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 $O\left(\frac{1}{k}\right)$ 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 $o\left(\frac{1}{k}\right)$ 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

$$\begin{aligned} \min_{x\in\mathbb{R}^p} \quad & \bar{f}(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x), \\ \text{where} \quad & f_i(x) := s_i(x) + r_i(x), \end{aligned} \tag{1}$$

W. Shi, Q. Ling, and G. Wu are with the Department of Automation, University of Science and Technology of China, Hefei, Anhui 230026, China. W. Yin is with the Department of Mathematics, University of California, Los Angeles, CA 90095, USA. Corresponding author: Q. Ling. Email: qingling@mail.ustc.edu.cn. Part of this paper appears in the 40th International Conference on Acoustics, Speech, and Signal Processing, Brisbane, Australia, April 19–25, 2015 [1]. The work of W. Yin is supported in part by NSF grant DMS-1317602.

and $s_i, r_i : \mathbb{R}^p \to \mathbb{R}$ 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 ∇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 ℓ_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 ℓ_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 ∞ 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 EXTRA [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 ℓ_1 minimization and has a similar restriction on the nonsmooth part. In addition, primal-dual type methods include [7], [27], which are based on DADMM. In this paper, we propose 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 $O\left(\frac{1}{k}\right)$, where k is the iteration number.

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 ∇s_i, the iterates of PG-EXTRA converge to a solution and the first-order optimality condition asymptotically holds at an ergodic rate of $O\left(\frac{1}{k}\right)$. The rate improves to non-ergodic $o\left(\frac{1}{k}\right)$ 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 $x_{(i)}^k$. We introduce an objective function that aggregates all the local terms as

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

where

$$\mathbf{x} \triangleq \begin{pmatrix} — & x_{(1)}^{\mathrm{T}} & — \\ — & x_{(2)}^{\mathrm{T}} & — \\ & \vdots & \\ — & x_{(n)}^{\mathrm{T}} & — \end{pmatrix} \in \mathbb{R}^{n\times p}.$$

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

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

By definition, $\mathbf{f}(\mathbf{x}) = \mathbf{s}(\mathbf{x}) + \mathbf{r}(\mathbf{x})$.

The gradient of $\mathbf{s}$ at $\mathbf{x}$ is given by

$$\nabla\mathbf{s}(\mathbf{x}) \triangleq \begin{pmatrix} — & \left(\nabla s_1(x_{(1)})\right)^{\mathrm{T}} & — \\ — & \left(\nabla s_2(x_{(2)})\right)^{\mathrm{T}} & — \\ & \vdots & \\ — & \left(\nabla s_n(x_{(n)})\right)^{\mathrm{T}} & — \end{pmatrix} \in \mathbb{R}^{n\times p}.$$

where $\nabla s_i(x_{(i)})$ is the gradient of s_i at x_i. We let $\tilde{\nabla}\mathbf{r}(\mathbf{x})$

denote a subgradient of $\tilde{\nabla}\mathbf{r}$ at $\mathbf{x}$:

$$\tilde{\nabla}\mathbf{r}(\mathbf{x}) \triangleq \begin{pmatrix} — & \left(\tilde{\nabla} r_1(x_{(1)})\right)^{\mathrm{T}} & — \\ — & \left(\tilde{\nabla} r_2(x_{(2)})\right)^{\mathrm{T}} & — \\ & \vdots & \\ — & \left(\tilde{\nabla} r_n(x_{(n)})\right)^{\mathrm{T}} & — \end{pmatrix} \in \mathbb{R}^{n\times p},$$

where $\tilde{\nabla} r_i(x_{(i)})$ is a subgradient of r_i at x_i. The ith row of $\mathbf{x}$, $\nabla\mathbf{s}(\mathbf{x})$, and $\tilde{\nabla}\mathbf{r}(\mathbf{x})$ 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 $\|A\|_{\mathrm{F}}$. Given a symmetric positive semidefinite matrix G, define the G–norm: $\|A\|_G \triangleq \sqrt{\mathrm{trace}(A^{\mathrm{T}}GA)}$. The largest singular value of a matrix A is denoted as $\sigma_{\max}(A)$. The largest and smallest eigenvalues of a symmetric matrix B are denoted as $\lambda_{\max}(B)$ and $\lambda_{\min}(B)$, respectively. The smallest *nonzero* eigenvalue of a symmetric positive semidefinite matrix B is denoted as $\tilde{\lambda}_{\min}(B)$. We have $\tilde{\lambda}_{\min}(B)\geq\lambda_{\min}(B)$. Let $\mathrm{null}\{A\}\triangleq\{x\in\mathbb{R}^n\big|Ax=0\}$ denote the null space of A, and $\mathrm{span}\{A\}\triangleq\{y\in\mathbb{R}^m\big|y=Ax,\forall x\in\mathbb{R}^n\}$ 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 $\mathbf{x}^0\in\mathbb{R}^{n\times p}$, that is, each agent i holds an arbitrary point $x^0_{(i)}$. The next point $\mathbf{x}^1$ is generated by a proximal gradient iteration

$$\mathbf{x}^{\frac{1}{2}} = W\mathbf{x}^0 - \alpha\nabla\mathbf{s}(\mathbf{x}^0), \tag{2a}$$

$$\mathbf{x}^1 = \arg\min_{\mathbf{x}}\ \mathbf{r}(\mathbf{x}) + \frac{1}{2\alpha}\|\mathbf{x}-\mathbf{x}^{\frac{1}{2}}\|_{\mathrm{F}}^2, \tag{2b}$$

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 $\mathbf{x}^2,\mathbf{x}^3,\ldots$ are obtained through the following update: $k=0,1,\ldots$

$$\mathbf{x}^{k+1+\frac{1}{2}} = W\mathbf{x}^{k+1} + \mathbf{x}^{k+\frac{1}{2}} - \tilde{W}\mathbf{x}^k - \alpha[\nabla\mathbf{s}(\mathbf{x}^{k+1}) - \nabla\mathbf{s}(\mathbf{x}^k)], \tag{3a}$$

$$\mathbf{x}^{k+2} = \arg\min_{\mathbf{x}}\ \mathbf{r}(\mathbf{x}) + \frac{1}{2\alpha}\|\mathbf{x}-\mathbf{x}^{k+1+\frac{1}{2}}\|_{\mathrm{F}}^2. \tag{3b}$$

In (3a), $\tilde{W}=[\tilde{w}_{ij}]\in\mathbb{R}^{n\times n}$ is another mixing matrix, which we typically set as $\frac{W+I}{2}$ though there are more general choices. With that typical choice, $\tilde{W}\mathbf{x}=\frac{W\mathbf{x}+\mathbf{x}}{2}$ can be easily computed from $W\mathbf{x}$. 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 $\tilde{W}\in\mathbb{R}^{n\times n}$;
Choose step size $\alpha>0$;

1. All agents $i=1,\ldots,n$
pick arbitrary initial $x^0_{(i)}\in\mathbb{R}^p$ and do

$$x^{\frac{1}{2}}_{(i)} = \sum_{j=1}^n w_{ij}x^0_{(j)} - \alpha\nabla s_i(x^0_{(i)});$$

$$x^1_{(i)} = \arg\min_x\ r_i(x) + \tfrac{1}{2\alpha}\|x - x^{\frac{1}{2}}_{(i)}\|_2^2;$$

2. **for** $k=0,1,\cdots$, all agents $i=1,\ldots,n$ **do**

$$x^{k+1+\frac{1}{2}}_{(i)} = \sum_{j=1}^n w_{ij}x^{k+1}_{(j)} + x^{k+\frac{1}{2}}_{(i)} - \sum_{j=1}^n \tilde{w}_{ij}x^k_{(j)} - \alpha\left[\nabla s_i(x^{k+1}_{(i)}) - \nabla s_i(x^k_{(i)})\right];$$

$$x^{k+2}_{(i)} = \arg\min_x\ r_i(x) + \tfrac{1}{2\alpha}\|x - x^{x+1+\frac{1}{2}}_{(i)}\|_2^2;$$

