EXTRAPUSH FOR CONVEX SMOOTH DECENTRALIZED OPTIMIZATION OVER DIRECTED NETWORKS *

JINSHAN ZENG † AND WOTAO YIN ‡

Abstract. In this note, we extend the existing algorithms Extra [13] and subgradient-push [10] to a new algorithm *ExtraPush* for convex consensus optimization over a directed network. When the network is stationary, we propose a simplified algorithm called *Normalized ExtraPush*. These algorithms use a fixed step size like in Extra and accept the column-stochastic mixing matrices like in subgradient-push. We present preliminary analysis for ExtraPush under a bounded sequence assumption. For Normalized ExtraPush, we show that it naturally produces a bounded, linearly convergent sequence provided that the objective function is strongly convex.

Key words. Dencentralized optimization, directed graph, consensus, non-doubly stochastic, EXTRA

1. Introduction.

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

$$\text{minimize}_{x\in\mathbf{R}^p} f(x) \triangleq \sum_{i=1}^{n} f_i(x), \tag{1.1}$$

where, for every agent i, f_i is a proper closed convex differentiable function only known to the agent.

The model (1.1) finds applications in decentralized averaging, learning, estimation, and control. For a stationary network with *bi-directional* communication, the existing algorithms include the (sub)gradient methods [2, 5, 8, 9, 13, 19], and the primal-dual domain methods such as the decentralized alternating direction method of multipliers (DADMM) [11, 12]. This note focuses on a *directed* network with *directional* communication, where the research of decentralized optimization is pioneered by the works [15, 16, 17]. With bi-directional communication, 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 latter case, the push-sum protocol [6] can be used to obtain a stationary distribution for the mixing matrix. In the former case, if the objective is Lipschitz-differentiable, the gradient-based algorithm Extra [13] converges at the rate of $O(1/t)$, where t is the iteration number. In the latter case, the best rate is $O(\ln t/\sqrt{t})$ from the subgradient-based algorithm [10]. It is not known how to take advantage of the gradient of a Lipschitz-differentiable objective. We make an attempt in this note to combine [10, 13] and present our preliminary results.

Specifically, we propose *ExtraPush*, which is a two-step iteration like Extra and incorporates the push-sum protocol. At each iteration, the Extra variables are approximately normalized by the current push-sum variables. When the network is stationary, we propose to first apply the push-sum protocol to obtain a stationary

*The work of W. Yin has been supported in part by the NSF grants DMS-1317602 and ECCS-1462398.

†College of Computer Information Engineering, Jiangxi Normal University, Nanchang, Jiangxi 330022, China (Email: jsh.zeng@gmail.com)

‡Department of Mathematics, University of California, Los Angeles, CA 90025, USA (Email: wotaoyin@ucla.edu)

distribution and then run the two-step iteration Extra; the resulting algorithm is called Normalized ExtraPush. At each of its iterations, the running variables are normalized by the stationary distribution.

Our algorithms are essentially the same as found in the recent work by Xi and Khan [18]. They present a nice attempt to prove convergence for a strongly convex objective function. They noticed that a certain matrix that is important to the analysis (as a part of their convergence metric) is positive semi-definite. Our analysis also uses this property. However, their analysis breaks down due to their strong assumptions; in fact, no function can satisfy all of their assumptions.

It is also worth noting that our algorithm can be generalized for time-varying networks; however, the proof in this note will also need significant generalization.

The rest of this note is organized as follows. Section 2 introduces the problem setup and preliminaries. Section 3 develops ExtraPush and Normalized ExtraPush. Section 4 establishes the optimality conditions for ExtraPush and shows its convergence under the boundedness assumption. Section 5 assumes that the objective is strongly convex and shows that Normalized ExtraPush produces a bounded sequence that converges linearly. We conclude this paper in section 6.

Notation: Let $\mathbf{I}_n$ denote an identity matrix with the size $n\times n$, and $\mathbf{1}_{n\times p}\in\mathbf{R}^{n\times p}$ denote the *matrix* with all entries equal to 1. We also use $\mathbf{1}_n\in\mathbf{R}^n$ as a vector of all 1's. For any *vector* x, we let x_i denote its ith component and $\mathbf{diag}(x)$ 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 $\|X\|\triangleq\sqrt{\langle X,X\rangle}=\sqrt{\sum_{i,j}X_{ij}^2}$ denotes its Frobenius norm. The largest and smallest eigenvalues of matrix X are denoted as $\lambda_{\max}(X)$ and $\lambda_{\min}(X)$, respectively. For any matrix $B\in\mathbf{R}^{n\times p}$, $\mathbf{null}(B)\triangleq\{x\in\mathbf{R}^p|Bx=0\}$ is the null space of B. The smallest *nonzero* eigenvalue of a symmetric positive semidefinite matrix $X\neq\mathbf{0}$ is denoted as $\tilde{\lambda}_{\min}(X)$, which is strictly positive. For any positive semidefinite matrix $G\in\mathbf{R}^{n\times n}$ (not necessarily symmetric in this paper), we use the notion $\|X\|_G^2\triangleq\langle X,GX\rangle$ for a matrix $X\in\mathbf{R}^{n\times p}$.

2. Problem reformulation.

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

$$\mathcal{N}_i^{\text{in}}\triangleq\{j:(j,i)\in E\}\cup\{i\},\quad \mathcal{N}_i^{\text{out}}\triangleq\{j:(i,j)\in E\}\cup\{i\},$$

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

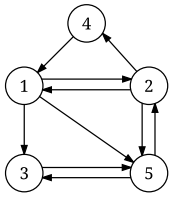
To illustrate a mixing matrix for a directed network, consider $A\in\mathbf{R}^{n\times n}$ where

$$\begin{cases} A_{ij}>0, & \text{if } j\in\mathcal{N}_i^{\text{in}}\\ A_{ij}=0, & \text{otherwise.}\end{cases}\tag{2.1}$$

The entries A_{ij} satisfy that, for each node j, $\sum_{i\in V}A_{ij}=1$; hence, for each j, $\{A_{ij}\}_{i\in V}$ is a discrete distribution. An example is the following mixing matrix

$$A_{ij}=\begin{cases} 1/d_j, & \text{if } j\in\mathcal{N}_i^{in}\\ 0, & \text{otherwise}\end{cases},\tag{2.2}$$

$i,j=1,\ldots,n$, which is used in the subgradient-push method [10]. See Fig. 1 for a directed graph $\mathcal{G}$ and an example of its mixing matrix A. The matrix A is column stochastic and asymmetric in general.



$$A = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{3} \\ \frac{1}{4} & 0 & \frac{1}{2} & 0 & \frac{1}{3} \\ 0 & \frac{1}{4} & 0 & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 & \frac{1}{3} \end{pmatrix}$$

FIG. 1. *A directed graph $\mathcal{G}$ (left) and its mixing matrix A (right).*

ASSUMPTION 1. *The graph $\mathcal{G}$ is strongly connected.*

PROPERTY 1. *Under Assumption 1, the followings hold (parts (i) and (iv) are results in [10, Corollary 2]):*

(i) Let $A^t = \overbrace{A \times A \cdots A}^{t}$ *for any* $t \in \mathbf{N}$. *Then*

$$A^t \to \phi \mathbf{1}_n^T \quad \text{geometrically fast as} \quad t \to \infty, \tag{2.3}$$

for some stationary distribution *vector* ϕ, *i.e.,* $\phi_i \geq 0$ *and* $\sum_i^n \phi_i = 1$.

