A Penalty Alternating Direction Method of Multipliers for Convex Composite Optimization over Decentralized Networks

Jiaojiao Zhang, Huikang Liu, Anthony Man-Cho So, and Qing Ling

***Abstract*—Consider the problem of minimizing a sum of convex composite functions over a decentralized network, where each agent in the network holds a private function consisting of a smooth part and a nonsmooth part, and it can only exchange information with its neighbors during the optimization process. One approach to tackling this problem is to study its penalized approximation. Although such an approximation becomes more accurate as the penalty parameter becomes smaller, it is well known that the penalized objective will also become more ill-conditioned, thereby causing the popular proximal gradient descent method to slow down substantially. To break this accuracy-speed tradeoff, we propose to solve the penalized approximation with the alternating direction method of multipliers (ADMM). We also exploit the composite structure of the private functions by linearizing the smooth parts and handling the nonsmooth parts with proximal operators, which allows us to further reduce the computational costs. The proposed penalty ADMM (abbreviated as PAD) is proven to be sublinearly convergent when the private functions are convex, and linearly convergent when in addition the smooth parts are strongly convex. We present numerical results to corroborate the theoretical analyses and to further demonstrate the advantages of PAD over existing state-of-the-art algorithms such as DL-ADMM, PG-EXTRA, and NIDS.**

***Index Terms*—Decentralized optimization, alternating direction method of multipliers (ADMM), composite optimization**

I. Introduction

This paper focuses on the following decentralized composite optimization problem defined over an undirected and connected network with n agents:

$$\hat{x}^* \in \arg\min_{x\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}\big(f_i(x)+g_i(x)\big). \tag{1}$$

Here, $f_i : \mathbb{R}^p \to \mathbb{R}$ and $g_i : \mathbb{R}^p \to \mathbb{R}\cup\{+\infty\}$ are two convex functions privately owned by agent i. We assume that f_i is continuously differentiable, while g_i is closed proper and its proximal mapping is easy to compute. Every agent aims to obtain an optimal solution $\hat{x}^*$ to (1) via local computation and communication with its neighbors. Such decentralized composite optimization problems appear in various application fields, including network optimization [1], [2], optimization-based cooperative control [3], and decentralized machine learning [4], [5].

Decentralized optimization methods have been extensively studied in the literature. They usually operate in either the primal or dual domain. For the primal domain methods, distributed gradient descent (DGD) and subsequent extensions are studied in [6]–[8]. With a fixed step size, DGD converges fast but only to a neighborhood of an optimal solution, and the size of the neighborhood is proportional to the step size [6], [7]. With a diminishing step size, DGD is able to converge to an optimal solution, but the speed is slow [8]. The inaccuracy is due to the fact that DGD is essentially a gradient descent method to solve a penalized approximation of (1), where the step size determines the penalty parameter and hence also the approximation error [7]. Second-order methods are also applicable to solving the penalized approximation with the same accuracy-speed tradeoff [9], [10].

The dual domain methods include decentralized alternating direction method of multipliers (ADMM) [11]–[13], decentralized linearized ADMM (DLM) [14], decentralized augmented Lagrangian-based algorithms [15], and dual accelerated schemes [16]. All these algorithms solve a reformulation of (1) with consensus constraints in the dual domain. The use of fixed step sizes enables them to have fast and exact convergence. When the local functions are smooth, some of these algorithms are proven to converge linearly to an exact optimal solution [12], [13], [15], [16]. When the local functions are nonsmooth and have the composite form as in (1), the proximal decentralized linearized ADMM (DL-ADMM) [17] is shown in [18] to have the sublinear ergodic convergence rate of $O(\frac{1}{k})$, where k is the iteration counter.

There are also other decentralized optimization methods that do not explicitly operate in the dual domain but are still able to converge to an exact optimal solution with fixed step sizes [19]–[26]. For instance, the EXTRA algorithm applies to the case where the local functions only have smooth parts [19]. When the local functions have the composite form as in (1), the PG-EXTRA algorithm proposed in [20] and the NIDS algorithm proposed in [21] will converge to an optimal solution at the rates of $O(\frac{1}{k})$ and $o(\frac{1}{k})$, respectively. The algorithms proposed in [22], [23] and [26] have a linear convergence rate, but the results assume that the nonsmooth parts are common across all the agents.

Jiaojiao Zhang, Huikang Liu and Anthony Man-Cho So are with Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong.

Qing Ling is with School of Data and Computer Science and Guangdong Province Key Laboratory of Computational Science, Sun Yat-Sen University. Qing Ling is supported in part by NSF China Grants 61573331 and 61973324, and Fundamental Research Funds for the Central Universities.

A short version of this paper appeared in IEEE International Conference on Acoustics, Speech, and Signal Processing, Barcelona, Spain, May 4-8, 2020.

TABLE I
CONVERGENCE RATES OF DIFFERENT ALGORITHMS FOR DECENTRALIZED CONVEX COMPOSITE OPTIMIZATION

Algorithm	Problem	Step size	Rate	Exactness
DL-ADMM [17]	$\min_x \sum_{i=1}^n (f_i(x) + g_i(x))$	$c < \frac{\alpha\lambda_{\min}(D+A)}{L_f^2/\mu_f^2}$ [1]	no rate	exact
PG-ADMM [18]	$\min_x \sum_{i=1}^n (f_i(x) + g_i(x))$	$c_i < \frac{1}{L_i+\alpha_i d_i}$ [2]	$O(\frac{1}{k})$	exact
PG-EXTRA [20]	$\min_x \sum_{i=1}^n (f_i(x) + g_i(x))$	$c \le \frac{2\lambda_{\min}((I+W)/2)}{L_f}$ [3]	$O(\frac{1}{k})$	exact
NIDS [21]	$\min_x \sum_{i=1}^n (f_i(x) + g_i(x))$	$c_i < \frac{2}{L_i}$ [4]	$o(\frac{1}{k})$	exact
PGA [22]	$\min_x \sum_{i=1}^n (f_i(x) + g(x))$	$c \le \frac{2(1-\lambda_{\max}(I-W))}{L_f+\mu_f}$ [5]	linear	exact
NEXT/SONATA [23]	$\min_x \sum_{i=1}^n (f_i(x) + g(x))$	$c < \min\{\frac{\mu_{\min}}{\sum_{i=1}^n L_i}, \alpha_1\}$ [6]	linear	exact
PAD (convex)	$\min_x \sum_{i=1}^n (f_i(x) + g_i(x))$	$c < \frac{1}{L_f+\alpha\lambda_{\max}(I-W)}$ [7]	$O(\frac{1}{k})$	inexact
PAD (strongly convex)	$\min_x \sum_{i=1}^n (f_i(x) + g_i(x))$	$c < \frac{1}{L_f^2/\mu_f+\alpha\lambda_{\max}(I-W)}$	linear	inexact

In this paper, we consider solving a penalized approximation of (1) just like the primal domain methods. However, unlike those methods, we allow the penalty parameter to take small values so as to obtain accurate approximations. It is well known that a small penalty parameter will lead to an ill-conditioned penalized objective. Consequently, the popular proximal gradient descent method will have to use a small step size, which makes it very slow. To overcome this difficulty, we propose to solve the penalized formulation with ADMM. We also exploit the composite structure of the private functions by linearizing the smooth parts and handling the nonsmooth parts with proximal operators, which allows us to further reduce the computational costs. The proposed penalty ADMM (abbreviated as PAD) is proven to be sublinearly convergent when the private functions are convex, and linearly convergent when in addition the smooth parts are strongly convex. A detailed comparison of our convergence results and the existing ones for decentralized convex composite optimization can be found in Table I. Different from the problem $\min_x \sum_{i=1}^n (f_i(x)+g_i(x))$ in (1), the problem $\min_x \sum_{i=1}^n (f_i(x) + g(x))$ in the table means that the nonsmooth function $g : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is common to all agents. Indeed, [27] shows that proximal decentralized algorithms cannot guarantee exact linear convergence when the nonsmooth functions g_i are different. In this case, linear convergence to an inexact solution has been studied in the context of penalty methods [28]–[31]. Our numerical results show that PAD allows the use of a small penalty parameter without sacrificing the convergence speed. In addition, we establish, for the first time, an upper bound on the distance between the original and penalized solutions, which is determined by the penalty parameter, for this nonsmooth formulation. We also demonstrate the connection between PAD and several dual domain methods, including EXTRA [19], DLM [14], and PG-EXTRA [20]. Lastly, we present numerical results to corroborate the theoretical analyses and to further demonstrate the advantages of PAD over existing state-of-the-art algorithms such as DL-ADMM, PG-EXTRA, and NIDS.

[1]Here, c and α are the step sizes of primal update and dual update, respectively; L_f is the Lipschitz constant of the gradients ∇f_i, μ_f is the strong convexity constant of the functions f_i, $D \in \mathbb{R}^{n\times n}$ is the degree matrix, $A \in \mathbb{R}^{n\times n}$ is the adjacency matrix, and $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of its argument.

[2]Here, c_i and α_i are the step sizes of primal update and dual update, respectively, on agent i; L_i is the Lipschitz constant of the gradient ∇f_i; d_i is the degree of agent i.

[3]Here, $W \in \mathbb{R}^{n\times n}$ is a mixing matrix; see the definition in Assumption 3.

[4]Here, NIDS can use uncoordinated step sizes. It has another parameter α that satisfies $I - \alpha\Gamma^{1/2}(I-W)\Gamma^{1/2} \succeq 0$ with $\Gamma = \mathrm{diag}\{c_i\}$.

[5]Here, $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of its argument. The step size is independent of the network topology.

[6]Here, $\tilde{\mu}_{\min} = \min_i \tilde{\mu}_i$, where $\tilde{\mu}_i$ is the strong convexity parameter of the successive convex approximation surrogate of f_i, and $\alpha_1 > 0$ is a constant.

[7]Here, $\alpha > 0$ is a constant. PAD converges to a neighborhood of an optimal solution, but the size of the neighborhood can be arbitrarily small. The same comment applies to the strongly convex case.

A. Paper Organization

This paper is organized as follows. Subsection I-B introduces the notations. The penalized approximation formulation is given in Subsection II-A and the proposed PAD algorithm is developed in Subsection II-B. The relationship between PAD and several dual domain methods are discussed in Subsection II-C. The convergence behavior of PAD under general convexity is studied in Subsection III-A and Subsection III-B. The linear convergence of PAD under a strong convexity assumption is established in Subsection III-C. The error caused by penalization is bounded in Subsection III-D. The effectiveness of the proposed method is verified through numerical experiments in Section IV. Finally, the conclusions are given in Section V.

B. Notations

In this paper, we consider an undirected and connected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N} \triangleq \{1, \ldots, n\}$ denotes the set of nodes and $\mathcal{E}$ denotes the set of edges. Let $\mathcal{N}_i \triangleq \{j \in \mathcal{N} : (i,j) \in \mathcal{E} \text{ or } (j,i) \in \mathcal{E}\} \cup \{i\}$ denote the set of neighboring nodes of i including itself. Each agent i holds a local variable $x_i \in \mathbb{R}^p$, whose value in the k-th iteration is denoted by x_i^k. We use $\mathbf{x}$ to denote the matrix obtained by stacking up the row vectors $x_1^T, \ldots, x_n^T$; i.e.,

$$\mathbf{x} \triangleq \begin{pmatrix} - & x_1^T & - \\ - & x_2^T & - \\ & \vdots & \\ - & x_n^T & - \end{pmatrix} \in \mathbb{R}^{n\times p}.$$

We say that $\mathbf{x}$ is consensual if all of its rows are identical; i.e., $x_1^T = \cdots = x_n^T$. Next, we denote the sums of the smooth and nonsmooth parts in (1) by

$$f(\mathbf{x}) \triangleq \sum_{i=1}^{n} f_i(x_i) \quad \text{and} \quad g(\mathbf{x}) \triangleq \sum_{i=1}^{n} g_i(x_i),$$

respectively. The gradient of f at $\mathbf{x}$ is given by

$$\nabla f(\mathbf{x}) \triangleq \begin{pmatrix} - & (\nabla f_1(x_1))^T & - \\ - & (\nabla f_2(x_2))^T & - \\ & \vdots & \\ - & (\nabla f_n(x_n))^T & - \end{pmatrix} \in \mathbb{R}^{n\times p},$$

where $\nabla f_i(x_i)$ is the gradient of f_i at x_i. A subgradient of g at $\mathbf{x}$ is given by

$$\partial g(\mathbf{x}) \triangleq \begin{pmatrix} - & (\partial g_1(x_1))^T & - \\ - & (\partial g_2(x_2))^T & - \\ & \vdots & \\ - & (\partial g_n(x_n))^T & - \end{pmatrix} \in \mathbb{R}^{n\times p},$$

where $\partial g_i(x_i)$ is a subgradient of g_i at x_i. The ith rows of $\mathbf{x}$, ∇f, and ∂g belong to agent i.

Given a vector v, $\|v\|$ and $\|v\|_1$ denote the ℓ_2 and ℓ_1 norms, respectively. Given a matrix A, a_{ij} denotes the element in the ith row and jth column, while a_i^T denotes the ith row of A. The Frobenius norm of a matrix A is denoted by $\|A\|_{\mathcal{F}}$. Given a symmetric matrix G, we use $G \succeq 0$ and $G \succ 0$ to mean that G is positive semidefinite and positive definite, respectively. For $G \succeq 0$, we define the inner product $\langle A, A'\rangle_G \triangleq \langle A, GA'\rangle$ and the induced norm $\|A\|_G \triangleq \sqrt{\mathrm{trace}\,(A^{\mathrm{T}}GA)}$. The largest and smallest eigenvalues of a symmetric matrix A are denoted by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$, respectively. We use I to denote the identity matrix; $\mathbf{1}_{m\times n}$ and $\mathbf{0}_{m\times n}$ to denote the $m \times n$ all-ones and $m \times n$ all-zeros matrices, respectively. Given a closed proper convex function $\tilde{g}$, the proximal mapping of $\tilde{g}$ is given by

$$\mathrm{prox}_{c\tilde{g}}(x) \triangleq \arg\min_y \left\{ \tilde{g}(y) + \frac{1}{2c}\|y - x\|^2 \right\},$$

where $c > 0$ is a scalar parameter. The norm is ℓ_2 when y is a vector, and Frobenius when y is a matrix.

II. Penalty ADMM

In this section, we give a penalized approximation formulation of (1) and derive the proposed PAD. After that, we discuss the relationship between PAD and several dual domain methods. In our algorithm development and convergence analysis, we make the following assumptions on the smooth and nonsmooth parts of the local functions.

Assumption 1. *The gradient ∇f_i is Lipschitz continuous; i.e.,*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \le L_f\|x - y\|,\ \forall x, y \in \mathbb{R}^p,$$

where $L_f > 0$ is the Lipschitz constant.

