Fenchel Dual Gradient Methods for Distributed Convex Optimization over Time-varying Networks

Xuyang Wu and Jie Lu

Abstract

In the large collection of existing distributed algorithms for convex multi-agent optimization, only a handful of them provide convergence rate guarantees on agent networks with time-varying topologies, which, however, restrict the problem to be unconstrained. Motivated by this, we develop a family of distributed Fenchel dual gradient methods for solving constrained, strongly convex but not necessarily smooth multi-agent optimization problems over time-varying undirected networks. The proposed algorithms are constructed based on the application of weighted gradient methods to the Fenchel dual of the multi-agent optimization problem, and can be implemented in a fully decentralized fashion. We show that the proposed algorithms drive all the agents to both primal and dual optimality asymptotically under a minimal connectivity condition and at sublinear rates under a standard connectivity condition. We also derive bounds on the convergence rate and the suboptimality when the dual gradient is inexactly evaluated at each iteration. Finally, the competent convergence performance of the Fenchel dual gradient methods is demonstrated via simulations.

Index Terms

Distributed optimization, multi-agent optimization, Fenchel duality.

I. Introduction

In many engineering scenarios, a network of agents often need to jointly make a decision so that a global cost consisting of their local costs is minimized and certain global constraints are satisfied. Such a multi-agent optimization problem has found a considerable number of

X. Wu and J. Lu are with the School of Information Science and Technology, ShanghaiTech University, 201210 Shanghai, China. Email: {wuxy, lujie}@shanghaitech.edu.cn.

This work has been supported by the National Natural Science Foundation of China under grant 61603254, the Shanghai Pujiang Program under grant 16PJ1406400, and the Natural Science Foundation of Shanghai under grant 16ZR1422500.

applications, such as estimation by sensor networks [1], network resource allocation [2], and cooperative control [3].

To address convex multi-agent optimization in an efficient, robust, and scalable way, distributed optimization algorithms have been substantially exploited, which allow each agent to reach an optimal or suboptimal decision by repeatedly exchanging its own information with neighbors [1]–[32]. One typical approach is to let the agents perform consensus operations so as to mix their decisions that are updated using first-order information of their local objectives (e.g., [4]–[14]). Recently, rates of convergence to optimality have been established for a few consensus-based algorithms. By assuming that the problem is unconstrained and smooth (i.e., the gradient of each local objective is Lipschitz) and that the network is fixed, the consensus-based multi-step gradient methods [8]–[11] are able to achieve sublinear rates of convergence, and also linear rates if the local objectives are further (restricted) strongly convex. Unlike these algorithms, the Subgradient-Push method [12], the Gradient-Push method [13], the DIGing algorithm [14], and the Push-DIGing algorithm [14] can be implemented over time-varying networks and still provide convergence rate guarantees. Specifically, Subgradient-Push is guaranteed to converge to optimality at a sublinear rate of $O(\ln k/\sqrt{k})$ for unconstrained, nonsmooth problems with bounded subgradients [12]. In addition, when the problem is unconstrained, strongly convex, and smooth, an $O(\ln k/k)$ rate is established for Gradient-Push [13], and linear rates are provided for DIGing and push-DIGing [14].

Another standard approach is to utilize dual decomposition techniques, which often lead to a dual problem with a decomposable structure, so that it can be solved in a distributed fashion by classic optimization methods including the gradient projection method, the accelerated gradient methods, the method of multipliers, and their variants (e.g., [2], [3], [15]–[24]). Compared with the aforementioned consensus-based primal methods, many distributed dual/primal-dual algorithms can handle more complicated coupling constraints, yet still manage to achieve sublinear rates of convergence to dual and primal optimality when the dual function is smooth, and achieve linear rates when the dual function is also strongly concave. Despite this advantage, most of such methods require a fixed network topology. Although the primal-dual subgradient methods in [19], the primal-dual perturbation method in [21], and the proximal-minimization-based method in [24] cope with time-varying agent networks, they only guarantee asymptotic convergence to optimality and no results on convergence rate are provided. In addition to the above two approaches, there are other lines of research on distributed optimization, including incremental

optimization methods (e.g., [1], [25], [26]), distributed Newton methods (e.g., [27]–[29]), and continuous-time distributed optimization algorithms (e.g., [30]–[32]).

This paper is motivated by the lack of distributed optimization algorithms in the literature that are able to address *constrained* convex multi-agent optimization at a guaranteed *convergence rate* over *time-varying* networks. We propose, in this paper, a family of distributed Fenchel dual gradient methods that are able to solve a class of constrained multi-agent optimization problems at sublinear rates on time-varying undirected networks, where the local objectives of the agents are strongly convex but not necessarily differentiable and the global constraint is the intersection of the nonidentical local convex constraints of the agents.

To develop such algorithms, we first derive the Fenchel dual of the multi-agent optimization problem, which consists of a separable, smooth dual function and a coupling linear constraint. Additionally, the gradient of the Fenchel dual function can be evaluated in parallel by the agents. We then utilize a class of weighted gradient methods to solve the Fenchel dual problem, which can be implemented over time-varying networks in a distributed fashion and can be viewed as a generalization of the distributed weighted gradient methods in [33], [34]. We show that the proposed Fenchel dual gradient algorithms asymptotically converge to both dual and primal optimality if the agents and their infinitely occurring interactions form a connected graph. We also show that the dual optimality is reached at an $O(1/k)$ rate and the primal optimality is achieved at an $O(1/\sqrt{k})$ rate if the underlying agent interaction graph during every B iterations is connected. Besides, the Fenchel dual gradient methods are shown to be more scalable to the network size and the number B of iterations to reach connectivity than the existing algorithms [12]–[14] that also have guaranteed convergence rates on time-varying networks. Further, we adapt the Fenchel dual gradient methods to the case where the Lipschitz constants of the dual gradients, which are used in algorithm parameter selection, are completely unknown, and obtain convergence rates on the same order. To reduce computational costs, we also allow the dual gradients to be inaccurately computed within an error ϵ, and prove that such inexact Fenchel dual gradient methods converge at an $O(1/\sqrt{k})$ rate to a suboptimal solution that is $O(\sqrt{\epsilon})$ away from primal optimality. Finally, the efficacy of the Fenchel dual gradient methods is illustrated via simulations.

The outline of the paper is as follows: Section II formulates the multi-agent optimization problem, and Section III develops the distributed Fenchel dual gradient methods. Section IV establishes the convergence results of the proposed algorithms. Section V involves the alternative

algorithms when the dual gradient information is imperfect. Section VI presents simulation results, and Section VII concludes the paper. All the proofs are included in the appendix. A 6-page conference version of this paper can be found in [35], which does not contain Sections IV-A, IV-C, V, VI-B, and the appendix.

Throughout the paper, we use $\|\cdot\|$ to represent the Euclidean norm and $\|\cdot\|_1$ the ℓ_1 norm. For any set $X \subseteq \mathbb{R}^d$, $\operatorname{int} X$ represents its interior and $|X|$ its cardinality. Let $P_X(x) = \arg\min_{y\in X} \|x-y\|$ denote the projection of $x \in \mathbb{R}^d$ onto X, which uniquely exists if X is closed and convex. The ball centered at $x \in \mathbb{R}^d$ with radius $r > 0$ is denoted by $B(x, r) := \{y \in \mathbb{R}^d : \|y - x\| \le r\}$. The floor of a real number is represented by $\lfloor\cdot\rfloor$. For any $\mathbf{x} \in \mathbb{R}^{nd}$, $\mathbf{x} = (x_1^T, \ldots, x_n^T)^T$ means the even partition of $\mathbf{x}$ into n blocks, i.e., $x_i \in \mathbb{R}^d$ $\forall i = 1, \ldots, n$. For any function $f : \mathbb{R}^d \to \mathbb{R}$, $\partial f(x)$ denotes any subgradient of f at $x \in \mathbb{R}^d$, i.e., $f(y) - f(x) - \partial f(x)^T(y - x) \ge 0$ $\forall y \in \mathbb{R}^d$. If f is differentiable, then $\nabla f(x)$ denotes the gradient of f at $x \in \mathbb{R}^d$. In addition, I_d is the $d \times d$ identity matrix, O_d is the $d \times d$ 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. For any matrices $M, M' \in \mathbb{R}^{n\times n}$, $M \preceq M'$ and $M' \succeq M$ both mean $M' - M$ is positive semidefinite. Also, $[M]_{ij}$ represents the (i, j)-entry of M, $\mathcal{R}(M)$ the range of M, and $\operatorname{Null}(M)$ the null space of M. If M is a block diagonal matrix with diagonal blocks $M_1, \ldots, M_m$, we write it as $M = \operatorname{diag}(M_1, \ldots, M_m)$. If M is symmetric positive semidefinite, we use $\lambda_i^{\downarrow}(M) \ge 0$ to denote its ith largest eigenvalue and $M^\dagger$ its Moore-Penrose pseudoinverse.

II. Problem Formulation

Consider a set $\mathcal{V} = \{1, 2, \ldots, n\}$ of agents, where each agent $i \in \mathcal{V}$ possesses a local objective function $f_i : \mathbb{R}^d \to \mathbb{R}$ and a local constraint set $X_i \subseteq \mathbb{R}^d$. All of the $n \ge 2$ agents attempt to solve the constrained optimization problem

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

which satisfies the following assumption.

Assumption 1. *(a) Each f_i, $i \in \mathcal{V}$ is strongly convex over X_i with convexity parameter $\theta_i > 0$, i.e., for any $x, y \in X_i$ and any subgradient $\partial f_i(x)$ of f_i at x, $f_i(y) - f_i(x) - \partial f_i(x)^T(y - x) \ge \theta_i\|y - x\|^2/2$.*
(b) $\mathbf{0}_d \in \operatorname{int} \bigcap_{i\in\mathcal{V}} X_i$.

Assumption 1 ensures the existence of a unique optimal solution $x^\star \in \bigcap_{i\in\mathcal{V}} X_i$ to problem (1). Notice that Assumption 1(a) is a common assumption for distributed optimization methods with convergence rate guarantees (e.g., [2], [3], [13], [14], [20], [22]). In addition, unlike many existing works that require each f_i to be continuously differentiable (e.g., [7]–[11], [13], [14], [17], [21], [22], [27]–[30], [32]), here each f_i is not necessarily differentiable. Also, Assumption 1(b) can always be replaced with the less restrictive condition $\operatorname{int}\bigcap_{i\in\mathcal{V}} X_i \neq \emptyset$, which is also assumed in [4]–[6], [24]. To see this, suppose $x' \in \operatorname{int}\bigcap_{i\in\mathcal{V}} X_i$ for some $x' \neq \mathbf{0}_d$. Consider the change of variable $z = x - x'$, and write each $f_i(x)$ and X_i as $f_i(z + x')$ and $\{z \in \mathbb{R}^d : z + x' \in X_i\}$, respectively. Then, the resulting new problem with the decision variable z is in the form of (1) and satisfies Assumption 1.

We model the n agents and their interactions as an undirected graph $\mathcal{G}^k = (\mathcal{V}, \mathcal{E}^k)$ with time-varying topologies, where $k \in \{0, 1, \ldots\}$ represents time, $\mathcal{V} = \{1, 2, \ldots, n\}$ is the set of nodes (i.e., the agents), and $\mathcal{E}^k \subseteq \{\{i, j\} : i, j \in \mathcal{V}, i \neq j\}$ is the set of links (i.e., the agent interactions) at time k. Without loss of generality, we assume that $\mathcal{E}^k \neq \emptyset\ \forall k \geq 0$. In addition, for each node $i \in \mathcal{V}$, we use $\mathcal{N}_i^k = \{j \in \mathcal{V} : \{i, j\} \in \mathcal{E}^k\}$ to denote the set of its neighbors (i.e., the nodes that it directly communicates with) at time k.

To enable cooperation of the nodes, we need to impose an assumption on network connectivity, so that the local decisions of the nodes can be mixed across the network. To this end, define $\mathcal{E}_\infty := \{\{i, j\} : \{i, j\} \in \mathcal{E}^k \text{ for infinite many } k \geq 0\}$. Then, consider the following assumption.

Assumption 2 (Infinite connectivity)**.** *The graph $(\mathcal{V}, \mathcal{E}_\infty)$ is connected.*

Assumption 2 is equivalent to the connectivity of the graph $(\mathcal{V}, \cup_{t=k}^{\infty}\mathcal{E}^t)$ for all $k \geq 0$. This is a minimal connectivity condition for distributed optimization algorithms to converge to optimality, which ensures every node to directly or indirectly influence any other nodes infinitely many times [4]. As Assumption 2 does not quantify how quickly the local decisions of the nodes diffuse throughout the network, we need a stronger connectivity condition in order to derive performance guarantees for the algorithms to be developed.

Assumption 3 (B-connectivity)**.** *There exists an integer $B > 0$ such that for any integer $k \geq 0$, the graph $(\mathcal{V}, \bigcup_{t=kB}^{(k+1)B-1} \mathcal{E}^t)$ is connected.*

Assumption 3 forces each node to have an impact on the others in the time intervals $[kB, (k+1)B-1]\ \forall k \geq 0$ of length B. Compared with Assumption 2, Assumption 3 is more restrictive but

more commonly adopted in the literature (e.g., [4]–[6], [12]–[14], [19], [21], [24], [26], [32]).

III. Fenchel Dual Gradient Algorithms

In this section, we develop a family of distributed algorithms to solve problem (1) based on Fenchel duality.

A. Fenchel Dual Problem

We first transform (1) into the following equivalent problem:

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^{nd}}{\text{minimize}} \quad & F(\mathbf{x}) := \textstyle\sum_{i\in\mathcal{V}} f_i(x_i) \\ \text{subject to} \quad & x_i \in X_i, \quad \forall i \in \mathcal{V}, \\ & \mathbf{x} \in S, \end{aligned} \tag{2}$$

where $\mathbf{x} = (x_1^T, \ldots, x_n^T)^T$ and $S := \{\mathbf{x} \in \mathbb{R}^{nd} : x_1 = x_2 = \cdots = x_n\}$. Note that problem (2) has a unique optimal solution $\mathbf{x}^\star = ((x^\star)^T, \ldots, (x^\star)^T)^T$, where $x^\star \in \bigcap_{i\in\mathcal{V}} X_i$ is the unique optimum of problem (1). In addition, its optimal value $F^\star$ is equal to that of problem (1).

Next, we construct the Fenchel dual problem [36] of (2). To this end, we introduce a function $q_i : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ for each $i \in \mathcal{V}$ defined as

$$q_i(x_i, w_i) = w_i^T x_i - f_i(x_i).$$

The conjugate convex function $d_i : \mathbb{R}^d \to \mathbb{R}$ is then given by

$$d_i(w_i) = \sup_{x_i\in X_i} q_i(x_i, w_i).$$

With the above, the Fenchel dual problem of (2) can be described as

$$\begin{aligned} \underset{\mathbf{w}\in\mathbb{R}^{nd}}{\text{maximize}} \quad & -D(\mathbf{w}) := -\textstyle\sum_{i\in\mathcal{V}} d_i(w_i) \\ \text{subject to} \quad & \mathbf{w} \in S^\perp, \end{aligned} \tag{3}$$

where $\mathbf{w} = (w_1^T, \ldots, w_n^T)^T$ and $S^\perp := \{\mathbf{w} \in \mathbb{R}^{nd} : w_1 + w_2 + \cdots + w_n = \mathbf{0}_d\}$ is the orthogonal complement of S. Note that (3) is a convex optimization problem. Also, with Assumption 1, it can be shown that strong duality between (2) and (3) holds, i.e., the optimal value $-D^\star$ of (3) equals $F^\star$, and that the optimal set of (3) is nonempty [36]. Moreover, $\mathbf{w}^\star = ((w_1^\star)^T, \ldots, (w_n^\star)^T)^T \in S^\perp$ is an optimal solution to (3) if and only if $\nabla d_i(w_i^\star) = \nabla d_j(w_j^\star)\ \forall i, j \in \mathcal{V}$ [34, Lemma 3.1], i.e., $\nabla D(\mathbf{w}^\star) \in S$.

Below we acquire a couple of properties regarding the Fenchel dual problem (3). Notice from Assumption 1(a) that for each $i \in \mathcal{V}$ and each $w_i \in \mathbb{R}^d$, there uniquely exists

$$\tilde{x}_i(w_i) := \arg\max_{x \in X_i} q_i(x, w_i). \tag{4}$$

Thus, d_i is differentiable [37] and

$$\nabla d_i(w_i) = \tilde{x}_i(w_i). \tag{5}$$

The following proposition shows that d_i is smooth, i.e., ∇d_i is Lipschitz.

Proposition 1. *[2, Lemma II.1] Suppose Assumption 1 holds. Then, for each $i \in \mathcal{V}$, ∇d_i is Lipschitz continuous with Lipschitz constant $L_i = 1/\theta_i$, where $\theta_i > 0$ is defined in Assumption 1, i.e., $\|\nabla d_i(u_i) - \nabla d_i(v_i)\| \leq L_i \|u_i - v_i\|$ $\forall u_i, v_i \in \mathbb{R}^d$.*

In fact, the strong convexity of f_i on X_i assumed in Assumption 1(a) is both sufficient and necessary for the smoothness of d_i [2].

Likewise, we can see that $D(\mathbf{w})$ is differentiable and

$$\nabla D(\mathbf{w}) = \tilde{\mathbf{x}}(\mathbf{w}) := (\tilde{x}_1(w_1)^T, \ldots, \tilde{x}_n(w_n)^T)^T. \tag{6}$$

According to (4) and (6), if each w_i is known to node i, then the gradient of the Fenchel dual function D can be evaluated in parallel by the nodes, while the Lagrange dual of (equivalent forms of) problem (2) does not have such a favorable feature when the network is time-varying and not necessarily connected at each time instance. Further, notice that $F(\mathbf{x})$ in problem (2) is strongly convex over $X_1 \times \cdots \times X_n$ with convexity parameter $\theta_{\min} := \min_{i \in \mathcal{V}} \theta_i$. Also note that $D(\mathbf{w}) = \sup_{\mathbf{x} \in X_1 \times \cdots \times X_n} \mathbf{w}^T \mathbf{x} - F(\mathbf{x})$. Like Proposition 1, we can establish the Lipschitz continuity of ∇D.

Corollary 1. *Suppose Assumption 1 holds. Then, ∇D is Lipschitz continuous with Lipschitz constant $L = 1/\theta_{\min}$.*

Finally, we show that the dual optimal set and the level sets of D on $S^\perp$ are bounded.

