On the Linear Convergence of Two Decentralized Algorithms

Yao Li[1] · Ming Yan[2]

Abstract

Decentralized algorithms solve multi-agent problems over a connected network, where the information can only be exchanged with the accessible neighbors. Though there exist several decentralized optimization algorithms, there are still gaps in convergence conditions and rates between decentralized and centralized algorithms. In this paper, we fill some gaps by considering two decentralized algorithms: EXTRA and NIDS. They both converge linearly with strongly convex objective functions. We will answer two questions regarding them. What are the optimal upper bounds for their stepsizes? Do decentralized algorithms require more properties on the functions for linear convergence than centralized ones? More specifically, we relax the required conditions for linear convergence of both algorithms. For EXTRA, we show that the stepsize is comparable to that of centralized algorithms. For NIDS, the upper bound of the stepsize is shown to be exactly the same as the centralized ones. In addition, we relax the requirement for the objective functions and the mixing matrices. We provide the linear convergence results for both algorithms under the weakest conditions.

Communicated by Zenon Mróz.

✉ Ming Yan
   myan@msu.edu

   Yao Li
   liyao6@msu.edu

[1] Department of Mathematics, Michigan State University, East Lansing, USA

[2] Department of Computational Mathematics, Science and Engineering and Department of Mathematics, Michigan State University, East Lansing, USA

⌖ Springer

1 Introduction

We consider a network of n agents minimizing $\bar{f}(x):=\frac{1}{n}\sum_{i=1}^{n} f_i(x)$ collaboratively. Each convex and differentiable function $f_i: \mathbb{R}^p \rightarrow \mathbb{R}$ is known only by the corresponding agent i. The whole system is decentralized in the sense that each agent has an estimation of the global variable and can only exchange the estimation with their accessible neighbors during each iteration. A symmetric mixing matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ is used to encode the communication weights between the agents and enforce the consensus. The minimum condition for $\mathbf{W}$ is having one eigenvalue 1 with the all-one vector $\mathbf{1}$ being a corresponding eigenvector. All other eigenvalues of $\mathbf{W}$ are less than 1.

Early decentralized methods based on decentralized gradient descent [8,16,18,23, 33] have sublinear convergence for strongly convex objective functions, because of the diminishing stepsize that is needed to obtain a consensual and optimal solution. This sublinear convergence rate is much slower than that for centralized ones. The first decentralized algorithm with linear convergence [25] is based on alternating direction method of multipliers (ADMM) [2,6]. Note that this type of algorithms has $O(1/k)$ rate for general convex functions [3,7,30]. After that, many linearly convergent algorithms are proposed. Some examples are EXTRA [26], NIDS [10], DIGing [19,22], ESOM [12], gradient tracking methods [5,13,19,21,22,32,37], exact diffusion [34,35], and dual optimal [24,29]. There are also works on composite functions, where each private function is the sum of a smooth and a non-smooth functions [1,4,10,27]. Another topic of interest is decentralized optimization over directed and dynamic graphs [11,14,17, 19,28,31,36]. Interested readers are referred to [15] and the references therein.

This paper focuses on two linear convergent algorithms: EXTRA and NIDS, and provides better theoretical convergence results. EXact firsT-ordeR Algorithm (EXTRA) was proposed in [26], and its iteration is described in (2). For the general convex case, where each f_i is convex and L-smooth (i.e., has a L-Lipschitz continuous gradient), the convergence condition in [26] is $\alpha \in \left(0, \frac{1+\lambda_{\min}(\mathbf{W})}{L}\right)$. Therefore, there is an implicit condition for $\mathbf{W}$ that the smallest eigenvalue of $\mathbf{W}$ is larger than -1. Later the condition is relaxed to $\alpha \in \left(0, \frac{5+3\lambda_{\min}(\mathbf{W})}{4L}\right)$ in [9], and the corresponding requirement for $\mathbf{W}$ is that the smallest eigenvalue of $\mathbf{W}$ is larger than $-5/3$. In addition, this condition for the stepsize is shown to be optimal, i.e., EXTRA may diverge if the condition is not satisfied. Though we can always manipulate $\mathbf{W}$ to change the smallest eigenvalue, the convergence speed of EXTRA depends on the matrix $\mathbf{W}$. In the numerical experiment, we will see that it is beneficial to choose small eigenvalues for EXTRA in certain scenarios.