The proposed algorithm is going to utilize the composite structure of the local functions. We will show its convergence and sublinear convergence rate under Assumption 1. To establish its linear convergence rate, we will need another assumption.

Assumption 2. *Each f_i is in addition strongly convex; i.e.,*

$$\langle x - y, \nabla f_i(x) - \nabla f_i(y)\rangle \ge \mu_f\|x - y\|^2,\ \forall x, y \in \mathbb{R}^p,$$

where $\mu_f > 0$ is the strong convexity constant.

We begin with a consensus-constrained reformulation of (1). Let $x_i \in \mathbb{R}^p$ be the local variable held by agent i. To characterize the consensual property $x_1^T = \cdots = x_n^T$, we introduce a weight w_{ij} between agents i and j and collect all the weights in a mixing matrix $W = [w_{ij}] \in \mathbb{R}^{n\times n}$. We make the following assumption on W, which is standard in the distributed optimization literature.

Assumption 3. *The mixing matrix W is symmetric and doubly stochastic; i.e., $W = W^T$ and $W\mathbf{1}_{n\times 1} = \mathbf{1}_{n\times 1}$. The null space of $I - W$ is $\mathrm{span}\,(\mathbf{1}_{n\times 1})$. Furthermore, if $j \notin \mathcal{N}_i$, then $w_{ij} = 0$; otherwise, $w_{ij} > 0$.*

When the underlying network is connected, a mixing matrix W satisfying Assumption 3 can be generated using the techniques introduced in [32]. According to the Perron-Frobenius theorem [33], Assumption 3 implies that the eigenvalues of W lie in $(-1, 1]$ and the multiplicity of eigenvalue 1 is one. Since the null space of $I - W$ is $\mathrm{span}\,(\mathbf{1}_{n\times 1})$, so is the null space of its square root $(I-W)^{\frac{1}{2}}$. Therefore, $(I-W)^{\frac{1}{2}}\mathbf{x} = 0$ if and only if $x_1^T = \cdots = x_n^T$. In particular, we see that (1) is equivalent to the constrained problem

$$\hat{\mathbf{x}}^* \in \arg\min_{\mathbf{x}}\ \{f(\mathbf{x}) + g(\mathbf{x})\} \quad \text{s.t. } (I - W)^{\frac{1}{2}}\mathbf{x} = 0, \quad (2)$$

where $\hat{\mathbf{x}}^* \in \mathbb{R}^{n\times p}$ contains n identical rows of $(\hat{x}^*)^T$.

A. Penalized Approximation Formulation

Instead of solving the constrained problem (2), we consider its penalized approximation in the form of

$$\mathbf{x}^* \in \arg\min_{\mathbf{x}}\ \left\{ f(\mathbf{x}) + g(\mathbf{x}) + \frac{1}{2\epsilon}\|(I - W)^{\frac{1}{2}}\mathbf{x}\|_{\mathcal{F}}^2 \right\}, \quad (3)$$

where $\epsilon > 0$ is the penalty parameter. The approximation error (i.e., the gap between $\hat{\mathbf{x}}^*$ and $\mathbf{x}^*$) is controlled by ϵ. When ϵ is sufficiently small, the approximation error is negligible.

Assumption 4. *The sets of minimizers of the original problem (1) and the penalized problem (3) are nonempty.*

When $g(\mathbf{x}) = 0$, problem (3) can be solved using gradient descent, whose iterates are generated by

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma\left(\nabla f(\mathbf{x}^k) + \frac{1}{\epsilon}(I - W)\mathbf{x}^k\right) \quad (4)$$

with $\gamma > 0$ being the step size. However, when ϵ is very small, the Lipschitz constant of the gradient $\nabla f(\mathbf{x}) + \frac{1}{\epsilon}(I - W)\mathbf{x}$ is on the order of $O(\frac{1}{\epsilon})$, which suggests that γ must be on the order of $O(\epsilon)$ to guarantee convergence [34]. In fact, setting $\gamma = \epsilon$ recovers the DGD update $\mathbf{x}^{k+1} = W\mathbf{x}^k - \epsilon\nabla f(\mathbf{x}^k)$ [6], [7]. Even though DGD will converge to a small neighborhood of $\hat{\mathbf{x}}^*$ when ϵ is small, the small step size will make it converge very slowly. This dilemma demonstrates the unfavorable

accuracy-speed tradeoff of DGD. The same statement holds true when we apply the proximal gradient method to handle the case where $g(\mathbf{x}) \neq 0$ [35].

B. Algorithm Development

To address this issue, we propose an ADMM-based algorithm to solve (3). It is worth noting that although ADMM is a popular dual domain method for decentralized optimization, here we apply it to the penalized approximation formulation, which appears in the primal domain. By introducing an auxiliary variable $\mathbf{z} \in \mathbb{R}^{n\times p}$, problem (3) is equivalent to

$$\begin{aligned}(\mathbf{x}^*, \mathbf{z}^*) = \arg\min_{\mathbf{x},\mathbf{z}} \ &\left\{ f(\mathbf{x}) + g(\mathbf{x}) + \frac{1}{2\epsilon}\|\mathbf{z}\|_{\mathcal{F}}^2 \right\}, \\ &\text{s.t. } (I-W)^{\frac{1}{2}}\mathbf{x} = \mathbf{z}.\end{aligned} \tag{5}$$

The augmented Lagrangian function of (5) is given by

$$\begin{aligned}L_\alpha(\mathbf{x},\mathbf{z},\Pi) \triangleq & f(\mathbf{x}) + g(\mathbf{x}) + \frac{1}{2\epsilon}\|\mathbf{z}\|_{\mathcal{F}}^2 + \left\langle \Pi, (I-W)^{\frac{1}{2}}\mathbf{x} - \mathbf{z}\right\rangle \\ &+ \frac{\alpha}{2}\|(I-W)^{\frac{1}{2}}\mathbf{x} - \mathbf{z}\|_{\mathcal{F}}^2,\end{aligned}$$

where $\Pi \in \mathbb{R}^{n\times p}$ is the Lagrange multiplier and $\alpha > 0$ is a constant. ADMM minimizes the augmented Lagrangian function over the primal variables $\mathbf{x}$ and $\mathbf{z}$ in an alternating manner, then updates the dual variable Π through dual gradient ascent. However, on account of the nonsmooth term g and the coefficient $(I-W)^{\frac{1}{2}}$ in the quadratic term in $L_\alpha(\mathbf{x},\mathbf{z},\Pi)$, the subproblem w.r.t. $\mathbf{x}$ does not have a closed-form solution. Thus, we utilize the composite structure to inexactly update $\mathbf{x}$. Specifically, we linearize the smooth part and calculate the proximal mapping of the nonsmooth part as follows.

Update of $\mathbf{x}$. We separate $L_\alpha(\mathbf{x},\mathbf{z}^k,\Pi^k)$ into a nonsmooth part $g(\mathbf{x})$ plus a smooth part $\tilde{L}_\alpha(\mathbf{x},\mathbf{z}^k,\Pi^k)$, where

$$\begin{aligned}\tilde{L}_\alpha(\mathbf{x},\mathbf{z}^k,\Pi^k) \triangleq & f(\mathbf{x}) + \frac{1}{2\epsilon}\|\mathbf{z}^k\|_{\mathcal{F}}^2 + \left\langle \Pi^k, (I-W)^{\frac{1}{2}}\mathbf{x} - \mathbf{z}^k\right\rangle \\ &+ \frac{\alpha}{2}\|(I-W)^{\frac{1}{2}}\mathbf{x} - \mathbf{z}^k\|_{\mathcal{F}}^2.\end{aligned}$$

The idea here is to replace the smooth part $\tilde{L}_\alpha(\mathbf{x},\mathbf{z}^k,\Pi^k)$ by the following quadratic approximation centered at $\mathbf{x}^k$:

$$\begin{aligned}\tilde{L}_\alpha(\mathbf{x},\mathbf{z}^k,\Pi^k) \approx & \tilde{L}_\alpha(\mathbf{x}^k,\mathbf{z}^k,\Pi^k) + \left\langle \nabla_{\mathbf{x}}\tilde{L}_\alpha(\mathbf{x}^k,\mathbf{z}^k,\Pi^k), \mathbf{x}-\mathbf{x}^k\right\rangle \\ &+ \frac{1}{2c}\|\mathbf{x}-\mathbf{x}^k\|_{\mathcal{F}}^2.\end{aligned}$$

Here, $c > 0$ is the step size and $\nabla_{\mathbf{x}}\tilde{L}_\alpha(\mathbf{x},\mathbf{z}^k,\Pi^k)$ is the gradient of $\tilde{L}_\alpha(\mathbf{x},\mathbf{z}^k,\Pi^k)$ w.r.t. $\mathbf{x}$. Using this approximation in $\min_{\mathbf{x}} L_\alpha(\mathbf{x},\mathbf{z}^k,\Pi^k)$ leads to the primal update

$$\begin{aligned}\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} \ &\Big\{ g(\mathbf{x}) + \tilde{L}_\alpha(\mathbf{x}^k,\mathbf{z}^k,\Pi^k) \\ &+ \left\langle \nabla_{\mathbf{x}}\tilde{L}_\alpha(\mathbf{x}^k,\mathbf{z}^k,\Pi^k), \mathbf{x}-\mathbf{x}^k\right\rangle + \frac{1}{2c}\|\mathbf{x}-\mathbf{x}^k\|_{\mathcal{F}}^2 \Big\},\end{aligned}$$

which is the proximal mapping of g and has the closed-form solution

$$\begin{aligned}\mathbf{x}^{k+1} = \mathrm{prox}_{cg}\Big(\mathbf{x}^k - c\big[&\nabla f(\mathbf{x}^k) + \alpha(I-W)^{\frac{1}{2}} \\ &((I-W)^{\frac{1}{2}}\mathbf{x}^k - \mathbf{z}^k + \frac{\Pi^k}{\alpha})\big]\Big).\end{aligned} \tag{6}$$

Update of $\mathbf{z}$. Update $\mathbf{z}^{k+1}$ from $\mathbf{z}^{k+1} = \arg\min_{\mathbf{z}}\{\frac{1}{2\epsilon}\|\mathbf{z}\|_{\mathcal{F}}^2 + \left\langle \Pi^k, (I-W)^{\frac{1}{2}}\mathbf{x}^{k+1} - \mathbf{z}\right\rangle + \frac{\alpha}{2}\|(I-W)^{\frac{1}{2}}\mathbf{x}^{k+1} - \mathbf{z}\|_{\mathcal{F}}^2\}$. It is equivalent to

$$\mathbf{z}^{k+1} = \frac{1}{\alpha + \frac{1}{\epsilon}}\big[\Pi^k + \alpha(I-W)^{\frac{1}{2}}\mathbf{x}^{k+1}\big]. \tag{7}$$

Update of Π. Finally, the update of Π^{k+1} is given by

$$\Pi^{k+1} = \Pi^k + \alpha\big[(I-W)^{\frac{1}{2}}\mathbf{x}^{k+1} - \mathbf{z}^{k+1}\big]. \tag{8}$$

According to (6)–(8), c and α are the primal and dual step sizes to update $\mathbf{x}$ and Π, respectively. Note that they are independent of ϵ, the penalty parameter. As we will show in the convergence analysis, even when ϵ is small, PAD can still use large step sizes α and c, which leads to its favorable convergence speed.

The updates in (6)–(8) can be simplified by introducing $\bar{\Pi}^k \triangleq (I-W)^{\frac{1}{2}}\Pi^k$ and $\bar{\mathbf{z}}^k \triangleq (I-W)^{\frac{1}{2}}\mathbf{z}^k$. With these notations and splitting the updates to the agents, we outline the proposed PAD in Algorithm 1. The implementation of Step 2 requires neighboring variables x_j^k from the previous iteration. The implementations of Steps 4 and 5 require current neighboring variables x_j^{k+1}, which become available through the variable exchange implemented in Step 3. This variable exchange also makes variables available for the update in Step 2 of the next iteration.

As shown in Algorithm 1, there is only one round of communication in each iteration of PAD. The proximal-linear operation for the x_i-subproblem reduces the computational complexity. Thus, both the communication and computational costs of PAD are low.

Algorithm 1 PAD run by agent i

Require: Choose the parameters ϵ, α, and c. Initialize $x_i^0 = 0$, $\bar{z}_i^0 = 0$, and $\bar{\pi}_i^0 = 0$.

1: **for** $k = 1, 2, \ldots$ **do**

2: Update local variable x_i^{k+1} by

$$\begin{aligned}x_i^{k+1} = \mathrm{prox}_{cg_i}\Bigg(x_i^k - c\Bigg[&\nabla f_i(x_i^k) \\ &+ \alpha\left(x_i^k - \sum_{j\in\mathcal{N}_i} w_{ij}x_j^k - \bar{z}_i^k\right) + \bar{\pi}_i^k\Bigg]\Bigg).\end{aligned}$$

3: Transmit x_i^{k+1}/ receive x_j^{k+1} from neighbors $j \in \mathcal{N}_i$.

4: Update local auxiliary variable $\bar{z}_i^{k+1}$ by

$$\bar{z}_i^{k+1} = \frac{1}{\alpha + \frac{1}{\epsilon}}\left[\bar{\pi}_i^k + \alpha\left(x_i^{k+1} - \sum_{j\in\mathcal{N}_i} w_{ij}x_j^{k+1}\right)\right].$$

5: Update local dual variable $\bar{\pi}_i^{k+1}$ by

$$\bar{\pi}_i^{k+1} = \bar{\pi}_i^k + \alpha\left[\left(x_i^{k+1} - \sum_{j\in\mathcal{N}_i} w_{ij}x_j^{k+1}\right) - \bar{z}_i^{k+1}\right].$$

6: **end for**

C. Connection between PAD and Dual Domain Methods

Although PAD solves the penalized approximation formulation (3) (or equivalently (5)), which is often associated with the

design of decentralized *primal domain* methods, here we show that PAD is closely related to several existing decentralized *dual domain* methods.

We begin with the scenario where $g(\mathbf{x}) = 0$ and show that the dual domain method EXTRA proposed in [19] is a special case of PAD. The update of EXTRA is

$$\begin{aligned}\mathbf{x}^{k+2} =&(I+W)\mathbf{x}^{k+1} - \frac{I+W}{2}\mathbf{x}^k \\ &- c\big[\nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k)\big],\end{aligned} \tag{9}$$

where $c > 0$ is the step size. To compare PAD with EXTRA, we eliminate the variables $\bar{\mathbf{z}}$ and $\bar{\Pi}$ in PAD as follows.

