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 composed 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:

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

Here, $f_i : \mathbb{R}^p \rightarrow \mathbb{R}$ and $g_i : \mathbb{R}^p \rightarrow \mathbb{R} \cup \{+\infty\}$ 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]–[9]. 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 $O(1/k)$, 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 $O(1/k)$ and $O(1/k^2)$, 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.
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 in (1), the problem \( \min_{x} \sum_{i=1}^{n} f_i(x) + g_i(x) \) in (1), the problem \( \min_{x} \sum_{i=1}^{n} f_i(x) + g_i(x) \) 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 further demonstrate the advantages of PAD over existing state-of-the-art algorithms such as DL-ADMM, PG-EXTRA, and NIDS.

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 \( G = (\mathcal{N}, \mathcal{E}) \), 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 \( x \) to denote the matrix obtained by stacking up the row vectors \( x_1^T, \ldots, x_n^T \); i.e.,

\[
\begin{pmatrix}
-x_1^T & - & - \\
-x_2^T & - & - \\
\vdots & \ddots & \\
-x_n^T & - & -
\end{pmatrix} \in \mathbb{R}^{n \times p}.
\]
We say that $x$ is consensual if all of its rows are identical; i.e., $x^T_1 = \cdots = x^T_n$. Next, we denote the sums of the smooth and nonsmooth parts in $\mathcal{F}$ by
\[
    f(x) \triangleq \sum_{i=1}^{n} f_i(x_i) \quad \text{and} \quad g(x) \triangleq \sum_{i=1}^{n} g_i(x_i),
\]
respectively. The gradient of $f$ at $x$ is given by
\[
    \nabla f(x) \triangleq \begin{pmatrix}
        - (\nabla f_1(x_1))^T \\
        - (\nabla f_2(x_2))^T \\
        \vdots \\
        - (\nabla f_n(x_n))^T
    \end{pmatrix} \in \mathbb{R}^{n \times p},
\]
where $\nabla f_i(x_i)$ is the gradient of $f_i$ at $x_i$. A subgradient of $g$ at $x$ is given by
\[
    \partial g(x) \triangleq \begin{pmatrix}
        - (\partial g_1(x_1))^T \\
        - (\partial g_2(x_2))^T \\
        \vdots \\
        - (\partial g_n(x_n))^T
    \end{pmatrix} \in \mathbb{R}^{n \times p},
\]
where $\partial g_i(x_i)$ is a subgradient of $g_i$ at $x_i$. The $i$th rows of $\nabla f$, $\partial g$, and $\partial y$ belong to agent $i$.

Given a vector $v$, $|v|$ and $\|v\|_1$ denote the $\ell_2$ and $\ell_1$ norms, respectively. Given a matrix $A$, $a_{ij}$ denotes the element in the $i$th row and $j$th column, while $a^T_i$ denotes the $i$th row of $A$. The Frobenius norm of a matrix $A$ is denoted by $\|A\|_F$. Given a symmetric matrix $A$, 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 $\langle A, A' \rangle_G \triangleq \text{trace}(A^T A)$ and the induced norm $\|A\|_G \triangleq \sqrt{\text{trace}(A^T A)}$. The largest and smallest eigenvalues of a symmetric matrix $A$ are denoted by $\lambda_{\text{max}}(A)$ and $\lambda_{\text{min}}(A)$, respectively. We use $I$ to denote the identity matrix; $I_{m \times n}$ and $0_{m \times n}$ denote the $m \times n$ all-ones and $m \times n$ all-zeros matrices, respectively. Given a closed proper convex function $g$, the proximal mapping of $g$ is given by
\[
    \text{prox}_{\gamma g}(x) \triangleq \arg\min_{y} \left\{ g(y) + \frac{1}{2\gamma} \|y - x\|^2 \right\},
\]
where $\gamma > 0$ is a scalar parameter. The norm is $\ell_2$ 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 $\mathcal{F}$ 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 $\nabla f_i$ is Lipschitz continuous; i.e.,
\[
    \|\nabla f_i(x) - \nabla f_i(y)\| \leq L_f \|x - y\|, \quad \forall x, y \in \mathbb{R}^p,
\]
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.,
\[
    \langle x - y, \nabla f_i(x) - \nabla f_i(y) \rangle \geq \mu_f \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^p,
\]
where $\mu_f > 0$ is the strong convexity constant.

We begin with a consensus-constrained reformulation of $\mathcal{F}$. Let $x_i \in \mathbb{R}^p$ be the local variable held by agent $i$. To characterize the consensual property $x^T_1 = \cdots = x^T_n$, 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 $W_{1 \times 1} = \mathbf{1}_{n \times 1}$. The null space of $I - W$ is span $\{\mathbf{1}_{n \times 1}\}$. Furthermore, if $j \notin N_i$, 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 span $\{\mathbf{1}_{n \times 1}\}$, so is the null space of its square root $(I - W)^{1/2}$. Therefore, $(I - W)^{1/2} x = 0$ if and only if $x^T_1 = \cdots = x^T_n$. In particular, we see that $\mathcal{F}$ is equivalent to the constrained problem
\[
    x^* \in \arg\min_x \left\{ f(x) + g(x) \right\} \quad \text{s.t.} \quad (I - W)^{1/2} x = 0, \tag{2}
\]
where $x^* \in \mathbb{R}^{n \times p}$ contains $n$ identical rows of $(\bar{x})^T$.

A. Penalized Approximation Formulation

Instead of solving the constrained problem (2), we consider its penalized approximation in the form of
\[
    x^* \in \arg\min_x \left\{ f(x) + g(x) + \frac{1}{2\epsilon} \|(I - W)^{1/2} x\|^2_F \right\}, \tag{3}
\]
where $\epsilon > 0$ is the penalty parameter. The approximation error (i.e., the gap between $\hat{x}^*$ and $x^*$) is controlled by $\epsilon$. When $\epsilon$ 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 $g(x) = 0$, problem (3) can be solved using gradient descent, whose iterates are generated by
\[
    x^{k+1} = x^k - \gamma \left( \nabla f(x^k) + \frac{1}{\epsilon} (I - W)x^k \right), \tag{4}
\]
with $\gamma > 0$ being the step size. However, when $\epsilon$ is very small, the Lipschitz constant of the gradient $\nabla f(x) + \frac{1}{\epsilon} (I - W)x$ is on the order of $O(\frac{1}{\epsilon})$, which suggests that $\gamma$ must be on the order of $O(\epsilon)$ to guarantee convergence [34]. In fact, setting $\gamma = \epsilon$ recovers the DGD update $x^{k+1} = Wx^k - \epsilon \nabla f(x^k)$ [6], [7]. Even though DGD will converge to a small neighborhood of $\hat{x}^*$ when $\epsilon$ 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 $g(x) \neq 0$ [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 $z \in \mathbb{R}^{n \times p}$, problem (3) is equivalent to

$$
(x^*, z^*) = \arg\min_{x, z} \left\{ f(x) + g(x) + \frac{1}{2\epsilon} \|z\|_F^2 \right\},
$$

s.t. $(I - W)^{\frac{1}{2}}x = z$.