The linear convergence of EXTRA requires additional conditions on the functions. There are mainly three types of conditions used in the literature: the strong convexity of $\bar{f}$ (and some weaker variants) [26], the strong convexity of each f_i (and some weaker variants) [9], and the strong convexity of one function f_i [35]. Note that the condition on $\bar{f}$ is much weaker than the other two; there are cases where $\bar{f}$ is strongly convex but none of f_i's is, e.g., $f_i = \|e_i^\top x\|_2^2$ for $p = n > 1$, where e_i is the vector whose ith component is 1 and all other components are 0. If $\bar{f}$ is (restricted) strongly convex with parameter $\mu_{\bar{f}}$, the linear convergence of EXTRA

is shown when $\alpha \in \left(0, \frac{\mu_{\bar{f}}(1+\lambda_{\min}(\mathbf{W}))}{L^2}\right)$ in [26]. The upper bound for the stepsize is very conservative, and the better performance with a larger stepsize was shown numerically in [26] without proof. If each f_i is strongly convex with parameter μ, the linear convergence is shown when $\alpha \in \left(0, \frac{1+\lambda_{\min}(\mathbf{W})}{L+\mu}\right)$ and $\alpha \in \left(0, \frac{5+3\lambda_{\min}(\mathbf{W})}{4L}\right)$ in [1] and [9], respectively. One contribution of this paper is to show the linear convergence of EXTRA under the (restricted) strong convexity of $\bar{f}$ and $\alpha \in \left(0, \frac{5+3\lambda_{\min}(\mathbf{W})}{4L}\right)$.

The algorithm NIDS (network-independent stepsize) was proposed in [10]. Though there is a small difference from EXTRA, NIDS can choose a stepsize that does not depend on the mixing matrices. The linear convergence of NIDS in [10] requires $\mathbf{I} \succcurlyeq \mathbf{W} \succ -\mathbf{I}$ and strong convexity of $\mathbf{f}(\mathbf{x})$. In this paper, we relax this condition for linear convergence to (restricted) strong convexity of $\bar{f}(x)$ and the relaxed mixing matrices with $\mathbf{I} \succcurlyeq \mathbf{W} \succ -(5/3)\mathbf{I}$.

In sum, we provide new and stronger linear convergence results for two state-of-the-art algorithms: EXTRA and NIDS. More specifically,

- We show the linear convergence of EXTRA under the strong convexity of $\bar{f}$ and the relaxed condition $\lambda_{\min}(\mathbf{W}) > -5/3$. The upper bound of the stepsize can be as large as $\frac{5+3\lambda_{\min}(\mathbf{W})}{4L}$, which is shown to be optimal in [9] for general convex problems;
- We show the linear convergence of NIDS under the same condition on $\bar{f}$ and $\mathbf{W}$ as EXTRA with any network-independent stepsize $\alpha \in (0, 2/L)$.

2 Notation

We let

$$\mathbf{f}(\mathbf{x}) := \sum_{i=1}^{n} f_i(x_i), \tag{1}$$

where each $x_i \in \mathbb{R}^p$ is the local copy of the global variable x and the kth iterated point is x_i^k. Since agent i has its own estimate x_i of the global variable x, we put them together and define

$$\mathbf{x} = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^{n \times p}.$$

The gradient of $\mathbf{f}$ is defined as

$$\nabla \mathbf{f}(\mathbf{x}) = [\nabla f_1(x_1), \nabla f_2(x_2), \ldots, \nabla f_n(x_n)]^\top \in \mathbb{R}^{n \times p}.$$

We say that $\mathbf{x}$ is consensual if $x_1 = x_2 = \cdots = x_n$, i.e., $\mathbf{x} = \mathbf{1}x^\top$, where $x \in \mathbb{R}^{p \times 1}$ and $\mathbf{1} = [1, 1, \ldots, 1]^\top \in \mathbb{R}^{n \times 1}$.

In this paper, we use $\|\cdot\|$ and $\langle\cdot, \cdot\rangle$ to denote the Frobenius norm and the corresponding inner product, respectively. For a given matrix $\mathbf{M} \in \mathbb{R}^{n \times p}$ and any positive (semi)definite matrix $\mathbf{H}$, which is denoted as $\mathbf{H} \succ 0$ ($\mathbf{H} \succcurlyeq 0$ for positive semidefinite),