(ii) $\mathbf{null}(\mathbf{I}_n - \phi \mathbf{1}_n^T) = \mathbf{null}(\mathbf{I}_n - A)$.

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

(iv) The quantity $\xi \triangleq \inf_t \min_{1 \leq i \leq n} (A^t \mathbf{1}_n)_i \geq \frac{1}{n^n} > 0$.

Proof. Part (iii) is obvious from (ii) since $\phi \in \mathbf{null}(\mathbf{I}_n - \phi \mathbf{1}_n^T)$. Next, we show part (ii). First, let $z \in \mathbf{null}(\mathbf{I}_n - \phi \mathbf{1}_n^T)$, which means $z = \phi \mathbf{1}_n^T z$ and thus $Az = A\phi \mathbf{1}_n^T z$. By (2.3), it is obvious that $A\phi \mathbf{1}_n^T = \phi \mathbf{1}_n^T$. Therefore, $Az = \phi \mathbf{1}_n^T z = z$ and hence $\mathbf{null}(\mathbf{I}_n - \phi \mathbf{1}_n^T) \subseteq \mathbf{null}(\mathbf{I}_n - A)$. On the other hand, any $z \in \mathbf{null}(\mathbf{I}_n - A)$, equivalently, $z = Az$, obeys $z = A^t z$ for any $t \geq 1$. Letting $t \to \infty$, it holds that $z = \phi \mathbf{1}_n^T z$, that is, $z \in \mathbf{null}(\mathbf{I}_n - \phi \mathbf{1}_n^T)$. Therefore, part (ii) holds. ∎

2.2. Problem with matrix notation. Let $x_{(i)} \in \mathbf{R}^p$ 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 (1.1) using local copies of the variable x:

$$\text{minimize}_{\mathbf{x}} \ \mathbf{1}_n^T \mathbf{f}(\mathbf{x}) \triangleq \sum_{i=1}^n f_i(x_{(i)}), \quad \text{subject to } x_{(i)} = x_{(j)}, \ \forall i, j \in E, \tag{2.4}$$

where $\mathbf{1}_n \in \mathbf{R}^n$ denotes the vector with all its entries equal to 1,

$$\mathbf{x} \triangleq \begin{pmatrix} — & x_{(1)}^T & — \\ — & x_{(2)}^T & — \\ & \vdots & \\ — & x_{(n)}^T & — \end{pmatrix} \in \mathbf{R}^{n \times p}, \quad \mathbf{f}(\mathbf{x}) \triangleq \begin{pmatrix} f_1(x_{(1)}) \\ f_2(x_{(2)}) \\ \vdots \\ f_n(x_{(n)}) \end{pmatrix} \in \mathbf{R}^n.$$

In addition, the gradient of $\mathbf{f}(\mathbf{x})$ is

$$\nabla \mathbf{f}(\mathbf{x}) \triangleq \begin{pmatrix} — & \nabla f_1(x_{(1)})^T & — \\ — & \nabla f_2(x_{(2)})^T & — \\ & \vdots & \\ — & \nabla f_n(x_{(n)})^T & — \end{pmatrix} \in \mathbf{R}^{n \times p}.$$

The ith rows of the above matrices $\mathbf{x}$ and $\nabla \mathbf{f}(\mathbf{x})$, and vector $\mathbf{f}(\mathbf{x})$, correspond to agent i. For simplicity, one can treat $p = 1$ throughout this paper.

For a vector $\bar{x} \in \mathbf{R}^n$, let $\bar{x}^{\text{ave}} \triangleq \frac{1}{n}(\sum_{i=1}^n \bar{x}_i) \in \mathbf{R}$. A special case of (2.4) is the well-known average consensus problem, where $f_i(x_{(i)}) = \frac{1}{2}|x_{(i)} - \bar{x}_i|^2$ for each node i and the solution is $x_{(i)} = \bar{x}^{\text{ave}}$ for all i.

3. Proposed algorithms.

3.1. Review of Extra and push-sum. Extra [13] is a "two-step" iterative algorithm for solving (2.4) over an undirected network. Let $W \in \mathbf{R}^{n\times n}$ be a symmetric and doubly stochastic mixing matrix, and $\bar{W} \triangleq \frac{\mathbf{I}_n+W}{2}$. The Extra iteration is

$$\mathbf{x}^{t+2} = (\mathbf{I}_n + W)\mathbf{x}^{t+1} - \bar{W}\mathbf{x}^t - \alpha(\nabla \mathbf{f}(\mathbf{x}^{t+1}) - \nabla \mathbf{f}(\mathbf{x}^t)),\ t = 0, 1, \cdots, \tag{3.1}$$

which starts with $\mathbf{x}^1 = \mathbf{x}^0 - \alpha\nabla \mathbf{f}(\mathbf{x}^0)$, any $\mathbf{x}^0 \in \mathbf{R}^{n\times p}$ and uses a properly bounded step size $\alpha > 0$. Extra converges at a rate $o(\frac{1}{t})$, measured by the best running violation to the first-order optimality condition, provided that f is Lipschitz differentiable. It improves to a linear rate of convergence if f is also (restricted) strongly convex.

The push-sum protocol (also known as *weighted gossip* or *sum-weight* algorithms) computes the average of $x^0_{(1)}, \ldots, x^0_{(n)} \in \mathbf{R}^p$ over a directed and possibly time-varying network with n nodes. It can be traced back to [6] and analyzed theoretically in [1], [7]. Let K be a column-stochastic mixing matrix. The push-sum iteration is

$$\begin{cases} \mathbf{z}^{t+1} = K\mathbf{z}^t \\ \mathbf{w}^{t+1} = K\mathbf{w}^t \\ \mathbf{x}^{t+1} = \mathbf{diag}(\mathbf{w}^{t+1})^{-1}\mathbf{z}^{t+1} \end{cases} \tag{3.2}$$

for $t = 0, 1, \cdots$, $\mathbf{z}^0 = \mathbf{x}^0$ and $\mathbf{w}^0 = \mathbf{1}_n$. By Property 1 (iv), $\mathbf{w}^t$ is uniformly positive as long as the underlying graph is strongly connected, ensuring the last step is well defined. Under assumptions, push-sum can converge to the average consensus geometrically fast [1].