Proposition 2. *Suppose Assumption 1 holds. For any optimal solution $\mathbf{w}^\star \in S^\perp$ of problem* (3)*,*

$$\|\mathbf{w}^\star\| \leq \frac{(\sum_{i \in \mathcal{V}} \max_{x_i \in B(\mathbf{0}_d, r_c)} f_i(x_i)) - F^\star}{r_c} < \infty, \tag{7}$$

where $r_c \in (0, \infty)$ is such that $B(\mathbf{0}_d, r_c) \subseteq \bigcap_{i \in \mathcal{V}} X_i$. In addition, for any $\mathbf{w} \in S^\perp$, the level set $S_0(\mathbf{w}) := \{\mathbf{w}' \in S^\perp : D(\mathbf{w}') \leq D(\mathbf{w})\}$ is compact.

Proof. See Appendix A. □

The boundedness of the dual optimal set relies on the nonemptyness of $\operatorname{int} \bigcap_{i\in\mathcal{V}} X_i$ assumed by Assumption 1(b), without which the dual optimal set can be unbounded (e.g., $X_i = \{(z_1, z_2)^T \in \mathbb{R}^2 : z_1 = 0\}\ \forall i \in \mathcal{V}$).

B. Algorithms

In [33], [34], a set of weighted gradient methods are proposed to solve a network resource allocation problem, which can be cast in the form of (3). Inspired by this, we consider a class of weighted gradient methods as follows: Starting from an arbitrary $\mathbf{w}^0 \in S^\perp$, the subsequent iterates are generated by

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \alpha^k (H_{\mathcal{G}^k} \otimes I_d) \nabla D(\mathbf{w}^k), \quad \forall k \geq 0, \tag{8}$$

where $\alpha^k > 0$ is the step-size and $H_{\mathcal{G}^k} \in \mathbb{R}^{n\times n}$ is the weight matrix that depends on the topology of $\mathcal{G}^k$, defined as

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

We require $h_{ij}^k = h_{ji}^k > 0\ \forall \{i,j\} \in \mathcal{E}^k\ \forall k \geq 0$. We also assume that there exists a finite interval $[\underline{h}, \bar{h}]$ such that

$$h_{ij}^k \in [\underline{h}, \bar{h}] \subset (0, \infty), \quad \forall k \geq 0,\ \forall i \in \mathcal{V},\ \forall j \in \mathcal{N}_i^k. \tag{10}$$

Since $\mathcal{E}^k \neq \emptyset$, $H_{\mathcal{G}^k} \neq O_n$ for any $k \geq 0$. Moreover, $H_{\mathcal{G}^k}$ is symmetric positive semidefinite and $H_{\mathcal{G}^k}\mathbf{1}_n = \mathbf{0}_n$. Thus, using the same rationale as [33], [34], the proposition below shows that as long as $\mathbf{w}^0$ is feasible, so are $\mathbf{w}^k\ \forall k \geq 1$.

Proposition 3. *Let* $(\mathbf{w}^k)_{k=0}^\infty$ *be the iterates generated by* (8)*. If* $\mathbf{w}^0 \in S^\perp$*, then* $(\mathbf{w}^k)_{k=0}^\infty \subseteq S^\perp$*.*

Remark 1. *The weighted gradient method* (8) *can be tuned to solve problems of minimizing* $\sum_{i\in\mathcal{V}} d_i(w_i)$ *subject to* $\sum_{i\in\mathcal{V}} w_i = c,\ \forall c \in \mathbb{R}^d$*. To do so, we can simply replace the initial condition* $\mathbf{w}^0 \in S^\perp$ *with* $\sum_{i\in\mathcal{V}} w_i^0 = c$*.*

Next, we introduce primal iterates to the weighted gradient method (8) that is intended for the Fenchel dual problem (3). Note from (9) and (6) that (8) can be written as

$$x_i^k = \tilde{x}_i(w_i^k), \quad \forall i \in \mathcal{V},$$
$$w_i^{k+1} = w_i^k - \alpha^k \sum_{j \in \mathcal{N}_i^k} h_{ij}^k (x_i^k - x_j^k), \quad \forall i \in \mathcal{V},$$

where $w_i^k \in \mathbb{R}^d$ is the ith d-dimensional block of $\mathbf{w}^k$ and $\tilde{x}_i(w_i^k)$ is defined in (4). We assign each w_i^k and x_i^k to node i as its dual and primal iterates, with x_i^k being node i's estimate on the optimal solution $x^\star$ of problem (1). Thus, the above algorithm with both dual and primal iterates can be implemented in a distributed and possibly asynchronous way on the time-varying network, as is shown in Algorithm 1.

Algorithm 1 Fenchel Dual Gradient Method

1: **Initialization:** Each node $i \in \mathcal{V}$ selects $w_i^0 \in \mathbb{R}^d$ so that $\sum_{j\in\mathcal{V}} w_j^0 = \mathbf{0}_d$ (or simply sets $w_i^0 = \mathbf{0}_d$), and sets $x_i^0 = \arg\max_{x\in X_i} (w_i^0)^T x - f_i(x)$.
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 to all $j \in \mathcal{N}_i^k$.
4: Upon receiving x_j^k $\forall j \in \mathcal{N}_i^k$, each node $i \in \mathcal{V}$ with $\mathcal{N}_i^k \neq \emptyset$ updates $w_i^{k+1} = w_i^k - \alpha^k \sum_{j\in\mathcal{N}_i^k} h_{ij}^k (x_i^k - x_j^k)$.
5: Each node $i \in \mathcal{V}$ with $\mathcal{N}_i^k \neq \emptyset$ computes $x_i^{k+1} = \arg\max_{x\in X_i} (w_i^{k+1})^T x - f_i(x)$.
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**

In Algorithm 1, the initial condition $\mathbf{w}^0 \in S^\perp$, i.e., $\sum_{j\in\mathcal{V}} w_j^0 = \mathbf{0}_d$, can simply be realized by setting $w_i^0 = \mathbf{0}_d$ $\forall i \in \mathcal{V}$. Subsequently at each iteration, every node i with at least one neighbor updates its dual iterate w_i^k via local interactions with its current neighbors and then updates its primal iterate x_i^k on its own.

To implement Algorithm 1, each node i needs to select the weights h_{ij}^k $\forall j \in \mathcal{N}_i^k$ that satisfy $h_{ij}^k = h_{ji}^k$ in a predetermined interval $[\underline{h}, \bar{h}] \subset (0, \infty)$, where $\underline{h}$ and $\bar{h}$ may or may not be related with $\mathcal{G}^k$ $\forall k \geq 0$. This can be done through inexpensive interactions between neighboring nodes. Two typical examples of $H_{\mathcal{G}^k}$ are the graph Laplacian matrix

$$[H_{\mathcal{G}^k}]_{ij} = [L_{\mathcal{G}^k}]_{ij} := \begin{cases} |\mathcal{N}_i^k|, & \text{if } i = j, \\ -1, & \text{if } \{i, j\} \in \mathcal{E}^k, \\ 0, & \text{otherwise,} \end{cases} \tag{11}$$

and the Metropolis weight matrix [33]

$$[H_{\mathcal{G}^k}]_{ij} = \begin{cases} \sum_{s\in\mathcal{N}_i^k} \frac{1}{\max\{|\mathcal{N}_i^k|L_i, |\mathcal{N}_s^k|L_s\}}, & \text{if } i=j, \\ -\frac{1}{\max\{|\mathcal{N}_i^k|L_i, |\mathcal{N}_j^k|L_j\}}, & \text{if } \{i,j\}\in\mathcal{E}^k, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

When $H_{\mathcal{G}^k}$ is set to (11), each node i does not need any additional efforts in computing the weights h_{ij}^k $\forall j \in \mathcal{N}_i^k$ since they are 1 by default. When $H_{\mathcal{G}^k}$ is set to (12), each node i only needs to obtain from every neighbor $j \in \mathcal{N}_i^k$ the product of node j's neighborhood size $|\mathcal{N}_j^k|$ and Lipschitz constant $L_j = 1/\theta_j$ of ∇d_j.

The remaining parameter to be determined is the step-size α^k. Later in Section IV, we will show that the following step-size condition is sufficient to guarantee the convergence of Algorithm 1: Suppose there is a finite interval $[\underline{\alpha}, \bar{\alpha}]$ such that

$$\alpha^k \in [\underline{\alpha}, \bar{\alpha}] \subset (0, 2/\delta), \quad \forall k \geq 0, \tag{13}$$

where $\delta > 0$ can be any positive constant satisfying

$$H_{\mathcal{G}^k} \preceq \delta \Lambda_L^{-1}, \quad \forall k \geq 0, \tag{14}$$

with $\Lambda_L := \mathrm{diag}(L_1, \ldots, L_n)$. Note that such δ always exists because Λ_L^{-1} is positive definite and $H_{\mathcal{G}^k}$ is positive semidefinite. For example, we may choose $\delta = L \sup_{k\geq 0} \lambda_1^{\downarrow}(H_{\mathcal{G}^k})$, where $L = 1/\theta_{\min} = \max_{i\in\mathcal{V}} L_i$. More conservatively, because $H_{\mathcal{G}^k} \preceq \bar{h} L_{\mathcal{G}^k}$ and $\lambda_1^{\downarrow}(L_{\mathcal{G}^k}) \leq n$, we can always let $\delta = L\bar{h}n$ and thus

$$[\underline{\alpha}, \bar{\alpha}] \subset (0, \frac{2}{L\bar{h}n}).$$

Since $\bar{h}$ can be predetermined and known to all the nodes, this condition only requires the nodes to obtain the global quantities n and $L = \max_{i\in\mathcal{V}} L_i$, which can be computed decentralizedly by some consensus schemes (e.g., [38]). Below, we provide less conservative step-size conditions for the two specific choices of $H_{\mathcal{G}^k}$ in (11) and (12), which also can be satisfied by the nodes without any centralized coordination.

Example 1. *When $H_{\mathcal{G}^k}$ is set to the graph Laplacian matrix $L_{\mathcal{G}^k}$ as in* (11)*, in addition to the aforementioned choice $\delta = L \sup_{k\geq 0} \lambda_1^{\downarrow}(L_{\mathcal{G}^k})$, another option for δ could be $\delta = 2 \sup_{k\geq 0} \max_{i\in\mathcal{V}} |\mathcal{N}_i^k| L_i$,*

so that $\delta\Lambda_L^{-1} - L_{\mathcal{G}^k}$ *is diagonally dominant and thus positive semidefinite for each* $k \geq 0$. *Therefore,* α^k *can be selected in the interval* $[\underline{\alpha}, \bar{\alpha}]$ *satisfying*

$$0 < \underline{\alpha} \leq \bar{\alpha} < \frac{1}{\min\{\frac{L}{2}\sup\limits_{k\geq 0}\lambda_1^{\downarrow}(L_{\mathcal{G}^k}), \sup\limits_{k\geq 0}\max\limits_{i\in\mathcal{V}}|\mathcal{N}_i^k|L_i\}}.$$

The above step-size condition can be simplified for some special interaction patterns. For instance, if the nodes interact in a gossiping pattern, i.e., each $\mathcal{E}^k$ *contains only one link, then we may let* $0 < \underline{\alpha} \leq \bar{\alpha} < 1/L$. *Even though the topologies of* $(\mathcal{G}^k)_{k=0}^{\infty}$ *are completely unknown, since* $\lambda_1^{\downarrow}(L_{\mathcal{G}^k}) \leq n$, *we can adopt a more conservative step-size condition* $0 < \underline{\alpha} \leq \bar{\alpha} < 2/(nL)$.

Example 2. *When* $H_{\mathcal{G}^k}$ *is set according to* (12), *we can simply take* $\delta = 2$, *because* $2\Lambda_L^{-1} - H_{\mathcal{G}^k}$ *is diagonally dominant and thus* $2\Lambda_L^{-1} \succeq H_{\mathcal{G}^k}$. *Hence, the step-sizes can be selected as*

$$0 < \underline{\alpha} \leq \alpha^k \leq \bar{\alpha} < 1, \quad \forall k \geq 0,$$

which requires no global information and is independent of the network and the problem.

The underlying weighted gradient method (8) in Algorithm 1 can be viewed as a generalization of the distributed weighted gradient methods in [33], [34]. By assuming the (directed) network to be time-invariant and connected, [33] proposes a class of weighted gradient methods in the form of (8) but with a constant weight matrix. It is also shown in [33] that if the time-invariant network is further undirected, the constant weight matrix can be determined in a distributed fashion via (11) or (12). The step-size conditions in [33] for fixed undirected networks and fixed weight matrices given by (11) and (12) are extended here in Examples 1 and 2 to handle time-varying networks and time-varying weight matrices. On the other hand, [34] considers time-varying undirected networks satisfying Assumption 3. By setting $H_{\mathcal{G}^k}$ to $L_{\mathcal{G}^k}$ in (11) and $\alpha^k = 1/(2nL)$ $\forall k \geq 0$, (8) reduces to the algorithm in [34]. Note from Example 1 that here we allow for a much broader step-size range for this particular weight matrix.

IV. Convergence Analysis

This section is dedicated to analyzing the convergence performance of Algorithm 1.

A. Asymptotic convergence under infinite connectivity

In this subsection, we show that Algorithm 1 asymptotically converges to the optimum of problem (1) under Assumption 2.

We first show that the step-size condition (13) ensures $(D(\mathbf{w}^k))_{k=0}^{\infty}$ to be non-increasing.

Lemma 1. *Suppose Assumption 1 holds. Let $(\mathbf{w}^k)_{k=0}^{\infty}$ be the dual iterates generated by Algorithm 1. If the step-sizes $(\alpha^k)_{k=0}^{\infty}$ satisfy* (13)*, then for each $k \geq 0$,*

$$D(\mathbf{w}^{k+1}) - D(\mathbf{w}^k) \leq -\rho \nabla D(\mathbf{w}^k)^T (H_{\mathcal{G}^k} \otimes I_d) \nabla D(\mathbf{w}^k),$$

where $\rho := \min\{\underline{\alpha} - \frac{\underline{\alpha}^2 \delta}{2}, \bar{\alpha} - \frac{\bar{\alpha}^2 \delta}{2}\} \in (0, \infty)$, with $\underline{\alpha}, \bar{\alpha} > 0$ in (13) *and $\delta > 0$ in* (14)*.*

Proof. See Appendix B. □

Lemma 1, along with Propositions 2 and 3, implies that for each $k \geq 0$, $\mathbf{w}^k \in S_0(\mathbf{w}^0)$ and $\|\mathbf{w}^k - \mathbf{w}^\star\| \leq M_0$, where $\mathbf{w}^\star$ is any optimum of problem (3) and

$$M_0 := \max_{\mathbf{w} \in S_0(\mathbf{w}^0),\ \mathbf{w}^\star \in S^\perp : D(\mathbf{w}^\star) = D^\star} \|\mathbf{w} - \mathbf{w}^\star\| \in [0, \infty). \tag{15}$$

Another important consequence of Lemma 1 is that the differences of the primal iterates along the time-varying links are vanishing. To see this, by adding the inequality in Lemma 1 from $k = 0$ to ∞, we obtain

$$\begin{aligned} \sum_{k=0}^{\infty} \langle \mathbf{x}^k, (H_{\mathcal{G}^k} \otimes I_d) \mathbf{x}^k \rangle &= \sum_{k=0}^{\infty} \langle \nabla D(\mathbf{w}^k), (H_{\mathcal{G}^k} \otimes I_d) \nabla D(\mathbf{w}^k) \rangle \\ &\leq (D(\mathbf{w}^0) - D^\star)/\rho < \infty, \end{aligned}$$

where $\mathbf{x}^k = ((x_1^k)^T, \ldots, (x_n^k)^T)^T$. This implies that $\langle \mathbf{x}^k, (H_{\mathcal{G}^k} \otimes I_d) \mathbf{x}^k \rangle \to 0$ as $k \to \infty$. Since $\langle \mathbf{x}^k, (H_{\mathcal{G}^k} \otimes I_d) \mathbf{x}^k \rangle = \sum_{\{i,j\} \in \mathcal{E}^k} h_{ij}^k \|x_i^k - x_j^k\|^2$ and $h_{ij}^k \geq \underline{h} > 0\ \forall \{i,j\} \in \mathcal{E}^k$, we have

$$\lim_{k \to \infty} \max_{\{i,j\} \in \mathcal{E}^k} \|x_i^k - x_j^k\| = 0. \tag{16}$$

Because $\mathcal{G}^k$ may not be connected at each $k \geq 0$, (16) alone is insufficient to assert that the primal iterates $x_i^k\ \forall i \in \mathcal{V}$ asymptotically reach a consensus. Nevertheless, by integrating (16) with Assumption 2, we are able to show in Lemma 2 below that such an assertion is indeed true. The main idea of proving this can be summarized as follows: By (16) we know that $\|x_i^k - x_j^k\|$ $\forall \{i,j\} \in \mathcal{E}^k$ can be arbitrarily small after some time $T \geq 0$. Then, instead of studying the differences $\|x_i^k - x_j^k\|\ \forall i, j \in \mathcal{V}$ across the entire network, we show that such differences within each connected component of the graph $(\mathcal{V}, \cup_{t=T}^k \mathcal{E}^t)$ become sufficiently small after some $k \geq T$. Finally, note from Assumption 2 that the graph $(\mathcal{V}, \cup_{t=T}^k \mathcal{E}^t)$ must be connected when $k \geq T$ is sufficiently large. The dissipation of the differences among all the x_i^k's can thus be concluded.

Lemma 2. *Suppose Assumptions 1 and 2 hold. Let* $(\mathbf{x}^k)_{k=0}^{\infty}$ *be the primal iterates generated by Algorithm 1. If the step-sizes* $(\alpha^k)_{k=0}^{\infty}$ *satisfy* (13)*, then* $\lim\limits_{k\to\infty}\max\limits_{i,j\in\mathcal{V}}\|x_i^k - x_j^k\| = 0$.

Proof. See Appendix C. □

Since $x_i^k \in X_i\ \forall i \in \mathcal{V}$, $\mathbf{x}^k$ is feasible if and only if $\mathbf{x}^k \in S$. Thus, $\|P_{S^\perp}(\mathbf{x}^k)\|$ can be used to quantify the infeasibility of $\mathbf{x}^k$. Note that $\|P_{S^\perp}(\mathbf{x}^k)\|^2 = \|\mathbf{x}^k - P_S(\mathbf{x}^k)\|^2 = \sum_{i\in\mathcal{V}}\|x_i^k - \frac{1}{n}\sum_{j\in\mathcal{V}}x_j^k\|^2 \leq \frac{1}{n}\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{V}}\|x_i^k - x_j^k\|^2$. It follows from Lemma 2 that $\|P_{S^\perp}(\mathbf{x}^k)\|^2 \to 0$ as $k \to \infty$. This can further be utilized to establish the asymptotic convergence to both dual and primal optimality, as is shown in the theorem below.