we define $\|\mathbf{M}\|_{\mathbf{H}} := \sqrt{\mathrm{tr}(\mathbf{M}^\top \mathbf{H} \mathbf{M})}$. The largest and the smallest eigenvalues of a matrix $\mathbf{A}$ are defined as $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$. For a symmetric positive semidefinite matrix $\mathbf{A}$, we let $\lambda_{\min}^+(\mathbf{A})$ be the smallest nonzero eigenvalue. $\mathbf{A}^\dagger$ is the pseudo-inverse of $\mathbf{A}$. For a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, we say a matrix $\mathbf{B} \in \mathbb{R}^{n \times p}$ is in $\mathbf{Ker}\{\mathbf{A}\}$ if $\mathbf{AB} = \mathbf{0}_{n \times p}$, and $\mathbf{B}$ is in $\mathbf{Range}\{\mathbf{A}\}$ if there exists $\mathbf{C} \in \mathbb{R}^{n \times p}$ such that $\mathbf{B} = \mathbf{AC}$. For simplicity, we may use $\mathbf{x}^+$ and $\mathbf{x}$ to replace $\mathbf{x}^{k+1}$ and $\mathbf{x}^k$, respectively, in the proofs.

3 Algorithms and Prerequisites

One iteration of EXTRA can be expressed as

$$\mathbf{x}^{k+2} = (\mathbf{I} + \mathbf{W})\mathbf{x}^{k+1} - \widetilde{\mathbf{W}}\mathbf{x}^k - \alpha[\nabla \mathbf{f}(\mathbf{x}^{k+1}) - \nabla \mathbf{f}(\mathbf{x}^k)]. \tag{2}$$

The stepsize $\alpha > 0$ and the symmetric matrices $\mathbf{W}$ and $\widetilde{\mathbf{W}}$ satisfy $\mathbf{I} + \mathbf{W} \succcurlyeq 2\widetilde{\mathbf{W}} \succcurlyeq 2\mathbf{W}$. The initial value $\mathbf{x}^0$ is chosen arbitrarily, and $\mathbf{x}^1 = \mathbf{W}\mathbf{x}^0 - \alpha \nabla \mathbf{f}(\mathbf{x}^0)$. In practice, we usually let $\widetilde{\mathbf{W}} = \frac{\mathbf{I} + \mathbf{W}}{2}$.

One iteration of NIDS is

$$\mathbf{x}^{k+2} = \frac{\mathbf{I} + \mathbf{W}}{2} \left[2\mathbf{x}^{k+1} - \mathbf{x}^k - \alpha(\nabla \mathbf{f}(\mathbf{x}^{k+1}) - \nabla \mathbf{f}(\mathbf{x}^k)) \right], \tag{3}$$

where $\alpha > 0$ is the stepsize. The initial value $\mathbf{x}^0$ is chosen arbitrarily, and $\mathbf{x}^1 = \frac{\mathbf{I} + \mathbf{W}}{2}[\mathbf{x}^0 - \alpha \nabla \mathbf{f}(\mathbf{x}^0)]$.

If we choose $\widetilde{\mathbf{W}} = \frac{\mathbf{I} + \mathbf{W}}{2}$ in (2), the difference between EXTRA and NIDS in the above mathematical forms happens only in the communicated data, i.e., whether we exchange the gradient information or not at each step. In practice, EXTRA can gain the advantage of time overlap by parallelizing communication and gradient evaluation, while NIDS evaluates the gradient and then communicates after the gradient is added. However, this small difference brings big changes in the convergence [10]. In order for both algorithms to converge, we have the following assumptions on $\mathbf{W}$ and $\widetilde{\mathbf{W}}$.

Assumption 1 (*Mixing matrix*) The connected network $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ consists of a set of nodes $\mathcal{V} = \{1, 2, \ldots, n\}$ and a set of undirected edges $\mathcal{E}$. An undirected edge $(i, j) \in \mathcal{E}$ means that there is a connection between agents i and j and both agents can exchange data. The mixing matrices $\mathbf{W} = [w_{ij}] \in \mathbb{R}^{n \times n}$ and $\widetilde{\mathbf{W}} = [\widetilde{w}_{ij}] \in \mathbb{R}^{n \times n}$ satisfy:

1. (Decentralized property): If $i \neq j$ and $(i, j) \notin \mathcal{E}$, then $w_{ij} = \widetilde{w}_{ij} = 0$.
2. (Symmetry): $\mathbf{W} = \mathbf{W}^\top$, $\widetilde{\mathbf{W}} = \widetilde{\mathbf{W}}^\top$.
3. (Null space property): $\mathbf{Null}\{\mathbf{W} - \widetilde{\mathbf{W}}\} = \mathbf{span}\{\mathbf{1}\} \subseteq \mathbf{Null}\{\mathbf{I} - \widetilde{\mathbf{W}}\}$.
4. (Spectral property): $\frac{\mathbf{I} + \mathbf{W}}{2} \succcurlyeq \widetilde{\mathbf{W}} \succ -\frac{1}{3}\mathbf{I}$, $\quad \widetilde{\mathbf{W}} \succcurlyeq \mathbf{W}$.

