A PROXIMAL GRADIENT ALGORITHM FOR DECENTRALIZED COMPOSITE OPTIMIZATION OVER DIRECTED NETWORKS

Tao He, Jinshan Zeng and Mingwen Wang

College of Computer Information Engineering, Jiangxi Normal University, Nanchang, China, 330022

ABSTRACT

This paper proposes a decentralized algorithm for solving a consensus optimization problem defined in a directed networked multi-agent system, where the local objective functions have the smooth+nonsmooth composite form, and are possibly nonconvex. Examples of such problems include decentralized compressed sensing and constrained quadratic programming problems, as well as many decentralized regularization problems. We extend the existing algorithms PG-EXTRA and ExtraPush to a new algorithm PG-ExtraPush for composite consensus optimization over a directed network. This algorithm takes advantage of the proximity operator like in PG-EXTRA to deal with the nonsmooth term, and employs the push-sum protocol like in ExtraPush to tackle the bias introduced by the directed network. We show that PG-ExtraPush converges to an optimal solution under the boundedness assumption. In numerical experiments, with a proper step size, PG-ExtraPush performs surprisingly linear rates in most of cases, even in some nonconvex cases.

Index Terms— Directed network, composite objective, decentralized optimization, proximal, nonsmooth.

1. INTRODUCTION

We consider the following consensus optimization problem defined on a directed, strongly connected network of $n$ agents:

\[
\text{minimize } f(x) \equiv \sum_{i=1}^{n} f_i(x),
\]

where \( f_i(x) = s_i(x) + r_i(x), \) (1)

and for every agent $i$, $f_i$ is a proper, coercive and possibly nonconvex function only known to the agent, $s_i$ is a smooth function, $r_i$ is generally nonsmooth and possibly nonconvex. We say that the objective has the smooth+nonsmooth composite structure.

The smooth+nonsmooth structure of the local objective 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 [5]; (ii) the compressed 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 $l_q$ (quasi)-norm with $0 \leq q \leq 1$ [6]. [8]; (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 [3], [7].

For a stationary network with bi-directional communication, the existing algorithms include the primal-dual domain methods such as the decentralized alternating direction method of multipliers (DADMM) [13], [14], and the primal domain methods including the distributed subgradient method (DSM) [15]. Both algorithms do not take advantage of the smooth+nonsmooth structure. While 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) [2], the proximal decentralized gradient descent method (Prox-DGD) [23], the distributed iterative soft thresholding algorithm (DISTA) [12], proximal gradient exact first-order algorithm (PG EXTRA) [16]. All these primal-domain methods consist of a gradient step for the smooth part and a proximal step for the nonsmooth part. Different from DPGM, Prox-DGD and DISTA, PG-EXTRA as an extension of EXTRA [15] has two interlaced sequences of iterates, whereas the proximal-gradient method just inherits the sequence of iterates in the gradient method.

This paper focuses on a directed network with directional communication, which is pioneered by the works [17], [18], [19]. When communication is bi-directional, algorithms can use a symmetric and doubly-stochastic mixing matrix to obtain a consensual solution; however, once the communication is directional, the mixing matrix becomes generally asymmetric and only column-stochastic. In the column-stochastic setting, the push-sum protocol [4] can be used to obtain a stationary distribution for the mixing matrix. Some recent decentralized algorithms over a directed network include Subgradient-Push [10], ExtraPush [22] (also called DEXTRA in [20]) and Push-DIGing [11]. The best rate of Subgradient-Push in the general convex case is $O(\ln t/\sqrt{t})$, where $t$ is the iteration number, and both ExtraPush and Push-DIGing perform linearly con-
vergent in the strongly convex case. However, all of these algorithms do not consider the smooth+nonsmooth structure as well as the nonconvex case as defined in problem (I).

In this paper, we extend the algorithms PG-EXTRA and ExtraPush to the composite consensus optimization problem with the smooth+nonsmooth part. At each iteration, each agent locally computes a gradient of the smooth part of its objective and a proximal map of the nonsmooth part, and exchanges information with its neighbors, then uses the push-sum protocol [4] to achieve the consensus. When the network is undirected, the proposed PG-ExtraPush reduces to PG-EXTRA, and when \( r_i = 0 \), PG-ExtraPush reduces to ExtraPush [22]. In numerical experiments, with a proper step size, the proposed algorithm PG-ExtraPush performs the linear rates in these convex cases, and is significantly faster than Subgradient-Push, even if the latter uses a hand-optimized step size. Moreover, we apply the proposed algorithm to the nonconvex decentralized \( \ell_q \) regularized least squares regression problems with \( 0 \leq q < 1 \). The experiment results also show the effectiveness of PG-ExtraPush even in such nonconvex cases.

