A Proximal Gradient Algorithm for Decentralized Composite Optimization

Wei Shi, Qing Ling, Gang Wu, Wotao Yin

Abstract—This paper proposes a decentralized algorithm for solving a consensus optimization problem defined in a static networked multi-agent system, where the local objective functions have the smooth+nonsmooth composite form. Examples of such problems include decentralized constrained quadratic programming and compressed sensing problems, as well as many regularization problems arising in inverse problems, signal processing, and machine learning, which have decentralized applications. This paper addresses the need for efficient decentralized algorithms that take advantages of proximal operations for the nonsmooth terms.

We propose a proximal gradient exact first-order algorithm (PG-EXTRA) that utilizes the composite structure and has the best known convergence rate. It is a nontrivial extension to the recent algorithm EXTRA. At each iteration, each agent locally computes a gradient of the smooth part of its objective and a proximal map of the nonsmooth part, as well as exchange information with its neighbors. The algorithm is “exact” in the sense that an exact consensus minimizer can be obtained with a fixed step size, whereas most previous methods must use diminishing step sizes. When the smooth part has Lipschitz gradients, PG-EXTRA has an ergodic convergence rate of $O(\frac{1}{k})$ in terms of the first-order optimality residual. When the smooth part vanishes, PG-EXTRA reduces to P-EXTRA, an algorithm that does not compute the gradients (so no “G” in the name), which has a slightly improved convergence rate at $o(\frac{1}{k})$ in a standard (non-ergodic) sense. Numerical experiments demonstrate effectiveness of PG-EXTRA and validate our convergence results.

Index Terms—Multi-agent network, decentralized optimization, composite objective, nonsmooth, regularization, proximal

I. INTRODUCTION

This paper considers a connected network of $n$ agents that cooperatively solve the consensus optimization problem in the form

$$\min_{x \in \mathbb{R}^p} \bar{f}(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x),$$

where

$$f_i(x) := s_i(x) + r_i(x),$$

and $s_i, r_i : \mathbb{R}^p \rightarrow \mathbb{R}$ are convex differentiable and possibly nondifferentiable functions, respectively, that are kept private by agent $i = 1, \ldots, n$. We say that the objective has the smooth+nonsmooth composite structure. We develop an algorithm for all the agents in the network to obtain a consensual solution to problem (1). In the algorithm, each agent $i$ locally computes the gradient $\nabla s_i$ and the so-called proximal operation of $r_i$ (see Section I-C for its definition) and performs one-hop communication with its neighbors. The iterations of the agents are synchronized.

The smooth+nonsmooth structure of the local objectives arises in a large number of signal processing, statistical inference, and machine learning problems. Specific examples include (i) the geometric median problem in which $s_i$ vanishes and $r_i$ is the $\ell_2$-norm [2], [3]; (ii) the compressive sensing problem, where $s_i$ is the data-fidelity term, which is often differentiable, and $r_i$ is a sparsity-promoting regularizer such as the $\ell_1$-norm [4], [5]; (iii) optimization problems with per-agent constraints, where $s_i$ is a differentiable objective function of agent $i$ and $r_i$ is the indicator function of the constraint set of agent $i$, that is, $r_i(x) = 0$ if $x$ satisfies the constraint and $\infty$ otherwise [6]–[8].

A. Background and Prior Art

Pioneered by the seminal work [9], [10] in 1980s, decentralized optimization, control, and decision-making in networked multi-agent systems have attracted increasing interest in recent years due to the rapid development of communication and computation technologies [11]–[13]. Different to centralized processing, which requires a fusion center to collect data, decentralized approaches rely on information exchange among neighbors in the network and autonomous optimization by all the individual agents, and are hence robust to failure of critical relaying agents and scalable to the network size. These advantages lead to successful applications of decentralized optimization in robotic networks [14], [15], wireless sensor networks [4], [16], smart grids [17], [18], and distributed machine learning systems [19], [20], just to name a few. In these applications, problem (1) appears as a generic model.

The existing algorithms that solve problem (1) include the primal-dual domain methods such as the de-
centralized alternating direction method of multipliers (DADMM) [16], [21] and the primal domain methods including the distributed subgradient method (DSM) [22]. DADMM reformulates problem (1) in a form to which ADMM becomes a decentralized algorithm. In this algorithm, each agent minimizes the sum of its local objective and a quadratic function that involves local variables from of its neighbors. DADMM does not take advantages of the smooth+nonsmooth structure. In DSM, each agent averages its local variable with those of its neighbors and moves along a negative subgradient direction of its local objective. DSM is computationally cheap but does not take advantages of the smooth+nonsmooth structure either. When the local objectives are Lipschitz differentiable, the recent exact first-order algorithm EXTR A [23] is much faster, yet it cannot handle nonsmooth terms.

The algorithms that consider smooth+nonsmooth objectives in the form of (1) include the following primal-domain methods: the (fast) distributed proximal gradient method (DPGM) [24] and the distributed iterative soft thresholding algorithm (DISTA) [25], [26]. Both DPGM and DISTA consist of a gradient step for the smooth part and a proximal step for the nonsmooth part. DPGM uses two loops where the inner one is dedicated for consensus. The nonsmooth terms of the objective functions of all agents must be the same. DISTA is exclusively designed for compressed sensing problems and is not a simpler algorithm that does not explicitly use any dual variable. We establish convergence under weaker conditions and show that the residual of the first-order optimality condition reduces at the rate of $O(\frac{1}{k})$. Clearly, PG-EXTRA extends EXTRA [23].

When $r_i \equiv 0$, the proposed algorithm PG-EXTRA reduces to EXTRA [23]. Clearly, PG-EXTRA extends EXTRA to handle nonsmooth objective terms. This extension is not the same as the extension from the gradient method to the proximal-gradient method. As the reader will see, PG-EXTRA will have two interlaced sequences of iterates, whereas the proximal-gradient method inherits the sequence of the iterates in the gradient method.

### B. Paper Organization and Contributions

Section II of this paper develops PG-EXTRA, which takes advantages of the smooth+nonsmooth structure of the objective functions. The details are given in Section II-A. The special cases of PG-EXTRA are discussed in Section II-B. In particular, it reduces to a new algorithm P-EXTRA when all $s_i \equiv 0$ and the gradient (or the “G”) steps are no longer needed.

Section III establishes the convergence and derives the rates for PG-EXTRA and P-EXTRA. Under the Lipschitz assumption of $\nabla s_i$, the iterates of PG-EXTRA converge to a solution and the first-order optimality condition asymptotically holds at an ergodic rate of $O(\frac{1}{k})$. The rate improves to non-ergodic $o(\frac{1}{k})$ for P-EXTRA.