Combining (8) and (7) to eliminate $(I-W)^{\frac{1}{2}}\mathbf{x}^{k+1}$, we get

$$\frac{1}{\epsilon}\mathbf{z}^{k+1} = \Pi^{k+1}. \tag{10}$$

When the variables are also initialized as $\frac{1}{\epsilon}\mathbf{z}^0 = \Pi^0$, we have $\frac{1}{\epsilon}\mathbf{z}^k = \Pi^k$, and consequently, $\frac{1}{\epsilon}\bar{\mathbf{z}}^k = \bar{\Pi}^k$. When $g(\mathbf{x}) = 0$, substituting $\frac{1}{\epsilon}\bar{\mathbf{z}}^k = \bar{\Pi}^k$ into (6) yields

$$\begin{aligned}\mathbf{x}^{k+1} =&\mathbf{x}^k - c\big[\nabla f(\mathbf{x}^k) \\ &+ \alpha(I-W)\mathbf{x}^k + (1-\alpha\epsilon)\bar{\Pi}^k\big]\end{aligned} \tag{11}$$

and

$$\begin{aligned}\mathbf{x}^{k+2} =&\mathbf{x}^{k+1} - c\big[\nabla f(\mathbf{x}^{k+1}) \\ &+ \alpha(I-W)\mathbf{x}^{k+1} + (1-\alpha\epsilon)\bar{\Pi}^{k+1}\big].\end{aligned} \tag{12}$$

Multiplying both sides of (12) by $(1+\alpha\epsilon)$, subtracting (11) from the resulting equality, and then eliminating $\bar{\Pi}^{k+1} - \bar{\Pi}^k$ by (8), we obtain

$$\begin{aligned}\mathbf{x}^{k+2} =& \frac{1}{1+\alpha\epsilon}\Big\{\big((2+\alpha\epsilon-2c\alpha)I + 2c\alpha W\big)\mathbf{x}^{k+1} \\ &- \big[I - c\alpha(I-W)\big]\mathbf{x}^k - c\big[(1+\alpha\epsilon)\nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k)\big]\Big\}.\end{aligned} \tag{13}$$

Comparing (9) with (13), we observe that the updates of PAD and EXTRA are equivalent if $c\alpha = \frac{1}{2}$ and $\epsilon = 0$. The parameter ϵ tunes the accuracy-speed tradeoff in PAD. In addition, $c\alpha$ can be set to values other than $\frac{1}{2}$ in PAD. These flexibilities enable PAD to converge faster than EXTRA and its proximal-gradient variant.

With (13), we further show that DLM [14] is also connected with PAD. Let $\epsilon = 0$, replace c with a diagonal matrix $(\rho I + 2\alpha D)^{-1}$ and set $I - W = L_o$, where $\rho > 0$ is an approximation parameter, D is the degree matrix, and L_o is the oriented Laplacian defined in [14]. Then, PAD becomes

$$\begin{aligned}\mathbf{x}^{k+2} =&\big(I - \alpha(\rho I + 2\alpha D)^{-1}L_o\big)(2\mathbf{x}^{k+1} - \mathbf{x}^k) \\ &- (\rho I + 2\alpha D)^{-1}\big[\nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k)\big],\end{aligned}$$

which is the iterate of DLM after eliminating the dual variable.

When $g(\mathbf{x}) \neq 0$, PAD utilizes the proximal mapping (6) to handle this nonsmooth term. In this scenario the proximal-gradient (PG-)EXTRA algorithm [20] modifies EXTRA to

$$\begin{aligned}\mathbf{y}^{k+1} =&\mathbf{y}^k - \mathbf{x}^k + \frac{I+W}{2}\Big(2\mathbf{x}^k - \mathbf{x}^{k-1}\Big) \\ &- c\big[\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1})\big], \\ \mathbf{x}^{k+1} =&\operatorname{prox}_{cg}(\mathbf{y}^{k+1}).\end{aligned}$$

Below we show that PAD can recover PG-EXTRA too. Let $\mathbf{y}^{k+1} = \mathbf{x}^k - c\big[\nabla f(\mathbf{x}^k) + \alpha(I-W)^{\frac{1}{2}}\big((I-W)^{\frac{1}{2}}\mathbf{x}^k - \mathbf{z}^k + \frac{\Pi^k}{\alpha}\big)\big]$. With $\bar{\mathbf{z}}^k = (I-W)^{\frac{1}{2}}\mathbf{z}^k$ and $\frac{1}{\epsilon}\bar{\mathbf{z}}^k = \bar{\Pi}^k$, we have

$$\begin{aligned}\mathbf{y}^{k+1} - \mathbf{y}^k =& \big[I - c\alpha(I-W)\big](\mathbf{x}^k - \mathbf{x}^{k-1}) \\ &- c\big(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1})\big) \\ &- c\alpha(1-\alpha\epsilon)\big[(I-W)\mathbf{x}^k \\ &- \frac{\epsilon}{\alpha\epsilon+1}\big(\bar{\Pi}^{k-1} + \alpha(I-W)\mathbf{x}^k\big)\big], \\ \mathbf{x}^{k+1} =& \operatorname{prox}_{cg}(\mathbf{y}^{k+1}).\end{aligned}$$

When $c\alpha = \frac{1}{2}$ and $\epsilon = 0$, the update is exactly PG-EXTRA. In our numerical experiments, we show that PAD converges faster than PG-EXTRA for the decentralized composite optimization problem; see Section IV.

III. Convergence Analysis of PAD

This section establishes the convergence of PAD and analyzes its convergence rates in the convex and strongly convex settings. Instead of considering the simplified sequences $\{\bar{\mathbf{z}}^k\}$ and $\{\bar{\Pi}^k\}$, we investigate the sequences $\{\mathbf{z}^k\}$ and $\{\Pi^k\}$ generated by (6)–(8), because their analyses are equivalent and the sequence $\{\mathbf{x}^k\}$ is identical.

A. Convergence Analysis under General Convexity

In this subsection, we will prove the convergence of PAD for the general convex case. Define the Lagrangian function of (5) as

$$L(\mathbf{x},\mathbf{z},\Pi) \triangleq f(\mathbf{x}) + g(\mathbf{x}) + \frac{1}{2\epsilon}\|\mathbf{z}\|_{\mathcal{F}}^2 + \big\langle \Pi, (I-W)^{\frac{1}{2}}\mathbf{x} - \mathbf{z}\big\rangle.$$

To simplify notations, let us define a triple $\mathbf{u}^k \triangleq (\mathbf{x}^k, \mathbf{z}^k, \Pi^k)$ of the primal and dual variables. Similarly, define $\mathbf{u}^* \triangleq (\mathbf{x}^*, \mathbf{z}^*, \Pi^*)$, where $\mathbf{x}^*$, $\mathbf{z}^*$ and Π^* are the optimal primal and dual solutions to (5), respectively. According to the Karush-Kuhn-Tucker (KKT) conditions of (5), $\mathbf{u}^*$ satisfies

$$-\nabla f(\mathbf{x}^*) - (I-W)^{\frac{1}{2}}\Pi^* \in \partial g(\mathbf{x}^*), \tag{14a}$$

$$\frac{1}{\epsilon}\mathbf{z}^* = \Pi^*, \tag{14b}$$

$$(I-W)^{\frac{1}{2}}\mathbf{x}^* = \mathbf{z}^*. \tag{14c}$$

To characterize the convergence of $\mathbf{u}^k$ to $\mathbf{u}^*$, define a matrix $Q \triangleq I - \alpha c(I-W)$ and a triple $P \triangleq (\frac{1}{2c}Q, \frac{\alpha}{2}I, \frac{1}{2\alpha}I)$ of matrices. Further, define $\|\mathbf{u}^k - \mathbf{u}^*\|_P^2 = \|\mathbf{x}^k - \mathbf{x}^*\|_{\frac{1}{2c}Q}^2 + \|\mathbf{z}^k - \mathbf{z}^*\|_{\frac{\alpha}{2}I}^2 + \|\Pi^k - \Pi^*\|_{\frac{1}{2\alpha}I}^2$ as the squared distance from $\mathbf{u}^k$ to $\mathbf{u}^*$. In the following lemma, we show that $\|\mathbf{u}^k - \mathbf{u}^*\|_P^2$ decreases sufficiently fast when the parameters are properly chosen.

Lemma 1. *Under Assumptions 1 and 3–4, if the parameters α and c are chosen such that $\frac{1}{2c}\big[I - \alpha c(I-W)\big] - \frac{L_f}{2}I \succ 0$, then it holds for all $k \geq 1$ that*

$$\|\mathbf{u}^k - \mathbf{u}^*\|_P^2 - \|\mathbf{u}^{k+1} - \mathbf{u}^*\|_P^2 \geq \beta\|\mathbf{u}^k - \mathbf{u}^{k+1}\|_P^2, \tag{15}$$

where $\beta > 0$ is a positive constant.

Proof. See Appendix A. □

With Lemma 1, the sequence $\{\|\mathbf{u}^k - \mathbf{u}^*\|_P^2\}$ converges to 0. The proof is the same as that of Theorem 3.2 in [36] and is omitted here. Since $Q = I - \alpha c(I-W) \succ 0$ given that $\frac{1}{2c}\big[I - \alpha c(I-W)\big] - \frac{L_f}{2} I \succ 0$, we conclude that $\{\mathbf{u}^k\}$ converges to $\mathbf{u}^*$. We summarize the results as follows.

Theorem 1. *Under Assumptions 1 and 3–4, if the parameters are chosen as in Lemma 1, then the sequence* $\{\mathbf{u}^k\}$ *generated by* (6)–(8) *from any initial point* $\mathbf{u}^0$ *converges to the optimal solution* $\mathbf{u}^*$ *to the penalized problem* (5)*.*

Theorem 1 establishes the convergence of PAD to an optimal solution to the penalized problem (5) (or equivalently (3)) *for any* ϵ. Therefore, we can choose a sufficiently small ϵ so that the approximation error between (2) and (3) is negligible. Theorem 1 also provides guidelines for setting the parameters c and α. The condition $\frac{1}{2c}\big[I - \alpha c(I-W)\big] - \frac{L_f}{2} I \succ 0$ implies that $c < \frac{1}{L_f + \alpha \lambda_{\max}(I-W)}$, where $\lambda_{\max}(I-W)$ is the largest eigenvalue of $I - W$. Therefore, Theorem 1 implies that PAD has guaranteed convergence when c and α are small enough.

B. Sublinear Convergence Rate under General Convexity

This subsection establishes the $O(\frac{1}{k})$ convergence rate of PAD in an ergodic sense under general convexity. Similar to the proof in [18], we define $\nu \triangleq (\mathbf{x}, \mathbf{z})$ and $\Phi(\nu) \triangleq f(\mathbf{x}) + g(\mathbf{x}) + \frac{1}{2\epsilon}\|\mathbf{z}\|_{\mathcal{F}}^2$.

Theorem 2. *Under Assumptions 1 and 3–4, let* $\{\mathbf{u}^k\} = \{(\mathbf{x}^k, \mathbf{z}^k, \Pi^k)\}$ *be the sequence generated by* (6)–(8)*. Define* $\tilde{\nu}^k \triangleq (\tilde{\mathbf{x}}^k, \tilde{\mathbf{z}}^k)$ *with* $\tilde{\mathbf{x}}^k \triangleq \frac{1}{k}\sum_{t=1}^k \mathbf{x}^t$ *and* $\tilde{\mathbf{z}}^k \triangleq \frac{1}{k}\sum_{t=1}^k \mathbf{z}^t$ *for* $k \geq 1$*. If the parameters* c *and* α *are chosen as in Lemma 1, then it holds for all* $k \geq 1$ *that*

$$\max\left\{|\Phi(\tilde{\nu}^k) - \Phi(\nu^*)|, \|\Pi^*\|_{\mathcal{F}}\|(I-W)^{\frac{1}{2}}\tilde{\mathbf{x}}^k - \tilde{\mathbf{z}}^k\|_{\mathcal{F}}\right\} \leq \frac{C_1}{k},$$

where $C_1 \triangleq \frac{1}{2c}\|\mathbf{x}^* - \mathbf{x}^0\|_Q^2 + \frac{\alpha}{2}\|\mathbf{z}^* - \mathbf{z}^0\|_{\mathcal{F}}^2 + \frac{4}{\alpha}\|\Pi^*\|_{\mathcal{F}}^2 + \frac{1}{\alpha}\|\Pi^0\|_{\mathcal{F}}^2$ *is a constant.*

Proof. See Appendix B. □

Theorem 2 further establishes the sublinear convergence rate of PAD under the same setting as that in Lemma 1 and Theorem 1. If the parameters α and c satisfy $c < \frac{1}{L_f + \alpha\lambda_{\max}(I-W)}$, both the optimality gap in terms of function value $|\Phi(\tilde{\nu}^k) - \Phi(\nu^*)|$ and the constraint violation $\|(I-W)^{\frac{1}{2}}\tilde{\mathbf{x}}^k - \tilde{\mathbf{z}}^k\|_{\mathcal{F}}$ converge sublinearly in an ergodic sense. This result holds for any $\epsilon > 0$ so that we can set ϵ sufficiently small to reach high accuracy. The analysis in the general convex case can be extended to $\epsilon = 0$ by modifying $\Phi(\nu) \triangleq f(\mathbf{x}) + g(\mathbf{x})$ and $C_1 \triangleq \frac{1}{2c}\|\mathbf{x}^* - \mathbf{x}^0\|_Q^2 + \frac{4}{\alpha}\|\Pi^*\|_{\mathcal{F}}^2 + \frac{1}{\alpha}\|\Pi^0\|_{\mathcal{F}}^2$.

C. Linear Convergence Rate under Strong Convexity

The following theorem shows that PAD is linearly convergent when the local functions are further assumed to be strongly convex.