The rest of paper is organized as follows. Section 2 introduces the problem setup. Section 3 develops the proposed algorithm and establishes its convergence under the boundedness assumption. Section 4 presents our numerical simulation results. We conclude this paper in Section 5.

Notation: Let \( I_n \) denote an identity matrix with the size \( n \times n \). We use \( 1_n \in \mathbb{R}^n \) as a vector of all 1’s. For any vector \( x \), we let \( x_i \) denote its \( i \)th component and \( \text{diag}(x) \) denote the diagonal matrix generated by \( x \). For any matrix \( X \), \( X^T \) denotes its transpose, \( X_{ij} \) denotes its \( (i,j) \)th component, and \( \| X \|_F \triangleq \sqrt{\langle X, X \rangle} = \sqrt{\sum_{i,j} X_{ij}^2} \) denotes its Frobenius norm. For any matrix \( B \in \mathbb{R}^{m \times n} \), \( \text{null}(B) \triangleq \{ x \in \mathbb{R}^n | Bx = 0 \} \) is the null space of \( B \). Given a matrix \( B \in \mathbb{R}^{m \times n} \), by \( Z \in \text{null}(B) \), we mean that each column of \( Z \) lies in \( \text{null}(B) \).

2. PROBLEM REFORMULATION

2.1. Network

Consider a directed network \( G = \{V, E\} \), where \( V \) is the vertex set and \( E \) is the edge set. Any edge \( (i,j) \in E \) represents a directed arc from node \( i \) to node \( j \). The sets of in-neighbors and out-neighbors of node \( i \) are

\[ N_{i}^{\text{in}} \triangleq \{ j : (j,i) \in E \} \cup \{ i \}, \quad N_{i}^{\text{out}} \triangleq \{ j : (i,j) \in E \} \cup \{ i \}, \]

respectively. Let \( d_i \triangleq |N_{i}^{\text{out}}| \) be the out-degree of node \( i \). In \( G \), each node \( i \) can only send information to its out-neighbors, not vice versa.

To illustrate a mixing matrix for a directed network, consider \( A \in \mathbb{R}^{n \times n} \) where

\[
\begin{cases}
    A_{ij} > 0, & \text{if } j \in N_{i}^{\text{in}} \\
    A_{ij} = 0, & \text{otherwise}.
\end{cases}
\]

![Fig. 1. A directed graph \( G \) (left) and its mixing matrix \( A \) (right).](image)

The entries \( A_{ij} \) satisfy that, for each node \( j \), \( \sum_{i \in V} A_{ij} = 1 \). An example is the following mixing matrix

\[
A_{ij} = \begin{cases} 1/d_j, & \text{if } j \in N_{i}^{\text{in}} \\ 0, & \text{otherwise} \end{cases},
\]

\( i,j = 1, \ldots, n \), which is used in the Subgradient-Push method [10]. See Fig. 1 for a directed graph \( G \) and an example of its mixing matrix \( A \). The matrix \( A \) is column stochastic and asymmetric in general.

**Property 1.** [22] Property 1 Assume that \( G \) is strongly connected, the followings hold:

(i) Let \( A^t = A \times A \cdots A \) for any \( t \in \mathbb{N} \). Then

\[
A^t \rightarrow \phi I_n^T \text{ geometrically fast as } t \rightarrow \infty,
\]

for some stationary distribution vector \( \phi \), i.e., \( \phi_i \geq 0 \) and \( \sum_{i} \phi_i = 1 \).

(ii) \( \text{null}(I_n - \phi I_n^T) = \text{null}(I_n - A) \).

(iii) \( A\phi = \phi \).

(iv) The quantity \( \xi \triangleq \inf_{1 \leq i \leq n} (A^t 1_n)_i \leq \frac{1}{n^r} > 0 \).