The performance of PG-EXTRA and P-EXTRA is numerically evaluated in Section IV, on a decentralized geometric median problem (Section IV-A), a decentralized compressive sensing problem (Section IV-B), and a decentralized quadratic program (Section IV-C). Simulation results confirm theoretical findings and validate the competitiveness of the proposed algorithms.

We have not yet found ways to further improve the convergence rates or theoretically grounded methods to relax our algorithms for stochastic or asynchronous steps, though some numerical experiments with modified algorithms appeared to be successful.

### C. Notation

Each agent $i$ hold a local variable $x(i) \in \mathbb{R}^p$, whose value at iteration $k$ is denoted by $x^k(i)$. We introduce an objective function that aggregates all the local terms as

$$f(x) \triangleq \sum_{i=1}^{n} f_i(x(i)),$$

where

$$x \triangleq \begin{pmatrix} -x^{T}_1 & \cdot & \cdot & \cdot & -x^{T}_n \end{pmatrix} \in \mathbb{R}^{n \times p}.$$

The $i$th row of $x$ corresponds to agent $i$. We say that $x$ is *consensual* if all of its rows are identical, i.e., $x(1) = \cdots = x(n)$. Similar to the definition of $f(x)$, we define

$$s(x) \triangleq \sum_{i=1}^{n} s_i(x(i)) \quad \text{and} \quad r(x) \triangleq \sum_{i=1}^{n} r_i(x(i)).$$

By definition, $f(x) = s(x) + r(x)$.

The gradient of $s$ at $x$ is given by

$$\nabla s(x) \triangleq \begin{pmatrix} - (\nabla s_1(x(1))^T & \cdots & - (\nabla s_n(x(n))^T \end{pmatrix} \in \mathbb{R}^{n \times p}.$$

where $\nabla s_i(x(i))$ is the gradient of $s_i$ at $x(i)$. We let $\nabla r(x)$
denote a subgradient of \( \nabla r \) at \( x \):

\[
\nabla r(x) \triangleq \begin{pmatrix}
- (\nabla r_1(x(1)))^T \\
- (\nabla r_2(x(2)))^T \\
\vdots \\
- (\nabla r_n(x(n)))^T
\end{pmatrix} \in \mathbb{R}^{n \times p},
\]

where \( \nabla r_i(x(i)) \) is a subgradient of \( r_i \) at \( x_i \). The \( i \)th row of \( x \), \( \nabla s(x) \), and \( \nabla r(x) \) belongs to agent \( i \).

In the proposed algorithm, agent \( i \) needs to compute the proximal map of \( r_i \) in the form

\[
\min_{x \in \mathbb{R}^p} r_i(x) + \frac{1}{2\alpha} \| x - y \|_2^2,
\]

where \( y \in \mathbb{R}^p \) is a proximal point and \( \alpha > 0 \) is a scalar. We assume that it is easy to compute the proximal map, which often has an explicit solution.

The Frobenius norm of a matrix \( A \) is denoted as \( ||A||_F \). Given a symmetric positive semidefinite matrix \( G \), define the \( G \)-norm: \( ||A||_G \triangleq \sqrt{\text{trace}(A^TG^A)} \). The largest singular value of a matrix \( A \) is denoted as \( \sigma_{\text{max}}(A) \). The largest and smallest eigenvalues of a symmetric matrix \( B \) are denoted as \( \lambda_{\text{max}}(B) \) and \( \lambda_{\text{min}}(B) \), respectively. The smallest nonzero eigenvalue of a symmetric positive semidefinite matrix \( B \) is denoted as \( \lambda_{\text{min}}(B) \). Let \( \text{null}(A) \triangleq \{ x \in \mathbb{R}^n | Ax = 0 \} \) denote the null space of \( A \), and \( \text{span}(A) \triangleq \{ y \in \mathbb{R}^n | y = Ax, \forall x \in \mathbb{R}^n \} \) denote the subspace spanned by the columns of \( A \).

II. ALGORITHM DEVELOPMENT

This section derives PG-EXTRA for problem (1) in Section II-A and discusses its special cases in Section II-B.

A. Proposed Algorithm: PG-EXTRA

PG-EXTRA starts from an arbitrary initial point \( x^0 \in \mathbb{R}^{n \times p} \), that is, each agent \( i \) holds an arbitrary point \( x_i^0 \). The next point \( x^1 \) is generated by a proximal gradient iteration

\[
x^k = W x^k - \alpha \nabla s(x^k), \quad (2a)
\]

\[
x^k = \arg \min_x \left( r(x) + \frac{1}{2\alpha} \| x - x^k \|_F^2 \right), \quad (2b)
\]

where \( \alpha \in \mathbb{R} \) is the step size and \( W = [w_{ij}] \in \mathbb{R}^{n \times n} \) is the mixing matrix which we will discuss later. All the subsequent points \( x^2, x^3, \ldots \) are obtained through the following update: \( k = 0, 1, \ldots \)

\[
x^{k+1} = W x^{k+1} + x^{k+\frac{1}{2}} - W x^k
\]

\[
- \alpha (\nabla s(x^{k+1}) - \nabla s(x^k)), \quad (3a)
\]

\[
x^{k+2} = \arg \min_x \left( r(x) + \frac{1}{2\alpha} \| x - x^{k+1+\frac{1}{2}} \|_F^2 \right). \quad (3b)
\]

In (3a), \( \tilde{W} = [\tilde{w}_{ij}] \in \mathbb{R}^{n \times n} \) is another mixing matrix, which we typically set as \( \frac{W + \frac{1}{2} I}{2} \) though there are more general choices. With that typical choice, \( \tilde{W} x = \frac{W x + x}{2} \) can be easily computed from \( W x \). PG-EXTRA is outlined in Algorithm 1, where the computation for all individual agents is presented.

Algorithm 1: PG-EXTRA

Set mixing matrices \( W \in \mathbb{R}^{n \times n} \) and \( \tilde{W} \in \mathbb{R}^{n \times n} \);

1. Choose step size \( \alpha > 0 \);

2. for \( k = 0, 1, \ldots, n \) do

3. \( x_i^{k+1} = \arg \min_{x_i} \left( r_i(x_i) + \frac{1}{2\alpha} \| x_i - x_i^k \|_2^2 \right) \)

4. end for

We will require \( w_{ij} = 0 \) and \( \tilde{w}_{ij} = 0 \) if \( i, j \) are not neighbors and \( i \neq j \). Then, the terms like \( \sum_{j=1}^n w_{ij} x_i^k \) and \( \sum_{j=1}^n \tilde{w}_{ij} x_i^k \) only involve \( x_i \), as well as \( x_j \) that are from the neighbors \( j \) of agent \( i \). All the other terms use only local information.