Remark 3.1 Parts 2–4 imply that the spectrum of $\mathbf{W}$ is enlarged to $(-\frac{5}{3}, 1]$, while the original assumption is $(-1, 1]$ for doubly stochastic matrices. Therefore, in our assumption, $\frac{\mathbf{I} + \mathbf{W}}{2}$ does not have to be positive definite. This assumption for $\mathbf{W}$ is strictly weaker than those in [26] and [10].

Remark 3.2 From [26, Proposition 2.2], **Null**$\{I - W\} = $ **span**$\{1\}$. It is a critical result for both algorithms.

Before showing their theoretical results, we reformulate both algorithms.
Reformulation of EXTRA: We reformulate EXTRA by introducing a variable $\mathbf{y} \in \mathbb{R}^{n \times p}$ as

$$\mathbf{x}^{k+1} = \widetilde{\mathbf{W}}\mathbf{x}^k + \mathbf{y}^k - \alpha \nabla \mathbf{f}(\mathbf{x}^k), \tag{4a}$$

$$\mathbf{y}^{k+1} = \mathbf{y}^k - (\widetilde{\mathbf{W}} - \mathbf{W})\mathbf{x}^{k+1}, \tag{4b}$$

with $\mathbf{y}^0 = -(\widetilde{\mathbf{W}} - \mathbf{W})\mathbf{x}^0$. Then, (4) is equivalent to EXTRA (2).

Proposition 3.1 *Let the* $\mathbf{x}$*-sequence generated by* (4) *with* $\mathbf{y}^0 = -(\widetilde{\mathbf{W}} - \mathbf{W})\mathbf{x}^0$ *be* $\{\mathbf{x}^k\}_{k=1}^{\infty}$, *then it is identical to the sequence generated by EXTRA* (2) *with the same initial point* $\mathbf{x}^0$.

Proof From (4a), we have

$$\begin{aligned}\mathbf{x}^1 &= \widetilde{\mathbf{W}}\mathbf{x}^0 + \mathbf{y}^0 - \alpha \nabla \mathbf{f}(\mathbf{x}^0) = \widetilde{\mathbf{W}}\mathbf{x}^0 - (\widetilde{\mathbf{W}} - \mathbf{W})\mathbf{x}^0 - \alpha \nabla \mathbf{f}(\mathbf{x}^0) \\ &= \mathbf{W}\mathbf{x}^0 - \alpha \nabla \mathbf{f}(\mathbf{x}^0).\end{aligned}$$

For $k \geq 0$, we have

$$\begin{aligned}\mathbf{x}^{k+2} &= \widetilde{\mathbf{W}}\mathbf{x}^{k+1} + \mathbf{y}^{k+1} - \alpha \nabla \mathbf{f}(\mathbf{x}^{k+1}) = \mathbf{W}\mathbf{x}^{k+1} + \mathbf{y}^k - \alpha \nabla \mathbf{f}(\mathbf{x}^{k+1}) \\ &= (\mathbf{I} + \mathbf{W})\mathbf{x}^{k+1} - \widetilde{\mathbf{W}}\mathbf{x}^k - \alpha [\mathbf{f}(\mathbf{x}^{k+1}) - \mathbf{f}(\mathbf{x}^k)],\end{aligned}$$

where the second and last equalities are from (4b) and (4a), respectively. □

Remark 3.3 By (4b) and the assumption of $\mathbf{y}^0$, $\mathbf{y}^k \in$ **Range**$\{\widetilde{\mathbf{W}} - \mathbf{W}\}$ for all k. Also, $\mathbf{x}^{k+1} = (\widetilde{\mathbf{W}} - \mathbf{W})^{\dagger}(\mathbf{y}^k - \mathbf{y}^{k+1}) + \mathbf{z}^{k+1}$ for some $\mathbf{z}^{k+1} \in$ **Ker**$\{\widetilde{\mathbf{W}} - \mathbf{W}\}$.

