EXTRA: AN EXACT FIRST-ORDER ALGORITHM FOR DECENTRALIZED CONSENSUS OPTIMIZATION[*]

WEI SHI , QING LING , GANG WU , AND WOTAO YIN

Abstract. Recently, there have been growing interests in solving consensus optimization problems in a multi-agent network. In this paper, we develop a decentralized algorithm for the consensus optimization problem

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

which is defined over a connected network of n agents, where each function f_i is held privately by agent i and encodes the agent's data and objective. All the agents shall collaboratively find the minimizer while each agent can only communicate with its neighbors. Such a computation scheme avoids a data fusion center or long-distance communication and offers better load balance to the network.

This paper proposes a novel decentralized exact first-order algorithm (abbreviated as EXTRA) to solve the consensus optimization problem. "Exact" means that it can converge to the exact solution. EXTRA can use a fixed large step size, which is independent of the network size, and has synchronized iterations. The local variable of every agent i converges uniformly and consensually to an exact minimizer of $\bar{f}$. In contrast, the well-known decentralized gradient descent (DGD) method must use diminishing step sizes in order to converge to an exact minimizer. EXTRA and DGD have the same choice of mixing matrices and similar per-iteration complexity. EXTRA, however, uses the gradients of last two iterates, unlike DGD which uses just that of last iterate.

EXTRA has the best known convergence rates among the existing first-order decentralized algorithms for decentralized consensus optimization with convex Lipschitz–differentiable objectives. Specifically, if f_i's are convex and have Lipschitz continuous gradients, EXTRA has an ergodic convergence rate $O\left(\frac{1}{k}\right)$ in terms of the first-order optimality residual. If $\bar{f}$ is also (restricted) strongly convex, EXTRA converges to an optimal solution at a linear rate $O(C^{-k})$ for some constant $C>1$.

Key words. Consensus optimization, decentralized optimization, gradient method, linear convergence

1. Introduction.

This paper focuses on *decentralized consensus optimization*, a problem defined on a connected network and solved by n agents cooperatively

$$\underset{x\in\mathbb{R}^p}{\text{minimize}}\ \ \bar{f}(x)=\frac{1}{n}\sum_{i=1}^{n}f_i(x), \tag{1.1}$$

over a common variable $x\in\mathbb{R}^p$, and for each agent i, $f_i:\mathbb{R}^p\rightarrow\mathbb{R}$ is a convex function privately known by the agent. We assume that f_i's are continuously differentiable and will introduce a novel first-order algorithm to solve (1.1) in a decentralized manner. We stick to the synchronous case in this paper, that is, all the agents carry out their iterations at the same time intervals.

Problems of the form (1.1) that require decentralized computation are found widely in various scientific and engineering areas including sensor network information processing,

[*] This work is supported by Chinese Scholarship Council (CSC) grants 201306340046 and 2011634506, NSFC grant 61004137, MOF/MIIT/MOST grant BB2100100015, and NSF grants DMS-0748839 and DMS-1317602.

multiple-agent control and coordination, as well as distributed machine learning. Examples and works include decentralized averaging [7,15,34], learning [9,22,26], estimation [1,2,16,18, 29], sparse optimization [19,35], and low-rank matrix completion [20] problems. Functions f_i can take forms of least squares [7, 15, 34], regularized least squares [1, 2, 9, 18, 22], as well as more general ones [26]. The solution x can represent, for example, the average temperature of a room [7,34], frequency-domain occupancy of spectra [1,2], states of a smart grid system [10, 16], sparse vectors [19, 35], and a matrix factor [20] and so on. In general, decentralized optimization fits the scenarios in which the data is collected and/or stored in a distributed network, a fusion center is either infeasible or not economical, and/or computing is required to be performed in a decentralized and collaborative manner by multiple agents.

1.1. Related Methods. Existing first-order decentralized methods for solving (1.1) include the (sub)gradient method [21, 25, 36], the (sub)gradient-push method [23, 24], the fast (sub)gradient method [5,14], and the dual averaging method [8]. Compared to classical centralized algorithms, decentralized algorithms encounter more restrictive assumptions and typically worse convergence rates. Most of the above algorithms are analyzed under the assumption of bounded (sub)gradients. Work [21] assumes bounded Hessian for strongly convex functions. Recent work [36] relaxes such assumptions for decentralized gradient descent. When (1.1) has additional constraints that force x in a bounded set, which also leads to bounded (sub)gradients and Hessian, projected first-order algorithms are applicable [27,37].

When using a fixed step size, these algorithms do not converge to a solution x^* of problem (1.1) but a point in its neighborhood no matter whether f_i's are differentiable or not [36]. This motivates the use of certain diminishing step sizes in [5, 8, 14] to guarantee convergence to x^*. The rates of convergence are generally weaker than their analogues in centralized computation. For the general convex case and under the bounded (sub)gradient (or Lipschitz–continuous objective) assumption, [5] shows that diminishing step sizes $\alpha_k = \frac{1}{\sqrt{k}}$ lead to a convergence rate of $O\left(\frac{\ln k}{\sqrt{k}}\right)$ in terms of the running best of objective error, and [8] shows that the dual averaging method has a rate of $O\left(\frac{\ln k}{\sqrt{k}}\right)$ in the ergodic sense in terms of objective error. For the general convex case, under assumptions of fixed step size and Lipschitz continuous, bounded gradient, [14] shows an outer–loop convergence rate of $O\left(\frac{1}{k^2}\right)$ in terms of objective error, utilizing Nesterov's acceleration, provided that the inner loop performs substantial consensus computation, without which diminishing step sizes $\alpha_k = \frac{1}{k^{1/3}}$ lead to a reduced rate of $O\left(\frac{\ln k}{k}\right)$. The (sub)gradient-push method [23] can be implemented in a dynamic digraph and, under the bounded (sub)gradient assumption and diminishing step sizes $\alpha_k = O\left(\frac{1}{\sqrt{k}}\right)$, has a rate of $O\left(\frac{\ln k}{\sqrt{k}}\right)$ in the ergodic sense in terms of objective error. A better rate of $O\left(\frac{\ln k}{k}\right)$ is proved for the (sub)gradient-push method in [24] under the strong convexity and Lipschitz gradient assumptions, in terms of expected objective error plus squared consensus residual.

Some of other related algorithms are as follows. For general convex functions and assuming closed and bounded feasible sets, the decentralized asynchronous ADMM [32] is proved to have a rate of $O\left(\frac{1}{k}\right)$ in terms of expected objective error and feasibility violation. The augmented Lagrangian based primal-dual methods have linear convergence under strong convexity and Lipschitz gradient assumptions [4,30] or under the positive-definite bounded Hessian assumption [12,13].

Our proposed algorithm is a synchronous gradient-based algorithm that has a rate of $O\left(\frac{1}{k}\right)$ for general convex objectives with Lipschitz differentials and has a linear rate once the sum of, rather than individual, functions f_i is also (restricted) strongly convex.

1.2. Notation. Throughout the paper, we let agent i hold a *local copy* of the global variable x, which is denoted by $x_{(i)} \in \mathbb{R}^p$; its value at iteration k is denoted by $x_{(i)}^k$. We introduce an aggregate objective function of the local variables

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

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 gradient of $\mathbf{f}(\mathbf{x})$ is defined by

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

Each row i of $\mathbf{x}$ and $\nabla\mathbf{f}(\mathbf{x})$ is associated with agent i. We say that $\mathbf{x}$ is *consensual* if all of its rows are identical, i.e., $x_{(1)} = \cdots = x_{(n)}$. The analysis and results of this paper hold for all $p \geq 1$. The reader can assume $p = 1$ for convenience (so $\mathbf{x}$ and $\nabla\mathbf{f}$ become vectors) without missing any major point.

Finally, for given matrix A and symmetric positive semidefinite matrix G, we define the G-matrix 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 $B \neq \mathbf{0}$ is denoted as $\tilde{\lambda}_{\min}(B)$, which is strictly positive. For a matrix $A \in \mathbb{R}^{m\times n}$, $\mathrm{null}\{A\} \triangleq \{x \in \mathbb{R}^n \big| Ax = 0\}$ is the null space of A and $\mathrm{span}\{A\} \triangleq \{y \in \mathbb{R}^m \big| y = Ax, \forall x \in \mathbb{R}^n\}$ is the linear span of all the columns of A.

1.3. Summary of Contributions. This paper introduces a novel gradient-based decentralized algorithm EXTRA, establishes its convergence conditions and rates, and presents numerical results in comparison to decentralized gradient descent. EXTRA can use a fixed step size independent of the network size and quickly converges to the solution to (1.1). It has a rate of convergence $O\left(\frac{1}{k}\right)$ in terms of best running violation to the first-order optimality condition when $\bar{f}$ is Lipschitz differentiable, and has a linear rate of convergence if $\bar{f}$ is also (restricted) strongly convex. Numerical simulations verify the theoretical results and demonstrate its competitive performance.

1.4. Paper Organization. The rest of this paper is organized as follows. Section 2 develops and interprets EXTRA. Section 3 presents its convergence results. Then, Section 4 presents three sets of numerical results. Finally, Section 5 concludes this paper.

2. Algorithm Development.

This section derives the proposed algorithm EXTRA. We start by briefly reviewing *decentralized gradient descent* (DGD) and discussing the dilemma that DGD converges slowly to an exact solution when it uses a sequence of diminishing step sizes, yet it converges faster using a fixed step size but stalls at an inaccurate solution. We then obtain the update formula of EXTRA by taking the difference of two formulas of the DGD update. Provided that the sequence generated by the new update formula with a fixed step size converges to a point, we argue that the point is consensual and optimal. Finally, we briefly discuss the choice of mixing matrices in EXTRA. Formal convergence results and proofs are left to Section 3.

2.1. Review of Decentralized Gradient Descent and Its Limitation.

DGD carries out the following iteration

$$x_{(i)}^{k+1} = \sum_{j=1}^{n} w_{ij} x_{(j)}^{k} - \alpha^k \nabla f_i(x_{(i)}^{k}), \quad \text{for agent } i = 1, \ldots, n. \tag{2.1}$$

Recall that $x_{(i)}^{k} \in \mathbb{R}^p$ is the local copy of x held by agent i at iteration k, $W = [w_{ij}] \in \mathbb{R}^{n\times n}$ is a symmetric mixing matrix satisfying $\text{null}\{I - W\} = \text{span}\{\mathbf{1}\}$ and $\sigma_{\max}(W - \frac{1}{n}\mathbf{1}\mathbf{1}^{\text{T}}) < 1$, and $\alpha^k > 0$ is a step size for iteration k. If two agents i and j are neither neighbors nor identical, then $w_{ij} = 0$. This way, the computation of (2.1) involves only local and neighbor information, and hence the iteration is decentralized.

Following our notation, we rewrite (2.1) for all the agents together as

$$\mathbf{x}^{k+1} = W\mathbf{x}^k - \alpha^k \nabla \mathbf{f}(\mathbf{x}^k). \tag{2.2}$$

With a fixed step size $\alpha^k \equiv \alpha$, DGD has *inexact convergence*. For each agent i, $x_{(i)}^{k}$ converges to a point in the $O(\alpha)$-neighborhood of a solution to (1.1), and these points for different agents can be different. On the other hand, properly reducing α^k enables *exact convergence*, namely, that each $x_{(i)}^{k}$ converges to the same exact solution. However, reducing α^k causes slower convergence, both in theory and in practice.

Paper [36] assumes that ∇f_i's are Lipschitz continuous, and studies DGD with a constant $\alpha^k \equiv \alpha$. Before the iterates reach the $O(\alpha)$-neighborhood, the objective value reduces at the rate $O\left(\frac{1}{k}\right)$, and this rate improves to linear if f_i's are also (restricted) strongly convex. In comparison, paper [14] studies DGD with diminishing $\alpha^k = \frac{1}{k^{1/3}}$ and assumes that

∇f_i's are Lipschitz continuous and bounded. The objective convergence rate slows down to $O\left(\frac{1}{k^{2/3}}\right)$. Paper [5] studies DGD with diminishing $\alpha^k = \frac{1}{k^{1/2}}$ and assumes that f_i's are Lipschitz continuous; a slower rate $O\left(\frac{\ln k}{\sqrt{k}}\right)$ is proved. A simple example of decentralized least squares in Section 4.1 gives a rough comparison of these three schemes (and how they compare to the proposed algorithm).

To see the cause of *inexact convergence* with a *fixed step size*, let $\mathbf{x}^\infty$ be the limit of $\mathbf{x}^k$ (assuming the step size is small enough to ensure convergence). Taking the limit over k on both sides of iteration (2.2) gives us

$$\mathbf{x}^\infty = W\mathbf{x}^\infty - \alpha\nabla\mathbf{f}(\mathbf{x}^\infty).$$

When α is fixed and nonzero, assuming the consensus of $\mathbf{x}^\infty$ (namely, it has identical rows $x_{(i)}^\infty$) will mean $\mathbf{x}^\infty = W\mathbf{x}^\infty$, as a result of $W\mathbf{1} = \mathbf{1}$, and thus $\nabla\mathbf{f}(\mathbf{x}^\infty) = \mathbf{0}$, which is equivalent to $\nabla f_i(x_{(i)}^\infty) = 0$, $\forall i$, i.e., the same point $x_{(i)}^\infty$ simultaneously minimizes f_i for all agents i. This is impossible in general and is different from our objective to find a point that minimizes $\sum_{i=1}^n f_i$.