We impose the following assumptions on \( W \) and \( \tilde{W} \).

Assumption 1 (Mixing matrices). Consider a connected network \( G = (V, E) \) consisting of a set of agents \( V = \{1, 2, \ldots, n\} \) and a set of undirected edges \( E \). An unordered pair \( (i, j) \in E \) if agents \( i \) and \( j \) have a direct communication link. The mixing matrices \( W = [w_{ij}] \in \mathbb{R}^{n \times n} \) and \( \tilde{W} = [\tilde{w}_{ij}] \in \mathbb{R}^{n \times n} \) satisfy

1. (Decentralization property) If \( i \neq j \) and \( (i, j) \notin E \), then \( w_{ij} = \tilde{w}_{ij} = 0 \).

2. (Symmetry property) \( W = W^T \), \( \tilde{W} = \tilde{W}^T \).

3. (Null space property) \( \text{null}(W - \tilde{W}) = \text{span}\{1\} \); \( \text{null}(I - W) \supseteq \text{span}\{1\} \).

4. (Spectral property) \( \tilde{W} > 0 \) and \( \frac{W + \frac{1}{2} I}{2} \succeq \tilde{W} \succeq W \).

The first two conditions together are standard (see [22], for example). The first condition alone ensures communications to occur between neighbor agents. All the four conditions together ensure that \( W \) satisfies \( \lambda_{\text{max}}(W) = 1 \) and its other eigenvalues lie in \(( -1, 1 ) \). Typical choices of \( W \) can be found in [23], [28]. If a matrix \( W \) satisfy the all conditions, then \( W = \frac{W + \frac{1}{2} I}{2} \) also satisfies the conditions.

B. Special Cases: EXTRA and P-EXTRA

When the possibly-nondifferentiable term \( r = 0 \), we have \( x^1 = x^2 \) in (2a) and (2b) and thus

\[
x^1 = W x^0 - \alpha \nabla s(x^0).
\]

(4)
In (3a) and (3b), we have \( x^{k+2} = x^{k+1} + \frac{1}{2} \) and thus
\[
x^{k+2} = Wx^{k+1} + x^{k+1} - Wx^k - \alpha [\nabla s(x^{k+1}) - \nabla s(x^k)].
\]

The updates (4) and (5) are known as EXTRA, a recent algorithm for decentralized differentiable optimization [23].

When the differentiable term \( s = 0 \), PG-EXTRA reduces to P-EXTRA by removing all gradient computation, which is given in **Algorithm 2**.

III. CONVERGENCE ANALYSIS

A. Preliminaries

Unless otherwise stated, the convergence results in this section are given under Assumptions 1–3.

**Assumption 2** (Convex objective functions and the smooth parts having Lipschitz gradients). For all \( i = 1, \ldots, n \), functions \( r_i \) and \( s_i \) are proper closed convex and \( s_i \) satisfy

\[
\| \nabla s_i(x) - \nabla s_i(y) \| \leq L_{s_i} \| x - y \|, \quad \forall x, y \in \mathbb{R}^p,
\]

where \( L_{s_i} > 0 \) are constant.

Following Assumption 2, \( f(x) = \sum_{i=1}^{n} f_i(x(i)) \) is proper closed convex and \( \nabla s \) satisfies

\[
\| \nabla s(x) - \nabla s(y) \| \leq L_s \| x - y \|, \quad \forall x, y \in \mathbb{R}^{n \times p},
\]

with constant \( L_s = \max_i \{ L_{s_i} \} \).

**Assumption 3** (Solution existence). The set of solution(s) \( X^* \) to problem (1) is nonempty.

We first give a lemma on the first-order optimality condition of problem (1).

**Lemma 1** (First-order optimality conditions). Given mixing matrices \( W \) and \( W_r \) and the economical-form singular value decomposition \( \hat{W} - W = VSVT \), define

\[
U \triangleq V S^{1/2} VT = (\hat{W} - W)^{1/2} \in \mathbb{R}^{n \times n}.
\]

Then, under Assumptions 1–3, the following two statements are equivalent

- \( x^* \in \mathbb{R}^{n \times n} \) is consensual, that is, \( x^*_{(1)} = x^*_{(2)} = \cdots = x^*_{(n)} \) and every \( x^*_{(i)} \) is optimal to problem (1);
- There exists \( q^* = U p \) for some \( p \in \mathbb{R}^{n \times p} \) and subgradient \( \nabla r(x^*) \in \partial r(x^*) \) such that

\[
\begin{align*}
U q^* + \alpha (\nabla s(x^*) + \nabla r(x^*)) &= 0, \quad (6a) \\
U x^* &= 0. \quad (6b)
\end{align*}
\]

**Proof**: According to Assumption 1 and the definition of \( U \), we have

\[
null(U) = null(V^T) = null(\hat{W} - W) = span \{ 1 \}.
\]

Hence \( x \) is consensual if and only if (6b) holds.

Next, any row of the consensual \( x^* \) is optimal if and only if \( 1^T (\nabla s(x^*)^T + \nabla r(x^*)) = 0 \). Since \( U \) is symmetric and \( U^T = 0 \), (6a) gives \( 1^T (\nabla s(x^*) + \nabla r(x^*)) = 0 \). Conversely, if \( 1^T (\nabla s(x^*) + \nabla r(x^*)) = 0 \), then \( \nabla s(x^*) + \nabla r(x^*) \in span \{ U \} \) follows from null \( U \) = (span \{ 1 \})\text{and thus } \alpha (\nabla s(x^*) + \nabla r(x^*)) = -U q \) for some \( q \). Let \( q^* = \text{Proj}_U q \). Then, \( U q^* = U q \) and (6a) holds.

Let \( x^* \) and \( q^* \) satisfy the optimality conditions (6a) and (6b). Introduce an auxiliary sequence

\[
q^k \triangleq \sum_{t=0}^{k} U x^t.
\]

The next lemma restates the updates of PG-EXTRA in terms of \( x^k, q^k, x^*, q^* \) for convergence analysis.

**Lemma 2** (Recursive relations of PG-EXTRA). In PG-EXTRA, the quadraple sequence \( \{ x^k, q^k, x^*, q^* \} \) obeys

\[
\begin{align*}
(I + W - 2\hat{W}) x^{k+1} + \hat{W} (x^{k+1} - x^k) &= -U q^{k+1} - \alpha \nabla s(x^k) - \alpha \nabla r(x^{k+1}) \quad (7) \\
(I + W - 2\hat{W}) (x^{k+1} - x^k) + \hat{W} (x^{k+1} - x^k) &= -U (q^{k+1} - q^*) - \alpha [ \nabla s(x^k) - \nabla s(x^*)] - \alpha [ \nabla r(x^{k+1}) - \nabla r(x^k)] \quad (8)
\end{align*}
\]

for any \( k = 0, 1, \ldots \).

