Dual Consensus Proximal Algorithm for Multi-Agent Sharing Problems

Sulaiman A. Alghunaim, Qi Lyu, Ming Yan, and Ali H. Sayed, *Fellow, IEEE*

***Abstract*—This work considers multi-agent sharing optimization problems, where each agent owns a local smooth function plus a non-smooth function, and the network seeks to minimize the sum of all local functions plus a coupling composite function (possibly non-smooth). For this non-smooth setting, centralized algorithms are known to converge linearly under certain conditions. On the other hand, decentralized algorithms have not been shown to achieve linear convergence under the same conditions. In this work, we propose a decentralized proximal primal-dual algorithm and establish its linear convergence under weaker conditions than existing decentralized works. Our result shows that decentralized algorithms match the linear rate of centralized algorithms without any extra condition. Finally, we provide numerical simulations that illustrate the theoretical findings and show the advantages of the proposed method.**

***Index Terms*—Multi-agent optimization, sharing problem, dual consensus, proximal algorithm, linear convergence.**

I. Introduction

We consider a network of K agents connected by some topology. The goal of agent k is to find its corresponding solution, denoted by $w_k^\star \in \mathbb{R}^{Q_k}$, of the following multi-agent optimization problem:

$$\min_{w_1,\ldots,w_K} \quad \sum_{k=1}^{K} \Big(J_k(w_k) + R_k(w_k) \Big) + h\left(\sum_{k=1}^{K} B_k w_k \right), \quad (1)$$

where each $J_k : \mathbb{R}^{Q_k} \to \mathbb{R}$ is a smooth convex function, $R_k : \mathbb{R}^{Q_k} \to \mathbb{R} \cup \{+\infty\}$ and $h : \mathbb{R}^{E} \to \mathbb{R} \cup \{+\infty\}$ are convex (possibly non-smooth) functions, and $B_k \in \mathbb{R}^{E \times Q_k}$. The two functions ($J_k$ and R_k) and the matrix B_k are known privately by agent k, while all agents know the function h. Problem (1) is the *sharing problem* [2], where the individual variables $\{w_k\}_{k=1}^K$ are coupled through a composite coupling function h. Formulation (1) arises in various engineering and machine learning applications, such as image processing [3], distributed basis pursuit [4], smart grids [5], [6], and learning problems over distributed models [2], [7], [8]. In this work, we study the linear convergence of *decentralized algorithms* (i.e., methods that only use local communications between directly connected agents) for problem (1).

Many methods can solve general problems of the form (1) – see [9]–[15] and references therein. Applying these methods directly to problem (1) result in *centralized implementations*, where a global communication step is needed to compute $\sum_{k=1}^K B_k w_k$. When h is *nonsmooth*, centralized algorithms have been shown to converge linearly if each $R_k = 0$, $\sum_{k=1}^K J_k(w_k)$ is strongly-convex, and the matrix $B = [B_1 \ \cdots \ B_k]$ has full row rank [9]–[11]. When h is *smooth*, centralized algorithms have also been shown to converge linearly if $\sum_{k=1}^K J_k(w_k)$ is strongly-convex [12]–[14].

Existing linear convergence results for decentralized algorithms solving the sharing problem (1) have only been established under special cases and require stronger assumptions compared to the ones used to establish linear convergence of centralized algorithms. In this work, we close this theoretical *linear convergence* gap between decentralized and centralized algorithms for problem (1). In particular, we propose a novel decentralized algorithm for problem (1) and establish its linear convergence under conditions matching the ones used for centralized algorithms.

A. Related Works

Sharing problems of the form (1) have been studied for many years [16], [17] – see the discussion in [2]. However, most works that study *decentralized methods* for the sharing problem consider different and/or special setups from this work. For example, the works [17]–[29] study the case where agents are coupled through equality constraints (i.e., $h(x) = 0$ if $x = 0$ and $h(x) = \infty$ otherwise). The works [30], [31] study inequality constrained sharing problems (i.e., $h(x) = 0$ if $x \leq 0$ and $h(x) = \infty$ otherwise). The works [32]–[35] consider a *smooth* coupling function h, and the work [36] considers conic coupling constraints. While decentralized algorithms for sharing problems have been studied before, their linear convergence under decentralized setups are not well established compared to centralized algorithms as we explain next.

For a general *non-smooth* function h, centralized algorithms are known to converge linearly when each $R_k(w_k) = 0$, $\sum_{k=1}^K J_k(w_k)$ is strongly-convex, and the matrix $B =$

Manuscript received November 17, 2020; revised June 14, 2021; accepted September 13, 2021. Date of publication September 24, 2021; date of current version October 15, 2021. The work of Q. Lyu and M. Yan are supported by National Science Foundation (NSF) under Grant DMS-2012439. A short preliminary conference version of this work appears in [1]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Subhro Das. *(Corresponding author: Sulaiman A. Alghunaim.)*

Sulaiman A. Alghunaim is with the Department of Electrical Engineering, Kuwait University, Kuwait 13060, Kuwait (e-mail: salghunaim@ucla.edu).

Qi Lyu is with the Department of Computational Mathematics, Science, and Engineering, Michigan State University (MSU), East Lansing, MI 48824 USA (e-mail: lyuqi1@msu.edu).

Ming Yan is with the Department of Computational Mathematics, Science, and Engineering, Michigan State University (MSU), East Lansing, MI 48824 USA, and also with the Department of Mathematics, Michigan State University (MSU), East Lansing, MI 48824 USA (e-mail: myan@msu.edu).

Ali H. Sayed is with the School of Engineering, Ecole Polytechnique Federale de Lausanne, 1015 Lausanne, Switzerland (e-mail: ali.sayed@epfl.ch).

Digital Object Identifier 10.1109/TSP.2021.3114978

TABLE I
COMPARISON WITH EXISTING *DECENTRALIZED METHODS* LINEAR CONVERGENCE RESULTS FOR PROBLEM (1). HERE, $B = [B_1 \ \cdots \ B_K]$, SC MEANS STRONGLY CONVEX, AND LC MEANS LINEAR CONSTRAINTS. THE COST J_k IS SMOOTH

Reference	J_k	R_k	B_k	h	**Additional comments**
[19]	SC	0	scalar $b_k \neq 0$	LC: $\sum_{k=1}^K b_k w_k = 0$	randomized coordinate updates
[20]	SC	0	I	LC: $\sum_{k=1}^K w_k = 0$	time-varying networks
[21], [22]	SC	0	I	LC: $\sum_{k=1}^K w_k - b_k = 0$	
[23]	SC	0	each B_k has full row rank	LC: $\sum_{k=1}^K B_k w_k - b_k = 0$	
[25]	SC	0	non-zero rows of B_k are linearly independent	LC: $\sum_{k=1}^K B_k w_k - b_k = 0$	
[34]	SC	0	any B_k	smooth	variance reduced method
[35]	0	non-smooth SC	any B_k	smooth	algorithm requires solving inner subproblems
This work	SC SC	0 non-smooth	B full row rank any B_k	non-smooth smooth	matches centralized algorithms conditions [9]–[14]

$[B_1 \ \cdots \ B_k]$ has full row rank [9]–[11]. On the other hand, existing *decentralized* linear convergence results have only been established for the special case of an equality constraint coupling function h (i.e., h is an indicator function of equality constraints) [19]–[23], [25], [26]. Moreover, these results require $B_k = I$ [19]–[22], [26] or each matrix B_k to have full row rank [23], [25].

For a *smooth* function h, centralized algorithms can achieve linear convergence if $\sum_{k=1}^K J_k(w_k)$ is strongly-convex [12]–[15]. Decentralized algorithms have been shown to achieve linear convergence if h is smooth, albeit under special and/or different cases from the centralized case. In particular, the work [34] established linear convergence for the case where h is smooth and each $R_k(w_k) = 0$. The work [35] studied problem (1) with $J_k(w_k) = 0$ and established linear convergence for smooth h and strongly-convex $R_k(w_k)$. Table I summarizes the conditions used to establish linear convergence of *decentralized algorithms* for problem (1).

We remark that problem (1) can be reformulated into an equivalent decentralized problem (see (12)) amenable to decentralized solutions. The same algorithms that solve (1) can also be used to solve the equivalent problem to get decentralized implementations. However, their linear convergence guarantees are not satisfied for the decentralized formulation, and the linear convergence results from [9]–[15] are not applicable to decentralized setups – see Remark 2.

Finally, note that if we choose h to be the indicator function of the consensus constraint: $w_1 = \cdots = w_K$, then formulation (1) recovers the "consensus problem," where the agents share a common variable – see e.g., [37]–[44] and references therein. Algorithms solving the consensus problem are not generally applicable to the sharing problem [2, Ch. 7]. This is because, decentralized consensus algorithms exploit the network sparsity structure of the matrix $[B_1, \ldots, B_K]$. In the sharing formulation (1), this matrix is not necessarily sparse; moreover, the matrix B_k is privately known by agent k alone.

B. Contribution

A natural question is whether decentralized algorithms can achieve linear convergence under the same conditions as centralized algorithms. In this work, we give a positive answer to this question. In particular, we propose a decentralized algorithm and establish its linear convergence to the *exact* solution of (1) under weaker conditions than existing decentralized algorithms – see Table I. Below, we list the main contributions of this work:

- We reformulate problem (1) into an equivalent saddle-point problem, which is amenable to decentralized algorithms, and propose a dual consensus proximal algorithm (DCPA) to solve this equivalent problem.
- For a *smooth* function h, we show that DCPA converges linearly if $\sum_{k=1}^K J_k(w_k)$ is strongly-convex in the presence of non-smooth R_k. This result matches the linear convergence of centralized algorithms for smooth h.
- For a *non-smooth* function h, we show that DCPA converges linearly when each $R_k(w_k) = 0$, $\sum_{k=1}^K J_k(w_k)$ is strongly-convex, and the matrix $B = [B_1 \ \cdots \ B_k]$ has full row rank. This result closes a major theoretical gap in linear convergence between centralized and decentralized algorithms for sharing problems of the form (1).

We note that the preliminary work [1] studied a different algorithm for a special case where each $R_k(w_k) = 0$. Moreover, the linear convergence result in [1] requires a stronger assumption that each matrix B_k has full row rank.

Notation: We let I_S denote the $S \times S$ identity matrix, while the symbol $\mathbb{1}_N$ denotes the $N \times 1$ vector with all entries equal to one. The subscripts are dropped when there is no confusion. The vector formed by stacking $x_1, \ldots, x_N$ on top of each other is denoted by $\mathrm{col}\{x_j\}_{j=1}^N$. The block diagonal matrix with diagonal blocks $\{X_j\}_{j=1}^N$ is denoted by $\mathrm{blkdiag}\{X_j\}_{j=1}^N$. For any matrix A, we let $\sigma_{\max}(A)$ and $\underline{\sigma}(A)$ denote the largest and smallest *non-zero* singular values of A, respectively. For a square matrix A, we let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the largest and smallest eigenvalues of A, respectively. We let $\|x\|_A^2 = x^\mathsf{T} A x$. Given a function $f : \mathbb{R}^M \to \mathbb{R}$, its subdifferential $\partial f(x)$ at $x \in \mathbb{R}^M$ is the set of all subgradients at x. Its proximal operator with step-size μ is $\mathrm{prox}_{\mu f}(x) = \arg\min_u f(u) + \frac{1}{2\mu}\|u - x\|^2$. Its conjugate with domain $\mathbb{R}^M$ is $f^*(v) = \sup_x \ v^\mathsf{T} x - f(x)$. The function f is δ-smooth ($\delta >$

0) if $\|\nabla f(x) - \nabla f(y)\| \leq \delta\|x-y\|$ for all $x, y \in \mathbb{R}^M$. It is ν-strongly-convex ($\nu > 0$) if $(x-y)^{\mathsf{T}}(\nabla f(x) - \nabla f(y)) \geq \nu\|x-y\|^2$ for all $x, y \in \mathbb{R}^M$.

II. Decentralized Saddle-Point Formulation

In this section, we show how problem (1) can be reformulated into an equivalent saddle-point problem that is amenable to decentralized solutions.

A. Saddle-Point Formulation

We start by rewriting the problem (1) in a compact form. We define the network quantities:

$$\mathcal{w} \triangleq \operatorname{col}\{w_1, \ldots, w_K\} \in \mathbb{R}^Q, \quad Q \triangleq \sum_{k=1}^K Q_k, \tag{2a}$$

$$\mathcal{J}(\mathcal{w}) \triangleq \sum_{k=1}^K J_k(w_k), \quad \mathcal{R}(\mathcal{w}) \triangleq \sum_{k=1}^K R_k(w_k), \tag{2b}$$

$$B \triangleq \begin{bmatrix} B_1 & \cdots & B_K \end{bmatrix} \in \mathbb{R}^{E\times Q}. \tag{2c}$$

Using the above notation, problem (1) becomes

$$\min_{\mathcal{w}} \quad \mathcal{J}(\mathcal{w}) + \mathcal{R}(\mathcal{w}) + h\left(B\mathcal{w}\right). \tag{3}$$

Throughout this work, the following assumption holds.