The augmented Lagrangian function of (5) is given by

$$L_\alpha(x, z, \Pi) \triangleq f(x) + g(x) + \frac{1}{2\epsilon} \|z\|_F^2 + \langle \Pi, (I - W)^{\frac{1}{2}}x - z \rangle + \frac{\alpha}{2} \| (I - W)^{\frac{1}{2}}x - z \|_F^2,$$

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 $x$ and $z$ in an alternating manner, then updates the dual variable $\Pi$ 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 $L_\alpha(x, z, \Pi)$, the subproblem w.r.t. $x$ does not have a closed-form solution. Thus, we use the composite structure to inexactly update $x$. Specifically, we linearize the smooth part and calculate the proximal mapping of the nonsmooth part as follows.

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

$$L_\alpha(x, z^k, \Pi^k) \triangleq f(x) + \frac{1}{2\epsilon} \|z^k\|_F^2 + \langle \Pi^k, (I - W)^{\frac{1}{2}}x - z^k \rangle + \frac{\alpha}{2} \| (I - W)^{\frac{1}{2}}x - z^k \|_F^2.$$

The idea here is to replace the smooth part $L_\alpha(x, z^k, \Pi^k)$ by the following quadratic approximation centered at $x^k$:

$$\tilde{L}_\alpha(x, z^k, \Pi^k) \approx L_\alpha(x^k, z^k, \Pi^k) + \langle \nabla_x \tilde{L}_\alpha(x^k, z^k, \Pi^k), x - x^k \rangle + \frac{1}{2\epsilon} \| x - x^k \|_F^2.$$

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

$$x^{k+1} = \arg\min_{x} \left\{ g(x) + \tilde{L}_\alpha(x^k, z^k, \Pi^k) + \langle \nabla_x \tilde{L}_\alpha(x^k, z^k, \Pi^k), x - x^k \rangle + \frac{1}{2\epsilon} \| x - x^k \|_F^2 \right\},$$

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

$$x^{k+1} = \text{prox}_{cg} \left( x^k - c \left[ \nabla f(x^k) + \alpha (I - W)^{\frac{1}{2}} \right] \right).$$

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

$$z^{k+1} = \frac{1}{\alpha + 1} \left[ \Pi^k + \alpha (I - W)^{\frac{1}{2}}x^{k+1} \right].$$

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

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

According to (6)–(8) can be simplified by introducing $\tilde{\Pi}^k \triangleq (I - W)^{\frac{1}{2}}\Pi^k$ and $\tilde{z}^k \triangleq (I - W)^{\frac{1}{2}}z^k$. 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 $x_i^k$ from the previous iteration. The implementations of Steps 4 and 5 require current neighboring variables $x_i^{k+1}$, 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 $\epsilon$, $\alpha$, and $c$. Initialize $x_i^0 = 0$, $\pi_i^0 = 0$, and $\bar{x}_i^0 = 0$.

1: for $k = 1, 2, \ldots$ do
2: Update local variable $x_i^{k+1}$ by

$$x_i^{k+1} \leftarrow \text{prox}_{cg} \left( x_i^k - c \left[ \nabla f_i(x_i^k) + \frac{1}{\alpha + 1} \left[ \bar{x}_i^k + \alpha \left( x_i^k - \sum_{j \in N_i} w_{ij}x_j^k - z_i^k \right) + \bar{\pi}_i^k \right] \right) \right).$$

3: Transmit $x_i^{k+1}$ to receive $x_j^{k+1}$ from neighbors $j \in N_i$.
4: Update local auxiliary variable $\bar{x}_i^{k+1}$ by

$$\bar{x}_i^{k+1} \leftarrow \frac{1}{\alpha + 1} \left[ \bar{\pi}_i^k + \alpha \left( x_i^k - \sum_{j \in N_i} w_{ij}x_j^{k+1} \right) \right].$$

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

$$\bar{\pi}_i^{k+1} \leftarrow \bar{\pi}_i^k + \alpha \left( x_i^{k+1} - \sum_{j \in N_i} w_{ij}x_j^{k+1} \right) - z_i^{k+1}.$$  
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 $g(x) = 0$ and show that the dual domain method EXTRA proposed in [19] is a special case of PAD. The update of EXTRA is

$$x^{k+2} = (I + W)x^{k+1} - \frac{I + W}{2}x^k + \epsilon_n\nabla f(x^k) - c\nabla f(x^k),$$

(9)

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

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

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

(10)

When the variables are also initialized as $\frac{1}{\epsilon}x^0 = \Pi^0$, we have $\frac{1}{\epsilon}x^k = \Pi^k$, and consequently, $\frac{1}{\epsilon}z^k = \Pi^k$. When $g(x) = 0$, substituting $\frac{1}{\epsilon}x^k = \Pi^k$ into (9) yields

$$x^{k+1} = x^k - c\nabla f(x^k) + \alpha(I - W)x^k + (1 - \alpha\epsilon)\Pi^k$$

(11)

and

$$x^{k+2} = x^{k+1} - c\nabla f(x^{k+1}) + \alpha(I - W)x^{k+1} + (1 - \alpha\epsilon)\Pi^{k+1}.$$ 

(12)

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

$$x^{k+2} = \frac{1}{1 + \alpha\epsilon} \left\{ (2 + 3\alpha - 2\alpha I + 2\alpha W)x^{k+1} - [I - \alpha\epsilon(I - W)]x^k - c[\nabla f(x^{k+1}) - \nabla f(x^k)] \right\}. $$

(13)

Comparing (9) with (13), we observe that the updates of PAD and EXTRA are equivalent if $\alpha = \frac{1}{2}$ and $\epsilon = 0$. The parameter $\epsilon$ tunes the accuracy-speed tradeoff in PAD. In addition, $\alpha\epsilon$ 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 $I - W = L_0$, where $\rho > 0$ is an approximation parameter. $D$ is the degree matrix, and $L_0$ is the oriented Laplacian defined in [14]. Then, PAD becomes

$$x^{k+2} = (I - \alpha\rho I + 2\alpha D^{-1})^{-1}L_0(2x^{k+1} - x^k) - \rho I + 2\alpha D^{-1} \left\{ \nabla f(x^{k+1}) - \nabla f(x^k) \right\}, $$

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

When $g(x) \neq 0$, PAD utilizes the proximal mapping [34] to handle this nonsmooth term. In this scenario the proximal-gradient (PG)-EXTRA algorithm [29] modifies EXTRA to

$$y^{k+1} = x^k - c\nabla f(x^k) + \alpha(I - W)^{1/2}(I - W)^{1/2}x^k - z^k, $$

(14)

and

$$x^{k+1} = \text{prox}_c (y^{k+1}).$$ 

(15)

Below we show that PAD can recover PG-EXTRA too. Let $y^{k+1} = x^k - c\nabla f(x^k) + \alpha(I - W)^{1/2}(I - W)^{1/2}x^k - z^k + \frac{\Pi^k}{\alpha\epsilon}$.