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 $\sum_{j=1}^n w_{ij}x_{(j)}$ and $\sum_{j=1}^n\tilde{w}_{ij}x_{(j)}$ 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 $\mathcal{G}=\{\mathcal{V},\mathcal{E}\}$ consisting of a set of agents $\mathcal{V}=\{1,2,\cdots,n\}$ and a set of undirected edges $\mathcal{E}$. An unordered pair $(i,j)\in\mathcal{E}$ 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 $(i,j)\notin\mathcal{E}$, then $w_{ij}=\tilde{w}_{ij}=0$.*
2) *(Symmetry property) $W=W^{\mathrm{T}}$, $\tilde{W}=\tilde{W}^{\mathrm{T}}$.*
3) *(Null space property) $\mathrm{null}\{W-\tilde{W}\}=\mathrm{span}\{\mathbf{1}\}$; $\mathrm{null}\{I-\tilde{W}\}\supseteq\mathrm{span}\{\mathbf{1}\}$.*
4) *(Spectral property) $\tilde{W}\succ 0$ and $\frac{I+W}{2}\succcurlyeq\tilde{W}\succcurlyeq W$.*

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 $\lambda_{\max}(W)=1$ and its other eigenvalues lie in $(-1,1)$. Typical choices of W can be found in [23], [28]. If a matrix W satisfy all the conditions, then $\tilde{W}=\frac{W+I}{2}$ also satisfies the conditions.

B. Special Cases: EXTRA and P-EXTRA

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

$$\mathbf{x}^1 = W\mathbf{x}^0 - \alpha\nabla\mathbf{s}(\mathbf{x}^0). \tag{4}$$

In (3a) and (3b), we have $\mathbf{x}^{k+2} = \mathbf{x}^{k+1+\frac{1}{2}}$ and thus

$$\begin{aligned}\mathbf{x}^{k+2} =& W\mathbf{x}^{k+1} + \mathbf{x}^{k+1} - \tilde{W}\mathbf{x}^k \\ & - \alpha[\nabla \mathbf{s}(\mathbf{x}^{k+1}) - \nabla \mathbf{s}(\mathbf{x}^k)].\end{aligned} \tag{5}$$

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

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

Algorithm 2: P-EXTRA

Set mixing matrices $W \in \mathbb{R}^{n\times n}$ and $\tilde{W} \in \mathbb{R}^{n\times n}$;
Choose step size $\alpha > 0$;

1. All agents $i = 1, \ldots, n$
pick arbitrary initial $x_{(i)}^0 \in \mathbb{R}^p$ and do

$$x_{(i)}^{\frac{1}{2}} = \sum_{j=1}^n w_{ij} x_{(j)}^0,$$

$$x_{(i)}^1 = \arg\min_x \; r_i(x) + \tfrac{1}{2\alpha}\|x - x_{(i)}^{\frac{1}{2}}\|_2^2.$$

2. **for** $k = 0, 1, \cdots$, all agents $i = 1, \ldots, n$ do

$$x_{(i)}^{k+1+\frac{1}{2}} = \sum_{j=1}^n w_{ij} x_{(j)}^{k+1} + x_{(i)}^{k+\frac{1}{2}} - \sum_{j=1}^n \tilde{w}_{ij} x_{(j)}^k,$$

$$x_{(i)}^{k+2} = \arg\min_x \; r_i(x) + \tfrac{1}{2\alpha}\|x - x_{(i)}^{x+1+\frac{1}{2}}\|_2^2.$$

end for

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*

$$\|\nabla s_i(x) - \nabla s_i(y)\|_2 \le L_{s_i}\|x - y\|_2, \quad \forall x, y \in \mathbb{R}^p,$$

where $L_{s_i} > 0$ are constant.

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

$$\|\nabla \mathbf{s}(\mathbf{x}) - \nabla \mathbf{s}(\mathbf{y})\|_{\mathrm{F}} \le L_{\mathbf{s}}\|\mathbf{x} - \mathbf{y}\|_{\mathrm{F}}, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^{n\times p}$$

with constant $L_{\mathbf{s}} = \max_i\{L_{s_i}\}$.

Assumption 3 (Solution existence). *The set of solution(s) $\mathcal{X}^*$ 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 $\tilde{W}$ and the economical-form singular value decomposition $\tilde{W} - W = VSV^{\mathrm{T}}$, define $U \triangleq VS^{1/2}V^{\mathrm{T}} = (\tilde{W} - W)^{1/2} \in \mathbb{R}^{n\times n}$. Then, under Assumptions 1–3, the following two statements are equivalent*

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

$$\begin{cases} U\mathbf{q}^* + \alpha(\nabla \mathbf{s}(\mathbf{x}^*) + \tilde{\nabla}\mathbf{r}(\mathbf{x}^*)) = \mathbf{0}, & \text{(6a)} \\ U\mathbf{x}^* = \mathbf{0}. & \text{(6b)} \end{cases}$$

Proof: According to Assumption 1 and the definition of U, we have

$$\mathrm{null}\{U\} = \mathrm{null}\{V^{\mathrm{T}}\} = \mathrm{null}\{\tilde{W} - W\} = \mathrm{span}\{\mathbf{1}\}.$$

Hence $\mathbf{x}$ is consensual if and only if (6b) holds.

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

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

$$\mathbf{q}^k \triangleq \sum_{t=0}^k U\mathbf{x}^t.$$

The next lemma restates the updates of PG-EXTRA in terms of $\mathbf{x}^k$, $\mathbf{q}^k$, $\mathbf{x}^*$, and $\mathbf{q}^*$ for convergence analysis.

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

$$\begin{aligned} & (I + W - 2\tilde{W})\mathbf{x}^{k+1} + \tilde{W}(\mathbf{x}^{k+1} - \mathbf{x}^k) \\ = \; & -U\mathbf{q}^{k+1} - \alpha\nabla \mathbf{s}(\mathbf{x}^k) - \alpha\tilde{\nabla}\mathbf{r}(\mathbf{x}^{k+1}) \end{aligned} \tag{7}$$

and

$$\begin{aligned} & (I + W - 2\tilde{W})(\mathbf{x}^{k+1} - \mathbf{x}^*) \\ & +\tilde{W}(\mathbf{x}^{k+1} - \mathbf{x}^k) \\ = \; & -U(\mathbf{q}^{k+1} - \mathbf{q}^*) - \alpha[\nabla \mathbf{s}(\mathbf{x}^k) - \nabla \mathbf{s}(\mathbf{x}^*)] \\ & -\alpha[\tilde{\nabla}\mathbf{r}(\mathbf{x}^{k+1}) - \tilde{\nabla}\mathbf{r}(\mathbf{x}^*)], \end{aligned} \tag{8}$$

for any $k = 0, 1, \cdots$.