**Proof**: By giving the first-order optimality conditions of the subproblems (2b) and (3b) and eliminating the auxiliary sequence \( x^{k+2} \) for \( k = 0, 1, \ldots \), we have the following equivalent subgradient recursions with respect to \( x^k \):

\[
x^1 &= W x^0 - \alpha \nabla s(x^0) - \alpha \nabla r(x^1), \\
x^{k+1} &= (I + W)x^k - W x^{k-1} - \alpha [ \nabla s(x^k) - \nabla s(x^{k-1})] - \alpha [ \nabla r(x^{k+1}) - \nabla r(x^k)], \quad k = 1, 2, \ldots .
\]
Summing these subgradient recursions over times 1 through \( k+1 \), we get
\[
\begin{align*}
\mathbf{x}^{k+1} &= \hat{W}\mathbf{x}^k - \sum_{t=0}^{k} (\hat{W} - W)\mathbf{x}^t \\
&\quad - \alpha \nabla s(\mathbf{x}^k) - \alpha \nabla r(\mathbf{x}^{k+1}).
\end{align*}
\] (9)

Using \( q^{k+1} = \sum_{t=0}^{k} U\mathbf{x}^t \) and the decomposition \( \hat{W} - W = U^2 \), (7) follows from (9) immediately.

Since \((I + W - 2\hat{W})1 = 0\), null\(U\) = \((\text{span}\{1\})\) and \(U\mathbf{x}^* = 0\), we have
\[
(I + W - 2\hat{W})\mathbf{x}^* = 0,
\] (10)
Subtracting (10) from (7) and adding \(0 = Uq^* + \alpha(\nabla s(x^*) + \nabla r(x^*))\) to (7), we obtain (8). \(\blacksquare\)

The recursive relations of P-EXTRA are shown in the following corollary of Lemma 2.

**Corollary 1 (Recursive relations of P-EXTRA).** In P-EXTRA, the quadruple sequence \(\{\mathbf{x}^k, q^k, x^*, q^k\}\) obeys
\[
\begin{align*}
(I + W - 2\hat{W})\mathbf{x}^{k+1} + W(\mathbf{x}^{k+1} - \mathbf{x}^k) &= -Uq^{k+1} - \alpha \nabla r(\mathbf{x}^{k+1}) \tag{11}
\end{align*}
\]
and
\[
\begin{align*}
(I + W - 2\hat{W})(\mathbf{x}^{k+1} - \mathbf{x}^k) &= -U(q^{k+1} - q^*) - \alpha(\nabla r(\mathbf{x}^{k+1}) - \nabla r(\mathbf{x}^*)) \tag{12}
\end{align*}
\]
for any \(k = 0, 1, \ldots\).

The convergence analysis of PG-EXTRA is based on the recursions (7) and (8) and that of P-EXTRA based on (11) and (12). Define
\[
\mathbf{z}^k \triangleq \begin{pmatrix} q^k \\ x^k \end{pmatrix}, \quad \mathbf{z}^* \triangleq \begin{pmatrix} q^* \\ x^* \end{pmatrix}, \quad G \triangleq \begin{pmatrix} I & 0 \\ 0 & \hat{W} \end{pmatrix}.
\]
For PG-EXTRA, we show that \(\mathbf{x}^k\) converges to a solution \(\mathbf{x}^*\) and the successive iterative difference \(\|\mathbf{z}^k - \mathbf{z}^{k+1}\|_G^2\) converges to 0 at an ergodic \(O\left(\frac{1}{k}\right)\) rate (see Theorem 2); the same ergodic rates hold for the first-order optimality residuals, which are defined in Theorem 2. For the special case, P-EXTRA, \(x^k\) also converges to an optimal solution \(x^*\); the progress \(\|z^k - z^{k+1}\|_G^2\) and the first-order optimality residuals converge to 0 at improved non-ergodic \(o\left(\frac{1}{k}\right)\) rates (see Theorem 3).

**B. Convergence and Convergence Rates of PG-EXTRA**

1) **Convergence of PG-EXTRA:** We first give a theorem that shows the contractive property of PG-EXTRA. This theorem provides a sufficient condition for PG-EXTRA to converge to a solution. In addition, it prepares for analyzing convergence rates of PG-EXTRA in subsection III-B2 and its limit case gives the contractive property of P-EXTRA (see Section III-C).

**Theorem 1.** Under Assumptions 1–3, if we set the step size \(\alpha \in \left(0, \frac{2\lambda_{\min}(\hat{W})}{L}\right)\), then the sequence \(\{z^k\}\)

\[
\|z^k - z^*\|^2_G - \|z^{k+1} - z^*\|^2_G \geq \zeta \|z^k - z^{k+1}\|^2_G,
\] (13)

\(k = 0, 1, \ldots\),

where \(\zeta = 1 - \frac{\alpha L_s}{2\lambda_{\min}(\hat{W})}\). Furthermore, \(z^k\) converges to an optimal \(z^*\).

**Proof:** By Assumption 2, \(s\) and \(r\) are convex, and \(\nabla s\) is Lipschitz continuous with constant \(L_s\), we have
\[
\begin{align*}
\frac{2\alpha}{L_s} \|\nabla s(x^k) - \nabla s(x^*)\|_F^2 &\leq 2\alpha(x^k - x^*, \nabla s(x^k) - \nabla s(x^*)) \\
&\quad + 2\alpha(x^k - x^{k+1}, \nabla s(x^k) - \nabla s(x^*)) \tag{14}
\end{align*}
\]
and
\[
0 \leq 2(x^k - x^*, \alpha(\nabla r(x^k) - \nabla r(x^*))).
\] (15)

Substituting (8) from Lemma 2 for \(\alpha(\nabla s(x^k) - \nabla s(x^*)) + \alpha(\nabla r(x^k) - \nabla r(x^*))\), it follows from (14) and (15) that
\[
\begin{align*}
\frac{2\alpha}{L_s} \|\nabla s(x^k) - \nabla s(x^*)\|_F^2 &\leq 2(x^k - x^*, \alpha(\nabla s(x^k) - \nabla s(x^*)) \\
&\quad + 2\alpha(x^k - x^{k+1}, \nabla s(x^k) - \nabla s(x^*)) \\
&\quad + 2\alpha(x^k - x^*, \nabla s(x^k) - \nabla s(x^*)) \\
&\quad + 2\alpha(x^k - x^{k+1}, \nabla s(x^k) - \nabla s(x^*))) \tag{16}
\end{align*}
\]