Assumption 1 (Objective Function): The function $\mathcal{J} : \mathbb{R}^Q \to \mathbb{R}$ is δ-smooth and convex. The functions $\mathcal{R} : \mathbb{R}^Q \to \mathbb{R} \cup \{+\infty\}$ and $h : \mathbb{R}^E \to \mathbb{R} \cup \{+\infty\}$ are proper lower semi-continuous and convex. There exists $\mathcal{w}$ in the relative interior domain of $\mathcal{R}$ such that $B\mathcal{w}$ belongs to the relative interior domain of h. Problem (3) has a solution $\mathcal{w}^\star$.

Under Assumption 1, problem (3) is equivalent to the saddle-point problem [45, Proposition 19.18]:

$$\boxed{\min_{\mathcal{w}} \max_{y} \quad \mathcal{J}(\mathcal{w}) + \mathcal{R}(\mathcal{w}) + y^{\mathsf{T}}B\mathcal{w} - h^*(y),} \tag{4}$$

where $y \in \mathbb{R}^E$ is the dual variable. Moreover, $(\mathcal{w}^\star, y^\star)$ is an optimal solution of (4) if, and only if, it satisfies [45, Proposition 19.18]:

$$-B^{\mathsf{T}}y^\star - \nabla\mathcal{J}(\mathcal{w}^\star) \in \partial\mathcal{R}(\mathcal{w}^\star), \tag{5a}$$

$$B\mathcal{w}^\star \in \partial h^*(y^\star). \tag{5b}$$

Note that the dual variable y in (4) is multiplied by B, which couples all agents. Therefore, algorithms directly solving (4) cannot be implemented in a *decentralized* manner. Next, we reformulate the problem into another equivalent problem that is amenable to decentralized solutions.

B. Decentralized Saddle-Point Formulation

Let $\overline{\mathcal{J}}(\mathcal{w}) \triangleq \mathcal{J}(\mathcal{w}) + \mathcal{R}(\mathcal{w})$, then the dual problem of (3) is [45], [46]:

$$\max_{y} \quad -\overline{\mathcal{J}}^*(-B^{\mathsf{T}}y) - h^*(y). \tag{6}$$

The above problem is a decentralized consensus problem since $\overline{\mathcal{J}}^*(-B^{\mathsf{T}}y) = \sum_{k=1}^K \overline{J}_k^*(-B_k^{\mathsf{T}}y)$ where $\overline{J}_k^*$ is the conjugate of $\overline{J}_k \triangleq J_k + R_k$. Note that the conjugate function $\overline{J}_k^*$ does not have a closed-form expression in general, and its gradient is expensive to obtain. Thus, it is infeasible to solve (6) in its current form. Next, we consider the dual problem as a consensus problem and let y_k denote a local copy of y available at agent k. For simplicity, we introduce the following network quantities:

$$\mathcal{y} \triangleq \operatorname{col}\{y_k\}_{k=1}^K \in \mathbb{R}^{EK}, \tag{7a}$$

$$\mathcal{H}^*(\mathcal{y}) \triangleq \frac{1}{K}\sum_{k=1}^K h^*(y_k), \tag{7b}$$

$$\mathcal{B}_d \triangleq \operatorname{blkdiag}\{B_k\}_{k=1}^K, \tag{7c}$$

and a symmetric matrix $\mathcal{L} \in \mathbb{R}^{EK\times EK}$ such that:

$$\mathcal{L}\mathcal{y} = 0 \iff y_1 = \cdots = y_K. \tag{8}$$

Then, the dual problem (6) is equivalent to:

$$\max_{\mathcal{y}} \quad -\overline{\mathcal{J}}^*(-\mathcal{B}_d^{\mathsf{T}}\mathcal{y}) - \mathcal{H}^*(\mathcal{y}), \quad \text{s.t. } \mathcal{L}\mathcal{y} = 0. \tag{9}$$

Note that we will later choose a specific matrix $\mathcal{L}$ that is related to the network. Introducing an additional variable $\theta = -\mathcal{B}_d^{\mathsf{T}}\mathcal{y}$, we derive the following Lagrange dual function of problem (9):

$$\begin{aligned}\sup_{\mathcal{y},\theta} & -\overline{\mathcal{J}}^*(\theta) - \mathcal{H}^*(\mathcal{y}) + \mathcal{w}^{\mathsf{T}}(\theta + \mathcal{B}_d^{\mathsf{T}}\mathcal{y}) + \mathcal{x}^{\mathsf{T}}\mathcal{L}\mathcal{y},\\ &= \sup_{\theta}(\mathcal{w}^{\mathsf{T}}\theta - \overline{\mathcal{J}}^*(\theta)) + \sup_{\mathcal{y}}\left((\mathcal{B}_d\mathcal{w} + \mathcal{L}\mathcal{x})^{\mathsf{T}}\mathcal{y} - \mathcal{H}^*(\mathcal{y})\right),\\ &= \overline{\mathcal{J}}(\mathcal{w}) + \mathcal{H}(\mathcal{B}_d\mathcal{w} + \mathcal{L}\mathcal{x}).\end{aligned} \tag{10}$$

Therefore, the dual problem of (9) is:

$$\min_{\mathcal{w},\mathcal{x}} \quad \overline{\mathcal{J}}(\mathcal{w}) + \mathcal{H}(\mathcal{B}_d\mathcal{w} + \mathcal{L}\mathcal{x}), \tag{11}$$

and the saddle-point reformulation of problem (11) is [45, Proposition 19.18]:

$$\boxed{\min_{\mathcal{w},\mathcal{x}} \max_{\mathcal{y}} \quad \mathcal{J}(\mathcal{w}) + \mathcal{R}(\mathcal{w}) + \mathcal{y}^{\mathsf{T}}\mathcal{B}_d\mathcal{w} + \mathcal{y}^{\mathsf{T}}\mathcal{L}\mathcal{x} - \mathcal{H}^*(\mathcal{y}).} \tag{12}$$

Lemma 1 (Saddle-Point): Suppose that Assumption 1 holds and let $(\mathcal{w}^\star, \mathcal{x}^\star, \mathcal{y}^\star)$ be a saddle-point of (12), i.e.,

$$-\mathcal{B}_d^{\mathsf{T}}\mathcal{y}^\star - \nabla\mathcal{J}(\mathcal{w}^\star) \in \partial\mathcal{R}(\mathcal{w}^\star), \tag{13a}$$

$$\mathcal{L}\mathcal{y}^\star = 0, \tag{13b}$$

$$\mathcal{B}_d\mathcal{w}^\star + \mathcal{L}\mathcal{x}^\star \in \partial\mathcal{H}^*(\mathcal{y}^\star). \tag{13c}$$

Then it holds that $\mathcal{y}^\star = \mathbb{1}_K \otimes y^\star$ and the point $(\mathcal{w}^\star, y^\star)$ satisfy the optimality condition (5).

Proof: See Appendix A. ■

Remark 1 (Existence of $\mathcal{x}^\star$): Suppose that $\mathcal{w}^\star$ and $\mathcal{y}^\star = \mathbb{1}_K \otimes y^\star$ are given such that $(\mathcal{w}^\star, y^\star)$ satisfies (5). From matrix algebra [47], we can decompose $\mathcal{B}_d\mathcal{w}^\star = \frac{1}{K}\mathbb{1}_K \otimes B\mathcal{w}^\star + \mathcal{L}\hat{\mathcal{x}}$ into the null space and range space of the symmetric matrix $\mathcal{L}$. Then $\mathcal{x}^\star = -\hat{\mathcal{x}}$ satisfies (13c) since

$$\mathcal{B}_d\mathcal{w}^\star + \mathcal{L}\mathcal{x}^\star = \tfrac{1}{K}\mathbb{1}_K \otimes B\mathcal{w}^\star \overset{(5b)}{\in} \partial\mathcal{H}^*(\mathcal{y}^\star).$$

Note that for any $\mathcal{w}^\star$ and $\mathcal{y}^\star = \mathbb{1}_K \otimes y^\star$, the value of $\mathcal{x}^\star$ is not unique because adding a vector from the null space of $\mathcal{L}$ does not change the optimality condition (13c).

From Lemma 1 and Remark 1, we see that problem (12) is equivalent to problem (4). However, unlike problem (4), problem (12) can be solved in a decentralized because the matrices $\mathcal{B}_d$ and $\mathcal{L}$ encode the network sparsity structure.

Remark 2 (Partial Strong-Convexity): The decentralized saddle-point formulation (12) is only strongly convex with

respect to w and not strongly-convex with respect to (w, x). Moreover, $\mathcal{B}_d$ is not assumed to have any rank condition. Therefore, existing linear convergence results [9]–[14] on general saddle-point problems of form (12) are not applicable.

III. Proposed Decentralized Solution

In this section, we introduce our proposed algorithm. To do that, we first select $\mathcal{L}$ based on the network graph.

A. Network Combination Matrix

We introduce the network combination weights $\{a_{sk}\}$, where a_{sk} is a scalar used by agent k to scale information coming from agent s. We let $a_{sk} = 0$ if $s \notin \mathcal{N}_k$, where $\mathcal{N}_k$ denotes the set of agents directly connected to agent k through an edge, including agent k itself. We also introduce the network combination matrices:

$$A = [a_{sk}], \quad \mathcal{A} \triangleq A \otimes I_E. \tag{14}$$

Assumption 2 (Combination Matrix): We assume that the network is static and undirected. Moreover, the matrix A is symmetric, doubly stochastic, and primitive.

We choose $\mathcal{L}^2$ as follows:

$$\mathcal{L}^2 = \tfrac{1}{2}(I - \mathcal{A}). \tag{15}$$

Note that under Assumption 2, the eigenvalues of the matrix $\mathcal{A}$ belong to $(-1, 1]$ – see [40, Lemma F.4]. Thus, it holds that $0 \leq \mathcal{L}^2 < I$ and $0 < \underline{\sigma}^2(\mathcal{L}) \leq \sigma^2_{\max}(\mathcal{L}) < 1$. Note that the matrix $\mathcal{L}$ is defined as the square root of $\frac{1}{2}(I - \mathcal{A})$, which is properly defined. To see this, let us introduce the eigen-decomposition of the positive semidefinite matrix $\mathcal{L}^2 = \frac{1}{2}(I - \mathcal{A}) = \mathcal{U}\mathcal{D}^2\mathcal{U}^{\mathsf{T}}$ then $\mathcal{L}$ exists and equal to $\mathcal{L} = \mathcal{U}\mathcal{D}\mathcal{U}^{\mathsf{T}}$.

B. Dual Consensus Proximal Algorithm

To solve (12), we propose the following dual consensus proximal algorithm (DCPA). Initialize w_{-1}, y_{-1} with arbitrary values and let $x_{-1} = 0$. Choose step-sizes $\mu_w, \mu_y, \mu_x > 0$ and repeat for $i \geq 0$:

$$w_i = \operatorname{prox}_{\mu_w \mathcal{R}}(w_{i-1} - \mu_w \nabla \mathcal{J}(w_{i-1}) - \mu_w \mathcal{B}_d^{\mathsf{T}} y_{i-1}), \tag{16a}$$

$$v_i = y_{i-1} - \mathcal{L}^2 y_{i-1} + \mu_y \mathcal{B}_d(2w_i - w_{i-1}) + \mathcal{L}x_{i-1}, \tag{16b}$$

$$x_i = x_{i-1} - \mu_x \mathcal{L} v_i, \tag{16c}$$

$$y_i = \operatorname{prox}_{\mu_y \mathcal{H}^*}(v_i). \tag{16d}$$

Recall that $\mathcal{L}^2 = \frac{1}{2}(I - \mathcal{A})$ has the network structure but $\mathcal{L}$ does not necessarily have the network structure. We can make a simple change of variable to transform DCPA (16) into an equivalent and fully decentralized recursion. In particular, if we let $z_i = \mathcal{L}(x_i - \mathcal{L}y_i)$ and multiply the update (16c) by $\mathcal{L}$, then we can rewrite (16) into the following equivalent recursion:

$$w_i = \operatorname{prox}_{\mu_w \mathcal{R}}(w_{i-1} - \mu_w \nabla \mathcal{J}(w_{i-1}) - \mu_w \mathcal{B}_d^{\mathsf{T}} y_{i-1}), \tag{17a}$$

$$v_i = y_{i-1} + \mu_y \mathcal{B}_d(2w_i - w_{i-1}) + z_{i-1}, \tag{17b}$$

$$y_i = \operatorname{prox}_{\mu_y \mathcal{H}^*}(v_i), \tag{17c}$$

$$z_i = z_{i-1} - \mathcal{L}^2(\mu_x v_i + y_i - y_{i-1}). \tag{17d}$$

Since only $\mathcal{L}^2$ appears in (17), the k-th block vector of w_i, y_i, z_i can be updated by agent k only as listed in Algorithm 1. The step (18d) requires agent k to send $(\mu_x v_{k,i} + y_{k,i} - y_{k,i-1})$ to its immediate neighbors $\mathcal{N}_k$.

