w})),\end{aligned}$$

which, together with (37) and $-F^\star=D^\star\leq D(\mathbf{w})$, results in

$$\begin{aligned}&F(\tilde{\mathbf{x}}_\gamma(\mathbf{w}))-F^\star\\ \leq&\langle\mathbf{w},P_{S^\perp}(\tilde{\mathbf{x}}_\gamma(\mathbf{w}))\rangle-(D_\gamma(\mathbf{w})-D(\mathbf{w}))\\ \leq&\langle\mathbf{w},P_{S^\perp}(\tilde{\mathbf{x}}_\gamma(\mathbf{w}))\rangle+\frac{\epsilon}{4}\\ \leq&(\|\mathbf{w}-\mathbf{w}^\star_{\gamma,\kappa}\|+\|\mathbf{w}^\star_{\gamma,\kappa}\|)\|P_{S^\perp}(\tilde{\mathbf{x}}_\gamma(\mathbf{w}))\|+\frac{\epsilon}{4}.\end{aligned} \tag{47}$$

From (43), we have

$$\begin{aligned}\|\mathbf{w}-\mathbf{w}^\star_{\gamma,\kappa}\| &\leq \frac{1}{\sqrt{\underline{\kappa}}}\|\mathbf{w}-\mathbf{w}^\star_{\gamma,\kappa}\|_{\Lambda_\kappa}\\ \leq&\frac{1}{\sqrt{\underline{\kappa}}}\sqrt{2(D_{\gamma,\kappa}(\mathbf{w})-D^\star_{\gamma,\kappa})}\\ =&2\bar{D}^W\sqrt{\frac{D_{\gamma,\kappa}(\mathbf{w})-D^\star_{\gamma,\kappa}}{\epsilon}},\end{aligned}$$

and due to (45) and $\|\mathbf{w}^\star\|^2_{\Lambda_\kappa}\leq\frac{\epsilon}{2(\underline{D}^W)^2}\|\mathbf{w}^\star\|^2\leq\frac{\epsilon}{2}$,

$$\begin{aligned}\|\mathbf{w}^\star_{\gamma,\kappa}\| &\leq \frac{1}{\sqrt{\underline{\kappa}}}\|\mathbf{w}^\star_{\gamma,\kappa}\|_{\Lambda_\kappa}\\ \leq&\frac{1}{\sqrt{\underline{\kappa}}}\sqrt{\|\mathbf{w}^\star\|^2_{\Lambda_\kappa}+\frac{\epsilon}{2}}\leq\sqrt{2}\bar{D}^W.\end{aligned}$$

Substituting these and (16) into (47) leads to (17).

Finally, due to (35), $-F^\star=D^\star$, and (36),

$$\begin{aligned}&F(\tilde{\mathbf{x}}_\gamma(\mathbf{w}))-F^\star\\ =&\langle\mathbf{w},\tilde{\mathbf{x}}_\gamma(\mathbf{w})\rangle-(D_\gamma(\mathbf{w})+\frac{1}{2}\|\tilde{\mathbf{x}}_\gamma(\mathbf{w})\|^2_{\Lambda_\gamma}-D^\star)\\ \geq&\langle\mathbf{w},\tilde{\mathbf{x}}_\gamma(\mathbf{w})\rangle-(D(\mathbf{w})+\frac{\epsilon}{4}-D^\star)\\ \geq&-\|\mathbf{w}\|\cdot\|P_{S^\perp}(\tilde{\mathbf{x}}_\gamma(\mathbf{w}))\|-(D(\mathbf{w})+\frac{\epsilon}{4}-D^\star)\\ \geq&-(\|\mathbf{w}-\mathbf{w}^\star_{\gamma,\kappa}\|+\|\mathbf{w}^\star_{\gamma,\kappa}\|)\|P_{S^\perp}(\tilde{\mathbf{x}}_\gamma(\mathbf{w}))\|\\ &-(D(\mathbf{w})-D^\star)-\frac{\epsilon}{4}.\end{aligned}$$

This, along with (44), (46), (16), and (15), implies (18).