Theorem 3. *Under Assumptions 1–4, suppose that* $\delta \triangleq 1 - \frac{L_f^2 c}{\lambda_{\min}(Q)\mu_f} > 0$ *and the variables are initialized as* $\frac{1}{\epsilon}\mathbf{z}^0 = \Pi^0$*. Then, the sequence* $\{(\mathbf{x}^k, \mathbf{z}^k)\}$ *generated by* (6)–(8) *converges to* $(\mathbf{x}^*, \mathbf{z}^*)$ *with*

$$C_2\|\mathbf{x}^* - \mathbf{x}^{k+1}\|_Q^2 + \|\mathbf{z}^* - \mathbf{z}^{k+1}\|_{\mathcal{F}}^2 \leq \eta\left(C_2\|\mathbf{x}^* - \mathbf{x}^k\|_Q^2 + \|\mathbf{z}^* - \mathbf{z}^k\|_{\mathcal{F}}^2\right),$$

where

$$\eta \triangleq \max\left\{\frac{1}{c_1 c + 1}, \frac{\alpha^2\epsilon^2 + 1}{\alpha\epsilon + \alpha^2\epsilon^2 + 1}\right\} < 1$$

and

$$c_1 = \frac{\mu_f \delta}{(1+\delta)\lambda_{\max}(Q)}, \quad C_2 = \frac{c_1 + \frac{1}{c}}{\frac{1}{\epsilon} + \alpha + \frac{1}{\alpha\epsilon^2}}.$$

Proof. See Appendix C. □

Theorem 3 requires $\delta \triangleq 1 - \frac{L_f^2 c}{\lambda_{\min}(Q)\mu_f} > 0$, which implies that $c < \frac{1}{L_f^2/\mu_f + \alpha\lambda_{\max}(I-W)}$. Since $\frac{L_f}{\mu_f} > 1$, this bound on the step size is smaller than the one required in Theorem 1 and 2. With this step size rule, Theorem 3 establishes the Q-linear convergence of the sequence $\{C_2\|\mathbf{x}^* - \mathbf{x}^k\|_Q^2 + \|\mathbf{z}^* - \mathbf{z}^k\|_{\mathcal{F}}^2\}$ to 0, from which we know that the sequence $\{\mathbf{x}^k\}$ converges to $\mathbf{x}^*$ and $\{\mathbf{z}^k\}$ converges to $\mathbf{z}^*$, both R-linearly. The analysis of linear rate under the strongly convex case no longer holds if $\epsilon = 0$. Otherwise, the result will contradict with that in [27]. With $\epsilon = 0$, the inequality (53) becomes $\frac{1}{c}\|\mathbf{x}^* - \mathbf{x}^k\|_Q^2 + \frac{1}{\alpha}\|\Pi^* - \Pi^k\|_{\mathcal{F}}^2 \geq \frac{\mu_f\delta}{1+\delta}\|\mathbf{x}^* - \mathbf{x}^{k+1}\|_{\mathcal{F}}^2 + \frac{1}{c}\|\mathbf{x}^* - \mathbf{x}^{k+1}\|_Q^2 + \frac{1}{\alpha}\|\Pi^* - \Pi^{k+1}\|_{\mathcal{F}}^2$. Since the coefficients of $\|\Pi^* - \Pi^k\|_{\mathcal{F}}^2$ and $\|\Pi^* - \Pi^{k+1}\|_{\mathcal{F}}^2$ are equal, we are unable to establish the linear rate.

Note that $\eta \to 1$ as $\epsilon \to 0$, meaning that the convergence of the sequence $\{C_2\|\mathbf{x}^* - \mathbf{x}^k\|_Q^2 + \|\mathbf{z}^* - \mathbf{z}^k\|_{\mathcal{F}}^2\}$ can be slow when ϵ is small. However, in the numerical experiments we will show that $\|\mathbf{x}^k - \hat{\mathbf{x}}^*\|_{\mathcal{F}}$ (and $\|\mathbf{x}^k - \mathbf{x}^*\|_{\mathcal{F}}$ too) itself converges fast under different ϵ until reaching a small neighborhood of $\hat{\mathbf{x}}^*$. This does not contradict with Theorem 3 since $C_2 = O(\epsilon^2)$.

D. Bound on $\|\mathbf{x}^* - \hat{\mathbf{x}}^*\|_{\mathcal{F}}$ *under Strong Convexity*

Observe that PAD converges to $\mathbf{x}^*$, the optimal solution to the penalized problem (3), but not $\hat{\mathbf{x}}^*$, the optimal solution to the original problem (2). This subsection bounds the gap between $\mathbf{x}^*$ and $\hat{\mathbf{x}}^*$ under the assumption of strong convexity. For simplicity, we define $\Psi(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$ and $\mathrm{null}(I - W) = \{\mathbf{x} \in \mathbb{R}^{n\times p} : (I-W)\mathbf{x} = \mathbf{0}_{n\times p}\}$ as the null space of $I - W$. We consider the orthogonal decomposition of $\mathbf{x}^*$ w.r.t. $\mathrm{null}(I-W)$ and its orthogonal complement $\mathrm{null}(I-W)^{\perp}$; i.e.,

$$\mathbf{x}^* = \mathbf{x}_1^* + \mathbf{x}_2^*, \tag{16}$$

where $\mathbf{x}_1^* \in \mathrm{null}(I-W)$ and $\mathbf{x}_2^* \in \mathrm{null}(I-W)^{\perp}$. We begin with the following proposition.

Proposition 1. *Under Assumptions 2–4, suppose that each* g_i *is a continuous function. Then, there exists a constant* $L_\Psi > 0$*, which is independent of* ϵ*, such that*

$$|\Psi(\mathbf{x}_1^*) - \Psi(\mathbf{x}^*)| \leq L_\Psi\|\mathbf{x}_1^* - \mathbf{x}^*\|_{\mathcal{F}}. \tag{17}$$

Proof. See Appendix D. □

Based on Proposition 1, the bound on $\|\mathbf{x}^* - \hat{\mathbf{x}}^*\|_{\mathcal{F}}$ is established as follows.

Theorem 4. *Under the same assumptions as in Proposition 1, there exists a constant $L_\Psi > 0$, which is independent of ϵ, such that the distance between the optimal solution $\mathbf{x}^*$ to the penalized problem* (3) *and the optimal solution $\hat{\mathbf{x}}^*$ to the constrained problem* (2) *is bounded by*

$$\|\mathbf{x}^* - \hat{\mathbf{x}}^*\|_{\mathcal{F}} \leq \frac{2L_\Psi \epsilon}{\tilde{\lambda}_{\min}^2} + \frac{2L_\Psi \epsilon^{\frac{1}{2}}}{\tilde{\lambda}_{\min}\mu_f^{\frac{1}{2}}}, \tag{18}$$

where $\tilde{\lambda}_{\min}$ is the smallest nonzero eigenvalue of $(I-W)^{\frac{1}{2}}$. In particular, if the network is not too poorly connected such that $\tilde{\lambda}_{\min} \gg \epsilon^{\frac{1}{2}}$, then we have

$$\|\mathbf{x}^* - \hat{\mathbf{x}}^*\|_{\mathcal{F}} = O(\epsilon^{\frac{1}{2}}). \tag{19}$$

Proof. See Appendix E. □

Theorem 4 implies that $\mathbf{x}^* \to \hat{\mathbf{x}}^*$ as $\epsilon \to 0$. To achieve a high accuracy, we can set ϵ to be sufficiently small. However, as we have mentioned in Subsection III-C, the analysis in Theorem 3 in the strongly convex case cannot be extended to the case of $\epsilon = 0$.

For the smooth case where $g_i = 0$, it is known [7], [37] that $\|\mathbf{x}^* - \hat{\mathbf{x}}^*\|_{\mathcal{F}} = O(\epsilon)$ when each f_i has a Lipschitz continuous gradient (i.e., Assumption 1 in our paper). Our result in Theorem 4 cannot recover that for the smooth case since we do not use Assumption 1 here.

IV. Numerical Experiments

In this section, we conduct numerical experiments — decentralized sparse logistic regression (Subsection IV-A) and decentralized quadratic programming (Subsection IV-B) — to demonstrate the effectiveness of PAD. Both problems satisfy Assumptions 1–4. The experiments are run over a randomly generated connected network with n agents and $\frac{\tau n(n-1)}{2}$ undirected edges, where $\tau \in (0,1]$ is the connectivity ratio. We compare PAD with several state-of-the-art algorithms—Proximal DGD [29], DL-ADMM [17], PG-EXTRA [20], and NIDS [22]. The mixing matrix W is generated with the Metropolis-Hastings rule. The performance is evaluated by the relative errors $\frac{\|\mathbf{x}^k - \hat{\mathbf{x}}^*\|_{\mathcal{F}}}{\|\mathbf{x}^0 - \hat{\mathbf{x}}^*\|_{\mathcal{F}}}$ and $\frac{C_2\|\mathbf{x}^k - \mathbf{x}^*\|_Q^2 + \|\mathbf{z}^k - \mathbf{z}^*\|_{\mathcal{F}}^2}{C_2\|\mathbf{x}^0 - \mathbf{x}^*\|_Q^2 + \|\mathbf{z}^0 - \mathbf{z}^*\|_{\mathcal{F}}^2}$, where the optimal solution $\hat{\mathbf{x}}^*$ to (2) and the optimal primal-dual solution $(\mathbf{x}^*, \mathbf{z}^*)$ to (5) are computed in advance. Note that all the decentralized algorithms numerically evaluated in this section have the same communication cost per iteration. Consequently, the amount of information exchange over the network is proportional to their numbers of iterations. We conduct the experiments with Matlab R2016b, running on a laptop with Intel(R) Core(TM) i7 CPU@1.80GHz, 16.0 GB RAM, and Windows 10 operating system.

A. Decentralized Sparse Logistic Regression

In this subsection, we consider the decentralized sparse logistic regression problem. Each agent i holds local training data $\left(M_{(i)j}, y_{(i)j}\right) \in \mathbb{R}^p \times \{-1, +1\}$, $j = 1, \ldots, m_i$, where $M_{(i)j}$ is the jth feature vector of agent i and $y_{(i)j}$ is the corresponding binary label. The entries of each $M_{(i)j}$ (except for the last one, which is set to 1) were generated according to the standard normal distribution. Each label $y_{(i)j}$ is generated according to the uniform distribution. The problem is $\min_x \frac{1}{n}\sum_{i=1}^n \left\{ \sum_{j=1}^{m_i} \ln\left(1+\exp(-(M_{(i)j}^T x) y_{(i)j})\right) + \frac{\zeta}{2}\|x\|^2 + \xi\|x\|_1 \right\}$, where $\zeta = 10^{-3}$ and $\xi = 10^{-3}$. We set $n = 30$, $\tau = 0.5$, $p = 10$, and $m_i = 5, \forall i$.

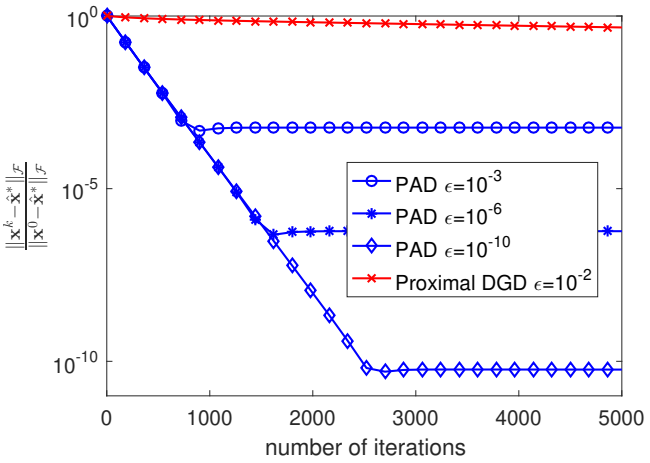

Fig. 1. $\frac{\|\mathbf{x}^k - \hat{\mathbf{x}}^*\|_{\mathcal{F}}}{\|\mathbf{x}^0 - \hat{\mathbf{x}}^*\|_{\mathcal{F}}}$ of proximal DGD (with $\epsilon = 10^{-2}$) and PAD (with $\epsilon = 10^{-3}$, $\epsilon = 10^{-6}$, and $\epsilon = 10^{-10}$, respectively) in the decentralized sparse logistic regression problem. For proximal DGD, its step size is $\gamma = \epsilon$. For PAD, $\alpha = 0.05$ and $c = 0.7$ are hand-tuned to get the best results.

As shown in Fig. 1, PAD and proximal DGD with a fixed step size both converge to a neighborhood of $\hat{\mathbf{x}}^*$. However, the step size of proximal DGD must be in the same order of ϵ. When ϵ is small, the convergence speed of DGD is very slow. For this reason we terminate DGD early and only show the first 5,000 iterations. PAD converges with fast speed under different values of ϵ. Smaller ϵ brings higher accuracy.

B. Decentralized Quadratic Programming

Each agent i has a local quadratic function $f_i(x) = \frac{1}{2}x^T Q_i x + h_i^T x$ and a local linear constraint $a_i^T x \leq b_i$, where $Q_i \in \mathbb{R}^{p\times p} \succ 0$, $h_i \in \mathbb{R}^p$, $a_i \in \mathbb{R}^p$, and $b_i \in \mathbb{R}$. The quadratic programming problem to solve takes the form $\min_x \sum_{i=1}^n \left(\frac{1}{2}x^T Q_i x + h_i^T x\right)$ s.t. $a_i^T x \leq b_i, \forall i = 1, \ldots, n$. It can be reformulated to the unconstrained form (1) by defining the nonsmooth function g_i as

$$g_i(x) = \begin{cases} 0 & \text{if } a_i^T x \leq b_i, \\ +\infty & \text{otherwise.} \end{cases}$$

We set $n = 10$, $\tau = 0.4$, and $p = 50$. The matrix Q_i is generated such that the Lipschitz and strong convexity constants of f_i are $L_f = 1$ and $\mu_f = 0.5$, respectively.

Fig. 2 shows that PAD has fast convergence. PAD needs less than 250 iterations to attain a relative error of 10^{-13}, while the other algorithms need more iterations to attain the same accuracy. PAD converges linearly to a neighborhood of the optimal solution, which is consistent with Theorem 3. PAD allows the penalty parameter ϵ to be sufficiently small without sacrificing the speed.

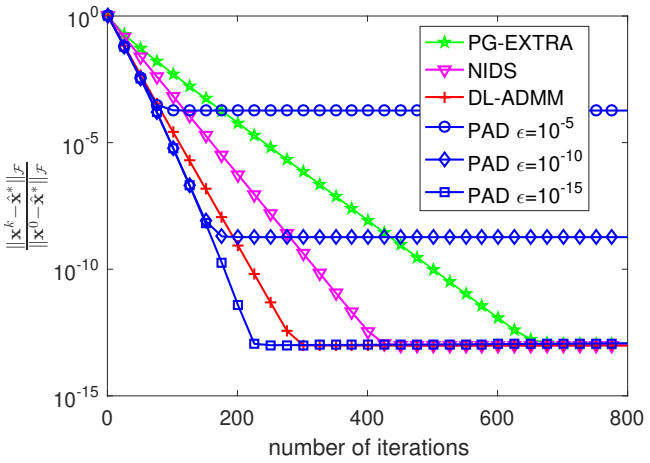

Fig. 2. $\frac{\|\mathbf{x}^k-\hat{\mathbf{x}}^*\|_{\mathcal{F}}}{\|\mathbf{x}^0-\hat{\mathbf{x}}^*\|_{\mathcal{F}}}$ of PG-EXTRA, NIDS, DL-ADMM, and PAD in the decentralized quadratic programming problem. The parameters of all the algorithms were hand-tuned to get the best results. For PAD, $\alpha = 3.18$ and $c = 0.3$ (with $\epsilon = 10^{-5}$, $\epsilon = 10^{-10}$, and $\epsilon = 10^{-15}$). For PG-EXTRA, the step size is $c = 0.54\lambda_{\min}((I+W)/2)/L_f$. For NIDS, $c_i = 0.27/L_i$ and $\alpha = 1/((1-\lambda_{\min}(W))\max_i c_i)$. For DL-ADMM, the step size of primal update is $c = 10$ and the step size of dual update is $\alpha = 0.4$.