Algorithm 1: Dual Consensus Proximal Algorithm (DCPA).

Setting: Let $C = \frac{1}{2}(I - A) = [c_{sk}]$. Choose step-sizes $\mu_w > 0, \mu_y > 0, \mu_x > 0$. Let $z_{k,-1} = y_{k,-1} = 0$ and arbitrary $w_{k,-1}$.

For every agent k, repeat for $i \geq 0$:

$$w_{k,i} = \operatorname*{prox}_{\mu_w R_k}\big(w_{k,i-1} - \mu_w \nabla J_k(w_{k,i-1}) - \mu_w B_k^{\mathsf{T}} y_{k,i-1}\big), \tag{18a}$$

$$v_{k,i} = y_{k,i-1} + \mu_y B_k(2w_{k,i} - w_{k,i-1}) + z_{k,i-1}, \tag{18b}$$

$$y_{k,i} = \operatorname*{prox}_{\frac{\mu_y}{K} h^*}(v_{k,i}), \tag{18c}$$

$$z_{k,i} = z_{k,i-1} - \sum_{s \in \mathcal{N}_k} c_{sk}(\mu_x v_{s,i} + y_{s,i} - y_{s,i-1}). \tag{18d}$$

Remark 3 (Intuition for the Update (16)): Algorithm (16) is not a typical proximal primal-dual method. The main difference lies in the update of x_i in (16c), where it uses the auxiliary variable v_i instead of the dual estimate y_i. This is inspired from [48] albeit for a different problem. This is a critical step that allows us to establish linear convergence when h is nonsmooth.

The term $\mathcal{B}_d(2w_i - w_i)$ is not necessary for our result and can be, for example, replaced by $\mathcal{B}_d w_i$. However, we use it here because algorithms using the form $\mathcal{B}_d(2w_i - w_i)$ instead of $\mathcal{B}_d w_i$ have stronger convergence guarantees under nonstrongly-convex settings – see [3].

The term $-\mathcal{L}^2 y_{i-1}$ allows us to establish linear convergence by only requiring B to have full row rank in Theorem 2 instead of requiring each B_k to have full row rank. Since $\mathcal{L}y = 0$, having this term is equivalent to adding $-1/2y\mathcal{L}^2 y$ to the saddle-point function in (12), which makes it an augmented Lagrangian formulation.

IV. Linear Convergence Results

In this section, we list our main linear convergence results. We begin with some auxiliary results.

A. Auxiliary Results

The next result shows the existence and optimality of the fixed points of (16).

Lemma 2 (Fixed Point of DCPA): A fixed point (w^o, x^o, y^o, v^o) of recursion (16) exists, i.e.,

$$0 \in \nabla \mathcal{J}(w^o) + \mathcal{B}_d^{\mathsf{T}} y^o + \partial \mathcal{R}(w^o), \tag{19a}$$

$$v^o = y^o + \mu_y \mathcal{B}_d w^o + \mathcal{L}x^o, \tag{19b}$$

$$0 = \mathcal{L}v^o, \tag{19c}$$

$$y^o = \operatorname{prox}_{\mu_y \mathcal{H}^*}(v^o), \tag{19d}$$

and $\mathcal{L}^2 y^o = 0$. Moreover, for any fixed-point (w^o, x^o, y^o, v^o), it holds that $y^o = \mathbb{1}_K \otimes y^o$ and (w^o, y^o) is an optimal point for problem (4). Consequently, w^o is an optimal solution of (3).

Proof: See Appendix B. ∎

Note that if $x_{-1}=0$, then from (16c), we have $x_1=-\mathcal{L}v_1$, which is in the range space of $\mathcal{L}$. As a consequence, the iterates $\{x_i\}_{i\geq 0}$ stay in the range space of $\mathcal{L}$. It can be shown that for a given point (w^o,y^o,v^o), there exists a unique x^o, denoted by $\bar{x}^o$, in the range space of $\mathcal{L}$ [49], [50]. To analyze algorithm (16), we consider the error quantities:

$$\widetilde{w}_i=w_i-w^o,\quad \widetilde{y}_i=y_i-y^o, \tag{20a}$$

$$\widetilde{v}_i=v_i-v^o,\quad \widetilde{x}_i=x_i-\bar{x}^o. \tag{20b}$$

From equations (16) and (19), and the definition of the proximal mapping, the error quantities evolve as:

$$\widetilde{w}_i=\widetilde{w}_{i-1}-\mu_w\left(\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^o)\right)-\mu_w\mathcal{B}_d^{\mathsf{T}}\widetilde{y}_{i-1} -\mu_w\left(\widehat{\partial\mathcal{R}}(w_i)-\widehat{\partial\mathcal{R}}(w^o)\right), \tag{21a}$$

$$\widetilde{v}_i=\widetilde{y}_{i-1}-\mathcal{L}^2\widetilde{y}_{i-1}+\mu_y\mathcal{B}_d(2\widetilde{w}_i-\widetilde{w}_{i-1})+\mathcal{L}\widetilde{x}_{i-1}, \tag{21b}$$

$$\widetilde{x}_i=\widetilde{x}_{i-1}-\mu_x\mathcal{L}\widetilde{v}_i, \tag{21c}$$

$$\widetilde{y}_i=\mathrm{prox}_{\mu_y\mathcal{H}^*}(v_i)-\mathrm{prox}_{\mu_y\mathcal{H}^*}(v^o), \tag{21d}$$

where $\widehat{\partial\mathcal{R}}(w)\in\partial\mathcal{R}(w)$. The following result will be useful in our analysis.

Lemma 3 (Inequality bound): Assume that the step-sizes μ_w, μ_y, and μ_x are strictly positive. Then, the iterates of the error recursion (21) satisfy:

$$\begin{aligned}c_y\|\widetilde{v}_i\|^2_{I-\mu_x\mathcal{L}^2}+\tfrac{c_y}{\mu_x}\|\widetilde{x}_i\|^2&\leq\left(1-\mu_x\underline{\sigma}^2(\mathcal{L})\right)\tfrac{c_y}{\mu_x}\|\widetilde{x}_{i-1}\|^2\\&\quad+c_y\|\widetilde{y}_{i-1}-\mathcal{L}^2\widetilde{y}_{i-1}+\mu_y\mathcal{B}_d\widetilde{w}_i\|^2\\&\quad+2\mu_w\mu_y\|\mathcal{B}_d(\widetilde{w}_i-\widetilde{w}_{i-1})\|^2+\mu_w\mu_y\|\mathcal{B}_d\widetilde{w}_i\|^2\\&\quad-\mu_w\mu_y\|\mathcal{B}_d\widetilde{w}_{i-1}\|^2+2\mu_w(\widetilde{w}_i-\widetilde{w}_{i-1})^{\mathsf{T}}\\&\quad\times\mathcal{B}_d^{\mathsf{T}}(\widetilde{y}_{i-1}-\mathcal{L}^2\widetilde{y}_{i-1}),\end{aligned} \tag{22}$$

where $c_y=\frac{\mu_w}{\mu_y}$.

Proof: See Appendix C. ■

B. Linear Convergence of DCPA

In this section, we establish the linear convergence of (16) under two different conditions listed in Assumption 3.

Assumption 3: The function $\mathcal{J}(w)$ is ν-strongly-convex and either one of the following two conditions is satisfied.

I: The function h is $\frac{1}{\nu_h}$-smooth.

II: $B=[B_1\ \cdots\ B_K]$ has full row rank and $\mathcal{R}=0$.

Remark 4: Assumption 3 is typically required for linearly convergent algorithms in solving the centralized saddle-point (4). Assumption 3-I implies that the conjugate function h^* is strongly-convex, and Assumption 3-II implies that $\mathcal{J}^*(-B^{\mathsf{T}}y)$ is strongly-convex. In other words, Assumption 3 implies that the centralized dual formulation (6) is strongly-concave.

We establish the linear convergence of (16) under Assumption 3-I and 3-II separately.

Theorem 1 (Linear Convergence I): Let Assumptions 1, 2, and 3-I hold. If the step-sizes μ_w,μ_y, and μ_x satisfy

$$\mu_w\leq\frac{1}{2\delta-\nu+(2\mu_y+\frac{1}{\nu_h})\sigma^2_{\max}(\mathcal{B}_d)}, \tag{23a}$$

$$\mu_y<\frac{\nu}{3\sigma^2_{\max}(\mathcal{B}_d)}, \tag{23b}$$

$$\mu_x<\frac{\mu_y\nu_h}{(1+\mu_y\nu_h)\sigma^2_{\max}(\mathcal{L})}, \tag{23c}$$

then it holds that $V_i\leq\gamma V_{i-1},\ i\geq 0$, where

$$V_i=\|\widetilde{w}_i\|^2+c_y(1+\mu_y\nu_h)^2\beta\|\widetilde{y}_i\|^2+\tfrac{c_y}{\mu_x}\|\widetilde{x}_i\|^2 \tag{24a}$$

$$\gamma=\max\left\{1-\mu_w\nu,\tfrac{1}{(1+\mu_y\nu_h)\beta},1-\mu_x\underline{\sigma}^2(\mathcal{L})\right\}<1, \tag{24b}$$

and $\beta=1-\mu_x\sigma^2_{\max}(\mathcal{L})$.

Proof: See Appendix D. ■

The above result shows that when h is smooth, DCPA converges linearly. The assumptions used to establish this result matches the ones in the centralized case [12]–[14]. The next result establishes the linear convergence for the case where h is non-smooth.

Theorem 2 (Linear Convergence II): Let Assumptions 1, 2, and 3-II hold. If the step-sizes μ_w,μ_y, and μ_x satisfy

$$\mu_w<\tfrac{1}{3\delta},\quad \mu_y\leq\tfrac{\nu(1-\sigma^2_{\max}(\mathcal{L}))}{2\sigma^2_{\max}(\mathcal{B}_d)}, \tag{25a}$$

$$\mu_x<\frac{\min\{1,\lambda_{\min}(\mu_w\mu_y\mathcal{B}_d\mathcal{B}_d^{\mathsf{T}}+\frac{1}{2}\mathcal{L}^2)\}}{\sigma^2_{\max}(\mathcal{L})}, \tag{25b}$$

then it holds that $V_i\leq\gamma V_{i-1},\ i\geq 0$, where

$$V_i=\|\widetilde{w}_i\|^2_{C_w}+c_y\beta\|\widetilde{y}_i\|^2+\tfrac{c_y}{\mu_x}\|\widetilde{x}_i\|^2 \tag{26a}$$

$$\gamma=\max\left\{1-\mu_w\nu(1-3\mu_w\delta),\tfrac{1-\lambda_{\min}(\mu_w\mu_y\mathcal{B}_d\mathcal{B}_d^{\mathsf{T}}+\frac{1}{2}\mathcal{L}^2)}{1-\mu_x\sigma^2_{\max}(\mathcal{L})},\ 1-\mu_x\underline{\sigma}^2(\mathcal{L})\right\}<1, \tag{26b}$$

with $C_w=I-\frac{2\mu_y\mu_w}{1-\sigma^2_{\max}(\mathcal{L})}\mathcal{B}_d^{\mathsf{T}}\mathcal{B}_d$ and $\beta=1-\mu_x\sigma^2_{\max}(\mathcal{L})$.

Proof: See Appendix D. ■

Theorem 2 shows that when h is nonsmooth, DCPA converges linearly to the exact solution if each $R_k=0$. To the best of our knowledge, this is the first result that establishes the linear convergence of a decentralized algorithm for non-smooth h and under the same conditions used to establish the linear convergence of centralized algorithms [9]–[11]. As shown in Remark 2, while the methods from [9]–[14] can be used to solve (12), their linear convergence requires stronger assumptions that are not satisfied here.

V. Numerical Experiments

In this section, we apply the DCPA algorithm to two numerical problems: the elastic net problem and a ridge regression problem. All experiments are performed using MATLAB R2019b on a laptop with Intel(R) Core(TM) i7-8750H CPU @ 2.20 GHz.