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

$$\begin{aligned} \mathbf{x}^1 = \; & W\mathbf{x}^0 - \alpha\nabla \mathbf{s}(\mathbf{x}^0) - \alpha\tilde{\nabla}\mathbf{r}(\mathbf{x}^1), \\ \mathbf{x}^{k+1} = \; & (I + W)\mathbf{x}^k - \tilde{W}\mathbf{x}^{k-1} \\ & -\alpha[\nabla \mathbf{s}(\mathbf{x}^k) - \nabla \mathbf{s}(\mathbf{x}^{k-1})] \\ & -\alpha[\tilde{\nabla}\mathbf{r}(\mathbf{x}^{k+1}) - \tilde{\nabla}\mathbf{r}(\mathbf{x}^k)], \quad k = 1, 2, \cdots. \end{aligned}$$

Summing these subgradient recursions over times 1 through $k+1$, we get

$$\begin{aligned}\mathbf{x}^{k+1} = \quad & \tilde{W}\mathbf{x}^k - \sum_{t=0}^{k}(\tilde{W}-W)\mathbf{x}^t \\ & -\alpha\nabla\mathbf{s}(\mathbf{x}^k) - \alpha\tilde{\nabla}\mathbf{r}(\mathbf{x}^{k+1}).\end{aligned} \tag{9}$$

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

Since $(I + W - 2\tilde{W})\mathbf{1} = 0$, $\text{null}\{U\} = (\text{span}\{\mathbf{1}\})^\perp$ and $U\mathbf{x}^* = \mathbf{0}$, we have

$$(I + W - 2\tilde{W})\mathbf{x}^* = \mathbf{0}, \tag{10}$$

Subtracting (10) from (7) and adding $\mathbf{0} = U\mathbf{q}^* + \alpha(\nabla\mathbf{s}(\mathbf{x}^*) + \tilde{\nabla}\mathbf{r}(\mathbf{x}^*))$ to (7), we obtain (8). ■

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* $\{\mathbf{x}^k, \mathbf{q}^k, \mathbf{x}^*, \mathbf{q}^*\}$ *obeys*

$$\begin{aligned}& (I + W - 2\tilde{W})\mathbf{x}^{k+1} + \tilde{W}(\mathbf{x}^{k+1} - \mathbf{x}^k) \\ = \quad & -U\mathbf{q}^{k+1} - \alpha\tilde{\nabla}\mathbf{r}(\mathbf{x}^{k+1})\end{aligned} \tag{11}$$

and

$$\begin{aligned}& (I + W - 2\tilde{W})(\mathbf{x}^{k+1} - \mathbf{x}^*) \\ & +\tilde{W}(\mathbf{x}^{k+1} - \mathbf{x}^k) \\ = \quad & -U(\mathbf{q}^{k+1} - \mathbf{q}^*) \\ & -\alpha[\tilde{\nabla}\mathbf{r}(\mathbf{x}^{k+1}) - \tilde{\nabla}\mathbf{r}(\mathbf{x}^*)],\end{aligned} \tag{12}$$

for any $k = 0, 1, \cdots$.

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

$$\mathbf{z}^k \triangleq \begin{pmatrix}\mathbf{q}^k \\ \mathbf{x}^k\end{pmatrix},\ \mathbf{z}^* \triangleq \begin{pmatrix}\mathbf{q}^* \\ \mathbf{x}^*\end{pmatrix},\ G \triangleq \begin{pmatrix}I & \mathbf{0} \\ \mathbf{0} & \tilde{W}\end{pmatrix}.$$

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, $\mathbf{x}^k$ also converges to an optimal solution $\mathbf{x}^*$; the progress $\|\mathbf{z}^k - \mathbf{z}^{k+1}\|_G^2$ 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* $\alpha \in \left(0, \frac{2\lambda_{\min}(\tilde{W})}{L_\mathbf{s}}\right)$, *then the sequence* $\{\mathbf{z}^k\}$ *generated by PG-EXTRA satisfies*

$$\|\mathbf{z}^k - \mathbf{z}^*\|_G^2 - \|\mathbf{z}^{k+1} - \mathbf{z}^*\|_G^2 \geq \zeta\|\mathbf{z}^k - \mathbf{z}^{k+1}\|_G^2, \tag{13}$$
$$k = 0, 1, \cdots,$$

where $\zeta = 1 - \frac{\alpha L_\mathbf{s}}{2\lambda_{\min}(\tilde{W})}$. *Furthermore,* $\mathbf{z}^k$ *converges to an optimal* $\mathbf{z}^*$.

Proof: By Assumption 2, $\mathbf{s}$ and $\mathbf{r}$ are convex, and $\nabla\mathbf{s}$ is Lipschitz continuous with constant $L_\mathbf{s}$, we have

$$\begin{aligned}& \tfrac{2\alpha}{L_\mathbf{s}}\|\nabla\mathbf{s}(\mathbf{x}^k) - \nabla\mathbf{s}(\mathbf{x}^*)\|_\text{F}^2 \\ \leq \quad & 2\alpha\langle\mathbf{x}^k - \mathbf{x}^*, \nabla\mathbf{s}(\mathbf{x}^k) - \nabla\mathbf{s}(\mathbf{x}^*)\rangle \\ = \quad & 2\langle\mathbf{x}^{k+1} - \mathbf{x}^*, \alpha[\nabla\mathbf{s}(\mathbf{x}^k) - \nabla\mathbf{s}(\mathbf{x}^*)]\rangle \\ & +2\alpha\langle\mathbf{x}^k - \mathbf{x}^{k+1}, \nabla\mathbf{s}(\mathbf{x}^k) - \nabla\mathbf{s}(\mathbf{x}^*)\rangle,\end{aligned} \tag{14}$$

and

$$0 \leq 2\langle\mathbf{x}^{k+1} - \mathbf{x}^*, \alpha[\tilde{\nabla}\mathbf{r}(\mathbf{x}^{k+1}) - \tilde{\nabla}\mathbf{r}(\mathbf{x}^*)]\rangle. \tag{15}$$

Substituting (8) from Lemma 2 for $\alpha[\nabla\mathbf{s}(\mathbf{x}^k) - \nabla\mathbf{s}(\mathbf{x}^*)] + \alpha[\tilde{\nabla}\mathbf{r}(\mathbf{x}^{k+1}) - \tilde{\nabla}\mathbf{r}(\mathbf{x}^*)]$, it follows from (14) and (15) that

$$\begin{aligned}& \tfrac{2\alpha}{L_\mathbf{s}}\|\nabla\mathbf{s}(\mathbf{x}^k) - \nabla\mathbf{s}(\mathbf{x}^*)\|_\text{F}^2 \\ = \quad & 2\langle\mathbf{x}^{k+1} - \mathbf{x}^*, \alpha[\nabla\mathbf{s}(\mathbf{x}^k) - \nabla\mathbf{s}(\mathbf{x}^*)] \\ & +\alpha[\tilde{\nabla}\mathbf{r}(\mathbf{x}^k) - \tilde{\nabla}\mathbf{r}(\mathbf{x}^*)]\rangle \\ & +2\alpha\langle\mathbf{x}^k - \mathbf{x}^{k+1}, \nabla\mathbf{s}(\mathbf{x}^k) - \nabla\mathbf{s}(\mathbf{x}^*)\rangle \\ \leq \quad & 2\langle\mathbf{x}^{k+1} - \mathbf{x}^*, U(\mathbf{q}^* - \mathbf{q}^{k+1})\rangle \\ & +2\langle\mathbf{x}^{k+1} - \mathbf{x}^*, \tilde{W}(\mathbf{x}^k - \mathbf{x}^{k+1})\rangle \\ & -2\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{I+W-2\tilde{W}}^2 \\ & +2\alpha\langle\mathbf{x}^k - \mathbf{x}^{k+1}, \nabla\mathbf{s}(\mathbf{x}^k) - \nabla\mathbf{s}(\mathbf{x}^*)\rangle.\end{aligned} \tag{16}$$