Letting,

\[
D \triangleq \xi \text{diag}(\phi).
\]

2.2. Problem with matrix notation

Let \( x_{(i)} \in \mathbb{R}^p \) denote the local copy of \( x \) at node \( i \), and \( x_{(i)}^t \) denote its value at the \( t \)th iteration. Throughout the note, we use the following equivalent form of the problem (I) using local copies of the variable \( x \):

\[
\begin{aligned}
\text{minimize}_{x} & \quad \frac{1}{n} \sum_{i=1}^{n} f_i(x_{(i)}), \\
\text{subject to} & \quad x_{(i)} = x_{(j)}, \quad \forall (i,j) \in E,
\end{aligned}
\]

where \( x_{(i)} \in \mathbb{R}^p \) denotes the vector with all its entries equal to 1, \( x \in \mathbb{R}^{n \times p} \), \( f(x) \in \mathbb{R}^n \), \( s(x) \in \mathbb{R}^n \) and \( r(x) \in \mathbb{R}^n \) with

\[
x \triangleq \begin{pmatrix}
    x_{(1)}^T \\
    x_{(2)}^T \\
    \vdots \\
    x_{(n)}^T
\end{pmatrix}, \quad f(x) \triangleq \begin{pmatrix}
    f_1(x_{(1)}) \\
    f_2(x_{(2)}) \\
    \vdots \\
    f_n(x_{(n)})
\end{pmatrix}.
\]
s(x) = \begin{pmatrix} s_1(x_1) \\ s_2(x_2) \\ \vdots \\ s_n(x_n) \end{pmatrix}, \quad r(x) = \begin{pmatrix} r_1(x_1) \\ r_2(x_2) \\ \vdots \\ r_n(x_n) \end{pmatrix}.

In addition, the gradient of s(x) is

\nabla s(x) = \begin{pmatrix} -\nabla s_1(x_1)^T \\ -\nabla s_2(x_2)^T \\ \vdots \\ -\nabla s_n(x_n)^T \end{pmatrix} \in \mathbb{R}^{n \times p},

and a subgradient of r(x) is

\tilde{\nabla} r(x) = \begin{pmatrix} -\tilde{\nabla} r_1(x_1)^T \\ -\tilde{\nabla} r_2(x_2)^T \\ \vdots \\ -\tilde{\nabla} r_n(x_n)^T \end{pmatrix} \in \mathbb{R}^{n \times p}.

The ith rows of the above matrices x, \nabla s(x) and \tilde{\nabla} r(x), and vector s(x), correspond to agent i. For simplicity, one can treat p = 1 throughout this paper. To deal with the nonsmooth part, given a parameter \alpha > 0, we introduce the proximity operator \text{prox}_{\alpha r}, associated with ri as follows

\text{prox}_{\alpha r_i}(z) = \arg\min_{w \in \mathbb{R}^p} \{r_i(u) + \frac{||u - z||^2}{2\alpha}\}.

For any \z \in \mathbb{R}^{n \times p}, define

\text{Prox}_{\alpha r}(\z) = \begin{pmatrix} \text{prox}_{\alpha r_1}(z_1) \\ \text{prox}_{\alpha r_2}(z_2) \\ \vdots \\ \text{prox}_{\alpha r_n}(z_n) \end{pmatrix}.

3. ALGORITHM AND CONVERGENCE ANALYSIS

3.1. Proposed Algorithm: PG-ExtraPush

The proposed algorithm PG-ExtraPush extends PG-EXTRA and ExtraPush to composite (smooth+nonsmooth) consensus optimization problem. Given a sequence of n-dimensional positive vectors \{w^t\} \in \mathbb{R}^n, we define a sequence of functions

r^t(x) = \text{diag}(w^t) r(\text{diag}(w^t)^{-1} x), \forall x \in \mathbb{R}^{n \times p}, t \in \mathbb{N}.