With $z^k = (I - W)^{1/2}z^k$ and $\frac{1}{\epsilon}z^k = \Pi^k$, we have

$$y^{k+1} - y^k = [I - \alpha\epsilon(I - W)](x^k - x^{k-1}) - c\nabla f(x^k) - \nabla f(x^{k-1}) - (\alpha\epsilon)(I - W)x^k - \alpha\epsilon(I - W)x^k - \alpha\epsilon(I - W)x^k,$$

$$x^{k+1} = \text{prox}_c (y^{k+1}).$$

When $\alpha = \frac{1}{2}$ 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 $\{z^k\}$ and $\{\Pi^k\}$, we investigate the sequences $\{z^k\}$ and $\{\Pi^k\}$ generated by (6)–(8), because their analyses are equivalent and the sequence $\{x^k\}$ 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

$$L(x, z, \Pi) \triangleq f(x) + g(x) + \frac{1}{2\epsilon}z^\top z + \langle \Pi, (I - W)^{1/2}x - z \rangle.$$ 

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

$$-\nabla f(x^*) - (I - W)^{1/2}\Pi^* \in \partial g(x^*),$$

(14a)

$$\frac{1}{\epsilon}z^* = \Pi^*,$$ 

(14b)

$$(I - W)^{1/2}x^* = z^*.$$ 

(14c)

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

**Lemma 1.** Under Assumptions 7 and 3, if the parameters $\alpha$ and $\epsilon$ are chosen such that $\frac{1}{\beta\epsilon} I - \alpha\epsilon(I - W) - \frac{\epsilon}{\beta\epsilon} I > 0$, then it holds for all $k \geq 1$ that

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

(15)

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

**Proof.** See Appendix A. □
With Lemma 1, the sequence \( \{\|u^k - u^*\|_2^2\} \) converges to 0. The proof is the same as that of Theorem 3.2 in [36] and is omitted here. Since \( Q = I - \alpha c (I - W) \geq 0 \) given that \( \frac{1}{\lambda_{\max}(I - W)} \geq 0 \) that \( \frac{1}{\lambda_{\max}(I - W)} \geq \frac{1}{\lambda_{\min}(Q)} \) and \( \frac{1}{\lambda_{\max}(I - W)} \) is the largest eigenvalue of \( I - W \).

Therefore, Theorem 1 implies that PAD has guaranteed convergence when \( \nu = \frac{1}{\lambda_{\max}(Q)} \) and \( \alpha = \frac{1}{\lambda_{\max}(Q)} \) are small enough.

### B. Sublinear Convergence Rate under General Convexity

This subsection establishes the \( O(\frac{1}{\nu}) \) convergence rate of PAD in an ergodic sense under general convexity. Similar to the proof in [18], we define \( \nu \triangleq \langle x, z \rangle \) and \( \Phi(\nu) \triangleq f(x) + g(x) + \frac{1}{2\nu} \|z\|_2^2 \).

**Theorem 2.** Under Assumptions 1 and 2, let \( \{u^k\} \triangleq \{x^k, z^k, \Pi^k\} \) be the sequence generated by (3)–(5). Define \( \tilde{\nu}^k \triangleq \langle \tilde{x}^k, \tilde{z}^k \rangle \) with \( \tilde{x}^k \triangleq \frac{1}{k} \sum_{i=1}^{k} x^i \) and \( \tilde{z}^k \triangleq \frac{1}{k} \sum_{i=1}^{k} z^i \) for \( k \geq 1 \). If the parameters \( c \) and \( \alpha \) are chosen as in Lemma 1 then it holds for all \( k \geq 1 \) that

\[
\max \left\{ |\Phi(\tilde{\nu}^k) - \Phi(\nu^*)|, ||\Pi||_F (I - W)^{\frac{1}{2}} \tilde{x}^k - \tilde{z}^k||_F \right\} \leq \frac{C_1}{k},
\]

where \( C_1 \triangleq \frac{1}{2\nu} \|x^* - x^0\|_2^2 + \frac{1}{2\nu} \|z^* - z^0\|_2^2 + \lambda_{\Pi^k} \|\Gamma\Pi^k\|_F^2 + \frac{1}{\alpha} \|\Pi^k\|_2^2 \) 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 \( \alpha \) and \( c \) satisfy \( c < \frac{1}{\lambda_{\max}(I - W)} \), then both the optimality gap in terms of function value \( |\Phi(\tilde{\nu}) - \Phi(\nu^*)| \) and the constraint violation \( ||(I - W)^{\frac{1}{2}} \tilde{x}^k - \tilde{z}^k||_F \) converge sublinearly in an ergodic sense. This result holds for any \( \nu > 0 \) so that we can set \( \epsilon \) sufficiently small to reach high accuracy. The analysis in the general convex case can be extended to \( \epsilon = 0 \) by modifying \( \Phi(\nu) \triangleq f(x) + g(x) \) and \( C_1 \triangleq \frac{1}{2\nu} \|x^* - x^0\|_2^2 + \frac{1}{2\nu} \|z^* - z^0\|_2^2 + \lambda_{\Pi^k} \|\Gamma\Pi^k\|_F^2 + \frac{1}{\alpha} \|\Pi^k\|_2^2 \).

### 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 and 2, suppose that \( \delta \triangleq \frac{1}{\lambda_{\min}(Q)} > 0 \) and the variables are initialized as \( \frac{1}{\epsilon} z^0 = \Pi^0 \).

Then, the sequence \( \{x^k, z^k\} \) generated by (3)–(5) converges to \( (x^*, z^*) \) with

\[
C_2 \|x^* - x^{k+1}\|_2^2 + \|z^* - z^{k+1}\|_2^2 \leq \eta(C_2 \|x^* - x^k\|_2^2 + \|z^* - z^k\|_2^2),
\]

where

\[
\eta \triangleq \max \left\{ \frac{1}{c_1 + \frac{\alpha}{c_1 + 1}} \frac{\alpha^2 c_1 + 1}{\epsilon}, \frac{1}{\epsilon} \right\}
\]

and

\[
c_1 = \frac{\mu_j \delta}{(1 + \delta) \lambda_{\max}(Q)}.
\]

Proof. See Appendix C.

**Theorem 3** requires \( \delta \leq \frac{1}{\lambda_{\max}(Q)} \), and 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 \( \{C_2 \|x^* - x^k\|_2^2 + \|z^* - z^k\|_2^2\} \) to \( 0 \), from which we know that the sequence \( \{x^k\} \) converges to \( x^* \) and \( \{z^k\} \) converges to \( z^* \), both R-linearly. The analysis of linear rate under the strongly convex case no longer holds when \( \epsilon = 0 \). Otherwise, the result will contradict with that in [27]. With \( \epsilon = 0 \), the inequality (53) becomes \( \frac{1}{\epsilon} \|x^* - x^k\|_2^2 + \frac{\lambda_{\Pi^k} \|\Gamma\Pi^k\|_F^2}{\epsilon} \geq \frac{\mu_j \delta}{(1 + \delta) \lambda_{\max}(Q)} \|x^* - x^k\|_2^2 + \frac{1}{\epsilon} \|z^* - z^k\|_2^2 \). Since the coefficients of \( \|\Pi^k - \Pi^*\|_2^2 \) and \( \|\Pi^k - \Pi^*\|_2^2 \) are equal, we are unable to establish the linear rate.