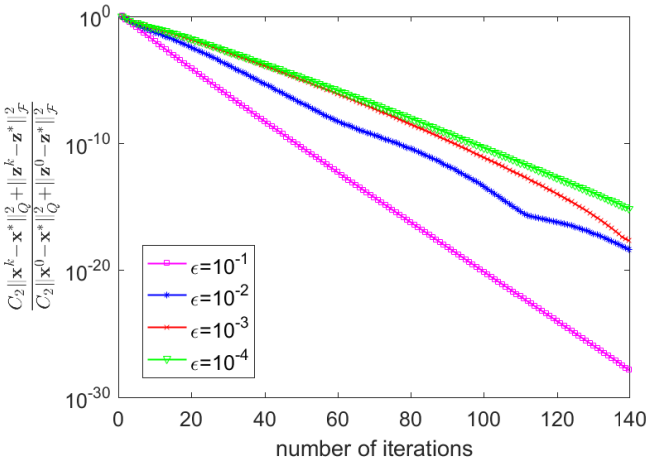

Fig. 3. $\frac{C_2\|\mathbf{x}^k-\mathbf{x}^*\|_Q^2+\|\mathbf{z}^k-\mathbf{z}^*\|_{\mathcal{F}}^2}{C_2\|\mathbf{x}^0-\mathbf{x}^*\|_Q^2+\|\mathbf{z}^0-\mathbf{z}^*\|_{\mathcal{F}}^2}$ of PAD in the decentralized quadratic programming problem. We set $\alpha = 3.18$ and $c = 0.3$ (with $\epsilon = 10^{-1}$, $\epsilon = 10^{-2}$, $\epsilon = 10^{-3}$, and $\epsilon = 10^{-4}$).

Fig. 3 records $\frac{C_2\|\mathbf{x}^k-\mathbf{x}^*\|_Q^2+\|\mathbf{z}^k-\mathbf{z}^*\|_{\mathcal{F}}^2}{C_2\|\mathbf{x}^0-\mathbf{x}^*\|_Q^2+\|\mathbf{z}^0-\mathbf{z}^*\|_{\mathcal{F}}^2}$ with the number of iterations, where $C_2 = O(\epsilon^2)$ is defined in Theorem 3. We can see that the linear convergence rate slows down when ϵ becomes smaller, which coincides with the result of Theorem 3.

V. Conclusion

In this paper, we proposed a consensus-based decentralized algorithm called PAD to solve the penalized approximation of the decentralized convex composite problem. Unlike existing penalized methods, the proposed PAD is fast even with a very small penalty parameter. Our numerical results show that PAD allows the use of a small penalty parameter without sacrificing the convergence speed. The proposed PAD is proven to be sublinearly convergent when the private functions are convex and linearly convergent when in addition the smooth parts are strongly convex. Our numerical results further demonstrated the advantages of PAD over existing state-of-the-art algorithms such as DL-ADMM, PG-EXTRA, and NIDS.

Appendix

A. Proof of Lemma 1

Proof. The optimality condition of (6) is

$$\mathbf{x}^k - c\Big[\nabla f(\mathbf{x}^k) + \alpha(I-W)^{\frac{1}{2}}\big((I-W)^{\frac{1}{2}}\mathbf{x}^k - \mathbf{z}^k + \frac{\Pi^k}{\alpha}\big)\Big] - \mathbf{x}^{k+1} \in c\partial g(\mathbf{x}^{k+1}). \tag{20}$$

By the definition of ∂g, we have

$$\Big\langle \mathbf{x}^* - \mathbf{x}^{k+1}, \mathbf{x}^k - \mathbf{x}^{k+1} - c\Big[\nabla f(\mathbf{x}^k) + \alpha(I-W)^{\frac{1}{2}}\Big((I-W)^{\frac{1}{2}}\mathbf{x}^k - \mathbf{z}^k + \frac{\Pi^k}{\alpha}\Big)\Big]\Big\rangle \le cg(\mathbf{x}^*) - cg(\mathbf{x}^{k+1}). \tag{21}$$

Substituting $\Pi^k = \Pi^{k+1} - \alpha\big[(I-W)^{\frac{1}{2}}\mathbf{x}^{k+1} - \mathbf{z}^{k+1}\big]$ into (21), we obtain

$$\begin{aligned} & g(\mathbf{x}^*) - g(\mathbf{x}^{k+1}) \\ \ge & \Big\langle \mathbf{x}^* - \mathbf{x}^{k+1}, \frac{1}{c}(\mathbf{x}^k - \mathbf{x}^{k+1}) \\ & - \big[\nabla f(\mathbf{x}^k) + \alpha(I-W)^{\frac{1}{2}}\big((I-W)^{\frac{1}{2}}(\mathbf{x}^k - \mathbf{x}^{k+1}) \\ & - (\mathbf{z}^k - \mathbf{z}^{k+1})\big) + (I-W)^{\frac{1}{2}}\Pi^{k+1}\big]\Big\rangle \\ = & \Big\langle \mathbf{x}^* - \mathbf{x}^{k+1}, \frac{1}{c}\underbrace{[I - \alpha c(I-W)]}_{\triangleq Q}(\mathbf{x}^k - \mathbf{x}^{k+1})\Big\rangle \\ & - \big\langle \mathbf{x}^* - \mathbf{x}^{k+1}, \nabla f(\mathbf{x}^k) + (I-W)^{\frac{1}{2}}\Pi^{k+1} \\ & - \alpha(I-W)^{\frac{1}{2}}(\mathbf{z}^k - \mathbf{z}^{k+1})\big\rangle. \end{aligned} \tag{22}$$

Using the identity $(v_2-v_1)^TQ(v_3-v_1) = \frac{1}{2}(\|v_2-v_1\|_Q^2 + \|v_3-v_1\|_Q^2 - \|v_2-v_3\|_Q^2)$ to rewrite the first term on the right-hand side of (22), we have

$$\begin{aligned} & g(\mathbf{x}^*) - g(\mathbf{x}^{k+1}) \\ \ge & \frac{1}{2c}\big(\|\mathbf{x}^* - \mathbf{x}^{k+1}\|_Q^2 + \|\mathbf{x}^k - \mathbf{x}^{k+1}\|_Q^2 - \|\mathbf{x}^* - \mathbf{x}^k\|_Q^2\big) \\ & \underbrace{-\big\langle \mathbf{x}^* - \mathbf{x}^{k+1}, \nabla f(\mathbf{x}^k)\big\rangle}_{(i)} - \big\langle \mathbf{x}^* - \mathbf{x}^{k+1}, (I-W)^{\frac{1}{2}}\Pi^{k+1}\big\rangle \\ & + \alpha\underbrace{\big\langle (I-W)^{\frac{1}{2}}(\mathbf{x}^* - \mathbf{x}^{k+1}), \mathbf{z}^k - \mathbf{z}^{k+1}\big\rangle}_{(ii)}. \end{aligned} \tag{23}$$

For the term (i) in (23), by Assumption 1, we have

$$\begin{aligned} & \big\langle \mathbf{x}^* - \mathbf{x}^{k+1}, \nabla f(\mathbf{x}^k)\big\rangle \\ = & \big\langle \mathbf{x}^* - \mathbf{x}^k, \nabla f(\mathbf{x}^k)\big\rangle + \big\langle \mathbf{x}^k - \mathbf{x}^{k+1}, \nabla f(\mathbf{x}^k)\big\rangle \\ \le & f(\mathbf{x}^*) - f(\mathbf{x}^k) + f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) + \frac{L_f}{2}\|\mathbf{x}^k - \mathbf{x}^{k+1}\|_{\mathcal{F}}^2 \\ = & f(\mathbf{x}^*) - f(\mathbf{x}^{k+1}) + \frac{L_f}{2}\|\mathbf{x}^k - \mathbf{x}^{k+1}\|_{\mathcal{F}}^2. \end{aligned} \tag{24}$$

For the term (ii) in (23), using the identity $(v_1-v_2)^T(v_3-v_4)=\frac{1}{2}(\|v_1-v_4\|^2-\|v_1-v_3\|^2)+\frac{1}{2}(\|v_2-v_3\|^2-\|v_2-v_4\|^2)$, we have

$$\begin{aligned}&\left\langle (I-W)^{\frac{1}{2}}(\mathbf{x}^*-\mathbf{x}^{k+1}),\mathbf{z}^k-\mathbf{z}^{k+1}\right\rangle \\ =&\frac{1}{2}\big(\|(I-W)^{\frac{1}{2}}\mathbf{x}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2-\|(I-W)^{\frac{1}{2}}\mathbf{x}^*-\mathbf{z}^k\|_{\mathcal{F}}^2\big)\\ +&\frac{1}{2}\big(\|(I-W)^{\frac{1}{2}}\mathbf{x}^{k+1}-\mathbf{z}^k\|_{\mathcal{F}}^2-\|(I-W)^{\frac{1}{2}}\mathbf{x}^{k+1}-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2\big).\end{aligned} \tag{25}$$

By substituting $(I-W)^{\frac{1}{2}}\mathbf{x}^{k+1}-\mathbf{z}^{k+1}=\frac{\Pi^{k+1}-\Pi^k}{\alpha}$ obtained from (8), the equation (25) becomes

$$\begin{aligned}&\left\langle (I-W)^{\frac{1}{2}}(\mathbf{x}^*-\mathbf{x}^{k+1}),(\mathbf{z}^k-\mathbf{z}^{k+1})\right\rangle \\ =&\frac{1}{2}\big(\|(I-W)^{\frac{1}{2}}\mathbf{x}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2-\|(I-W)^{\frac{1}{2}}\mathbf{x}^*-\mathbf{z}^k\|_{\mathcal{F}}^2\big)\\ &+\frac{1}{2}\left(\left\|\frac{\Pi^{k+1}-\Pi^k}{\alpha}+\mathbf{z}^{k+1}-\mathbf{z}^k\right\|_{\mathcal{F}}^2-\left\|\frac{\Pi^{k+1}-\Pi^k}{\alpha}\right\|_{\mathcal{F}}^2\right)\\ =&\frac{1}{2}\big(\|(I-W)^{\frac{1}{2}}\mathbf{x}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2-\|(I-W)^{\frac{1}{2}}\mathbf{x}^*-\mathbf{z}^k\|_{\mathcal{F}}^2\big)\\ &+\frac{1}{2}\left(2\left\langle\frac{\Pi^{k+1}-\Pi^k}{\alpha},\mathbf{z}^{k+1}-\mathbf{z}^k\right\rangle+\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_{\mathcal{F}}^2\right).\end{aligned} \tag{26}$$

Substituting (24) and (26) into (23), followed by substituting (23) into (22), we get

$$\begin{aligned}&g(\mathbf{x}^*)-g(\mathbf{x}^{k+1})\\ \geq&\frac{1}{2c}\big(\|\mathbf{x}^*-\mathbf{x}^{k+1}\|_Q^2+\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_Q^2-\|\mathbf{x}^*-\mathbf{x}^k\|_Q^2\big)\\ &-f(\mathbf{x}^*)+f(\mathbf{x}^{k+1})-\frac{L_f}{2}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathcal{F}}^2\\ &-\left\langle\mathbf{x}^*-\mathbf{x}^{k+1},(I-W)^{\frac{1}{2}}\Pi^{k+1}\right\rangle\\ &+\frac{\alpha}{2}\big(\|(I-W)^{\frac{1}{2}}\mathbf{x}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2-\|(I-W)^{\frac{1}{2}}\mathbf{x}^*-\mathbf{z}^k\|_{\mathcal{F}}^2\big)\\ &+\frac{\alpha}{2}\left(2\left\langle\frac{\Pi^{k+1}-\Pi^k}{\alpha},\mathbf{z}^{k+1}-\mathbf{z}^k\right\rangle+\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_{\mathcal{F}}^2\right).\end{aligned} \tag{27}$$

Further substituting (10) into (27) and rearranging the terms, we have

$$\begin{aligned}&g(\mathbf{x}^*)+f(\mathbf{x}^*)-g(\mathbf{x}^{k+1})-f(\mathbf{x}^{k+1})\\ &+\left\langle\mathbf{x}^*-\mathbf{x}^{k+1},(I-W)^{\frac{1}{2}}\Pi^{k+1}\right\rangle\\ \geq&\frac{1}{2c}\big(\|\mathbf{x}^*-\mathbf{x}^{k+1}\|_Q^2+\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_Q^2-\|\mathbf{x}^*-\mathbf{x}^k\|_Q^2\big)\\ &+\frac{\alpha}{2}\big(\|(I-W)^{\frac{1}{2}}\mathbf{x}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2-\|(I-W)^{\frac{1}{2}}\mathbf{x}^*-\mathbf{z}^k\|_{\mathcal{F}}^2\big)\\ &-\frac{L_f}{2}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathcal{F}}^2+\left(\frac{\alpha}{2}+\frac{1}{\epsilon}\right)\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_{\mathcal{F}}^2.\end{aligned} \tag{28}$$

Now, by direct calculation, we have

$$\begin{aligned}&\frac{1}{2\epsilon}\|\mathbf{z}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2+\left\langle\mathbf{z}^*-\mathbf{z}^{k+1},\frac{1}{\epsilon}\mathbf{z}^{k+1}\right\rangle\\ =&\frac{1}{2\epsilon}\|\mathbf{z}^*\|_{\mathcal{F}}^2-\frac{1}{2\epsilon}\|\mathbf{z}^{k+1}\|_{\mathcal{F}}^2.\end{aligned} \tag{29}$$

Substituting (10) into (29), we get

$$\begin{aligned}&\frac{1}{2\epsilon}\|\mathbf{z}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2+\left\langle\mathbf{z}^*-\mathbf{z}^{k+1},\Pi^{k+1}\right\rangle\\ =&\frac{1}{2\epsilon}\|\mathbf{z}^*\|_{\mathcal{F}}^2-\frac{1}{2\epsilon}\|\mathbf{z}^{k+1}\|_{\mathcal{F}}^2.\end{aligned} \tag{30}$$

Substituting (30) into (28) then yields

$$\begin{aligned}&\underbrace{g(\mathbf{x}^*)+f(\mathbf{x}^*)+\frac{1}{2\epsilon}\|\mathbf{z}^*\|_{\mathcal{F}}^2-g(\mathbf{x}^{k+1})-f(\mathbf{x}^{k+1})-\frac{1}{2\epsilon}\|\mathbf{z}^{k+1}\|_{\mathcal{F}}^2}_{(iii)}\\ &+\underbrace{\left\langle\mathbf{x}^*-\mathbf{x}^{k+1},(I-W)^{\frac{1}{2}}\Pi^{k+1}\right\rangle-\left\langle\mathbf{z}^*-\mathbf{z}^{k+1},\Pi^{k+1}\right\rangle}_{(iv)}\\ &-\frac{1}{2c}\big(\|\mathbf{x}^*-\mathbf{x}^{k+1}\|_Q^2-\|\mathbf{x}^*-\mathbf{x}^k\|_Q^2\big)\\ &-\frac{\alpha}{2}\big(\|(I-W)^{\frac{1}{2}}\mathbf{x}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2-\|(I-W)^{\frac{1}{2}}\mathbf{x}^*-\mathbf{z}^k\|_{\mathcal{F}}^2\big)\\ \geq&\frac{1}{2c}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_Q^2-\frac{L_f}{2}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathcal{F}}^2\\ &+\left(\frac{\alpha}{2}+\frac{1}{\epsilon}\right)\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_{\mathcal{F}}^2+\frac{1}{2\epsilon}\|\mathbf{z}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2.\end{aligned} \tag{31}$$