Let \bar{A} \triangleq \frac{A + L}{2}. Specifically, the proposed algorithm can be described as follows: for all agents \(i = 1, \ldots, n\), set arbitrary \(z^0_{i} \in \mathbb{R}^p, w^0_{i} = 1, x^0_{i} = z^0_{i}, z^{1/2}_{i} = \sum_{j=1}^{n} A_{ij} z^0_{j} - \alpha \nabla s_i(z^0_{i}), w^{1}_{i} = \sum_{j=1}^{n} A_{ij} w^0_{j}, z^{1/2}_{i} = \text{prox}_{\alpha r_i}(z^{1/2}_{i}),

x^{1}_{i} = \frac{z^{1}_{i}}{w^{1}_{i}}. \text{ For } t = 1, 2, \ldots, \text{ perform }
\begin{align*}
    z^{t+1/2}_{i} &= \sum_{j=1}^{n} A_{ij} z^{t}_{j} + z^{t-1/2}_{i} - \sum_{j=1}^{n} \bar{A}_{ij} z^{t-1}_{j} - \alpha(\nabla s_i(x^{t}_{i}) - \nabla s_i(x^{t-1}_{i})), \\
    w^{t+1}_{i} &= \sum_{j=1}^{n} A_{ij} w^{t}_{j}, \quad z^{t+1}_{i} = \text{prox}_{\alpha r_i}(z^{t+1/2}_{i}), \\
    x^{t+1}_{i} &= \frac{z^{t+1}_{i}}{w^{t+1}_{i}}. \quad (8)
\end{align*}

The matrix form of the algorithm can be described as follows: set arbitrary \(z^0 \in \mathbb{R}^{n \times p}, w^0 = 1_n, x^0 = z^0, z^{1/2} = A z^0 - \alpha \nabla s(z^0), w^1 = A w^0, z^1 = \text{Prox}_{\alpha r}(z^{1/2}), x^1 = \text{diag}(w^1)^{-1} z^1. \text{ For } t = 1, 2, \ldots, \text{ perform }
\begin{align*}
    z^{t+1/2} &= Az^{t} + z^{t-1/2} - \bar{A} z^{t-1} - \alpha(\nabla s(x^{t}) - \nabla s(x^{t-1})), \\
    w^{t+1} &= A w^{t}, \quad z^{t+1} = \text{Prox}_{\alpha r}(z^{t+1/2}), \\
    x^{t+1} &= \text{diag}(w^{t+1})^{-1} z^{t+1}. \quad (9)
\end{align*}

3.2. Special Cases: PG-EXTRA, ExtraPush, P-ExtraPush

When the network is undirected, then the weight sequence \(w^t \equiv 1_n, \text{ thus, the function } r^t \equiv r \text{ and the sequence } x^t = z^t. \text{ Therefore, PG-ExtraPush reduces to PG-EXTRA, a recent algorithm for composite consensus optimization over undirected networks.}

When the possibly-nondifferentiable term r \equiv 0, we have \(z^1 = z^{1/2}, \text{ and thus, } z^t = A z^0 - \alpha \nabla s(z^0). \text{ In the third update of (9), } z^{t+1} = z^{t+1/2}, \text{ and thus }
\begin{align*}
    z^{t+1} &= Az^t + z^{t-1/2} - \bar{A} z^{t-1} - \alpha(\nabla s(x^t) - \nabla s(x^{t-1})). \quad (10)
\end{align*}

With these, in this case, PG-ExtraPush reduces to ExtraPush, a recent algorithm for decentralized differentiable optimization over directed networks.

When the differentiable term s \equiv 0, PG-ExtraPush reduces to P-ExtraPush by removing all gradient computation, which is given as follows: set arbitrary \(z^0 \in \mathbb{R}^{n \times p}, w^0 = 1_n, x^0 = z^0; z^{1/2} = A z^0, w^1 = A w^0, z^1 = \text{Prox}_{\alpha r}(z^{1/2}), x^1 = \text{diag}(w^1)^{-1} z^1. \text{ For } t = 1, 2, \ldots, \text{ perform }
\begin{align*}
    z^{t+1/2} &= Az^t + z^{t-1/2} - \bar{A} z^{t-1}, \\
    w^{t+1} &= A w^{t}, \quad z^{t+1} = \text{Prox}_{\alpha r}(z^{t+1/2}), \\
    x^{t+1} &= \text{diag}(w^{t+1})^{-1} z^{t+1}. \quad (11)
\end{align*}

3.3. Convergence Analysis

In this subsection, we first develop the first-order optimality conditions for the problem (6) and then provide the convergence of PG-ExtraPush under the boundedness assumption.