A. Elastic Net

We first consider the elastic net problem [51]:

$$\min_{w_1,\ldots,w_K}\quad f(w), \tag{27}$$

where

$$f(w)\triangleq\sum_{k=1}^K\left(\frac{1}{2}\|w_k\|^2+\beta_k\|w_k\|_1\right)+\frac{\gamma}{2}\left\|\sum_{k=1}^K B_kw_k-b\right\|^2.$$

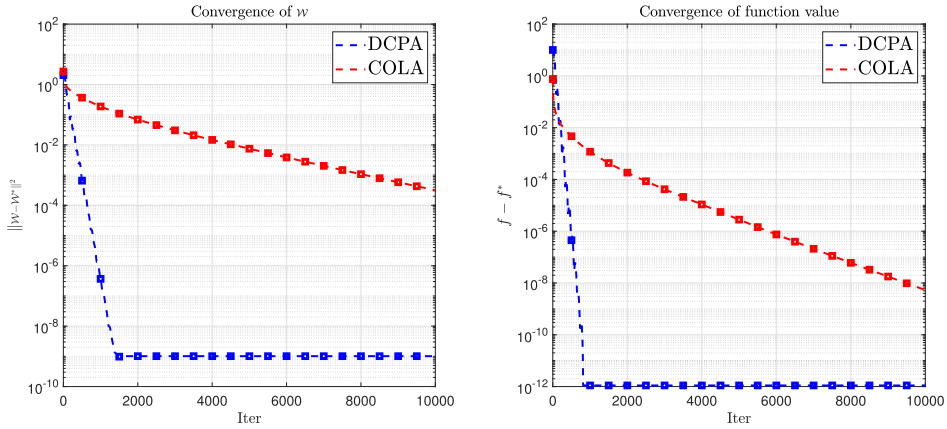

Fig. 1. Comparison of COLA [35] and DCPA. Left: the distance of w_i to the optimal solution $w^\star$, which is defined by $\|w_i - w^\star\|^2$. Right: the objective function value at each iteration minus the optimal function value. Both algorithms require one communication step per iteration.

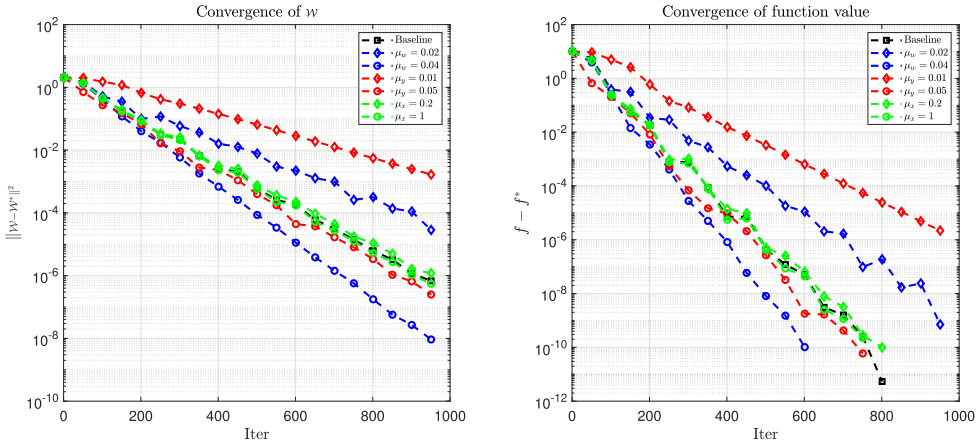

Fig. 2. The convergence rate plot of w and function value, which is defined by $\|w_i - w^\star\|^2$ and $f(w_i) - f^\star$. The black one is the baseline: $\mu_w = 0.03$, $\mu_y = 0.03$, and $\mu_x = 0.5$. The blue lines only change μ_w from the baseline values, while the red lines change μ_y and the green lines change μ_x.

The above problem satisfies Assumption 3-I because $h(\cdot) \triangleq \frac{\gamma}{2}\|\cdot - b\|^2$ is γ-smooth. We choose $K = 10$, $w_k \in \mathbb{R}^{20}$ for all k, $B_k \in \{0,1\}^{20\times 20}$ with entries randomly drafted from {0,1} for all k, and $b \in \mathbb{R}^{20}$, whose entries are chosen from the standard normal distribution independently. Since there are K agents, there can be at most $\frac{K(K-1)}{2}$ edges. We define the connectivity ratio as the actual edges divided by $\frac{K(K-1)}{2}$. The network graph is generated using the same way as [49] with connectivity ratio 0.4. We set the parameters $\gamma = 0.1$ and $\beta_k = 1$ for all k. The optimal function value $f^\star = f(w^\star)$ is estimated by the CVX toolbox [52].

1) Comparison With COLA: We compare the performance of our proposed algorithm and COLA [35] on problem (27). The parameter setting of DCPA is $\mu_w = 0.1$, $\mu_y = 0.003$ and $\mu_x = 0.3$. The COLA parameters are chosen as recommended in [35]. The results are shown in Fig. 1. Both algorithms have linear convergence, and DCPA has a faster convergence rate than COLA in terms of number of iterations (or communication rounds). One reason for this superiority is that COLA requires solving inner optimization sub-problems at each iteration that do not have analytical solutions and can only be approximated. In our simulations, we used FISTA [53] to approximate the solutions of these sub problems. Hence, the computational time for each iteration of COLA may be larger than that of DCPA, because several FISTA iterations are required at each COLA iteration.

2) Step-Sizes μ_w, μ_x, and μ_y: We simulate DCPA for different step-size parameters. We define the baseline parameters as $\mu_w = 0.03$, $\mu_y = 0.03$, and $\mu_x = 0.5$, since it gives a reasonable performance. We fix two step-size parameters and change the third one. The result is shown in Fig. 2. We see that increasing μ_w or μ_y from the baseline value results in faster convergence speed.

To see how the step-sizes affects the convergence of the algorithm, we numerically test different step-size parameters. For any fixed μ_w and μ_y, there exist an upper bound for the parameter μ_x to make the algorithm converge. The contour lines (or level sets) of this upper bound is shown in Fig. 3. The figure shows that μ_w and μ_y have an inverse relation; moreover, increasing μ_x decreases the range of step-sizes μ_w and μ_y.

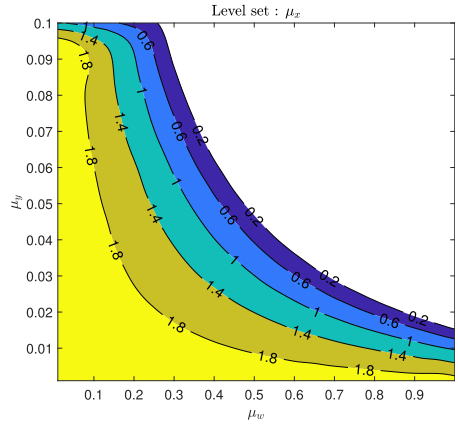

Fig. 3. The contour of the upper bound of μ_x to make DCPA converge.

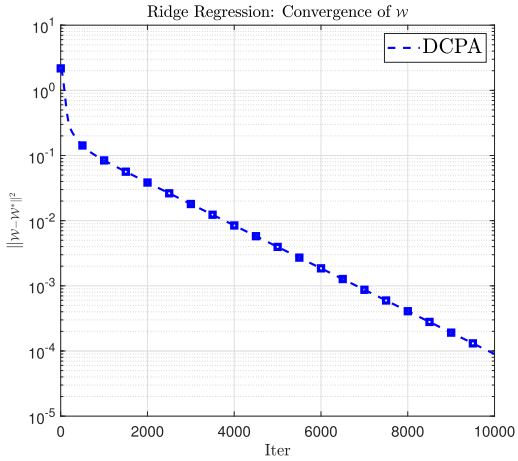

Fig. 4. The convergence of w_i to the optimal solution $w^\star$, $\|w_i - w^\star\|^2$, for ridge regression problem.

B. Ridge Regression

In this section, we test our algorithm for the case where h is non-smooth. We consider the ridge regression problem

$$\min_{w_1,\ldots,w_K} \quad f(w) \triangleq \frac{1}{2}\sum_{k=1}^{K}\|w_k\|^2,$$
$$\text{subject to} \quad \left\|\sum_{k=1}^{K} B_k w_k - b\right\| \leq \sigma, \tag{28}$$

which can be rewritten as

$$\min_{w_1,\ldots,w_K} \quad \frac{1}{2}\sum_{k=1}^{K}\|w_k\|_2^2 + h\left(\sum_{k=1}^{K} B_k w_k\right), \tag{29}$$

with h being the indicator function that returns zero for all $\|x-b\| \leq \sigma$ and $+\infty$ otherwise. The above problem satisfies Assumption 3-II if $B = [B_1, \ldots, B_K]$ has a full row rank. We let $\sigma = 0.1$ in the numerical experiment. As in the previous experiment, we generate a random graph with $K = 20$ agents and connectivity ratio 0.3. We let $w_k \in \mathbb{R}^{10}$ for all k. The matrix $B_k \in \{0,1\}^{20\times 10}$ and the vector $b \in \mathbb{R}^{20}$ are constructed using the same way as in the previous experiment. The result is shown in Fig. 4. As expected DCPA achieves linear convergence under Assumption 3-II.

VI. Conclusion

We studied the linear convergence of decentralized algorithms for the multi-agent sharing optimization problem (1) with a general coupling function (possibly non-smooth). To solve the problem in a decentralized manner, we reformulated it into the equivalent *decentralized* saddle-point problem (12). We proposed a decentralized algorithm that solves problem (12) (hence, (1)) and established its exact global linear convergence. Our conditions are weaker than the conditions used to establish linear convergence of existing decentralized algorithms and match the standard conditions used to establish linear convergence for centralized implementations. Finally, we provided numerical simulations that illustrate our theory and show the advantages of the proposed method.

Appendix A
Proof of Lemma 1

It holds that $y^\star = \mathbb{1}_K \otimes v^\star$ for some $v^\star$, which follows from equations (8) and (13b). Thus, substituting $-\mathcal{B}_d^\mathsf{T} y^\star = -B^\mathsf{T} v^\star$ into (13a), we have:

$$-B^\mathsf{T} v^\star - \nabla \mathcal{J}(w^\star) \in \partial \mathcal{R}(w^\star). \tag{30}$$

Multiplying (13c) by $\mathbb{1}_K^\mathsf{T} \otimes I_E$ on the left, we get:

$$(\mathbb{1}_K^\mathsf{T} \otimes I_E)\mathcal{B}_d w^\star + (\mathbb{1}_K^\mathsf{T} \otimes I_E)\mathcal{L}x^\star \in (\mathbb{1}_K^\mathsf{T} \otimes I_E)\partial \mathcal{H}^*(y^\star)$$
$$\Longrightarrow Bw^\star \in \partial h^*(v^\star), \tag{31}$$

where we used the fact that $(\mathbb{1}_K^\mathsf{T} \otimes I_E)\mathcal{L} = 0$ and $y^\star = \mathbb{1}_K \otimes v^\star$. Equations (30) and (31) are the same conditions as (5). Thus, the point $(w^\star, y^\star)$ with $y^\star = v^\star$ is optimal. ■

Appendix B
Proof of Lemma 2

Suppose that an optimal point $(w^\star, y^\star)$ of of (4) is given, which satisfies (5). We define $w^o \triangleq w^\star$ and $y^o \triangleq y^\star = \mathbb{1}_K \otimes y^\star$. Then, (19a) is satisfied due to (13a). We define

$$v^o = \mathbb{1}_K \otimes v^o \triangleq \mathbb{1}_K \otimes \left(y^\star + \frac{\mu_y}{K} B w^\star\right), \tag{32}$$

which satisfies condition (19c). Note that equation (19d) is equivalent to $v^o - y^o \in \mu_y \partial \mathcal{G}^*(y^o)$. Hence, from the definition of $\mathcal{G}^*$, (5b) and (32), the condition (19d) is satisfied. It remains to show the existence of x^o such that (19b) holds. Note that

$$(\mathbb{1}_K^\top \otimes I_E)(v^o - y^o - \mu_y \mathcal{B}_d w^o)$$
$$= K(v^o - y^o) - \mu_y B w^o \overset{(32)}{=} 0. \tag{33}$$

This means that $v^o - y^o - \mu_y \mathcal{B}_d w^o$ is in null space of $\mathbb{1}_K^\top \otimes I_E$, consequently, it is in the range space of $\mathcal{L}$. Hence, there exists x^o such that (19b) holds.