Theorem 1. *Suppose Assumptions 1 and 2 hold. Let* $(\mathbf{w}^k)_{k=0}^{\infty}$ *and* $(\mathbf{x}^k)_{k=0}^{\infty}$ *be the dual and primal iterates generated by Algorithm 1, respectively. If the step-sizes* $(\alpha^k)_{k=0}^{\infty}$ *satisfy* (13)*, then* $\lim_{k\to\infty}\|P_{S^\perp}(\mathbf{x}^k)\| = 0$*,* $\lim_{k\to\infty} D(\mathbf{w}^k) = D^\star$*,* $\lim_{k\to\infty} F(\mathbf{x}^k) = F^\star$*, and* $\lim_{k\to\infty}\mathbf{x}^k = \mathbf{x}^\star$.

Proof. See Appendix D. □

B. Convergence rates under B-connectivity

In this subsection, we offer sublinear rates of convergence for Algorithm 1 under Assumption 3.

Inspired from [34], we first provide a bound on the accumulative drop in the value of D over each time interval $[tB, (t+1)B-1]$, $t \in \{0, 1, \ldots\}$, which depends only on the dual iterate at time tB and the underlying interaction graph during these B iterations. To this end, for each $k \geq 0$, let $\tilde{\mathcal{G}}^k = (\mathcal{V}, \tilde{\mathcal{E}}^k)$ be *any* spanning subgraph of $(\mathcal{V}, \bigcup_{t=k}^{k+B-1}\mathcal{E}^t)$, which, owing to Assumption 3, is chosen to be connected at $k \in \{0, B, 2B, \ldots\}$. Also let ϖ^k be the maximum degree of $\tilde{\mathcal{G}}^k$ and $\bar{\varpi} := \sup_{t\in\{0,1,\ldots\}}\varpi^{tB}$. Clearly, $1 \leq \varpi^{tB} \leq \bar{\varpi} \leq n-1\ \forall t \in \{0, 1, \ldots\}$.

Lemma 3. *Suppose Assumptions 1 and 3 hold. Let* $(\mathbf{w}^k)_{k=0}^{\infty}$ *be the dual iterates generated by Algorithm 1. If the step-sizes* $(\alpha^k)_{k=0}^{\infty}$ *satisfy* (13)*, then for each* $k \in \{0, B, 2B, \ldots\}$*,*

$$\sum_{t=k}^{k+B-1}\nabla D(\mathbf{w}^t)^T(H_{\mathcal{G}^t}\otimes I_d)\nabla D(\mathbf{w}^t) \geq \nabla D(\mathbf{w}^k)^T(L_{\tilde{\mathcal{G}}^k}\otimes I_d)\nabla D(\mathbf{w}^k)/\eta, \tag{17}$$

where $\eta := 3B\bar{\varpi}\bar{\alpha}^2\delta L + 3/\underline{h} \in (0, \infty)$*, with* $\bar{\alpha} > 0$ *in* (13)*,* $\delta > 0$ *in* (14)*,* $L > 0$ *in Corollary 1, and* $\underline{h} > 0$ *in* (10).

Proof. See Appendix E. □

When $H_{\mathcal{G}^k} = L_{\mathcal{G}^k}$ and $\alpha^k = 1/(2nL)$, [34, Lemma A.9] provides a similar bound to (17) with η replaced by $3B/2$ and $\tilde{\mathcal{G}}^k$ being a spanning tree. Lemma 3 improves this bound since $\eta \leq 3B/4 + 3$ for such a particular choice of $H_{\mathcal{G}^k}$ and α^k, allows for more general selections of $H_{\mathcal{G}^k}$ and α^k, and sheds light on how the network topologies come into play.

Lemma 1 and Lemma 3 together bound the decrease in the value of D during every B iterations, with which we are able to provide a rate for $D(\mathbf{w}^k) \to D^\star$. Prior to doing that, we define a sequence $(\tilde{M}_k)_{k=0}^\infty$ as follows: Let $\tilde{M}_0 \in \mathbb{R}$ be any positive constant and define

$$\tilde{M}_k = \max_{t=0,\ldots,k-1} \min_{\mathbf{w}^\star \in S^\perp : D(\mathbf{w}^\star) = D^\star} \|\mathbf{w}^{tB} - \mathbf{w}^\star\|, \ \forall k \geq 1. \tag{18}$$

Notice that $0 \leq \tilde{M}_k \leq M_0 < \infty$, where M_0 is given by (15).

Theorem 2. *Suppose Assumptions 1 and 3 hold. Let $(\mathbf{w}^k)_{k=0}^\infty$ be the dual iterates generated by Algorithm 1. If the step-sizes $(\alpha^k)_{k=0}^\infty$ satisfy* (13)*, then for each $k \geq 0$,*

$$D(\mathbf{w}^k) - D^\star \leq \frac{\eta \tilde{M}_{\lfloor k/B \rfloor}^2 (D(\mathbf{w}^0) - D^\star)}{\eta \tilde{M}_{\lfloor k/B \rfloor}^2 + \rho \underline{\lambda} (D(\mathbf{w}^0) - D^\star) \lfloor k/B \rfloor}, \tag{19}$$

where $\tilde{M}_{\lfloor k/B \rfloor} \in [0, M_0]$ is defined in (18) *with $M_0 \geq 0$ in* (15)*, $\underline{\lambda} := \inf_{t \in \{0,1,\ldots\}} \lambda_{n-1}^\downarrow(L_{\tilde{\mathcal{G}}^{tB}}) \in (0, \infty)$, and $\eta, \rho > 0$ are given in Lemma 3 and Lemma 1, respectively.*

Proof. See Appendix F. □

Theorem 2 says that Algorithm 1, or equivalently, the underlying weighted gradient method (8), converges to the optimal value $D^\star$ of problem (3) at an $O(1/k)$ rate. The derivation of this result requires each d_i to be smooth and the dual optimal set to be compact. These two conditions on problem (3) may not hold if Assumption 1 is not satisfied (cf. Section III-A). Note that without the compactness of the dual optimal set, (19) still holds, but we cannot guarantee $(\mathbf{w}^k)_{k=0}^\infty$ and thus $\tilde{M}_{\lfloor k/B \rfloor}$ $\forall k \geq 0$ to be bounded.

The distributed weighted gradient methods in [33], [34] also require the above two conditions on problem (3) to establish their convergence to $D^\star$. By imposing an additional assumption that the Hessian matrices of d_i $\forall i \in \mathcal{V}$ are positive definite, the methods in [33] are proved to achieve linear convergence rates on fixed networks. In contrast, Theorems 1 and 2 allow for time-varying networks and do not even require the existence of the Hessian matrices of d_i $\forall i \in \mathcal{V}$. The algorithm in [34] is shown to asymptotically drive $D(\mathbf{w}^k)$ to $D^\star$ and satisfy $\min_{t=1,\ldots,k} \|P_{S^\perp}(\nabla D(\mathbf{w}^{tB}))\|^2 \leq C \cdot n^3 B/k$ for some $C > 0$. Our results in Theorems 1 and 2

for the more general algorithm (8) are still stronger. We show that $\lim_{k\to\infty} D(\mathbf{w}^k) = D^\star$ under the less restrictive Assumption 2, and that $D(\mathbf{w}^k)$ converges to $D^\star$ at an $O(1/k)$ rate under Assumption 3. Also, since $\nabla D(\mathbf{w}^k) = \mathbf{x}^k$, the first inequality in Theorem 3 below is comparable to and slightly stronger than the aforementioned convergence rate in [34].

Based on Theorem 2, below we show that the primal errors $\|\mathbf{x}^k - \mathbf{x}^\star\|$ and $|F(\mathbf{x}^k) - F^\star|$ in optimality and $\|P_{S^\perp}(\mathbf{x}^k)\|$ in feasibility all converge to zero at rates of $O(1/\sqrt{k})$. Like many Lagrange dual gradient methods (e.g., [3], [37]), we do so by relating such primal errors with the dual error $D(\mathbf{w}^k) - D^\star$.

Theorem 3. *Suppose Assumptions 1 and 3 hold. Let $(\mathbf{x}^k)_{k=0}^\infty$ be the primal iterates generated by Algorithm 1. If the step-sizes $(\alpha^k)_{k=0}^\infty$ satisfy* (13)*, then for each $k \geq 0$,*

$$\|P_{S^\perp}(\mathbf{x}^k)\| \leq \|\mathbf{x}^k - \mathbf{x}^\star\| \leq \sqrt{\frac{2L\eta\tilde{M}^2_{\lfloor k/B\rfloor}(D(\mathbf{w}^0) - D^\star)}{\eta\tilde{M}^2_{\lfloor k/B\rfloor} + \rho\underline{\lambda}(D(\mathbf{w}^0) - D^\star)\lfloor k/B\rfloor}},$$

$$F(\mathbf{x}^k) - F^\star \leq \|\mathbf{w}^k\|\sqrt{\frac{2L\eta\tilde{M}^2_{\lfloor k/B\rfloor}(D(\mathbf{w}^0) - D^\star)}{\eta\tilde{M}^2_{\lfloor k/B\rfloor} + \rho\underline{\lambda}(D(\mathbf{w}^0) - D^\star)\lfloor k/B\rfloor}},$$

$$F(\mathbf{x}^k) - F^\star \geq -\|\mathbf{w}^\star\|\sqrt{\frac{2L\eta\tilde{M}^2_{\lfloor k/B\rfloor}(D(\mathbf{w}^0) - D^\star)}{\eta\tilde{M}^2_{\lfloor k/B\rfloor} + \rho\underline{\lambda}(D(\mathbf{w}^0) - D^\star)\lfloor k/B\rfloor}},$$

where $\mathbf{w}^\star$ is any optimal solution of problem (3)*, L is given in Corollary 1, and the remaining constants have been introduced in Theorem 2.*

Proof. See Appendix G. □

Since $\mathbf{w}^k \in S_0(\mathbf{w}^0)\ \forall k \geq 0$ and $S_0(\mathbf{w}^0)$ is compact, the term $\|\mathbf{w}^k\|$ that appears in the convergence rate of $F(\mathbf{x}^k) - F^\star$ is uniformly bounded above by $M_0 + \|\mathbf{w}^\star\|$. Consequently, the primal convergence rates of Algorithm 1 in Theorem 3 are all of order $O(1/\sqrt{k})$, which commensurate with the convergence rate of the classic (centralized) subgradient projection method [39].

C. Discussions on convergence rates

Theorems 2 and 3 reveal the joint impact of network topologies, algorithm parameters, and problem characteristics on the convergence rates of Algorithm 1. Further discussions regarding those convergence rates are as follows.

1) Scalability with respect to n and B: All the rates in Theorems 2 and 3 scale polynomially in the network size n and the number B of iterations to achieve connectivity.[1] This can be seen by noticing $\underline{\lambda} \geq \inf_{k\geq 0} 4/(n\,\mathrm{diam}(\tilde{\mathcal{G}}^{kB})) \geq 4/(n(n-1))$ [40] and $\bar{\varpi} \leq n-1$, where $\mathrm{diam}(\tilde{\mathcal{G}}^{kB}) \leq n-1$ is the diameter of $\tilde{\mathcal{G}}^{kB}$. In particular, when the weight matrices and the step-sizes are appropriately selected (e.g., $H_{\mathcal{G}^k}$ in (11), $\alpha^k = 1/(nL)$, or $H_{\mathcal{G}^k}$ in (12), $\alpha^k = 1/2$), it can be shown that in the worst case the dual rate in Theorem 2 grows with $O(n^3B^2)$ and the primal rates in Theorem 3 grow with $O(n^{1.5}B)$. Hence, the iteration complexity of Algorithm 1 to reach ϵ-accuracy in primal optimality and feasibility is $O(n^3B^2/\epsilon^2)$. Such an $O(n^3B^2)$ scalability regarding n and B may be improved if further information on the agent interactions is known. For instance, if $(\mathcal{V}, \bigcup_{t=kB}^{(k+1)B-1}\mathcal{E}^t)$ for each $k \geq 0$ has a spanning subgraph as a line or a ring, then $\bar{\varpi} = 2$ and thus the scalability reduces to $O(n^2B^2)$; If it contains certain types of spanning trees [41], then $\underline{\lambda}$ is bounded below by some positive constant that is independent of n and B, so that the iteration complexity is linear in n.

2) On-line estimation of primal and dual errors: We can derive more conservative convergence rate bounds than Theorems 2 and 3 in order to estimate the errors $D(\mathbf{w}^k) - D^\star$ in dual optimality, $\|\mathbf{x}^k - \mathbf{x}^\star\|$ and $|F(\mathbf{x}^k) - F^\star|$ in primal optimality, and $\|P_{S^\perp}(\mathbf{x}^k)\|$ in primal feasibility at every iteration $k \geq 0$. Suppose n and B (or their upper bounds) are known. Then, according to the discussions on the algorithm parameter selections in Section III-B and the bounds of $\underline{\lambda}$ and $\bar{\varpi}$ in Section IV-C1, we can figure out lower bounds on ρ and $\underline{\lambda}$ as well as an upper bound on η in terms of known quantities such as n, B, and the L_i's. Also, since $-D^\star = F^\star$, the term $D(\mathbf{w}^0) - D^\star$ can be replaced with $D(\mathbf{w}^0) + \sum_{i\in\mathcal{V}} f_i(x')$ for an arbitrary $x' \in \cap_{i\in\mathcal{V}} X_i$. Further, at each $k \geq B$, we may bound $\tilde{M}_{\lfloor k/B\rfloor}$ as $\tilde{M}_{\lfloor k/B\rfloor} \leq \max\limits_{t=0,\ldots,\lfloor k/B\rfloor-1} \|\mathbf{w}^{tB}\| + \|\mathbf{w}^\star\|$ for any dual optimum $\mathbf{w}^\star$. The first term on the right-hand side can be directly determined from the history of the dual iterates, and the second term $\|\mathbf{w}^\star\|$ can be bounded via (7), in which $-F^\star$ can be replaced by $D(\mathbf{w})$ for any $\mathbf{w} \in S^\perp$.

3) Comparison with related distributed optimization algorithms: We compare the primal convergence rates of Algorithm 1 with those of the existing distributed optimization algorithms that also have guaranteed *convergence rates* over *time-varying* networks, including Subgradient-Push [12], Gradient-Push [13], DIGing [14], and Push-DIGing [14]. Different from Algorithm 1 that is

[1]We exclude the pathological selections of $[\underline{h}, \bar{h}]$ in (10), δ in (14), and $[\underline{\alpha}, \bar{\alpha}]$ in (13) that make $1/\rho$ and η exponential in n and B, which, according to Section III-B, are unnecessary at all.

developed by applying distributed weighted gradient methods to the Fenchel dual, Subgradient-Push and Gradient-Push are constructed by incorporating the subgradient method and the stochastic gradient descent method into the Push-Sum consensus protocol [42], DIGing is designed by combining a distributed inexact gradient method with a gradient tracking technique, and Push-DIGing is derived by introducing Push-Sum into DIGing.

The convergence rates of the aforementioned algorithms are all established under Assumption 3.[2] For each of these algorithms, Table I lists its assumptions, convergence rate, and scalability of iteration complexity with respect to n and B. Observe that only Algorithm 1 is capable of solving problems with different local constraints of the agents, while the remaining algorithms all require the problem to be unconstrained and their extensions to constrained problems are still open challenges. Also, Gradient-Push, DIGing, and Push-DIGing require both strong convexity and smoothness of the f_i's, leading to faster convergence rates than the $O(1/\sqrt{k})$ rate of Algorithm 1. This is natural because we assume a weaker condition on f_i $\forall i \in \mathcal{V}$, which allows the strongly convex f_i's to be nonsmooth. Subgradient-Push needs neither strong convexity nor smoothness of each f_i, and the resulting convergence rate $O(\ln k/\sqrt{k})$ is slower than our $O(1/\sqrt{k})$ result. Note that the assumption on the f_i's for Algorithm 1 is not necessarily more restrictive than that for Subgradient-Push, since Subgradient-Push requires the subgradients of each f_i to be uniformly bounded over $\mathbb{R}^d$ but Algorithm 1 does not. Unlike Subgradient-Push, Gradient-Push, and Push-DIGing that admit directed links, DIGing and Algorithm 1 are only applicable to undirected graphs. With that said, Algorithm 1 is guaranteed to converge to the optimum with the minimal connectivity condition, i.e., Assumption 2, while the other methods have no such convergence results. The last column of Table I shows that the iteration complexities of Subgradient-Push, Gradient-Push, and Push-DIGing exhibit exponential dependence of n and B, yet those of DIGing and Algorithm 1 are polynomial in n and B, with $O(n^3B^2)$ for Algorithm 1 shown in Section IV-C1 the best.

V. Inexact Dual Gradient

The implementation of Algorithm 1 requires that the Lipschitz constant L_i of each ∇d_i is accessible to node i. Also, the convergence results in Section IV are established based on the exact

[2]When it comes to Subgradient-Push, Gradient-Push, and Push-DIGing, "connected" in Assumption 3 is indeed "strongly connected" since they consider directed networks.

Algorithm	unconstrained problem	strongly convex	Lipschitz gradient	bounded subgradient	undirected links	convergence rate	scalability w.r.t. n and B
Subgradient-Push [12]	$\surd$			$\surd$		$O(\ln k/\sqrt{k})$	$O(n^{2nB})$
Gradient-Push [13]	$\surd$	$\surd$	$\surd$			$O(\ln k/k)$	$O(n^{2nB})$
DIGing [14]	$\surd$	$\surd$	$\surd$		$\surd$	$O(q^k)$, $0<q<1$	$O(n^{4.5}B^3)$
Push-DIGing [14]	$\surd$	$\surd$	$\surd$			$O(q^k)$, $0<q<1$	$O(n^{n^2B^2})$
Algorithm 1		$\surd$			$\surd$	$O(1/\sqrt{k})$	$O(n^3B^2)$

TABLE I

Comparison of Algorithm 1 and related methods in assumptions, convergence rate, and scalability. Here, $\surd$ means the assumption is required.

evaluation of ∇d_i at each dual iterate w_i^k. In practice, these operations could be computationally expensive. In this section, we adapt Algorithm 1 to alleviate such issues.

A. Adaptive Fenchel dual gradient methods

We first consider the case where the Lipschitz constant L_i of each ∇d_i (or equivalently, the convexity parameter θ_i of each f_i) can hardly be determined by node i. This may cause difficulty in selecting proper algorithm parameters that guarantee the convergence of Algorithm 1.