Note that \( \eta \to 1 \) as \( \epsilon \to 0 \) since, meaning that the convergence of the sequence \( \{C_2 \|x^* - x^k\|_2^2 + \|z^* - z^k\|_2^2\} \) can be slow when \( \epsilon \) is small. However, in the numerical experiments we will show that \( \|x^k - x^*\|_2 \) (and \( \|x^k - x^*\|_F \)) too itself converges fast under different \( \epsilon \) until reaching a small neighborhood of \( x^* \). This does not contradict with Theorem 3 since \( C_2 = O(\epsilon^2) \).

### D. Bound on \( \|x^* - \hat{x}^*\|_2 \) under Strong Convexity

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

\[
x^* = x_1^* + x_2^* \tag{16}
\]

where \( x_1^* \in \text{null}(I - W) \) and \( x_2^* \in \text{null}(I - W)⊥ \). We begin with the following proposition.

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

\[
|\Psi(x_1^*) - \Psi(x^*)| \leq L_\Psi \|x_1^* - x^*\|_2 \tag{17}
\]

Proof. See Appendix D.
Based on Proposition 1, the bound on $\|x^* - \hat{x}^*\|_F$ is established as follows.

**Theorem 4.** Under the same assumptions as in Proposition 7 there exists a constant $L_\Psi > 0$, which is independent of $\epsilon$, such that the distance between the optimal solution $x^*$ to the penalized problem (3) and the optimal solution $\hat{x}^*$ to the constrained problem (2) is bounded by

$$
\|x^* - \hat{x}^*\|_F \leq \frac{2L_\Psi \epsilon}{\lambda_{\text{min}}^2} + \frac{2L_\Psi \epsilon}{\lambda_{\text{min}} H_f^2},
$$

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

$$
\|x^* - \hat{x}^*\|_F = O(\epsilon^2).
$$

**Proof.** See Appendix E.

Theorem 4 implies that $x^* \to \hat{x}^*$ as $\epsilon \to 0$. To achieve a high accuracy, we can set $\epsilon$ 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 $\|x^* - \hat{x}^*\|_F = O(\epsilon)$ 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 $\frac{\|x^* - \hat{x}^*\|_F}{\|x^*\|_F}$ and $\frac{\|z^* - \hat{z}^*\|_F}{\|z^*\|_F}$, where the optimal solution $x^*$ to (2) and the optimal primal-dual solution $(x^*, z^*)$ 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 $(M_{(i)}j, y_{(i)}j) \in \mathbb{R}^p \times \{-1, +1\}$, $j = 1, \ldots, m_i$, where $M_{(i)}j$ is the $j$th feature vector of agent $i$ and $y_{(i)}j$ is the corresponding binary label. The entries of each $\{M_{(i)}j\}_{j=1}^m$ are zero or one. PAD and proximal DGD with a fixed step size both converge to a neighborhood of $\hat{x}^*$. However, the step size of proximal DGD must be in the same order of magnitude as the Lipschitz constant $L_f$ to reach the optimal solution, which is consistent with Theorem 3. PAD allows the penalty parameter $\epsilon$ to be sufficiently small without sacrificing the speed.

![Fig. 1](image.png)

**Fig. 1.** Plots of $\|x^* - \hat{x}^*\|_F$ 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 $\epsilon = 0.7$ are hand-tuned to get the best results.

**B. Decentralized Quadratic Programming**

Each agent $i$ has a local quadratic function $f_i(x) = \frac{1}{2}x^TQ_ix + h_i^Tx$ and a local linear constraint $a_i^Tx \leq b_i$, where $Q_i \in \mathbb{R}^{p \times p}$, $h_i \in \mathbb{R}^p$, $a_i \in \mathbb{R}^p$, and $b_i \in \mathbb{R}$. The quadratic programming problem to solve takes the form

$$
\min_{x} \sum_{i=1}^{n} \left( \frac{1}{2}x^TQ_ix + h_i^Tx \right) s.t. a_i^Tx \leq b_i, \forall i = 1, \ldots, n.
$$

It can be reformulated to the unconstrained form (1) by defining the nonsmooth function $g_i$ as

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

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 $\epsilon$ to be sufficiently small without sacrificing the speed.
Theorem 3. We can see that the linear convergence rate slows down and linearly convergent when in addition the smooth parts are sublinearly convergent when the private functions are convex, allowing the use of a small penalty parameter without sacrificing penalized methods, the proposed PAD is fast even with a very small penalty parameter for the decentralized convex composite problem. Unlike existing 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

\[ x^k - c \left[ \nabla f(x^k) + \alpha (I - W) \tilde{\frac{1}{2}} (I - W) \tilde{\frac{1}{2}} x^k \right. \]
\[ \left. - z^k + \Pi^k \right] - x^{k+1} \in c \partial g(x^{k+1}). \]  

By the definition of \( \partial g \), we have

\[ \langle x^* - x^{k+1}, x^k - x^{k+1} - c \left[ \nabla f(x^k) + \alpha (I - W) \tilde{\frac{1}{2}} (I - W) \tilde{\frac{1}{2}} x^k - z^k + \Pi^k \right] \rangle \leq cg(x^*) - cg(x^{k+1}). \]  

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

\[ g(x^*) - g(x^{k+1}) \]
\[ \geq \left\langle x^* - x^{k+1}, \frac{1}{c} (x^k - x^{k+1}) \right\rangle - \left[ \nabla f(x^k) + \alpha (I - W) \tilde{\frac{1}{2}} (I - W) \tilde{\frac{1}{2}} (x^k - x^{k+1}) \right. \]
\[ - (z^k - z^{k+1}) \left. \right] + (I - W) \tilde{\frac{1}{2}} \Pi^{k+1} \]
\[ = \left\langle x^* - x^{k+1}, \frac{1}{c} [I - \alpha c (I - W)](x^k - x^{k+1}) \right\rangle \]
\[ - (x^* - x^{k+1}, \nabla f(x^k) + (I - W) \tilde{\frac{1}{2}} \Pi^{k+1} \tilde{\frac{1}{2}} x^{k+1} \]
\[ - \alpha (I - W) \tilde{\frac{1}{2}} (z^k - z^{k+1}). \]

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

\[ g(x^*) - g(x^{k+1}) \]
\[ \geq \left\langle x^* - x^{k+1}, \frac{1}{c} (x^k - x^{k+1}) \right\rangle - \frac{1}{2c} (\|x^* - x^{k+1}\|_Q^2 + \|x^k - x^{k+1}\|_Q^2 - \|x^* - x^{k+1}\|_Q^2) \]
\[ - (x^* - x^{k+1}, \nabla f(x^k) + (I - W) \tilde{\frac{1}{2}} \Pi^{k+1} \tilde{\frac{1}{2}} x^{k+1} \tilde{\frac{1}{2}} x^{k+1}) \]
\[ + \alpha (I - W) \tilde{\frac{1}{2}} (x^* - x^{k+1}, z^k - z^{k+1}). \]