C. Proof of Theorem 2

Because of Theorem 1 and because $1-\frac{2\underline{\kappa}\rho\underline{\lambda}_T}{\eta}\leq e^{-\frac{2\underline{\kappa}\rho\underline{\lambda}_T}{\eta}}$, we derive that if

$$k\geq K' := \frac{\eta B}{2\underline{\kappa}\rho\underline{\lambda}_T}\ln\frac{D_{\gamma,\kappa}(\mathbf{w}^0)-D^\star_{\gamma,\kappa}}{\sigma}, \tag{48}$$

then $D_{\gamma,\kappa}(\mathbf{w}^k)-D^\star_{\gamma,\kappa}\leq\sigma$. Since $H_{\mathcal{G}^k}\preceq\bar{h}L_{\mathcal{G}^k}$, (10) holds when δ is set to be $L_{\gamma,\kappa}\bar{h}\bar{\lambda}$. Therefore, there always exists $\delta\in(0,L_{\gamma,\kappa}\bar{h}\bar{\lambda}]$ satisfying (10). Combining these with $\underline{\kappa}=\frac{\epsilon}{2(\bar{D}^W)^2}$ and $L_{\gamma,\kappa}\leq\frac{4n\bar{D}^X}{\epsilon}+\frac{\epsilon}{2(\underline{D}^W)^2}$, we have

$$\frac{\eta}{\underline{\kappa}\rho}\leq\frac{3(\bar{D}^W)^2(B\bar{\varpi}\bar{c}^2+\bar{\lambda}\bar{h}/\underline{h})}{\max\{\underline{c}-\underline{c}^2/2,\bar{c}-\bar{c}^2/2\}}$$

$$\cdot \Big(\frac{1}{(\underline{D}^W)^2}+\frac{8n\bar{D}^X}{\epsilon^2}\Big). \quad (49)$$

In addition, due to (35), (13), and (38), we have

$$D_{\gamma,\kappa}(\mathbf{w}^0)=D_\gamma(\mathbf{w}^0)+\frac{1}{2}\|\mathbf{w}^0\|^2_{\Lambda_\kappa}$$
$$\le D(\mathbf{w}^0)+\frac{\epsilon}{4(\underline{D}^W)^2}\|\mathbf{w}^0\|^2,$$
$$D^\star_{\gamma,\kappa}\ge D(\mathbf{w}^\star_{\gamma,\kappa})-\frac{\epsilon}{4}\ge D^\star-\frac{\epsilon}{4},$$

which further leads to

$$D_{\gamma,\kappa}(\mathbf{w}^0)-D^\star_{\gamma,\kappa}\le D(\mathbf{w}^0)-D^\star+\frac{\epsilon}{4}\left(\frac{\|\mathbf{w}^0\|^2}{(\underline{D}^W)^2}+1\right). \quad (50)$$

Substituting (50) and (49) into (48), we obtain that for any $k\ge K_{\epsilon,\sigma}\ge K'$, where $K_{\epsilon,\sigma}$ is given by (19), $D_{\gamma,\kappa}(\mathbf{w}^k)-D^\star_{\gamma,\kappa}\le\sigma$. This, along with $L_{\gamma,\kappa}\le\frac{4n\bar{D}^X}{\epsilon}+\frac{\epsilon}{2(\underline{D}^W)^2}$ and Lemma 1, implies (20)–(23). Finally, it is straightforward to see that if we take $\sigma=O(\epsilon^3)$, then (20)–(23) are all on the order of $O(\epsilon)$ and $K_{\epsilon,\sigma}=O(\frac{1}{\epsilon^2}\ln\frac{1}{\epsilon})$.