For the terms on the right-hand side of (16), we have

$$\begin{aligned}& 2\langle\mathbf{x}^{k+1} - x^*, U(\mathbf{q}^* - \mathbf{q}^{k+1})\rangle \\ = \quad & 2\langle\mathbf{q}^{k+1} - \mathbf{q}^k, \mathbf{q}^* - \mathbf{q}^{k+1}\rangle,\end{aligned} \tag{17}$$

$$\begin{aligned}& 2\langle\mathbf{x}^{k+1} - \mathbf{x}^*, \tilde{W}(\mathbf{x}^k - \mathbf{x}^{k+1})\rangle \\ = \quad & 2\langle\mathbf{x}^{k+1} - \mathbf{x}^k, \tilde{W}(\mathbf{x}^* - \mathbf{x}^{k+1})\rangle,\end{aligned} \tag{18}$$

and

$$\begin{aligned}& 2\alpha\langle\mathbf{x}^k - \mathbf{x}^{k+1}, \nabla\mathbf{s}(\mathbf{x}^k) - \nabla\mathbf{s}(\mathbf{x}^*)\rangle \\ \leq \quad & \tfrac{\alpha L_\mathbf{s}}{2}\|\mathbf{x}^k - \mathbf{x}^{k+1}\|_\text{F}^2 + \tfrac{2\alpha}{L_\mathbf{s}}\|\nabla\mathbf{s}(\mathbf{x}^k) - \nabla\mathbf{s}(\mathbf{x}^*)\|_\text{F}^2.\end{aligned} \tag{19}$$

Plugging (17)–(19) into (16), we have

$$\begin{aligned}& \tfrac{2\alpha}{L_\mathbf{s}}\|\nabla\mathbf{s}(\mathbf{x}^k) - \nabla\mathbf{s}(\mathbf{x}^*)\|_\text{F}^2 \\ \leq \quad & 2\langle\mathbf{q}^{k+1} - \mathbf{q}^k, \mathbf{q}^* - \mathbf{q}^{k+1}\rangle \\ & +2\langle\mathbf{x}^{k+1} - \mathbf{x}^k, \tilde{W}(\mathbf{x}^* - \mathbf{x}^{k+1})\rangle \\ & -2\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{I+W-2\tilde{W}}^2 + \tfrac{\alpha L_\mathbf{s}}{2}\|\mathbf{x}^k - \mathbf{x}^{k+1}\|_\text{F}^2 \\ & +\tfrac{2\alpha}{L_\mathbf{s}}\|\nabla\mathbf{s}(\mathbf{x}^k) - \nabla\mathbf{s}(\mathbf{x}^*)\|_\text{F}^2.\end{aligned} \tag{20}$$

Using the definitions of $\mathbf{z}^k$, $\mathbf{z}^*$ and G, (20) is equivalent to

$$\begin{aligned}0 \leq \quad & 2\langle\mathbf{z}^{k+1} - \mathbf{z}^k, G(\mathbf{z}^* - \mathbf{z}^{k+1})\rangle \\ & -2\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{I+W-2\tilde{W}}^2 + \tfrac{\alpha L_\mathbf{s}}{2}\|\mathbf{x}^k - \mathbf{x}^{k+1}\|_\text{F}^2.\end{aligned} \tag{21}$$

Applying the basic equality

$$\begin{aligned} & 2\langle \mathbf{z}^{k+1}-\mathbf{z}^k, G(\mathbf{z}^*-\mathbf{z}^{k+1})\rangle \\ = \; & \|\mathbf{z}^k-\mathbf{z}^*\|_G^2-\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_G^2-\|\mathbf{z}^k-\mathbf{z}^{k+1}\|_G^2 \end{aligned} \tag{22}$$

to (21), we have

$$\begin{aligned} 0 \le \; & \|\mathbf{z}^k-\mathbf{z}^*\|_G^2-\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_G^2-\|\mathbf{z}^k-\mathbf{z}^{k+1}\|_G^2 \\ & -2\|\mathbf{x}^{k+1}-\mathbf{x}^*\|^2_{I+W-2\tilde{W}}+\tfrac{\alpha L_{\mathbf{s}}}{2}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathrm{F}}^2. \end{aligned} \tag{23}$$

By Assumption 1, in particular, $I+W-2\tilde{W}\succcurlyeq 0$, we have $\|\mathbf{x}^{k+1}-\mathbf{x}^*\|^2_{I+W-2\tilde{W}}\ge 0$ and thus

$$\begin{aligned} & \|\mathbf{z}^k-\mathbf{z}^*\|_G^2-\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_G^2 \\ \ge \; & \|\mathbf{z}^k-\mathbf{z}^{k+1}\|_G-\tfrac{\alpha L_{\mathbf{s}}}{2}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathrm{F}}^2 \\ \ge \; & \zeta\|\mathbf{z}^k-\mathbf{z}^{k+1}\|_G, \end{aligned} \tag{24}$$

where $\zeta=1-\frac{\alpha L_{\mathbf{s}}}{2\lambda_{\min}(\tilde{W})}>0$. The last inequality holds since $\alpha<\frac{2\lambda_{\min}(\tilde{W})}{L_{\mathbf{s}}}$.

It shows from (24) that for any optimal solution $\mathbf{z}^*$, $\|\mathbf{z}^k-\mathbf{z}^*\|_G^2$ is bounded and contractive. Therefore, $\|\mathbf{z}^k-\mathbf{z}^*\|_G^2$ is converging as long as $\|\mathbf{z}^k-\mathbf{z}^{k+1}\|_G^2\to 0$. The convergence of $\mathbf{z}^k$ to an optimal solution $\mathbf{z}^*$ 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) $a_k\ge 0$ *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) $\min_{t\le k}\{a_t\}=o\left(\frac{1}{k}\right)$; *If the sequence $\{a_k\}$ further obeys:* (3) $a_{k+1}\le a_k$, *then in addition, we have:* (iv) $a_k=o\left(\frac{1}{k}\right)$.

Proof: Part (i) is obvious. Let $b_k\triangleq\frac{1}{k}\sum_{t=1}^{k}a_t$. By the assumptions, kb_k is uniformly bounded and obeys

$$\lim_{k\to\infty}kb_k<\infty,$$

from which part (ii) follows. Since $c_k\triangleq\min_{t\le k}\{a_t\}$ is monotonically non-increasing, we have

$$kc_{2k}=k\min_{t\le 2k}\{a_t\}\le\sum_{t=k+1}^{2k}a_t.$$

This and the fact that $\lim_{k\to\infty}\sum_{t=k+1}^{2k}a_t\to 0$ give us $c_k=o\left(\frac{1}{k}\right)$ or part (iii).

If a_k is further monotonically non-increasing, we have

$$ka_{2k}\le\sum_{t=k+1}^{2k}a_t.$$

This and the fact that $\lim_{k\to\infty}\sum_{t=k+1}^{2k}a_t\to 0$ 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:

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

(ii) Running-best successive difference:

$$\min_{t\le k}\left\{\|\mathbf{z}^t-\mathbf{z}^{t+1}\|_G^2\right\}=o\left(\frac{1}{k}\right);$$

(iii) Running-average optimality residuals:

$$\frac{1}{k}\sum_{t=1}^{k}\|U\mathbf{q}^t+\alpha(\nabla\mathbf{s}(\mathbf{x}^t)+\tilde{\nabla}\mathbf{r}(\mathbf{x}^{t+1}))\|^2_{\tilde{W}}=O\left(\frac{1}{k}\right),$$

$$\frac{1}{k}\sum_{t=1}^{k}\|U\mathbf{x}^t\|_{\mathrm{F}}^2=O\left(\frac{1}{k}\right);$$

(iv) Running-best optimality residuals:

$$\begin{aligned} & \min_{t\le k}\left\{\|U\mathbf{q}^t+\alpha(\nabla\mathbf{s}(\mathbf{x}^t)+\tilde{\nabla}\mathbf{r}(\mathbf{x}^{t+1}))\|^2_{\tilde{W}}\right\} \\ = \; & o\left(\frac{1}{k}\right), \end{aligned}$$

$$\min_{t\le k}\left\{\|U\mathbf{x}^t\|_{\mathrm{F}}^2\right\}=o\left(\frac{1}{k}\right).$$

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 $\|\mathbf{z}^k-\mathbf{z}^*\|_G^2$ converges to 0 when k goes to ∞, we are able to sum (13) in Theorem 1 over $k=0$ through ∞ and apply the telescopic cancellation, which yields

$$\begin{aligned} & \sum_{t=0}^{\infty}\|\mathbf{z}^t-\mathbf{z}^{t+1}\|_G^2 \\ = \; & \tfrac{1}{\zeta}\sum_{t=0}^{\infty}\left(\|\mathbf{z}^t-\mathbf{z}^*\|_G^2-\|\mathbf{z}^{t+1}-\mathbf{z}^*\|_G^2\right) \\ = \; & \tfrac{\|\mathbf{z}_0-\mathbf{z}^*\|_G^2}{\zeta} \\ < \; & \infty. \end{aligned} \tag{25}$$

Then, the results follow from Proposition 1 immediately.