Reformulation of NIDS: We adopt the following reformulation from [10]:

$$\mathbf{d}^{k+1} = \mathbf{d}^k + \frac{\mathbf{I} - \mathbf{W}}{2\alpha}[\mathbf{x}^k - \alpha \nabla \mathbf{f}(\mathbf{x}^k) - \alpha \mathbf{d}^k], \tag{5a}$$

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha \nabla \mathbf{f}(\mathbf{x}^k) - \alpha \mathbf{d}^{k+1}, \tag{5b}$$

with $\mathbf{d}^0 = \mathbf{0}$. The equivalence is shown in [10].

To establish the linear convergence of EXTRA and NIDS, we need the following two assumptions.

Assumption 2 (*Uniqueness*) There is a unique minimizer x^* for $\bar{f}(x)$.

Assumption 3 (*L-smoothness and restricted strong convexity*) Each function f_i is a proper, closed and convex function with a Lipschitz continuous gradient:

$$\|\nabla f_i(x) - \nabla f_i(\widetilde{x})\| \leq L\|x - \widetilde{x}\|, \ \forall x, \ \widetilde{x} \in \mathbb{R}^p, \tag{6}$$

where $L > 0$ is the Lipschitz constant. Furthermore, $\bar{f}(x)$ is (restricted) strongly convex with respect to x^*:

$$\langle x - x^*, \nabla \bar{f}(x) - \nabla \bar{f}(x^*) \rangle \geq \mu_{\bar{f}} \|x - x^*\|^2, \ \forall x \in \mathbb{R}^p. \tag{7}$$

From [20, Theorem 2.1.5], the inequality (6) is equivalent to, for any $\mathbf{x}, \ \tilde{\mathbf{x}} \in \mathbb{R}^{n \times p}$,

$$\langle \mathbf{x} - \tilde{\mathbf{x}}, \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\tilde{\mathbf{x}}) \rangle \geq L^{-1} \|\nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\tilde{\mathbf{x}})\|^2. \tag{8}$$

Proposition 3.2 ([26, Appendix A]) *The following two statements are equivalent:*

1. *$\bar{f}(x)$ is (restricted) strongly convex with respect to x^*;*
2. *For any $\eta > 0$, $\mathbf{g}(\mathbf{x}) := \mathbf{f}(\mathbf{x}) + \frac{\eta}{2} \|\mathbf{x}\|_{\mathbf{I}-\mathbf{W}}^2$ is $\mu_{\mathbf{g}}$ (restricted) strongly convex with respect to $\mathbf{x}^* = \mathbf{1}(x^*)^\top$. Specially, we can characterize*

$$\mu_{\mathbf{g}} = \min \left\{ \frac{\mu_{\bar{f}}}{2}, \frac{\mu_{\bar{f}}^2 \lambda_{\min}^+ (\mathbf{I} - \mathbf{W})}{\mu_{\bar{f}}^2 + 16L^2} \eta \right\}.$$

This proposition shows

$$\langle \mathbf{x} - \mathbf{x}^*, \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{x}^*) \rangle + \eta \|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{I}-\mathbf{W}}^2 \geq \mu_{\mathbf{g}} \|\mathbf{x} - \mathbf{x}^*\|^2 \tag{9}$$

for any $\mathbf{x} \in \mathbb{R}^{n \times p}$.

4 New Linear Convergence Results for EXTRA and NIDS

Throughout this section, we assume that Assumptions 1–3 hold. Two techniques are used to show the linear convergence: (a) Proposition 3.2 serves as a bridge to connect $\bar{f}(x)$ and $\mathbf{f}(\mathbf{x})$. It is the key to the weaker assumption on objective functions. (b) Both algorithms are equivalent to the extended proximal alternating predictor–corrector (PAPC) in [9], and this equivalence is the key to relaxing the conditions on the mixing matrices $\mathbf{W}$ and $\widetilde{\mathbf{W}}$.

4.1 Linear Convergence of EXTRA

When $\widetilde{\mathbf{W}} = \frac{\mathbf{I}+\mathbf{W}}{2}$, EXTRA is recovered by applying the extended PAPC in [9] to the following dual form of the decentralized consensus problem

$$\text{minimize}_{\mathbf{y}} \ \mathbf{f}^*(\sqrt{\mathbf{I} - \mathbf{W}} \mathbf{y}),$$

where $\mathbf{f}^*$ is the conjugate function of $\mathbf{f}$ and $\mathbf{y}$ is the dual variable. In this case, EXTRA has the optimal bound of the stepsize over the relaxed mixing matrix $\mathbf{W} \succ -(5/3)\mathbf{I}$. This fact enlightens us on the critical Lemma 4.3.