Theorem 1 (first-order optimality conditions). Suppose that graph \(G\) is strongly connected. Then \(x^*\) is consensual and
\[ x^*_t(1) \equiv x^*_z(2) \equiv \cdots \equiv x^*_x(n) \ \text{is an optimal solution of (1) if and only if, for some } \alpha > 0, \text{ there exist } z^* \in \text{null}(I_n - A) \text{ and } y^* \in \text{null}(I^n_n) \text{ such that the following conditions hold}
\]
\[
\begin{cases}
y^* + \alpha(\nabla s(x^*) + \nabla r(x^*)) = 0, \\
x^* = D^{-1}z^*.
\end{cases}
\] (12)

(We let \( \mathcal{E}_+ \) denote the set of triples \((z^*, y^*, x^*)\) satisfying the above conditions.)

**Proof.** Assume that \( x^* \) is consensual and \( x^*_t(1) \equiv x^*_z(2) \equiv \cdots \equiv x^*_x(n) \) is optimal. Let \( z^* = n\text{diag}(\phi)x^* = n(\phi x^*_T)^T \). Then \( \phi I^n_n z^* = \phi I^n_n n(\phi x^*_T)^T = n(\phi x^*_T)^T = z^* \). It implies that \( z^* \in \text{null}(I_n - \phi I^n_n) \). By Property (i), it follows that \( z^* \in \text{null}(I_n - A) \). Moreover, letting \( y^* = -\alpha(\nabla s(x^*) + \nabla r(x^*)) \), it holds that \( I^n_n y^* = -\alpha I^n_n(\nabla s(x^*) + \nabla r(x^*)) = 0 \), thus \( y^* \in \text{null}(I^n_n) \).

On the other hand, assume that (12) holds. By Property (ii), it follows that \( z^* = \phi I^n_n x^* \). Plugging \( x^* = D^{-1}z^* \) gives \( x^* = \frac{1}{n}I_n D^{-1}z^* \), which implies that \( x^* \) is consensual. Moreover, by \( y^* + \alpha(\nabla s(x^*) + \nabla r(x^*)) = 0 \) and \( y^* \in \text{null}(I^n_n) \), it holds \( I^n_n(\nabla s(x^*) + \nabla r(x^*)) = -\frac{1}{n}I_n y^* = 0 \), which implies that \( x^* \) is optimal.

**Lemma 1** (recursion of PG-ExtraPush). Introducing the sequence
\[
y^t = \sum_{k=0}^{t} (A - A)x^k,
\] (13)
the iteration (9) of PG-ExtraPush can be rewritten as
\[
\begin{cases}
\tilde{A}z^{t+1} = \tilde{A}z^t - \alpha\nabla r(x^{t+1}) - \alpha \nabla s(x^t) - y^{t+1}, \\
y^{t+1} = y^t + (A - A)z^{t+1}, \\
w^{t+1} = Aw^t, \\
x^{t+1} = \text{diag}(w^{t+1})^{-1}z^{t+1}.
\end{cases}
\] (14)

**Proof.** By the definitions of \( r^{t+1} \) and \( \text{prox}_{\alpha r^{t+1}} \) and the x-update in (9), it follows
\[
z^{t+1/2} = z^{t+1} + \alpha \nabla r(x^{t+1}), \forall t \in \mathbb{N}.
\] (15)
Then the first update of (9) imply
\[
z^t = \tilde{A}z^t - \tilde{A}(z^t - z^{t-1}) - \alpha(\nabla s(x^t) - \nabla s(x^{t-1})) - \alpha(\nabla r(x^{t+1}) - \nabla r(x^t)),
\] (16)
for \( t = 1, 2, \ldots \). Moreover, observe that
\[
z = Az^0 - \alpha \nabla s(x^0) - \alpha \nabla r(x^t).
\]
Summing these subgradient recursions over times 1 through \( t + 1 \), we get
\[
z^{t+1} = \tilde{A}z^t + \sum_{k=0}^{t} (A - \tilde{A})z^k - \alpha \nabla s(x^t) - \alpha \nabla r(x^{t+1}).
\]
Furthermore, adding \((A - \tilde{A})z^{t+1}\) into both sides of the above equation and noting \( A + I_n = 2A \), we get
\[
\tilde{A}z^{t+1} = \tilde{A}z^t - \alpha \tilde{A}r(x^{t+1}) - \alpha \nabla s(x^t) - y^{t+1}. 
\] (17)

**Theorem 2** (convergence under boundedness assumption). Suppose that the sequence \( \{x^t\} \) generated by PG-ExtraPush (9) and the sequence \( \{y^t\} \) defined in (13) are bounded. Then, any limit point of \( \{\langle z^t, y^t, x^t\rangle\} \), denoted by \((z^*, y^*, x^*)\), satisfies the optimality conditions (12).