2.2. Development of EXTRA. The next proposition provides simple conditions for the consensus and optimality for problem (1.1).

PROPOSITION 2.1. *Assume* $\mathrm{null}\{I - W\} = \mathrm{span}\{\mathbf{1}\}$. *If*

$$\mathbf{x}^* \triangleq \begin{pmatrix} — & x_{(1)}^{*\mathrm{T}} & — \\ — & x_{(2)}^{*\mathrm{T}} & — \\ & \vdots & \\ — & x_{(n)}^{*\mathrm{T}} & — \end{pmatrix} \tag{2.3}$$

satisfies conditions:

1. $\mathbf{x}^* = W\mathbf{x}^*$ *(consensus),*
2. $\mathbf{1}^{\mathrm{T}}\nabla\mathbf{f}(\mathbf{x}^*) = 0$ *(optimality),*

then $x^* = x_{(i)}^*$, *for any* i, *is a solution to the consensus optimization problem* (1.1).

Proof. Since $\mathrm{null}\{I - W\} = \mathrm{span}\{\mathbf{1}\}$, $\mathbf{x}$ is consensual if and only if condition 1 holds, i.e., $\mathbf{x}^* = W\mathbf{x}^*$. Since $\mathbf{x}^*$ is consensual, we have $\mathbf{1}^{\mathrm{T}}\nabla\mathbf{f}(\mathbf{x}^*) = \sum_{i=1}^n \nabla f_i(x^*)$, so condition 2 means optimality. □

Next, we construct the update formula of EXTRA, following which the iterate sequence will converge to a point satisfying the two conditions in Proposition 2.1.

Consider the DGD update (2.2) written at iterations $k+1$ and k as follows

$$\mathbf{x}^{k+2} = W\mathbf{x}^{k+1} - \alpha\nabla\mathbf{f}(\mathbf{x}^{k+1}), \tag{2.4}$$

$$\mathbf{x}^{k+1} = \tilde{W}\mathbf{x}^{k} - \alpha\nabla\mathbf{f}(\mathbf{x}^{k}), \tag{2.5}$$

where the former uses the mixing matrix W and the latter uses

$$\tilde{W} = \frac{I + W}{2}.$$

The choice of $\tilde{W}$ will be generalized later. The update formula of EXTRA is simply their difference, subtracting (2.5) from (2.4):

$$\mathbf{x}^{k+2} - \mathbf{x}^{k+1} = W\mathbf{x}^{k+1} - \tilde{W}\mathbf{x}^{k} - \alpha\nabla\mathbf{f}(\mathbf{x}^{k+1}) + \alpha\nabla\mathbf{f}(\mathbf{x}^{k}). \tag{2.6}$$

Given $\mathbf{x}^k$ and $\mathbf{x}^{k+1}$, the next iterate $\mathbf{x}^{k+2}$ is generated by (2.6).

Let us assume that $\{\mathbf{x}^k\}$ converges for now and let $\mathbf{x}^* = \lim_{k\to\infty} \mathbf{x}^k$. Let us also assume that $\nabla\mathbf{f}$ is continuous. We first establish condition 1 of Proposition 2.1. Taking $k \to \infty$ in (2.6) gives us

$$\mathbf{x}^* - \mathbf{x}^* = (W - \tilde{W})\mathbf{x}^* - \alpha\nabla\mathbf{f}(\mathbf{x}^*) + \alpha\nabla\mathbf{f}(\mathbf{x}^*), \tag{2.7}$$

from which it follows that

$$W\mathbf{x}^* - \mathbf{x}^* = 2(W - \tilde{W})\mathbf{x}^* = \mathbf{0}. \tag{2.8}$$

Therefore, $\mathbf{x}^*$ is consensual.

Provided that $\mathbf{1}^{\mathrm{T}}(W - \tilde{W}) = 0$, we show that $\mathbf{x}^*$ also satisfies condition 2 of Proposition 2.1. To see this, adding the first update $\mathbf{x}^1 = W\mathbf{x}^0 - \alpha\nabla\mathbf{f}(x^0)$ to the subsequent updates following the formulas of $(\mathbf{x}^2 - \mathbf{x}^1), (\mathbf{x}^3 - \mathbf{x}^2), \ldots, (\mathbf{x}^{k+2} - \mathbf{x}^{k+1})$ given by (2.6) and then applying telescopic cancellation, we obtain

$$\mathbf{x}^{k+2} = \tilde{W}\mathbf{x}^{k+1} - \alpha\nabla\mathbf{f}(\mathbf{x}^{k+1}) + \sum_{t=0}^{k+1}(W - \tilde{W})\mathbf{x}^t, \tag{2.9}$$

or equivalently,

$$\mathbf{x}^{k+2} = W\mathbf{x}^{k+1} - \alpha\nabla\mathbf{f}(\mathbf{x}^{k+1}) + \sum_{t=0}^{k}(W - \tilde{W})\mathbf{x}^t. \tag{2.10}$$

Taking $k \to \infty$, from $\mathbf{x}^* = \lim_{k\to\infty} \mathbf{x}^k$ and $\mathbf{x}^* = \tilde{W}\mathbf{x}^* = W\mathbf{x}^*$, it follows that

$$\alpha\nabla\mathbf{f}(\mathbf{x}^*) = \sum_{t=0}^{\infty}(W - \tilde{W})\mathbf{x}^t. \tag{2.11}$$

Left-multiplying $\mathbf{1}^{\mathrm{T}}$ on both sides of (2.11), in light of $\mathbf{1}^{\mathrm{T}}(W - \tilde{W}) = 0$, we obtain the condition 2 of Proposition 2.1:

$$\mathbf{1}^{\mathrm{T}}\nabla\mathbf{f}(\mathbf{x}^*) = 0. \tag{2.12}$$

To summarize, provided that $\mathrm{null}\{I - W\} = \mathrm{span}\{\mathbf{1}\}$, $\tilde{W} = \frac{I+W}{2}$, $\mathbf{1}^{\mathrm{T}}(W - \tilde{W}) = 0$, and the continuity of $\nabla\mathbf{f}$, if a sequence following EXTRA (2.6) converges to a point $\mathbf{x}^*$, then by Proposition 2.1, $\mathbf{x}^*$ is consensual and any of its identical row vectors solves problem (1.1).

2.3. The Algorithm EXTRA and its Assumptions. We present EXTRA — an exact first-order algorithm for decentralized consensus optimization — in Algorithm 1.

Algorithm 1: EXTRA

```
Choose $\alpha > 0$ and mixing matrices $W \in \mathbb{R}^{n\times n}$ and $\tilde{W} \in \mathbb{R}^{n\times n}$;
Pick any $\mathbf{x}^0 \in \mathbb{R}^{n\times p}$;
1. $\mathbf{x}^1 \leftarrow W\mathbf{x}^0 - \alpha\nabla\mathbf{f}(\mathbf{x}^0)$;
2. for $k = 0, 1, \cdots$ do
     $\mathbf{x}^{k+2} \leftarrow (I+W)\mathbf{x}^{k+1} - \tilde{W}\mathbf{x}^k - \alpha\left[\nabla\mathbf{f}(\mathbf{x}^{k+1}) - \nabla\mathbf{f}(\mathbf{x}^k)\right]$;
   end for
```

on Breaking to the individual agents, Step 1 of EXTRA performs updates

$$x_{(i)}^1 = \sum_{j=1}^{n} w_{ij} x_{(j)}^0 - \alpha\nabla f_i(x_{(i)}^0), \quad i = 1, \ldots, n,$$

and Step 2 at each iteration k performs updates

$$x_{(i)}^{k+2} = x_{(i)}^{k+1} + \sum_{j=1}^{n} w_{ij} x_{(j)}^{k+1} - \sum_{j=1}^{n} \tilde{w}_{ij} x_{(j)}^{k} - \alpha\left[\nabla f_i(x_{(i)}^{k+1}) - \nabla f_i(x_{(i)}^{k})\right], \quad i = 1, \ldots, n.$$

Each agent computes $\nabla f_i(x_{(i)}^k)$ once for each k and uses it twice for $x_{(i)}^{k+1}$ and $x_{(i)}^{k+2}$. For our recommended choice of $\tilde{W} = (W+I)/2$, each agent computes $\sum_{j=1}^{n} w_{ij} x_{(j)}^k$ once as well.

Here we formally give the assumptions on the mixing matrices W and $\tilde{W}$ for EXTRA. All of them will be used in the convergence analysis in the next section.

Assumption 1 (Mixing matrix). *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* $\mathcal{E}$. *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. *(Decentralized property) If* $i \neq j$ *and* $(i,j) \notin \mathcal{E}$, *then* $w_{ij} = \tilde{w}_{ij} = 0$.
2. *(Symmetry)* $W = W^{\mathrm{T}}$, $\tilde{W} = \tilde{W}^{\mathrm{T}}$.
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* $\frac{I+W}{2} \succcurlyeq \tilde{W} \succcurlyeq W$.

We claim that Parts 2–4 of Assumption 1 imply $\mathrm{null}\{I - W\} = \mathrm{span}\{\mathbf{1}\}$ and the eigenvalues of W lie in $(-1, 1]$, which are commonly assumed for DGD. Therefore, the additional assumptions are merely on $\tilde{W}$. In fact, EXTRA can use the same W used in DGD and simply take $\tilde{W} = \frac{I+W}{2}$, which satisfies Part 4. It is also worth noting that the recent work push-DGD [23] relaxes the symmetry condition, yet such relaxation for EXTRA is not trivial and is our future work.

Proposition 2.2. *Parts 2–4 of Assumption 1 imply* $\mathrm{null}\{I - W\} = \mathrm{span}\{\mathbf{1}\}$ *and that the eigenvalues of* W *lie in* $(-1, 1]$.

Proof. From part 4, we have $\frac{I+W}{2} \succcurlyeq \tilde{W} \succ 0$ and thus $W \succ -I$ and $\lambda_{\min}(W) > -1$. Also from part 4, we have $\frac{I+W}{2} \succcurlyeq W$ and thus $I \succeq W$, which means $\lambda_{\max}(W) \leq 1$. Hence, all eigenvalues of W (and those of $\tilde{W}$) lie in $(-1, 1]$.

Now, we show $\text{null}\{I-W\}=\text{span}\{\mathbf{1}\}$. Consider a *nonzero* vector $\mathbf{v}\in\text{null}\{I-W\}$, which satisfies $(I-W)\mathbf{v}=0$ and thus $\mathbf{v}^T(I-W)\mathbf{v}=0$ and $\mathbf{v}^T\mathbf{v}=\mathbf{v}^TW\mathbf{v}$. From $\frac{I+W}{2}\succcurlyeq\tilde{W}$ (part 4), we get $\mathbf{v}^T\mathbf{v}=\mathbf{v}^T(\frac{I+W}{2})\mathbf{v}\geq\mathbf{v}^T\tilde{W}\mathbf{v}$, while from $\tilde{W}\succcurlyeq W$ (part 4) we also get $\mathbf{v}^T\tilde{W}\mathbf{v}\geq\mathbf{v}^TW\mathbf{v}=\mathbf{v}^T\mathbf{v}$. Therefore, we have $\mathbf{v}^T\tilde{W}\mathbf{v}=\mathbf{v}^T\mathbf{v}$ or equivalently $(\tilde{W}-I)\mathbf{v}=0$, adding which to $(I-W)\mathbf{v}=0$ yields $(\tilde{W}-W)\mathbf{v}=0$. In light of $\text{null}\{W-\tilde{W}\}=\text{span}\{\mathbf{1}\}$ (part 3), we must have $\mathbf{v}\in\text{span}\{\mathbf{1}\}$ and thus $\text{null}\{I-W\}=\text{span}\{\mathbf{1}\}$. □

2.4. Mixing Matrices. In EXTRA, the mixing matrices W and $\tilde{W}$ diffuse information throughout the network.

The role of W is the similar as that in DGD [5, 31, 36] and average consensus [33]. It has a few common choices, which can significantly affect performance.

(i) Symmetric doubly stochastic matrix [5, 31, 36]: $W=W^{\mathrm{T}}$, $W\mathbf{1}=\mathbf{1}$, and $w_{ij}\geq 0$. Special cases of such matrices include parts (ii) and (iii) below.

(ii) Laplacian-based constant edge weight matrix [28, 33],

$$W=I-\frac{L}{\tau},$$

where L is the Laplacian matrix of the graph $\mathcal{G}$ and $\tau>\frac{1}{2}\lambda_{\max}(L)$ is a scaling parameter. Denote $\deg(i)$ as the degree of agent i. When $\lambda_{\max}(L)$ is not available, $\tau=\max_{i\in\mathcal{V}}\{\deg(i)\}+\epsilon$ for some small $\epsilon>0$, say $\epsilon=1$, can be used.

(iii) Metropolis constant edge weight matrix [3, 34],

$$w_{ij}=\begin{cases}\frac{1}{\max\{\deg(i),\deg(j)\}+\epsilon}, & \text{if } (i,j)\in\mathcal{E},\\ 0, & \text{if } (i,j)\notin\mathcal{E} \text{ and } i\neq j,\\ 1-\sum\limits_{k\in\mathcal{V}}w_{ik}, & \text{if } i=j,\end{cases}$$

for some small positive $\epsilon>0$.