3.2. Proposed: ExtraPush. ExtraPush combines the above two algorithms. Specifically, set arbitrary $\mathbf{z}^0$ and $\mathbf{w}^0 = \mathbf{1}_n$; set $\mathbf{x}^0 = \mathbf{z}^0$; for $t = 1$, set $\mathbf{w}^1 = A\mathbf{w}^0$, $\mathbf{z}^1 = A\mathbf{z}^0 - \alpha\nabla \mathbf{f}(\mathbf{x}^0)$, and $\mathbf{x}^1 = \mathbf{diag}(\mathbf{w}^1)^{-1}\mathbf{z}^1$. Letting $\bar{A} \triangleq \frac{\mathbf{I}_n+A}{2}$, for $t = 2, 3, \ldots$, perform

$$\begin{cases} \mathbf{z}^t = (A + \mathbf{I}_n)\mathbf{z}^{t-1} - \bar{A}\mathbf{z}^{t-2} - \alpha(\nabla \mathbf{f}(\mathbf{x}^{t-1}) - \nabla \mathbf{f}(\mathbf{x}^{t-2})), \\ \mathbf{w}^t = A\mathbf{w}^{t-1}, \\ \mathbf{x}^t = \mathbf{diag}(\mathbf{w}^t)^{-1}\mathbf{z}^t. \end{cases} \tag{3.3}$$

By the structure of A, each node i broadcasts its $z_{(i)}$ to its out-neighbors at each ExtraPush iteration. The step size $\alpha > 0$ needs to be properly set.

3.3. Proposed: Normalized ExtraPush. In this algorithm, we first compute the stationary distribution ϕ of A and save each ϕ_i at node i. Then, in the main iteration, we can eliminate the $\mathbf{w}$-step from (3.3) and use $n \cdot \phi$ instead of $\mathbf{w}^t$ to obtain $\mathbf{x}^t$. As such, the main iteration of Normalized ExtraPush simplifies (3.3). Letting,

$$D \triangleq n\mathbf{diag}(\phi).$$

we describe the iteration of Normalized ExtraPush as follows: set arbitrary $\mathbf{z}^0$ and $\mathbf{x}^0 = D^{-1}\mathbf{z}^0$; for $t = 1$, set $\mathbf{z}^1 = A\mathbf{z}^0 - \alpha\nabla \mathbf{f}(\mathbf{x}^0)$ and $\mathbf{x}^1 = D^{-1}\mathbf{z}^1$. For $t = 2, 3, \ldots$, perform

$$\begin{cases} \mathbf{z}^t = (A + \mathbf{I})\mathbf{z}^{t-1} - \bar{A}\mathbf{z}^{t-2} - \alpha(\nabla \mathbf{f}(\mathbf{x}^{t-1}) - \nabla \mathbf{f}(\mathbf{x}^{t-2})), \\ \mathbf{x}^t = D^{-1}\mathbf{z}^t. \end{cases} \tag{3.4}$$

Next, we present two equivalent forms of Normalized ExtraPush. Letting $\mathbf{f}_\phi(\mathbf{z}) \triangleq D\mathbf{f}(D^{-1}\mathbf{z})$, we have $\nabla\mathbf{f}_\phi(\mathbf{z}) = \nabla\mathbf{f}(D^{-1}\mathbf{z})$. Substituting the $\mathbf{x}$-step of (3.4) into its $\mathbf{z}$-step yields the single-value iteration:

$$\mathbf{z}^t = (A+\mathbf{I})\mathbf{z}^{t-1} - \bar{A}\mathbf{z}^{t-2} - \alpha(\nabla\mathbf{f}_\phi(\mathbf{z}^{t-1}) - \nabla\mathbf{f}_\phi(\mathbf{z}^{t-2})). \tag{3.5}$$

Upon stopping, one shall return $\mathbf{x}^t = D^{-1}\mathbf{z}^t$. The iteration (3.5) is nearly identical to the Extra iteration (3.1) except that (3.1) must use a doubly-stochastic matrix.

Letting $A_\phi \triangleq D^{-1}AD$ and $\bar{A}_\phi \triangleq \frac{1}{2}(\mathbf{I}_n + A_\phi)$, which are row stochastic matrices, gives another equivalent form of normalized Extra

$$\mathbf{x}^t = (A_\phi+\mathbf{I})\mathbf{x}^{t-1} - \bar{A}_\phi\mathbf{x}^{t-2} - \alpha D^{-1}(\nabla\mathbf{f}(\mathbf{x}^{t-1}) - \nabla\mathbf{f}(\mathbf{x}^{t-2})), \tag{3.6}$$

which, compared to the Extra iteration (3.1), has the extra diagonal matrix D^{-1}. Indeed, this iteration generalizes Extra to use row-stochastic matrices A_ϕ and $\bar{A}$.

4. Preliminary analysis of ExtraPush. In this section, we first develop the first-order optimality conditions for the problem (2.4) and then provide the convergence of ExtraPush under the boundedness assumption.

THEOREM 1. **(First-order optimality conditions)** *Suppose that graph $\mathcal{G}$ is strongly connected. Then $\mathbf{x}^*$ is consensual and $x^*_{(1)} \equiv x^*_{(2)} \equiv \cdots \equiv x^*_{(n)}$ is an optimal solution of (1.1) if and only if, for some $\alpha > 0$, there exist $\mathbf{z}^* \in \mathbf{null}(\mathbf{I}_n - A)$ and $\mathbf{y}^* \in \mathbf{null}(\mathbf{1}_n)$ such that the following conditions hold*

$$\begin{cases} \mathbf{y}^* + \alpha\nabla\mathbf{f}(\mathbf{x}^*) = 0, \\ \mathbf{x}^* = D^{-1}\mathbf{z}^*. \end{cases} \tag{4.1}$$

(We let $\mathcal{L}^*$ denote the set of triples $(\mathbf{z}^*, \mathbf{y}^*, \mathbf{x}^*)$ satisfying the above conditions.)