D. Proof of Theorem 3

Note that $L_{\gamma,\kappa}\le\frac{1}{\underline{\theta}}+\frac{\epsilon}{2(\underline{D}^W)^2}$. Similar to (49), we have

$$\frac{\eta}{\underline{\kappa}\rho}\le\frac{6(\bar{D}^W)^2(\frac{1}{\underline{\theta}}+\frac{\epsilon}{2(\underline{D}^W)^2})(B\bar{\varpi}\bar{c}^2+\bar{\lambda}\bar{h}/\underline{h})}{\epsilon\max\{\underline{c}-\underline{c}^2/2,\bar{c}-\bar{c}^2/2\}}. \quad (51)$$

Substituting (50) and (51) into (48), we obtain that for any $k\ge\tilde{K}_{\epsilon,\sigma}$, where $\tilde{K}_{\epsilon,\sigma}$ is given by (24), $D_{\gamma,\kappa}(\mathbf{w}^k)-D^\star_{\gamma,\kappa}\le\sigma$. This, along with $L_{\gamma,\kappa}\le\frac{1}{\underline{\theta}}+\frac{\epsilon}{2(\underline{D}^W)^2}$ and Lemma 1, implies (25)–(28). Finally, it is straightforward to see that if we set $\sigma=O(\epsilon^2)$, then (25)–(28) are all on the order of $O(\epsilon)$ and $\tilde{K}_{\epsilon,\sigma}=O(\frac{1}{\epsilon}\ln\frac{1}{\epsilon})$.

E. Proof of Proposition 1

For any $\mathbf{x}\in X$,

$$\langle P_{S^\perp}(g(\mathbf{x}^\star))-g(\mathbf{x}^\star),\mathbf{x}-\mathbf{x}^\star\rangle=-\langle P_S(g(\mathbf{x}^\star)),\mathbf{x}-\mathbf{x}^\star\rangle$$
$$=-\langle P_S(g(\mathbf{x}^\star)),P_S(\mathbf{x})-\mathbf{x}^\star\rangle=-\langle g(\mathbf{x}^\star),P_S(\mathbf{x})-\mathbf{x}^\star\rangle\le 0.$$

This implies that $\mathbf{x}^\star$ is the optimal solution of $\max_{\mathbf{x}\in X}\langle P_{S^\perp}(g(\mathbf{x}^\star)),\mathbf{x}\rangle-F(\mathbf{x})$, i.e.,

$$D(P_{S^\perp}(g(\mathbf{x}^\star)))=\langle P_{S^\perp}(g(\mathbf{x}^\star)),\mathbf{x}^\star\rangle-F(\mathbf{x}^\star)=-F^\star=D^\star.$$

Therefore, $P_{S^\perp}(g(\mathbf{x}^\star))\in W^\star$.

REFERENCES

[1] S.-H. Son, M. Chiang, S. R. Kulkarni, and S. C. Schwartz, "The value of clustering in distributed estimation for sensor networks," in *Proc. International Conference on Wireless Networks, Communications and Mobile Computing*, Princeton, USA, 2005, pp. 969–974.

[2] A. Beck, A. Nedić, A. Ozdaglar, and M. Teboulle, "An $O(1/k)$ gradient method for network resource allocation problems," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 1, pp. 64–73, 2014.

[3] P. Forero, A. Cano, and G. Giannakis, "Consensus-based distributed support vector machines," *Journal of Machine Learning Research*, vol. 11, no. May, pp. 1663–1707, 2010.

[4] J. Lu, C. Y. Tang, P. R. Regier, and T. D. Bow, "Gossip algorithms for convex consensus optimization over networks," *IEEE Transactions on Automatic Control*, vol. 56, no. 12, pp. 2917–2923, 2011.

[5] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 601– 615, 2015.

[6] ——, "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 3936–3947, 2016.

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

[8] D. Yuan, Y. Hong, D. W. C. Ho, and G. Jiang, "Optimal distributed stochastic mirror descent for strongly convex optimization," *Automatica*, vol. 90, pp. 196–203, 2018.

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

[10] S. Liang, L. Wang, and G. Yin, "Dual averaging push for distributed convex optimization over time-varying directed graph," accepted to *IEEE Transactions on Automatic Control*, 2019.

[11] G. Scutari and Y. Sun, "Distributed nonconvex constrained optimization over time-varying digraphs," *Mathematical Programming*, vol. Series B, no. 176, pp. 497–544, 2019.

[12] A. Nedić, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, 2010.