Parts (iii) and (iv): Using the basic inequality $\|\mathbf{a}+\mathbf{b}\|_{\mathrm{F}}^2 \geq \frac{1}{\rho}\|\mathbf{a}\|_{\mathrm{F}}^2 - \frac{1}{\rho-1}\|\mathbf{b}\|_{\mathrm{F}}^2$ which holds for any $\rho > 1$ and any matrices $\mathbf{a}$ and $\mathbf{b}$ of the same size, it follows that

$$\begin{aligned} & \|\mathbf{z}^k - \mathbf{z}^{k+1}\|_G^2 \\ = \; & \|\mathbf{q}^k - \mathbf{q}^{k+1}\|_{\mathrm{F}}^2 + \|\mathbf{x}^k - \mathbf{x}^{k+1}\|_{\tilde{W}}^2 \\ = \; & \|\mathbf{x}^{k+1}\|_{\tilde{W}-W}^2 \\ & + \|(I-\tilde{W})\mathbf{x}^k + U\mathbf{q}^k + \alpha(\nabla \mathbf{s}(\mathbf{x}^k) + \tilde{\nabla}\mathbf{r}(\mathbf{x}^{k+1}))\|_{\tilde{W}}^2 \\ \geq \; & \|\mathbf{x}^{k+1}\|_{\tilde{W}-W}^2 \\ & + \tfrac{1}{\rho}\|U\mathbf{q}^k + \alpha(\nabla \mathbf{s}(\mathbf{x}^k) + \tilde{\nabla}\mathbf{r}(\mathbf{x}^{k+1}))\|_{\tilde{W}}^2 \\ & - \tfrac{1}{\rho-1}\|(I-\tilde{W})\mathbf{x}^k\|_{\tilde{W}}^2. \end{aligned} \tag{26}$$

Since $\tilde{W} - W$ and $(I-\tilde{W})\tilde{W}(I-\tilde{W})$ are symmetric and

$$\mathrm{null}\{\tilde{W}-W\} \subseteq \mathrm{null}\{(I-\tilde{W})\tilde{W}(I-\tilde{W})\},$$

there exists a bounded $\upsilon > 0$ such that

$$\|(I-\tilde{W})\mathbf{x}^k\|_{\tilde{W}}^2 = \|\mathbf{x}^k\|_{(I-\tilde{W})\tilde{W}(I-\tilde{W})}^2 \leq \upsilon\|\mathbf{x}^k\|_{\tilde{W}-W}^2.$$

It follows from (26) that

$$\begin{aligned} & \tfrac{1}{k}\sum_{t=1}^k \|\mathbf{z}^t - \mathbf{z}^{t+1}\|_G^2 + \tfrac{1}{k}\|\mathbf{x}^1\|_{\tilde{W}-W}^2 \\ \geq \; & \tfrac{1}{k}\sum_{t=1}^k \left(\|\mathbf{x}^{t+1}\|_{\tilde{W}-W}^2 - \tfrac{\upsilon}{\rho-1}\|\mathbf{x}^t\|_{\tilde{W}-W}^2\right) \\ & + \tfrac{1}{k}\|\mathbf{x}^1\|_{\tilde{W}-W}^2 \\ & + \tfrac{1}{k}\sum_{t=1}^k \tfrac{1}{\rho}\|U\mathbf{q}^t + \alpha(\nabla \mathbf{s}(\mathbf{x}^k) + \tilde{\nabla}\mathbf{r}(\mathbf{x}^{k+1}))\|_{\tilde{W}}^2 \\ & (\text{let } \rho > \upsilon + 1) \\ = \; & \tfrac{1}{k}\sum_{t=1}^k (1 - \tfrac{\upsilon}{\rho-1})\|U\mathbf{x}^t\|_{\mathrm{F}}^2 + \tfrac{1}{k}\|\mathbf{x}^{k+1}\|_{\tilde{W}-W}^2 \\ & + \tfrac{1}{k}\sum_{t=1}^k \tfrac{1}{\rho}\|U\mathbf{q}^t + \alpha(\nabla \mathbf{s}(\mathbf{x}^k) + \tilde{\nabla}\mathbf{r}(\mathbf{x}^{k+1}))\|_{\tilde{W}}^2. \end{aligned} \tag{27}$$

As part (i) shows that $\frac{1}{k}\sum_{t=1}^k \|\mathbf{z}^t - \mathbf{z}^{t+1}\|_G^2 = O\left(\frac{1}{k}\right)$, we have $\frac{1}{k}\sum_{t=1}^k \|U\mathbf{q}^t + \alpha(\nabla \mathbf{s}(\mathbf{x}^t) + \tilde{\nabla}\mathbf{r}(\mathbf{x}^{t+1}))\|_{\tilde{W}}^2 = O\left(\frac{1}{k}\right)$ and $\frac{1}{k}\sum_{t=1}^k \|U\mathbf{x}^t\|_{\mathrm{F}}^2 = O\left(\frac{1}{k}\right)$.

From (27) and (25), we see that both $\|U\mathbf{q}^k + \alpha(\nabla \mathbf{s}(\mathbf{x}^k) + \tilde{\nabla}\mathbf{r}(\mathbf{x}^{k+1}))\|_{\tilde{W}}^2$ and $\|U\mathbf{x}^k\|_{\mathrm{F}}^2$ 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 $\|\mathbf{z}^k - \mathbf{z}^{k+1}\|_G^2$ 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 $L_{\mathbf{s}} \to 0^+$, the sequence $\{\|\mathbf{z}^k - \mathbf{z}^{k+1}\|_G^2\}$ is summable. Second, the sequence $\{\|\mathbf{z}^k - \mathbf{z}^{k+1}\|_G^2\}$ 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* $\mathbf{s}(\mathbf{x}) = 0$*, for any step size* $\alpha > 0$*, the sequence* $\{\mathbf{z}^k\}$ *generated by P-EXTRA satisfies*

$$\begin{aligned} & \|\mathbf{z}^{k+1} - \mathbf{z}^{k+2}\|_G^2 \leq \|\mathbf{z}^k - \mathbf{z}^{k+1}\|_G^2, \\ & k = 0, 1, \cdots. \end{aligned} \tag{28}$$

Proof: To simplify the description of the proof, define $\Delta\mathbf{x}^{k+1} \triangleq \mathbf{x}^k - \mathbf{x}^{k+1}$, $\Delta\mathbf{q}^{k+1} \triangleq \mathbf{q}^k - \mathbf{q}^{k+1}$, $\Delta\mathbf{z}^{k+1} \triangleq \mathbf{z}^k - \mathbf{z}^{k+1}$, and $\Delta\tilde{\nabla}\mathbf{r}(\mathbf{x}^{k+1}) \triangleq \tilde{\nabla}\mathbf{r}(\mathbf{x}^k) - \tilde{\nabla}\mathbf{r}(\mathbf{x}^{k+1})$. By convexity of $\mathbf{r}$ in Assumption 2, we have

$$\langle \Delta\mathbf{x}^{k+1}, \Delta\tilde{\nabla}\mathbf{r}(\mathbf{x}^{k+1})\rangle \geq 0. \tag{29}$$