(iv) Symmetric fastest distributed linear averaging (FDLA) matrix. It is a symmetric W that achieves fastest information diffusion and can be obtained by a semidefinite program [33].

It is worth noting that the optimal choice for average consensus, FDLA, no longer appears optimal in decentralized consensus optimization, which is more general.

When W is chosen following any strategy above, $\tilde{W}=\frac{I+W}{2}$ is found to be very efficient.

2.5. EXTRA as Corrected DGD. We rewrite (2.10) as

$$\underbrace{\mathbf{x}^{k+1}=W\mathbf{x}^k-\alpha\nabla\mathbf{f}(\mathbf{x}^k)}_{\text{DGD}}+\underbrace{\sum_{t=0}^{k-1}(W-\tilde{W})\mathbf{x}^t}_{\text{correction}},\quad k=0,1,\cdots. \tag{2.13}$$

An EXTRA update is, therefore, a DGD update with a cumulative correction term. In subsection 2.1, we have argued that the DGD update cannot reach consensus asymptotically unless α asymptotically vanishes. Since $\alpha\nabla\mathbf{f}(\mathbf{x}^k)$ with a fixed $\alpha>0$ cannot vanish in general, it must be corrected, or otherwise $\mathbf{x}^{k+1}-W\mathbf{x}^k$ does not vanish, preventing $\mathbf{x}^k$ from being asymptotically consensual. Provided that (2.13) converges, the role of the cumulative term $\sum_{t=0}^{k-1}(W-\tilde{W})\mathbf{x}^t$ is to *neutralize* $-\alpha\nabla\mathbf{f}(\mathbf{x}^k)$ in $(\mathrm{span}\{\mathbf{1}\})^{\perp}$, the subspace orthogonal to $\mathbf{1}$. If a vector $\mathbf{v}$ obeys $\mathbf{v}^{\mathrm{T}}(W-\tilde{W})=0$, then the convergence of (2.13) means the vanishing of $\mathbf{v}^{\mathrm{T}}\nabla\mathbf{f}(\mathbf{x}^k)$ in the limit. We need $\mathbf{1}^{\mathrm{T}}\nabla\mathbf{f}(\mathbf{x}^k)=0$ for consensus optimality. The correction term in (2.13) is the simplest that we could find so far. In particular, the summation is necessary since each individual term $(W-\tilde{W})\mathbf{x}^t$ is asymptotically vanishing. The terms must work cumulatively.

3. Convergence Analysis.

To establish convergence of EXTRA, this paper makes two additional but common assumptions as follows. Unless otherwise stated, the results in this section are given under Assumptions 1–3.

Assumption 2. ***(Convex objective with Lipschitz continuous gradient)*** *Objective functions f_i are proper closed convex and Lipschitz differentiable:*

$$\|\nabla f_i(x_a)-\nabla f_i(x_b)\|_2 \le L_{f_i}\|x_a-x_b\|_2, \quad \forall x_a, x_b \in \mathbb{R}^p,$$

where $L_{f_i}\ge 0$ are constant.

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

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

with constant $L_{\mathbf{f}}=\max_i\{L_{f_i}\}$.

Assumption 3. ***(Solution existence)*** *Problem (1.1) has a nonempty set of optimal solutions: $\mathcal{X}^*\neq\emptyset$.*

3.1. Preliminaries.

We first state a lemma that gives the first-order optimality conditions of (1.1).

Lemma 3.1 (First-order optimality conditions). *Given mixing matrices W and $\tilde{W}$, define $U=(\tilde{W}-W)^{1/2}$ by letting $U\triangleq VS^{1/2}V^{\mathrm{T}}\in\mathbb{R}^{n\times n}$ where $VSV^{\mathrm{T}}=\tilde{W}-W$ is the economical-form singular value decomposition. Then, under Assumptions 1–3, $\mathbf{x}^*$ is consensual and $x^*_{(1)}\equiv x^*_{(2)}\equiv\cdots\equiv x^*_{(n)}$ is optimal to problem (1.1) if and only if there exists $\mathbf{q}^*=U\mathbf{p}$ for some $\mathbf{p}\in\mathbb{R}^{n\times p}$ such that*

$$\begin{cases} U\mathbf{q}^*+\alpha\nabla\mathbf{f}(\mathbf{x}^*)=\mathbf{0}, & (3.1)\\ U\mathbf{x}^*=\mathbf{0}. & (3.2)\end{cases}$$

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 from Proposition 2.1, condition 1, $\mathbf{x}^*$ is consensual if and only if (3.2) holds.

Next, following Proposition 2.1, condition 2, $\mathbf{x}$ is optimal if and only if $\mathbf{1}^{\mathrm{T}}\nabla\mathbf{f}(\mathbf{x}^*)=0$. Since U is symmetric and $U^{\mathrm{T}}\mathbf{1}=0$, (3.1) gives $\mathbf{1}^{\mathrm{T}}\nabla\mathbf{f}(\mathbf{x}^*)=0$. Conversely, if $\mathbf{1}^{\mathrm{T}}\nabla\mathbf{f}(\mathbf{x}^*)=0$, then $\nabla \mathbf{f}(\mathbf{x}^*) \in \mathrm{span}\{U\}$ follows from $\mathrm{null}\{U\} = (\mathrm{span}\{\mathbf{1}\})^\perp$ and thus $\alpha\nabla \mathbf{f}(\mathbf{x}^*) = -U\mathbf{q}$ for some $\mathbf{q}$. Let $\mathbf{q}^* = \mathrm{Proj}_U \mathbf{q}$. Then, $U\mathbf{q}^* = U\mathbf{q}$ and (3.1) holds. ∎

Let $\mathbf{x}^*$ and $\mathbf{q}^*$ satisfy the optimality conditions (3.1) and (3.2). Introduce auxiliary sequence

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

and for each k,

$$\mathbf{z}^k = \begin{pmatrix} \mathbf{q}^k \\ \mathbf{x}^k \end{pmatrix}, \quad \mathbf{z}^* = \begin{pmatrix} \mathbf{q}^* \\ \mathbf{x}^* \end{pmatrix}, \quad G = \begin{pmatrix} I & \mathbf{0} \\ \mathbf{0} & \tilde{W} \end{pmatrix}. \tag{3.3}$$

The next lemma establishes the relations among $\mathbf{x}^k$, $\mathbf{q}^k$, $\mathbf{x}^*$, and $\mathbf{q}^*$.

Lemma 3.2. *In EXTRA, the quadruple sequence* $\{\mathbf{x}^k, \mathbf{q}^k, \mathbf{x}^*, \mathbf{q}^*\}$ *obeys*

$$\begin{array}{rl} & (I + W - 2\tilde{W})(\mathbf{x}^{k+1} - \mathbf{x}^*) + \tilde{W}(\mathbf{x}^{k+1} - \mathbf{x}^k) \\ = & -U(\mathbf{q}^{k+1} - \mathbf{q}^*) - \alpha[\nabla \mathbf{f}(\mathbf{x}^k) - \nabla \mathbf{f}(\mathbf{x}^*)], \end{array} \tag{3.4}$$

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

Proof. Similar to how (2.9) is derived, summing EXTRA iterations 1 through $k+1$

$$\begin{gathered} \mathbf{x}^1 = W\mathbf{x}^0 - \alpha\nabla \mathbf{f}(\mathbf{x}^0), \\ \mathbf{x}^2 = (I + W)\mathbf{x}^1 - \tilde{W}\mathbf{x}^0 - \alpha\nabla \mathbf{f}(\mathbf{x}^1) + \alpha\nabla \mathbf{f}(\mathbf{x}^0), \\ \cdots, \\ \mathbf{x}^{k+1} = (I + W)\mathbf{x}^k - \tilde{W}\mathbf{x}^{k-1} - \alpha\nabla \mathbf{f}(\mathbf{x}^k) + \alpha\nabla \mathbf{f}(\mathbf{x}^{k-1}), \end{gathered}$$

we get

$$\mathbf{x}^{k+1} = \tilde{W}\mathbf{x}^k - \sum_{t=0}^{k} (\tilde{W} - W)\mathbf{x}^t - \alpha\nabla \mathbf{f}(\mathbf{x}^k). \tag{3.5}$$

Using $\mathbf{q}^{k+1} = \sum_{t=0}^{k+1} U\mathbf{x}^t$ and the decomposition $\tilde{W} - W = U^2$, it follows from (3.5) that

$$(I + W - 2\tilde{W})\mathbf{x}^{k+1} + \tilde{W}(\mathbf{x}^{k+1} - \mathbf{x}^k) = -U\mathbf{q}^{k+1} - \alpha\nabla \mathbf{f}(\mathbf{x}^k). \tag{3.6}$$

Since $(I + W - 2\tilde{W})\mathbf{1} = 0$, we have

$$(I + W - 2\tilde{W})\mathbf{x}^* = \mathbf{0}. \tag{3.7}$$

Subtracting (3.7) from (3.6) and adding $\mathbf{0} = U\mathbf{q}^* + \alpha\nabla \mathbf{f}(\mathbf{x}^*)$ to (3.6), we obtain (3.4). ∎

The convergence analysis is based on the recursion (3.4). Below we will show that $\mathbf{x}^k$ converges to a solution $\mathbf{x}^* \in \mathcal{X}^*$ and $\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_{\tilde{W}}^2$ converges to 0 at a rate of $O\left(\frac{1}{k}\right)$ in an ergodic sense. Further assuming (restricted) strong convexity, we obtain the Q-linear convergence of $\|\mathbf{z}^k - \mathbf{z}^*\|_G^2$ to 0, which implies the R-linear convergence of $\mathbf{x}^k$ to $\mathbf{x}^*$.

3.2. Convergence and Rate. Let us first interpret the step size condition

$$\alpha < \frac{2\lambda_{\min}(\tilde{W})}{L_{\mathbf{f}}}, \tag{3.8}$$

which is assumed by Theorem 3.3 below. First of all, let W satisfy Assumption 1. It is easy to ensure $\lambda_{\min}(W) \geq 0$ since otherwise, we can replace W by $\frac{I+W}{2}$. In light of part 4 of Assumption 1, if we let $\tilde{W} = \frac{I+W}{2}$, then we have $\lambda_{\min}(\tilde{W}) \geq \frac{1}{2}$, which simplifies the bound (3.8) to

$$\alpha < \frac{1}{L_{\mathbf{f}}},$$

which is independent of any network property (size, diameter, etc.). Furthermore, if L_{f_i} $(i = 1, \ldots, n)$ are in the same order, the bound $\frac{1}{L_{\mathbf{f}}}$ has the same order as the bound $1/(\frac{1}{n}\sum_{i=1}^{n} L_{f_i})$, which is used in the (centralized) gradient descent method. In other words, a fixed and rather large step size is permitted by EXTRA.

THEOREM 3.3. *Under Assumptions 1–3, if α satisfies $0 < \alpha < \frac{2\lambda_{\min}(\tilde{W})}{L_{\mathbf{f}}}$, then*

$$\|\mathbf{z}^k - \mathbf{z}^*\|_G^2 - \|\mathbf{z}^{k+1} - \mathbf{z}^*\|_G^2 \geq \zeta \|\mathbf{z}^k - \mathbf{z}^{k+1}\|_G^2, \quad k = 0, 1, \ldots, \tag{3.9}$$

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

Proof. Following Assumption 2, $\nabla \mathbf{f}$ is Lipschitz continuous and thus we have

$$\begin{aligned} & \tfrac{2\alpha}{L_{\mathbf{f}}} \|\nabla \mathbf{f}(\mathbf{x}^k) - \nabla \mathbf{f}(\mathbf{x}^*)\|_{\mathrm{F}}^2 \\ \leq\ & 2\alpha \langle \mathbf{x}^k - \mathbf{x}^*, \nabla \mathbf{f}(\mathbf{x}^k) - \nabla \mathbf{f}(\mathbf{x}^*) \rangle \\ =\ & 2\langle \mathbf{x}^{k+1} - \mathbf{x}^*, \alpha[\nabla \mathbf{f}(\mathbf{x}^k) - \nabla \mathbf{f}(\mathbf{x}^*)] \rangle + 2\alpha \langle \mathbf{x}^k - \mathbf{x}^{k+1}, \nabla \mathbf{f}(\mathbf{x}^k) - \nabla \mathbf{f}(\mathbf{x}^*) \rangle. \end{aligned} \tag{3.10}$$

Substituting (3.4) from Lemma 3.2 for $\alpha[\nabla \mathbf{f}(\mathbf{x}^k) - \nabla \mathbf{f}(\mathbf{x}^*)]$, it follows from (3.10) that

$$\begin{aligned} & \tfrac{2\alpha}{L_{\mathbf{f}}} \|\nabla \mathbf{f}(\mathbf{x}^k) - \nabla \mathbf{f}(\mathbf{x}^*)\|_{\mathrm{F}}^2 \\ \leq\ & 2\langle \mathbf{x}^{k+1} - \mathbf{x}^*, U(\mathbf{q}^* - \mathbf{q}^{k+1}) \rangle + 2\langle \mathbf{x}^{k+1} - \mathbf{x}^*, \tilde{W}(\mathbf{x}^k - \mathbf{x}^{k+1}) \rangle \\ & -2\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{I+W-2\tilde{W}}^2 + 2\alpha \langle \mathbf{x}^k - \mathbf{x}^{k+1}, \nabla \mathbf{f}(\mathbf{x}^k) - \nabla \mathbf{f}(\mathbf{x}^*) \rangle. \end{aligned} \tag{3.11}$$