Now, suppose that (w^o, x^o, y^o, v^o) is a fixed-point of (16). It follows that $v^o = \mathbb{1}_K \otimes v^o$ due to equation (19c). As a result, (19d) implies that $y^o = \mathbb{1}_K \otimes y^o$. We also have $\mathcal{B}_d w^o + \frac{1}{\mu_y}\mathcal{L}x^o \in \partial \mathcal{G}^*(y^o)$, which holds from (19b) and (19d). Thus, using Lemma 1, the point (w^o, y^o) satisfies the optimality condition (5). ■

APPENDIX C
PROOF OF LEMMA 3

Taking the norm squares of (21b) and (21c), we have:

$$\begin{aligned}\|\widetilde{\nu}_i\|^2 =& \|\widetilde{y}_{i-1}-\mathcal{L}^2\widetilde{y}_{i-1}+\mu_y\mathcal{B}_d(2\widetilde{w}_i-\widetilde{w}_{i-1})\|^2+\|\mathcal{L}\widetilde{x}_{i-1}\|^2\\ &+2\widetilde{x}_{i-1}^{\mathsf{T}}\mathcal{L}(\widetilde{y}_{i-1}-\mathcal{L}^2\widetilde{y}_{i-1}+\mu_y\mathcal{B}_d(2\widetilde{w}_i-\widetilde{w}_{i-1})),\end{aligned} \tag{34}$$

and

$$\begin{aligned}\|\widetilde{x}_i\|^2 =& \|\widetilde{x}_{i-1}\|^2+\mu_x^2\|\mathcal{L}\widetilde{\nu}_i\|^2-2\mu_x\widetilde{x}_{i-1}^{\mathsf{T}}\mathcal{L}\left(\widetilde{y}_{i-1}-\mathcal{L}^2\widetilde{y}_{i-1}\right.\\ &\left.+\mu_y\mathcal{B}_d(2\widetilde{w}_i-\widetilde{w}_{i-1})+\mathcal{L}\widetilde{x}_{i-1}\right).\end{aligned} \tag{35}$$

Dividing (35) by μ_x and adding it to (34) give us

$$\begin{aligned}&\|\widetilde{\nu}_i\|^2_{I-\mu_x\mathcal{L}^2}+\tfrac{1}{\mu_x}\|\widetilde{x}_i\|^2\\ &=\tfrac{1}{\mu_x}\|\widetilde{x}_{i-1}\|^2_{I-\mu_x\mathcal{L}^2}+\|\widetilde{y}_{i-1}-\mathcal{L}^2\widetilde{y}_{i-1}\\ &\quad+\mu_y\mathcal{B}_d(2\widetilde{w}_i-\widetilde{w}_{i-1})\|^2\\ &=\tfrac{1}{\mu_x}\|\widetilde{x}_{i-1}\|^2_{I-\mu_x\mathcal{L}^2}+\|\widetilde{y}_{i-1}-\mathcal{L}^2\widetilde{y}_{i-1}+\mu_y\mathcal{B}_d\widetilde{w}_i\|^2\\ &\quad+\mu_y^2\|\mathcal{B}_d(\widetilde{w}_i-\widetilde{w}_{i-1})\|^2\\ &\quad+2\mu_y(\widetilde{w}_i-\widetilde{w}_{i-1})^{\mathsf{T}}\mathcal{B}_d^{\mathsf{T}}\left(\widetilde{y}_{i-1}-\mathcal{L}^2\widetilde{y}_{i-1}+\mu_y\mathcal{B}_d\widetilde{w}_i\right).\end{aligned} \tag{36}$$

We can rewrite the last term on the right hand side of (36) as

$$\begin{aligned}&2\mu_y(\widetilde{w}_i-\widetilde{w}_{i-1})^{\mathsf{T}}\mathcal{B}_d^{\mathsf{T}}(\widetilde{y}_{i-1}-\mathcal{L}^2\widetilde{y}_{i-1}+\mu_y\mathcal{B}_d\widetilde{w}_i)\\ &=2\mu_y^2(\mathcal{B}_d\widetilde{w}_i-\mathcal{B}_d\widetilde{w}_{i-1})^{\mathsf{T}}\mathcal{B}_d\widetilde{w}_i\\ &\quad+2\mu_y(\widetilde{w}_i-\widetilde{w}_{i-1})^{\mathsf{T}}\mathcal{B}_d^{\mathsf{T}}(\widetilde{y}_{i-1}-\mathcal{L}^2\widetilde{y}_{i-1})\\ &=\mu_y^2\|\mathcal{B}_d(\widetilde{w}_i-\widetilde{w}_{i-1})\|^2+\mu_y^2\|\mathcal{B}_d\widetilde{w}_i\|^2-\mu_y^2\|\mathcal{B}_d\widetilde{w}_{i-1}\|^2\\ &\quad+2\mu_y(\widetilde{w}_i-\widetilde{w}_{i-1})^{\mathsf{T}}\mathcal{B}_d^{\mathsf{T}}(\widetilde{y}_{i-1}-\mathcal{L}^2\widetilde{y}_{i-1}),\end{aligned} \tag{37}$$

where we used $2x^{\mathsf{T}}y=\|x\|^2+\|y\|^2-\|x-y\|^2$ in the last step. Substituting the above equation into (36) and multiplying by $c_y=\frac{\mu_w}{\mu_y}$, we get

$$\begin{aligned}&c_y\|\widetilde{\nu}_i\|^2_{I-\mu_x\mathcal{L}^2}+\tfrac{c_y}{\mu_x}\|\widetilde{x}_i\|^2\\ &=\tfrac{c_y}{\mu_x}\|\widetilde{x}_{i-1}\|^2_{I-\mu_x\mathcal{L}^2}+c_y\|\widetilde{y}_{i-1}-\mathcal{L}^2\widetilde{y}_{i-1}+\mu_y\mathcal{B}_d\widetilde{w}_i\|^2\\ &\quad+2\mu_w\mu_y\|\mathcal{B}_d(\widetilde{w}_i-\widetilde{w}_{i-1})\|^2+\mu_w\mu_y\|\mathcal{B}_d\widetilde{w}_i\|^2\\ &\quad-\mu_w\mu_y\|\mathcal{B}_d\widetilde{w}_{i-1}\|^2\\ &\quad+2\mu_w(\widetilde{w}_i-\widetilde{w}_{i-1})^{\mathsf{T}}\mathcal{B}_d^{\mathsf{T}}(\widetilde{y}_{i-1}-\mathcal{L}^2\widetilde{y}_{i-1}).\end{aligned} \tag{38}$$

Now, since $x_{-1}=0$ and $\bar{x}^o$ belong to the range space of $\mathcal{L}$, the error quantity $\widetilde{x}_{i-1}$ always belongs to the range space $\mathcal{L}$. This implies that [50, Lemma 1]:

$$\|\widetilde{x}_{i-1}\|^2_{\mathcal{L}^2}\geq\underline{\sigma}^2(\mathcal{L})\|\widetilde{x}_{i-1}\|^2.$$

Therefore,

$$\|\widetilde{x}_{i-1}\|^2_{I-\mu_x\mathcal{L}^2}\leq\left(1-\mu_x\underline{\sigma}^2(\mathcal{L})\right)\|\widetilde{x}_{i-1}\|^2. \tag{39}$$

Using this bound in (38) we get our result. ■

APPENDIX D
THEOREM 1 AND 2 PROOFS

Lemma 4: Under Assumption 1, the iterates of the error recursion in (21) satisfy:

$$\begin{aligned}&\|\widetilde{w}_i\|^2-\mu_w\mu_y\|\mathcal{B}_d\widetilde{w}_i\|^2+\underbrace{c_y\|\widetilde{\nu}_i\|^2_{I-\mu_x\mathcal{L}^2}}_{(A)}+\tfrac{c_y}{\mu_x}\|\widetilde{x}_i\|^2\\ &\leq\|\widetilde{w}_{i-1}\|^2-\mu_w\mu_y\|\mathcal{B}_d\widetilde{w}_{i-1}\|^2+c_y\|\widetilde{y}_{i-1}\|^2-2c_y\|\widetilde{y}_{i-1}\|^2_{\mathcal{L}^2}\\ &\quad+\left(1-\mu_x\underline{\sigma}^2(\mathcal{L})\right)\tfrac{c_y}{\mu_x}\|\widetilde{x}_{i-1}\|^2-\|\widetilde{w}_i-\widetilde{w}_{i-1}\|^2\\ &\quad+2\mu_w\mu_y\|\mathcal{B}_d(\widetilde{w}_i-\widetilde{w}_{i-1})\|^2\\ &\quad\underbrace{-2\mu_w(\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^o))^{\top}\widetilde{w}_i}_{(B)}\\ &\quad+\underbrace{c_y\|\mu_y\mathcal{B}_d\widetilde{w}_i-\mathcal{L}^2\widetilde{y}_{i-1}\|^2}_{(C)}\\ &\quad+\underbrace{2\mu_w(\widetilde{w}_i-\widetilde{w}_{i-1})^{\mathsf{T}}\mathcal{B}_d^{\mathsf{T}}(\widetilde{y}_{i-1}-\mathcal{L}^2\widetilde{y}_{i-1})}_{(D)}.\end{aligned} \tag{40}$$

Proof: Since the subgradient of a convex function is monotone, it holds that

$$\begin{aligned}0\leq&\,2\mu_w\left(\widehat{\partial\mathcal{R}}(w_i)-\widehat{\partial\mathcal{R}}(w^o)\right)^{\top}\widetilde{w}_i\\ \overset{(21a)}{=}&\,2\Big(\widetilde{w}_{i-1}-\widetilde{w}_i-\mu_w(\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^o))\\ &-\mu_w\mathcal{B}_d^{\top}\widetilde{y}_{i-1}\Big)^{\top}\widetilde{w}_i\\ =&\,2(\widetilde{w}_{i-1}-\widetilde{w}_i)^{\top}\widetilde{w}_i-2\mu_w\left(\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^o)\right)^{\top}\widetilde{w}_i\\ &-2\mu_w\widetilde{y}_{i-1}^{\top}\mathcal{B}_d\widetilde{w}_i.\end{aligned} \tag{41}$$

Using

$$2(\widetilde{w}_{i-1}-\widetilde{w}_i)^{\top}\widetilde{w}_i=\|\widetilde{w}_{i-1}\|^2-\|\widetilde{w}_i\|^2-\|\widetilde{w}_i-\widetilde{w}_{i-1}\|^2$$

and

$$\begin{aligned}&-2\mu_w\widetilde{y}_{i-1}^{\top}\mathcal{B}_d\widetilde{w}_i\\ &=c_y\|\widetilde{y}_{i-1}\|^2+c_y\|\mu_y\mathcal{B}_d\widetilde{w}_i-\mathcal{L}^2\widetilde{y}_{i-1}\|^2\\ &\quad-2c_y\|\widetilde{y}_{i-1}\|^2_{\mathcal{L}^2}-c_y\|\widetilde{y}_{i-1}+\mu_y\mathcal{B}_d\widetilde{w}_i-\mathcal{L}^2\widetilde{y}_{i-1}\|^2,\end{aligned}$$

with $c_y=\mu_w/\mu_y$, and rearranging (41), we get:

$$\begin{aligned}\|\widetilde{w}_i\|^2\leq&\,\|\widetilde{w}_{i-1}\|^2-\|\widetilde{w}_i-\widetilde{w}_{i-1}\|^2\\ &-2\mu_w\left(\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^o)\right)^{\top}\widetilde{w}_i\\ &+c_y\|\widetilde{y}_{i-1}\|^2+c_y\|\mu_y\mathcal{B}_d\widetilde{w}_i-\mathcal{L}^2\widetilde{y}_{i-1}\|^2\\ &-2c_y\|\widetilde{y}_{i-1}\|^2_{\mathcal{L}^2}-c_y\|\widetilde{y}_{i-1}-\mathcal{L}^2\widetilde{y}_{i-1}+\mu_y\mathcal{B}_d\widetilde{w}_i\|^2.\end{aligned}$$

Adding the above to (22) in Lemma 3 and rearranging the terms yield our result. ■

To prove Theorems 1 and 2, we upper bound the four terms $((A),(B),(C),(D))$ for each case.