Proof. Assume that $\mathbf{x}^*$ is consensual and $x^*_{(1)} \equiv x^*_{(2)} \equiv \cdots \equiv x^*_{(n)}$ is optimal. Let $\mathbf{z}^* = n\mathbf{diag}(\phi)\mathbf{x}^* = n(\phi x^{*T}_{(1)})$. Then $\phi\mathbf{1}_n^T\mathbf{z}^* = \phi\mathbf{1}_n^T n\phi x^{*T}_{(1)} = n\phi x^{*T}_{(1)} = \mathbf{z}^*$. It implies that $\mathbf{z}^* \in \mathbf{null}(\mathbf{I} - \phi\mathbf{1}_n^T)$. By Property 1 (ii), it follows that $\mathbf{z}^* \in \mathbf{null}(\mathbf{I}_n - A)$. Moreover, letting $\mathbf{y}^* = -\alpha\nabla\mathbf{f}(\mathbf{x}^*)$, it holds that $\mathbf{1}_n^T\mathbf{y}^* = -\alpha\mathbf{1}_n^T\nabla\mathbf{f}(\mathbf{x}^*) = 0$, that is, $\mathbf{y}^* \in \mathbf{null}(\mathbf{1}_n)$.

On the other hand, assume (4.1) holds. By Property 1 (ii), it follows that $\mathbf{z}^* = \phi\mathbf{1}_n^T\mathbf{z}^*$. Plugging $\mathbf{x}^* = D^{-1}\mathbf{z}^*$ gives $\mathbf{x}^* = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\mathbf{z}^*$, which implies that $\mathbf{x}^*$ is consensual. Moreover, by $\mathbf{y}^* + \alpha\nabla\mathbf{f}(\mathbf{x}^*) = 0$ and $\mathbf{y}^* \in \mathbf{null}(\mathbf{1}_n)$, it holds $\mathbf{1}_n^T\nabla\mathbf{f}(\mathbf{x}^*) = -\frac{1}{\alpha}\mathbf{1}_n^T\mathbf{y}^* = 0$, which implies that $\mathbf{x}^*$ is optimal. ∎

Introducing the sequence

$$\mathbf{y}^t \triangleq \sum_{k=0}^{t}(\bar{A} - A)\mathbf{z}^k, \tag{4.2}$$

the iteration (3.3) of ExtraPush can be rewritten as

$$\begin{cases} \bar{A}\mathbf{z}^{t+1} = \bar{A}\mathbf{z}^t - \alpha\nabla\mathbf{f}(\mathbf{x}^t) - \mathbf{y}^{t+1}, \\ \mathbf{y}^{t+1} = \mathbf{y}^t + (\bar{A} - A)\mathbf{z}^{t+1}, \\ \mathbf{w}^{t+1} = A\mathbf{w}^t, \\ \mathbf{x}^{t+1} = \mathbf{diag}(\mathbf{w}^{t+1})^{-1}\mathbf{z}^{t+1}. \end{cases} \tag{4.3}$$

THEOREM 2. *Suppose that the sequence $\{(\mathbf{z}^t, \mathbf{x}^t)\}$ generated by ExtraPush (3.3) and the sequence $\{\mathbf{y}^t\}$ defined in (4.2) are bounded. Then, any limit point of $\{(\mathbf{z}^t, \mathbf{y}^t, \mathbf{x}^t)\}$, $(\mathbf{z}^*, \mathbf{y}^*, \mathbf{x}^*)$, satisfies the optimality conditions (4.1).*

Proof. By Property 1, $\{\mathbf{w}^t\}$ is also bounded. Hence, there exists a convergent subsequence $\{(\mathbf{z},\mathbf{y},\mathbf{w},\mathbf{x})^{t_j}\}_{j=1}^{\infty}$. Let $(\mathbf{z}^*,\mathbf{y}^*,\mathbf{w}^*,\mathbf{x}^*)$ be its limit. By (2.3), we know that $\mathbf{w}^* = n\phi$ and thus that $\mathbf{x}^* = D^{-1}\mathbf{z}^*$. Letting $t\to\infty$ in the second equation of (4.3) gives $\mathbf{z}^* = A\mathbf{z}^*$, or equivalently $\mathbf{z}^* \in \mathbf{null}(\mathbf{I}_n - A)$. Similarly, letting $t\to\infty$ in the first equation of (4.3) yields $\mathbf{y}^* + \alpha\nabla\mathbf{f}(\mathbf{x}^*) = 0$. Moreover, from the definition (4.2) of $\mathbf{y}^t$ and the facts that both A and $\bar{A}$ are column stochastic, it follows that $\mathbf{1}_n^T\mathbf{y}^* = 0$ and $\mathbf{1}_n^T\nabla\mathbf{f}(\mathbf{x}^*) = 0$. Therefore, $(\mathbf{z}^*,\mathbf{y}^*,\mathbf{x}^*)$ satisfies the optimality conditions (4.1). ∎

5. Convergence of Normalized ExtraPush. In this section, we show the linear convergence of Normalized ExtraPush under the smoothness and strong convexity assumptions of the objective function. Similar to (4.3), introducing a new sequence $\mathbf{y}^t = \sum_{k=0}^{t}(\bar{A}-A)\mathbf{z}^k$, the iterative formula (3.4) of Normalized ExtraPush implies

$$\begin{cases} \bar{A}\mathbf{z}^{t+1} = \bar{A}\mathbf{z}^t - \alpha\nabla\mathbf{f}(\mathbf{x}^t) - \mathbf{y}^{t+1}, \\ \mathbf{y}^{t+1} = \mathbf{y}^t + (\bar{A}-A)\mathbf{z}^{t+1}, \\ \mathbf{x}^{t+1} = D^{-1}\mathbf{z}^{t+1}. \end{cases} \tag{5.1}$$

THEOREM 3. *Suppose that sequence $\{(\mathbf{z}^t,\mathbf{x}^t)\}$ generated by Normalized ExtraPush (3.4) is bounded, and that the sequence $\{\mathbf{y}^t\}$ is also bounded, then any limit point of $\{(\mathbf{z}^t,\mathbf{y}^t,\mathbf{x}^t)\}_{t=0}^{\infty}$, $(\mathbf{z}^*,\mathbf{y}^*,\mathbf{x}^*)$ satisfies the first-order optimality conditions (4.1).*

The proof is very similar to that of Theorem 2. It only needs to replace the sequence $\{\mathbf{w}^t\}$ with its limitation $n\phi$ in the proof procedure, thus we omit it here. From Theorem 3, it shows that Normalized ExtraPush has subsequence convergence to an optimal solution of the considered optimization problem under the boundedness assumption. To obtain the linear convergence of Normalized ExtraPush, we still need the following assumptions.