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

\[ (x^* - x^{k+1}, \nabla f(x^k)) \]
\[ = (x^* - x^{k+1}, \nabla f(x^k)) + (x^k - x^{k+1}, \nabla f(x^k)) \]
\[ \leq f(x^*) - f(x^k) + f(x^k) - f(x^{k+1}) + \frac{L_f}{2} \|x^k - x^{k+1}\|_2^2 \]
\[ = f(x^*) - f(x^{k+1}) + \frac{L_f}{2} \|x^k - x^{k+1}\|_2^2. \]
For the term (ii) in (23), using the identity \((v_1 - v_2)^2 = \frac{1}{2}((v_1 - v_2)^2 - \|v_1 - v_3\|^2) + \frac{1}{2}(\|v_2 - v_3\|^2 - \|v_2 - v_4\|^2)\), we have

\[
\langle (I - W)^{\frac{1}{2}}(x^* - x^{k+1}), z - z^{k+1} \rangle = \frac{1}{2}(\|(I - W)^{\frac{1}{2}}x^* - z^{k+1}\|_\mathcal{F}^2 - \|(I - W)^{\frac{1}{2}}z - z^{k}\|_\mathcal{F}^2) + \frac{1}{2}(\|(I - W)^{\frac{1}{2}}x^{k+1} - z^{k}\|_\mathcal{F}^2 - \|(I - W)^{\frac{1}{2}}z^k - z^{k+1}\|_\mathcal{F}^2).
\]

By substituting \((I - W)^{\frac{1}{2}}x^{k+1} - z^{k+1} = \frac{\Pi_{k+1} - \Pi_k}{\alpha}\) obtained from (3), the equation (25) becomes

\[
\langle (I - W)^{\frac{1}{2}}(x^* - x^{k+1}), (z - z^{k+1}) \rangle = \frac{1}{2}(\|(I - W)^{\frac{1}{2}}x^* - z^{k+1}\|_\mathcal{F}^2 - \|(I - W)^{\frac{1}{2}}z^* - z^{k}\|_\mathcal{F}^2) + \frac{1}{2}(\|(I - W)^{\frac{1}{2}}x^{k+1} - z^k\|_\mathcal{F}^2 - \|(I - W)^{\frac{1}{2}}z^k - z^{k+1}\|_\mathcal{F}^2).
\]

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

\[
g(x^*) - g(x^{k+1}) \geq \frac{1}{2\varepsilon}(\|x^* - x^{k+1}\|_Q^2 + \|x^k - x^{k+1}\|_Q^2 - \|x^* - x^k\|_Q^2)
- f(x^*) + f(x^{k+1}) - \frac{L_f}{2}\|x^k - x^{k+1}\|_F^2
- \langle x^* - x^{k+1}, (I - W)^{\frac{1}{2}}\Pi_{k+1} \rangle
+ \frac{\alpha}{2}(\|(I - W)^{\frac{1}{2}}x^* - z^{k+1}\|_\mathcal{F}^2 - \|(I - W)^{\frac{1}{2}}z^* - z^{k}\|_\mathcal{F}^2)
+ \frac{\alpha}{2}(\|(I - W)^{\frac{1}{2}}x^{k+1} - z^k\|_\mathcal{F}^2 - \|(I - W)^{\frac{1}{2}}z^k - z^{k+1}\|_\mathcal{F}^2).
\]

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

\[
g(x^*) + f(x^*) - g(x^{k+1}) - f(x^{k+1}) + \langle x^* - x^{k+1}, (I - W)^{\frac{1}{2}}\Pi_{k+1} \rangle \geq \frac{1}{2\varepsilon}(\|x^* - x^{k+1}\|_Q^2 + \|x^k - x^{k+1}\|_Q^2 - \|x^* - x^k\|_Q^2)
- \frac{L_f}{2}\|x^k - x^{k+1}\|_F^2 + \frac{\alpha}{2}(\|x^* - z^{k+1}\|_\mathcal{F}^2 - \|x^* - z^k\|_\mathcal{F}^2)
- \frac{\alpha}{2}(\|x^{k+1} - z^k\|_\mathcal{F}^2 - \|x^{k+1} - z^{k+1}\|_\mathcal{F}^2).
\]

Now, by direct calculation, we have

\[
\frac{1}{2\varepsilon}\|z^* - z^{k+1}\|_\mathcal{F}^2 + \langle z^* - z^{k+1}, \frac{1}{\varepsilon}z^{k+1} \rangle = \frac{1}{2\varepsilon}\|z^*\|_\mathcal{F}^2 - \frac{1}{2\varepsilon}\|z^{k+1}\|_\mathcal{F}^2.
\]

Substituting (10) into (29), we get

\[
\frac{1}{2\varepsilon}\|z^* - z^{k+1}\|_\mathcal{F}^2 + \langle z^* - z^{k+1}, \Pi_{k+1} \rangle = \frac{1}{2\varepsilon}\|z^*\|_\mathcal{F}^2 - \frac{1}{2\varepsilon}\|z^{k+1}\|_\mathcal{F}^2.
\]

Substituting (30) into (28) then yields

\[
g(x^*) + f(x^*) + \frac{1}{2\varepsilon}\|z^*\|_\mathcal{F}^2 - g(x^{k+1}) - f(x^{k+1}) - \frac{1}{2\varepsilon}\|z^{k+1}\|_\mathcal{F}^2 - \langle x^* - x^{k+1}, (I - W)^{\frac{1}{2}}\Pi_{k+1} \rangle
- \langle x^* - x^{k+1}, (I - W)^{\frac{1}{2}}\Pi_{k+1} \rangle
- \frac{1}{2\varepsilon}\|z^* - x^{k+1}\|_Q^2 - \|z^* - x^k\|_Q^2
- \frac{1}{2\varepsilon}\|x^* - x^{k+1}\|_Q^2 + \frac{\alpha}{2}\|z^* - z^{k+1}\|_\mathcal{F}^2
- \alpha\|z^* - z^{k+1}\|_\mathcal{F}^2 + \frac{1}{2\alpha}\|\Pi_{k+1} - \Pi_k\|_\mathcal{F}^2
- \frac{1}{2\alpha}\|\Pi_{k+1} - \Pi_k\|_\mathcal{F}^2 + \frac{1}{2\varepsilon}\|z^* - z^{k+1}\|_\mathcal{F}^2,
\]

Next, we proceed to simplify (31). By the update of \(\Pi_{k+1}\) in (3), we have

\[
0 = \langle \Pi - \Pi_{k+1}, \frac{\Pi_{k+1} - \Pi_k}{\alpha} \rangle = \langle \Pi - \Pi_{k+1}, \frac{\Pi_{k+1} - \Pi_k}{\alpha} \rangle = \langle \Pi - \Pi_{k+1}, (I - W)^{\frac{1}{2}}x^{k+1} + z^{k+1} \rangle
- \langle \Pi - \Pi_{k+1}, \frac{\Pi_{k+1} - \Pi_k}{\alpha} \rangle, \forall \Pi.
\]