A. Proof of Theorem 1

We consider term (A) first. From equations (16d) and (19d), it holds that:

$$\widetilde{\nu}_i = \widetilde{\nu}_i + \mu_y \left(\widehat{\partial \mathcal{H}^*}(\nu_i) - \widehat{\partial \mathcal{H}^*}(\nu^\star) \right). \tag{42}$$

Left-multiplying both sides of the previous equation by $\widetilde{\nu}_i^{\mathsf{T}}$, we obtain:

$$\begin{aligned}\widetilde{\nu}_i^{\mathsf{T}}\widetilde{\nu}_i &= \|\widetilde{\nu}_i\|^2 + \mu_y \widetilde{\nu}_i^{\mathsf{T}} \left(\widehat{\partial \mathcal{H}^*}(\nu_i) - \widehat{\partial \mathcal{H}^*}(\nu^\star) \right) \\ &\geq (1+\mu_y\nu_h)\|\widetilde{\nu}_i\|^2, \end{aligned} \tag{43}$$

where the last inequality comes from the ν_h-strong convexity of $\mathcal{H}^*$. Hence, by using the Cauchy-Schwarz inequality, the bound

$$(1+\mu_y\nu_h)\|\widetilde{\nu}_i\| \leq \|\widetilde{\nu}_i\|$$

holds. We can thus bound

$$\begin{aligned}(A) = c_y\|\widetilde{\nu}_i\|^2_{I-\mu_x\mathcal{L}^2} &\geq c_y\left(1-\mu_x\sigma^2_{\max}(\mathcal{L})\right)\|\widetilde{\nu}_i\|^2 \\ &\geq c_y(1+\mu_y\nu_h)^2\beta\|\widetilde{\nu}_i\|^2. \end{aligned} \tag{44}$$

where

$$\beta \triangleq 1-\mu_x\sigma^2_{\max}(\mathcal{L}).$$

Since $\mathcal{J}$ is ν-strongly convex and δ-smooth, we have

$$\begin{aligned}(B) &= 2\mu_w\nabla\mathcal{J}(w^o)^\top(w_i-w^o) \\ &\quad + 2\mu_w\nabla\mathcal{J}(w_{i-1})^\top(w^o-w_{i-1}) \\ &\quad + 2\mu_w\nabla\mathcal{J}(w_i)^\top(w_{i-1}-w_i) \\ &\quad + 2\mu_w\left(\nabla\mathcal{J}(w_i)-\nabla\mathcal{J}(w_{i-1})\right)^\top(w_i-w_{i-1}) \\ &\leq 2\mu_w\mathcal{J}(w_i) - 2\mu_w\mathcal{J}(w^o) - \mu_w\nu\|\widetilde{w}_i\|^2 \\ &\quad + 2\mu_w\mathcal{J}(w^o) - 2\mu_w\mathcal{J}(w_{i-1}) - \mu_w\nu\|\widetilde{w}_{i-1}\|^2 \\ &\quad + 2\mu_w\mathcal{J}(w_{i-1}) - 2\mu_w\mathcal{J}(w_i) - \mu_w\nu\|\widetilde{w}_i-\widetilde{w}_{i-1}\|^2 \\ &\quad + 2\mu_w\delta\|\widetilde{w}_i-\widetilde{w}_{i-1}\|^2 \\ &= -\mu_w\nu\|\widetilde{w}_i\|^2 - \mu_w\nu\|\widetilde{w}_{i-1}\|^2 \\ &\quad - (\mu_w\nu+2\mu_w\delta)\|\widetilde{w}_i-\widetilde{w}_{i-1}\|^2. \end{aligned} \tag{45}$$

For term (C), we have

$$\begin{aligned}(C) &\leq 2c_y\mu_y^2\|\mathcal{B}_d\widetilde{w}_i\|^2 + 2c_y\|\widetilde{\nu}_{i-1}\|^2_{\mathcal{L}^4} \\ &\leq 2\mu_w\mu_y\|\mathcal{B}_d\widetilde{w}_i\|^2 + 2c_y\|\widetilde{\nu}_{i-1}\|^2_{\mathcal{L}^2}, \end{aligned} \tag{46}$$

where we used $\|x+y\|^2 \leq 2\|x\|^2+2\|y\|^2$ and $0 \leq \mathcal{L}^2 < I$. Using $2x^{\mathsf{T}}y \leq \frac{1}{\epsilon}\|x\|^2 + \epsilon\|y\|^2$ for any $\epsilon > 0$, the fourth term (D) can bounded by

$$(D) \leq \frac{\mu_w}{\nu_h}\|\mathcal{B}_d(\widetilde{w}_i-\widetilde{w}_{i-1})\|^2 + \mu_w\nu_h\|(I-\mathcal{L}^2)\widetilde{\nu}_{i-1}\|^2. \tag{47}$$

Substituting all four bounds on $((A),(B),(C),(D))$ into (40), we obtain

$$\begin{aligned}&(1+\mu_w\nu)\|\widetilde{w}_i\|^2 - 3\mu_w\mu_y\|\mathcal{B}_d\widetilde{w}_i\|^2 \\ &\quad + c_y(1+\mu_y\nu_h)^2\beta\|\widetilde{\nu}_i\|^2 + \tfrac{c_y}{\mu_x}\|\widetilde{x}_i\|^2 \\ &\leq (1-\mu_w\nu)\|\widetilde{w}_{i-1}\|^2 - \mu_w\mu_y\|\mathcal{B}_d\widetilde{w}_{i-1}\|^2 + c_y\|\widetilde{\nu}_{i-1}\|^2 \\ &\quad + \mu_w\nu_h\|(I-\mathcal{L}^2)\widetilde{\nu}_{i-1}\|^2 + \left(1-\mu_x\underline{\sigma}^2(\mathcal{L})\right)\tfrac{c_y}{\mu_x}\|\widetilde{x}_{i-1}\|^2 \\ &\quad + \left(2\mu_w\mu_y + \tfrac{\mu_w}{\nu_h}\right)\|\mathcal{B}_d(\widetilde{w}_i-\widetilde{w}_{i-1})\|^2 \\ &\quad - (1+\mu_w\nu-2\mu_w\delta)\|\widetilde{w}_i-\widetilde{w}_{i-1}\|^2. \end{aligned} \tag{48}$$

We can ensure that

$$\begin{aligned}&\left(2\mu_w\mu_y + \tfrac{\mu_w}{\nu_h}\right)\|\mathcal{B}_d(\widetilde{w}_i-\widetilde{w}_{i-1})\|^2 \\ &\quad - (1+\mu_w\nu-2\mu_w\delta)\|\widetilde{w}_i-\widetilde{w}_{i-1}\|^2 \leq 0, \end{aligned} \tag{49}$$

under the condition on μ_w given in (23a). We can also ensure that

$$\|\widetilde{w}_i\|^2 \leq (1+\mu_w\nu)\|\widetilde{w}_i\|^2 - 3\mu_w\mu_y\|\mathcal{B}_d\widetilde{w}_i\|^2 \tag{50}$$

if the conditions on μ_y in (23b) holds. Note that

$$\begin{aligned}&(1-\mu_w\nu)\|\widetilde{w}_{i-1}\|^2 - \mu_w\mu_y\|\mathcal{B}_d\widetilde{w}_{i-1}\|^2 \\ &\quad \leq (1-\mu_w\nu)\|\widetilde{w}_{i-1}\|^2. \end{aligned} \tag{51}$$

Using the condition $0 \leq \mathcal{L}^2 < I$, it holds that

$$\begin{aligned}&c_y\|\widetilde{\nu}_{i-1}\|^2 + \mu_w\nu_h\|(I-\mathcal{L}^2)\widetilde{\nu}_{i-1}\|^2 \\ &\quad \leq (c_y+\mu_w\nu_h)\|\widetilde{\nu}_{i-1}\|^2 \\ &\quad = \gamma_2 c_y(1+\mu_y\nu_h)^2\left(1-\mu_x\sigma^2_{\max}(\mathcal{L})\right)\|\widetilde{\nu}_{i-1}\|^2, \end{aligned} \tag{52}$$

where

$$\gamma_2 = \frac{1}{(1+\mu_y\nu_h)(1-\mu_x\sigma^2_{\max}(\mathcal{L}))} < 1, \tag{53}$$

under the condition (23c). Therefore, substituting the previous bounds into (48), it holds that

$$\begin{aligned}&\|\widetilde{w}_i\|^2 + c_y(1+\mu_y\nu_h)^2\beta\|\widetilde{\nu}_i\|^2 + \tfrac{c_y}{\mu_x}\|\widetilde{x}_i\|^2 \\ &\quad \leq \gamma\left(\|\widetilde{w}_{i-1}\|^2 + c_y(1+\mu_y\nu_h)^2\beta\|\widetilde{\nu}_{i-1}\|^2 + \tfrac{c_y}{\mu_x}\|\widetilde{x}_{i-1}\|^2\right) \end{aligned} \tag{54}$$

with

$$\gamma = \max\left\{1-\mu_w\nu, \tfrac{1+\mu_y\nu_h}{(1+\mu_y\nu_h)^2\beta}, 1-\mu_x\underline{\sigma}^2(\mathcal{L})\right\} < 1, \tag{55}$$

under the conditions (23). The theorem is proved. ■

B. Proof of Theorem 2

Note that

$$\beta \triangleq 1-\mu_x\sigma^2_{\max}(\mathcal{L}) > 0$$

for $\mu_x < \frac{1}{\sigma^2_{\max}(\mathcal{L})}$. We can lower bound (A) by

$$\begin{aligned}\beta c_y\|\widetilde{\nu}_i\|^2 &= \beta c_y\|\mathrm{prox}_{\mu_y\mathcal{H}^*}(\nu_i) - \mathrm{prox}_{\mu_y\mathcal{H}^*}(\nu^o)\|^2 \\ &\leq \beta c_y\|\widetilde{\nu}_i\|^2 \leq (A). \end{aligned} \tag{56}$$

We have the following upper bound for the term (B):

$$\begin{aligned}(B) &= -2\mu_w(\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^o))^\top\widetilde{w}_{i-1} \\ &\quad - 2\mu_w(\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^o))^\top(w_i-w_{i-1}) \\ &= \|\widetilde{w}_{i-1} - \mu_w(\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^o))\|^2 - \|\widetilde{w}_{i-1}\|^2 \\ &\quad - \|\mu_w(\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^o))\|^2 \\ &\quad - 2\mu_w\left(\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^o)\right)^{\mathsf{T}}(w_i-w_{i-1}) \\ &= \|\widetilde{w}_{i-1} - \mu_w(\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^o))\|^2 \\ &\quad - \|\widetilde{w}_{i-1}\|^2 + \|w_i-w_{i-1}\|^2 \\ &\quad - \|\mu_w(\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^o)) + w_i - w_{i-1}\|^2 \\ &= \|\widetilde{w}_{i-1} - \mu_w(\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^o))\|^2 - \|\widetilde{w}_{i-1}\|^2 \\ &\quad - \|\mu_w\mathcal{B}_d^\top\widetilde{\nu}_{i-1}\|^2 + \|w_i-w_{i-1}\|^2 \end{aligned}$$

$$\leq -\mu_w(2-\mu_w\delta)\widetilde{w}_{i-1}^\top(\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^o)) - \|\mu_w\mathcal{B}_d^\top\widetilde{y}_{i-1}\|^2+\|w_i-w_{i-1}\|^2, \quad (57)$$

where the last equality holds from (21a) since $\mathcal{R}=0$. The last inequality holds because $\mathcal{J}$ is convex and δ-smooth. Using Jensen's inequality $\|x+y\|^2\leq\frac{1}{1-t}\|x\|^2+\frac{1}{t}\|y\|^2$ for any $t<1$, the term (C) can be upper bounded by:

$$(C)\leq\frac{c_y\mu_y^2}{1-\sigma_{\max}^2(\mathcal{L})}\|\mathcal{B}_d\widetilde{w}_i\|^2+\frac{c_y}{\sigma_{\max}^2(\mathcal{L})}\|\widetilde{y}_{i-1}\|_{\mathcal{L}^4}^2 \leq\frac{\mu_w\mu_y}{1-\sigma_{\max}^2(\mathcal{L})}\|\mathcal{B}_d\widetilde{w}_i\|^2+c_y\|\widetilde{y}_{i-1}\|_{\mathcal{L}^2}^2. \quad (58)$$

Last, we can upper bound the term (D) by:

$$\begin{aligned}(D)=&-\mu_w^2\|\mathcal{B}_d^\mathsf{T}(\widetilde{y}_{i-1}-\mathcal{L}^2\widetilde{y}_{i-1})\|^2-\|\widetilde{w}_i-\widetilde{w}_{i-1}\|^2\\&+\mu_w^2\|\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^\star)+\mathcal{B}_d^\top\mathcal{L}^2\widetilde{y}_{i-1}\|^2\\\leq&-\|\widetilde{w}_i-\widetilde{w}_{i-1}\|^2\\&+2\mu_w^2\delta\widetilde{w}_{i-1}^\top(\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^\star))\\&+2\mu_w^2\|\widetilde{y}_{i-1}\|^2_{\mathcal{L}^2\mathcal{B}_d\mathcal{B}_d^\top\mathcal{L}^2}, \end{aligned}\quad (59)$$

where the equality holds due to $2x^\mathsf{T}y=-\|x\|^2-\|y\|^2+\|x+y\|^2$ and (21a) (with $\mathcal{R}=0$). The last step holds since

$$\begin{aligned}&\mu_w^2\|\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^\star)+\mathcal{B}_d^\mathsf{T}\mathcal{L}^2\widetilde{y}_{i-1}\|^2\\&\leq2\mu_w^2\|\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^\star)\|^2+2\mu_w^2\|\mathcal{B}_d^\mathsf{T}\mathcal{L}^2\widetilde{y}_{i-1}\|^2\\&\leq2\mu_w^2\delta\widetilde{w}_{i-1}^\top(\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^\star))\\&\quad+2\mu_w^2\|\widetilde{y}_{i-1}\|^2_{\mathcal{L}^2\mathcal{B}_d\mathcal{B}_d^\top\mathcal{L}^2},\end{aligned}\quad (60)$$

and

$$-\mu_w^2\|\mathcal{B}_d^\mathsf{T}\widetilde{y}_{i-1}-\mathcal{B}_d^\mathsf{T}\mathcal{L}^2\widetilde{y}_{i-1}\|^2\leq0.$$