For simplicity, we introduce some notations. Because of part 4 of Assumption 1, given the mixing matrices $\mathbf{W}$ and $\widetilde{\mathbf{W}}$, there is a constant

$$\theta \in \left(\frac{3}{4}, \min\left\{\frac{1}{1 - \lambda_{\min}(\widetilde{\mathbf{W}})}, 1\right\}\right]$$

such that

$$\overline{\mathbf{W}} := \theta\widetilde{\mathbf{W}} + (1 - \theta)\mathbf{I} \succ \mathbf{0}, \tag{10}$$

$$\mathbf{H} := \overline{\mathbf{W}} + (\theta - \tfrac{1}{2})(\mathbf{I} - \widetilde{\mathbf{W}}) = \tfrac{\mathbf{I} + \widetilde{\mathbf{W}}}{2} \succ \mathbf{0}, \tag{11}$$

$$\mathbf{M} := (\widetilde{\mathbf{W}} - \mathbf{W})^{\dagger} \succcurlyeq \mathbf{0}, \tag{12}$$

$$\mathbf{G} := \mathbf{W} + \mathbf{I} - 2\widetilde{\mathbf{W}} \succcurlyeq \mathbf{0}. \tag{13}$$

Based on (10), we have

$$\widetilde{\mathbf{W}} = \overline{\mathbf{W}} - (1 - \theta)(\mathbf{I} - \widetilde{\mathbf{W}}). \tag{14}$$

Let $(\mathbf{x}^*, \mathbf{y}^*)$ be a fixed point of (4), it is straightforward to show that

$$(\widetilde{\mathbf{W}} - \mathbf{W})\mathbf{x}^* = \mathbf{0}. \tag{15}$$

Part 3 of Assumption 1 shows that $\mathbf{x}^*$ is consensual, i.e., $\mathbf{x}^* = \mathbf{1}(x^*)^{\top}$ for certain $x^* \in \mathbb{R}^p$. The $\mathbf{y}$-iteration in (4b) and the initialization of $\mathbf{y}^0$ show $\mathbf{y}^k \in \mathbf{Range}\{\widetilde{\mathbf{W}} - \mathbf{W}\} = \mathbf{Ker}\{\mathbf{1}^{\top}\}$. Then, we have $\mathbf{1}^{\top}\mathbf{y}^* = \alpha\mathbf{1}^{\top}\nabla\mathbf{f}(\mathbf{x}^*) = 0$. Thus, x^* is the unique minimizer of $\bar{f}(x)$.

Lemma 4.1 (Norm over range space [10, Lemma 3]) *For any symmetric positive (semi)definite matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ with rank r ($r \leq n$), let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ be its r eigenvalues. Then, $\mathbf{Range}\{\mathbf{A}\}$ is a rp-dimensional subspace in $\mathbb{R}^{n \times p}$ and has a norm defined by $\|\mathbf{x}\|^2_{\mathbf{A}^{\dagger}} := \langle\mathbf{x}, \mathbf{A}^{\dagger}\mathbf{x}\rangle$, where $\mathbf{A}^{\dagger}$ is the pseudo-inverse of $\mathbf{A}$. In addition, $\lambda_1^{-1}\|\mathbf{x}\|^2 \leq \|\mathbf{x}\|^2_{\mathbf{A}^{\dagger}} \leq \lambda_r^{-1}\|\mathbf{x}\|^2$ for all $\mathbf{x} \in \mathbf{Range}\{\mathbf{A}\}$.*

For simplicity, we let $\mathbf{x}^+$ and $\mathbf{x}$ stand for $\mathbf{x}^{k+1}$ and $\mathbf{x}^k$, respectively, in the proofs. The same simplification applies to $\mathbf{y}^k$.

Lemma 4.2 (Norm equality) *Let $\{(\mathbf{x}^k, \mathbf{y}^k)\}_{k=1}^{\infty}$ be the sequence generated by (4), then it satisfies*

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2_{\widetilde{\mathbf{W}} - \mathbf{W}} = \|\mathbf{y}^k - \mathbf{y}^{k+1}\|^2_{\mathbf{M}}. \tag{16}$$