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

\[
g(x^*) + f(x^*) + \frac{1}{2\varepsilon}\|z^*\|_\mathcal{F}^2 - g(x^{k+1}) - f(x^{k+1}) - \frac{1}{2\varepsilon}\|z^{k+1}\|_\mathcal{F}^2 - \langle x^* - x^{k+1}, (I - W)^{\frac{1}{2}}\Pi_{k+1} \rangle
- \langle x^* - x^{k+1}, (I - W)^{\frac{1}{2}}\Pi_{k+1} \rangle
- \frac{1}{2\varepsilon}\|z^* - x^{k+1}\|_Q^2 - \|z^* - x^k\|_Q^2
- \frac{1}{2\varepsilon}\|x^* - x^{k+1}\|_Q^2 + \frac{\alpha}{2}\|z^* - z^{k+1}\|_\mathcal{F}^2
- \alpha\|z^* - z^{k+1}\|_\mathcal{F}^2 + \frac{1}{2\alpha}\|\Pi_{k+1} - \Pi_k\|_\mathcal{F}^2
- \frac{1}{2\alpha}\|\Pi_{k+1} - \Pi_k\|_\mathcal{F}^2 + \frac{1}{2\varepsilon}\|z^* - z^{k+1}\|_\mathcal{F}^2
\]

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

\[
g(x^*) + f(x^*) + \frac{1}{2\varepsilon}\|z^*\|_\mathcal{F}^2 - g(x^{k+1}) - f(x^{k+1}) - \frac{1}{2\varepsilon}\|z^{k+1}\|_\mathcal{F}^2 - \langle x^* - x^{k+1}, (I - W)^{\frac{1}{2}}\Pi_{k+1} \rangle
- \langle x^* - x^{k+1}, (I - W)^{\frac{1}{2}}\Pi_{k+1} \rangle
- \frac{1}{2\varepsilon}\|z^* - x^{k+1}\|_Q^2 - \|z^* - x^k\|_Q^2
- \frac{1}{2\varepsilon}\|x^* - x^{k+1}\|_Q^2 + \frac{\alpha}{2}\|z^* - z^{k+1}\|_\mathcal{F}^2
- \alpha\|z^* - z^{k+1}\|_\mathcal{F}^2 + \frac{1}{2\alpha}\|\Pi_{k+1} - \Pi_k\|_\mathcal{F}^2
- \frac{1}{2\alpha}\|\Pi_{k+1} - \Pi_k\|_\mathcal{F}^2 + \frac{1}{2\varepsilon}\|z^* - z^{k+1}\|_\mathcal{F}^2,
\]

\[\forall \Pi.\]
When $\Pi = \Pi^*$, (33) becomes
\[
g(x^*) + f(x^*) + \frac{1}{2c}\|z^*\|^2 - g(x^{k+1}) - f(x^{k+1}) \tag{37}
\]
\[-\frac{1}{2c}\|z^{k+1}\|^2 + \langle \Pi^*, -(I - W)^{\frac{1}{2}}z^{k+1} \rangle = L(x^*, z^*, \Pi^*) - L(x^{k+1}, z^{k+1}, \Pi^*) \leq 0.
\]
Finally, taking $\Pi = \Pi^*$ and substituting (37) into (36), we have
\[
\frac{1}{2c}\|x^* - x^k\|^2 - \frac{1}{2c}\|x^* - x^{k+1}\|^2 + \frac{\alpha}{2}\|z^* - z^{k+1}\|^2 + \frac{\alpha}{2}\|\Pi^* - \Pi^{k+1}\|^2 
\geq \frac{1}{2c}\|x^* - x^{k+1}\|^2 - \frac{L_f}{2}\|x^* - x^{k+1}\|^2 + \left(\frac{\alpha}{2} + \frac{1}{2c}\right)\|z^{k+1} - z^*\|^2 
+ \frac{1}{2\alpha}\|\Pi^{k+1} - \Pi^k\|^2 + \frac{1}{2c}\|z^* - z^{k+1}\|^2.
\]
Choosing $\frac{1}{2c}Q - \frac{L_f}{2}I > 0$, or equivalently, $\frac{1}{2c}\left[1 - \alpha c\max(1 - |I - W|^{-1})\right] > \frac{L_f}{2} > 0$, we can find $\beta > 0$ such that $\frac{1}{2c}Q - \frac{L_f}{2}I > \beta I$, and it holds for all $k \geq 1$ that $\|u^k - u^1\|^2 - \|u^{k+1} - u^1\|^2 \geq \beta\|u^k - u^{k+1}\|^2$. This completes the proof.

\section*{B. Proof of Theorem \ref{thm:main2}}

\textbf{Proof.} Since $L(\nu, \Pi^*) \geq L(\nu^*, \Pi^*)$ holds for all $\nu$, by letting $\nu = \tilde{\nu}^k$, we have
\[
\Phi(\tilde{\nu}^k) + \langle \Pi^*, (I - W)^{\frac{1}{2}}\tilde{x}^k - \tilde{z}^k \rangle \geq \Phi(\nu^*). \tag{39}
\]
By convexity, we have
\[
\Phi\left(\frac{\nu^1 + \cdots + \nu^k}{k}\right) \leq \frac{1}{k}\Phi(\nu^1) + \cdots + \Phi(\nu^k).
\]
Thus, we know that
\[
\Phi(\tilde{\nu}^k) - \Phi(\nu^*) + \langle \Pi^*, (I - W)^{\frac{1}{2}}\tilde{x}^k - \tilde{z}^k \rangle \leq \Phi(\nu^1) + \cdots + \Phi(\nu^k) - \Phi(\nu^*) + \langle \Pi^*, (I - W)^{\frac{1}{2}}\tilde{x}^k - \tilde{z}^k \rangle \tag{40}
\]
\[
\leq \frac{1}{k}\left[\Phi(\nu^1) + \cdots + \Phi(\nu^k) - \Phi(\nu^*) + \langle \Pi^*, (I - W)^{\frac{1}{2}}\tilde{x}^k - \tilde{z}^k \rangle\right]
\leq \frac{1}{k}\left[\frac{1}{2c}\|x^* - x^1\|^2 - \frac{1}{c}\|x^* - x^{k+1}\|^2 + \alpha\|z^* - z^{k+1}\|^2 + \frac{\alpha}{2}\|\Pi^* - \Pi^{k+1}\|^2ight]

In (40), then the left-hand side of (40) becomes
\[
\Phi(\tilde{\nu}^k) - \Phi(\nu^*) + \langle \Pi^*, (I - W)^{\frac{1}{2}}\tilde{x}^k - \tilde{z}^k \rangle = \Phi(\tilde{\nu}^k) - \Phi(\nu^*) + 2\|\Pi^*\|_{x^*}(I - W)^{\frac{1}{2}}\tilde{x}^k - \tilde{z}^k \|_{x^*}
\]
For the right-hand side of (40), substituting $\|\Pi^*\|_{x^*} = 2\|\Pi^*\|_{x^*}$ and using the triangle inequality, we bound
\[
\|\Pi - \Pi^0\|^2_{x^*} \leq 8\|\Pi^*\|^2_{x^*} + 2\|\Pi^0\|^2_{x^*}. \tag{42}
\]
Combining (40)–(42), we have
\[
\Phi(\tilde{\nu}^k) - \Phi(\nu^*) \leq C_1 \frac{1}{k} - 2\|\Pi^*\|_{x^*}(I - W)^{\frac{1}{2}}\tilde{x}^k - \tilde{z}^k \|_{x^*},
\]
where $C_1 \equiv \frac{1}{2c}\|x^* - x^0\|^2 + \frac{\alpha}{2}\|z^* - z^0\|^2 + \frac{\alpha}{2}\|\Pi^*\|^2_{x^*} + \frac{1}{2}\|\Pi^0\|^2_{x^*}$.