To address this issue, below we propose an adaptive version of the Fenchel dual gradient methods. For each $i \in \mathcal{V}$ and $k \geq 0$, let $\hat{L}_i^k$ be node i's estimate on L_i at time k, whose initial value $\hat{L}_i^0$ can be arbitrarily set. At each $k \geq 0$, we determine $H_{\mathcal{G}^k}$ with L_i replaced by $\hat{L}_i^k$ if any L_i appears in the definition of $H_{\mathcal{G}^k}$ (e.g., (12)). We select the step-size α^k that satisfies the following condition:

$$\alpha^k \geq \underline{\alpha} > 0, \quad \alpha^k - 2/\delta^k \leq \sup_{t\geq 0}(\alpha^t - 2/\delta^t) < 0,$$

where $\delta^k > 0$ satisfies $H_{\mathcal{G}^k} \preceq \delta^k(\Lambda_{\hat{L}}^k)^{-1}$ with $\Lambda_{\hat{L}}^k := \mathrm{diag}(\hat{L}_1^k, \ldots, \hat{L}_n^k)$. This can be done similarly to Section III-B. With $H_{\mathcal{G}^k}$ and α^k selected based on $\hat{L}_i^k$ $\forall i \in \mathcal{V}$, each node $i \in \mathcal{V}$ with a nonempty neighborhood computes w_i^{k+1} and x_i^{k+1} in the same way as Algorithm 1. Then, it checks, on its own, whether the inequality

$$d_i(w_i^{k+1}) - d_i(w_i^k) - \langle \nabla d_i(w_i^k), w_i^{k+1} - w_i^k \rangle \leq \hat{L}_i^k \|w_i^{k+1} - w_i^k\|^2/2 \tag{20}$$

holds, where $\nabla d_i(w_i^k) = x_i^k$. If (20) does not hold, node i increases the value of $\hat{L}_i^k$. Then, the nodes repeat the above process of selecting $H_{\mathcal{G}^k}$ and α^k with the latest $\hat{L}_i^k$ $\forall i \in \mathcal{V}$ and computing the w_i^{k+1}'s and the x_i^{k+1}'s until (20) is satisfied for all $i \in \mathcal{V}$. Finally, set $\hat{L}_i^{k+1} = \hat{L}_i^k$ and move on to the next iteration $k+1$. Following the analysis in Section IV-B, it can be shown that under

Assumption 3, such adaptive Fenchel dual gradient methods achieve convergence rates on the same order as those of Algorithm 1 established in Theorems 2 and 3.

B. Inexact Fenchel dual gradient methods

In Algorithm 1, each node $i \in \mathcal{V}$ needs to maximize the concave function $q_i(\cdot, w_i^k)$ over X_i, in order to compute x_i^k, or equivalently, $\nabla d_i(w_i^k)$. This, however, could be very costly especially when f_i is nonsmooth and X_i has a complicated structure. To alleviate such computational efforts, below we develop and analyze inexact Fenchel dual gradient methods, where the dual gradients are computed within a certain error.

Start from any $\mathbf{w}^0 \in S^\perp$ (or simply set $w_i^0 = \mathbf{0}_d \ \forall i \in \mathcal{V}$). Then, at every $k \geq 0$, each node $i \in \mathcal{V}$ updates as follows:

$$\begin{aligned} &\hat{x}_i(w_i^k) \approx \arg\max_{x_i \in X_i} q_i(x_i, w_i^k),\ \hat{x}_i(w_i^k) \in X_i, \\ &w_i^{k+1} = w_i^k - \alpha^k \textstyle\sum_{j \in \mathcal{N}_i^k} h_{ij}^k (\hat{x}_i(w_i^k) - \hat{x}_j(w_j^k)). \end{aligned} \tag{21}$$

Same as Algorithm 1, here we still ensure $\mathbf{w}^k \in S^\perp \ \forall k \geq 0$. Different from Algorithm 1, (21) uses $\hat{x}_i(w_i^k) \in X_i$ as an approximation of $\tilde{x}_i(w_i^k) = \nabla d_i(w_i^k)$ to update the dual iterates. To quantify the error caused by each $\hat{x}_i(w_i^k)$, define

$$\bar{\epsilon} := \sup_{i \in \mathcal{V},\, k \geq 0} q_i(\tilde{x}_i(w_i^k), w_i^k) - q_i(\hat{x}_i(w_i^k), w_i^k) > 0.$$

In addition, a couple of useful inequalities in terms of such approximations are provided below.

Lemma 4. *Suppose Assumption 1 holds. For any $u_i, v_i \in \mathbb{R}^d$,*

$$\begin{aligned} &d_i(u_i) \leq d_i(v_i) + \langle \hat{x}_i(v_i), u_i - v_i \rangle + L_i \|u_i - v_i\|^2 + q_i(\tilde{x}_i(v_i), v_i) - q_i(\hat{x}_i(v_i), v_i), \\ &\frac{1}{2L_i} \|\hat{x}_i(v_i) - \tilde{x}_i(v_i)\|^2 \leq q_i(\tilde{x}_i(v_i), v_i) - q_i(\hat{x}_i(v_i), v_i). \end{aligned}$$

Proof. The lemma can be proved from [43, sec 3.2]. □

To establish the convergence results for the inexact Fenchel dual gradient methods that take the form of (21), we impose the following assumption on network connectivity.

Assumption 4. *There exists an integer $B > 0$ such that for any $k \geq 0$, the graph $(\mathcal{V}, \bigcup_{t=k}^{k+B-1} \mathcal{E}^t)$ is connected.*

Assumption 4 is very similar to but slightly more restrictive than Assumption 3, as the latter requires the connectivity of $(\mathcal{V}, \bigcup_{t=k}^{k+B-1} \mathcal{E}^t)$ only when k is a multiple of B. Further, due to

Assumption 4, we let $\tilde{\mathcal{G}}^k$ be any *connected* spanning subgraph of $(\mathcal{V}, \bigcup_{t=k}^{k+B-1} \mathcal{E}^t)$ with maximum degree $\varpi^k \in [1, n-1]$ throughout Section V-B.

We now provide counterparts of Lemmas 1 and 3 based on Lemma 4.

Lemma 5. *Suppose Assumption 1 holds. Let $(\mathbf{w}^k)_{k=0}^\infty$ be the dual iterates generated by Algorithm* (21)*. If*

$$\alpha^k \in [\underline{\alpha}, \bar{\alpha}] \in (0, 1/\delta), \quad \forall k \geq 0, \tag{22}$$

where δ is defined in (14)*, then for each $k \geq 0$,*

$$D(\mathbf{w}^{k+1}) - D(\mathbf{w}^k) \leq -\hat{\rho}\langle \hat{\mathbf{x}}(\mathbf{w}^k), (H_{\mathcal{G}^k} \otimes I_d)\hat{\mathbf{x}}(\mathbf{w}^k)\rangle + n\bar{\epsilon},$$

where $\hat{\rho} := \min\{\bar{\alpha} - \bar{\alpha}^2\delta, \underline{\alpha} - \underline{\alpha}^2\delta\} \in (0, \infty)$ and $\hat{\mathbf{x}}(\mathbf{w}^k) = (\hat{x}_1(w_1^k)^T, \ldots, \hat{x}_n(w_n^k)^T)^T \in X_1 \times \cdots \times X_n$.

Proof. See Appendix H. □

Lemma 6. *Suppose Assumptions 1 and 4 hold. Let $(\hat{\mathbf{x}}(\mathbf{w}^k))_{k=0}^\infty$ and $(\mathbf{w}^k)_{k=0}^\infty$ be the primal and dual iterates generated by Algorithm* (21)*. If the step-sizes $(\alpha^k)_{k=0}^\infty$ satisfy* (22)*, then*

$$\hat{\mathbf{x}}(\mathbf{w}^k)^T(L_{\tilde{\mathcal{G}}^k} \otimes I_d)\hat{\mathbf{x}}(\mathbf{w}^k) \leq \gamma_1\Big(\sum_{t=k}^{k+B-1} \hat{\mathbf{x}}(\mathbf{w}^t)^T(H_{\mathcal{G}^t} \otimes I_d)\hat{\mathbf{x}}(\mathbf{w}^t)\Big) + \gamma_2\bar{\epsilon}, \quad \forall k \geq 0,$$

where $\gamma_1 = 9B\hat{\varpi}\bar{\alpha}^2\delta L + 3/\underline{h} \in (0, \infty)$ and $\gamma_2 = 36\hat{\varpi}\sum_{i\in\mathcal{V}} L_i \in (0, \infty)$, with $\hat{\varpi} := \sup_{k\geq 0} \varpi^k \in [1, n-1]$.

Proof. See Appendix I. □

According to Lemma 5, with inexact dual gradients, we can no longer ensure the monotonicity of $(D(\mathbf{w}^k))_{k=0}^\infty$ and thus the boundedness of $(\mathbf{w}^k)_{k=0}^\infty$. However, if the error bound $\bar{\epsilon}$ is sufficiently small, we can utilize Lemmas 5 and 6 to show in the following lemma that $(\mathbf{w}^k)_{k=0}^\infty$ is bounded.

Lemma 7. *Suppose Assumptions 1 and 4 and the step-size condition* (22) *hold. Also suppose*

$$\sqrt{(\gamma_1 Bn/\hat{\rho} + \gamma_2)\bar{\epsilon}} \leq \sqrt{\lambda_a}\beta r_c \tag{23}$$

for some $\beta \in (0, 1)$, with $\lambda_a := \inf_{t\geq 0} \lambda_{n-1}^\downarrow(L_{\tilde{\mathcal{G}}^t}) \in (0, \infty)$, $\gamma_1, \gamma_2 > 0$ in Lemma 6, $\hat{\rho} > 0$ in Lemma 5, and $r_c \in (0, \infty)$ in Proposition 2. Let $(\mathbf{w}^k)_{k=0}^\infty$ be the dual iterates generated by Algorithm (21)*. Also let $\bar{D} := \max\{D(\mathbf{w}^0), \max\{D(\mathbf{w}) : \|\mathbf{w}\| \leq \frac{\bar{F} - \underline{F} + n\bar{\epsilon}}{r_c(1-\beta)},\ \mathbf{w} \in S^\perp\}\} + Bn\bar{\epsilon}$ and*

$\hat{M} := \max\{\|\mathbf{w}\| : D(\mathbf{w}) \leq \bar{D},\ \mathbf{w} \in S^{\perp}\} \in [0, \infty)$, *where* $\bar{F} = \sup\{F(\mathbf{x}) : x_i \in B(0, r_c)\ \forall i \in \mathcal{V}\} < \infty$ *and* $\underline{F} = \inf\{F(\mathbf{x}) : x_i \in X_i\ \forall i \in \mathcal{V}\} > -\infty$. *Then,* $\forall k \geq 0$, $D(\mathbf{w}^k) \leq \bar{D}$ *and* $\|\mathbf{w}^k\| \leq \hat{M}$.

Proof. See Appendix J. □

To guarantee the accuracy condition (23) for updating the $\hat{x}_i(w_i^k)$'s, every node needs to know r_c, λ_a, $\hat{\rho}$ (or their lower bounds), as well as B, n, γ_1, γ_2 (or their upper bounds). From the definitions of $\hat{\rho}$, γ_1, and γ_2, they indeed depend on B, $\underline{h}$, $\hat{\varpi}$, δ, $\underline{\alpha}$, $\bar{\alpha}$, L, and $\sum_{i\in\mathcal{V}} L_i$. We assume that B and $[\underline{h}, \bar{h}]$ in (10) are known *a priori* to the nodes. Similar to Section III-B, δ satisfying (14) and $[\underline{\alpha}, \bar{\alpha}]$ satisfying (22) can be easily determined in a distributed way. Also, we may adopt $\lambda_a \geq \frac{4}{n(n-1)}$ [40] and $\hat{\varpi} \leq n-1$ or better bounds on λ_a and $\hat{\varpi}$ if we know more about the graph sequence $(\mathcal{G}^k)_{k=0}^{\infty}$. Further, r_c can be taken as $\min_{i\in\mathcal{V}} r_i$, where $r_i > 0$ satisfying $B(\mathbf{0}_d, r_i) \subseteq X_i$ can be found by node i itself. The minimal r_i, along with the remaining constants (e.g., n, L, $\sum_{i\in\mathcal{V}} L_i$), can all be estimated in a decentralized manner with the help of certain consensus schemes (e.g., [38]).

Lemma 7 plays a vital role in establishing the following theorem, which provides convergence guarantees with respect to the running averages of the dual and primal iterates in (21), i.e.,

$$\bar{\mathbf{w}}^k := \frac{1}{k+1}\sum_{t=0}^{k} \mathbf{w}^t, \quad \bar{\mathbf{x}}^k := \frac{1}{k+1}\sum_{t=0}^{k} \hat{\mathbf{x}}(\mathbf{w}^t).$$

Theorem 4. *Suppose Assumptions 1 and 4 hold. Also suppose* (22) *and* (23) *are satisfied. Let* $(\hat{\mathbf{x}}(\mathbf{w}^k))_{k=0}^{\infty}$ *and* $(\mathbf{w}^k)_{k=0}^{\infty}$ *be the primal and dual iterates generated by Algorithm* (21)*. Then, for each* $k \geq 0$*,*

$$D(\bar{\mathbf{w}}^k) - D^\star \leq (\|\hat{N}_k\| + \|\mathbf{w}^\star\|)\sqrt{\frac{2\gamma_1 B(\bar{D} - D^\star)}{\hat{\rho}\lambda_a(k+1)}} + (\|\hat{N}_k\| + \|\mathbf{w}^\star\|)\sqrt{\Big(2\Big(\frac{\gamma_1 Bn}{\hat{\rho}\lambda_a} + \frac{\gamma_2}{\lambda_a}\Big) + 4\sum_{i\in\mathcal{V}} L_i\Big)\bar{\epsilon}}, \tag{24}$$

$$F(\bar{\mathbf{x}}^k) - F^\star \leq \hat{N}_k\sqrt{\frac{\gamma_1 B(\bar{D} - D^\star)}{\hat{\rho}\lambda_a(k+1)}} + \hat{N}_k\sqrt{\Big(\frac{\gamma_1 Bn}{\hat{\rho}\lambda_a} + \frac{\gamma_2}{\lambda_a}\Big)\bar{\epsilon}} + n\bar{\epsilon}, \tag{25}$$

$$F(\bar{\mathbf{x}}^k) - F^\star \geq -\|\mathbf{w}^\star\|\left(\sqrt{\frac{\gamma_1 B(\bar{D} - D^\star)}{\hat{\rho}\lambda_a(k+1)}} + \sqrt{\Big(\frac{\gamma_1 Bn}{\hat{\rho}\lambda_a} + \frac{\gamma_2}{\lambda_a}\Big)\bar{\epsilon}}\right), \tag{26}$$

$$\|P_{S^\perp}(\bar{\mathbf{x}}^k)\| \leq \sqrt{\frac{\gamma_1 B(\bar{D} - D^\star)}{\hat{\rho}\lambda_a(k+1)}} + \sqrt{\Big(\frac{\gamma_1 Bn}{\hat{\rho}\lambda_a} + \frac{\gamma_2}{\lambda_a}\Big)\bar{\epsilon}}, \tag{27}$$

where $\hat{N}_k := \max_{t=0,\ldots,k} \|\mathbf{w}^t\| \leq \hat{M} < \infty$ *with* $\hat{M}$ *in Lemma 7, and* $\mathbf{w}^\star \in S^\perp$ *is any optimal solution of problem* (3)*. In addition,* $\bar{D} \geq D^\star$ *and* $\lambda_a > 0$ *are defined in Lemma 7,* $\gamma_1, \gamma_2 > 0$ *in Lemma 6,* $\hat{\rho} > 0$ *in Lemma 5, and* L_i $\forall i \in \mathcal{V}$ *in Proposition 1.*

Proof. See Appendix K. □

Theorem 4 states that with the inexact Fenchel dual gradient methods in the form of (21), the running averages of $\mathbf{w}^k$ and $\hat{\mathbf{x}}(\mathbf{w}^k)$ converge at rates of $O(1/\sqrt{k})$ to suboptimality of $O(\sqrt{\bar{\epsilon}})$. Like Theorem 3, (24)–(27) also increase with n and B on the order of $O(n^{1.5}B)$ in the worst case for appropriately selected weight matrices and step-sizes. Furthermore, note that one way to bound $\bar{D}$ is to utilize the inequality

$$D(\mathbf{w}) \leq D(\mathbf{0}_{nd}) + \nabla D(\mathbf{0}_{nd})^T \mathbf{w} + \frac{L}{2}\|\mathbf{w}\|^2 \leq D(\mathbf{0}_{nd}) + \|\nabla D(\mathbf{0}_{nd})\| \cdot \|\mathbf{w}\| + \frac{L}{2}\|\mathbf{w}\|^2, \ \forall \mathbf{w} \in S^\perp.$$

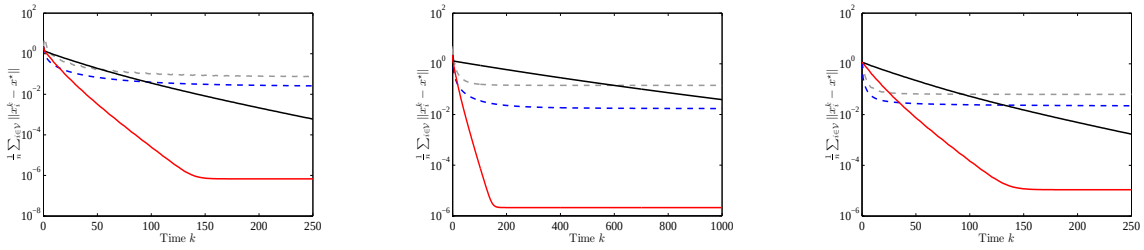
Then, following the discussions below Lemma 7 and in Section IV-C2, we can derive less tight bounds than (24)–(27), so that the dual and primal errors can be evaluated during the execution of Algorithm (21).

VI. Numerical Examples

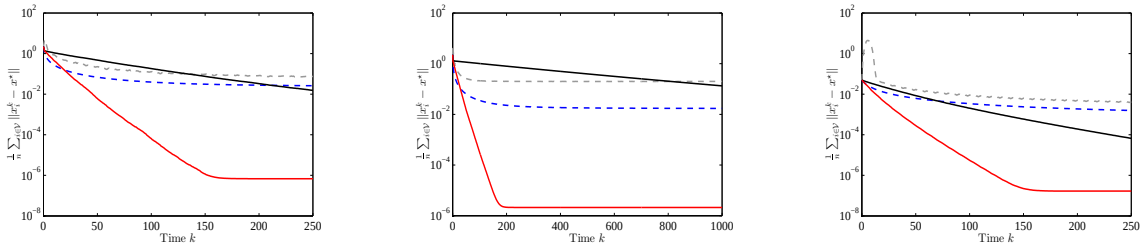
In this section, we demonstrate the competent convergence performance of the proposed distributed Fenchel dual gradient methods by comparing them with a number of existing distributed optimization algorithms via simulations.