For the terms on the right-hand-side of (3.11), we have

$$\begin{aligned} 2\langle \mathbf{x}^{k+1} - x^*, U(\mathbf{q}^* - \mathbf{q}^{k+1}) \rangle &= 2\langle U(\mathbf{x}^{k+1} - \mathbf{x}^*), \mathbf{q}^* - \mathbf{q}^{k+1} \rangle \\ (\because U\mathbf{x}^* = \mathbf{0}) &= 2\langle U\mathbf{x}^{k+1}, \mathbf{q}^* - \mathbf{q}^{k+1} \rangle \\ &= 2\langle \mathbf{q}^{k+1} - \mathbf{q}^k, \mathbf{q}^* - \mathbf{q}^{k+1} \rangle, \end{aligned} \tag{3.12}$$

$$2\langle \mathbf{x}^{k+1} - \mathbf{x}^*, \tilde{W}(\mathbf{x}^k - \mathbf{x}^{k+1}) \rangle = 2\langle \mathbf{x}^{k+1} - \mathbf{x}^k, \tilde{W}(\mathbf{x}^* - \mathbf{x}^{k+1}) \rangle, \tag{3.13}$$

and

$$\begin{aligned} & 2\alpha \langle \mathbf{x}^k - \mathbf{x}^{k+1}, \nabla \mathbf{f}(\mathbf{x}^k) - \nabla \mathbf{f}(\mathbf{x}^*) \rangle \\ \leq\ & \tfrac{\alpha L_{\mathbf{f}}}{2} \|\mathbf{x}^k - \mathbf{x}^{k+1}\|_{\mathrm{F}}^2 + \tfrac{2\alpha}{L_{\mathbf{f}}} \|\nabla \mathbf{f}(\mathbf{x}^k) - \nabla \mathbf{f}(\mathbf{x}^*)\|_{\mathrm{F}}^2. \end{aligned} \tag{3.14}$$

Plugging (3.12)–(3.14) into (3.11) and recalling the definitions of $\mathbf{z}^k$, $\mathbf{z}^*$, and G, we have

$$\begin{aligned} & \tfrac{2\alpha}{L_{\mathbf{f}}}\|\nabla\mathbf{f}(\mathbf{x}^k)-\nabla\mathbf{f}(\mathbf{x}^*)\|_{\mathrm{F}}^2 \\ \leq\ & 2\langle\mathbf{q}^{k+1}-\mathbf{q}^k,\mathbf{q}^*-\mathbf{q}^{k+1}\rangle+2\langle\mathbf{x}^{k+1}-\mathbf{x}^k,\tilde{W}(\mathbf{x}^*-\mathbf{x}^{k+1})\rangle \\ & -2\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_{I+W-2\tilde{W}}^2+\tfrac{\alpha L_{\mathbf{f}}}{2}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathrm{F}}^2+\tfrac{2\alpha}{L_{\mathbf{f}}}\|\nabla\mathbf{f}(\mathbf{x}^k)-\nabla\mathbf{f}(\mathbf{x}^*)\|_{\mathrm{F}}^2, \end{aligned} \tag{3.15}$$

that is

$$0\leq 2\langle\mathbf{z}^{k+1}-\mathbf{z}^k,G(\mathbf{z}^*-\mathbf{z}^{k+1})\rangle-2\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_{I+W-2\tilde{W}}^2+\tfrac{\alpha L_{\mathbf{f}}}{2}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathrm{F}}^2. \tag{3.16}$$

Apply the basic equality $2\langle\mathbf{z}^{k+1}-\mathbf{z}^k,G(\mathbf{z}^*-\mathbf{z}^{k+1})\rangle=\|\mathbf{z}^k-\mathbf{z}^*\|_G^2-\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_G^2-\|\mathbf{z}^k-\mathbf{z}^{k+1}\|_G^2$ to (3.16), we have

$$\begin{aligned} 0\ \leq\ & \|\mathbf{z}^k-\mathbf{z}^*\|_G^2-\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_G^2-\|\mathbf{z}^k-\mathbf{z}^{k+1}\|_G^2 \\ & -2\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_{I+W-2\tilde{W}}^2+\tfrac{\alpha L_{\mathbf{f}}}{2}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathrm{F}}^2. \end{aligned} \tag{3.17}$$

Define

$$G'=\begin{pmatrix} I & \mathbf{0} \\ \mathbf{0} & \tilde{W}-\frac{\alpha L_{\mathbf{f}}}{2}I \end{pmatrix}.$$

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

$$\|\mathbf{z}^k-\mathbf{z}^*\|_G^2-\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_G^2\geq\|\mathbf{z}^k-\mathbf{z}^{k+1}\|_G-\frac{\alpha L_{\mathbf{f}}}{2}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathrm{F}}^2=\|\mathbf{z}^k-\mathbf{z}^{k+1}\|_{G'}.$$

Since $\alpha<\frac{2\lambda_{\min}(\tilde{W})}{L_{\mathbf{f}}}$, we have $G'\succ 0$ and

$$\|\mathbf{z}^k-\mathbf{z}^{k+1}\|_{G'}^2\geq\zeta\|\mathbf{z}^k-\mathbf{z}^{k+1}\|_G^2, \tag{3.18}$$

which gives (3.9).

It shows from (3.9) that for any optimal $\mathbf{z}^*$, $\|\mathbf{z}^k-\mathbf{z}^*\|_G^2$ is bounded and contractive, so $\|\mathbf{z}^k-\mathbf{z}^*\|_G^2$ is converging as $\|\mathbf{z}^k-\mathbf{z}^{k+1}\|_G^2\to 0$. The convergence of $\mathbf{z}^k$ to a solution $\mathbf{z}^*$ follows from the standard analysis for contraction methods; see, for example, Theorem 3 in [11]. □

To estimate the rate of convergence, we need the following result.

Proposition 3.4. *If a sequence $\{a_k\}\subset\mathbb{R}$ obeys: $a_k\geq 0$ and $\sum_{t=1}^{\infty}a_t<\infty$, then we have*[1]*:* (i) $\lim_{k\to\infty}a_k=0$*;* (ii) $\frac{1}{k}\sum_{t=1}^{k}a_t=O\left(\frac{1}{k}\right)$*;* (iii) $\min_{t\leq k}\{a_t\}=o\left(\frac{1}{k}\right)$*.*

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

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

[1] Part (iii) is due to [6].

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

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

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). □

THEOREM 3.5. *In the same setting of Theorem 3.3, the following rates hold:*

(1) *Running-average progress:*

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

(2) *Running-best progress:*

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

(3) *Running-average optimality residuals:*

$$\frac{1}{k}\sum_{t=1}^{k}\|U\mathbf{q}^t+\alpha\nabla\mathbf{f}(\mathbf{x}^t)\|_{\tilde{W}}^2 = O\left(\frac{1}{k}\right) \;\; \textit{and} \;\; \frac{1}{k}\sum_{t=1}^{k}\|U\mathbf{x}^t\|_{\mathrm{F}}^2 = O\left(\frac{1}{k}\right);$$

(4) *Running-best optimality residuals:*

$$\min_{t\le k}\left\{\|U\mathbf{q}^t+\alpha\nabla\mathbf{f}(\mathbf{x}^t)\|_{\tilde{W}}^2\right\} = o\left(\frac{1}{k}\right) \;\; \textit{and} \;\; \min_{t\le k}\left\{\|U\mathbf{x}^t\|_{\mathrm{F}}^2\right\} = o\left(\frac{1}{k}\right);$$

Proof. Parts (1) and (2): Since the individual terms $\|\mathbf{z}^k-\mathbf{z}^*\|_G^2$ converge to 0, we are able to sum (3.9) in Theorem 3.3 over $k=0$ through ∞ and apply the telescopic cancellation, i.e.,

$$\sum_{t=0}^{\infty}\|\mathbf{z}^t-\mathbf{z}^{t+1}\|_G^2 = \tfrac{1}{\delta}\sum_{t=0}^{\infty}\left(\|\mathbf{z}^t-\mathbf{z}^*\|_G^2-\|\mathbf{z}^{t+1}-\mathbf{z}^*\|_G^2\right) = \tfrac{\|\mathbf{z}_0-\mathbf{z}^*\|_G^2}{\delta} < \infty. \tag{3.19}$$

Then, the results follow from Proposition 3.4 immediately.

Parts (3) and (4): The progress $\|\mathbf{z}^k-\mathbf{z}^{k+1}\|_G^2$ can be interpreted as the residual to the first-order optimality condition. In light of the first-order optimality conditions (3.1) and (3.2) in Lemma 3.1, the *optimality residuals* are defined as $\|U\mathbf{q}^k+\alpha\nabla\mathbf{f}(\mathbf{x}^k)\|_{\tilde{W}}^2$ and $\|U\mathbf{x}^k\|_{\mathrm{F}}^2$. Furthermore, $\|\frac{1}{\alpha}\mathbf{1}^{\mathrm{T}}(U\mathbf{q}^k+\alpha\nabla\mathbf{f}(\mathbf{x}^k))\|_2^2 = \|\nabla f_1(x_{(1)}^k)+\ldots+\nabla f_n(x_{(n)}^k)\|_2^2$ is the violation to the first-order optimality of (1.1), while $\|U\mathbf{x}^k\|_{\mathrm{F}}^2$ is the violation of consensus. Below we obtain the convergence rates of the optimality residuals.

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

$$\begin{aligned}
\|\mathbf{z}^k-\mathbf{z}^{k+1}\|_G^2 &= \|\mathbf{q}^k-\mathbf{q}^{k+1}\|_{\mathrm{F}}^2+\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\tilde{W}}^2 \\
&= \|\mathbf{x}^{k+1}\|_{\tilde{W}-W}^2+\|(I-\tilde{W})\mathbf{x}^k+U\mathbf{q}^k+\alpha\nabla\mathbf{f}(\mathbf{x}^k)\|_{\tilde{W}}^2 \\
&\ge \|\mathbf{x}^{k+1}\|_{\tilde{W}-W}^2+\tfrac{1}{\rho}\|U\mathbf{q}^k+\alpha\nabla\mathbf{f}(\mathbf{x}^k)\|_{\tilde{W}}^2-\tfrac{1}{\rho-1}\|(I-\tilde{W})\mathbf{x}^k\|_{\tilde{W}}^2.
\end{aligned} \tag{3.20}$$

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 (3.20) that

$$\begin{aligned}
&\tfrac{1}{k}\sum_{t=1}^{k}\|\mathbf{z}^t-\mathbf{z}^{t+1}\|_G^2+\tfrac{1}{k}\|\mathbf{x}^1\|_{\tilde{W}-W}^2\\
\geq\ &\tfrac{1}{k}\sum_{t=1}^{k}\left(\|\mathbf{x}^{t+1}\|_{\tilde{W}-W}^2-\tfrac{\upsilon}{\rho-1}\|\mathbf{x}^t\|_{\tilde{W}-W}^2\right)+\tfrac{1}{k}\|\mathbf{x}^1\|_{\tilde{W}-W}^2\\
&+\tfrac{1}{k}\sum_{t=1}^{k}\tfrac{1}{\rho}\|U\mathbf{q}^t+\alpha\nabla\mathbf{f}(\mathbf{x}^t)\|_{\tilde{W}}^2\ (\text{Set }\rho>\upsilon+1)\\
=\ &\tfrac{1}{k}\sum_{t=1}^{k}(1-\tfrac{\upsilon}{\rho-1})\|U\mathbf{x}^t\|_{\mathrm{F}}^2+\tfrac{1}{k}\|\mathbf{x}^{k+1}\|_{\tilde{W}-W}^2+\tfrac{1}{k}\sum_{t=1}^{k}\tfrac{1}{\rho}\|U\mathbf{q}^t+\alpha\nabla\mathbf{f}(\mathbf{x}^t)\|_{\tilde{W}}^2.
\end{aligned}\tag{3.21}$$

As part (1) 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 $\frac{1}{k}\sum_{t=1}^{k}\|U\mathbf{q}^t+\alpha\nabla\mathbf{f}(\mathbf{x}^t)\|_{\tilde{W}}^2=O\left(\frac{1}{k}\right)$ and $\frac{1}{k}\sum_{t=1}^{k}\|U\mathbf{x}^t\|_{\mathrm{F}}^2=O\left(\frac{1}{k}\right)$.

From (3.21) and (3.19), we see that both $\|U\mathbf{q}^t+\alpha\nabla\mathbf{f}(\mathbf{x}^t)\|_{\tilde{W}}^2$ and $\|U\mathbf{x}^t\|_{\mathrm{F}}^2$ are summable. Again, by Proposition 3.4, we have part (4), the $o\left(\frac{1}{k}\right)$ rate of running best first-order optimality residuals. □

It is open whether $\|\mathbf{z}^k-\mathbf{z}^{k+1}\|_G^2$ is monotonic or not. If one can show its monotonicity, then the convergence rates will hold for the last point in the running sequence.

3.3. Linear Convergence under Restricted Strong Convexity.

In this subsection we prove that EXTRA with a proper step size reaches linear convergence if the original objective $\bar{f}$ is restricted strongly convex.