Using (44) and (43), we have
\[
\max\left\{\|\Phi(\tilde{\nu}^k) - \Phi(\nu^*)\|, \|\Pi^*\|_{x^*}(I - W)^{\frac{1}{2}}\tilde{x}^k - \tilde{z}^k \|_{x^*}\right\} \leq \frac{C_1}{k},
\]
which completes the proof.

\section*{C. Proof of Theorem \ref{thm:main3}}

\textbf{Proof.} The proof is modified from that of Lemma \ref{lem:main1}.

By the $\mu_f$-strong convexity of $f$, (24) becomes
\[
\langle x^* - x^{k+1}, \nabla f(x^*) \rangle \tag{45}
\]
\[
= \langle x^* - x^{k+1}, \nabla f(x^{k+1}) \rangle 
+ \langle x^* - x^{k+1}, \nabla f(x^{k+1}) - \nabla f(x^*) \rangle 
\leq \langle x^* - x^{k+1}, \nabla f(x^{k+1}) - \frac{H_f}{2}\|x^* - x^{k+1}\|^2 \rangle 
+ \langle x^* - x^{k+1}, \nabla f(x^*) - \nabla f(x^{k+1}) \rangle.
\]

Following the derivation in the proof of Lemma \ref{lem:main1} the inequality (38) is modified to
\[
\frac{1}{2c}\|x^* - x^k\|^2 - \frac{1}{2c}\|x^* - x^{k+1}\|^2 + \frac{\alpha}{2}\|z^* - z^{k+1}\|^2 
- \frac{\alpha}{2}\|\Pi^* - \Pi^{k+1}\|^2_{x^*} - \frac{1}{2\alpha}\|\Pi^* - \Pi^{k+1}\|^2_{x^*} \tag{46}
\geq \frac{1}{2c}\|x^* - x^{k+1}\|^2_{x^*} + \left(\frac{\alpha}{2} + \frac{1}{2c}\right)\|z^{k+1} - z^*\|^2_{x^*} 
+ \frac{1}{2\alpha}\|\Pi^{k+1} - \Pi^k\|^2_{x^*} + \frac{1}{2c}\|z^* - z^{k+1}\|^2_{x^*}.
\]

The strong convexity of $f$ leads to the term $\frac{\mu_f}{2}\|x^* - x^{k+1}\|^2_{x^*}$ in (46), which enables us to establish the linear convergence result. Indeed, by expanding $\|\sqrt{\theta}(x^* - x^{k+1}) - \frac{1}{\sqrt{\theta}}(\nabla f(x^*) - \nabla f(x^{k+1}))\|^2_{x^*},$ where $\theta > 0$ is any constant, and using the Lipschitz continuity of $\nabla f$, we have
\[
-2\langle x^* - x^{k+1}, \nabla f(x^*) - \nabla f(x^{k+1}) \rangle \tag{47}
\geq -\theta\|x^* - x^{k+1}\|^2 - \frac{L_f^2}{\theta}\|x^* - x^{k+1}\|^2_{x^*},
\]
Substituting (47) into (46) yields
\[ \frac{1}{2c} \| x^* - x^k \|^2_Q - \frac{1}{2c} \| x^* - x^{k+1} \|^2_Q + \frac{\alpha}{2} \| z^* - z^k \|^2_F \geq \frac{1}{2c} \| x^k - x^{k+1} \|^2_Q - \frac{L_j^2}{2\theta} \| z^k - z^{k+1} \|^2_F + \left( \alpha + \frac{1}{\alpha \epsilon^2} \right) \| z^k - z^{k+1} \|^2_F. \] (48)

Substituting \( \frac{1}{2c} \| x^k - x^{k+1} \|^2_Q \) into (48), we get
\[ \frac{1}{c} \| x^* - x^k \|^2_Q \geq \frac{1}{c} \| x^* - x^{k+1} \|^2_Q + \left( \alpha + \frac{1}{\alpha \epsilon^2} \right) \| z^* - z^k \|^2_F + \left( \frac{\alpha}{2} + \frac{1}{\alpha \epsilon^2} \right) \| z^k - z^{k+1} \|^2_F. \] (49)

Differentiating (47) with respect to \( k \), we have
\[ \frac{1}{c} \| x^* - x^k \|^2_Q - \frac{1}{c} \| x^* - x^{k+1} \|^2_Q \geq \left( \frac{\alpha}{2} + \frac{1}{\alpha \epsilon^2} \right) \| z^k - z^{k+1} \|^2_F + \left( \alpha + \frac{1}{\alpha \epsilon^2} \right) \| z^k - z^{k+1} \|^2_F. \] (51)

After rearranging, the inequality (52) becomes
\[ \frac{1}{c} \| x^* - x^k \|^2_Q + \left( \alpha + \frac{1}{\alpha \epsilon^2} \right) \| z^* - z^k \|^2_F \geq \frac{\mu_1 \delta}{1 + \delta} \| x^* - x^{k+1} \|^2_Q + \frac{1}{c} \| x^* - x^{k+1} \|^2_Q. \] (53)

To complete the proof, we require the existence of a constant \( c_1 > 0 \) such that
\[ \frac{\mu_1 \delta}{1 + \delta} \| x^* - x^{k+1} \|^2_Q \geq c_1 \| x^* - x^{k+1} \|^2_Q. \] (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
\[ c_1 = \frac{\mu_1 \delta}{(1 + \delta) \lambda_{\max}(Q)}. \]

Substituting (54) into (53), we get
\[ \frac{1}{c} \| x^* - x^k \|^2_Q - \frac{1}{c} \| x^* - x^{k+1} \|^2_Q \geq \frac{c_1}{c} \| x^* - x^{k+1} \|^2_Q + \| z^* - z^{k+1} \|^2_F \]
\[ \leq \frac{c_1}{c} \| x^* - x^k \|^2_Q + \| z^* - z^{k+1} \|^2_F, \] (55)

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

\[ D. \text{ Proof of Proposition 1} \]