A. Constrained case

We first compare the convergence performance of a consensus-based subgradient projection method [4], a proximal-minimization-based method [24], and Algorithm 1 with $H_{\mathcal{G}^k}$ given by the graph Laplacian matrix (11) and the Metropolis weight matrix (12), respectively, in solving constrained distributed optimization problems in the form of (1). It has been proved that when each local constraint X_i is compact, the consensus-based subgradient projection method and the proximal-minimization-based method, with diminishing step-sizes (e.g., $1/k$), asymptotically converge to an optimum over time-varying networks satisfying Assumption 3



(a) $n = 50$, $B = 10$, $2 < \theta_i < 3$ (b) $n = 500$, $B = 10$, $2 < \theta_i < 3$ (c) $n = 50$, $B = 10$, $0.2 < \theta_i < 0.4$



(d) $n = 50$, $B = 50$, $2 < \theta_i < 3$ (e) $n = 500$, $B = 50$, $2 < \theta_i < 3$ (f) $n = 50$, $B = 10$, $5 < \theta_i < 10$

Fig. 1. Primal errors in solving problem (28) (The grey dashed, blue dashed, black solid, and red solid curves correspond to the consensus-based subgradient projection method, the proximal-minimization-based method, Algorithm 1 with $H_{\mathcal{G}^k}$ in (11), and Algorithm 1 with $H_{\mathcal{G}^k}$ in (12), respectively.).

[6], [24]. Thus, consider the following multi-agent ℓ_1-regularization problem that often arises in machine learning:

$$\begin{array}{ll} \underset{x\in\mathbb{R}^5}{\text{minimize}} & \sum_{i\in\mathcal{V}}(x^T A_i x + b_i^T x + \frac{1}{n}\|x\|_1) \\ \text{subject to} & x \in \bigcap_{i\in\mathcal{V}}\{x \in \mathbb{R}^5 : p_i \leq x \leq q_i\}, \end{array} \tag{28}$$

where each $A_i \in \mathbb{R}^{5\times 5}$ is symmetric positive definite, $b_i \in \mathbb{R}^5$, and $p_i \leq x \leq q_i$ with $p_i, q_i \in \mathbb{R}^5$ means an elementwise inequality. In addition, for each $i \in \mathcal{V}$, the convexity parameter of its local objective is $\theta_i = \lambda_5^{\downarrow}(A_i) > 0$.

For Algorithm 1, we adopt $\alpha^k = 1/(Ln)$ for $H_{\mathcal{G}^k}$ in (11) and $\alpha^k = 1/2$ for $H_{\mathcal{G}^k}$ in (12) to satisfy the step-size condition (13). For the other two methods, we adopt the diminishing step-size $1/k$ and the local (unweighted) averaging operation as the consensus scheme to guarantee convergence. We also let the algorithms all start from the same initial primal iterate.

Figure 1 presents the average primal errors produced by the aforementioned algorithms with different values of n, B and θ_i $\forall i \in \mathcal{V}$. Observe that Algorithm 1 with the Metropolis weight matrix (12) outperforms the others in all six cases. Moreover, although at early stage the subgradient projection method and the proximal minimization method converge faster than

Algorithm 1 with the Laplacian weight matrix (11), their convergence gradually becomes much slower due to the diminishing nature of the step-size. By comparing Figure 1(a) versus 1(d) and Figure 1(b) versus 1(e), we can see that smaller B leads to faster convergence of Algorithm 1, which is consistent with our convergence analysis in Section IV, while the impact of B on the subgradient projection method and the proximal minimization method is not apparent. Besides, Figure 1(a) versus 1(b) and Figure 1(d) versus 1(e) suggest that Algorithm 1 with $H_{\mathcal{G}^k}$ in (12) is more scalable to the network size n than the others. Also, by comparing Figures 1(c) and 1(f) with Figure 1(a), it can be inferred that the larger the θ_i's are, the better Algorithm 1 performs.

B. Unconstrained case

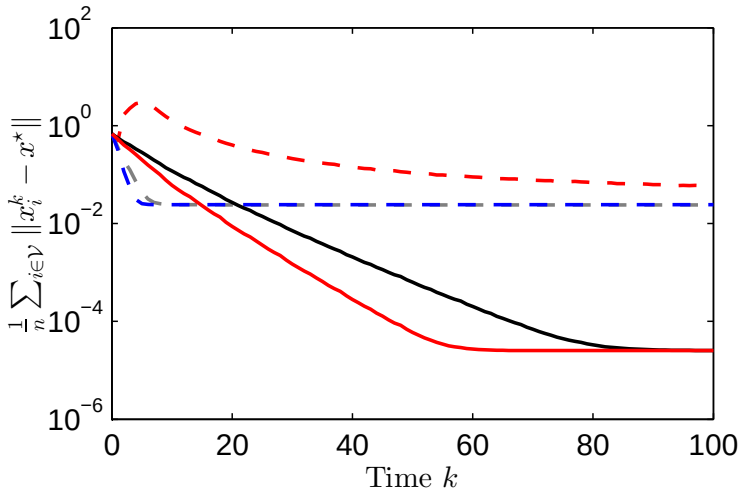
In Section IV-C3, we have compared Algorithm 1 versus Subgradient-Push [12], Gradient-Push [13], DIGing [14], and Push-DIGing [14] in the theoretical aspects. Here, we compare, via simulation, their convergence performance in solving the following unconstrained quadratic program that satisfies all the assumptions in [13], [14]:

$$\text{minimize}_{x\in\mathbb{R}^5} \sum_{i\in\mathcal{V}}(x^T A_i x + b_i^T x), \tag{29}$$

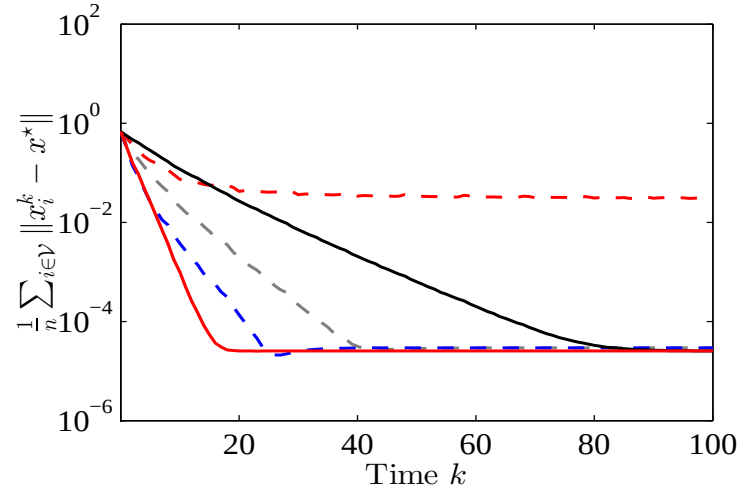
where we let $\theta_i = \lambda_5^{\downarrow}(A_i) \in (2,3)\ \forall i \in \mathcal{V}$ and $(n, B) = (50, 10)$. For fair comparison, we assume there is no stochastic error in gradient evaluation for Gradient-Push. Then, Gradient-Push and Subgradient-Push have the same algorithmic form when the local objectives are differentiable, and below we omit Subgradient-Push.

Figure 2(a) plots the evolution of the average primal error for Gradient-Push, DIGing, Push-DIGing, and Algorithm 1 with the Laplacian weight matrix (11) and with the Metropolis weight matrix (12). We adopt the same step-sizes for Algorithm 1 as in Section VI-A. For the other three methods, we fine-tune the step-sizes while satisfying the step-size conditions in [13], [14] that theoretically ensure their convergence rates. Observe that Gradient-Push, DIGing, and Push-DIGing almost stop making progress after a few iterations with a non-negligible primal error, while Algorithm 1 achieves much better accuracy with the above two choices of $H_{\mathcal{G}^k}$.

As all the convergence rate results in [13], [14] and this paper are derived from worst-case analysis, the theoretical step-size conditions could be very conservative. Thus, in Figure 2(b) we empirically choose the step-sizes for these algorithms, whose values may violate the theoretical conditions but speed up convergence. After some tuning, we select the step-sizes to be $1/(nL)$, 1.7, $0.15/k$, 0.05, and 0.04 for Algorithm 1 with $H_{\mathcal{G}^k}$ in (11), Algorithm 1 with $H_{\mathcal{G}^k}$ in (12),



(a) Theoretically-selected step-sizes



(b) Empirically-selected step-sizes

Fig. 2. Primal errors in solving problem (29) (The red dashed, grey dashed, blue dashed, black solid, and red solid curves correspond to Gradient-Push, DIGing, Push-DIGing, Algorithm 1 with $H_{\mathcal{G}^k}$ in (11), and Algorithm 1 with $H_{\mathcal{G}^k}$ in (12), respectively.).

Gradient-Push, DIGing, and Push-DIGing, respectively. Note that for Algorithm 1 with $H_{\mathcal{G}^k}$ in (11), the empirical step-size coincides with the theoretical one in Figure 2(a). By comparing Figure 2(b) with Figure 2(a), we can observe that with the above empirically-selected step-sizes, Gradient-Push slightly accelerates its convergence, DIGing and Push-DIGing exhibit prominently improved convergence performance, yet Algorithm 1 with $H_{\mathcal{G}^k}$ in (12) still performs best.

VII. Conclusion

We have constructed a family of distributed Fenchel dual gradient methods for solving multi-agent optimization problems with strongly convex local objectives and nonidentical local constraints over time-varying networks. The proposed algorithms have been proved to asymptotically converge to the optimal solution under a minimal connectivity condition, and have an $O(1/\sqrt{k})$ convergence rate under a standard connectivity condition. We have also provided alternatives to the proposed algorithms when the Lipschitz constants of the dual gradients are unknown or when the dual gradient evaluations are inexact. Simulation results have illustrated the competitive performance of the distributed Fenchel dual gradient methods by comparing them with related algorithms. In future, this work may be extended in a number of directions such as problems with general convex objective functions and networks with directed links.

Appendix

A. Proof of Proposition 2

Let $\mathbf{w}^\star = ((w_1^\star)^T, \ldots, (w_n^\star)^T)^T$ be an optimal solution of problem (3). Since Assumption 1(b) assumes $\mathbf{0}_d \in \operatorname{int} \bigcap_{i\in\mathcal{V}} X_i$, there exists $r_c \in (0,\infty)$ such that $B(\mathbf{0}_d, r_c) \subseteq \bigcap_{i\in\mathcal{V}} X_i$. For each $i \in \mathcal{V}$, if $w_i^\star \neq \mathbf{0}_d$, let $x_i' = r_c \frac{w_i^\star}{\|w_i^\star\|}$; otherwise let $x_i' = \mathbf{0}_d$. Clearly, $x_i' \in B(\mathbf{0}_d, r_c)$. Consequently,

$$\begin{aligned} D^\star = D(\mathbf{w}^\star) &= \sum_{i\in\mathcal{V}} \Big(\sup_{x_i\in X_i} (w_i^\star)^T x_i - f_i(x_i) \Big) \\ &\geq \sum_{i\in\mathcal{V}} \Big((w_i^\star)^T x_i' - f_i(x_i') \Big) = r_c \sum_{i\in\mathcal{V}} \|w_i^\star\| - \sum_{i\in\mathcal{V}} f_i(x_i'). \end{aligned}$$

This, along with $\|\mathbf{w}^\star\| \leq \sum_{i\in\mathcal{V}} \|w_i^\star\|$ and $D^\star = -F^\star$, implies that $\|\mathbf{w}^\star\| \leq \big((\sum_{i\in\mathcal{V}} f_i(x_i')) - F^\star \big)/r_c$. Note that $\sum_{i\in\mathcal{V}} f_i(x_i') \leq \sum_{i\in\mathcal{V}} \max_{x_i\in B(\mathbf{0}_d, r_c)} f_i(x_i)$, where $F^\star \leq \sum_{i\in\mathcal{V}} \max_{x_i\in B(\mathbf{0}_d,r_c)} f_i(x_i) < \infty$ because $B(\mathbf{0}_d, r_c)$ is compact. Therefore, (7) holds, which suggests that the optimal set of problem (3) is compact. Then, due to the convexity of D and $S^\perp$, the level sets $S_0(\mathbf{w})$ $\forall \mathbf{w} \in S^\perp$ are also compact [44, proposition 1.4.5].

B. Proof of Lemma 1

For convenience, let $\mathbf{y}^k = (H_{\mathcal{G}^k} \otimes I_d)\nabla D(\mathbf{w}^k)$. Due to the Descent Lemma [36] and (8),

$$\begin{aligned} D(\mathbf{w}^{k+1}) - D(\mathbf{w}^k) &\leq \langle \nabla D(\mathbf{w}^k), \mathbf{w}^{k+1} - \mathbf{w}^k \rangle + (\mathbf{w}^{k+1} - \mathbf{w}^k)^T \frac{\Lambda_L \otimes I_d}{2} (\mathbf{w}^{k+1} - \mathbf{w}^k) \\ &= -\alpha^k \langle \nabla D(\mathbf{w}^k), \mathbf{y}^k \rangle + (\alpha^k)^2 (\mathbf{y}^k)^T \frac{\Lambda_L \otimes I_d}{2} \mathbf{y}^k. \end{aligned} \tag{30}$$

Then, consider the following lemma.

Lemma 8. *Suppose $M, \bar{M} \in \mathbb{R}^{n\times n}$ are symmetric positive semidefinite and $M \preceq \bar{M}$. Then, for any $\mathbf{x} \in \mathbb{R}^{nd}$ and any $\mathbf{y} \in \mathcal{R}(M \otimes I_d)$,*

$$\langle \mathbf{x}, (M \otimes I_d)\mathbf{x}\rangle \geq \langle (M \otimes I_d)\mathbf{x}, (\bar{M}^\dagger \otimes I_d)(M \otimes I_d)\mathbf{x}\rangle.$$

Proof. Let $\mathbf{x} \in \mathbb{R}^{nd}$. Then,

$$\langle \mathbf{x}, (M \otimes I_d)\mathbf{x}\rangle - \langle (M \otimes I_d)\mathbf{x}, (\bar{M}^\dagger \otimes I_d)(M \otimes I_d)\mathbf{x}\rangle = \mathbf{x}^T[(M - M\bar{M}^\dagger M) \otimes I_d]\mathbf{x}. \quad (31)$$

In addition, by Schur complement condition, $M \succeq O_n$ and $\bar{M} \succeq M$ implies

$$\begin{pmatrix} M & M \\ M & \bar{M} \end{pmatrix} \succeq O_{2n}$$

and the inequality above leads to $M - M\bar{M}^\dagger M \succeq O_n$. Combining this with (31), we complete the proof. □

From Lemma 8, $(\mathbf{y}^k)^T(\Lambda_L \otimes I_d)\mathbf{y}^k \leq \delta\langle \nabla D(\mathbf{w}^k), \mathbf{y}^k\rangle$. Combining this with (30) leads to

$$D(\mathbf{w}^{k+1}) - D(\mathbf{w}^k) \leq (\frac{(\alpha^k)^2\delta}{2} - \alpha^k)\langle \nabla D(\mathbf{w}^k), \mathbf{y}^k\rangle.$$

This, along with (13), completes the proof.

C. Proof of Lemma 2

We first consider the following optimization problem: For any $\mathcal{I} \subseteq \mathcal{V}$, $\mathcal{I} \neq \emptyset$ and any $c \in \mathbb{R}^d$,

$$\begin{array}{ll} \underset{w_i \in \mathbb{R}^d \ \forall i \in \mathcal{I}}{\text{minimize}} & \sum_{i\in\mathcal{I}} d_i(w_i) \\ \text{subject to} & \sum_{i\in\mathcal{I}} w_i = c. \end{array} \quad (32)$$

Similar to problem (3), $w_i' \ \forall i \in \mathcal{I}$ compose an optimum to (32) if and only if for any $i, j \in \mathcal{I}$, $\nabla d_i(w_i') = \nabla d_j(w_j')$ [34, Lemma 3.1], or equivalently, $\tilde{x}_i(w_i') = \tilde{x}_j(w_j')$. With the above setting, consider the following lemma.

Lemma 9. *Suppose Assumption 1 and the step-size condition* (13) *hold. Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{nd}$ be two feasible solutions of problem* (3) *such that $u_i \ \forall i \in \mathcal{I}$ and $v_i \ \forall i \in \mathcal{I}$ are feasible to problem* (32)*. Suppose $\|\tilde{x}_i(v_i) - \tilde{x}_j(v_j)\| \leq \epsilon' \ \forall i, j \in \mathcal{I}$ for some $\epsilon' > 0$, $\sum_{i\in\mathcal{I}} d_i(u_i) \leq \sum_{i\in\mathcal{I}} d_i(v_i)$, and $D(\mathbf{v}) \leq D(\mathbf{w}^0)$, where $\mathbf{w}^0 \in S^\perp$ is the initial dual iterate of Algorithm 1. Then,*

$$\|\tilde{x}_i(u_i) - \tilde{x}_j(u_j)\| \leq 4\sqrt{LM_0(|\mathcal{I}| - 1)\epsilon'}, \quad \forall i, j \in \mathcal{I},$$

where M_0 is defined in (15).