Taking difference of (7) at the k-th and $(k+1)$-th iterations yields

$$\begin{aligned} & \alpha\Delta\tilde{\nabla}\mathbf{r}(\mathbf{x}^{k+1}) + U\Delta\mathbf{q}^{k+1} + (I + W - 2\tilde{W})\Delta\mathbf{x}^{k+1} \\ & + \tilde{W}(\Delta\mathbf{x}^{k+1} - \Delta\mathbf{x}^k) = 0. \end{aligned} \tag{30}$$

Combine (29) and (30) it follows that

$$\begin{aligned} & \langle \Delta\mathbf{x}^{k+1}, -U\Delta\mathbf{q}^{k+1}\rangle \\ & + \langle \Delta\mathbf{x}^{k+1}, -\tilde{W}(\Delta\mathbf{x}^{k+1} - \Delta\mathbf{x}^k)\rangle \\ \geq \; & \|\Delta\mathbf{x}^{k+1}\|_{I+W-2\tilde{W}}^2. \end{aligned} \tag{31}$$

Using the definition of $\mathbf{q}^k$, $\Delta\mathbf{q}^{k+1} = -U\mathbf{x}^{k+1}$. Thus, we have

$$\Delta\mathbf{q}^k - \Delta\mathbf{q}^{k+1} = -U\Delta\mathbf{x}^{k+1}. \tag{32}$$

Substituting (32) into (31) yields

$$\begin{aligned} & \langle \Delta\mathbf{q}^k - \Delta\mathbf{q}^{k+1}, \Delta\mathbf{q}^{k+1}\rangle \\ & + \langle \Delta\mathbf{x}^{k+1}, -\tilde{W}(\Delta\mathbf{x}^{k+1} - \Delta\mathbf{x}^k)\rangle \\ \geq \; & \|\Delta\mathbf{x}^{k+1}\|_{I+W-2\tilde{W}}^2, \end{aligned} \tag{33}$$

or equivalently

$$\langle \Delta\mathbf{z}^k - \Delta\mathbf{z}^{k+1}, G\Delta\mathbf{z}^{k+1}\rangle \geq \|\Delta\mathbf{x}^{k+1}\|_{I+W-2\tilde{W}}^2. \tag{34}$$

By applying the basic equality $2\langle \Delta\mathbf{z}^k - \Delta\mathbf{z}^{k+1}, G\Delta\mathbf{z}^{k+1}\rangle = \|\Delta\mathbf{z}^k\|_G^2 - \|\Delta\mathbf{z}^{k+1}\|_G^2 - \|\Delta\mathbf{z}^k - \Delta\mathbf{z}^{k+1}\|_G^2$ to (34), we finally have

$$\begin{aligned} & \|\Delta\mathbf{z}^k\|_G^2 - \|\Delta\mathbf{z}^{k+1}\|_G^2 \\ \geq \; & \|\Delta\mathbf{z}^k - \Delta\mathbf{z}^{k+1}\|_G^2 + 2\|\Delta\mathbf{x}^{k+1}\|_{I+W-2\tilde{W}}^2 \\ \geq \; & 0, \end{aligned} \tag{35}$$

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 $\{\|\mathbf{z}^k - \mathbf{z}^{k+1}\|_G^2\}$ has been shown monotonic in P-EXTRA. Invoking fact

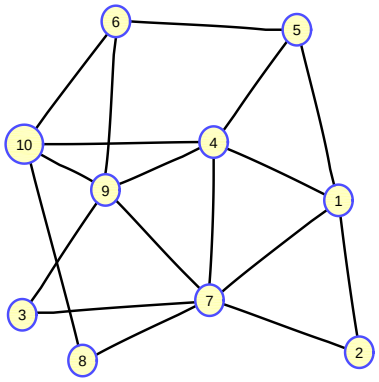

Fig. 1. The underlying network for the experiments.

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

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

(i) Successive difference:

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

(ii) First-order optimality residuals:

$$\|U\mathbf{q}^k + \alpha\tilde{\nabla}\mathbf{r}(\mathbf{x}^k)\|_{\tilde{W}^{-1}}^2 = o\left(\frac{1}{k}\right),$$

$$\|U\mathbf{x}^k\|_{\mathrm{F}}^2 = o\left(\frac{1}{k}\right).$$

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 showing in Fig. 1, which has $n = 10$ agents and $\frac{0.4n(n-1)}{2} = 18$ edges.

In the numerical experiments, we use the relative error $\frac{\|\mathbf{x}^k - \mathbf{x}^*\|_{\mathrm{F}}^2}{\|\mathbf{x}^0 - \mathbf{x}^*\|_{\mathrm{F}}^2}$ and the successive difference $\|\mathbf{x}^k - \mathbf{x}^{k+1}\|_{\mathrm{F}}^2$ 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 $\mathbf{x}^{k+2} \leftarrow \arg\min_{\mathbf{x}} \mathbf{f}(\mathbf{x}) + \frac{1}{2\alpha}\|\mathbf{x} - \mathbf{x}^{k+1+\frac{1}{2}}\|_{\mathrm{F}}^2$ has an explicit solution

$$x_{(i)}^{k+2} = y_{(i)} - \frac{e_{(i)}^{k+1+\frac{1}{2}}}{\|e_{(i)}^{k+1+\frac{1}{2}}\|_2}\left(\|e_{(i)}^{k+1+\frac{1}{2}}\|_2 - \alpha\right)_+,$$

where $e_{(i)}^{k+1+\frac{1}{2}} \triangleq y_{(i)} - x_{(i)}^{k+1+\frac{1}{2}}$ 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 $y_{(i)}$ are generated following the uniform distribution in $[-200, 200] \times [-200, 200] \times [-200, 200]$. The algorithm starts from $x_{(i)}^0 = y_{(i)}$, $\forall i$. We use the Metropolis constant edge weight for W and $\tilde{W} = \frac{I+W}{2}$, as well as a constant step size α.

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 $\alpha^k = O(k^{-\frac{1}{2}})$. 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.

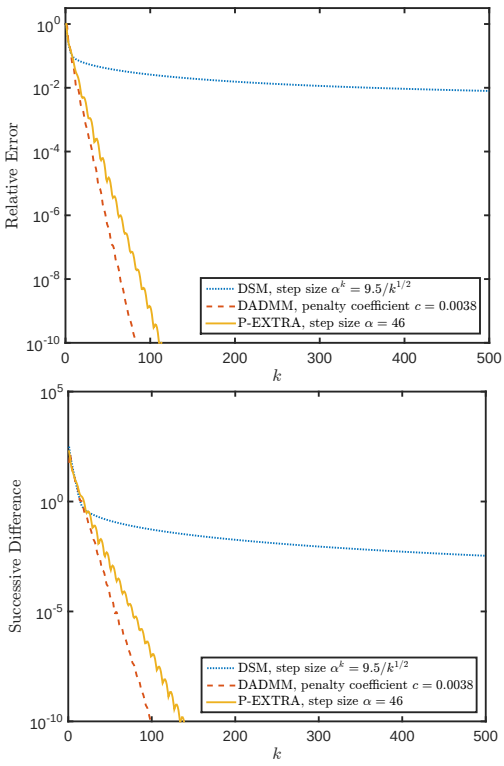

Fig. 2. Relative error $\frac{\|\mathbf{x}^k-\mathbf{x}^*\|_F}{\|\mathbf{x}^0-\mathbf{x}^*\|_F}$ and successive difference $\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_F^2$ of P-EXTRA, DSM, and DADMM in the decentralized geometric median problem.

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.