For the terms on the right-hand side of (16), we have
\[
\begin{align*}
2(x^k - x^*, U(q^* - q^{k+1})) &= 2(x^k - x^*, U(q^* - q^{k+1})) \\
&= 2(x^k - x^*, \hat{W}(x^k - x^{k+1})) \\
&= 2(x^k - x^*, \hat{W}(x^k - x^{k+1})) \\
&= 2(x^k - x^*, \hat{W}(x^k - x^{k+1})) \tag{17}
\end{align*}
\]
and
\[
2\alpha(x^k - x^{k+1}, \nabla s(x^k) - \nabla s(x^*)) \leq \frac{\alpha L_s}{2} \|x^k - x^{k+1}\|^2_F + \frac{\alpha L_s}{2} \|\nabla s(x^k) - \nabla s(x^*)\|^2_F. \tag{19}
\]

Plugging (17) into (16), we have
\[
\frac{2\alpha}{L_s} \|\nabla s(x^k) - \nabla s(x^*)\|_F^2 \leq 2(x^k - x^*, U(q^* - q^{k+1})) + 2\alpha(x^k - x^{k+1}, \hat{W}(x^k - x^{k+1})) - 2\|x^k - x^{k+1}\|^2_{\hat{W}} - 2\alpha L_s \|x^k - x^{k+1}\|^2_F + 2\alpha L_s \|\nabla s(x^k) - \nabla s(x^*)\|^2_F. \tag{20}
\]

Using the definitions of \(z^k, z^*\) and \(G\), (20) is equivalent to
\[
0 \leq 2(z^k - z^*, G(z^* - z^{k+1})) - 2\|z^k - z^*\|^2_{\hat{W}} - 2\alpha L_s \|x^k - x^{k+1}\|^2_F. \tag{21}
\]
Applying the basic equality
\[2(z^{k+1} - z^k, G(z^* - z^{k+1})) = \|z^k - z^*\|_G^2 - \|z^{k+1} - z^*\|_G^2 - \|z^k - z^{k+1}\|_G^2,\]
to (21), we have
\[0 \leq \|z^k - z^*\|_G^2 - \|z^{k+1} - z^*\|_G^2 - \|z^k - z^{k+1}\|_G^2 - 2\|x^{k+1} - x^k\|_G^2 + \frac{\alpha_L}{2}\|z^k - x^{k+1}\|_G^2.\]

By Assumption 1, in particular, \(I + W - 2\tilde{W} \geq 0\), we have
\[\|x^{k+1} - x^k\|_G^2 \geq 0\] and thus
\[\|z^k - z^*\|_G^2 - \|z^{k+1} - z^*\|_G^2 \geq \zeta\|z^k - z^{k+1}\|_G^2,\]
where \(\zeta = 1 - \frac{\alpha_L}{2\lambda_{\text{min}}(W)} > 0\). The last inequality holds since \(\alpha < \frac{2\lambda_{\text{min}}(W)}{\lambda_L}\).

It shows from (24) that for any optimal solution \(z^*\), \(\|z^k - z^*\|^2_G\) is bounded and contractive. Therefore, \(\|z^k - z^*\|^2_G\) is converging as long as \(\|z^k - z^{k+1}\|^2_G \to 0\). The convergence of \(z^k\) to an optimal solution \(z^*\) follows from the standard analysis for contraction methods; see, for example, Theorem 3 in [29].

2) Ergodic \(O \left( \frac{1}{k} \right)\) Rates of PG-EXTRA: To establish rate of convergence, we need the following proposition. Parts of it appeared in recent works [30], [31].

**Proposition 1.** If a sequence \(\{a_k\} \subset \mathbb{R}\) obeys: (1) \(a_k \geq 0\) and (2) \(\sum_{t=1}^{\infty} a_t < \infty\), then we have: (i) \(\lim_{k \to \infty} a_k = 0\); (ii) \(\frac{1}{k} \sum_{t=1}^{k} a_t = O \left( \frac{1}{k} \right)\); (iii) \(\min_{t \leq k} \{a_t\} = o \left( \frac{1}{k} \right)\).

If the sequence \(\{a_k\}\) further obeys: (3) \(a_{k+1} \leq a_k\), then in addition, we have: (iv) \(a_k = O \left( \frac{1}{k} \right)\).

**Proof:** Part (i) is obvious. Let \(b_k = \frac{1}{k} \sum_{t=1}^{k} a_t\). By the assumptions, \(kb_k\) is uniformly bounded and obeys
\[\lim_{k \to \infty} kb_k < \infty,\]
from which part (ii) follows. Since \(c_k = \min_{t \leq k} \{a_t\}\) is monotonically non-increasing, we have
\[kc_{2k} = k \min_{t \leq 2k} \{a_t\} \leq \sum_{t=k+1}^{2k} a_t.\]
This and the fact that \(\lim_{k \to \infty} \sum_{t=k+1}^{2k} a_t = 0\) give us \(c_k = o \left( \frac{1}{k} \right)\) or part (iii).

If \(a_k\) is further monotonically non-increasing, we have
\[ka_{2k} \leq \sum_{t=k+1}^{2k} a_t.\]
This and the fact that \(\lim_{k \to \infty} \sum_{t=k+1}^{2k} a_t = 0\) gives us part (iv).

This proposition serves for the proof of Theorem 2, as well as that of Theorem 3 appearing in Section III-C. We give the ergodic \(O \left( \frac{1}{k} \right)\) convergence rates of PG-EXTRA below.

**Theorem 2.** In the same setting of Theorem 1, the following rates hold for PG-EXTRA:

(i) Running-average successive difference:
\[\frac{1}{k} \sum_{t=1}^{k} \|z^t - z^{t+1}\|^2_G = O \left( \frac{1}{k} \right);\]

(ii) Running-best successive difference:
\[\min_{t \leq k} \{\|z^t - z^{t+1}\|^2_G\} = o \left( \frac{1}{k} \right);\]

(iii) Running-average optimality residuals:
\[\frac{1}{k} \sum_{t=1}^{k} \|Uq_t + \alpha(\nabla s(x^t) + \nabla r(x^{t+1}))\|^2_W = O \left( \frac{1}{k} \right);\]

(iv) Running-best optimality residuals:
\[\min_{t \leq k} \{\|Uq^t + \alpha(\nabla s(x^t) + \nabla r(x^{t+1}))\|^2_W\} = o \left( \frac{1}{k} \right).\]

Before proving the theorem, let us explain the rates. The first two rates on the squared successive difference are used to deduce the last two rates on the optimality residuals. Since our algorithm does not guarantee to reduce objective functions in a monotonic manner, we choose to establish our convergence rates in terms of optimality residuals, which show how quickly the residuals to the KKT system (6) reduce. Note that the rates are given on the standard squared quantities since they are summable and naturally appear in the convergence analysis. With particular note, these \(\frac{1}{k}\) rates match those on the squared successive difference and optimality residual in the classical (centralized) gradient-descent method.