Next, we proceed to simplify (31). By the update of Π^{k+1} in (8), we have

$$\begin{aligned}0=&\left\langle\Pi-\Pi^{k+1},\frac{\Pi^k-\Pi^{k+1}}{\alpha}\right\rangle-\left\langle\Pi-\Pi^{k+1},\frac{\Pi^k-\Pi^{k+1}}{\alpha}\right\rangle\\ =&\left\langle\Pi-\Pi^{k+1},-(I-W)^{\frac{1}{2}}\mathbf{x}^{k+1}+\mathbf{z}^{k+1}\right\rangle\\ &\underbrace{-\left\langle\Pi-\Pi^{k+1},\frac{\Pi^k-\Pi^{k+1}}{\alpha}\right\rangle}_{(v)},\forall\Pi.\end{aligned} \tag{32}$$

On the other hand, adding $\left\langle\Pi,-(I-W)^{\frac{1}{2}}\mathbf{x}^{k+1}+\mathbf{z}^{k+1}\right\rangle$ to the term (iii) in (31) gives

$$\begin{aligned}&g(\mathbf{x}^*)+f(\mathbf{x}^*)+\frac{1}{2\epsilon}\|\mathbf{z}^*\|_{\mathcal{F}}^2-g(\mathbf{x}^{k+1})-f(\mathbf{x}^{k+1})\\ &-\frac{1}{2\epsilon}\|\mathbf{z}^{k+1}\|_{\mathcal{F}}^2+\left\langle\Pi,-(I-W)^{\frac{1}{2}}\mathbf{x}^{k+1}+\mathbf{z}^{k+1}\right\rangle\\ =&L(\mathbf{x}^*,\mathbf{z}^*,\Pi)-L(\mathbf{x}^{k+1},\mathbf{z}^{k+1},\Pi),\forall\Pi.\end{aligned} \tag{33}$$

To the term (iv) in (31) we add $\left\langle\Pi^{k+1},(I-W)^{\frac{1}{2}}\mathbf{x}^{k+1}-\mathbf{z}^{k+1}\right\rangle$ and use $(I-W)^{\frac{1}{2}}\mathbf{x}^*=\mathbf{z}^*$ to get

$$\begin{aligned}&\left\langle\mathbf{x}^*-\mathbf{x}^{k+1},(I-W)^{\frac{1}{2}}\Pi^{k+1}\right\rangle-\left\langle\mathbf{z}^*-\mathbf{z}^{k+1},\Pi^{k+1}\right\rangle\\ &+\left\langle\Pi^{k+1},(I-W)^{\frac{1}{2}}\mathbf{x}^{k+1}-\mathbf{z}^{k+1}\right\rangle=0.\end{aligned} \tag{34}$$

For the term (v) in (32), adding $\frac{1}{2\alpha}\|\Pi^{k+1}-\Pi^k\|_{\mathcal{F}}^2$, we have

$$\begin{aligned}&\frac{1}{2\alpha}\|\Pi^{k+1}-\Pi^k\|_{\mathcal{F}}^2-\left\langle\Pi-\Pi^{k+1},\frac{\Pi^k-\Pi^{k+1}}{\alpha}\right\rangle\\ =&\frac{1}{2\alpha}\|\Pi-\Pi^k\|_{\mathcal{F}}^2-\frac{1}{2\alpha}\|\Pi-\Pi^{k+1}\|_{\mathcal{F}}^2,\forall\Pi.\end{aligned} \tag{35}$$

Now, substituting (32)–(35) into (31) and using the fact that $(I-W)^{\frac{1}{2}}\mathbf{x}^*=\mathbf{z}^*$, we have

$$\begin{aligned}&g(\mathbf{x}^*)+f(\mathbf{x}^*)+\frac{1}{2\epsilon}\|\mathbf{z}^*\|_{\mathcal{F}}^2-g(\mathbf{x}^{k+1})-f(\mathbf{x}^{k+1})\\ &-\frac{1}{2\epsilon}\|\mathbf{z}^{k+1}\|_{\mathcal{F}}^2+\left\langle\Pi,-(I-W)^{\frac{1}{2}}\mathbf{x}^{k+1}+\mathbf{z}^{k+1}\right\rangle\\ &+\frac{1}{2c}\|\mathbf{x}^*-\mathbf{x}^k\|_Q^2-\frac{1}{2c}\|\mathbf{x}^*-\mathbf{x}^{k+1}\|_Q^2+\frac{\alpha}{2}\|\mathbf{z}^*-\mathbf{z}^k\|_{\mathcal{F}}^2\\ &-\frac{\alpha}{2}\|\mathbf{z}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2+\frac{1}{2\alpha}\|\Pi-\Pi^k\|_{\mathcal{F}}^2-\frac{1}{2\alpha}\|\Pi-\Pi^{k+1}\|_{\mathcal{F}}^2\\ \geq&\frac{1}{2c}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_Q^2-\frac{L_f}{2}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathcal{F}}^2+(\frac{\alpha}{2}+\frac{1}{\epsilon})\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_{\mathcal{F}}^2\\ &+\frac{1}{2\alpha}\|\Pi^{k+1}-\Pi^k\|_{\mathcal{F}}^2+\frac{1}{2\epsilon}\|\mathbf{z}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2,\forall\Pi.\end{aligned} \tag{36}$$

When $\Pi = \Pi^*$, (33) becomes

$$\begin{aligned} & g(\mathbf{x}^*) + f(\mathbf{x}^*) + \frac{1}{2\epsilon}\|\mathbf{z}^*\|_{\mathcal{F}}^2 - g(\mathbf{x}^{k+1}) - f(\mathbf{x}^{k+1}) \quad (37) \\ & - \frac{1}{2\epsilon}\|\mathbf{z}^{k+1}\|_{\mathcal{F}}^2 + \left\langle \Pi^*, -(I-W)^{\frac{1}{2}}\mathbf{x}^{k+1} + \mathbf{z}^{k+1}\right\rangle \\ =& L(\mathbf{x}^*, \mathbf{z}^*, \Pi^*) - L(\mathbf{x}^{k+1}, \mathbf{z}^{k+1}, \Pi^*) \leq 0. \end{aligned}$$

Finally, taking $\Pi = \Pi^*$ and substituting (37) into (36), we have

$$\begin{aligned} & \frac{1}{2c}\|\mathbf{x}^* - \mathbf{x}^k\|_Q^2 - \frac{1}{2c}\|\mathbf{x}^* - \mathbf{x}^{k+1}\|_Q^2 + \frac{\alpha}{2}\|\mathbf{z}^* - \mathbf{z}^k\|_{\mathcal{F}}^2 \quad (38) \\ & - \frac{\alpha}{2}\|\mathbf{z}^* - \mathbf{z}^{k+1}\|_{\mathcal{F}}^2 + \frac{1}{2\alpha}\|\Pi^* - \Pi^k\|_{\mathcal{F}}^2 - \frac{1}{2\alpha}\|\Pi^* - \Pi^{k+1}\|_{\mathcal{F}}^2 \\ \geq & \frac{1}{2c}\|\mathbf{x}^k - \mathbf{x}^{k+1}\|_Q^2 - \frac{L_f}{2}\|\mathbf{x}^k - \mathbf{x}^{k+1}\|_{\mathcal{F}}^2 + (\frac{\alpha}{2} + \frac{1}{\epsilon})\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_{\mathcal{F}}^2 \\ & + \frac{1}{2\alpha}\|\Pi^{k+1} - \Pi^k\|_{\mathcal{F}}^2 + \frac{1}{2\epsilon}\|\mathbf{z}^* - \mathbf{z}^{k+1}\|_{\mathcal{F}}^2. \end{aligned}$$

Choosing $\frac{1}{2c}Q - \frac{L_f}{2}I \succ 0$, or equivalently, $\frac{1}{2c}\left[1 - \alpha c\lambda_{\max}(I-W)\right] - \frac{L_f}{2} > 0$, we can find $\beta > 0$ such that $\frac{1}{2c}Q - \frac{L_f}{2}I \succ \beta I$, and it holds for all $k \geq 1$ that $\|\mathbf{u}^k - \mathbf{u}^*\|_P^2 - \|\mathbf{u}^{k+1} - \mathbf{u}^*\|_P^2 \geq \beta\|\mathbf{u}^k - \mathbf{u}^{k+1}\|_P^2$. This completes the proof. □

B. Proof of Theorem 2

Proof. Since $L(\nu, \Pi^*) \geq L(\nu^*, \Pi^*)$ holds for all ν, by letting $\nu = \tilde{\nu}^k$, we have

$$\Phi(\tilde{\nu}^k) + \left\langle \Pi^*, (I-W)^{\frac{1}{2}}\tilde{\mathbf{x}}^k - \tilde{\mathbf{z}}^k\right\rangle \geq \Phi(\nu^*). \quad (39)$$

By convexity, we have

$$\Phi\left(\frac{\nu^1 + \cdots + \nu^k}{k}\right) \leq \frac{\Phi(\nu^1) + \cdots + \Phi(\nu^k)}{k}.$$

Thus, we know that

$$\begin{aligned} & \Phi(\tilde{\nu}^k) - \Phi(\nu^*) + \left\langle \Pi, (I-W)^{\frac{1}{2}}\tilde{\mathbf{x}}^k - \tilde{\mathbf{z}}^k\right\rangle \quad (40) \\ \leq & \frac{\Phi(\nu^1) + \cdots + \Phi(\nu^k)}{k} - \Phi(\nu^*) + \left\langle \Pi, (I-W)^{\frac{1}{2}}\tilde{\mathbf{x}}^k - \tilde{\mathbf{z}}^k\right\rangle \\ = & \frac{1}{k}\sum_{t=0}^{k-1}\left[- \Phi(\nu^*) + \Phi(\nu^{t+1}) + \left\langle \Pi, (I-W)^{\frac{1}{2}}\mathbf{x}^{t+1} - \mathbf{z}^{t+1}\right\rangle\right] \\ \leq & \frac{1}{2k}\sum_{t=0}^{k-1}\Big[\frac{1}{c}\|\mathbf{x}^* - \mathbf{x}^t\|_Q^2 - \frac{1}{c}\|\mathbf{x}^* - \mathbf{x}^{t+1}\|_Q^2 + \alpha\|\mathbf{z}^* - \mathbf{z}^t\|_{\mathcal{F}}^2 \\ & - \alpha\|\mathbf{z}^* - \mathbf{z}^{t+1}\|_{\mathcal{F}}^2 + \frac{1}{\alpha}\|\Pi - \Pi^t\|_{\mathcal{F}}^2 - \frac{1}{\alpha}\|\Pi - \Pi^{t+1}\|_{\mathcal{F}}^2\Big] \\ = & \frac{1}{2k}\Big[\frac{1}{c}\|\mathbf{x}^* - \mathbf{x}^0\|_Q^2 - \frac{1}{c}\|\mathbf{x}^* - \mathbf{x}^k\|_Q^2 + \alpha\|\mathbf{z}^* - \mathbf{z}^0\|_{\mathcal{F}}^2 \\ & - \alpha\|\mathbf{z}^* - \mathbf{z}^k\|_{\mathcal{F}}^2 + \frac{1}{\alpha}\|\Pi - \Pi^0\|_{\mathcal{F}}^2 - \frac{1}{\alpha}\|\Pi - \Pi^k\|_{\mathcal{F}}^2\Big] \\ \leq & \frac{1}{2k}\left[\frac{1}{c}\|\mathbf{x}^* - \mathbf{x}^0\|_Q^2 + \alpha\|\mathbf{z}^* - \mathbf{z}^0\|_{\mathcal{F}}^2 + \frac{1}{\alpha}\|\Pi - \Pi^0\|_{\mathcal{F}}^2\right], \forall \Pi. \end{aligned}$$

Here, the second inequality comes from (36) and the results in Lemma 1. Specifically, if the parameters α and c are chosen as in Lemma 1, then the right-hand side of (36) is nonnegative.

Suppose that we further set

$$\Pi = \frac{2\|\Pi^*\|_{\mathcal{F}}\big((I-W)^{\frac{1}{2}}\tilde{\mathbf{x}}^k - \tilde{\mathbf{z}}^k\big)}{\|(I-W)^{\frac{1}{2}}\tilde{\mathbf{x}}^k - \tilde{\mathbf{z}}^k\|_{\mathcal{F}}}$$

in (40). Then, the left-hand side of (40) becomes

$$\begin{aligned} & \Phi(\tilde{\nu}^k) - \Phi(\nu^*) + \left\langle \Pi, (I-W)^{\frac{1}{2}}\tilde{\mathbf{x}}^k - \tilde{\mathbf{z}}^k\right\rangle \quad (41) \\ = & \Phi(\tilde{\nu}^k) - \Phi(\nu^*) + 2\|\Pi^*\|_{\mathcal{F}}\|(I-W)^{\frac{1}{2}}\tilde{\mathbf{x}}^k - \tilde{\mathbf{z}}^k\|_{\mathcal{F}}. \end{aligned}$$

For the right-hand side of (40), substituting $\|\Pi\|_{\mathcal{F}} = 2\|\Pi^*\|_{\mathcal{F}}$ and using the triangle inequality, we bound

$$\|\Pi - \Pi^0\|_{\mathcal{F}}^2 \leq 8\|\Pi^*\|_{\mathcal{F}}^2 + 2\|\Pi^0\|_{\mathcal{F}}^2. \quad (42)$$

Combining (40)–(42), we have

$$\Phi(\tilde{\nu}^k) - \Phi(\nu^*) \leq \frac{C_1}{k} - 2\|\Pi^*\|_{\mathcal{F}}\|(I-W)^{\frac{1}{2}}\tilde{\mathbf{x}}^k - \tilde{\mathbf{z}}^k\|_{\mathcal{F}}, \quad (43)$$

where $C_1 \triangleq \frac{1}{2c}\|\mathbf{x}^* - \mathbf{x}^0\|_Q^2 + \frac{\alpha}{2}\|\mathbf{z}^* - \mathbf{z}^0\|_{\mathcal{F}}^2 + \frac{4}{\alpha}\|\Pi^*\|_{\mathcal{F}}^2 + \frac{1}{\alpha}\|\Pi^0\|_{\mathcal{F}}^2$. By the Cauchy-Schwarz inequality, we have

$$\Phi(\tilde{\nu}^k) - \Phi(\nu^*) \geq -\|\Pi^*\|_{\mathcal{F}}\|(I-W)^{\frac{1}{2}}\tilde{\mathbf{x}}^k - \tilde{\mathbf{z}}^k\|_{\mathcal{F}}. \quad (44)$$