B. Decentralized Compressed Sensing

Consider a decentralized compressed sensing problem. Each agent $i \in \{1, \cdots, n\}$ holds its own measurement equations, $y_{(i)} = M_{(i)}x + e_{(i)}$, where $y_{(i)} \in \mathbb{R}^{m_i}$ is a measurement vector, $M_{(i)} \in \mathbb{R}^{m_i \times p}$ is a sensing matrix, $x \in \mathbb{R}^p$ is an unknown *sparse* signal, and $e_{(i)} \in \mathbb{R}^{m_i}$ is an i.i.d. 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 unknowns p, which fails the ordinary least squares. We instead solve an ℓ_1-regularized least squares problem

$$\min_x \bar{s}(x) + \bar{r}(x) = \frac{1}{n}\sum_{i=1}^n s_i(x) + \frac{1}{n}\sum_{i=1}^n r_i(x),$$

where

$$s_i(x) = \frac{1}{2}\|M_{(i)}x - y_{(i)}\|_2^2, \quad r_i(x) = \lambda_{(i)}\|x\|_1,$$

and $\lambda_{(i)}$ is the regularization parameter on agent i.

The decentralized compressed sensing is a special case of the general distributed compressed sensing [34] where its intra-signal correlation is consensus. 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.

Considering 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 noises vectors $e_{(i)}$ are randomly generated with their elements 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 $x_{(i)}^0 = 0, \ \forall i$. In PG-EXTRA, we use the Metropolis constant edge weight for W and $\tilde{W} = \frac{I+W}{2}$, and constant step size α.

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 τ 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 $i \in \{1, \cdots, n\}$ has a local quadratic objective $\frac{1}{2}x^{\mathrm{T}}Q_i x + h_i^{\mathrm{T}}x$ and a local linear constraint $a_i^{\mathrm{T}}x \leq b_i$, where the symmetric positive semidefinite matrix $Q_i \in \mathbb{R}^{p\times p}$, the vectors $h_i \in \mathbb{R}^p$ 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:

$$\begin{aligned} \min_x \quad & \tfrac{1}{n}\sum_{i=1}^n s_i(x) = \tfrac{1}{n}\sum_{i=1}^n \left(\tfrac{1}{2}x^{\mathrm{T}}Q_i x + h_i^{\mathrm{T}}x\right), \\ s.t. \quad & a_i^{\mathrm{T}}x \leq b_i, \ \forall i = 1, ..., n. \end{aligned}$$

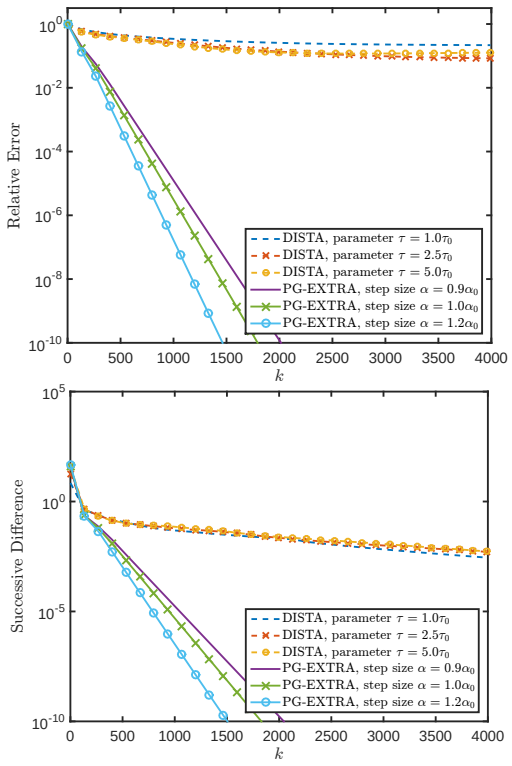

Fig. 3. Relative error $\frac{\|\mathbf{x}^k-\mathbf{x}^*\|_{\mathrm{F}}}{\|\mathbf{x}^0-\mathbf{x}^*\|_{\mathrm{F}}}$ and successive difference $\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathrm{F}}^2$ of PG-EXTRA and DISTA in the decentralized compressive sensing problem. Constant $\alpha_0 = 0.82193$ is the step size given in Theorem 1 for PG-EXTRA. Constant $\tau_0 = \|M_{(i)}\|_2^{-2} = 1$ is a parameter of DISTA.

We recast it as

$$\min_x \frac{1}{n}\sum_{i=1}^n f_i(x) = \frac{1}{n}\sum_{i=1}^n \left(\tfrac{1}{2}x^{\mathrm{T}}Q_i x + h_i^{\mathrm{T}}x + \mathcal{I}(a_i^{\mathrm{T}}x - b_i)\right), \tag{36}$$

where

$$\mathcal{I}(c) = \begin{cases} 0, & \text{if } c \le 0, \\ +\infty, & \text{otherwise,} \end{cases}$$

is an indicator function. Setting $s_i(x) = \frac{1}{2}x^{\mathrm{T}}Q_i x + h_i^{\mathrm{T}}x$ and $r_i(x) = \mathcal{I}(a_i^{\mathrm{T}}x - b_i)$, it has the form of (1) and can be solved by PG-EXTRA. The minimization subproblem in PG-EXTRA $\mathbf{x}^{k+2} \leftarrow \arg\min_{\mathbf{x}} \mathbf{f}(\mathbf{x}) + \frac{1}{2\alpha}\|\mathbf{x} - \mathbf{x}^{k+1+\frac{1}{2}}\|_{\mathrm{F}}^2$ has an explicit solution. Indeed, for agent i, the solution is

$$x_{(i)}^{k+2} = \begin{cases} x_{(i)}^{k+1+\frac{1}{2}}, & \text{if } a_i^{\mathrm{T}}x_{(i)}^{k+1+\frac{1}{2}} \le b_i, \\ x_{(i)}^{k+1+\frac{1}{2}} + \frac{(b_i - a_i^{\mathrm{T}}x_{(i)}^{k+1+\frac{1}{2}})a_i}{\|a_i\|_2^2}, & \text{otherwise.} \end{cases}$$

In this experiment, we set $p = 50$. Each Q_i is generated by a p-by-p matrix, whose elements follow and i.i.d. Gaussian distribution, multiplying its transpose. Each h_i's elements are generated following and i.i.d. Gaussian distribution. We also randomly generate the constraints 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 $a_i^{\mathrm{T}}x \le b_i$, $\forall i$. In this way we can make sure at least one of the constraints $a_i^{\mathrm{T}}x \le b_i$, $\forall i$ is activated. In PG-EXTRA, we use the Metropolis constant edge weight for W and $\tilde{W} = \frac{I+W}{2}$, and constant step size α.

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 $a_i^{\mathrm{T}}x \le b_i$, $\forall i$ so that DSPM can be applied to solve (36). The iteration of DSPM1 at each agent i is $x_i^{k+1} = \mathcal{P}_\Omega\left(\sum_{j=1}^n w_{ij}x_j^k - \alpha^k \nabla s_i(x_i^k)\right)$ where the set $\Omega = \{\tilde{x} \big| a_i^{\mathrm{T}}\tilde{x} \le b_i, \ \forall i\}$ and $\mathcal{P}_\Omega(\cdot)$ stands for projection on to Ω. 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 $x_i^{k+1} = \mathcal{P}_{\Omega_i}\left(\sum_{j=1}^n w_{ij}x_j^k - \alpha^k \nabla s_i(x_i^k)\right)$ where $\Omega_i = \{\tilde{x} \big| a_i^{\mathrm{T}}\tilde{x} \le b_i\}$. DSPM2 is likely to be convergent but has no theoretical guarantee. Both DSPM1 and DSPM2 use diminishing step size $\alpha_k = O(k^{-\frac{1}{2}})$ 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+nonsmooth 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+nonsmooth problems. It remains open to improve the rate to $O\left(\frac{1}{k^2}\right)$ with Nesterov acceleration techniques, and to extend our method to asynchronous and stochastic settings.

References

[1] W. Shi, Q. Ling, G. Wu, and W. Yin, "A Proximal Gradient Algorithm for Decentralized Nondifferentiable Optimization," in *Proceedings of the 40th IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2015.