Note that

$$\mu_w\mu_y\|\mathcal{B}_d\widetilde{w}_i\|^2\leq\tfrac{\mu_w\mu_y}{1-\sigma_{\max}^2(\mathcal{L})}\|\mathcal{B}_d\widetilde{w}_i\|^2$$

since $0\leq\sigma_{\max}^2(\mathcal{L})<1$. Using this, and substituting all four bounds on $((A),(B),(C),(D))$ into (40), we get

$$\begin{aligned}&\|\widetilde{w}_i\|_{C_w}^2+c_y\beta\|\widetilde{y}_i\|^2+\frac{c_y}{\mu_x}\|\widetilde{x}_i\|^2\\&\leq\|\widetilde{w}_{i-1}\|^2-\mu_w\mu_y\|\mathcal{B}_d\widetilde{w}_{i-1}\|^2+c_y\|\widetilde{y}_{i-1}\|^2-c_y\|\widetilde{y}_{i-1}\|^2_{\mathcal{L}^2}\\&\quad+\left(1-\mu_x\underline{\sigma}^2(\mathcal{L})\right)\tfrac{c_y}{\mu_x}\|\widetilde{x}_{i-1}\|^2\\&\quad-\|\widetilde{w}_i-\widetilde{w}_{i-1}\|^2+2\mu_w\mu_y\|\mathcal{B}_d(\widetilde{w}_i-\widetilde{w}_{i-1})\|^2\\&\quad-\mu_w(2-3\mu_w\delta)\widetilde{w}_{i-1}^\top(\nabla\mathcal{J}(w_{i-1})-\nabla\mathcal{J}(w^o))\\&\quad-\|\mu_w\mathcal{B}_d^\top\widetilde{y}_{i-1}\|^2+2\mu_w^2\|\widetilde{y}_{i-1}\|^2_{\mathcal{L}^2\mathcal{B}_d\mathcal{B}_d^\top\mathcal{L}^2}\\&\leq(1-\mu_w\nu(2-3\mu_w\delta))\|\widetilde{w}_{i-1}\|^2-\mu_w\mu_y\|\mathcal{B}_d\widetilde{w}_{i-1}\|^2\\&\quad+c_y\|\widetilde{y}_{i-1}\|^2-c_y\|\widetilde{y}_{i-1}\|^2_{\mathcal{L}^2}-\|\mu_w\mathcal{B}_d^\top\widetilde{y}_{i-1}\|^2\\&\quad+2\mu_w^2\|\widetilde{y}_{i-1}\|^2_{\mathcal{L}^2\mathcal{B}_d\mathcal{B}_d^\top\mathcal{L}^2}+\left(1-\mu_x\underline{\sigma}^2(\mathcal{L})\right)\tfrac{c_y}{\mu_x}\|\widetilde{x}_{i-1}\|^2\\&\quad-\|\widetilde{w}_i-\widetilde{w}_{i-1}\|^2+2\mu_w\mu_y\|\mathcal{B}_d(\widetilde{w}_i-\widetilde{w}_{i-1})\|^2,\end{aligned}\quad (61)$$

where

$$C_w\triangleq I-\tfrac{2\mu_y\mu_w}{1-\sigma_{\max}^2(\mathcal{L})}\mathcal{B}_d^\mathsf{T}\mathcal{B}_d$$

and we used strong-convexity in the last inequality. Note that

$$-\|\widetilde{w}_i-\widetilde{w}_{i-1}\|^2+2\mu_w\mu_y\|\mathcal{B}_d(\widetilde{w}_i-\widetilde{w}_{i-1})\|^2\leq0, \quad (62)$$

for $\mu_w\mu_y\leq\frac{1}{2\sigma_{\max}^2(\mathcal{B}_d)}$. Moreover,

$$\begin{aligned}&(1-\mu_w\nu(2-3\mu_w\delta))\|\widetilde{w}_{i-1}\|^2-\mu_w\mu_y\|\mathcal{B}_d\widetilde{w}_{i-1}\|^2\\&=\gamma_1\|\widetilde{w}_{i-1}\|^2_{C_w}-\mu_w\nu\|\widetilde{w}_{i-1}\|^2\\&\quad+\gamma_1\|\widetilde{w}_{i-1}\|^2_{\frac{2\mu_y\mu_w}{1-\sigma_{\max}^2(\mathcal{L})}\mathcal{B}_d^\mathsf{T}\mathcal{B}_d}-\mu_w\mu_y\|\mathcal{B}_d\widetilde{w}_{i-1}\|^2\\&\leq\gamma_1\|\widetilde{w}_{i-1}\|^2_{C_w}-\mu_w\nu\|\widetilde{w}_{i-1}\|^2+\tfrac{2\mu_y\mu_w\sigma_{\max}^2(\mathcal{B}_d)}{1-\sigma_{\max}^2(\mathcal{L})}\|\widetilde{w}_{i-1}\|^2\\&\leq\gamma_1\|\widetilde{w}_{i-1}\|^2_{C_w}-\mu_w\left(\nu-\tfrac{2\mu_y}{1-\sigma_{\max}^2(\mathcal{L})}\sigma_{\max}^2(\mathcal{B}_d)\right)\|\widetilde{w}_{i-1}\|^2\\&\leq\gamma_1\|\widetilde{w}_{i-1}\|^2_{C_w},\end{aligned}\quad (63)$$

where $\gamma_1\triangleq1-\mu_w\nu(1-3\mu_w\delta)<1$ for $\mu_w<\frac{1}{3\delta}$ and the last step holds for

$$\mu_y\leq\frac{\nu(1-\sigma_{\max}^2(\mathcal{L}))}{2\sigma_{\max}^2(\mathcal{B}_d)}.$$

Also note that

$$\begin{aligned}&c_y\|\widetilde{y}_{i-1}\|^2-c_y\|\widetilde{y}_{i-1}\|^2_{\mathcal{L}^2}-\|\mu_w\mathcal{B}_d^\top\widetilde{y}_{i-1}\|^2+2\mu_w^2\|\widetilde{y}_{i-1}\|^2_{\mathcal{L}^2\mathcal{B}_d\mathcal{B}_d^\top\mathcal{L}^2}\\&\leq c_y\|\widetilde{y}_{i-1}\|^2-c_y\|\widetilde{y}_{i-1}\|^2_{\mathcal{L}^2}-\|\mu_w\mathcal{B}_d^\top\widetilde{y}_{i-1}\|^2\\&\quad+2\mu_w^2\sigma_{\max}^2(\mathcal{B}_d^\mathsf{T}\mathcal{L})\|\widetilde{y}_{i-1}\|^2_{\mathcal{L}^2}\\&=c_y\|\widetilde{y}_{i-1}\|^2_{I-\frac{1}{2}\mathcal{L}^2-\mu_w\mu_y\mathcal{B}_d\mathcal{B}_d^\mathsf{T}}\\&\quad-(\tfrac{1}{2}-2\mu_w\mu_y\sigma_{\max}^2(\mathcal{B}_d^\mathsf{T}\mathcal{L}))c_y\|\widetilde{y}_{i-1}\|_{\mathcal{L}^2}\\&\leq c_y\|\widetilde{y}_{i-1}\|^2_{I-\frac{1}{2}\mathcal{L}^2-\mu_w\mu_y\mathcal{B}_d\mathcal{B}_d^\mathsf{T}},\end{aligned}\quad (64)$$

where the last step holds under the condition $\frac{1}{2}-2\mu_w\mu_y\sigma_{\max}^2(\mathcal{B}_d^\mathsf{T}\mathcal{L})>0$. It holds that (see Appendix E)

$$0<\frac{1}{2}\mathcal{L}^2+\mu_w\mu_y\mathcal{B}_d\mathcal{B}_d^\mathsf{T}<I$$

for

$$\mu_w\mu_y<\frac{1-\frac{1}{2}\sigma_{\max}(\mathcal{L}^2)}{\sigma_{\max}^2(\mathcal{B}_d)}.$$

Therefore, we can further bound the last inequality by

$$\begin{aligned}&c_y\|\widetilde{y}_{i-1}\|^2-c_y\|\widetilde{y}_{i-1}\|^2_{\mathcal{L}^2}-\|\mu_w\mathcal{B}_d^\top\widetilde{y}_{i-1}\|^2+2\mu_w^2\|\widetilde{y}_{i-1}\|^2_{\mathcal{L}^2\mathcal{B}_d\mathcal{B}_d^\top\mathcal{L}^2}\\&\leq c_y\|\widetilde{y}_{i-1}\|^2_{I-\frac{1}{2}\mathcal{L}^2-\mu_w\mu_y\mathcal{B}_d\mathcal{B}_d^\mathsf{T}}\\&\leq\left(1-\lambda_{\min}\left(\mu_w\mu_y\mathcal{B}_d\mathcal{B}_d^\mathsf{T}+\tfrac{1}{2}\mathcal{L}^2\right)\right)\|\widetilde{y}_{i-1}\|^2.\end{aligned}\quad (65)$$

Note that

$$\gamma_2\triangleq\frac{1-\lambda_{\min}(\mu_w\mu_y\mathcal{B}_d\mathcal{B}_d^\mathsf{T}+\frac{1}{2}\mathcal{L}^2)}{1-\mu_x\sigma_{\max}^2(\mathcal{L})}<1 \quad (66)$$

for $\mu_x<\frac{\lambda_{\min}(\mu_w\mu_y\mathcal{B}_d\mathcal{B}_d^\mathsf{T}+\frac{1}{2}\mathcal{L}^2)}{\sigma_{\max}^2(\mathcal{L})}$. Combining the previous bounds into (61), and noting that $\beta=1-\mu_x\sigma_{\max}^2(\mathcal{L})$, the following holds

$$\begin{aligned}&\|\widetilde{w}_i\|_{C_w}^2+c_y\beta\|\widetilde{y}_i\|^2+\tfrac{c_y}{\mu_x}\|\widetilde{x}_i\|^2\\&\leq\gamma\left(\|\widetilde{w}_{i-1}\|_{C_w}^2+c_y\beta\|\widetilde{y}_{i-1}\|^2+\tfrac{c_y}{\mu_x}\|\widetilde{x}_{i-1}\|^2\right),\end{aligned}\quad (67)$$

where

$$\gamma = \max\left\{1-\mu_w\nu(1-3\mu_w\delta), \frac{1-\lambda_{\min}(\mu_w\mu_y\mathcal{B}_d\mathcal{B}_d^{\mathsf{T}}+\frac{1}{2}\mathcal{L}^2)}{\beta}, 1-\mu_x\underline{\sigma}^2(\mathcal{L})\right\} < 1. \quad (68)$$

Collecting all step-size conditions, our bound holds for the sufficient conditions given in (25). The linear convergence is proved. ■

APPENDIX E
POSITIVE DEFINITENESS OF $\mu_w\mu_y\mathcal{B}_d\mathcal{B}_d^{\top}+\frac{1}{2}\mathcal{L}^2$

In this section, we will show that if $B=[B_1\ \cdots\ B_K]$ has full row rank, then the matrix

$$\mathcal{G} \triangleq \tfrac{1}{2}\mathcal{L}^2+\mu_w\mu_y\mathcal{B}_d\mathcal{B}_d^{\top}$$

is positive definite. To see this, note that $\frac{1}{2}\mathcal{L}^2$ and $\mu_w\mu_y\mathcal{B}_d\mathcal{B}_d^{\top}$ are positive semi-definite matrices. This means that $y^{\mathsf{T}}\mathcal{G}y\geq 0$ and is equal to zero if and only if $y^{\mathsf{T}}\mathcal{L}^2y=0$ and $\mu_w\mu_y y^{\mathsf{T}}\mathcal{B}_d\mathcal{B}_d^{\top}y=0$. To show that $\mathcal{G}$ is positive definite, it remains to show that $y^{\mathsf{T}}\mathcal{G}y=0$ only for $y=0$. Assume that there exist y such that $y^{\top}\mathcal{G}y=0$. This means that $\mathcal{L}y=0$ and $\mathcal{B}_d^{\top}y=0$. Note that $\mathcal{L}y=0$ implies that y is consensual, $y=\mathbb{1}_K\otimes y$ for some y. Using this, means that $\mathcal{B}_d^{\top}y=B^{\top}y$. Because B has a full row rank, then $B^{\top}$ has a full column rank, and $B^{\top}y=0$ only for $y=0$. Thus, the matrix $\mathcal{G}$ is positive definite and $\lambda_{\min}(\mathcal{G})>0$. To ensure that $\lambda_{\max}(\mu_w\mu_y\mathcal{B}_d\mathcal{B}_d^{\mathsf{T}}+\frac{1}{2}\mathcal{L}^2)<1$, we use Weyl's Theorem to bound the maximum eigenvalue [54]:

$$\lambda_{\max}(\mu_w\mu_y\mathcal{B}_d\mathcal{B}_d^{\mathsf{T}}+\tfrac{1}{2}\mathcal{L}^2) \leq \mu_w\mu_y\lambda_{\max}(\mathcal{B}_d\mathcal{B}_d^{\mathsf{T}})+\tfrac{1}{2}\lambda_{\max}(\mathcal{L}^2), \quad (69)$$

which is less than one for

$$\mu_w\mu_y < \frac{1-\frac{1}{2}\lambda_{\max}(\mathcal{L}^2)}{\lambda_{\max}(\mathcal{B}_d\mathcal{B}_d^{\mathsf{T}})}.$$

Note that $\lambda_{\max}(\mathcal{B}_d\mathcal{B}_d^{\mathsf{T}})=\sigma^2_{\max}(\mathcal{B}_d)$ and $\lambda_{\max}(\mathcal{L}^2)=\sigma^2_{\max}(\mathcal{L})$.