Proof From Remark 3.3, we have

$$\mathbf{x}^+ = \mathbf{M}(\mathbf{y} - \mathbf{y}^+) + \mathbf{z}^+ \tag{17}$$

for $\mathbf{z}^+ \in \mathbf{Ker}\{\widetilde{\mathbf{W}} - \mathbf{W}\}$. This equality and (15) give

$$
\begin{aligned}
\|\mathbf{x}^+ - \mathbf{x}^*\|_{\widetilde{\mathbf{W}}-\mathbf{W}}^2 &= \langle \mathbf{x}^+ - \mathbf{x}^*, (\widetilde{\mathbf{W}} - \mathbf{W})(\mathbf{x}^+ - \mathbf{x}^*) \rangle = \langle \mathbf{x}^+, (\widetilde{\mathbf{W}} - \mathbf{W})\mathbf{x}^+ \rangle \\
&= \langle \mathbf{M}(\mathbf{y} - \mathbf{y}^+), \mathbf{y} - \mathbf{y}^+ \rangle = \|\mathbf{y} - \mathbf{y}^+\|_{\mathbf{M}}^2,
\end{aligned}
$$

where the third equality holds because of (12), (17), and $\mathbf{y} - \mathbf{y}^+ \in \mathbf{Range}(\widetilde{\mathbf{W}} - \mathbf{W})$.
□

Lemma 4.3 (A key inequality for EXTRA) *Let* $\{(\mathbf{x}^k, \mathbf{y}^k)\}_{k=1}^{\infty}$ *be the sequence generated by* (4)*, then we have*

$$
\begin{aligned}
&\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{\mathbf{H}}^2 + \|\mathbf{y}^{k+1} - \mathbf{y}^*\|_{\mathbf{M}}^2 \\
&\leq \|\mathbf{x}^k - \mathbf{x}^*\|_{\mathbf{H}}^2 + \|\mathbf{y}^k - \mathbf{y}^*\|_{\mathbf{M}}^2 - \|\mathbf{x}^k - \mathbf{x}^{k+1}\|_{(\theta-\frac{3}{4})(\mathbf{I}-\widetilde{\mathbf{W}})}^2 - \|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{\mathbf{G}}^2 \\
&\quad - \|\mathbf{x}^k - \mathbf{x}^{k+1}\|_{\overline{\mathbf{W}}}^2 - 2\alpha\langle \mathbf{x}^{k+1} - \mathbf{x}^*, \nabla \mathbf{f}(\mathbf{x}^k) - \nabla \mathbf{f}(\mathbf{x}^*) \rangle.
\end{aligned}
\tag{18}
$$

Proof The iteration (4) and equation (14) show

$$
\begin{aligned}
&2\alpha\langle \mathbf{x}^+ - \mathbf{x}^*, \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{x}^*) \rangle \\
&= 2\langle \mathbf{x}^+ - \mathbf{x}^*, \widetilde{\mathbf{W}}(\mathbf{x} - \mathbf{x}^+) + \widetilde{\mathbf{W}}(\mathbf{x}^+ - \mathbf{x}^*) - (\mathbf{x}^+ - \mathbf{x}^*) + (\mathbf{y} - \mathbf{y}^*) \rangle \\
&= 2\langle \mathbf{x}^+ - \mathbf{x}^*, \widetilde{\mathbf{W}}(\mathbf{x} - \mathbf{x}^+) + (\widetilde{\mathbf{W}} - \mathbf{I})(\mathbf{x}^+ - \mathbf{x}^*) \\
&\quad + (\widetilde{\mathbf{W}} - \mathbf{W})(\mathbf{x}^+ - \mathbf{x}^*) + \mathbf{y}^+ - \mathbf{y} + \mathbf{y} - \mathbf{y}^* \rangle \\
&= 2\langle \mathbf{x}^+ - \mathbf{x}^*, \widetilde{\mathbf{W}}(\mathbf{x} - \mathbf{x}^+) \rangle + 2\langle \mathbf{x}^+ - \mathbf{x}^*, \mathbf{y}^+ - \mathbf{y}^* \rangle - 2\|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{G}}^2 \\
&= 2\langle \mathbf{x}^+ - \mathbf{x}^*, \overline{\mathbf{W}}(\mathbf{x} - \mathbf{x}^+) \rangle - 2\langle \mathbf{x}^+ - \mathbf{x}^*, (1-\theta)(\mathbf{I} - \widetilde{\mathbf{W}})(\mathbf{x} - \mathbf{x}^+) \rangle \\
&\quad + 2\langle \mathbf{x}^+ - \mathbf{x}^*, \mathbf{y}^+ - \mathbf{y}^* \rangle - 2\|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{G}}^2,
\end{aligned}
\tag{19}
$$