**Proof.** By Property (i), \( \{w^t\} \) is bounded. By the last update of (13) and the boundedness of both \( \{x^t\} \) and \( \{w^t\} \), \( \{x^t\} \) is bounded. Hence, there exists a convergent subsequence \( \{\langle z^t, y^t, x^t\rangle\}_j \). Let \( \langle z^*, y^*, x^*\rangle \) be its limit. By (4), we know that \( w^* = n\phi \) and thus that \( x^* = D^{-1}z^* \). Letting \( t \to \infty \) in the second equation of (14) gives \( z^* = Az^* \), or equivalently \( z^* \in \text{null}(I_n - A) \). Similarly, letting \( t \to \infty \) in the first equation of (14) yields \( y^* + \alpha(\nabla s(x^*) + \nabla r(x^*)) = 0 \). Moreover, from the definition of (9) and the facts that both \( A \) and \( \tilde{A} \) are column stochastic, it follows that \( I^n_n y^* = 0 \) and \( I^n_n(\nabla s(x^*) + \nabla r(x^*)) = 0 \). Therefore, \((z^*, y^*, x^*)\) satisfies the optimality conditions (12).

4. NUMERICAL EXPERIMENTS

In this section, we provide a series of numerical experiments to show the effectiveness of the proposed algorithms via comparing to Subgradient-Push algorithm. In these experiments, the connected network and its corresponding mixing matrix \( A \) are generated randomly.

4.1. Decentralized Geometric Median

Consider a decentralized geometric median problem. Each agent \( i \in \{1, \ldots, n\} \) holds a vector \( b_{i} \), and all the agents collaboratively calculate the geometric median \( x \in \mathbb{R}^p \) of all \( b_{i} \). This task can be formulated as solving the following minimization problem:
\[
x^* \leftarrow \arg \min_{x \in \mathbb{R}^p} f(x) = \sum_{i=1}^{n} \|x - b_{i}\|_2.
\] (18)

The geometric median problem is solved by P-ExtraPush over directed networks. The proximity operator \( \text{prox}_{\alpha r_i} \) has an explicit solution, for any \( u \in \mathbb{R}^p \),
\[
\text{prox}_{\alpha r_i}(u) = \frac{b_{i} - u}{\|b_{i} - u\|^2} \max\{\|b_{i} - u\|_2 - 2 - \alpha, 0\}.
\]
We set \( n = 10 \) and \( p = 256 \), that is, each point \( b_{i} \) is drawn from i.i.d. Gaussian distribution. The algorithm starts from \( x_0(x) = b_{i}, \forall i \). We use three different step sizes \( \alpha \) to show the effect of the step size. The
4.2. Decentralized $\ell_1$ Regularized Least Squares Regression

We consider the following decentralized $\ell_1$ regularized least squares regression problem, i.e.,

$$
x^* \leftarrow \arg\min_{x \in \mathbb{R}^n} f(x) = \sum_{i=1}^n f_i(x),
$$

where $f_i(x) = \frac{1}{2} \| B_{(i)} x - b_{(i)} \|_2^2 + \lambda_i \| x \|_1$, $B_{(i)} \in \mathbb{R}^{m_i \times p}$, $b_{(i)} \in \mathbb{R}^{m_i}$ for $i = 1, \ldots, n$, $\| x \|_1 = \sum_{i=1}^n |x_i|$. In this experiment, we take $n = 10$, $p = 256$, and $m_i = 150$ for $i = 1, \ldots, n$. In this case, the proximity operator of $\ell_1$-norm is the soft shrinkage function. The experiment result is illustrated in Fig. 3.