A convex function $h:\mathbb{R}^p\to\mathbb{R}$ is *strongly convex* if there exists $\mu>0$ such that

$$\langle\nabla h(x_a)-\nabla h(x_b),x_a-x_b\rangle\geq\mu\|x_a-x_b\|^2,\quad\forall x_a,x_b\in\mathbb{R}^p.$$

h is *restricted strongly convex*[2] with respect to point $\tilde{x}$ if there exists $\mu>0$ such that

$$\langle\nabla h(x)-\nabla h(\tilde{x}),x-\tilde{x}\rangle\geq\mu\|x-\tilde{x}\|_2^2,\quad\forall x\in\mathbb{R}^p.$$

For proof convenience, we introduce function

$$\mathbf{g}(\mathbf{x})\triangleq\mathbf{f}(\mathbf{x})+\frac{1}{4\alpha}\|\mathbf{x}\|_{\tilde{W}-W}^2$$

and claim that $\bar{f}$ is restricted strongly convex with respect to its solution x^* if, and only if, $\mathbf{g}$ is so with respect to $\mathbf{x}^*=\mathbf{1}(x^*)^{\mathrm{T}}$.

PROPOSITION 3.6. *Under Assumptions 1 and 2, the following two statements are equivalent:*

(i) The original objective $\bar{f}(x)=\frac{1}{n}\sum_{i=1}^{n}f_i(x)$ is restricted strongly convex with respect to x^*;

[2] There are different definitions of restricted strong convexity. Ours is derived from [17].

(ii) The penalized function $\mathbf{g}(\mathbf{x}) = \mathbf{f}(\mathbf{x}) + \frac{1}{4\alpha}\|\mathbf{x}\|^2_{\tilde{W}-W}$ *is restricted strongly convex with respect to* $\mathbf{x}^*$*.*

In addition, the strong convexity constant of $\mathbf{g}$ *is no less than that of* $\bar{f}$*.*

See Appendix A for its proof.

THEOREM 3.7. *If* $\mathbf{g}(\mathbf{x}) \triangleq \mathbf{f}(\mathbf{x}) + \frac{1}{4\alpha}\|\mathbf{x}\|^2_{\tilde{W}-W}$ *is restricted strongly convex with respect to* $\mathbf{x}^*$ *with constant* $\mu_{\mathbf{g}} > 0$*, then with proper step size* $\alpha < \frac{2\mu_{\mathbf{g}}\lambda_{\min}(\tilde{W})}{L_{\mathbf{f}}{}^2}$*, there exists* $\delta > 0$ *such that the sequence* $\{\mathbf{z}^k\}$ *generated by EXTRA satisfies*

$$\|\mathbf{z}^k - \mathbf{z}^*\|^2_G \geq (1+\delta)\|\mathbf{z}^{k+1} - \mathbf{z}^*\|^2_G. \tag{3.22}$$

That is, $\|\mathbf{z}^k - \mathbf{z}^*\|^2_G$ *converges to* 0 *at the Q-linear rate* $O\big((1+\delta)^{-k}\big)$*. Consequently,* $\|\mathbf{x}^k - \mathbf{x}^*\|^2_{\tilde{W}}$ *converges to* 0 *at the R-linear rate* $O\big((1+\delta)^{-k}\big)$*.*

Proof. **Toward a lower bound of** $\|\mathbf{z}^k - \mathbf{z}^*\|^2_G - \|\mathbf{z}^{k+1} - \mathbf{z}^*\|^2_G$**:** From the definition of $\mathbf{g}$ and its restricted strong convexity, we have

$$\begin{array}{rcl} 2\alpha\mu_{\mathbf{g}}\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2_{\mathrm{F}} & \leq & 2\alpha\langle \mathbf{x}^{k+1} - \mathbf{x}^*, \nabla\mathbf{g}(\mathbf{x}^{k+1}) - \nabla\mathbf{g}(\mathbf{x}^*)\rangle \\ & = & \|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2_{\tilde{W}-W} + 2\alpha\langle \mathbf{x}^{k+1} - \mathbf{x}^*, \nabla\mathbf{f}(\mathbf{x}^{k+1}) - \nabla\mathbf{f}(\mathbf{x}^k)\rangle \\ & & +2\langle \mathbf{x}^{k+1} - \mathbf{x}^*, \alpha[\nabla\mathbf{f}(\mathbf{x}^k) - \nabla\mathbf{f}(\mathbf{x}^*)]\rangle. \end{array} \tag{3.23}$$

Using Lemma 3.2 for $\alpha[\nabla\mathbf{f}(\mathbf{x}^k) - \nabla\mathbf{f}(\mathbf{x}^*)]$ in (3.23), we get

$$\begin{array}{rl} & 2\alpha\mu_{\mathbf{g}}\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2_{\mathrm{F}} \\ \leq & \|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2_{\tilde{W}-W} + 2\alpha\langle \mathbf{x}^{k+1} - \mathbf{x}^*, \nabla\mathbf{f}(\mathbf{x}^{k+1}) - \nabla\mathbf{f}(\mathbf{x}^k)\rangle - 2\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2_{I+W-2\tilde{W}} \\ & +2\langle \mathbf{x}^{k+1} - \mathbf{x}^*, U(\mathbf{q}^* - \mathbf{q}^{k+1})\rangle + 2\langle \mathbf{x}^{k+1} - \mathbf{x}^*, \tilde{W}(\mathbf{x}^k - \mathbf{x}^{k+1})\rangle \\ = & \|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2_{(\tilde{W}-W)-2(I+W-2\tilde{W})} + 2\alpha\langle \mathbf{x}^{k+1} - \mathbf{x}^*, \nabla\mathbf{f}(\mathbf{x}^{k+1}) - \nabla\mathbf{f}(\mathbf{x}^k)\rangle \\ & +2\langle \mathbf{x}^{k+1} - \mathbf{x}^*, U(\mathbf{q}^* - \mathbf{q}^{k+1})\rangle + 2\langle \mathbf{x}^{k+1} - \mathbf{x}^*, \tilde{W}(\mathbf{x}^k - \mathbf{x}^{k+1})\rangle. \end{array} \tag{3.24}$$

For the last three terms on the right-hand side of (3.24), we have from Young's inequality

$$\begin{array}{rl} & 2\alpha\langle \mathbf{x}^{k+1} - \mathbf{x}^*, \nabla\mathbf{f}(\mathbf{x}^{k+1}) - \nabla\mathbf{f}(\mathbf{x}^k)\rangle \\ \leq & \alpha\eta\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2_{\mathrm{F}} + \frac{\alpha}{\eta}\|\nabla\mathbf{f}(\mathbf{x}^{k+1}) - \nabla\mathbf{f}(\mathbf{x}^k)\|^2_{\mathrm{F}} \\ \leq & \alpha\eta\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2_{\mathrm{F}} + \frac{\alpha L_{\mathbf{f}}{}^2}{\eta}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2_{\mathrm{F}}, \end{array} \tag{3.25}$$

where $\eta > 0$ is a tunable parameter and

$$2\langle \mathbf{x}^{k+1} - \mathbf{x}^*, U(\mathbf{q}^* - \mathbf{q}^{k+1})\rangle = 2\langle \mathbf{q}^{k+1} - \mathbf{q}^k, \mathbf{q}^* - \mathbf{q}^{k+1}\rangle, \tag{3.26}$$

and

$$2\langle \mathbf{x}^{k+1} - \mathbf{x}^*, \tilde{W}(\mathbf{x}^k - \mathbf{x}^{k+1})\rangle = 2\langle \mathbf{x}^{k+1} - \mathbf{x}^k, \tilde{W}(\mathbf{x}^* - \mathbf{x}^{k+1})\rangle. \tag{3.27}$$

Plugging (3.25)–(3.27) into (3.24) and recalling the definition of $\mathbf{z}^k$, $\mathbf{z}^*$, and G, we obtain

$$\begin{array}{rl} & 2\alpha\mu_{\mathbf{g}}\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2_{\mathrm{F}} \\ \leq & \|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2_{(\tilde{W}-W)-2(I+W-2\tilde{W})} + \alpha\eta\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2_{\mathrm{F}} \\ & +\frac{\alpha L_{\mathbf{f}}{}^2}{\eta}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2_{\mathrm{F}} + 2\langle \mathbf{z}^{k+1} - \mathbf{z}^k, G(\mathbf{z}^* - \mathbf{z}^{k+1})\rangle. \end{array} \tag{3.28}$$

By $2\langle \mathbf{z}^{k+1}-\mathbf{z}^k, G(\mathbf{z}^*-\mathbf{z}^{k+1})\rangle = \|\mathbf{z}^k-\mathbf{z}^*\|_G^2 - \|\mathbf{z}^{k+1}-\mathbf{z}^*\|_G^2 - \|\mathbf{z}^k-\mathbf{z}^{k+1}\|_G^2$, (3.28) turns into

$$\begin{aligned}\|\mathbf{z}^k-\mathbf{z}^*\|_G^2 - \|\mathbf{z}^{k+1}-\mathbf{z}^*\|_G^2 \quad \geq \quad & \|\mathbf{x}^{k+1}-\mathbf{x}^*\|^2_{\alpha(2\mu_{\mathbf{g}}-\eta)I-(\tilde{W}-W)+2(I+W-2\tilde{W})} \\ & +\|\mathbf{z}^k-\mathbf{z}^{k+1}\|_G^2 - \tfrac{\alpha L_{\mathbf{f}}{}^2}{\eta}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathrm{F}}^2.\end{aligned} \tag{3.29}$$

A critical inequality: In order to establish (3.22), in light of (3.29), it remains to show

$$\begin{aligned} & \|\mathbf{x}^{k+1}-\mathbf{x}^*\|^2_{\alpha(2\mu_{\mathbf{g}}-\eta)I-(\tilde{W}-W)+2(I+W-2\tilde{W})} + \|\mathbf{z}^k-\mathbf{z}^{k+1}\|_G^2 - \tfrac{\alpha L_{\mathbf{f}}{}^2}{\eta}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathrm{F}}^2 \\ \geq \quad & \delta\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_G^2.\end{aligned} \tag{3.30}$$

With the terms of $\mathbf{z}^k$, $\mathbf{z}^*$ in (3.30) expanded and from $\|\mathbf{x}^{k+1}-\mathbf{x}^*\|^2_{\tilde{W}-W} = \|U(\mathbf{x}^{k+1}-\mathbf{x}^*)\|_{\mathrm{F}}^2 = \|U\mathbf{x}^{k+1}\|_{\mathrm{F}}^2 = \|\mathbf{q}^{k+1}-\mathbf{q}^k\|_{\mathrm{F}}^2$, (3.30) is equivalent to

$$\|\mathbf{x}^{k+1}-\mathbf{x}^*\|^2_{\alpha(2\mu_{\mathbf{g}}-\eta)I+2(I+W-2\tilde{W})-\delta\tilde{W}} + \|\mathbf{x}^k-\mathbf{x}^{k+1}\|^2_{\tilde{W}-\frac{\alpha L_{\mathbf{f}}{}^2}{\eta}I} \geq \delta\|\mathbf{q}^{k+1}-\mathbf{q}^*\|_{\mathrm{F}}^2, \tag{3.31}$$

which is *what remains to be shown below.* That is, we must find a upper bound for $\|\mathbf{q}^{k+1}-\mathbf{q}^*\|_{\mathrm{F}}^2$ in terms of $\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_{\mathrm{F}}^2$ and $\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathrm{F}}^2$.