Proof. Let $\mathbf{w}' = (w_1'^T, \ldots, w_n'^T)^T \in \mathbb{R}^{nd}$ be such that $w_i' \in \mathbb{R}^d \ \forall i \in \mathcal{I}$ compose an optimal solution to (32) and $w_j' = v_j \ \forall j \notin \mathcal{I}$. Due to the convexity of each d_i and (5),

$$\sum_{i\in\mathcal{I}} d_i(v_i) - \sum_{i\in\mathcal{I}} d_i(w_i') \leq \sum_{i\in\mathcal{I}} \langle \tilde{x}_i(v_i), v_i - w_i' \rangle.$$

Let $\bar{x}_v := \frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} \tilde{x}_i(v_i)$. Since $w_i' \ \forall i \in \mathcal{I}$ and $v_i \ \forall i \in \mathcal{I}$ are feasible to (32), we have $\sum_{i\in\mathcal{I}} w_i' = \sum_{i\in\mathcal{I}} v_i$, which gives

$$\sum_{i\in\mathcal{I}} \langle \tilde{x}_i(v_i), v_i - w_i' \rangle = \sum_{i\in\mathcal{I}} \langle \tilde{x}_i(v_i) - \bar{x}_v, v_i - w_i' \rangle \leq \sum_{i\in\mathcal{I}} \|\tilde{x}_i(v_i) - \bar{x}_v\| \cdot \|v_i - w_i'\|.$$

Also note that for each $i \in \mathcal{I}$, $\|\tilde{x}_i(v_i) - \bar{x}_v\| = \frac{1}{|\mathcal{I}|} \| \sum_{j\in\mathcal{I}} (\tilde{x}_i(v_i) - \tilde{x}_j(v_j)) \| \leq \frac{|\mathcal{I}|-1}{|\mathcal{I}|} \epsilon'$. Combining the above,

$$\sum_{i\in\mathcal{I}} d_i(v_i) - \sum_{i\in\mathcal{I}} d_i(w_i') \leq \frac{|\mathcal{I}|-1}{|\mathcal{I}|} \epsilon' \sum_{i\in\mathcal{I}} \|v_i - w_i'\| \leq (|\mathcal{I}|-1)\epsilon' \sqrt{\sum_{i\in\mathcal{I}} \|v_i - w_i'\|^2}. \tag{33}$$

Since $\sum_{i\in\mathcal{I}} d_i(w_i') \leq \sum_{i\in\mathcal{I}} d_i(v_i)$ and $w_j' = v_j \ \forall j \notin \mathcal{I}$, we have $D(\mathbf{w}') \leq D(\mathbf{v}) \leq D(\mathbf{w}^0)$, implying that $\mathbf{w}', \mathbf{v} \in S_0(\mathbf{w}^0)$ and that for any optimum $\mathbf{w}^\star$ of problem (3),

$$\|\mathbf{w}' - \mathbf{v}\| \leq \|\mathbf{w}' - \mathbf{w}^\star\| + \|\mathbf{v} - \mathbf{w}^\star\| \leq 2M_0.$$

This inequality and (33) together yield

$$\sum_{i\in\mathcal{I}} d_i(v_i) - \sum_{i\in\mathcal{I}} d_i(w_i') \leq 2M_0(|\mathcal{I}|-1)\epsilon'. \tag{34}$$

Due to the optimality of $w_i' \ \forall i \in \mathcal{I}$ with respect to (32), we have $\nabla d_i(w_i') = \nabla d_j(w_j') \ \forall i, j \in \mathcal{I}$. Also, because of the feasibility of $u_i \ \forall i \in \mathcal{I}$, $\sum_{i\in\mathcal{I}} u_i = \sum_{i\in\mathcal{I}} w_i'$. Therefore, $\sum_{i\in\mathcal{I}} \langle \nabla d_i(w_i'), u_i - w_i' \rangle = 0$. This, along with (5), (34), and the inequality $d_i(u_i) - d_i(w_i') \geq \langle \nabla d_i(w_i'), u_i - w_i' \rangle + \frac{1}{2L} \|\nabla d_i(w_i') - \nabla d_i(u_i)\|^2$ [39, Theorem 2.1.5], implies

$$\begin{aligned}\sum_{i\in\mathcal{I}} \|\tilde{x}_i(u_i) - \tilde{x}_i(w_i')\|^2 &\leq 2L \sum_{i\in\mathcal{I}} (d_i(u_i) - d_i(w_i')) \\ &\leq 2L \sum_{i\in\mathcal{I}} (d_i(v_i) - d_i(w_i')) \leq 4LM_0(|\mathcal{I}|-1)\epsilon'.\end{aligned}$$

Hence, for any $i, j \in \mathcal{I}$, we have $\|\tilde{x}_i(u_i) - \tilde{x}_j(u_j)\| \leq \|\tilde{x}_i(u_i) - \tilde{x}_i(w_i')\| + \|\tilde{x}_j(u_j) - \tilde{x}_j(w_j')\| \leq 4\sqrt{LM_0(|\mathcal{I}|-1)\epsilon'}$, where the first inequality is from the optimality of $w_i' \ \forall i \in \mathcal{I}$ and (5). □

Next, we define the following: Arbitrarily pick $\epsilon > 0$. Due to (16), $\exists T_\epsilon \geq 0$ such that

$$\|x_i^k - x_j^k\| \leq \epsilon, \quad \forall \{i, j\} \in \mathcal{E}^k, \ \forall k \geq T_\epsilon. \tag{35}$$

Then, for each $i \in \mathcal{V}$, let $\mathcal{C}_{i,\epsilon}^k = \emptyset \ \forall k \in [0, T_\epsilon)$. For each $k \geq T_\epsilon$, let

$$\mathcal{C}_{i,\epsilon}^k = \{i\} \cup \{j \in \mathcal{V} : \text{There exists a path between } i \text{ and } j \text{ in the graph } (\mathcal{V}, \cup_{t=T_\epsilon}^k \mathcal{E}^t)\} \subseteq \mathcal{V}.$$

For each $k \geq T_\epsilon$, observe that in the graph $(\mathcal{V}, \cup_{t=T_\epsilon}^k \mathcal{E}^t)$, the subgraph induced by $\mathcal{C}_{i,\epsilon}^k$ is the largest connected component that contains node i. Thus, for any two nodes i and j, $i \neq j$, $\mathcal{C}_{i,\epsilon}^k$ and $\mathcal{C}_{j,\epsilon}^k$ are either identical or disjoint. Additionally, for every $s \in \mathcal{C}_{i,\epsilon}^{k+1}$, $\mathcal{C}_{s,\epsilon}^k$ is always contained in $\mathcal{C}_{i,\epsilon}^{k+1}$. This implies that the number of distinct sets in the collection $\{\mathcal{C}_{i,\epsilon}^k\}_{i\in\mathcal{V}}$ is non-increasing with k over $[T_\epsilon, \infty)$. In particular, from each k to $k+1$, $\mathcal{C}_{i,\epsilon}^{k+1}$ either equals $\mathcal{C}_{i,\epsilon}^k$ or is the union of $\mathcal{C}_{i,\epsilon}^k$ and some other $\mathcal{C}_{j,\epsilon}^k$'s that are disjoint from $\mathcal{C}_{i,\epsilon}^k$. Also due to Assumption 2, there exists $K_\epsilon \in [T_\epsilon, \infty)$ such that $\mathcal{C}_{i,\epsilon}^k = \mathcal{V} \ \forall i \in \mathcal{V} \ \forall k \geq K_\epsilon$. By means of the $\mathcal{C}_{i,\epsilon}^k$'s and Lemma 9, below we show that $\forall i \in \mathcal{V}$, $\forall k \geq T_\epsilon$,

$$\max_{j,\ell \in \mathcal{C}_{i,\epsilon}^k} \|x_j^k - x_\ell^k\| \leq \Phi_i^k(\epsilon). \tag{36}$$

Here, $\Phi_i^k(\epsilon) \ \forall i \in \mathcal{V} \ \forall k \geq T_\epsilon$ are defined recursively as follows: Initially at $k = T_\epsilon$, $\Phi_i^k(\epsilon) = (|\mathcal{C}_{i,\epsilon}^k| - 1)\epsilon$. At each subsequent $k \geq T_\epsilon + 1$,

$$\Phi_i^k(\epsilon) = \begin{cases} 4\sqrt{LM_0(|\mathcal{C}_{i,\epsilon}^k| - 1)\Phi_i^{t^k}(\epsilon)}, & \text{if } \mathcal{C}_{i,\epsilon}^k = \mathcal{C}_{i,\epsilon}^{k-1}, \\ (1 + 2L\bar{\alpha}\bar{h}n)|\mathcal{C}_{i,\epsilon}^k|\epsilon + \sum_{s\in\mathcal{C}_{i,\epsilon}^k} \Phi_s^{k-1}(\epsilon), & \text{otherwise,} \end{cases}$$

where $t^k := \max\{t \in [T_\epsilon, k] : \mathcal{C}_{i,\epsilon}^t \neq \mathcal{C}_{i,\epsilon}^{t-1}\}$. Note that $\mathcal{C}_{i,\epsilon}^k = \mathcal{C}_{i,\epsilon}^t \ \forall t \in [t^k, k]$.

We prove (36) by induction. At time $k = T_\epsilon$, for each $i \in \mathcal{V}$, if $|\mathcal{C}_{i,\epsilon}^k| = 1$, then $\max_{j,\ell\in\mathcal{C}_{i,\epsilon}^k} \|x_j^k - x_\ell^k\| = \Phi_i^k(\epsilon) = 0$, i.e., (36) is satisfied; otherwise for any $j, \ell \in \mathcal{C}_{i,\epsilon}^k$, $j \neq \ell$, there exists a path of length at most $|\mathcal{C}_{i,\epsilon}^k| - 1$ connecting j and ℓ. It follows from (35) that $\|x_j^k - x_\ell^k\| \leq (|\mathcal{C}_{i,\epsilon}^k| - 1)\epsilon = \Phi_i^k(\epsilon)$, i.e., (36) also holds. Next, suppose $\max_{j,\ell\in\mathcal{C}_{i,\epsilon}^t} \|x_j^t - x_\ell^t\| \leq \Phi_i^t(\epsilon) \ \forall i \in \mathcal{V} \ \forall t \in [T_\epsilon, k-1]$ for some $k \geq T_\epsilon + 1$. For each $i \in \mathcal{V}$, to show that (36) holds, consider the following two cases.

Case i: $\mathcal{C}_{i,\epsilon}^k = \mathcal{C}_{i,\epsilon}^{k-1}$. In this case, we have $T_\epsilon \leq t^k \leq k-1$. Also, $\forall t \in [t^k + 1, k]$, $\forall j \in \mathcal{C}_{i,\epsilon}^{t-1}$, we have $\mathcal{N}_j^t \subseteq \mathcal{C}_{i,\epsilon}^{t-1} = \mathcal{C}_{i,\epsilon}^k$. Hence, using the same arguments as the proofs of Proposition 3 and Lemma 1, it can be shown that $\sum_{s\in\mathcal{C}_{i,\epsilon}^k} w_s^k = \sum_{s\in\mathcal{C}_{i,\epsilon}^k} w_s^{k-1} = \cdots = \sum_{s\in\mathcal{C}_{i,\epsilon}^k} w_s^{t^k}$ and that $\sum_{s\in\mathcal{C}_{i,\epsilon}^k} d_s(w_s^k) \leq \sum_{s\in\mathcal{C}_{i,\epsilon}^k} d_s(w_s^{k-1}) \leq \cdots \leq \sum_{s\in\mathcal{C}_{i,\epsilon}^k} d_s(w_s^{t^k})$. Let $\mathcal{I} = \mathcal{C}_{i,\epsilon}^k$ and $c = \sum_{s\in\mathcal{C}_{i,\epsilon}^k} w_s^{t^k}$ in problem (32). It then follows from Lemma 1 and Lemma 9 with $\epsilon' = \Phi_i^{t^k}(\epsilon)$, $\mathbf{u} = \mathbf{w}^k$, and $\mathbf{v} = \mathbf{w}^{t^k}$ that (36) holds.

Case ii: $\mathcal{C}_{i,\epsilon}^k \neq \mathcal{C}_{i,\epsilon}^{k-1}$. Pick any $j, \ell \in \mathcal{C}_{i,\epsilon}^k$, $j \neq \ell$ and consider the following two subcases.

Subcase ii(a): $\mathcal{C}_{j,\epsilon}^{k-1} = \mathcal{C}_{\ell,\epsilon}^{k-1}$. Then, $\|x_j^k - x_\ell^k\| \le \|x_j^k - x_j^{k-1}\| + \|x_j^{k-1} - x_\ell^{k-1}\| + \|x_\ell^{k-1} - x_\ell^k\| \le \|x_j^k - x_j^{k-1}\| + \|x_\ell^k - x_\ell^{k-1}\| + \Phi_j^{k-1}(\epsilon)$. Also, from (5), Proposition 1, (8), and (35), we have

$$\|x_p^k - x_p^{k-1}\| \le L_p \|w_p^k - w_p^{k-1}\| \le L\bar{\alpha} \| \sum_{q \in \mathcal{N}_p^{k-1}} h_{pq}^{k-1}(x_p^{k-1} - x_q^{k-1}) \|$$

$$\le L\bar{\alpha}\bar{h} \sum_{q \in \mathcal{N}_p^{k-1}} \|x_p^{k-1} - x_q^{k-1}\| \le L\bar{\alpha}\bar{h}n\epsilon, \quad \forall p \in \mathcal{V}.$$

Consequently, $\|x_j^k - x_\ell^k\| \le 2L\bar{\alpha}\bar{h}n\epsilon + \Phi_j^{k-1}(\epsilon)$.

Subcase ii(b): $\mathcal{C}_{j,\epsilon}^{k-1} \cap \mathcal{C}_{\ell,\epsilon}^{k-1} = \emptyset$. Then, there exists a path from j to ℓ belonging to the subgraph induced in the graph $(\mathcal{V}, \cup_{t=T_\epsilon}^k \mathcal{E}^t)$ by $\mathcal{C}_{i,\epsilon}^k$. Along the path are nodes $p_1 = j, s_1, p_2, s_2, \ldots, p_\tau, s_\tau = \ell$ such that (1) $\mathcal{C}_{p_r,\epsilon}^{k-1} = \mathcal{C}_{s_r,\epsilon}^{k-1}$ $\forall r = 1, \ldots, \tau$; (2) $\mathcal{C}_{p_r,\epsilon}^{k-1}$ $\forall r \in \{1, \ldots, \tau\}$ are disjoint from each other; and (3) $\{s_r, p_{r+1}\} \in \mathcal{E}^k$ $\forall r \in \{1, \ldots, \tau - 1\}$. Here, $\tau \in \{2, \ldots, |\mathcal{C}_{i,\epsilon}^k|\}$ is an integer whose value is no more than the number of distinct sets in the collection $\{\mathcal{C}_{s,\epsilon}^{k-1}\}_{s \in \mathcal{C}_{i,\epsilon}^k}$. Hence, $\|x_j^k - x_\ell^k\| \le \|x_{p_1}^k - x_{s_1}^k\| + \sum_{r=1}^{\tau-1} \big(\|x_{s_r}^k - x_{p_{r+1}}^k\| + \|x_{p_{r+1}}^k - x_{s_{r+1}}^k\| \big)$. For each $r = 1, \ldots, \tau$, since $p_r, s_r \in \mathcal{C}_{p_r,\epsilon}^{k-1}$, we obtain from *Subcase ii(a)* that $\|x_{p_r}^k - x_{s_r}^k\| \le 2L\bar{\alpha}\bar{h}n\epsilon + \Phi_{p_r}^{k-1}(\epsilon)$. It then follows from (35) that $\|x_j^k - x_\ell^k\| \le (\tau - 1)\epsilon + 2\tau L\bar{\alpha}\bar{h}n\epsilon + \sum_{r=1}^{\tau} \Phi_{p_r}^{k-1}(\epsilon) \le (1 + 2L\bar{\alpha}\bar{h}n)|\mathcal{C}_{i,\epsilon}^k|\epsilon + \sum_{s \in \mathcal{C}_{i,\epsilon}^k} \Phi_s^{k-1}(\epsilon)$.

Combining the above two subcases, we obtain (36). This completes the proof of (36) for all $i \in \mathcal{V}$ and all $k \ge T_\epsilon$. Further, notice that for each $i \in \mathcal{V}$, $\Phi_i^k(\epsilon)$ is updated only if either $\mathcal{C}_{i,\epsilon}^k$ or $\mathcal{C}_{i,\epsilon}^{k-1}$ is changed. Also note that $\mathcal{C}_{i,\epsilon}^k$ can be expanded at most n times and remains unchanged since time K_ϵ. Therefore, for any $k \ge K_\epsilon + 1$,

$$\max_{i,j \in \mathcal{V}} \|x_i^k - x_j^k\| = \max_{i \in \mathcal{V}} \max_{j,\ell \in \mathcal{C}_{i,\epsilon}^k} \|x_j^k - x_\ell^k\| \le \max_{i \in \mathcal{V}} \Phi_i^k(\epsilon) \le O(\epsilon^{1/2^n}),$$

which implies $\max_{i,j \in \mathcal{V}} \|x_i^k - x_j^k\| \to 0$ as $k \to \infty$.

D. Proof of Theorem 1

Let $\mathbf{w}^\star$ be an optimal solution to the dual problem (3). Due to the convexity of D, (5), and Proposition 3,

$$D(\mathbf{w}^k) - D^\star \le \langle \nabla D(\mathbf{w}^k), \mathbf{w}^k - \mathbf{w}^\star \rangle = \langle \mathbf{x}^k, \mathbf{w}^k - \mathbf{w}^\star \rangle$$

$$\le \|P_{S^\perp}(\mathbf{x}^k)\| \cdot \|\mathbf{w}^k - \mathbf{w}^\star\| \le M_0 \|P_{S^\perp}(\mathbf{x}^k)\|,$$

where M_0 is defined in (15). As $k \to \infty$, we have shown in the paragraph below Lemma 2 that $\|P_{S^\perp}(\mathbf{x}^k)\| \to 0$. This, along with the above inequality, implies $D(\mathbf{w}^k) \to D^\star$. In addition, since

Assumption 1 guarantees zero duality gap, we have $F(\mathbf{x}^k) \to F^\star$. Finally, for any $\mathbf{w} \in S^\perp$, due to Corollary 1, [39, Theorem 2.1.5], and (6),

$$D(\mathbf{w}) - D^\star \geq \langle \nabla D(\mathbf{w}^\star), \mathbf{w} - \mathbf{w}^\star \rangle + \frac{1}{2L}\|\nabla D(\mathbf{w}) - \nabla D(\mathbf{w}^\star)\|^2 = \frac{1}{2L}\|\tilde{\mathbf{x}}(\mathbf{w}) - \mathbf{x}^\star\|^2, \quad (37)$$

where the last equality is because $\nabla D(\mathbf{w}^\star) = \mathbf{x}^\star \in S$ and $\mathbf{w}, \mathbf{w}^\star \in S^\perp$. Thus, because $\lim_{k\to\infty} D(\mathbf{w}^k) - D^\star = 0$ and $L > 0$, $\|\mathbf{x}^k - \mathbf{x}^\star\|^2 \to 0$ as $k \to \infty$.