**Proof:** Parts (i) and (ii): Since \(\|z^k - z^*\|^2_G\) converges to 0 when \(k\) goes to \(\infty\), we are able to sum (13) in Theorem 1 over \(k = 0\) through \(\infty\) and apply the telescopic cancellation, which yields
\[\sum_{t=0}^{\infty} \|z^t - z^{t+1}\|^2_G = \frac{1}{\zeta} \sum_{t=0}^{\infty} \left(\|z^t - z^*\|^2_G - \|z^{t+1} - z^*\|^2_G\right)\]
\[= \frac{\|z_0 - z^*\|^2_G}{\zeta} < \infty.\]

Then, the results follow from Proposition 1 immediately.
Parts (iii) and (iv): Using the basic inequality \( \|a + b\|_2^2 \geq \frac{1}{\rho} \|a\|_2^2 - \frac{1}{\rho} \|b\|_2^2 \), which holds for any \( \rho > 1 \) and any matrices \( a \) and \( b \) of the same size, it follows that
\[
\begin{align*}
\|z^k - z^{k+1}\|_G^2 &= \|q^k - q^{k+1}\|_2^2 + \|x^k - x^{k+1}\|_W^2 \\
&= \|x^{k+1}\|_W^2 - W + \|(I - W)x^k + Uq^k + \alpha(\nabla s(x^k) + \nabla r(x^{k+1}))\|_W^2 \\
&\geq \|x^{k+1}\|_W^2 - W + \frac{1}{\rho} \|Uq^k + \alpha(\nabla s(x^k) + \nabla r(x^{k+1}))\|_W^2 \\
&= \frac{1}{\rho - 1} \|(I - W)x^k\|_W^2.
\end{align*}
\]

(26)

Since \( W - W \) and \( (I - W) \tilde{W}(I - W) \) are symmetric and
\[
\text{null} \{W - W\} \subseteq \text{null} \{(I - \tilde{W})(I - \tilde{W})\},
\]
there exists a bounded \( v > 0 \) such that
\[
\|(I - \tilde{W})x^k\|_W^2 = \|x^k\|_{(I-\tilde{W})(I-\tilde{W})}^2 \leq v\|x^k\|_{W-W}^2.
\]

It follows from (26) that
\[
\begin{align*}
\frac{1}{k} \sum_{t=1}^{k} \|z^t - z^{t+1}\|_G^2 &= \|x^t\|_{(I-\tilde{W})(I-\tilde{W})}^2 - v\frac{1}{k} \|x^t\|_{W-W}^2 \\
\geq \frac{1}{k} \sum_{t=1}^{k} \left( \|x^t\|_{W-W}^2 - \frac{v}{\rho - 1} \|x^t\|_{W-W}^2 \right) \\
&\geq \frac{1}{k} \sum_{t=1}^{k} \frac{1}{\rho} \|Uq^t + \alpha(\nabla s(x^t) + \nabla r(x^{t+1}))\|_{W-W}^2
\end{align*}
\]

(27)

As part (i) shows that
\[
\frac{1}{k} \sum_{t=1}^{k} \|z^t - z^{t+1}\|_G = O \left( \frac{1}{k} \right),
\]
we have
\[
\frac{1}{k} \sum_{t=1}^{k} \|Uq^t + \alpha(\nabla s(x^t) + \nabla r(x^{t+1}))\|_{W-W}^2 = O \left( \frac{1}{k} \right)
\]

From (27) and (25), we see that both \( \|Uq^k + \alpha(\nabla s(x^k) + \nabla r(x^{k+1}))\|_{W-W}^2 \) and \( \|Ux^k\|_W^2 \) are summable. Again, by Proposition 1, we have part (iv), the \( o \left( \frac{1}{k} \right) \) rates of the running best first-order optimality residuals.

The monotonicity of \( \|z^k - z^{k+1}\|_G^2 \) is an open question. If it holds, then \( o \left( \frac{1}{k} \right) \) convergence rates will apply to the sequence itself.

C. Convergence Rates of P-EXTRA

Convergence of P-EXTRA follows from that of PG-EXTRA directly. Since P-EXTRA is a special case of PG-EXTRA and its updates are free of gradient steps, it enjoys slightly better convergence rates: non-ergodic \( o \left( \frac{1}{k} \right) \). Let us brief on our steps. First, as a special case of Theorem 1 by letting \( L_s \to 0^+ \), the sequence \( \{\|z^k - z^{k+1}\|_G^2\} \) is summable. Second, the sequence \( \{\|z^k - z^{k+1}\|_G^2\} \) of P-EXTRA is shown to be monotonic in Lemma 3. Based on these results, the non-ergodic \( o \left( \frac{1}{k} \right) \) convergence rates are then established for successive difference and first-order optimality residuals.

Lemma 3. Under the same assumptions of Theorem 1 except \( s(x) = 0 \), for any step size \( \alpha > 0 \), the sequence \( \{z^k\} \) generated by P-EXTRA satisfies
\[
\|z^{k+2} - z^{k+1}\|_G^2 \leq \|z^k - z^{k+1}\|_G^2, \quad k = 0, 1, \cdots.
\]

(28)

Proof: To simplify the description of the proof, define \( \Delta x^{k+1} = x^k - x^{k+1} \), \( \Delta q^{k+1} = q^k - q^{k+1} \), \( \Delta z^{k+1} = z^k - z^{k+1} \), and \( \Delta \nabla r(x^{k+1}) = \nabla r(x^k) - \nabla r(x^{k+1}) \). By convexity of \( r \) in Assumption 2, we have
\[
\langle \Delta x^{k+1}, \Delta \nabla r(x^{k+1}) \rangle \geq 0.
\]

(29)

Taking difference of (7) at the \( k \)-th and \( (k + 1) \)-th iterations yields
\[
\alpha \Delta \nabla r(x^{k+1}) + U \Delta q^{k+1} + (I + W - 2\tilde{W}) \Delta x^{k+1} + W(\Delta x^{k+1} - \Delta x^k) = 0.
\]

(30)

Combine (29) and (30) it follows that
\[
\langle \Delta x^{k+1} - U \Delta q^{k+1}, \Delta x^{k+1} - U \Delta q^{k+1} \rangle
\]
\[
+ (\Delta x^{k+1} - U \Delta q^{k+1} - I + W - 2\tilde{W})(\Delta x^{k+1} - \Delta x^k) \geq \|\Delta x^{k+1}\|_{I+W-2\tilde{W}}^2.
\]