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

[14] N. S. Aybat and E. Y. Hamedani, "A primal-dual method for conic constrained distributed optimization problems," in *Proc. International Conference on Neural Information Processing Systems*, Barcelona, Spain, 2016, pp. 5056–5064.

[15] E. Y. Hamedani and N. S. Aybat, "Multi-agent constrained optimization of a strongly convex function over time-varying directed networks," in *Proc. Annual Allerton Conference*, Illinois, MA, 2017, pp. 518–525.

[16] C. H. J. Pang, "Distributed deterministic asynchronous algorithms in time-varying graphs through Dykstra splitting," *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 484–510, 2019.

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

[18] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson, "A survey of distributed optimization," *Annual Reviews in Control*, vol. 47, pp. 278–305, 2019.

[19] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA: Athena Scientific, 1999.

[20] X. Wu, K. C. Sou, and J. Lu, "Fenchel dual gradient methods enabling a smoothing technique for nonsmooth distributed convex optimization," in *Proc. IEEE Conference on Decision and Control*, Miami, USA, 2018, pp. 1757–1762.

[21] O. Devolder, F. Glineur, and Y. Nesterov, "Double smoothing technique for large-scale linearly constrained convex optimization," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 702–707, 2012.

[22] H. Lakshmanan and D. P. de Farias, "Decentralized resource allocation in dynamic networks of agents," *SIAM Journal on Optimization*, vol. 19, no. 2, pp. 911–940, 2008.

[23] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak Mathematical Journal*, vol. 23, no. 98, pp. 298–305, 1973.

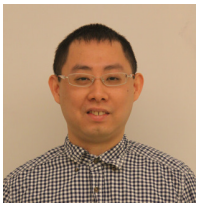
[24] L. Xiao and S. Boyd, "Optimal scaling of a gradient method for distributed resource allocation," *Journal of Optimization Theory and Applications*, vol. 129, no. 3, pp. 469–488, 2006.

[25] B. Mohar, Y. Alavi, G. Chartrand, and O. Oellermann, "The Laplacian spectrum of graphs," *Graph theory, combinatorics, and applications*, vol. 2, pp. 871–898, 1991.

[26] J.-Y. Chen, G. Pandurangan, and D. Xu, "Robust computation of aggregates in wireless sensor networks: Distributed randomized algorithms and analysis," *IEEE Transactions on Parallel and Distributed Systems*, vol. 17, no. 9, pp. 987–1000, 2006.

[27] Y. Nesterov, *Introductory lectures on Convex Optimization: A Basic Course*. Norwell, MA: Kluwer Academic Publishers, 2004.

Xuyang Wu received the B.S. degree in Information and Computing Science from Northwestern Polytechnical University, Xi'an, China, in 2015. He is currently pursuing his Ph.D. degree in the School of Information Science and Technology at ShanghaiTech University, Shanghai, China. His research interests include distributed optimization and large-scale optimization algorithms.



Kin Cheong Sou received a Ph.D. degree in Electrical Engineering and Computer Science at Massachusetts Institute of Technology in 2008. From 2008 to 2010 he was a postdoctoral researcher at Lund University, Lund, Sweden. From 2010 to 2012 he was a postdoctoral researcher at KTH Royal Institute of Technology, Stockholm, Sweden. Between 2013 and 2016 he was an assistant professor with the department of Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg, Sweden. Since 2017 he has been an assistant professor with the department of Electrical Engineering at the National Sun Yat-sen University in Taiwan.

Jie Lu (SM'08-M'13) received the B.S. degree in Information Engineering from Shanghai Jiao Tong University, China, in 2007, and the Ph.D. degree in Electrical and Computer Engineering from the University of Oklahoma, USA, in 2011. From 2012 to 2015 she was a postdoctoral researcher with KTH Royal Institute of Technology, Stockholm, Sweden, and with Chalmers University of Technology, Gothenburg, Sweden. Since 2015, she has been an assistant professor in the School of Information Science and Technology at ShanghaiTech University, Shanghai, China. Her research interests include distributed optimization, optimization theory and algorithms, and networked dynamical systems.