ASSUMPTION 2. **(existence of solution)** *Let $\mathcal{X}^*$ be the optimal solution set of problem (1.1), and assume that $\mathcal{X}^*$ is nonempty.*

ASSUMPTION 3. *For each agent i, its objective function f_i satisfies the following:*

(i) **(Lipschitz differentiability)** *f_i is differentiable, and its gradient ∇f_i is L_i-Lipschitz continuous, i.e., $\|\nabla f_i(x) - \nabla f_i(y)\| \le L_i\|x-y\|, \forall x,y\in\mathbf{R}^p$;*

(ii) **(quasi-strong convexity)** *f_i is quasi-strongly convex, and there exists a positive constant S_i such that $S_i\|x^*-y\|^2 \le \langle\nabla f_i(x^*) - \nabla f_i(y), x^*-y\rangle$ for any $y\in\mathbf{R}^p$ and some optimal value $x^*\in\mathcal{X}^*$.*

Following Assumption 3, there hold for any $\mathbf{x},\mathbf{y}\in\mathbf{R}^{n\times p}$ and some $\mathbf{x}^* \equiv \mathbf{1}_n(x^*)^T$

$$\|\nabla\mathbf{f}(\mathbf{x}) - \nabla\mathbf{f}(\mathbf{y})\| \le L_f\|\mathbf{x}-\mathbf{y}\|, \tag{5.2}$$

$$S_f\|\mathbf{x}^*-\mathbf{y}\|^2 \le \langle\nabla\mathbf{f}(\mathbf{x}^*) - \nabla\mathbf{f}(\mathbf{y}), \mathbf{x}^*-\mathbf{y}\rangle, \tag{5.3}$$

where the constants $L_f \triangleq \max_i L_i$ and $S_f \triangleq \min_i S_i$.

ASSUMPTION 4. **(Positive Definiteness)** $D^{-1}\bar{A} + \bar{A}^T D^{-1} \succ 0$.

Note that $\bar{A}_{ii} > \sum_{j\neq i}\bar{A}_{ij}$ for each i. It means that $\bar{A}$ is strictly column-diagonal dominant. To ensure the positive definiteness of $\bar{A}+\bar{A}^T$, each node j can be more "selfish" and take a sufficiently large A_{jj}.

Before presenting the main result, we introduce the following notation. For each t, introducing $\mathbf{u}^t = \sum_{k=0}^{t}\mathbf{z}^k$, then the Normalized ExtraPush iteration (3.4) gives

$$\begin{cases} \bar{A}\mathbf{z}^{t+1} = \bar{A}\mathbf{z}^t - \alpha\nabla\mathbf{f}(\mathbf{x}^t) - (\bar{A}-A)\mathbf{u}^{t+1} \\ \mathbf{u}^{t+1} = \mathbf{u}^t + \mathbf{z}^{t+1} \\ \mathbf{x}^{t+1} = D^{-1}\mathbf{z}^{t+1}. \end{cases} \tag{5.4}$$

Let $(\mathbf{z}^*, \mathbf{y}^*, \mathbf{x}^*) \in \mathcal{L}^*$, where $\mathbf{x}^*$ has been specified in (5.3). Let $\mathbf{u}^*$ be any matrix that satisfies $(\bar{A} - A)\mathbf{u}^* = \mathbf{y}^*$. For simplicity, we introduce

$$\mathbf{v}^t = \begin{pmatrix} \mathbf{z}^t \\ \mathbf{u}^t \end{pmatrix},\ \mathbf{v}^* = \begin{pmatrix} \mathbf{z}^* \\ \mathbf{u}^* \end{pmatrix},\ G = \begin{pmatrix} N^T & \\ & M \end{pmatrix},\ S = \begin{pmatrix} & M \\ -M^T & \end{pmatrix}, \tag{5.5}$$

where $N = D^{-1}\bar{A}$, $M = D^{-1}(\bar{A} - A)$. Let $\mathbf{f}_D(\mathbf{z}) \triangleq \mathbf{f}(D^{-1}\mathbf{z})$ and $\bar{\mathbf{f}}(\mathbf{v}) \triangleq \mathbf{f}_D(\mathbf{z})$. Then $\nabla\bar{\mathbf{f}}(\mathbf{v}) = [\nabla\mathbf{f}_D(\mathbf{z}), 0]$. By (5.2) and (5.3), there hold

$$\|\nabla\bar{\mathbf{f}}(\mathbf{v}_1) - \nabla\bar{\mathbf{f}}(\mathbf{v}_2)\| = \|\nabla\mathbf{f}_D(\mathbf{z}_1) - \nabla\mathbf{f}_D(\mathbf{z}_2)\| \le \bar{L}\|\mathbf{z}_1 - \mathbf{z}_2\|, \tag{5.6}$$

$$\bar{\mu}\|\mathbf{z}^* - \mathbf{z}\|^2 \le \langle\nabla\mathbf{f}_D(\mathbf{z}^*) - \nabla\mathbf{f}_D(\mathbf{z}), \mathbf{z}^* - \mathbf{z}\rangle = \langle\nabla\bar{\mathbf{f}}(\mathbf{v}^*) - \nabla\bar{\mathbf{f}}(\mathbf{v}), \mathbf{v}^* - \mathbf{v}\rangle, \tag{5.7}$$

where $\bar{L} \triangleq \frac{L_f}{\sigma^2_{\min}(D)}$ and $\bar{\mu} \triangleq \frac{S_f}{\sigma^2_{\max}(D)}$. By (5.4) and (5.5), the Normalized ExtraPush iteration (3.4) implies

$$G^T(\mathbf{v}^{t+1} - \mathbf{v}^t) = -S\mathbf{v}^{t+1} - \alpha\nabla\bar{\mathbf{f}}(\mathbf{v}^t). \tag{5.8}$$

Next, we will show that $G + G^T$ is positive semidefinite, which by Assumption 4 implies that $N + N^T$ is positive definite. It is sufficient to show that $M + M^T$ is positive semidefinite. Note that

$$\begin{aligned} M + M^T &= \frac{D^{-1}(\mathbf{I}_n - A)}{2} + \frac{(\mathbf{I}_n - A)^T D^{-1}}{2} \\ &= D^{-1/2}\left(\mathbf{I}_n - \frac{D^{1/2}A^T D^{-1/2} + D^{-1/2}AD^{1/2}}{2}\right)D^{-1/2} \triangleq D^{-1/2}\Lambda D^{-1/2}, \end{aligned}$$

and by Property 1 (iii), $n\phi^T$ is the left eigenvector of A^T corresponding to eigenvalue 1, and thus, Λ is the Laplacian of a certain directed graph $\mathcal{G}'$ with A^T being its corresponding transition probability matrix [3]. It follows that $0 = \lambda_1(\Lambda) \le \lambda_2(\Lambda) \le \cdots \le \lambda_n(\Lambda)$, where $\lambda_i(\Lambda)$ denotes the ith eigenvalue of Λ. Therefore, $M + M^T$ is positive semidefinite, and the following property holds

$$\|x\|_G^2 = \frac{1}{2}\|x\|^2_{G+G^T} \ge 0, \quad \forall x \in \mathbf{R}^n.$$

Let $c_1 = \frac{\lambda_{\max}(MM^T)}{\lambda_{\min}(M^TM)}$, $c_2 = \frac{\lambda_{\max}(\frac{M+M^T}{2})}{\lambda_{\min}(M^TM)}$, and $c_3 = \lambda_{\max}(NN^T) + 3c_1\lambda_{\max}(N^TN)$. Let $\Delta_1 = (\bar{\mu} - \frac{\eta}{2})^2 - 6c_1\bar{L}^2$, and $\Delta_2 = \frac{\bar{L}^4}{4\eta^2} - 3c_1\bar{L}^2\sigma(c_3\sigma - \lambda_{\min}(N^T + N))$ for some appropriate tunable parameters η and σ. Then we describe our main result as follows.