(31)

Using the definition of \( q^k \), \( \Delta q^{k+1} = -U\Delta x^{k+1} \). Thus, we have
\[
\Delta q^{k+1} = -U\Delta x^{k+1}.
\]

(32)

Substituting (32) into (31) yields
\[
\langle \Delta q^{k+1} - \Delta q^{k+1}, \Delta q^{k+1} \rangle
\]
\[
+ (\Delta x^{k+1} - U \Delta q^{k+1} - I + W - 2\tilde{W})(\Delta x^{k+1} - \Delta x^k) \geq \|\Delta x^{k+1}\|_{I+W-2\tilde{W}}^2.
\]

(33)

or equivalently
\[
\langle \Delta x^{k+1} - \Delta x^k, G \Delta x^{k+1} \rangle \geq \|\Delta x^{k+1}\|_{I+W-2\tilde{W}}^2.
\]

(34)

By applying the basic equality
\[
2(\Delta x^k - \Delta x^{k+1}, G \Delta z^{k+1}) = \|\Delta x^k\|_G^2 - \|\Delta z^{k+1}\|_G^2 - \|\Delta z^k - \Delta z^{k+1}\|_G^2, \quad \text{to (34)},
\]

we finally have
\[
\|\Delta x^k\|_G^2 - \|\Delta z^{k+1}\|_G^2 \geq \|\Delta z^k - \Delta z^{k+1}\|_G^2 + 2\|\Delta x^{k+1}\|_{I+W-2\tilde{W}}^2 \geq 0,
\]

(35)

which implies (28) and completes the proof.

The next theorem gives the \( o \left( \frac{1}{k} \right) \) convergence rates of P-EXTRA. Its proof is omitted as it is similar to that of Theorem 2. The only difference is that \( \{\|z^k - z^{k+1}\|_G^2\} \) has been shown monotonic in P-EXTRA. Invoking fact
Fig. 1. The underlying network for the experiments.

(iv) in Proposition 1, the rates are improved from ergodic $O\left(\frac{1}{\sqrt{t}}\right)$ to non-ergodic $o\left(\frac{1}{\sqrt{t}}\right)$.

**Theorem 3.** In the same setting of Lemma 3, the following rates hold for P-EXTRA:

(i) Successive difference:

$$\|z^k - z^{k+1}\|^2_G = O\left(\frac{1}{k}\right);$$

(ii) First-order optimality residuals:

$$\|Uq^k + \alpha \nabla r(x^k)\|^2_{W^{-1}} = O\left(\frac{1}{k}\right),$$

$$\|Ux^k\|^2 = O\left(\frac{1}{k}\right).$$

**Remark 1** (Less restriction on step size). We can see from Section III-C that P-EXTRA accepts a larger range of step size than PG-EXTRA.

### IV. Numerical Experiments

In this section, we provide three numerical experiments, decentralized geometric median, decentralized compressed sensing, and decentralized quadratic programming, to demonstrate the effectiveness of the proposed algorithms. All the experiments are conducted over a randomly generated connected network shown in Fig. 1, which has $n = 10$ agents and $\frac{3n(n-1)}{2} = 18$ edges.

In the numerical experiments, we use the relative error $\|x^k - \bar{x}\|_2$ and the successive difference $\|x^k - x^{k+1}\|^2_F$ as performance metrics; the former is a standard metric to assess the solution optimality and the later evaluates the bounds of the rates proved in this paper.

**A. Decentralized Geometric Median**

Consider a decentralized geometric median problem. Each agent $i \in \{1, \cdots, n\}$ holds a vector $y(i) \in \mathbb{R}^p$, and all the agents collaboratively calculate the geometric median $x \in \mathbb{R}^p$ of all $y(i)$’s. This task can be formulated as solving the following minimization problem:

$$\min_{x} \bar{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \|x - y(i)\|_2.$$

Computing decentralized geometric medians have interesting applications: (i) in [2], the multi-agent system locates a facility to minimize the cost of transportation in a decentralized manner; (ii) in cognitive robotics [32], a group of collaborative robots sets up a rally point such that the overall moving cost is minimal; (iii) in distributed robust Bayesian learning [33], decentralized geometric median is also an important subproblem.

The above problem can further be generalized as the group least absolute deviations problem

$$\min_{x} \bar{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \|M(i)x - y(i)\|_2$$

($M(i)$ is the measurement matrix on agent $i$), which can be considered as a variant of cooperative least squares estimation while being capable of detecting anomalous agents and maintaining the system out of harmful effect caused by agents of collapse.

The geometric median problem is solved by P-EXTRA. The minimization subproblem in P-EXTRA

$$x^{k+2} \leftarrow \arg \min_{x} f(x) + \frac{1}{2\alpha} \|x - x^{k+1}\|^2_F$$

has an explicit solution

$$x^{k+2}_{(i)} = y(i) - \frac{\epsilon^{k+1+\frac{1}{2}}(i)}{\|\epsilon^{k+1+\frac{1}{2}}(i)\|_2} \left(\epsilon^{k+1+\frac{1}{2}}(i)\right)_+,$$

where $\epsilon^{k+1+\frac{1}{2}}(i) \triangleq y(i) - x^{k+1+\frac{1}{2}}(i)$ and $(a)_+ \triangleq \max\{a, 0\}$, $\forall a \in \mathbb{R}$, $\forall i$.

We set $p = 3$, that is, each point $y(i) \in \mathbb{R}^3$. Data $y(i)$ is generated following the uniform distribution in $[-200, 200] \times [-200, 200] \times [-200, 200]$. The algorithm starts from $x^{0}_{(i)} = y(i)$, $\forall i$. We use the Metropolis constant edge weight for $W$ and $\tilde{W} = \frac{1}{2} I + W$, as well as a constant step size $\alpha$.

We compare P-EXTRA to DSM [22] and DADMM [21]. In DSM, at each iteration, each agent combines the local variables from its neighbors with a Metropolis constant edge weight and performs a subgradient step along the negative subgradient of its own objective with a diminishing step size $\alpha = O(k^{-\frac{3}{2}})$. In DADMM, at each iteration, each agent updates its primal local copy by solving an optimization problem and then updates its local dual variable with simple vector operations.
which fails the ordinary least squares. We instead solve
measurement vector, $\sum_i y_i$ equations, where $y_i = M(i)x + e(i)$, and $e(i)$ is a Gaussian noise vector. The goal is to estimate the sparse vector $x$. The number of total measurements $\sum_{i=1}^n m_i$ is often less than the number of unknown $p$, which fails the ordinary least squares. We instead solve an $\ell_1$-regularized least squares problem
\[
\min_x \frac{1}{2} \|x^+ - x^-\|_F \quad \text{subject to} \quad \|x\|_1 \leq \tau.
\]
where $x^+$ and $x^-$ are the positive and negative parts of $x$, respectively. We use PG-EXTRA to solve this problem.