REFERENCES

[1] S. A. Alghunaim, M. Yan, and A. H. Sayed, "A multi-agent primal-dual strategy for composite optimization over distributed features," in *Proc. IEEE 28th Eur. Signal Process. Conf.*, Amsterdam, The Netherlands, 2021, pp. 1–5.

[2] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via alternating direction method of multipliers," *Found. Trends Mach. Lear.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[3] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, 2011.

[4] J. F. Mota, J. M. Xavier, P. M. Aguiar, and M. Puschel, "Distributed basis pursuit," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1942–1956, Apr. 2012.

[5] F. Guo, C. Wen, J. Mao, and Y.-D. Song, "Distributed economic dispatch for smart grids with random wind power," *IEEE Trans. Smart Grid*, vol. 7, no. 3, pp. 1572–1583, May 2016.

[6] R. Halvgaard, L. Vandenberghe, N. K. Poulsen, H. Madsen, and J. B. Jorgensen, "Distributed model predictive control for smart energy systems," *IEEE Trans. Smart Grid*, vol. 7, no. 3, pp. 1675–1682, Apr. 2016.

[7] V. Smith, S. Forte, M. Chenxin, M. Takac, M. I. Jordan, and M. Jaggi, "CoCoA: A general framework for communication-efficient distributed optimization," *J. Mach. Learn. Res.*, vol. 18, 2018, Art. no. 230.

[8] J. Chen, Z. J. Towfic, and A. H. Sayed, "Dictionary learning over distributed models," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 1001–1016, Feb. 2015.

[9] P. Chen, J. Huang, and X. Zhang, "A primal-dual fixed point algorithm for convex separable minimization with applications to image restoration," *Inverse Problems*, vol. 29, no. 2, Jan. 2013, Art. no. 025011.

[10] N. K. Dhingra, S. Z. Khong, and M. R. Jovanovic, "The proximal augmented lagrangian method for nonsmooth composite optimization," *IEEE Trans. Autom. Control*, vol. 64, no. 7, pp. 2861–2868, Jul. 2019.

[11] W. Deng and W. Yin, "On the global and linear convergence of the generalized alternating direction method of multipliers," *J. Sci. Comput.*, vol. 66, no. 3, pp. 889–916, 2016.

[12] A. Chambolle and T. Pock, "On the ergodic convergence rates of a first-order primal-dual algorithm," *Math. Program.*, vol. 159, no. 1-2, pp. 253–287, Sep. 2016.

[13] R. I. Bot, E. R. Csetnek, A. Heinrich, and C. Hendrich, "On the convergence rate improvement of a primal-dual splitting algorithm for solving monotone inclusion problems," *Math. Program.*, vol. 150, no. 2, pp. 251–279, 2015.

[14] M. Yan, "A new primal-dual algorithm for minimizing the sum of three functions with a linear operator," *J. Sci. Comput.*, vol. 76, no. 3, pp. 1698–1717, 2018.

[15] S. S. Du and W. Hu, "Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, Naha, Okinawa, Japan, 2019, pp. 196–205.

[16] L. Walras, *Elements Deconomie Politique Pure, ou, Theorie de la Richesse Sociale*. F. Rouge, 1896.

[17] Y. Ho, L. Servi, and R. Suri, "A class of center-free resource allocation algorithms," *IFAC Proc. Volumes*, vol. 13, no. 6, pp. 475–482, 1980.

[18] H. Lakshmanan and D. P. De Farias, "Decentralized resource allocation in dynamic networks of agents," *SIAM J. Optim.*, vol. 19, no. 2, pp. 911–940, 2008.

[19] I. Necoara, "Random coordinate descent algorithms for multi-agent convex optimization over networks," *IEEE Trans. Autom. Control*, vol. 58, no. 8, pp. 2001–2012, Aug. 2013.

[20] T. T. Doan and A. Olshevsky, "Distributed resource allocation on dynamic networks in quadratic time," *Syst. Control Lett.*, vol. 99, pp. 57–63, 2017.

[21] A. Nedić, A. Olshevsky, and W. Shi, "Improved convergence rates for distributed resource allocation," in *Proc. IEEE Conf. Decis. Control*, Miami Beach, FL, USA, 2018, pp. 172–177.

[22] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "A dual splitting approach for distributed resource allocation with regularization," *IEEE Control Netw. Syst.*, vol. 6, no. 1, pp. 403–414, Mar. 2019.

[23] T.-H. Chang, M. Hong, and X. Wang, "Multi-agent distributed optimization via inexact consensus ADMM," *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 482–497, Jan. 2015.

[24] T.-H. Chang, "A proximal dual consensus ADMM method for multi-agent constrained optimization," *IEEE Trans. Signal Process.*, vol. 64, no. 14, pp. 3719–3734, Jul. 2016.

[25] S. A. Alghunaim, K. Yuan, and A. H. Sayed, "A proximal diffusion strategy for multiagent optimization with sparse affine constraints," *IEEE Trans. Autom. Control*, vol. 65, no. 11, pp. 4554–4567, Nov. 2020.

[26] T. Yang *et al.*, "Distributed energy resources coordination over time-varying directed communication networks," *IEEE Control Netw. Syst.*, vol. 6, no. 3, pp. 1124–1134, Sep. 2019.

[27] Y. Xu, T. Han, K. Cai, Z. Lin, G. Yan, and M. Fu, "A distributed algorithm for resource allocation over dynamic digraphs," *IEEE Trans. Signal Process.*, vol. 65, no. 10, pp. 2600–2612, May 2017.

[28] J. Zhang, K. You, and K. Cai, "Distributed dual gradient tracking for resource allocation in unbalanced networks," *IEEE Trans. Signal Process.*, vol. 68, pp. 2186–2198, 2020, doi: 10.1109/TSP.2020.2981762.

[29] A. Falsone, I. Notarnicola, G. Notarstefano, and M. Prandini, "Tracking-ADMM for distributed constraint-coupled optimization," *Automatica*, vol. 117, 2020, Art. no. 108962.

[30] I. Notarnicola and G. Notarstefano, "Constraint-coupled distributed optimization: Relaxation and duality approach," *IEEE Control Netw. Syst.*, vol. 7, no. 1, pp. 483–492, Mar. 2020.

[31] A. Falsone, K. Margellos, S. Garatti, and M. Prandini, "Dual decomposition for multi-agent distributed optimization with coupling constraints," *Automatica*, vol. 84, pp. 149–158, Oct. 2017.

[32] T.-H. Chang, A. Nedić, and A. Scaglione, "Distributed constrained optimization by consensus-based primal-dual perturbation method," *IEEE Trans. Autom. Control*, vol. 59, no. 6, pp. 1524–1538, Jun. 2014.

[33] S. Sundhar Ram, A. Nedić, and V. V. Veeravalli, "A new class of distributed optimization algorithms: Application to regression of distributed data," *Optim. Methods Softw.*, vol. 27, no. 1, pp. 71–88, 2012.

[34] B. Ying, K. Yuan, and A. H. Sayed, "Supervised learning under distributed features," *IEEE Trans. Signal Process.*, vol. 67, no. 4, pp. 977–992, Feb. 2019.

[35] L. He, A. Bian, and M. Jaggi, "COLA: Decentralized linear learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, Canada, 2018, pp. 4536–4546.

[36] N. S. Aybat and E. Y. Hamedani, "A distributed ADMM-like method for resource sharing over time-varying networks," *SIAM J. Optim.*, vol. 29, no. 4, pp. 3036–3068, 2019.

[37] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. 31, no. 9, pp. 803–812, Sep. 1986.
[38] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.
[39] A. H. Sayed, "Adaptive networks," *Proc. IEEE*, vol. 102, no. 4, pp. 460–497, Apr. 2014.
[40] A. H. Sayed, "Adaptation, learning, and optimization over neworks," *Found. Trends Mach. Learn*, vol. 7, no. 4-5, pp. 311–801, 2014.
[41] Z. Li, W. Shi, and M. Yan, "A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates," *IEEE Trans. Signal Process.*, vol. 67, no. 17, pp. 4494–4506, Sep. 2019.
[42] S. A. Alghunaim, E. K. Ryu, K. Yuan, and A. H. Sayed, "Decentralized proximal gradient algorithms with linear convergence rates," *IEEE Trans. Autom. Control*, vol. 66, no. 6, pp. 2787–2794, Jun. 2021.
[43] R. Xin, S. Pu, A. Nedić, and U. A. Khan, "A general framework for decentralized optimization with first-order methods," *Proc. IEEE*, vol. 108, no. 11, pp. 1869–1889, Nov. 2020.
[44] Y. Sun, A. Daneshmand, and G. Scutari, "Distributed optimization based on gradient-tracking revisited: Enhancing convergence rate via surrogation," May 2019, *arXiv:1905.02637*.
[45] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Berlin, Germany: Springer, 2011, vol. 408.
[46] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[47] A. J. Laub, *Matrix Analysis for Scientists and Engineers*. Philadelphia, PA, USA: SIAM, 2004.
[48] S. A. Alghunaim, K. Yuan, and A. H. Sayed, "A linearly convergent proximal gradient algorithm for decentralized optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, Canada, 2019, pp. 2844–2854.
[49] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the ADMM in decentralized consensus optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1750–1761, Apr. 2014.
[50] S. A. Alghunaim and A. H. Sayed, "Linear convergence of primal-dual gradient methods and their performance in distributed optimization," *Automatica*, vol. 117, Jul. 2020, Art. no. 109003.
[51] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Stat. Soc.: Ser. B. (Stat. Methodol.)*, vol. 67, no. 2, pp. 301–320, 2005.
[52] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," [Online]. Available: http://cvxr.com/cvx, Mar. 2014.
[53] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.
[54] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

Sulaiman A. Alghunaim received the B.S. degree in electrical engineering from Kuwait University, Kuwait City, Kuwait, in 2013, and the M.S. degree in electrical engineering and the Ph.D. degree in electrical and computer engineering from the University of California, Los Angeles, Los Angeles, CA, USA, in 2016 and 2020, respectively. He is currently an Assistant Professor with Electrical Engineering Department, Kuwait University. His research interests include optimization, machine learning, signal processing, and control. His current research interests include the development and analysis of distributed optimization and learning algorithms.

Qi Lyu received the B.S. degree from Xi'an Jiaotong University, Xi'an, China, and the M.S. degree from Michigan State University, East Lansing, MI, USA, in 2021. He was with the Department of Computational Mathematics, Science, and Engineering, during his M.S. study. His research is about optimization and deep learning coding.

Ming Yan received the B.S. and M.S. degrees from the University of Science and Technology of China, Hefei, China, and the Ph.D. degree from the University of California, Los Angeles, Los Angeles, CA, USA, in 2012. He is currently an Associate Professor with the Department of Computational Mathematics, Science, and Engineering and the Department of Mathematics, Michigan State University, East Lansing, MI, USA. His current research interests include optimization methods and their applications in sparse recovery and regularized inverse problems, variational methods for image processing, and parallel and distributed algorithms for solving Big Data problems.

Ali H. Sayed (Fellow, IEEE) is currently the Dean of engineering with EPFL, Switzerland. He has served before as a Distinguished Professor and the Chairman of electrical engineering with the University of California, Los Angeles, Los Angeles, CA, USA. He has authored more than 550 scholarly publications and six books. His research interests include adaptation and learning theories, network and data sciences, statistical inference, and distributed optimization. He is recognized as a Highly Cited Researcher, and is a Member of the US National Academy of Engineering. He was the President of the IEEE Signal Processing Society in 2018 and 2019. His work was the recipient of several main awards.