E. Proof of Lemma 3

Let $k \in \{0, B, 2B, \ldots\}$. For each $\{i,j\} \in \tilde{\mathcal{E}}^k$, let $t^k_{\{i,j\}} \in \{k, \ldots, k+B-1\}$ be such that $\{i,j\} \in \mathcal{E}^{t^k_{\{i,j\}}}$. Then, note from Proposition 1 that

$$\begin{aligned}
&\|\nabla d_i(w_i^k) - \nabla d_i(w_i^{t^k_{\{i,j\}}})\|^2 = \|\sum_{t=k}^{t^k_{\{i,j\}}-1} (\nabla d_i(w_i^{t+1}) - \nabla d_i(w_i^t))\|^2 \\
&\leq B \sum_{t=k}^{k+B-1} \|\nabla d_i(w_i^{t+1}) - \nabla d_i(w_i^t)\|^2 \leq L_i^2 B \sum_{t=k}^{k+B-1} \|w_i^{t+1} - w_i^t\|^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sum_{\{i,j\}\in\tilde{\mathcal{E}}^k} (\|\nabla d_i(w_i^k) - \nabla d_i(w_i^{t^k_{\{i,j\}}})\|^2 + \|\nabla d_j(w_j^{t^k_{\{i,j\}}}) - \nabla d_j(w_j^k)\|^2) \\
\leq & B \sum_{\{i,j\}\in\tilde{\mathcal{E}}^k} \sum_{t=k}^{k+B-1} (L_i^2\|w_i^{t+1} - w_i^t\|^2 + L_j^2\|w_j^{t+1} - w_j^t\|^2) \\
\leq & B\bar{\varpi} \sum_{t=k}^{k+B-1} \sum_{i\in\mathcal{V}} L_i^2\|w_i^{t+1} - w_i^t\|^2 \\
\leq & B\bar{\varpi}\bar{\alpha}^2 \sum_{t=k}^{k+B-1} \langle \nabla D(\mathbf{w}^t), ((H_{\mathcal{G}^t}\Lambda_L^2 H_{\mathcal{G}^t}) \otimes I_d)\nabla D(\mathbf{w}^t)\rangle.
\end{aligned}$$

Note that $H_{\mathcal{G}^t}\Lambda_L^2 H_{\mathcal{G}^t} \preceq L H_{\mathcal{G}^t}\Lambda_L H_{\mathcal{G}^t}$. Also, from (14) and Lemma 8, $H_{\mathcal{G}^t}\Lambda_L H_{\mathcal{G}^t} \preceq \delta H_{\mathcal{G}^t}$. Hence,

$$\begin{aligned}
&\sum_{\{i,j\}\in\tilde{\mathcal{E}}^k} (\|\nabla d_i(w_i^k) - \nabla d_i(w_i^{t^k_{\{i,j\}}})\|^2 + \|\nabla d_j(w_j^{t^k_{\{i,j\}}}) - \nabla d_j(w_j^k)\|^2) \\
&\leq B\bar{\varpi}\bar{\alpha}^2\delta L \sum_{t=k}^{k+B-1} \nabla D(\mathbf{w}^t)^T (H_{\mathcal{G}}^t \otimes I_d)\nabla D(\mathbf{w}^t). \quad (38)
\end{aligned}$$

In addition,

$$\sum_{\{i,j\}\in\tilde{\mathcal{E}}^k} \|\nabla d_i(w_i^{t^k_{\{i,j\}}}) - \nabla d_j(w_j^{t^k_{\{i,j\}}})\|^2 \leq \frac{1}{\underline{h}} \sum_{t=k}^{k+B-1} \sum_{\{i,j\}\in\mathcal{E}^t} h_{ij}^k \|\nabla d_i(w_i^t) - \nabla d_j(w_j^t)\|^2$$

$$\leq \frac{1}{\underline{h}} \sum_{t=k}^{k+B-1} \nabla D(\mathbf{w}^t)^T (H_{\mathcal{G}}^t \otimes I_d) \nabla D(\mathbf{w}^t). \tag{39}$$

It follows from (38) and (39) that

$$\begin{aligned}
&\nabla D(\mathbf{w}^k)^T (L_{\tilde{\mathcal{G}}^k} \otimes I_d) \nabla D(\mathbf{w}^k) = \sum_{\{i,j\} \in \tilde{\mathcal{E}}^k} \|\nabla d_i(w_i^k) - \nabla d_j(w_j^k)\|^2 \\
&\leq 3 \sum_{\{i,j\} \in \tilde{\mathcal{E}}^k} (\|\nabla d_i(w_i^k) - \nabla d_i(w_i^{t_{\{i,j\}}^k})\|^2 + \|\nabla d_j(w_j^{t_{\{i,j\}}^k}) - \nabla d_j(w_j^k)\|^2 + \|\nabla d_i(w_i^{t_{\{i,j\}}^k}) - \nabla d_j(w_j^{t_{\{i,j\}}^k})\|^2) \\
&\leq \eta \sum_{t=k}^{k+B-1} \nabla D(\mathbf{w}^t)^T (H_{\mathcal{G}}^t \otimes I_d) \nabla D(\mathbf{w}^t).
\end{aligned}$$

F. Proof of Theorem 2

Let $k \geq 0$. By Lemmas 1 and 3,

$$\begin{aligned}
&\big(D(\mathbf{w}^{(k+1)B}) - D^\star\big) - \big(D(\mathbf{w}^{kB}) - D^\star\big) = \sum_{t=kB}^{(k+1)B-1} (D(\mathbf{w}^{t+1}) - D(\mathbf{w}^t)) \\
&\leq -\rho \sum_{t=kB}^{(k+1)B-1} \nabla D(\mathbf{w}^t)^T (H_{\mathcal{G}^t} \otimes I_d) \nabla D(\mathbf{w}^t) \leq -\frac{\rho}{\eta} \nabla D(\mathbf{w}^{kB})^T (L_{\tilde{\mathcal{G}}^{kB}} \otimes I_d) \nabla D(\mathbf{w}^{kB}) \\
&\leq -\frac{\rho \underline{\lambda}}{\eta} \|P_{S^\perp}(\nabla D(\mathbf{w}^{kB}))\|^2,
\end{aligned} \tag{40}$$

where the last inequality is because $\tilde{\mathcal{G}}^{kB}$ is connected and thus $\mathrm{Null}(L_{\tilde{\mathcal{G}}^{kB}} \otimes I_d) = S$. Also, since $\tilde{\mathcal{G}}^{tB}$ $\forall t = 0, 1, \ldots$ are connected, we have $\underline{\lambda} > 0$. From Proposition 3, we know that $\mathbf{w}^{kB} \in S^\perp$. Also, for any optimal solution $\mathbf{w}^\star$ to (3), because $\mathbf{w}^\star \in S^\perp$, we have $\mathbf{w}^{kB} - \mathbf{w}^\star \in S^\perp$. Then,

$$\begin{aligned}
D(\mathbf{w}^{kB}) - D^\star \leq& \langle \nabla D(\mathbf{w}^{kB}), \mathbf{w}^{kB} - \mathbf{w}^\star \rangle = \langle P_{S^\perp}(\nabla D(\mathbf{w}^{kB})), \mathbf{w}^{kB} - \mathbf{w}^\star \rangle \\
\leq& \|P_{S^\perp}(\nabla D(\mathbf{w}^{kB}))\| \cdot \|\mathbf{w}^{kB} - \mathbf{w}^\star\|.
\end{aligned}$$

This, along with (40), gives

$$\big(D(\mathbf{w}^{(k+1)B}) - D^\star\big) - \big(D(\mathbf{w}^{kB}) - D^\star\big) \leq -\rho \underline{\lambda} \big(D(\mathbf{w}^{kB}) - D^\star\big)^2 / (\eta \min_{\mathbf{w}^\star \in S^\perp : D(\mathbf{w}^\star) = D^\star} \|\mathbf{w}^{kB} - \mathbf{w}^\star\|^2).$$

Finally, using Lemma 6 in [45, Sec. 2.2.1], we obtain

$$\begin{aligned}
D(\mathbf{w}^{kB}) - D^\star \leq& \frac{D(\mathbf{w}^0) - D^\star}{1 + \frac{\rho \underline{\lambda} (D(\mathbf{w}^0) - D^\star)}{\eta} \sum_{t=0}^{k-1} (\min_{\mathbf{w}^\star \in S^\perp : D(\mathbf{w}^\star) = D^\star} \|\mathbf{w}^{tB} - \mathbf{w}^\star\|^2)^{-1}} \\
\leq& \frac{D(\mathbf{w}^0) - D^\star}{1 + \rho \underline{\lambda} (D(\mathbf{w}^0) - D^\star) k / (\eta \tilde{M}_k^2)}.
\end{aligned}$$

Note that the above inequality is equivalent to (19) since $(D(\mathbf{w}^k))_{k=0}^\infty$ is non-increasing.

G. Proof of Theorem 3

Let $\mathbf{w} \in S^{\perp}$. Note that $\|P_{S^{\perp}}(\tilde{\mathbf{x}}(\mathbf{w}))\| = \|\tilde{\mathbf{x}}(\mathbf{w}) - P_S(\tilde{\mathbf{x}}(\mathbf{w}))\| \leq \|\tilde{\mathbf{x}}(\mathbf{w}) - \mathbf{x}^{\star}\|$. It follows from (37) that

$$\|P_{S^{\perp}}(\tilde{\mathbf{x}}(\mathbf{w}))\| \leq \|\tilde{\mathbf{x}}(\mathbf{w}) - \mathbf{x}^{\star}\| \leq \sqrt{2L(D(\mathbf{w}) - D^{\star})}. \tag{41}$$

Also note that

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

On the other hand, for any dual optimum $\mathbf{w}^{\star} \in S^{\perp}$, we have $-F^{\star} = D^{\star} \geq \langle \mathbf{w}^{\star}, \tilde{\mathbf{x}}(\mathbf{w}) \rangle - F(\tilde{\mathbf{x}}(\mathbf{w}))$, which leads to

$$F(\tilde{\mathbf{x}}(\mathbf{w})) - F^{\star} \geq \langle \mathbf{w}^{\star}, P_{S^{\perp}}(\tilde{\mathbf{x}}(\mathbf{w})) \rangle.$$

As a result,

$$-\|\mathbf{w}^{\star}\| \cdot \|P_{S^{\perp}}(\tilde{\mathbf{x}}(\mathbf{w}))\| \leq F(\tilde{\mathbf{x}}(\mathbf{w})) - F^{\star} \leq \|\mathbf{w}\| \cdot \|P_{S^{\perp}}(\tilde{\mathbf{x}}(\mathbf{w}))\|. \tag{42}$$

Combining (41) and (42) with Proposition 3 and Theorem 2 completes the proof.

H. Proof of Lemma 5

From Lemma 4,

$$\begin{aligned} D(\mathbf{w}^{k+1}) - D(\mathbf{w}^k) \leq & \langle \hat{\mathbf{x}}(\mathbf{w}^k), \mathbf{w}^{k+1} - \mathbf{w}^k \rangle + (\mathbf{w}^{k+1} - \mathbf{w}^k)^T (\Lambda_L \otimes I_d)(\mathbf{w}^{k+1} - \mathbf{w}^k) + n\bar{\epsilon} \\ \leq & -\hat{\rho} \langle \hat{\mathbf{x}}(\mathbf{w}^k), (H_{\mathcal{G}^k} \otimes I_d) \hat{\mathbf{x}}(\mathbf{w}^k) \rangle + n\bar{\epsilon}, \end{aligned}$$

where the last inequality can be derived in a similar way as in the proof of Lemma 1.

I. Proof of Lemma 6

Let $k \geq 0$. For each $\{i,j\} \in \tilde{\mathcal{E}}^k$, let $t^k_{\{i,j\}} \in \{k, \ldots, k+B-1\}$ be such that $\{i,j\} \in \mathcal{E}^{t^k_{\{i,j\}}}$. Then, from Lemma 4, Proposition 1, and (5),

$$\begin{aligned} & \|\hat{x}_i(w_i^k) - \hat{x}_i(w_i^{t^k_{\{i,j\}}})\|^2 \\ \leq & 3\|\hat{x}_i(w_i^k) - \tilde{x}_i(w_i^k)\|^2 + 3\|\tilde{x}_i(w_i^k) - \tilde{x}_i(w_i^{t^k_{\{i,j\}}})\|^2 + 3\|\tilde{x}_i(w_i^{t^k_{\{i,j\}}}) - \hat{x}_i(w_i^{t^k_{\{i,j\}}})\|^2 \\ \leq & 12L_i\bar{\epsilon} + 3\|\tilde{x}_i(w_i^k) - \tilde{x}_i(w_i^{t^k_{\{i,j\}}})\|^2 \\ \leq & 12L_i\bar{\epsilon} + 3B \sum_{t=k}^{k+B-1} \|\tilde{x}_i(w_i^{t+1}) - \tilde{x}_i(w_i^t)\|^2 \end{aligned}$$

$$\leq 12L_i\bar{\epsilon} + 3BL_i^2 \sum_{t=k}^{k+B-1} \|w_i^{t+1} - w_i^t\|^2.$$

Hence, similar to (38) and (39), we derive

$$\begin{aligned}
&\sum_{\{i,j\}\in\tilde{\mathcal{E}}^k} (\|\hat{x}_i(w_i^k) - \hat{x}_i(w_i^{t^k_{\{i,j\}}})\|^2 + \|\hat{x}_j(w_j^k) - \hat{x}_j(w_j^{t^k_{\{i,j\}}})\|^2) \\
\leq & 3B \sum_{\{i,j\}\in\tilde{\mathcal{E}}^k} \sum_{t=k}^{k+B-1} (L_i^2\|w_i^{t+1} - w_i^t\|^2 + L_j^2\|w_j^{t+1} - w_j^t\|^2) + 12\bar{\epsilon} \sum_{\{i,j\}\in\tilde{\mathcal{E}}^k} (L_i + L_j) \\
\leq & 3B\hat{\varpi}\bar{\alpha}^2\delta L\Big(\sum_{t=k}^{k+B-1} \hat{\mathbf{x}}(\mathbf{w}^t)^T (H_{\mathcal{G}^t} \otimes I_d)\hat{\mathbf{x}}(\mathbf{w}^t)\Big) + 12\hat{\varpi} \sum_{i\in\mathcal{V}} L_i\bar{\epsilon}
\end{aligned}$$

and $\sum_{\{i,j\}\in\tilde{\mathcal{E}}^k} \|\hat{x}_i(w_i^{t^k_{\{i,j\}}}) - \hat{x}_j(w_j^{t^k_{\{i,j\}}})\|^2 \leq \sum_{t=k}^{k+B-1} \hat{\mathbf{x}}(\mathbf{w}^t)^T (H_{\mathcal{G}}^t \otimes I_d)\hat{\mathbf{x}}(\mathbf{w}^t)/\underline{h}$. The above two inequalities, together with

$$\begin{aligned}
&\langle \hat{\mathbf{x}}(\mathbf{w}^k), (L_{\tilde{\mathcal{G}}^k} \otimes I_d)\hat{\mathbf{x}}(\mathbf{w}^k)\rangle = \sum_{\{i,j\}\in\tilde{\mathcal{E}}^k} \|\hat{x}_i(w_i^k) - \hat{x}_j(w_j^k)\|^2 \\
\leq & 3 \sum_{\{i,j\}\in\tilde{\mathcal{E}}^k} (\|\hat{x}_i(w_i^k) - \hat{x}_i(w_i^{t^k_{\{i,j\}}})\|^2 + \|\hat{x}_j(w_j^k) - \hat{x}_j(w_j^{t^k_{\{i,j\}}})\|^2) + 3 \sum_{\{i,j\}\in\tilde{\mathcal{E}}^k} \|\hat{x}_i(w_i^{t^k_{\{i,j\}}}) - \hat{x}_j(w_j^{t^k_{\{i,j\}}})\|^2,
\end{aligned}$$

complete the proof.

J. Proof of Lemma 7

According to Assumption 1, there exists $r_c \in (0,\infty)$ such that $B(0,r_c) \subseteq \bigcap_{i=1}^n X_i$. In addition, since $B(0,r_c)$ is compact and each f_i is strongly convex, $-\infty < \underline{F} \leq \bar{F} < \infty$, which, along with Proposition 2, implies $0 \leq \hat{M} < \infty$. We first prove that $D(\mathbf{w}^k) \leq \bar{D}\ \forall k \geq 0$ by induction. For $k = 0, 1, \ldots, B-1$, Lemma 5 suggests that $D(\mathbf{w}^k) \leq D(\mathbf{w}^0) + Bn\bar{\epsilon} \leq \bar{D}$. Now suppose $D(\mathbf{w}^{k_1}) \leq \bar{D}$ for some $k_1 \geq 0$. If $D(\mathbf{w}^{k_1+B}) \leq D(\mathbf{w}^{k_1})$, then clearly $D(\mathbf{w}^{k_1+B}) \leq \bar{D}$. Otherwise, note from Lemma 5 that $0 \leq D(\mathbf{w}^{k_1+B}) - D(\mathbf{w}^{k_1}) \leq -\hat{\rho}(\sum_{t=k_1}^{k_1+B-1} \langle \hat{\mathbf{x}}(\mathbf{w}^t), (H_{\mathcal{G}^t} \otimes I_d)\hat{\mathbf{x}}(\mathbf{w}^t)\rangle) + Bn\bar{\epsilon}$. This, along with Lemma 6, implies $\langle \hat{\mathbf{x}}(\mathbf{w}^{k_1}), (L_{\tilde{\mathcal{G}}^{k_1}} \otimes I_d)\hat{\mathbf{x}}(\mathbf{w}^{k_1})\rangle \leq (\gamma_1 Bn/\hat{\rho} + \gamma_2)\bar{\epsilon}$. Furthermore, due to (23), $\|P_{S^\perp}(\hat{\mathbf{x}}(\mathbf{w}^{k_1}))\| \leq \sqrt{\frac{1}{\lambda_a}\langle \hat{\mathbf{x}}(\mathbf{w}^{k_1}), (L_{\tilde{\mathcal{G}}^{k_1}} \otimes I_d)\hat{\mathbf{x}}(\mathbf{w}^{k_1})\rangle} \leq \sqrt{\frac{(\gamma_1 Bn/\hat{\rho} + \gamma_2)\bar{\epsilon}}{\lambda_a}} \leq \beta r_c$, where $\lambda_a > 0$ because $\tilde{\mathcal{G}}^k\ \forall k \geq 0$ are connected. Then, consider the lemma below.