**Proof.** Since \( x^* \) is the minimizer of the penalized problem (3), we have
\[ \Psi(x^*) + \frac{1}{2\epsilon} \| (I - W)^{1/2} x^* \|^2_F \leq \Psi(x), \]
\[ \leq \Psi(x^*) + \frac{1}{2\epsilon} \| (I - W)^{1/2} x^* \|^2_F = \Psi(x^*), \] (56)

where we use \( (I - W)^{1/2} x^* = 0 \). Thus, it follows that
\[ \| (I - W)^{1/2} x^* \|^2_F \leq 2\epsilon (\Psi(x^*) - \Psi(x^*)) \leq \epsilon C_3, \] (57)

where \( C_3 \) is a constant independent of \( \epsilon \) such that \( 2(\Psi(x^*) - \Psi(x^*)) \leq C_3 \). We need to verify the existence of such a constant. Under Assumption 3, \( \Psi(x^*) \) is finite and independent of \( \epsilon \). Below we show that \( \Psi(x^*) \) has a lower bound that is finite and independent of \( \epsilon \). Indeed, since \( \Psi(x) \) is strongly convex with constant \( \mu_f \), we have
\[ \Psi(x^*) \geq \Psi(0) + \langle \nabla \Psi(0), x^* \rangle + \frac{\mu_f}{2} \| x^* \|^2_F, \] (58)

where
\[ \tilde{\nabla} \Psi(0) = \left( \begin{array}{c} -\langle \nabla f_1(0) + \nabla g_1(0), x^* \rangle \\ \vdots \\ -\langle \nabla f_n(0) + \nabla g_n(0), x^* \rangle \end{array} \right) \in \mathbb{R}^{n \times p}. \]
and $\nabla g_i(0) \in \partial g_i(0), \forall i$. Observe that
\[
\Psi(0) + \langle \nabla \Psi(0), x^* \rangle + \frac{\mu I}{2} \|x^*\|^2_F \geq \min_x \left\{ \Psi(0) + \langle \nabla \Psi(0), x \rangle + \frac{\mu I}{2} \|x\|^2_F \right\}
\]
\[= \Psi(0) - \frac{1}{2\mu_f} \|\nabla \Psi(0)\|^2_F. \tag{59} \]
Substituting (59) into (58), we have
\[
\Psi(x^*) \geq \Psi(0) - \frac{1}{2\mu_f} \|\nabla \Psi(0)\|^2_F. \tag{60} \]
Without loss of generality, we assume that $0 \in \text{dom}(\Psi)$ such that $\Psi(0) < \infty$. Otherwise, we can replace 0 by any solution in dom(\Psi) and (60) still holds. Thus, we can choose $C_3 = 2 \|\Psi(x^*) - \Psi(0) + \frac{1}{2\mu_f} \|\nabla \Psi(0)\|^2_F\|$, which is finite and independent of $\epsilon$.

Then, according to the orthogonal decomposition of $x^*$ in (16), we have
\[
\|(I - W)^\frac{3}{2} x^*\|_F = \|(I - W)^\frac{3}{2} x_1\|_F \geq \hat{\lambda}_{\text{min}} \|x_1\|_F, \tag{61} \]
where $\hat{\lambda}_{\text{min}}$ is the smallest nonzero eigenvalue of $(I - W)^\frac{3}{2}$. Substituting (57) into (61) leads to
\[
\|x_1\|_F \leq (\epsilon C_3)^{1/2} \hat{\lambda}_{\text{min}}. \tag{62} \]
By (62), the distance between $x^*$ and $x_1^*$ (i.e., $\|x_1\|_F$) is bounded. Since $x^*$ is optimal for the penalized problem (1), we have
\[
\Psi(x^*) \leq \Psi(0) - \frac{1}{2\epsilon} \|(I - W)^\frac{3}{2} x^*\|_F \leq \Psi(0). \tag{63} \]
Thus, $x^* \in C_0(\Psi)$, where $C_0(\Psi) = \{x : \Psi(x) \leq \Psi(0)\}$ is a sub-level set of the function $\Psi$. By the strong convexity of $\Psi$, we know that the sub-level set $C_0(\Psi)$ is compact [38]. By (63) and (62), we know both $x^*$ and $x_1^*$ belong to a compact set that is independent of $\epsilon$. Note that a finite convex function is Lipschitz on any compact set. Consequently, there exists a constant $L_\Psi > 0$, which is independent of $\epsilon$, such that (17) holds. This completes the proof.

### E. Proof of Theorem 4

**Proof.** According to the triangle inequality, we have $\|x^* - \hat{x}^*\|_F \leq \|x_1^* - \hat{x}^*\|_F + \|x_2\|_F$. We will bound $\|x_1^* - \hat{x}^*\|_F$ and $\|x_2\|_F$ in the following, respectively.

For $\|x_2\|_F$, although we have had an $O(\epsilon^{1/2})$ bound in (62), we will establish a better $O(\epsilon)$ bound using Proposition 1.

Since $\hat{x}^*$ is the minimizer of the constrained problem (2) and $(I - W)^\frac{3}{2} x_1^* = 0$, we have
\[
\Psi(\hat{x}^*) \leq \Psi(x_1^*). \tag{64} \]
Combining (64) with (64) and then substituting (17) in Proposition 1 we have
\[
\frac{1}{2\epsilon} \|(I - W)^\frac{3}{2} x^*\|_F^2 \leq \Psi(x_1^*) - \Psi(\hat{x}^*) \leq L_\Psi \|x_2\|_F. \tag{65} \]
Further combining (61) with (65), we have
\[
\|x_2\|_F \leq \frac{2\epsilon L_\Psi}{\lambda_{\text{min}}} \tag{66} \]
For $\|x_1^* - \hat{x}^*\|_F$, with (17) in Proposition 1 (56), and (66), we have
\[
\Psi(x_1^*) \leq \Psi(x^*) + L_\Psi \|x^* - x_1^*\|_F \leq \Psi(\hat{x}^*) + L_\Psi \|\hat{x}^* - x_1^*\|_F \leq \Psi(\hat{x}^*) + \frac{2\epsilon L_\Psi}{\lambda_{\text{min}}}, \tag{67} \]
By the strong convexity of $\Psi$, we have
\[
\Psi(x_1^*) \geq \Psi(\hat{x}^*) + \langle \nabla \Psi(\hat{x}^*), x_1^* - \hat{x}^* \rangle + \frac{\mu I}{2} \|x_1^* - \hat{x}^*\|_F^2 \tag{68} \]
where
\[
\nabla \Psi(\hat{x}^*) = - (\nabla f_1(\hat{x}^*) + \nabla g_1(\hat{x}^*))^T - \cdots - (\nabla f_n(\hat{x}^*) + \nabla g_n(\hat{x}^*))^T \in \mathbb{R}^{n \times p}. \]
and $\nabla g_i(\hat{x}^*) \in \partial g_i(\hat{x}^*), \forall i$ satisfy the optimality condition of (1), namely, $\sum_{i=1}^n \left( \nabla f_i(\hat{x}^*) + \nabla g_i(\hat{x}^*) \right) = 0$. The equality in (68) holds because $x_1^* - \hat{x}^*$ contains $n$ identical rows. Combining (68) with (67), we get
\[
\|x_1^* - \hat{x}^*\|_F \leq \frac{2L_\Psi \epsilon^{1/2}}{\lambda_{\text{min}}} \tag{69} \]
With (66) and (69), we have
\[
\|x^* - \hat{x}^*\|_F \leq \|x_2\|_F + \|x_1^* - \hat{x}^*\|_F \leq \frac{2L_\Psi \epsilon}{\lambda_{\text{min}}} + \frac{2L_\Psi \epsilon^{1/2}}{\lambda_{\text{min}}}. \tag{70} \]
Consequently, if the network is not too poorly connected so that $\hat{\lambda}_{\text{min}} \gg \epsilon^{1/2}$, we have
\[
\|x^* - \hat{x}^*\|_F = O(\epsilon^{1/2}), \tag{71} \]
which completes the proof. \hfill \Box

### REFERENCES