Using (44) and (43), we then have

$$\max\left\{|\Phi(\tilde{\nu}^k) - \Phi(\nu^*)|, \|\Pi^*\|_{\mathcal{F}}\|(I-W)^{\frac{1}{2}}\tilde{\mathbf{x}}^k - \tilde{\mathbf{z}}^k\|_{\mathcal{F}}\right\} \leq \frac{C_1}{k},$$

which completes the proof. □

C. Proof of Theorem 3

Proof. The proof is modified from that of Lemma 1. By the μ_f-strong convexity of f, (24) becomes

$$\begin{aligned} & \left\langle \mathbf{x}^* - \mathbf{x}^{k+1}, \nabla f\left(\mathbf{x}^k\right)\right\rangle \quad (45) \\ = & \left\langle \mathbf{x}^* - \mathbf{x}^{k+1}, \nabla f\left(\mathbf{x}^{k+1}\right)\right\rangle \\ & + \left\langle \mathbf{x}^* - \mathbf{x}^{k+1}, \nabla f\left(\mathbf{x}^k\right) - \nabla f\left(\mathbf{x}^{k+1}\right)\right\rangle \\ \leq & f(\mathbf{x}^*) - f\left(\mathbf{x}^{k+1}\right) - \frac{\mu_f}{2}\left\|\mathbf{x}^* - \mathbf{x}^{k+1}\right\|_{\mathcal{F}}^2 \\ & + \left\langle \mathbf{x}^* - \mathbf{x}^{k+1}, \nabla f\left(\mathbf{x}^k\right) - \nabla f\left(\mathbf{x}^{k+1}\right)\right\rangle. \end{aligned}$$

Following the derivation in the proof of Lemma 1, the inequality (38) is modified to

$$\begin{aligned} & \frac{1}{2c}\|\mathbf{x}^* - \mathbf{x}^k\|_Q^2 - \frac{1}{2c}\|\mathbf{x}^* - \mathbf{x}^{k+1}\|_Q^2 + \frac{\alpha}{2}\|\mathbf{z}^* - \mathbf{z}^k\|_{\mathcal{F}}^2 \quad (46) \\ & - \frac{\alpha}{2}\|\mathbf{z}^* - \mathbf{z}^{k+1}\|_{\mathcal{F}}^2 + \frac{1}{2\alpha}\|\Pi^* - \Pi^k\|_{\mathcal{F}}^2 - \frac{1}{2\alpha}\|\Pi^* - \Pi^{k+1}\|_{\mathcal{F}}^2 \\ \geq & \frac{1}{2c}\|\mathbf{x}^k - \mathbf{x}^{k+1}\|_Q^2 + \left(\frac{\alpha}{2} + \frac{1}{\epsilon}\right)\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_{\mathcal{F}}^2 \\ & + \frac{\mu_f}{2}\|\mathbf{x}^* - \mathbf{x}^{k+1}\|_{\mathcal{F}}^2 - \left\langle \mathbf{x}^* - \mathbf{x}^{k+1}, \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k+1})\right\rangle \\ & + \frac{1}{2\alpha}\|\Pi^{k+1} - \Pi^k\|_{\mathcal{F}}^2 + \frac{1}{2\epsilon}\|\mathbf{z}^* - \mathbf{z}^{k+1}\|_{\mathcal{F}}^2. \end{aligned}$$

The strong convexity of f leads to the term $\frac{\mu_f}{2}\|\mathbf{x}^* - \mathbf{x}^{k+1}\|_{\mathcal{F}}^2$ in (46), which enables us to establish the linear convergence result. Indeed, by expanding $\|\sqrt{\theta}(\mathbf{x}^* - \mathbf{x}^{k+1}) - \frac{1}{\sqrt{\theta}}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k+1}))\|_{\mathcal{F}}^2$, where $\theta > 0$ is any constant, and using the Lipschitz continuity of ∇f, we have

$$\begin{aligned} & - 2\left\langle \mathbf{x}^* - \mathbf{x}^{k+1}, \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k+1})\right\rangle \quad (47) \\ \geq & - \theta\|\mathbf{x}^* - \mathbf{x}^{k+1}\|_{\mathcal{F}}^2 - \frac{L_f^2}{\theta}\|\mathbf{x}^k - \mathbf{x}^{k+1}\|_{\mathcal{F}}^2. \end{aligned}$$

Substituting (47) into (46) yields

$$\begin{aligned}
&\frac{1}{2c}\|\mathbf{x}^*-\mathbf{x}^k\|_Q^2-\frac{1}{2c}\|\mathbf{x}^*-\mathbf{x}^{k+1}\|_Q^2+\frac{\alpha}{2}\|\mathbf{z}^*-\mathbf{z}^k\|_{\mathcal{F}}^2 \qquad (48)\\
&-\frac{\alpha}{2}\|\mathbf{z}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2+\frac{1}{2\alpha}\|\Pi^*-\Pi^k\|_{\mathcal{F}}^2-\frac{1}{2\alpha}\|\Pi^*-\Pi^{k+1}\|_{\mathcal{F}}^2\\
\geq&\frac{1}{2c}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_Q^2-\frac{L_f^2}{2\theta}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathcal{F}}^2\\
&+\left(\frac{\alpha}{2}+\frac{1}{\epsilon}\right)\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_{\mathcal{F}}^2+\frac{\mu_f-\theta}{2}\|\mathbf{x}^*-\mathbf{x}^{k+1}\|_{\mathcal{F}}^2\\
&+\frac{1}{2\alpha}\|\Pi^{k+1}-\Pi^k\|_{\mathcal{F}}^2+\frac{1}{2\epsilon}\|\mathbf{z}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2.
\end{aligned}$$

Substituting $\frac{1}{\epsilon}\mathbf{z}^{k+1}=\Pi^{k+1}$ into (48), we get

$$\begin{aligned}
&\frac{1}{c}\|\mathbf{x}^*-\mathbf{x}^k\|_Q^2-\frac{1}{c}\|\mathbf{x}^*-\mathbf{x}^{k+1}\|_Q^2 \qquad (49)\\
&+\left(\alpha+\frac{1}{\alpha\epsilon^2}\right)\|\mathbf{z}^*-\mathbf{z}^k\|_{\mathcal{F}}^2-\left(\alpha+\frac{1}{\alpha\epsilon^2}\right)\|\mathbf{z}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2\\
\geq&\frac{1}{c}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_Q^2-\frac{L_f^2}{\theta}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathcal{F}}^2\\
&+\left(\alpha+\frac{2}{\epsilon}+\frac{1}{\alpha\epsilon^2}\right)\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_{\mathcal{F}}^2\\
&+(\mu_f-\theta)\|\mathbf{x}^*-\mathbf{x}^{k+1}\|_{\mathcal{F}}^2+\frac{1}{\epsilon}\|\mathbf{z}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2\\
\geq&\left(\frac{\lambda_{\min}(Q)}{c}-\frac{L_f^2}{\theta}\right)\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathcal{F}}^2\\
&+\left(\alpha+\frac{2}{\epsilon}+\frac{1}{\alpha\epsilon^2}\right)\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_{\mathcal{F}}^2\\
&+(\mu_f-\theta)\|\mathbf{x}^*-\mathbf{x}^{k+1}\|_{\mathcal{F}}^2+\frac{1}{\epsilon}\|\mathbf{z}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2.
\end{aligned}$$

To prove the linear convergence of PAD, the parameters in (49) are required to satisfy

$$\begin{cases}\dfrac{\lambda_{\min}(Q)}{c}-\dfrac{L_f^2}{\theta}>0,\\ \mu_f-\theta>0,\end{cases} \qquad (50)$$

which is attainable when

$$\delta\triangleq 1-\frac{L_f^2 c}{\lambda_{\min}(Q)\mu_f}>0. \qquad (51)$$

When $\delta>0$, then (50) holds true if we choose $\theta=\frac{\mu_f}{1+\delta}$. Substituting this specific θ and the definition of δ, we can rewrite (49) as

$$\begin{aligned}
&\frac{1}{c}\|\mathbf{x}^*-\mathbf{x}^k\|_Q^2-\frac{1}{c}\|\mathbf{x}^*-\mathbf{x}^{k+1}\|_Q^2 \qquad (52)\\
&+\left(\alpha+\frac{1}{\alpha\epsilon^2}\right)\|\mathbf{z}^*-\mathbf{z}^k\|_{\mathcal{F}}^2-\left(\alpha+\frac{1}{\alpha\epsilon^2}\right)\|\mathbf{z}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2\\
\geq&\frac{\lambda_{\min}(Q)\delta^2}{c}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathcal{F}}^2+\frac{\mu_f\delta}{1+\delta}\|\mathbf{x}^*-\mathbf{x}^{k+1}\|_{\mathcal{F}}^2\\
&+\frac{1}{\epsilon}\|\mathbf{z}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2+\left(\alpha+\frac{2}{\epsilon}+\frac{1}{\alpha\epsilon^2}\right)\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_{\mathcal{F}}^2\\
\geq&\frac{\mu_f\delta}{1+\delta}\|\mathbf{x}^*-\mathbf{x}^{k+1}\|_{\mathcal{F}}^2+\frac{1}{\epsilon}\|\mathbf{z}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2.
\end{aligned}$$

After rearranging, the inequality (52) becomes

$$\begin{aligned}
&\frac{1}{c}\|\mathbf{x}^*-\mathbf{x}^k\|_Q^2+\left(\alpha+\frac{1}{\alpha\epsilon^2}\right)\|\mathbf{z}^*-\mathbf{z}^k\|_{\mathcal{F}}^2 \qquad (53)\\
\geq&\frac{\mu_f\delta}{1+\delta}\|\mathbf{x}^*-\mathbf{x}^{k+1}\|_{\mathcal{F}}^2+\frac{1}{c}\|\mathbf{x}^*-\mathbf{x}^{k+1}\|_Q^2\\
&+\left(\frac{1}{\epsilon}+\alpha+\frac{1}{\alpha\epsilon^2}\right)\|\mathbf{z}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2.
\end{aligned}$$

To complete the proof, we require the existence of a constant $c_1>0$ such that

$$\frac{\mu_f\delta}{1+\delta}\|\mathbf{x}^*-\mathbf{x}^{k+1}\|_{\mathcal{F}}^2\geq c_1\|\mathbf{x}^*-\mathbf{x}^{k+1}\|_Q^2, \qquad (54)$$

which can be satisfied when c_1 is sufficiently small because the eigenvalues of $Q=I-\alpha c(I-W)$ are upper bounded. One possible choice of c_1 is

$$c_1=\frac{\mu_f\delta}{(1+\delta)\lambda_{\max}(Q)}.$$

Substituting (54) into (53), we get

$$\begin{aligned}
&\frac{c_1+\frac{1}{c}}{\frac{1}{\epsilon}+\alpha+\frac{1}{\alpha\epsilon^2}}\|\mathbf{x}^*-\mathbf{x}^{k+1}\|_Q^2+\|\mathbf{z}^*-\mathbf{z}^{k+1}\|_{\mathcal{F}}^2 \qquad (55)\\
\leq&\eta\left(\frac{c_1+\frac{1}{c}}{\frac{1}{\epsilon}+\alpha+\frac{1}{\alpha\epsilon^2}}\|\mathbf{x}^*-\mathbf{x}^k\|_Q^2+\|\mathbf{z}^*-\mathbf{z}^k\|_{\mathcal{F}}^2\right),
\end{aligned}$$

where we define $\eta=\max\left\{\frac{\frac{1}{c}}{c_1+\frac{1}{c}},\frac{\alpha+\frac{1}{\alpha\epsilon^2}}{\frac{1}{\epsilon}+\alpha+\frac{1}{\alpha\epsilon^2}}\right\}$. By substituting the definition of C_2 in Theorem 3, we complete the proof. □

D. Proof of Proposition 1

Proof. Since $\mathbf{x}^*$ is the minimizer of the penalized problem (3), we have

$$\begin{aligned}
&\Psi(\mathbf{x}^*)+\frac{1}{2\epsilon}\|(I-W)^{\frac{1}{2}}\mathbf{x}^*\|_{\mathcal{F}}^2 \qquad (56)\\
\leq&\Psi(\hat{\mathbf{x}}^*)+\frac{1}{2\epsilon}\|(I-W)^{\frac{1}{2}}\hat{\mathbf{x}}^*\|_{\mathcal{F}}^2\\
=&\Psi(\hat{\mathbf{x}}^*),
\end{aligned}$$

where we use $(I-W)^{\frac{1}{2}}\hat{\mathbf{x}}^*=0$. Thus, it follows that

$$\|(I-W)^{\frac{1}{2}}\mathbf{x}^*\|_{\mathcal{F}}^2\leq 2\epsilon(\Psi(\hat{\mathbf{x}}^*)-\Psi(\mathbf{x}^*))\leq\epsilon C_3, \qquad (57)$$

where C_3 is a constant independent of ϵ such that $2(\Psi(\hat{\mathbf{x}}^*)-\Psi(\mathbf{x}^*))\leq C_3$. We need to verify the existence of such a constant. Under Assumption 4, $\Psi(\hat{\mathbf{x}}^*)$ is finite and independent of ϵ. Below we show that $\Psi(\mathbf{x}^*)$ has a lower bound that is finite and independent of ϵ. Indeed, since $\Psi(\mathbf{x})$ is strongly convex with constant μ_f, we have

$$\Psi(\mathbf{x}^*)\geq\Psi(0)+\langle\tilde{\nabla}\Psi(0),\mathbf{x}^*\rangle+\frac{\mu_f}{2}\|\mathbf{x}^*\|_{\mathcal{F}}^2, \qquad (58)$$

where

$$\tilde{\nabla}\Psi(0)=\begin{pmatrix}- & \left(\nabla f_1(0)+\tilde{\nabla}g_1(0)\right)^T & -\\ & \vdots & \\ - & \left(\nabla f_n(0)+\tilde{\nabla}g_n(0)\right)^T & -\end{pmatrix}\in\mathbb{R}^{n\times p}$$

and $\tilde{\nabla} g_i(0) \in \partial g_i(0), \forall i$. Observe that

$$\begin{aligned}&\Psi(0) + \langle \tilde{\nabla}\Psi(0), \mathbf{x}^*\rangle + \frac{\mu_f}{2}\|\mathbf{x}^*\|_{\mathcal{F}}^2 \\ \geq &\min_{\mathbf{x}} \left\{\Psi(0) + \langle \tilde{\nabla}\Psi(0), \mathbf{x}\rangle + \frac{\mu_f}{2}\|\mathbf{x}\|_{\mathcal{F}}^2\right\} \\ = &\Psi(0) - \frac{1}{2\mu_f}\|\tilde{\nabla}\Psi(0)\|_{\mathcal{F}}^2.\end{aligned} \tag{59}$$

Substituting (59) into (58), we have

$$\Psi(\mathbf{x}^*) \geq \Psi(0) - \frac{1}{2\mu_f}\|\tilde{\nabla}\Psi(0)\|_{\mathcal{F}}^2. \tag{60}$$

Without loss of generality, we assume that $0 \in \mathrm{dom}(\Psi)$ such that $\Psi(0) < \infty$. Otherwise, we can replace 0 by any solution in $\mathrm{dom}(\Psi)$ and (60) still holds. Thus, we can choose $C_3 = 2\left(\Psi(\hat{\mathbf{x}}^*) - \Psi(0) + \frac{1}{2\mu_f}\|\tilde{\nabla}\Psi(0)\|_{\mathcal{F}}^2\right)$, which is finite and independent of ϵ.