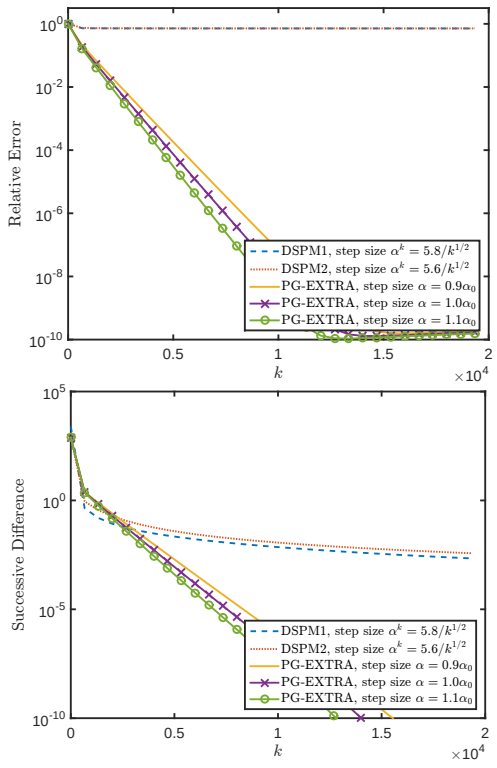

Fig. 4. Relative error $\frac{\|\mathbf{x}^k-\mathbf{x}^*\|_{\mathrm{F}}}{\|\mathbf{x}^0-\mathbf{x}^*\|_{\mathrm{F}}}$ and successive difference $\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathrm{F}}^2$ of PG-EXTRA, DPSM1, and DPSM2 in the decentralized quadratic programming problem. Constant $\alpha_0 = 0.82193$ is the critical step size given in Theorem 1 for PG-EXTRA.

[2] H. Eiselt and V. Marianov, Eds., *Foundations of Location Analysis*, Springer, 2011.

[3] H. Lopuhaa and P. Rousseeuw, "Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices," *The Annals of Statistics*, vol. 19, no. 1, pp. 229–248, 1991.

[4] Q. Ling and Z. Tian, "Decentralized Sparse Signal Recovery for Compressive Sleeping Wireless Sensor Networks," *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3816–3827, 2010.

[5] G. Mateos, J. Bazerque, and G. Giannakis, "Distributed Sparse Linear Regression," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5262–5276, 2010.

[6] S. Lee and A. Nedic, "Distributed Random Projection Algorithm for Convex Optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 2, pp. 221–229, 2013.

[7] T. Chang, M. Hong, and X. Wang, "Multi-Agent Distributed Optimization via Inexact Consensus ADMM," *IEEE Transactions on Signal Processing*, vol. 63, no. 2, pp. 482–497, 2015.

[8] C. Pang, "Set Intersection Problems: Supporting Hyperplanes and Quadratic Programming," *Mathematical Programming*, vol. 149, no. 1-2, pp. 329–359, 2015.

[9] J. Tsitsiklis, *Problems in Decentralized Decision Making and Computation*, Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1984.

[10] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed Asynchronous Deterministic and Stochastic Gradient Optimization Algorithms," *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 803–812, 1986.

[11] B. Johansson, *On Distributed Optimization in Networked Systems*, Ph.D. thesis, School of Electrical Engineering, Royal Institute of Technology, 2008.

[12] Y. Cao, W. Yu, W. Ren, and G. Chen, "An Overview of Recent Progress in the Study of Distributed Multi-agent Coordination," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 1, pp. 427–438, 2013.

[13] A. Sayed, "Adaptation, Learning, and Optimization over Networks," *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, 2014.

[14] F. Bullo, J. Cortes, and S. Martinez, *Distributed Control of Robotic Networks*, Princeton University Press, 2009.

[15] K. Zhou and S. Roumeliotis, "Multirobot Active Target Tracking with Combinations of Relative Observations," *IEEE Transactions on Robotics*, vol. 27, no. 4, pp. 678–695, 2010.

[16] I. Schizas, A. Ribeiro, and G. Giannakis, "Consensus in Ad Hoc WSNs with Noisy Links–Part I: Distributed Estimation of Deterministic Signals," *IEEE Transactions on Signal Processing*, vol. 56, no. 1, pp. 350–364, 2008.

[17] V. Kekatos and G. Giannakis, "Distributed Robust Power System State Estimation," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1617–1626, 2013.

[18] G. Giannakis, V. Kekatos, N. Gatsis, S. Kim, H. Zhu, and B. Wollenberg, "Monitoring and Optimization for Power Grids: A Signal Processing Perspective," *IEEE Signal Processing Magazine*, vol. 30, no. 5, pp. 107–128, 2013.

[19] P. Forero, A. Cano, and G. Giannakis, "Consensus-Based Distributed Support Vector Machines," *Journal of Machine Learning Research*, vol. 11, pp. 1663–1707, 2010.

[20] F. Yan, S. Sundaram, S. Vishwanathan, and Y. Qi, "Distributed Autonomous Online Learning: Regrets and Intrinsic Privacy-Preserving Properties," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 11, pp. 2483–2493, 2013.

[21] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the Linear Convergence of the ADMM in Decentralized Consensus Optimization," *IEEE Transactions on Signal Processing*, vol. 62, no. 7, pp. 1750–1761, 2014.

[22] A. Nedic and A. Ozdaglar, "Distributed Subgradient Methods for Multi-agent Optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[23] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An Exact First-Order Algorithm for Decentralized Consensus Optimization," arXiv preprint arXiv:1404.6264, 2014.

[24] A. Chen, "Fast Distributed First-Order Methods," M.S. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2012.

[25] C. Ravazzi, S. Fosson, and E. Magli, "Distributed Iterative Thresholding for ℓ_0/ℓ_1-Regularized Linear Inverse Problems," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 2081–2100, 2015.

[26] C. Ravazzi, S. M. Fosson, and E. Magli, "Distributed Soft Thresholding for Sparse Signal Recovery," in *Proceedings of IEEE Global Communications Conference (GLOBECOM)*, 2013, pp. 3429–3434.

[27] P. Bianchi, W. Hachem, and F. Iutzeler, "A Stochastic Primal-Dual Algorithm for Distributed Asynchronous Composite Optimization," in *Proceedings of the 2nd Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 732–736.

[28] S. Boyd, P. Diaconis, and L. Xiao, "Fastest Mixing Markov Chain on a Graph," *SIAM Review*, vol. 46, no. 4, pp. 667–689, 2004.

[29] B. He, "A New Method for A Class of Linear Variational Inequalities," *Mathematical Programming*, vol. 66, no. 1-3, pp. 137–144, 1994.

[30] Wei Deng, Ming-Jun Lai, Zhimin Peng, and Wotao Yin, "Parallel multi-block ADMM with $o(1/k)$ convergence," .

[31] Damek Davis and Wotao Yin, "Convergence rate analysis of several splitting schemes," .

[32] R. Ravichandran, G. Gordon, and S. Goldstein, "A Scalable Distributed Algorithm for Shape Transformation in Multi-Robot Systems," in *Proceedings of the IEEE/RSJ International Confer-*

ence on Intelligent Robots and Systems (IROS), 2007, pp. 4188–4193.

[33] S. Minsker, S. Srivastava, L. Lin, and D. Dunson, "Scalable and Robust Bayesian Inference via the Median Posterior," in *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014, pp. 1656–1664.

[34] D. Baron, M. Duarte, M. Wakin, S. Sarvotham, and R. Baraniuk, "Distributed Compressive Sensing," *arXiv preprint arXiv:0901.3403*, 2009.

[35] S. Ram, A. Nedic, and V. Veeravalli, "Distributed Stochastic Subgradient Projection Algorithms for Convex Optimization," *Journal of Optimization Theory and Applications*, vol. 147, no. 3, pp. 516–545, 2010.