Lemma 10. *Suppose Assumption 1 and* (22) *hold. For any* $\mathbf{w} \in S^\perp$*, if* $\|P_{S^\perp}(\hat{\mathbf{x}}(\mathbf{w}))\| < r_c$*, then*

$$\|\mathbf{w}\| \leq \frac{\bar{F} - \underline{F} + n\bar{\epsilon}}{r_c - \|P_{S^\perp}(\hat{\mathbf{x}}(\mathbf{w}))\|}. \tag{43}$$

Proof. Let $\mathbf{w} \in S^\perp$. Due to (4), we have $q_i(\tilde{x}_i(w_i), w_i) \geq q_i(x_i, w_i)\ \forall x_i \in X_i\ \forall i \in \mathcal{V}$, which leads to $\langle \tilde{\mathbf{x}}(\mathbf{w}), \mathbf{w}\rangle - F(\tilde{\mathbf{x}}(\mathbf{w})) \geq \langle \mathbf{x}, \mathbf{w}\rangle - F(\mathbf{x})\ \forall \mathbf{x} \in X_1 \times X_2 \times \ldots \times X_n$. Let $\mathbf{x}' = (x_1'^T, \ldots, x_n'^T)^T$, where $x_i' = r_c \frac{w_i}{\|w_i\|}$ if $w_i \neq \mathbf{0}_d$ and $x_i' = \mathbf{0}_d$ otherwise. Note that $x_i' \in B(\mathbf{0}_d, r_c) \subseteq \bigcap_{j=1}^n X_j$ $\forall i \in \mathcal{V}$. Thus, $\langle \tilde{\mathbf{x}}(\mathbf{w}), \mathbf{w}\rangle - F(\tilde{\mathbf{x}}(\mathbf{w})) \geq \langle \mathbf{x}', \mathbf{w}\rangle - F(\mathbf{x}') \geq r_c\|\mathbf{w}\| - \bar{F}$. This, along with $\langle \hat{\mathbf{x}}(\mathbf{w}), \mathbf{w}\rangle - F(\hat{\mathbf{x}}(\mathbf{w})) \geq \langle \tilde{\mathbf{x}}(\mathbf{w}), \mathbf{w}\rangle - F(\tilde{\mathbf{x}}(\mathbf{w})) - n\bar{\epsilon}$ and $F(\hat{\mathbf{x}}(\mathbf{w})) \geq \underline{F}$, implies $\langle \hat{\mathbf{x}}(\mathbf{w}), \mathbf{w}\rangle - \underline{F} \geq r_c\|\mathbf{w}\| - \bar{F} - n\bar{\epsilon}$. Further, because $\mathbf{w} \in S^\perp$, we have $\langle \hat{\mathbf{x}}(\mathbf{w}), \mathbf{w}\rangle \leq \|P_{S^\perp}(\hat{\mathbf{x}}(\mathbf{w}))\| \cdot \|\mathbf{w}\|$. It then follows that (43) is satisfied. □

From Lemma 10, we have $\|\mathbf{w}^{k_1}\| \leq \frac{\bar{F}-\underline{F}+n\bar{\epsilon}}{r_c(1-\beta)}$. This implies $D(\mathbf{w}^{k_1}) \leq \max\{D(\mathbf{w}) : \|\mathbf{w}\| \leq \frac{\bar{F}-\underline{F}+n\bar{\epsilon}}{r_c(1-\beta)},\ \mathbf{w} \in S^\perp\}$. Due again to Lemma 5, we have $D(\mathbf{w}^{k_1+B}) \leq \bar{D}$. This completes the proof of $D(\mathbf{w}^k) \leq \bar{D}\ \forall k \geq 0$. As a result, $\|\mathbf{w}^k\| \leq \hat{M}$.

K. Proof of Theorem 4

Let $k \geq 0$ and $\mathbf{w}^\star$ be any optimal solution to (3). Note that $\|\mathbf{w}^k - \mathbf{w}^\star\| \leq \|\mathbf{w}^k\| + \|\mathbf{w}^\star\| \leq \hat{N}_k + \|\mathbf{w}^\star\|$. Moreover, due to Lemma 7, $\hat{N}_k \leq \hat{M} < \infty\ \forall k \geq 0$. It follows from Lemma 4 that

$$\begin{aligned}
&D(\mathbf{w}^k) - D^\star \leq \langle \nabla D(\mathbf{w}^k), \mathbf{w}^k - \mathbf{w}^\star\rangle \\
\leq &\langle \hat{\mathbf{x}}(\mathbf{w}^k), \mathbf{w}^k - \mathbf{w}^\star\rangle + \langle \tilde{\mathbf{x}}(\mathbf{w}^k) - \hat{\mathbf{x}}(\mathbf{w}^k), \mathbf{w}^k - \mathbf{w}^\star\rangle \\
\leq &\|P_{S^\perp}(\hat{\mathbf{x}}(\mathbf{w}^k))\| \cdot \|\mathbf{w}^k - \mathbf{w}^\star\| + \|\tilde{\mathbf{x}}(\mathbf{w}^k) - \hat{\mathbf{x}}(\mathbf{w}^k)\| \cdot \|\mathbf{w}^k - \mathbf{w}^\star\| \\
\leq &(\hat{N}_k + \|\mathbf{w}^\star\|)\|P_{S^\perp}(\hat{\mathbf{x}}(\mathbf{w}^k))\| + (\hat{N}_k + \|\mathbf{w}^\star\|)\sqrt{2\sum_{i\in\mathcal{V}} L_i\bar{\epsilon}}.
\end{aligned}$$

This, together with the convexity of D, implies that

$$\begin{aligned}
\Big(D(\bar{\mathbf{w}}^k) - D^\star\Big)^2 &\leq \Big(\frac{1}{k+1}\sum_{t=0}^k (D(\mathbf{w}^t) - D^\star)\Big)^2 \leq \frac{1}{k+1}\sum_{t=0}^k (D(\mathbf{w}^t) - D^\star)^2 \\
&\leq \frac{2(\hat{N}_k + \|\mathbf{w}^\star\|)^2}{k+1}\Big(\sum_{t=0}^k \|P_{S^\perp}(\hat{\mathbf{x}}(\mathbf{w}^t))\|^2\Big) + 4(\hat{N}_k + \|\mathbf{w}^\star\|)^2 \sum_{i\in\mathcal{V}} L_i\bar{\epsilon}. \qquad (44)
\end{aligned}$$

We then provide an upper bound on the term $\sum_{t=0}^k \|P_{S^\perp}(\hat{\mathbf{x}}(\mathbf{w}^t))\|^2$ in (44). From Lemmas 6 and 5, we have

$$\begin{aligned}
\sum_{t=0}^k \|P_{S^\perp}(\hat{\mathbf{x}}(\mathbf{w}^t))\|^2 &\leq \frac{1}{\lambda_a}\sum_{t=0}^k \langle \hat{\mathbf{x}}(\mathbf{w}^t), (L_{\tilde{\mathcal{G}}^t} \otimes I_d)\hat{\mathbf{x}}(\mathbf{w}^t)\rangle \\
\leq &\frac{1}{\lambda_a}\Big(\sum_{t=0}^k \gamma_1 \sum_{s=t}^{t+B-1} \langle \hat{\mathbf{x}}(\mathbf{w}^s), (H_{\mathcal{G}^s} \otimes I_d)\hat{\mathbf{x}}(\mathbf{w}^s)\rangle\Big) + \frac{k+1}{\lambda_a}\gamma_2\bar{\epsilon}
\end{aligned}$$

$$\leq \frac{\gamma_1}{\hat{\rho}\lambda_a}\sum_{t=0}^{k}(D(\mathbf{w}^t)-D(\mathbf{w}^{t+B}))+\frac{k+1}{\lambda_a}(\frac{\gamma_1 Bn}{\hat{\rho}}+\gamma_2)\bar{\epsilon}.$$

Moreover, it can be shown that $\sum_{t=0}^{k}(D(\mathbf{w}^t)-D(\mathbf{w}^{t+B}))\leq B(\bar{D}-D^\star)$. To see this, note from Lemma 7 that when $k\geq B-1$, $\sum_{t=0}^{k}(D(\mathbf{w}^t)-D(\mathbf{w}^{t+B}))=\sum_{s=0}^{B-1}(D(\mathbf{w}^s)-D(\mathbf{w}^{B+k-s}))\leq \sum_{s=0}^{B-1}(\bar{D}-D^\star)=B(\bar{D}-D^\star)$. When $k\leq B-1$, we still have $\sum_{t=0}^{k}(D(\mathbf{w}^t)-D(\mathbf{w}^{t+B}))\leq (k+1)(\bar{D}-D^\star)\leq B(\bar{D}-D^\star)$. Therefore,

$$\sum_{t=0}^{k}\|P_{S^\perp}(\hat{\mathbf{x}}(\mathbf{w}^t))\|^2\leq\frac{\gamma_1}{\hat{\rho}\lambda_a}B(\bar{D}-D^\star)+\frac{k+1}{\lambda_a}(\frac{\gamma_1 Bn}{\hat{\rho}}+\gamma_2)\bar{\epsilon}. \tag{45}$$

Combining (45) with (44) results in (24). To prove (25), note that $\langle\hat{\mathbf{x}}(\mathbf{w}^t),\mathbf{w}^t\rangle-F(\hat{\mathbf{x}}(\mathbf{w}^t))\geq\langle\tilde{\mathbf{x}}(\mathbf{w}^t),\mathbf{w}^t\rangle-F(\tilde{\mathbf{x}}(\mathbf{w}^t))-n\bar{\epsilon}\geq -F^\star-n\bar{\epsilon}$. Thus, from Lemma 7, $F(\hat{\mathbf{x}}(\mathbf{w}^t))-F^\star\leq\langle\hat{\mathbf{x}}(\mathbf{w}^t),\mathbf{w}^t\rangle+n\bar{\epsilon}\leq\|\mathbf{w}^t\|\cdot\|P_{S^\perp}(\hat{\mathbf{x}}(\mathbf{w}^t))\|+n\bar{\epsilon}$. Then, due to the convexity of F,

$$\begin{aligned}F(\bar{\mathbf{x}}^k)-F^\star&\leq\frac{1}{k+1}\sum_{t=0}^{k}(F(\hat{\mathbf{x}}(\mathbf{w}^t))-F^\star)\leq\frac{\hat{N}_k}{k+1}(\sum_{t=0}^{k}\|P_{S^\perp}(\hat{\mathbf{x}}(\mathbf{w}^t))\|)+n\bar{\epsilon}\\&\leq\frac{\hat{N}_k}{\sqrt{k+1}}\sqrt{\sum_{t=0}^{k}\|P_{S^\perp}(\hat{\mathbf{x}}(\mathbf{w}^t))\|^2}+n\bar{\epsilon}.\end{aligned}$$

It follows from (45) that (25) holds. Also, since $\frac{1}{k+1}\sum_{t=0}^{k}P_S(\hat{\mathbf{x}}(\mathbf{w}^t)\in S$,

$$\begin{aligned}\|P_{S^\perp}(\bar{\mathbf{x}}^k)\|^2&=\|\bar{\mathbf{x}}^k-P_S(\bar{\mathbf{x}}^k)\|^2\leq\|\bar{\mathbf{x}}^k-\frac{1}{k+1}\sum_{t=0}^{k}P_S(\hat{\mathbf{x}}(\mathbf{w}^t))\|^2\\&\leq\frac{1}{k+1}\sum_{t=0}^{k}\|\hat{\mathbf{x}}(\mathbf{w}^t)-P_S(\hat{\mathbf{x}}(\mathbf{w}^t))\|^2=\frac{1}{k+1}\sum_{t=0}^{k}\|P_{S^\perp}(\hat{\mathbf{x}}(\mathbf{w}^t))\|^2.\end{aligned}$$

Due again to (45), we obtain (27). Finally, since $\bar{\mathbf{x}}^k\in X_1\times\cdots\times X_n$, $-F^\star=D^\star\geq\langle\mathbf{w}^\star,\bar{\mathbf{x}}^k\rangle-F(\bar{\mathbf{x}}^k)$, i.e., $F(\bar{\mathbf{x}}^k)-F^\star\geq-\|\mathbf{w}^\star\|\cdot\|P_{S^\perp}(\bar{\mathbf{x}}^k)\|$. This and (27) yield (26).

REFERENCES

[1] M. G. Rabbat and R. D. Nowak, "Distributed optimization in sensor networks," in *Proc. International Symposium on Information Processing in Sensor Networks*, Berkeley, CA, 2004, pp. 20–27.

[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. Giselsson, M. D. Doan, T. Keviczky, B. Schutter, and A. Rantzer, "Accelerated gradient methods and dual decomposition in distributed model predictive control," *Automatica*, vol. 49, no. 3, pp. 829–833, 2013.

[4] 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.

[5] S. Lee and A. Nedić, "Distributed random projection algorithm for convex optimization," *IEEE Journal of Selected Topics in Signal Processing, a special issue on Adaptation and Learning over Complex Networks*, vol. 7, no. 2, pp. 221–229, 2013.

[6] P. Lin, W. Ren, and Y. Song, "Distributed multi-agent optimization subject to nonidentical constraints and communication delays," *Automatica*, vol. 65, pp. 120–131, 2016.

[7] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Transactions on Control of Network Systems*, 2017.

[8] D. Jakovetić, J. Xavier, and J. Moura, "Fast distributed gradient methods," *IEEE Transactions on Automatic Control*, vol. 59, no. 5, pp. 1131–1146, 2014.

[9] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: an exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[10] G. Qu and N. Li, "Accelerated distributed Nesterov gradient descent," *arXiv preprint arXiv:1705.07176*, 2017.

[11] C. Xi and U. Khan, "DEXTRA: A fast algorithm for optimization over directed graphs," *IEEE Transactions on Automatic Control*, 2017.

[12] 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.

[13] ——, "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.

[14] 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.

[15] B. Johansson, P. Soldati, and M. Johansson, "Mathematical decomposition techniques for distributed cross-layer optimization of data networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1535–1547, 2006.

[16] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[17] J. Koshal, A. Nedić, and U. V. Shanbhag, "Multiuser optimization: Distributed algorithms and error analysis," *SIAM Journal on Optimization*, vol. 21, no. 3, pp. 1046–1081, 2011.

[18] J. Duchi, A. Agarwal, and M. Wainwright, "Dual averaging for distributed optimization: Convergence and network scaling," *IEEE Transactions on Automatic Control*, vol. 57, no. 3, pp. 592–606, 2012.

[19] M. Zhu and S. Martínez, "On distributed convex optimization under inequality and equality constraints," *IEEE Transactions on Automatic Control*, vol. 57, no. 1, pp. 151–164, 2012.

[20] P. Patrinos and A. Bemporad, "An accelerated dual gradient-projection algorithm for embedded linear model predictive control," *IEEE Transactions on Automatic Control*, vol. 59, no. 1, pp. 18 – 33, 2013.

[21] T. Chang, A. Nedić, and A. Scaglione, "Distributed constrained optimization by consensus-based primal-dual perturbation method," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1524–1538, 2014.

[22] I. Necoara and V. Nedelcu, "Rate analysis of inexact dual first-order methods application to dual decomposition," *IEEE Transactions on Automatic Control*, vol. 59, no. 5, pp. 1232–1243, 2014.

[23] P. Bianchi, W. Hachem, and F. Iutzeler, "A coordinate descent primal-dual algorithm and application to distributed asynchronous optimization," *IEEE Transactions on Automatic Control*, vol. 61, no. 10, pp. 2947–2957, 2016.

[24] K. Margellos, A. Falsone, S. Garatti, and M. Prandini, "Proximal minimization based distributed convex optimization," in *Proc.American Control Conference*, Boston, MA, 2016, pp. 2466–2471.

[25] B. Johansson, M. Rabi, and M. Johansson, "A randomized incremental subgradient method for distributed optimization in networked systems," *SIAM Journal on Optimization*, vol. 20, no. 3, pp. 1157–1170, 2009.

[26] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Incremental stochastic subgradient algorithms for convex optimization," *SIAM Journal on Optimization*, vol. 20, no. 2, pp. 691–717, 2009.

[27] E. Wei, A. Ozdaglar, and A. Jadbabaie, "A distributed newton method for network utility maximization–i: Algorithm," *IEEE Transactions on Automatic Control*, vol. 58, no. 9, pp. 2162–2175, 2013.

[28] ——, "A distributed newton method for network utility maximization–part ii: Convergence," *IEEE Transactions on Automatic Control*, vol. 58, no. 9, pp. 2176–2188, 2013.

[29] D. Varagnolo, F. Zanella, A. Cenedese, G. Pillonetto, and L. Schenato, "Newton-Raphson consensus for distributed convex optimization," *IEEE Transactions on Automatic Control*, vol. 61, no. 4, pp. 994–1009, 2016.

[30] J. Lu and C. Y. Tang, "Zero-gradient-sum algorithms for distributed convex optimization: The continuous-time case," *IEEE Transactions on Automatic Control*, vol. 57, no. 9, pp. 2348–2354, 2012.

[31] S. S. Kia, J. Cortés, and S. Martínez, "Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication," *Automatica*, vol. 55, pp. 254–264, 2015.

[32] Y. Lou, Y. Hong, and S. Wang, "Distributed continuous-time approximate projection protocols for shortest distance optimization problems," *Automatica*, vol. 69, pp. 289–297, 2016.

[33] 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.

[34] H. Lakshmanan and D. P. de Farias, "Decentralized resource allocation in dynamic networks of agents," *SIAM Journal on Optimization*, vol. 19, no. 2, p. 911940, 2008.

[35] X. Wu and J. Lu, "Fenchel dual gradient methods for distributed convex optimization over time-varying networks," in *Proc. IEEE Conference on Decision and Control*, Melbourne, Australia, 2017, pp. 2894–2899.

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

[37] J. Lu and M. Johansson, "Convergence analysis of approximate primal solutions in dual first-order methods," *SIAM Journal on Optimization*, vol. 26, no. 4, pp. 2430–2467, 2016.

[38] 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.

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

[40] 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.

[41] X. Yuan, J. Shao, and L. Zhang, "The six classes of trees with the largest algebraic connectivity," *Discrete Applied Mathematics*, vol. 156, pp. 757–769, 2008.

[42] D. Kempe, A. Dobra, and J. Gehrke, "Gossip-based computation of aggregate information," in *Proc. IEEE Symposium on Foundations of Computer Science*, Cambridge, MA, 2003, pp. 482–491.

[43] O. Devolder, F. Glineur, and Y. Nesterov, "First-order methods of smooth convex optimization with inexact oracle," *Mathematical Programming*, vol. 146, no. 1-2, pp. 37–75, 2014.

[44] D. P. Bertsekas, *Convex optimization theory*. Belmont, MA: Athena Scientific, 2009.

[45] B. T. Polyak, *Introduction to Optimization*. New York, NY: Optimization Software, Inc., 1987.

Xuyang Wu (SM'17) received the B.S. degree in Information and Computing Sciences 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.

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.