From Fig. 3 $\alpha = 0.038$ is a critical value of step size in the sense that the algorithm will diverge once $\alpha$ is bigger than this value, and with this proper step size, PG-ExtraPush converges linearly and is faster than Subgradient-Push. Moreover, a smaller step size generally implies a slower convergence rate.

4.3. Decentralized Quadratic Programming

We use decentralized quadratic programming as an example to show that how PG-ExtraPush 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 $h_i \in \mathbb{R}^p$ 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:

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

where

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

is an indicator function. Setting $s_i(x) = \frac{1}{2} x^T Q_i x + h_i^T x$ and $r_i(x) = \mathcal{I}(a_i^T x - b_i)$, it has the form of (11) and can be solved by PG-ExtraPush. The proximity operator $\text{prox}_{\alpha r_i}$ has an explicit solution

$$
\text{prox}_{\alpha r_i}(u) = \begin{cases} 
u, & \text{if } a_i^T u \leq b_i \leq 0, \\
u + \frac{(b_i - a_i^T u) \alpha}{\| a_i \|^2}, & \text{otherwise}, 
\end{cases}
$$

Consider $n = 10$ and $p = 256$. For any agent $i$, $Q_i$ is a positive semidefinite symmetric matrix, $h_i$, $a_i$ and $b_i$ are generated from i.i.d. Gaussian distribution. Three different step sizes are used to show the effect of the step size. The experiment result is presented in Fig. 4. Since Subgradient-Push is not appropriately used to solve this problem, so we show the performance of PG-ExtraPush only without any comparison in this case. As show in Fig. 4 PG-ExtraPush can also adopt a large range of step size parameter and $\alpha = 5.5$ is a critical value in this case in the sense that PG-ExtraPush may diverge if a larger step size is adopted. With a proper step size (say, $\alpha = 4$), PG-ExtraPush performs the similar linear convergence rate when all $Q_i$ are positive semidefinite.

4.4. Nonconvex Decentralized Composite Optimization

We apply the proposed algorithm to solve the following nonconvex decentralized $\ell_q$ ($0 \leq q < 1$) regularized least squares
Fig. 4. Experiment results for decentralized quadratic programming with symmetric positive semidefinite $Q_i$’s. Trends of $\|x^t - x^*\|_F$, where $x^*$ is the limitation of $x^t$, which is taken as the iterate at $t = 10000$.

regression problem, i.e.,

$$x^* \leftarrow \arg\min_{x \in \mathbb{R}^p} f(x) = \sum_{i=1}^{n_i} f_i(x),$$  \hspace{1cm} (21)

where $f_i(x) = \frac{1}{2}\|B_i x - b_i\|_2^2 + \lambda_i \|x\|_q^q$, $B_i \in \mathbb{R}^{m_i \times p}$, $b_i \in \mathbb{R}^{m_i}$ for $i = 1, \ldots, n$, $\|x\|_q^q = \sum_{i=1}^p |x_i|^q$ for $0 < q < 1$, and when $q = 0$, $\|x\|_q^q$ denotes the number of nonzero components of $x$. Similar to Subsection 5.2, we take $n = 10$, $p = 256$, and $m_i = 150$ for $i = 1, \ldots, n$. We take different $q = 0, 1/2, 2/3$ since their proximity operators have explicit forms and can be easily computed. In all cases, $\lambda_i = 0.5$ for each agent $i$ and four different step sizes are used. The experiment results are illustrated in Fig. 5.

From Fig. 5 when applied to these nonconvex cases, the choice of step size is more sensitive. The optimal step sizes for $q = 0, 1/2$ and $2/3$ are $0.035, 0.012$ and $0.04$, respectively. With these proper step sizes, when $q = 0$, PG-ExtraPush performs linearly convergent, while for both $q = 1/2$ and $2/3$, PG-ExtraPush decays sublinearly at the first several iterations, and then performs linearly.

5. CONCLUSION

In this paper, we propose a decentralized algorithm called PG-ExtraPush, for solving decentralized composite consensus optimization problems over directed networks. The algorithm uses a fixed step size and the proximal map of the nonsmooth part. We show that PG-ExtraPush converges to an optimal solution under the boundedness assumption. The convergence as well as the rate of convergence of PG-ExtraPush should be justified in the future. With a proper step size, PG-ExtraPush converges linearly and is significantly faster than Subgradient-Push in numerical experiments. Moreover, we show the potential of PG-ExtraPush for solving the decentralized nonconvex regularized optimization problems.