THEOREM 4. *Under Assumptions 1-4, if the step size parameter α satisfies*

$$\frac{\bar{\mu} - \frac{\eta}{2} - \sqrt{\Delta_1}}{3c_1\bar{L}^2\sigma} < \alpha < \min\Big\{\frac{\bar{\mu} - \frac{\eta}{2} + \sqrt{\Delta_1}}{3c_1\bar{L}^2\sigma}, \frac{-\frac{\bar{L}^2}{2\eta} + \sqrt{\Delta_2}}{3c_1\bar{L}^2\sigma}\Big\} \tag{5.9}$$

for some appropriate η and σ as specified in (5.25) *and* (5.26)*, respectively, then the sequence $\{\mathbf{v}^t\}$ defined in* (5.5) *satisfies*

$$\|\mathbf{v}^t - \mathbf{v}^*\|_G^2 \ge (1 + \delta)\|\mathbf{v}^{t+1} - \mathbf{v}^*\|_G^2, \tag{5.10}$$

for $\delta > 0$ obeying

$$0 < \delta \le \min\left\{\frac{-\frac{1}{\sigma} + (\bar{\mu} - \frac{\eta}{2})\alpha - \frac{3}{2}c_1\bar{L}^2\sigma\alpha^2}{\lambda_{\max}(\frac{N+N^T}{2}) + 3c_2\alpha^2\bar{L}^2}, \frac{\lambda_{\min}(\frac{N^T+N}{2}) - \frac{c_3\sigma}{2} - \frac{\bar{L}^2\alpha}{2\eta} - \frac{3}{2}c_1\bar{L}^2\sigma\alpha^2}{3c_2(\lambda_{\max}(N^TN) + \alpha^2\bar{L}^2)}\right\}.$$

From this theorem, the sequence $\{\mathbf{v}^t\}$ converges to $\mathbf{v}^*$ at a linear rate in the sense of "G-norm". By the definition of $\mathbf{v}^*$ in (5.5), $\mathbf{v}^*$ is indeed defined by some optimal value $(\mathbf{z}^*, \mathbf{y}^*, \mathbf{x}^*)$. Roughly speaking, bigger δ means faster convergence rate. As specified in Theorem 4, δ is affected by many factors. Generally, δ decreases with respect to both $\lambda_{\max}(\frac{N+N^T}{2})$ and $\lambda_{\max}(N^TN)$, which potentially implies that if all nodes are more "selfish", that is, they hold more information for themselves than sending to their out-neighbors. Consequently, the information mixing speed of the network will get smaller, and thus the convergence of Normalized ExtraPush becomes slower. Therefore, we suggest a more democratic rule (such as the matrix A specified in (2.2)) for faster convergence in practice. To ensure $\delta > 0$, it requires that the step size α lie in an appropriate interval.

To prove Theorem 4, we need the following lemmas.

LEMMA 1. *For any $(\mathbf{z}^*, \mathbf{y}^*, \mathbf{x}^*) \in \mathcal{L}^*$, let $\mathbf{u}^*$ satisfy $(\bar{A} - A)\mathbf{u}^* = \mathbf{y}^*$. Then there hold*

$$M\mathbf{z}^* = \mathbf{0}, \tag{5.11}$$

$$M^T\mathbf{z}^* = \mathbf{0}, \tag{5.12}$$

$$S\mathbf{v}^* + \alpha\nabla\bar{\mathbf{f}}(\mathbf{v}^*) = 0. \tag{5.13}$$

Proof. By the optimality of $(\mathbf{z}^*, \mathbf{y}^*, \mathbf{x}^*)$, the followings hold: (i) $(\bar{A} - A)\mathbf{z}^* = 0$, and thus $M\mathbf{z}^* = 0$; (ii) $D^{-1}\mathbf{z}^* = \mathbf{x}^*$ is consensual; from the column stochasticity of both A and $\bar{A}$, it follows $M^T\mathbf{z}^* = (\bar{A} - A)^T\mathbf{x}^* = 0$; (iii) $M\mathbf{u}^* + \alpha\nabla\mathbf{f}_D(\mathbf{z}^*) = D^{-1}\mathbf{y}^* + \alpha D^{-1}\nabla\mathbf{f}(\mathbf{x}^*) = 0$, with $M^T\mathbf{z}^* = 0$, which imply $S\mathbf{v}^* + \alpha\nabla\bar{\mathbf{f}}(\mathbf{v}^*) = 0$. ∎

LEMMA 2. *For any $t \in \mathbf{N}$, it holds*

$$N(\mathbf{z}^{t+1} - \mathbf{z}^t) = -M(\mathbf{u}^{t+1} - \mathbf{u}^*) - \alpha[\nabla\mathbf{f}_D(\mathbf{z}^t) - \nabla\mathbf{f}_D(\mathbf{z}^*)]. \tag{5.14}$$

This lemma follows from (5.4) and the fact $M\mathbf{u}^* + \alpha\nabla\mathbf{f}_D(\mathbf{z}^*) = 0$ in the last lemma.

LEMMA 3. *Let $\{\mathbf{v}^t\}$ be a sequence generated by the iteration (5.8) and $\mathbf{v}^*$ be defined in (5.5). Then, it holds*

$$\begin{aligned}\|\mathbf{v}^{t+1} - \mathbf{v}^*\|_G^2 - \|\mathbf{v}^t - \mathbf{v}^*\|_G^2 \le &-\|\mathbf{v}^{t+1} - \mathbf{v}^t\|_G^2 + \|\mathbf{z}^{t+1} - \mathbf{z}^t\|^2_{\frac{\sigma}{2}NN^T + \frac{\alpha\bar{L}^2}{2\eta}\mathbf{I}_n} \\ &- \|\mathbf{z}^* - \mathbf{z}^{t+1}\|^2_{(\alpha\bar{\mu} - \frac{\alpha\eta}{2} - \frac{1}{\sigma})\mathbf{I}_n} + \frac{\sigma}{2}\|\mathbf{u}^* - \mathbf{u}^{t+1}\|^2_{MM^T},\end{aligned} \tag{5.15}$$

where $\sigma, \eta > 0$ are two tunable parameters.