Then, according to the orthogonal decomposition of $\mathbf{x}^\star$ in (16), we have

$$\|(I-W)^{\frac{1}{2}}\mathbf{x}^*\|_{\mathcal{F}} = \|(I-W)^{\frac{1}{2}}\mathbf{x}_2^*\|_{\mathcal{F}} \geq \tilde{\lambda}_{\min}\|\mathbf{x}_2^*\|_{\mathcal{F}}, \tag{61}$$

where $\tilde{\lambda}_{\min}$ is the smallest nonzero eigenvalue of $(I-W)^{\frac{1}{2}}$. Substituting (57) into (61) leads to

$$\|\mathbf{x}_2^*\|_{\mathcal{F}} \leq \frac{(\epsilon C_3)^{1/2}}{\tilde{\lambda}_{\min}}. \tag{62}$$

By (62), the distance between $\mathbf{x}^*$ and $\mathbf{x}_1^*$ (i.e., $\|\mathbf{x}_2^*\|_{\mathcal{F}}$) is bounded. Since $\mathbf{x}^*$ is optimal for the penalized problem (3), we have

$$\Psi(\mathbf{x}^*) \leq \Psi(0) - \frac{1}{2\epsilon}\|(I-W)^{\frac{1}{2}}\mathbf{x}^*\|_{\mathcal{F}} \leq \Psi(0). \tag{63}$$

Thus, $\mathbf{x}^* \in \mathcal{C}_0(\Psi)$, where $\mathcal{C}_0(\Psi) = \{\mathbf{x} : \Psi(\mathbf{x}) \leq \Psi(0)\}$ is a sub-level set of the function Ψ. By the strong convexity of Ψ, we know that the sub-level set $\mathcal{C}_0(\Psi)$ is compact [38]. By (63) and (62), we know both $\mathbf{x}^*$ and $\mathbf{x}_1^*$ belong to a compact set that is independent of ϵ. Note that a finite convex function is Lipschitz on any compact set. Consequently, there exists a constant $L_\Psi > 0$, which is independent of ϵ, such that (17) holds. This completes the proof. □

E. Proof of Theorem 4

Proof. According to the triangle inequality, we have $\|\mathbf{x}^* - \hat{\mathbf{x}}^*\|_{\mathcal{F}} \leq \|\mathbf{x}_1^* - \hat{\mathbf{x}}^*\|_{\mathcal{F}} + \|\mathbf{x}_2^*\|_{\mathcal{F}}$. We will bound $\|\mathbf{x}_1^* - \hat{\mathbf{x}}^*\|_{\mathcal{F}}$ and $\|\mathbf{x}_2^*\|_{\mathcal{F}}$ in the following, respectively.

For $\|\mathbf{x}_2^*\|_{\mathcal{F}}$, although we have had an $O(\epsilon^{1/2})$ bound in (62), we will establish a better $O(\epsilon)$ bound using Proposition 1. Since $\hat{\mathbf{x}}^*$ is the minimizer of the constrained problem (2) and $(I-W)^{\frac{1}{2}}\mathbf{x}_1^* = 0$, we have

$$\Psi(\hat{\mathbf{x}}^*) \leq \Psi(\mathbf{x}_1^*). \tag{64}$$

Combining (56) with (64) and then substituting (17) in Proposition 1, we have

$$\frac{1}{2\epsilon}\|(I-W)^{\frac{1}{2}}\mathbf{x}^*\|_{\mathcal{F}}^2 \leq \Psi(\mathbf{x}_1^*) - \Psi(\mathbf{x}^*) \leq L_\Psi\|\mathbf{x}_2^*\|_{\mathcal{F}}. \tag{65}$$

Further combining (61) with (65), we have

$$\|\mathbf{x}_2^*\|_{\mathcal{F}} \leq \frac{2\epsilon L_\Psi}{\tilde{\lambda}_{\min}^2}. \tag{66}$$

For $\|\mathbf{x}_1^* - \hat{\mathbf{x}}^*\|_{\mathcal{F}}$, with (17) in Proposition 1, (56), and (66), we have

$$\begin{aligned}\Psi(\mathbf{x}_1^*) &\leq \Psi(\mathbf{x}^*) + L_\Psi\|\mathbf{x}^* - \mathbf{x}_1^*\|_{\mathcal{F}} \\ &\leq \Psi(\hat{\mathbf{x}}^*) + L_\Psi\|\mathbf{x}_2^*\|_{\mathcal{F}} \\ &\leq \Psi(\hat{\mathbf{x}}^*) + \frac{2\epsilon L_\Psi^2}{\tilde{\lambda}_{\min}^2}.\end{aligned} \tag{67}$$

By the strong convexity of Ψ, we have

$$\begin{aligned}\Psi(\mathbf{x}_1^*) &\geq \Psi(\hat{\mathbf{x}}^*) + \langle \tilde{\nabla}\Psi(\hat{\mathbf{x}}^*), \mathbf{x}_1^* - \hat{\mathbf{x}}^*\rangle + \frac{\mu_f}{2}\|\mathbf{x}_1^* - \hat{\mathbf{x}}^*\|_{\mathcal{F}}^2 \\ &= \Psi(\hat{\mathbf{x}}^*) + \frac{\mu_f}{2}\|\mathbf{x}_1^* - \hat{\mathbf{x}}^*\|_{\mathcal{F}}^2,\end{aligned} \tag{68}$$

where

$$\tilde{\nabla}\Psi(\hat{\mathbf{x}}^*) = \begin{pmatrix} - & \left(\nabla f_1(\hat{x}^*) + \tilde{\nabla} g_1(\hat{x}^*)\right)^T & - \\ & \vdots & \\ - & \left(\nabla f_n(\hat{x}^*) + \tilde{\nabla} g_n(\hat{x}^*)\right)^T & - \end{pmatrix} \in \mathbb{R}^{n\times p}$$

and $\tilde{\nabla} g_i(\hat{x}^*) \in \partial g_i(\hat{x}^*), \forall i$ satisfy the optimality condition of (1), namely, $\sum_{i=1}^n \left(\nabla f_i(\hat{x}^*) + \tilde{\nabla} g_i(\hat{x}^*)\right) = 0$. The equality in (68) holds because $\mathbf{x}_1^* - \hat{\mathbf{x}}^*$ contains n identical rows. Combining (68) with (67), we get

$$\|\mathbf{x}_1^* - \hat{\mathbf{x}}^*\|_{\mathcal{F}} \leq \frac{2L_\Psi\epsilon^{1/2}}{\tilde{\lambda}_{\min}\mu_f^{1/2}}. \tag{69}$$

With (66) and (69), we have

$$\begin{aligned}\|\mathbf{x}^* - \hat{\mathbf{x}}^*\|_{\mathcal{F}} &\leq \|\mathbf{x}_2^*\|_{\mathcal{F}} + \|\mathbf{x}_1^* - \hat{\mathbf{x}}^*\|_{\mathcal{F}} \\ &\leq \frac{2L_\Psi\epsilon}{\tilde{\lambda}_{\min}^2} + \frac{2L_\Psi\epsilon^{1/2}}{\tilde{\lambda}_{\min}\mu_f^{1/2}}.\end{aligned} \tag{70}$$

Consequently, if the network is not too poorly connected so that $\tilde{\lambda}_{\min} \gg \epsilon^{1/2}$, we have

$$\|\mathbf{x}^* - \hat{\mathbf{x}}^*\|_{\mathcal{F}} = O(\epsilon^{1/2}), \tag{71}$$

which completes the proof. □

REFERENCES

[1] S. Pu, W. Shi, J. Xu, and A. Nedić, "A push-pull gradient method for distributed optimization in networks," in *IEEE Conference on Decision and Control*, 2018, pp. 3385–3390.

[2] K. Scaman, F. Bach, S. Bubeck, L. Massoulié, and Y. T. Lee, "Optimal algorithms for non-smooth distributed optimization in networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 2740–2749.

[3] R. Dutta, L. Sun, and D. Pack, "A decentralized formation and network connectivity tracking controller for multiple unmanned systems," *IEEE Transactions on Control Systems Technology*, vol. 26, no. 6, pp. 2206–2213, 2017.

[4] C. Zhang, P. Zhao, S. Hao, Y. C. Soh, B. S. Lee, C. Miao, and S. C. Hoi, "Distributed multi-task classification: a decentralized online learning approach," *Machine Learning*, vol. 107, no. 4, pp. 727–747, 2018.

[5] A. Koppel, S. Paternain, C. Richard, and A. Ribeiro, "Decentralized online learning with kernels," *IEEE Transactions on Signal Processing*, vol. 66, no. 12, pp. 3240–3255, 2018.

[6] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[7] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1835–1854, 2016.

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

[9] A. Mokhtari, Q. Ling, and A. Ribeiro, "Network Newton distributed optimization methods," *IEEE Transactions on Signal Processing*, vol. 65, no. 1, pp. 146–161, 2016.

[10] D. Bajovic, D. Jakovetic, N. Krejic, and N. K. Jerinkic, "Newton-like method with diagonal correction for distributed optimization," *SIAM Journal on Optimization*, vol. 27, no. 2, pp. 1171–1203, 2017.

[11] J. F. Mota, J. M. Xavier, P. M. Aguiar, and M. Püschel, "D-ADMM: A communication-efficient distributed algorithm for separable optimization," *IEEE Transactions on Signal Processing*, vol. 61, no. 10, pp. 2718–2723, 2013.

[12] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the ADMM in decentralized consensus optimization," *IEEE Transactions on Signal Processing*, vol. 62, no. 7, pp. 1750–1761, 2014.

[13] M. Maros and J. Jaldén, "On the Q-linear convergence of distributed generalized ADMM under non-strongly convex function components," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 5, no. 3, pp. 442 – 453, 2018.

[14] Q. Ling, W. Shi, G. Wu, and A. Ribeiro, "DLM: Decentralized linearized alternating direction method of multipliers," *IEEE Transactions on Signal Processing*, vol. 63, no. 15, pp. 4051–4064, 2015.

[15] D. Jakovetić, J. M. Moura, and J. Xavier, "Linear convergence rate of a class of distributed augmented lagrangian algorithms," *IEEE Transactions on Automatic Control*, vol. 60, no. 4, pp. 922–936, 2014.

[16] K. Scaman, F. Bach, S. Bubeck, Y. T. Lee, and L. Massoulié, "Optimal algorithms for smooth and strongly convex distributed optimization in networks," in *International Conference on Machine Learning*, 2017, pp. 3027–3036.

[17] T. H. Chang, M. Hong, and X. Wang, "Multi-agent distributed optimization via inexact consensus ADMM," *IEEE Transactions on Signal Processing*, vol. 63, no. 2, pp. 482–497, 2014.

[18] N. S. Aybat, Z. Wang, T. Lin, and S. Ma, "Distributed linearized alternating direction method of multipliers for composite convex consensus optimization," *IEEE Transactions on Automatic Control*, vol. 63, no. 1, pp. 5–20, 2017.

[19] 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.

[20] ——, "A proximal gradient algorithm for decentralized composite optimization," *IEEE Transactions on Signal Processing*, vol. 63, no. 22, pp. 6013–6023, 2015.

[21] Z. Li, W. Shi, and M. Yan, "A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates," *IEEE Transactions on Signal Processing*, vol. 67, no. 17, pp. 4494–4506, 2019.

[22] S. Alghunaim, K. Yuan, and A. H. Sayed, "A linearly convergent proximal gradient algorithm for decentralized optimization," in *Advances in Neural Information Processing Systems*, 2019, pp. 2848–2858.

[23] Y. Sun, A. Daneshmand, and G. Scutari, "Convergence rate of distributed optimization algorithms based on gradient tracking," *arXiv preprint arXiv:1905.02637*, 2019.

[24] P. Di Lorenzo and G. Scutari, "NEXT: In-network nonconvex optimization," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 2, pp. 120–136, 2016.

[25] G. Scutari and Y. Sun, "Distributed nonconvex constrained optimization over time-varying digraphs," *Mathematical Programming*, vol. 176, no. 1-2, pp. 497–544, 2019.

[26] J. Xu, Y. Tian, Y. Sun, and G. Scutari, "Distributed algorithms for composite optimization: Unified and tight convergence analysis," *arXiv preprint arXiv:2002.11534*, 2020.

[27] S. A. Alghunaim, E. Ryu, K. Yuan, and A. H. Sayed, "Decentralized proximal gradient algorithms with linear convergence rates," *IEEE Transactions on Automatic Control*, 2020.

[28] S. Vlaski and A. H. Sayed, "Proximal diffusion for stochastic costs with non-differentiable regularizers," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2015, pp. 3352–3356.

[29] N. Bastianello, A. Ajalloeian, and E. Dall'Anese, "Distributed and inexact proximal gradient method for online convex optimization," *arXiv preprint arXiv:2001.00870*, 2020.

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

[31] S. Vlaski, L. Vandenberghe, and A. H. Sayed, "Regularized diffusion adaptation via conjugate smoothing," *arXiv preprint arXiv:1909.09417*, 2019.

[32] S. Boyd, P. Diaconis, and L. Xiao, "Fastest mixing Markov chain on a graph," *SIAM Review*, vol. 46, no. 4, pp. 667–689, 2004.

[33] S. U. Pillai, T. Suel, and S. Cha, "The Perron-Frobenius theorem: some of its applications," *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp. 62–75, 2005.

[34] D. P. Bertsekas, "Nonlinear programming," *Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, 1997.

[35] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[36] S. Ma, "Alternating proximal gradient method for convex minimization," *Journal of Scientific Computing*, vol. 68, no. 2, pp. 546–572, 2016.

[37] A. Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani, "An exact quantized decentralized gradient descent algorithm," *IEEE Transactions on Signal Processing*, vol. 67, no. 19, pp. 4934–4947, 2019.

[38] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.