Fig. 5. Experiment results for the decentralized $\ell_q(0 \leq q < 1)$ regularized least squares regression. Trends of $\|x^t - x^*\|_F$, where $x^*$ is the limitation of $x^t$, which is taken as the iterate at $t = 10000$.

6. REFERENCES

[1] F. Benezit, V. Blondel, P. Thiran, J. Tsitsiklis and M. Vetterli, Weighted Gossip: distributed averaging using non-doubly stochastic matrices, in Proc. IEEE Conf. Decision Control, pp: 1753-1757, 2010.

[2] I. Chen, Fast Distributed First-Order Methods, Masters thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 2012.

[3] T. Chang, M. Hong and X. Wang, Multi-agent distributed optimization via inexact consensus ADMM, IEEE Trans. Signal Process., 63(2): 482-497, 2015.

[4] D. Kempe, A. Dobra and J. Gehrke, Gossip-based computation of aggregate information, in 44th Annual IEEE Symposium on Foundations of Computer Science, pp: 482-491, 2003.

[5] H. Lopuhaa and P. Rousseuw, Breakdown points of affine equivariant estimators of multivariate location and covariance matrices, Ann. Statist., 19(1): 229-248, 1991.

[6] Q. Ling and Z. Tian, Decentralized sparse signal recovery for compressive sleeping wireless sensor networks, IEEE Trans. Signal Process., 58(7): 3816-3827, 2010.
[7] S. Lee and A. Nedic, Distributed random projection algorithm for convex optimization, *IEEE J. Sel. Topics Signal Process.*, 7(2): 221-229, 2013.

[8] G. Mateos, J. Bazerque and G. Giannakis, Distributed sparse linear regression, *IEEE Trans. Signal Process.*, 58(10): 5262-5276, 2010.

[9] A. Nedic and A. Ozdaglar, Distributed subgradient methods for multi-agent optimization, *IEEE Trans. Automatic Control*, 54: 48-61, 2009.

[10] A. Nedic and A. Olshevsky, Distributed optimization over time-varying directed graphs, *IEEE Trans. Automatic Control*, 60(3): 601-615, 2015.

[11] A. Nedic, A. Olshevsky and W. Shi, Achieving geometric convergence for distributed optimization over time-varying graphs, arXiv:1607.03218, 2016.

[12] C. Ravazzi, S. Fosson and E. Magli, Distributed iterative thresholding for $\ell_0/\ell_1$-regularized linear inverse problems, *IEEE Trans. Inf. Theory*, 61(4): 2081-2100, 2015.

[13] I. Schizas, A. Ribeiro and G. Giannakis, Consensus in ad hoc WSNs with noisy links-part I: Distributed estimation of deterministic signals, *IEEE Trans. Signal Process.*, 56(1): 350-364, 2008.

[14] W. Shi, Q. Ling, K. Yuan, G. Wu and W. Yin, On the linear convergence of the ADMM in decentralized consensus optimization, *IEEE Trans. Signal Process.*, 62(7): 1750-1761, 2014.

[15] W. Shi, Q. Ling, G. Wu and W. Yin, EXTRA: an exact first-order algorithm for decentralized consensus optimization, *SIAM J. Optimization*, 25(2): 944-966, 2015.

[16] W. Shi, Q. Ling, G. Wu and W. Yin, A Proximal Gradient Algorithm for Decentralized Composite Optimization, *IEEE Transactions on Signal Processing*, 63(22): 6013-6023, 2015.

[17] K. Tsianos, S. Lawlor and M. Rabbat, Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning, in *Proc. 50th Allerton Conf. Commun., Control, Comp.*, pp: 1543-1550, 2012.

[18] K. Tsianos, S. Lawlor and M. G. Rabbat, Push-sum distributed dual averaging for convex optimization, in *Proc. IEEE Conf. Decision Control*, pp: 5453-5458, 2012.

[19] K. Tsianos, *The role of the Network in Distributed Optimization Algorithms: Convergence Rates, Scalability, Communication/Computation Tradeoffs and Communication Delays*, PhD thesis, Dept. Elect. Comp. Eng., McGill Univ., Montreal, QC, Canada, 2013.