Proof. Note that

$$\begin{aligned}\|\mathbf{v}^{t+1} - \mathbf{v}^*\|_G^2 - \|\mathbf{v}^t - \mathbf{v}^*\|_G^2 = &-\|\mathbf{v}^{t+1} - \mathbf{v}^t\|_G^2 + \langle\mathbf{v}^* - \mathbf{v}^{t+1}, G(\mathbf{v}^t - \mathbf{v}^{t+1})\rangle \\ &+ \langle\mathbf{v}^* - \mathbf{v}^{t+1}, G^T(\mathbf{v}^t - \mathbf{v}^{t+1})\rangle.\end{aligned} \tag{5.16}$$

In the following, we analyze the two inner-product terms:

$$\begin{aligned}\langle\mathbf{v}^* - \mathbf{v}^{t+1}, G(\mathbf{v}^t - \mathbf{v}^{t+1})\rangle &= \langle\mathbf{z}^* - \mathbf{z}^{t+1}, N^T(\mathbf{z}^t - \mathbf{z}^{t+1})\rangle + \langle M^T(\mathbf{u}^* - \mathbf{u}^{t+1}), \mathbf{u}^t - \mathbf{u}^{t+1}\rangle \\ (\because (5.11), M\mathbf{z}^* = \mathbf{0}) &= \langle\mathbf{z}^* - \mathbf{z}^{t+1}, N^T(\mathbf{z}^t - \mathbf{z}^{t+1})\rangle + \langle M^T(\mathbf{u}^* - \mathbf{u}^{t+1}), \mathbf{z}^* - \mathbf{z}^{t+1}\rangle \\ &\le \frac{\sigma}{2}\|\mathbf{z}^t - \mathbf{z}^{t+1}\|^2_{NN^T} + \frac{1}{\sigma}\|\mathbf{z}^* - \mathbf{z}^{t+1}\|^2 + \frac{\sigma}{2}\|\mathbf{u}^* - \mathbf{u}^{t+1}\|^2_{MM^T},\end{aligned} \tag{5.17}$$

and

$$\begin{aligned}
\langle \mathbf{v}^* - \mathbf{v}^{t+1}, G^T(\mathbf{v}^t - \mathbf{v}^{t+1})\rangle &= \langle \mathbf{v}^* - \mathbf{v}^{t+1}, S\mathbf{v}^{t+1} + \alpha\nabla\bar{\mathbf{f}}(\mathbf{v}^t)\rangle \ (\because (5.5)) \\
&= \langle \mathbf{v}^* - \mathbf{v}^{t+1}, S(\mathbf{v}^{t+1} - \mathbf{v}^*) + \alpha(\nabla\bar{\mathbf{f}}(\mathbf{v}^t) - \nabla\bar{\mathbf{f}}(\mathbf{v}^*))\rangle \ (\because (5.13)) \\
(\because S = -S^T) \ &= \alpha\langle \mathbf{v}^* - \mathbf{v}^{t+1}, \nabla\bar{\mathbf{f}}(\mathbf{v}^t) - \nabla\bar{\mathbf{f}}(\mathbf{v}^*)\rangle \\
&= \alpha\langle \mathbf{v}^* - \mathbf{v}^{t+1}, \nabla\bar{\mathbf{f}}(\mathbf{v}^{t+1}) - \nabla\bar{\mathbf{f}}(\mathbf{v}^*)\rangle + \alpha\langle \mathbf{v}^* - \mathbf{v}^{t+1}, \nabla\bar{\mathbf{f}}(\mathbf{v}^t) - \nabla\bar{\mathbf{f}}(\mathbf{v}^{t+1})\rangle \\
&\le -\alpha\bar{\mu}\|\mathbf{z}^{t+1} - \mathbf{z}^*\|^2 + \frac{\alpha\eta}{2}\|\mathbf{z}^* - \mathbf{z}^{t+1}\|^2 + \frac{\alpha\bar{L}^2}{2\eta}\|\mathbf{z}^t - \mathbf{z}^{t+1}\|^2. \qquad (5.18)
\end{aligned}$$

Substituting (5.17) and (5.18) into (5.16), then we can conclude the lemma. ∎

Proof. **(for Theorem 4)** In order to establish (5.10) for some constant $\delta > 0$, in light of Lemma 3, it is sufficient to show that the right-hand side of (5.15) is no more than $-\delta\|\mathbf{v}^{t+1} - \mathbf{v}^*\|_G^2$, which implies

$$\|\mathbf{z}^{t+1} - \mathbf{z}^*\|_P^2 + \|\mathbf{z}^{t+1} - \mathbf{z}^t\|_Q^2 \ge \|\mathbf{u}^{t+1} - \mathbf{u}^*\|_R^2, \qquad (5.19)$$

where $P = (\alpha\bar{\mu} - \frac{\alpha\eta}{2} - \frac{1}{\sigma})\mathbf{I}_n - \delta\frac{N+N^T}{2}$, $Q = \frac{N^T+N}{2} - \frac{\sigma}{2}NN^T - \frac{\alpha\bar{L}^2}{2\eta}\mathbf{I}_n$ and $R = \frac{\sigma}{2}MM^T + \delta(\frac{M+M^T}{2})$.

Establishing (5.19)*: Step 1.* From Lemma 2, there holds

$$\begin{aligned}
\|\mathbf{u}^* - \mathbf{u}^{t+1}\|_{M^TM}^2 &= \|M(\mathbf{u}^* - \mathbf{u}^{t+1})\|^2 \\
&= \|N(\mathbf{z}^{t+1} - \mathbf{z}^t) + \alpha[\nabla\mathbf{f}_D(\mathbf{z}^{t+1}) - \nabla\mathbf{f}_D(\mathbf{z}^*)] + \alpha[\nabla\mathbf{f}_D(\mathbf{z}^t) - \nabla\mathbf{f}_D(\mathbf{z}^{t+1})]\|^2 \\
&\le 3\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_{N^TN}^2 + 3\alpha^2\bar{L}^2\|\mathbf{z}^{t+1} - \mathbf{z}^*\|^2 + 3\alpha^2\bar{L}^2\|\mathbf{z}^{t+1} - \mathbf{z}^t\|^2 \\
&= \|\mathbf{z}^{t+1} - \mathbf{z}^*\|_{3\alpha^2\bar{L}^2\mathbf{I}_n}^2 + \|\mathbf{z}^{t+1} - \mathbf{z}^t\|_{3(N^TN+\alpha^2\bar{L}^2\mathbf{I}_n)}^2. \qquad (5.20)
\end{aligned}$$