The decentralized compressed sensing is a special case of the general distributed compressed sensing [34] where its intra-signal correlation is ignored. This case appears specifically in cooperative spectrum sensing for cognitive radio networks. More of its applications can be found in [5] and the references therein.

Consider the smooth-nonsmooth structure, PG-EXTRA is applied. In this experiment, each agent $i$ holds $m_i = 3$ measurements and the dimension of $x$ is $p = 50$. Measurement matrices $M(i)$ and noise vectors $e(i)$ are randomly generated, following an i.i.d. Gaussian distribution and $M(i)$ have been normalized to have $\|M(i)\|_2 = 1$. The signal $x$ is randomly generated and has a sparsity of 0.8 (containing 10 nonzero elements). The algorithm starts from $x_0 = 0$, $\forall i$. In PG-EXTRA, we use the Metropolis constant edge weight for $W$ and $\tilde{W} = \frac{W + \tilde{W}}{2}$, and constant step size $\alpha$.

We compare PG-EXTRA with the recent work DISTA [25], [26], which has two free parameters: temperature parameter $q \in (0, 1)$ and $\tau \leq \|M(i)\|_2^2$, $\forall i$. We have hand optimized $q$ and show the effect of $\tau$ in our experiment.

The numerical results are illustrated in Fig. 3. It shows that the relative errors of PG-EXTRA drops to $10^{-5}$ in 1000 iterations while DISTA still has a relative error larger than $10^{-2}$ when it is terminated at 4000 iterations. Both PG-EXTRA and DISTA utilize the smooth-nonsmooth structure of the problem but PG-EXTRA achieves faster convergence.

C. Decentralized Quadratic Programming

We use decentralized quadratic programming as an example to show that how PG-EXTRA solves a constrained optimization problem. Each agent $i \in \{1, \ldots, n\}$ has a local quadratic objective $\frac{1}{2} x^T Q_i x + h_i^T x$ and a local linear constraint $a_i^T x \leq b_i$, where the symmetric positive semidefinite matrix $Q_i \in \mathbb{R}^{p \times p}$, the vectors $b_i \in \mathbb{R}^p$ and $a_i \in \mathbb{R}^p$, and the scalar $b_i \in \mathbb{R}$ are stored at agent $i$. The agents collaboratively minimize the average of the local objectives subject to all local constraints. The quadratic program is:
\[
\min_x \frac{1}{n} \sum_{i=1}^n s_i(x) = \frac{1}{n} \sum_{i=1}^n \left( \frac{1}{2} x^T Q_i x + h_i^T x \right),
\]
s.t. $a_i^T x \leq b_i$, $\forall i = 1, \ldots, n$. 

DADMM has a penalty coefficient $c$ as its parameter. We have hand-optimized this parameter and the step sizes for DADMM, DSM, and P-EXTRA.

The numerical results are illustrated in Fig. 2. It shows that the relative errors of DADMM and P-EXTRA both drop to $10^{-8}$ in 100 iterations while DSM has a relative error of larger than $10^{-2}$ before 400 iterations. P-EXTRA is better than DSM because it utilizes the problem structure, which is ignored by DSM. In this case, both P-EXTRA and DADMM can be considered as proximal point algorithms and thus have similar convergence performance.
data $a_i$ and $b_i$ but guarantee that the feasible set is nonempty and the optimal solution to the problem (36) is different to the optimization problem with the same objective of (36) but without constraints $a_i^T x \leq b_i$, $\forall i$. In this way we can make sure at least one of the constraints $a_i^T x \leq b_i$, $\forall i$ is activated. In PG-EXTRA, we use the Metropolis constant edge weight for $W$ and $\tilde{W} = \frac{1}{2} W$, and constant step size $\alpha$.

The numerical experiment result is illustrated in Fig. 4. We compare PG-EXTRA with two distributed subgradient projection methods (DSPMs) [35] (denoted as DSPM1 and DSPM2 in Fig. 4). DSPM1 assumes that each agent knows all the constraints $a_i^T x \leq b_i$, $\forall i$ so that DSPM can be applied to solve (36). The iteration of DSPM1 at each agent $i$ is $x_i^{k+1} = P_\Omega \left( \sum_{j=1}^n w_{ij} x_j^k - \alpha k \nabla s_i(x_i^k) \right)$ where the set $\Omega = \{ \tilde{x} | a_i^T \tilde{x} \leq b_i, \forall i \}$ and $P_\Omega (\cdot)$ stands for projection on to $\Omega$. The projection step employs the alternating projection method [8] and its computation cost is high. To address this issue, we modify DSPM1 to DSPM2 with $x_i^{k+1} = P_\Omega \left( \sum_{j=1}^n w_{ij} x_j^k - \alpha k \nabla s_i(x_i^k) \right)$, where $\Omega_i = \{ \tilde{x} | a_i^T \tilde{x} \leq b_i \}$. DSPM2 is likely to be convergent but has no theoretical guarantee. Both DSPM1 and DSPM2 use diminishing step size $\alpha_k = O(k^{-\frac{3}{2}})$ and we hand-optimize the initial step sizes.

It is shown in Fig. 4 that the relative errors of PG-EXTRA drops to $10^{-4}$ in 4000 iterations while DSPM1 and DSPM2 still have relative errors of larger than $10^{-1}$ when they are terminated at 20000 iterations. PG-EXTRA is better than DSPM1 and DSPM2 because it utilizes the specific problem structure.

V. CONCLUSION

This paper attempts to solve a broad class of decentralized optimization problems with local objectives in the smooth-non-smooth form by extending the recent method EXTRA, which integrates gradient descent with consensus averaging. We proposed PG-EXTRA, which inherits most properties of EXTRA and can take advantages of easy proximal operations on many nonsmooth functions. We proved its convergence and established its $O\left(\frac{1}{k}\right)$ convergence rate. The preliminary numerical results demonstrate its competitiveness, especially over the subgradient and double-loop algorithms on the tested smooth-non-smooth problems. It remains open to improve the rate to $O\left(\frac{1}{\sqrt{k}}\right)$ with Nesterov acceleration techniques, and to extend our method to asynchronous and stochastic settings.

REFERENCES

Fig. 4. Relative error \( \|x^k - x^*\|_F \) and successive difference \( \|x^k - x^{k+1}\|_F^2 \) of PG-EXTRA, DPSM1, and DPSM2 in the decentralized quadratic programming problem. Constant \( \alpha_0 = 0.82193 \) is the critical step size given in Theorem 1 for PG-EXTRA.