Establishing (3.31), Step 1: From Lemma 3.2 we have

$$\begin{aligned} & \|U(\mathbf{q}^{k+1}-\mathbf{q}^*)\|_{\mathrm{F}}^2 \\ = \quad & \|(I+W-2\tilde{W})(\mathbf{x}^{k+1}-\mathbf{x}^*)+\tilde{W}(\mathbf{x}^{k+1}-\mathbf{x}^k)+\alpha[\nabla\mathbf{f}(\mathbf{x}^k)-\nabla\mathbf{f}(\mathbf{x}^*)]\|_{\mathrm{F}}^2 \\ = \quad & \|(I+W-2\tilde{W})(\mathbf{x}^{k+1}-\mathbf{x}^*)+\alpha[\nabla\mathbf{f}(\mathbf{x}^{k+1})-\nabla\mathbf{f}(\mathbf{x}^*)] \\ & +\tilde{W}(\mathbf{x}^{k+1}-\mathbf{x}^k)+\alpha[\nabla\mathbf{f}(\mathbf{x}^k)-\nabla\mathbf{f}(\mathbf{x}^{k+1})]\|_{\mathrm{F}}^2.\end{aligned} \tag{3.32}$$

From the inequality $\|\mathbf{a}+\mathbf{b}+\mathbf{c}+\mathbf{d}\|_{\mathrm{F}}^2 \leq \theta\left(\frac{\beta}{\beta-1}\|\mathbf{a}\|_{\mathrm{F}}^2+\beta\|\mathbf{b}\|_{\mathrm{F}}^2\right)+\frac{\theta}{\theta-1}\left(\frac{\gamma}{\gamma-1}\|\mathbf{c}\|_{\mathrm{F}}^2+\gamma\|\mathbf{d}\|_{\mathrm{F}}^2\right)$, which holds for any $\theta>1$, $\beta>1$, $\gamma>1$ and any matrices $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{d}$ of the same dimensions, it follows that

$$\begin{aligned} & \|\mathbf{q}^{k+1}-\mathbf{q}^*\|^2_{\tilde{W}-W} \\ \leq \quad & \theta\left(\tfrac{\beta}{\beta-1}\|\mathbf{x}^{k+1}-\mathbf{x}^*\|^2_{(I+W-2\tilde{W})^2}+\beta\alpha^2\|\nabla\mathbf{f}(\mathbf{x}^{k+1})-\nabla\mathbf{f}(\mathbf{x}^*)\|_{\mathrm{F}}^2\right) \\ & +\tfrac{\theta}{\theta-1}\left(\tfrac{\gamma}{\gamma-1}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|^2_{\tilde{W}^2}+\gamma\alpha^2\|\nabla\mathbf{f}(\mathbf{x}^k)-\nabla\mathbf{f}(\mathbf{x}^{k+1})\|_{\mathrm{F}}^2\right).\end{aligned} \tag{3.33}$$

By Lemma 3.1 and the definition of $\mathbf{q}^k$, all the columns of $\mathbf{q}^*$ and $\mathbf{q}^{k+1}$ lie in the column space of $\tilde{W}-W$. This together with the Lipschitz continuity of $\nabla\mathbf{f}(\mathbf{x})$ turns (3.33) into

$$\begin{aligned} & \|\mathbf{q}^{k+1}-\mathbf{q}^*\|_{\mathrm{F}}^2 \\ \leq \quad & \tfrac{\theta}{\tilde{\lambda}_{\min}(\tilde{W}-W)}\left(\tfrac{\beta\lambda_{\max}((I+W-2\tilde{W})^2)}{\beta-1}+\beta\alpha^2L_{\mathbf{f}}{}^2\right)\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_{\mathrm{F}}^2 \\ & +\tfrac{\theta}{(\theta-1)\tilde{\lambda}_{\min}(\tilde{W}-W)}\left(\tfrac{\gamma\lambda_{\max}(\tilde{W}^2)}{\gamma-1}+\gamma\alpha^2L_{\mathbf{f}}{}^2\right)\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathrm{F}}^2,\end{aligned} \tag{3.34}$$

where $\tilde{\lambda}_{\min}(\cdot)$ gives the smallest *nonzero* eigenvalue. To make a rather tight bound, we choose $\gamma = 1+\frac{\sigma_{\max}(\tilde{W})}{\alpha L_{\mathbf{f}}}$ and $\beta = 1+\frac{\sigma_{\max}(I+W-2\tilde{W})}{\alpha L_{\mathbf{f}}}$ in (3.34) and obtain

$$\begin{aligned} & \|\mathbf{q}^{k+1}-\mathbf{q}^*\|_{\mathrm{F}}^2 \\ \leq \quad & \tfrac{\theta(\sigma_{\max}(I+W-2\tilde{W})+\alpha L_{\mathbf{f}})^2}{\tilde{\lambda}_{\min}(\tilde{W}-W)}\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_{\mathrm{F}}^2+\tfrac{\theta(\sigma_{\max}(\tilde{W})+\alpha L_{\mathbf{f}})^2}{(\theta-1)\tilde{\lambda}_{\min}(\tilde{W}-W)}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathrm{F}}^2.\end{aligned} \tag{3.35}$$

Establishing (3.31), Step 2: In order to establish (3.31), with (3.35), it only remains to show

$$\begin{aligned} &\|\mathbf{x}^{k+1}-\mathbf{x}^*\|^2_{\alpha(2\mu_{\mathbf{g}}-\eta)I+2(I+W-2\tilde{W})-\delta\tilde{W}} + \|\mathbf{x}^k-\mathbf{x}^{k+1}\|^2_{\tilde{W}-\frac{\alpha L_{\mathbf{f}}{}^2}{\eta}I} \\ \geq\quad & \delta\left(\tfrac{\theta(\sigma_{\max}(I+W-2\tilde{W})+\alpha L_{\mathbf{f}})^2}{\tilde{\lambda}_{\min}(\tilde{W}-W)}\|\mathbf{x}^{k+1}-\mathbf{x}^*\|^2_{\mathrm{F}} + \tfrac{\theta(\sigma_{\max}(\tilde{W})+\alpha L_{\mathbf{f}})^2}{(\theta-1)\tilde{\lambda}_{\min}(\tilde{W}-W)}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|^2_{\mathrm{F}}\right). \end{aligned} \tag{3.36}$$

To validate (3.36), we need

$$\left\{\begin{array}{c} \alpha(2\mu_{\mathbf{g}}-\eta)I+2(I+W-2\tilde{W})-\delta\tilde{W} \succcurlyeq \frac{\delta\theta(\sigma_{\max}(I+W-2\tilde{W})+\alpha L_{\mathbf{f}})^2}{\tilde{\lambda}_{\min}(\tilde{W}-W)}I, \\ \tilde{W}-\frac{\alpha L_{\mathbf{f}}{}^2}{\eta}I \succcurlyeq \frac{\delta\theta(\sigma_{\max}(\tilde{W})+\alpha L_{\mathbf{f}})^2}{(\theta-1)\tilde{\lambda}_{\min}(\tilde{W}-W)}I, \end{array}\right. \tag{3.37}$$

which holds as long as

$$\delta \leq \min\left\{\frac{\alpha(2\mu_{\mathbf{g}}-\eta)\tilde{\lambda}_{\min}(\tilde{W}-W)}{\theta(\sigma_{\max}(I+W-2\tilde{W})+\alpha L_{\mathbf{f}})^2+\lambda_{\max}(\tilde{W})\tilde{\lambda}_{\min}(\tilde{W}-W)}, \frac{(\theta-1)(\eta\lambda_{\min}(\tilde{W})-\alpha L_{\mathbf{f}}{}^2)\tilde{\lambda}_{\min}(\tilde{W}-W)}{\theta\eta(\sigma_{\max}(\tilde{W})+\alpha L_{\mathbf{f}})^2}\right\}. \tag{3.38}$$

To ensure $\delta > 0$, the following conditions are what we finally need:

$$\eta \in (0, 2\mu_{\mathbf{g}}) \quad \text{and} \quad \alpha \in \left(0, \frac{\eta\lambda_{\min}(\tilde{W})}{L_{\mathbf{f}}{}^2}\right) \triangleq \mathcal{S}. \tag{3.39}$$

Obviously set $\mathcal{S}$ is nonempty. Therefore, with a proper step size $\alpha \in \mathcal{S}$, the sequences $\|\mathbf{z}^k-\mathbf{z}^*\|^2_G$ is Q-linearly convergent to 0 at the rate $O\big((1+\delta)^{-k}\big)$. Since the definition of G-norm implies $\|\mathbf{x}^k-\mathbf{x}^*\|^2_{\tilde{W}} \leq \|\mathbf{z}^k-\mathbf{z}^*\|^2_G$, $\|\mathbf{x}^k-\mathbf{x}^*\|^2_{\tilde{W}}$ is R-linearly convergent to 0 at the same rate. □

REMARK 1 (Strong convexity condition for linear convergence). *The restricted strong convexity assumption in Theorem 3.7 is imposed on* $\mathbf{g}(\mathbf{x}) = \mathbf{f}(\mathbf{x}) + \frac{1}{4\alpha}\|\mathbf{x}\|^2_{\tilde{W}-W}$*, not on* $\mathbf{f}(\mathbf{x})$*. In other words, the linear convergence of EXTRA does not require all* f_i *to be individually (restricted) strongly convex.*

REMARK 2 (Acceleration by overshooting $\tilde{W}$). *For conciseness, we used Assumption 1 for both Theorems 3.5 and 3.7. In fact, for Theorem 3.7, the condition* $\frac{I+W}{2} \succcurlyeq \tilde{W}$ *in part 4 of Assumption 1 can be relaxed, thanks to* $\mu_{\mathbf{g}}$ *in (3.37). Certain* $\tilde{W} \succcurlyeq \frac{I+W}{2}$*, such as* $\tilde{W} = \frac{1.5I+W}{2.5}$*, can still give linear convergence. In fact, we observed that such an "overshot" choice of* $\tilde{W}$ *can slightly accelerate the convergence of EXTRA.*

REMARK 3 (Step size optimization). *We tried deriving an optimal step size and corresponding explicit linear convergence rate by optimizing certain quantities that appear in the proof, but it becomes quite tricky and messy. For the special case* $\tilde{W} = \frac{I+W}{2}$*, by taking* $\eta \to \mu_{\mathbf{g}}$*, we get a satisfactory step size* $\alpha \to \frac{\mu_{\mathbf{g}}(1+\lambda_{\min}(W))}{4L_{\mathbf{f}}{}^2}$.

REMARK 4 (Step size for ensuring linear convergence). *Interestingly, the critical step size,* $\alpha < \frac{2\mu_{\mathbf{g}}\lambda_{\min}(\tilde{W})}{L_{\mathbf{f}}{}^2} = O\left(\frac{\mu_{\mathbf{g}}}{L_{\mathbf{f}}{}^2}\right)$*, in (3.39) for ensuring the linear convergence, and the parameter* $\alpha = \frac{\tilde{\lambda}_{\min}(\tilde{W}-W)}{2(1+\frac{1}{\gamma^2})(\mu_{\bar{f}}-2L_{\mathbf{f}}\gamma)} = O\left(\frac{\mu_{\mathbf{g}}}{L_{\mathbf{f}}{}^2}\right)$ *in (A.7) for ensuring the restricted strong convexity with* $O(\mu_{\mathbf{g}}) = O(\mu_{\bar{f}})$*, have the same order.*

On the other hand, we numerically observed that a step size as large as $O\left(\frac{1}{L_{\mathbf{f}}}\right)$ *still leads to linear convergence, and EXTRA becomes faster with this larger step size. It remains an open question to prove linear convergence under this larger step size.*

3.4. Decentralized implementation. We shall explain how to perform EXTRA with only local computation and neighbor communication. EXTRA's formula is formed by $\nabla \mathbf{f}(\mathbf{x})$, $W\mathbf{x}$ and $\tilde{W}\mathbf{x}$, and α. By definition $\nabla \mathbf{f}(\mathbf{x})$ is local computation. Assumption 1 part 1 ensures that $W\mathbf{x}$ and $\tilde{W}\mathbf{x}$ can be computed with local and neighbor information. Following our convergence theorems above, determining α requires the bounds on $L_{\mathbf{f}}$ and $\lambda_{\min}(\tilde{W})$, as well as that of $\mu_{\mathbf{g}}$ in the (restricted) strongly convex case. As we have argued at the beginning of Subsection 3.2, it is easy to ensure $\lambda_{\min}(\tilde{W}) \geq \frac{1}{2}$, so $\lambda_{\min}(\tilde{W})$ can be conservatively set as $\frac{1}{2}$. To obtain $L_{\mathbf{f}} = \max_i\{L_{f_i}\}$, a maximum consensus algorithm is needed. On the other hand, it is tricky to determine $\mu_{\mathbf{g}}$ or its lower bound $\mu_{\bar{f}}$, except in the case that each f_i is (restricted) strongly convex, we can conservatively use $\min_i\{\mu_{f_i}\}$. When no bound $\mu_{\mathbf{g}}$ is available in the (restricted) strongly convex case, setting α according to the general convex case (subsection 3.2) often still leads to linear convergence.

4. Numerical Experiments.

4.1. Decentralized Least Squares.

Consider a decentralized sensing problem: each agent $i \in \{1, \cdots, n\}$ holds its own measurement equation, $y_{(i)} = M_{(i)}x + e_{(i)}$, where $y_{(i)} \in \mathbb{R}^{m_i}$ and $M_{(i)} \in \mathbb{R}^{m_i \times p}$ are measured data, $x \in \mathbb{R}^p$ is unknown signal, and $e_{(i)} \in \mathbb{R}^{m_i}$ is unknown noise. The goal is to estimate x. We apply the least squares loss and try to solve

$$\underset{x}{\operatorname{minimize}}\ \bar{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\|M_{(i)}x - y_{(i)}\|_2^2.$$

The network in this experiment is randomly generated with connectivity ratio $r = 0.5$, where r is defined as the number of edges divided by $\frac{L(L-1)}{2}$, the number of all possible ones. We set $n = 10$, $m_i = 1, \forall i$, $p = 5$. Data $y_{(i)}$ and $M_{(i)}$, as well as noise $e_{(i)}$, $\forall i$, are generated following the standard normal distribution. We normalize the data so that $L_{\mathbf{f}} = 1$. The algorithm starts from $x^0_{(i)} = 0, \forall i$, and $\|x^* - x^0_{(i)}\| = 300$.

We use the same matrix W by strategy (iv) in Section 2.4 for both DGD and EXTRA. For EXTRA, we simply use the aforementioned matrix $\tilde{W} = \frac{I+W}{2}$. We run DGD with a fixed step size α, a diminishing one $\frac{\alpha}{k^{1/3}}$ [14], a diminishing one $\frac{\alpha^0}{k^{1/3}}$ with hand-optimized α^0, a diminishing one $\frac{\alpha}{k^{1/2}}$ [5], and a diminishing one $\frac{\alpha^0}{k^{1/2}}$ with hand-optimized α^0, where α is the theoretical critical step size given in [36]. We let EXTRA use the same fixed step size α.

The numerical results are illustrated in Fig. 4.1. In this experiment, we observe that both DGD with the fixed step size and EXTRA show similar linear convergence in the first 200 iterations. Then DGD with the fixed step size begins to slow down and eventually stall, and EXTRA continues its progress.