Note that

$$\frac{\|\mathbf{u}^* - \mathbf{u}^{t+1}\|_{\frac{\sigma}{2}MM^T+\delta M}^2}{\frac{\sigma\lambda_{\max}(MM^T)}{2} + \delta\lambda_{\max}(\frac{M+M^T}{2})} \le \|\mathbf{u}^* - \mathbf{u}^{t+1}\|^2 \le \frac{\|\mathbf{u}^* - \mathbf{u}^{t+1}\|_{M^TM}^2}{\tilde{\lambda}_{\min}(M^TM)}.$$

If the following conditions hold

$$\begin{cases} P \succeq 3(\frac{1}{2}c_1\sigma + c_2\delta)\alpha^2\bar{L}^2\mathbf{I}_n \\ Q \succeq 3(\frac{1}{2}c_1\sigma + c_2\delta)(N^TN + \alpha^2\bar{L}^2\mathbf{I}_n) \end{cases}, \qquad (5.21)$$

then (5.19) holds. To show (5.21), it is sufficient to prove

$$\begin{cases} (\lambda_{\max}(\frac{N+N^T}{2}) + 3c_2\alpha^2\bar{L}^2)\delta \le -\frac{1}{\sigma} + (\bar{\mu} - \frac{\eta}{2})\alpha - \frac{3}{2}c_1\bar{L}^2\sigma\alpha^2 \\ 3c_2(\lambda_{\max}(N^TN) + \alpha^2\bar{L}^2)\delta \le \lambda_{\min}(\frac{N^T+N}{2}) - \frac{c_3\sigma}{2} - \frac{\bar{L}^2\alpha}{2\eta} - \frac{3}{2}c_1\bar{L}^2\sigma\alpha^2 \end{cases}. \qquad (5.22)$$

Let $c_4 \triangleq (\bar{\mu} - \frac{\eta}{2}) + \sqrt{\Delta_1}$, $c_5 \triangleq \frac{\bar{L}^2}{\eta}$, $c_6 \triangleq \frac{2c_4c_5+12c_1\bar{L}^2}{c_4^2}$, $c_7 \triangleq \frac{\lambda_{\min}^2(N^T+N)}{4c_3}$, $c_8 \triangleq a(c_7+2)-(2-c_7)$ for some positive constant $a \in (0,1)$, $\Delta_3 \triangleq \lambda_{\min}^2(N^T+N) - 4c_3c_6$. After reduction, we claim that if the following conditions hold

$$\frac{2-c_7}{2+c_7} < a < 1, \qquad (5.23)$$

$$\bar{\mu} > (\sqrt{\frac{6c_1}{1-a^2}} + \frac{1}{c_8}\sqrt{\frac{1-a^2}{6c_1}})\bar{L}, \qquad (5.24)$$

$$\bar{\mu}\big(1-\sqrt{1-\frac{4\bar{L}^2}{c_8\bar{\mu}^2}}\big)<\eta<\min\{\bar{\mu}\big(1+\sqrt{1-\frac{4\bar{L}^2}{c_8\bar{\mu}^2}}\big),2(\bar{\mu}-\sqrt{\frac{6c_1}{1-a^2}}\bar{L})\}, \tag{5.25}$$

$$\frac{\lambda_{\min}(N^T+N)-\sqrt{\Delta_3}}{2c_3}<\sigma<\frac{\lambda_{\min}(N^T+N)+\sqrt{\Delta_3}}{2c_3}, \tag{5.26}$$

$$\frac{\bar{\mu}-\frac{\eta}{2}-\sqrt{\Delta_1}}{3c_1\bar{L}^2\sigma}<\alpha<\min\big\{\frac{\bar{\mu}-\frac{\eta}{2}+\sqrt{\Delta_1}}{3c_1\bar{L}^2\sigma},\frac{-\frac{\bar{L}^2}{2\eta}+\sqrt{\Delta_2}}{3c_1\bar{L}^2\sigma}\big\}, \tag{5.27}$$

then (5.22) holds for some positive constant δ. Therefore, we end the proof of this theorem. □

6. Conclusion. In this note, we propose a decentralized algorithm called ExtraPush, as well as its simplified version called Normalized ExtraPush, for solving distributed consensus optimization problems over directed graphs. The algorithms use column stochastic and asymmetric mixing matrices. We show that Normalized ExtraPush converges at a linear rate if the objective function is smooth and strongly convex. In additional, we develop the first-order optimality conditions and provide the convergence of ExtraPush under the boundedness assumption. The convergence as well as the rate of convergence of ExtraPush should be justified in the future. Moreover, when applied to a directed time-varying network, the performance of the proposed algorithms will also be studied in the future. Another line of future research is to generalize ExtraPush to handle the sum of smooth and proximable (possibly nonsmooth) functions as done in [14] to generalize Extra this way.

REFERENCES

[1] F. Benezit, V. Blondel, P. Thiran, J. Tsitsiklis, and M. Vetterli, *Weighted Gossip: distributed averaging using non-doubly stochastic matrices*, ISIT, 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] F.Chung, *Laplancians and the cheeger inequality for directed graphs*, Annals of Combinatorics, 9(1):1-19, 2005.

[4] J. Duchi, A. Agarwal, and M. Wainwright, *Dual averaging for distributed optimization: Convergence analysis and network scaling*, IEEE Transactions on Automatic Control, 57: 592-606, 2012.

[5] D. Jakovetic, J. Xavier, and J. Moura, *Fast distributed gradient methods*, IEEE Transactions on Automatic Control, 59: 1131-1146, 2014.

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

[7] F. Iutzeler, P. Ciblat, and W. Hachem, *Analysis of sum-weight-like algorithms for averaging in wireless sensor networks*, IEEE Transactions on Signal Processing, 61(11): 2802-2814, 2013.

[8] I. Matei and J. Baras, *Performance evaluation of the consensus-based distributed subgradient method under random communication topologies*, IEEE J. Sel. Top. Signal Process., 5:754-771, 2011.

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

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

[11] 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.

[12] 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.

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

[14] 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.

[15] K. I. Tsianos, S. Lawlor, and M. G. 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.

[16] K. I. 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.

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

[18] C. Xi, and U.A. Khan, *On the linear convergence of distributed optimization over directed graphs*, preprint, arXiv:1510.02149, 2015.

[19] K. Yuan, Q. Ling, and W. Yin, *On the Convergence of Decentralized Gradient Descent*, preprint, arXiv:1310.7063, 2013.