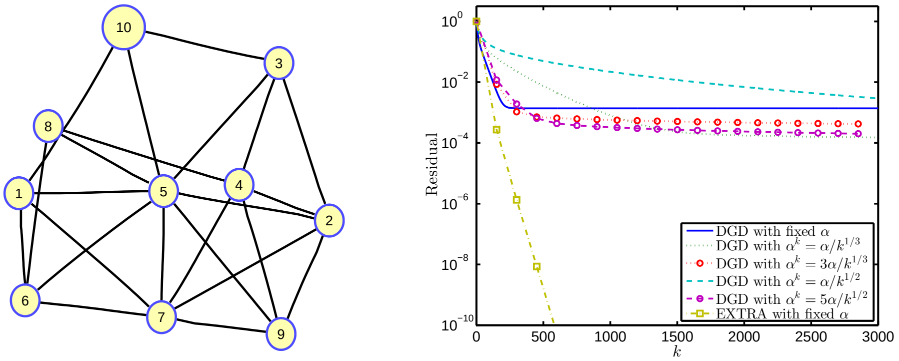

Fig. 4.1. *Plot of residuals* $\frac{\|\mathbf{x}^k-\mathbf{x}^*\|_{\mathrm{F}}}{\|\mathbf{x}^0-\mathbf{x}^*\|_{\mathrm{F}}}$. *Constant* $\alpha = 0.5276$ *is the theoretical critical step size given for DGD in [36]. For DGD with diminishing step sizes* $O(1/k^{1/3})$ *and* $O(1/k^{1/2})$, *we have* hand-optimized *their initial step sizes as* 3α *and* 5α, *respectively.*

4.2. Decentralized Robust Least Squares. Consider the same decentralized sensing setting and network as in Section 4.1. In this experiment, we use the Huber loss, which is known to be robust to outliers, and it allows us to observe both sublinear and linear convergence. We call the problem as decentralized robust least squares:

$$\underset{x}{\operatorname{minimize}}\ \bar{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{m_i} H_\xi(M_{(i)j}x - y_{(i)j})\right\},$$

where $M_{(i)j}$ is the j-th row of matrix $M_{(i)}$ and $y_{(i)j}$ is the j-th entry of vector $y_{(i)}$. The Huber loss function H_ξ is defined as

$$H_\xi(a) = \begin{cases} \frac{1}{2}a^2, & \text{for } |a| \le \xi, \quad (\ell_2^2 \text{ zone}), \\ \xi(|a| - \frac{1}{2}\xi), & \text{otherwise}, \quad (\ell_1 \text{ zone}). \end{cases}$$

We set $\xi = 2$. The optimal solution x^* is artificially set in the ℓ_2^2 zone while $x_{(i)}^0$ is set in the ℓ_1 zone at all agents i.

Except for new hand-optimized initial step sizes for DGD's diminishing step sizes, all other algorithmic parameters remain unchanged from the last test.

The numerical results are illustrated in Fig. 4.2. EXTRA has sublinear convergence for the fist 1000 iterations and then begins linear convergence, as $x_{(i)}^k$ for most i enter the ℓ_2^2 zone.

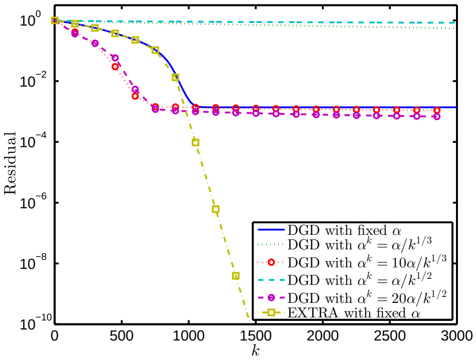

FIG. 4.2. *Plot of residuals* $\frac{\|\mathbf{x}^k-\mathbf{x}^*\|_{\mathrm{F}}}{\|\mathbf{x}^0-\mathbf{x}^*\|_{\mathrm{F}}}$. *Constant* $\alpha = 0.5276$ *is the theoretical critical step size given for DGD in [36]. For DGD with diminishing step sizes* $O(1/k^{1/3})$ *and* $O(1/k^{1/2})$, *we have* hand-optimized *their initial step sizes as* 10α *and* 20α, *respectively. The initial large step sizes have helped them (the red and purple curves) realize faster convergence initially.*

4.3. Decentralized Logistic Regression. Consider the decentralized logistic regression problem:

$$\underset{x}{\operatorname{minimize}}\ \bar{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{m_i}\sum_{j=1}^{m_i}\ln\left(1+\exp\left(-(M_{(i)j}x)y_{(i)j}\right)\right)\right\},$$

where every agent i holds its training date $\left(M_{(i)j}, y_{(i)j}\right) \in \mathbb{R}^p \times \{-1,+1\},\ j = 1,\cdots,m_i$, including explanatory/feature variables $M_{(i)j}$ and binary output/outcome $y_{(i)j}$. To simplify the notation, we set the last entry of every $M_{(i)j}$ to 1 thus the last entry of x will yield the offset parameter of the logistic regression model.

We show a decentralized logistic regression problem solved by DGD and EXTRA over a medium-scale network. The settings are as follows. The connected network is randomly generated with $n = 200$ agents and connectivity ratio $r = 0.2$. Each agent holds 10 samples, i.e., $m_i = 10, \forall i$. The agents shall collaboratively obtain $p = 20$ coefficients via logistic regression. All the 2000 samples are randomly generated, and the reference (ground true) logistic classifier x^* is pre-computed with a centralized method. As it is easy to implement in practice, we use the Metropolis constant edge weight matrix W, which is mentioned by strategy (iii) in Section 2.4, with $\epsilon = 1$, and we use $\tilde{W} = \frac{I+W}{2}$. The numerical results are illustrated in Fig. 4.3. EXTRA outperforms DGD, showing linear and exact convergence to the reference logistic classifier x^*.

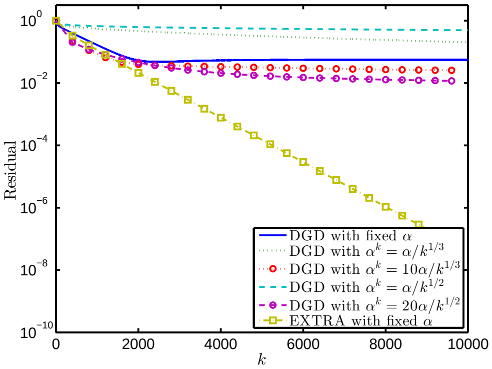

FIG. 4.3. *Plot of residuals* $\frac{\|\mathbf{x}^k-\mathbf{x}^*\|_{\mathrm{F}}}{\|\mathbf{x}^0-\mathbf{x}^*\|_{\mathrm{F}}}$. *Constant* $\alpha = 0.0059$ *is the theoretical critical step size given for DGD in [36]. For DGD with diminishing step sizes* $O(1/k^{1/3})$ *and* $O(1/k^{1/2})$*, we have* hand-optimized *their initial step sizes as* 10α *and* 20α*, respectively.*

5. Conclusion.

As one of the fundamental method, gradient descent has been adapted to decentralized optimization, giving rise to simple and elegant iterations. In this paper, we attempted to address a dilemma or deficiency of the current decentralized gradient descent method: to obtain an accurate solution, it works slowly as it must use a small step size or iteratively diminish the step size; a large step size will lead to faster convergence to, however, an inaccurate solution. Our solution is an exact first-order algorithm, EXTRA, which uses a fixed large step size and quickly returns an accurate solution. The claim is supported by both theoretical convergence and preliminary numerical results. On the other hand, EXTRA is far from perfect, and more work is needed to adapt it to the asynchronous and dynamic network settings. They are interesting open questions for future work.

Appendix A. Proof of Proposition 3.6.

Proof.

"(ii) ⇒ (i)": By definition of restricted strong convexity, there exists $\mu_{\mathbf{g}} > 0$ so that for any $\mathbf{x}$,

$$\begin{aligned}\mu_{\mathbf{g}}\|\mathbf{x}-\mathbf{x}^*\|_{\mathrm{F}}^2 &\leq \langle \nabla\mathbf{g}(\mathbf{x}) - \nabla\mathbf{g}(\mathbf{x}^*), \mathbf{x}-\mathbf{x}^*\rangle \\ &= \langle \nabla\mathbf{f}(\mathbf{x}) - \nabla\mathbf{f}(\mathbf{x}^*), \mathbf{x}-\mathbf{x}^*\rangle + \tfrac{1}{2\alpha}\|\mathbf{x}-\mathbf{x}^*\|_{\tilde{W}-W}^2.\end{aligned} \tag{A.1}$$

For any $x \in \mathbb{R}^p$, set $\mathbf{x} = \mathbf{1}x^T$, and from the above inequality, we get

$$\begin{aligned}\mu_{\mathbf{g}}\|x-x^*\|_2^2 &\leq \tfrac{1}{n}\sum_{i=1}^{n}\langle \nabla f_i(x) - \nabla f_i(x^*), x-x^*\rangle \\ &= \langle \nabla\bar{f}(x) - \nabla\bar{f}(x^*), x-x^*\rangle.\end{aligned} \tag{A.2}$$

Therefore, $\bar{f}(x)$ is restricted strongly convex with a constant $\mu_{\bar{f}} \triangleq \mu_{\mathbf{g}}$.

"(i) ⇒ (ii)": For any $\mathbf{x} \in \mathbb{R}^{n\times p}$, decompose

$$\mathbf{x} = \mathbf{u} + \mathbf{v}$$

so that every column of $\mathbf{u}$ belongs to $\mathrm{span}\{\mathbf{1}\}$ (i.e., $\mathbf{u}$ is consensual) while that of $\mathbf{v}$ belongs to $\mathrm{span}\{\mathbf{1}\}^{\perp}$. Such an *orthogonal* decomposition obviously satisfies $\|\mathbf{x}\|_{\mathrm{F}}^2 = \|\mathbf{u}\|_{\mathrm{F}}^2 + \|\mathbf{v}\|_{\mathrm{F}}^2$. Since solution $\mathbf{x}^*$ is consensual and thus $\langle \mathbf{u}-\mathbf{x}^*, \mathbf{v}\rangle = 0$, we also have $\|\mathbf{x}-\mathbf{x}^*\|_{\mathrm{F}}^2 = \|\mathbf{u}-\mathbf{x}^*\|_{\mathrm{F}}^2 + \|\mathbf{v}\|_{\mathrm{F}}^2$. In addition, being consensual, $\mathbf{u} = \mathbf{1}u^T$ for some $u \in \mathbb{R}^p$. From the inequalities

$$\begin{aligned}
\langle \nabla \mathbf{f}(\mathbf{u}) - \nabla \mathbf{f}(\mathbf{x}^*), \mathbf{u} - \mathbf{x}^*\rangle &= n\frac{1}{n}\sum_{i=1}^{n}\langle \nabla f_i(u) - \nabla f_i(x^*), u - x^*\rangle \\
&\geq n\mu_{\bar f}\|u - x^*\|_2^2 = \mu_{\bar f}\|\mathbf{u} - \mathbf{x}^*\|_{\mathrm{F}}^2, \\
\langle \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{u}), \mathbf{x} - \mathbf{u}\rangle &\geq 0, \\
\langle \nabla \mathbf{f}(\mathbf{u}) - \nabla \mathbf{f}(\mathbf{x}^*), \mathbf{x} - \mathbf{u}\rangle &\geq -L_{\mathbf{f}}\|\mathbf{u} - \mathbf{x}^*\|_{\mathrm{F}}\|\mathbf{v}\|_{\mathrm{F}}, \\
\langle \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{u}), \mathbf{u} - \mathbf{x}^*\rangle &\geq -L_{\mathbf{f}}\|\mathbf{v}\|_{\mathrm{F}}\|\mathbf{u} - \mathbf{x}^*\|_{\mathrm{F}},
\end{aligned}$$

we get

$$\begin{aligned}
& \langle \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle \\
= \;& \langle \nabla \mathbf{f}(\mathbf{u}) - \nabla \mathbf{f}(\mathbf{x}^*), \mathbf{u} - \mathbf{x}^*\rangle + \langle \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{u}), \mathbf{x} - \mathbf{u}\rangle \\
& + \langle \nabla \mathbf{f}(\mathbf{u}) - \nabla \mathbf{f}(\mathbf{x}^*), \mathbf{x} - \mathbf{u}\rangle + \langle \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{u}), \mathbf{u} - \mathbf{x}^*\rangle \\
\geq \;& \mu_{\bar f}\|\mathbf{u} - \mathbf{x}^*\|_{\mathrm{F}}^2 - 2L_{\mathbf{f}}\|\mathbf{u} - \mathbf{x}^*\|_{\mathrm{F}}\|\mathbf{v}\|_{\mathrm{F}}.
\end{aligned} \tag{A.3}$$

In addition, from the fact that $\mathbf{u} - \mathbf{x}^* \in \mathrm{null}\{\tilde W - W\}$ and $\mathbf{v} \in \mathrm{span}\{\tilde W - W\}$, it follows that

$$\tfrac{1}{2\alpha}\|\mathbf{x} - \mathbf{x}^*\|_{\tilde W - W}^2 = \tfrac{1}{2\alpha}\|\mathbf{v}\|_{\tilde W - W}^2 \geq \tfrac{\tilde\lambda_{\min}(\tilde W - W)}{2\alpha}\|\mathbf{v}\|_{\mathrm{F}}^2, \tag{A.4}$$

where $\tilde\lambda_{\min}(\cdot)$ gives the smallest *nonzero* eigenvalue of a positive semidefinite matrix.

Pick any $\gamma > 0$. When $\|\mathbf{v}\|_{\mathrm{F}} \leq \gamma\|\mathbf{u} - \mathbf{x}^*\|_{\mathrm{F}}$, it follows that

$$\begin{aligned}
& \langle \nabla \mathbf{g}(\mathbf{x}) - \nabla \mathbf{g}(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle \\
= \;& \langle \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle + \tfrac{1}{2\alpha}\|\mathbf{x} - \mathbf{x}^*\|_{\tilde W - W}^2 \\
\geq \;& \mu_{\bar f}\|\mathbf{u} - \mathbf{x}^*\|_{\mathrm{F}}^2 - 2L_{\mathbf{f}}\|\mathbf{u} - \mathbf{x}^*\|_{\mathrm{F}}\|\mathbf{v}\|_{\mathrm{F}} + \tfrac{\tilde\lambda_{\min}(\tilde W - W)}{2\alpha}\|\mathbf{v}\|_{\mathrm{F}}^2 \quad \text{(by (A.3) and (A.4))} \\
\geq \;& (\mu_{\bar f} - 2L_{\mathbf{f}}\gamma)\|\mathbf{u} - \mathbf{x}^*\|_{\mathrm{F}}^2 + \tfrac{\tilde\lambda_{\min}(\tilde W - W)}{2\alpha}\|\mathbf{v}\|_{\mathrm{F}}^2 \\
\geq \;& \min\left\{\mu_{\bar f} - 2L_{\mathbf{f}}\gamma, \tfrac{\tilde\lambda_{\min}(\tilde W - W)}{2\alpha}\right\}\|\mathbf{x} - \mathbf{x}^*\|_{\mathrm{F}}^2.
\end{aligned} \tag{A.5}$$

When $\|\mathbf{v}\|_{\mathrm{F}} \geq \gamma\|\mathbf{u} - \mathbf{x}^*\|_{\mathrm{F}}$, it follows that

$$\begin{aligned}
& \langle \nabla \mathbf{g}(\mathbf{x}) - \nabla \mathbf{g}(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle \\
= \;& \langle \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle + \tfrac{1}{2\alpha}\|\mathbf{x} - \mathbf{x}^*\|_{\tilde W - W}^2 \\
\geq \;& 0 + \tfrac{\tilde\lambda_{\min}(\tilde W - W)}{2\alpha}\|\mathbf{v}\|_{\mathrm{F}}^2 \quad \text{(applied convexity of } \mathbf{f} \text{ and (A.4))} \\
\geq \;& \tfrac{\tilde\lambda_{\min}(\tilde W - W)}{2\alpha(1+\frac{1}{\gamma^2})}\|\mathbf{v}\|_{\mathrm{F}}^2 + \tfrac{\tilde\lambda_{\min}(\tilde W - W)}{2\alpha(1+\frac{1}{\gamma^2})}\|\mathbf{u} - \mathbf{x}^*\|_{\mathrm{F}}^2 \\
= \;& \tfrac{\tilde\lambda_{\min}(\tilde W - W)}{2\alpha(1+\frac{1}{\gamma^2})}\|\mathbf{x} - \mathbf{x}^*\|_{\mathrm{F}}^2.
\end{aligned} \tag{A.6}$$

Finally, in all conditions,

$$\begin{aligned}
& \langle \nabla \mathbf{g}(\mathbf{x}) - \nabla \mathbf{g}(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle \\
\geq \;& \min\left\{\mu_{\bar f} - 2L_{\mathbf{f}}\gamma, \tfrac{\tilde\lambda_{\min}(\tilde W - W)}{2\alpha(1+\frac{1}{\gamma^2})}\right\}\|\mathbf{x} - \mathbf{x}^*\|_{\mathrm{F}}^2 \triangleq \mu_{\mathbf{g}}\|\mathbf{x} - \mathbf{x}^*\|_{\mathrm{F}}^2.
\end{aligned} \tag{A.7}$$

By, for example, setting $\gamma = \frac{\mu_{\bar{f}}}{4L_{\mathbf{f}}}$, we have $\mu_{\mathbf{g}} > 0$. Hence, function $\mathbf{g}$ is restricted strongly convex for any $\alpha > 0$ as long as function $\bar{f}$ is restricted strongly convex. $\square$

In the direction of "(ii) ⇒ (i)", we find $\mu_{\mathbf{g}} < \mu_{\bar{f}}$, unlike the more pleasant $\mu_{\bar{f}} = \mu_{\mathbf{g}}$ in the other direction. However, from (A.7), we have

$$\sup_{\gamma,\alpha} \mu_{\mathbf{g}} = \lim_{\gamma \to 0^+} \mu_{\mathbf{g}} \Big|_{\alpha = \frac{\tilde{\lambda}_{\min}(\tilde{W}-W)}{2(1+\frac{1}{\gamma^2})(\mu_{\bar{f}}-2L_{\mathbf{f}}\gamma)}} = \mu_{\bar{f}},$$

which means that $\mu_{\mathbf{g}}$ can be arbitrarily close to $\mu_{\bar{f}}$ as α goes to zero. On the other hand, just to have $O(\mu_{\mathbf{g}}) = O(\mu_{\bar{f}})$, we can set $\gamma = O\left(\frac{\mu_{\bar{f}}}{L_{\mathbf{f}}}\right)$ and $\alpha = \frac{\tilde{\lambda}_{\min}(\tilde{W}-W)}{2(1+\frac{1}{\gamma^2})(\mu_{\bar{f}}-2L_{\mathbf{f}}\gamma)} = O\left(\frac{\mu_{\bar{f}}}{{L_{\mathbf{f}}}^2}\right) = O\left(\frac{\mu_{\mathbf{g}}}{{L_{\mathbf{f}}}^2}\right)$. This order of α coincides, in terms of order of magnitude, with the critical step size for ensuring the linear convergence.

REFERENCES

[1] J. Bazerque and G. Giannakis, *Distributed Spectrum Sensing for Cognitive Radio Networks by Exploiting Sparsity*, IEEE Transactions on Signal Processing, 58 (2010), pp. 1847–1862. 1

[2] J. Bazerque, G. Mateos, and G. Giannakis, *Group-Lasso on Splines for Spectrum Cartography*, IEEE Transactions on Signal Processing, 59 (2011), pp. 4648–4663. 1

[3] S. Boyd, P. Diaconis, and L. Xiao, *Fastest Mixing Markov Chain on a Graph*, SIAM Review, 46 (2004), pp. 667–689. 2.4

[4] T. Chang, M. Hong, and X. Wang, *Multi-Agent Distributed Optimization via Inexact Consensus ADMM*, arXiv preprint arXiv:1402.6065, (2014). 1.1

[5] I. Chen, *Fast Distributed First-Order Methods*, master's thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2012. 1.1, 2.1, 2.4, 4.1

[6] D. Davis and W. Yin, *Convergence Rates of Splitting Algorithms for Optimization.* arXiv preprint arXiv:1406.4834, 2014. 1

[7] A. Dimakis, S. Kar, M. Rabbat J. Moura, and A. Scaglione, *Gossip Algorithms for Distributed Signal Processing*, Proceedings of the IEEE, 98 (2010), pp. 1847–1864. 1

[8] J. Duchi, A. Agarwal, and M. Wainwright, *Dual Averaging for Distributed Optimization: Convergence Analysis and Network Scaling*, IEEE Transactions on Automatic Control, 57 (2012), pp. 592–606. 1.1

[9] P. Forero, A. Cano, and G. Giannakis, *Consensus-Based Distributed Support Vector Machines*, Journal of Machine Learning Research, 59 (2010), pp. 1663–1707. 1

[10] L. Gan, U. Topcu, and S. Low, *Optimal Decentralized Protocol for Electric Vehicle Charging*, IEEE Transactions on Power Systems, 28 (2013), pp. 940–951. 1

[11] B. He, *A New Method for A Class of Linear Variational Inequalities*, Mathematical Programming, 66 (1994), pp. 137–144. 3.2

[12] F. Iutzeler, P. Bianchi, P. Ciblat, and W. Hachem, *Explicit Convergence Rate of a Distributed Alternating Direction Method of Multipliers*, arXiv preprint arXiv:1312.1085, (2013). 1.1

[13] D. Jakovetic, J. Moura, and J. Xavier, *Linear Convergence Rate of Class of Distributed Augmented Lagrangian Algorithms*, arXiv preprint arXiv:1307.2482, (2013). 1.1

[14] D. Jakovetic, J. Xavier, and J. Moura, *Fast Distributed Gradient Methods*, IEEE Transactions on Automatic Control, 59 (2014), pp. 1131–1146. 1.1, 2.1, 4.1

[15] B. Johansson, *On Distributed Optimization in Networked Systems*, PhD thesis, KTH, 2008. 1

[16] V. Kekatos and G. Giannakis, *Distributed Robust Power System State Estimation*, IEEE Transactions on Power Systems, 28 (2013), pp. 1617–1626. 1

[17] M. Lai and W. Yin, *Augmented ℓ_1 and Nuclear-Norm Models with a Globally Linearly Convergent Algorithm*, SIAM Journal on Imaging Sciences, 6 (2013), pp. 1059–1091. 2

[18] Q. Ling and Z. Tian, *Decentralized Sparse Signal Recovery for Compressive Sleeping Wireless Sensor Networks*, IEEE Transactions on Signal Processing, 58 (2010), pp. 3816–3827. 1

[19] Q. Ling, Z. Wen, and W. Yin, *Decentralized Jointly Sparse Recovery by Reweighted ℓ_q Minimization*, IEEE Transactions on Signal Processing, 61 (2013), pp. 1165–70. 1

[20] Q. Ling, Y. Xu, W. Yin, and Z. Wen, *Decentralized Low-rank Matrix Completion*, in Proceedings of the 37th IEEE International Conference on Acoustics, Speech, and Signal Processing, 2012, pp. 2925–2928. 1

[21] I. Matei and J. Baras, *Performance Evaluation of the Consensus-Based Distributed Subgradient Method under Random Communication Topologies*, IEEE Journal of Selected Topics in Signal Processing, 5 (2011), pp. 754–771. 1.1

[22] G. Mateos, J. Bazerque, and G. Giannakis, *Distributed Sparse Linear Regression*, IEEE Transactions on Signal Processing, 58 (2010), pp. 5262–5276. 1

[23] A. Nedic and A. Olshevsky, *Distributed Optimization over Time-Varying Directed Graphs*, in The 52nd IEEE Annual Conference on Decision and Control, 2013, pp. 6855–6860. 1.1, 2.3

[24] ———, *Stochastic Gradient-Push for Strongly Convex Functions on Time-Varying Directed Graphs*, arXiv preprint arXiv:1406.2075, (2014). 1.1

[25] A. Nedic and A. Ozdaglar, *Distributed Subgradient Methods for Multi-agent Optimization*, IEEE Transactions on Automatic Control, 54 (2009), pp. 48–61. 1.1

[26] J. Predd, S. Kulkarni, and H. Poor, *A Collaborative Training Algorithm for Distributed Learning*, IEEE Transactions on Information Theory, 55 (2009), pp. 1856–1871. 1

[27] S. Ram, A. Nedic, and V. Veeravalli, *Distributed Stochastic Subgradient Projection Algorithms for Convex Optimization*, Journal of Optimization Theory and Applications, 147 (2010), pp. 516–545. 1.1

[28] A. Sayed, *Diffusion Adaptation over Networks*, arXiv preprint arXiv:1205.4220, (2012). 2.4

[29] 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, 56 (2008), pp. 350–364. 1

[30] 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, 62 (2014), pp. 1750–1761. 1.1

[31] J. Tsitsiklis, *Problems in Decentralized Decision Making and Computation*, PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1984. 2.4

[32] E. Wei and A. Ozdaglar, *On the $O(1/k)$ Convergence of Asynchronous Distributed Alternating Direction Method of Multipliers*, arXiv preprint arXiv:1307.8254, (2013). 1.1

[33] L. Xiao and S. Boyd, *Fast Linear Iterations for Distributed Averaging*, Systems and Control Letters, 53 (2004), pp. 65–78. 2.4

[34] L. Xiao, S. Boyd, and S. Kim, *Distributed Average Consensus with Least-mean-square Deviation*, Journal of Parallel and Distributed Computing, 67 (2007), pp. 33–46. 1, 2.4

[35] K. Yuan, Q. Ling, A. Ribeiro, and W. Yin, *A Linearized Bregman Algorithm for Decentralized Basis Pursuit*, in Proceedings of the 21st European Signal Processing Conference, 2013, pp. 1–5. 1

[36] K. Yuan, Q. Ling, and W. Yin, *On the Convergence of Decentralized Gradient Descent*, arXiv preprint arXiv:1310.7063, (2013). 1.1, 2.1, 2.4, 4.1, 4.1, 4.2, 4.3

[37] M. Zhu and S. Martinez, *On Distributed Convex Optimization under Inequality and Equality Constraints*, IEEE Transactions on Automatic Control, 57 (2012), pp. 151–164. 1.1