where the first equality comes from (4a), the second one follows (4b), and the last one is from (14). From Remark 3.3, $\mathbf{x}^+ - \mathbf{x}^* = \mathbf{M}(\mathbf{y} - \mathbf{y}^+) + \mathbf{z}^+ - \mathbf{x}^*$ for some $\mathbf{z}^+ \in \mathbf{Ker}\{\widetilde{\mathbf{W}} - \mathbf{W}\}$. Thus, $\langle \mathbf{z}^+ - \mathbf{x}^*, \mathbf{y}^+ - \mathbf{y}^* \rangle = 0$, and the equality (19) can be rewritten as

$$
\begin{aligned}
&2\alpha\langle \mathbf{x}^+ - \mathbf{x}^*, \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{x}^*) \rangle \\
&= 2\langle \mathbf{x}^+ - \mathbf{x}^*, \overline{\mathbf{W}}(\mathbf{x} - \mathbf{x}^+) \rangle - 2\langle \mathbf{x}^+ - \mathbf{x}^*, (1-\theta)(\mathbf{I} - \widetilde{\mathbf{W}})(\mathbf{x} - \mathbf{x}^+) \rangle \\
&\quad + 2\langle \mathbf{M}(\mathbf{y} - \mathbf{y}^+), \mathbf{y}^+ - \mathbf{y}^* \rangle - 2\|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{G}}^2.
\end{aligned}
$$

Using the basic equality $2\langle a - b, b - c \rangle = \|a - c\|^2 - \|a - b\|^2 - \|b - c\|^2$ and Lemma 4.2, we have

$$
\begin{aligned}
&\|\mathbf{x}^+ - \mathbf{x}^*\|_{\overline{\mathbf{W}}}^2 - \|\mathbf{x}^+ - \mathbf{x}^*\|_{(1-\theta)(\mathbf{I}-\widetilde{\mathbf{W}})}^2 + \|\mathbf{y}^+ - \mathbf{y}^*\|_{\mathbf{M}}^2 \\
&= \|\mathbf{x} - \mathbf{x}^*\|_{\overline{\mathbf{W}}}^2 - \|\mathbf{x} - \mathbf{x}^*\|_{(1-\theta)(\mathbf{I}-\widetilde{\mathbf{W}})}^2 + \|\mathbf{y} - \mathbf{y}^*\|_{\mathbf{M}}^2 \\
&\quad - \|\mathbf{x} - \mathbf{x}^+\|_{\overline{\mathbf{W}}}^2 + \|\mathbf{x} - \mathbf{x}^+\|_{(1-\theta)(\mathbf{I}-\widetilde{\mathbf{W}})}^2 - \|\mathbf{x}^+ - \mathbf{x}^*\|_{\widetilde{\mathbf{W}}-\mathbf{W}}^2
\end{aligned}
$$

$$- 2\|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{G}}^2 - 2\alpha\langle\mathbf{x}^+ - \mathbf{x}^*, \nabla\mathbf{f}(\mathbf{x}) - \nabla\mathbf{f}(\mathbf{x}^*)\rangle. \tag{20}$$

Note that the following inequality holds,

$$\tfrac{1}{2}\|\mathbf{x}^+ - \mathbf{x}^*\|_{\widetilde{\mathbf{W}}-\mathbf{W}}^2 \leq \|\mathbf{x}^+ - \mathbf{x}^*\|_{\widetilde{\mathbf{W}}-\mathbf{W}}^2 + \tfrac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_{\widetilde{\mathbf{W}}-\mathbf{W}}^2 - \tfrac{1}{4}\|\mathbf{x} - \mathbf{x}^+\|_{\widetilde{\mathbf{W}}-\mathbf{W}}^2.$$

Adding it onto both sides of (20), we have

$$\begin{aligned}
&\|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{H}}^2 - \tfrac{1}{2}\|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{G}}^2 + \|\mathbf{y}^+ - \mathbf{y}^*\|_{\mathbf{M}}^2 \\
\leq &\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{H}}^2 - \tfrac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{G}}^2 + \|\mathbf{y} - \mathbf{y}^*\|_{\mathbf{M}}^2 \\
&- \|\mathbf{x} - \mathbf{x}^+\|_{\widetilde{\mathbf{W}}}^2 - \|\mathbf{x} - \mathbf{x}^+\|_{(\theta-\frac{3}{4})(\mathbf{I}-\widetilde{\mathbf{W}})}^2 + \tfrac{1}{4}\|\mathbf{x} - \mathbf{x}^+\|_{\mathbf{G}}^2 \\
&- 2\|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{G}}^2 - 2\alpha\langle\mathbf{x}^+ - \mathbf{x}^*, \nabla\mathbf{f}(\mathbf{x}) - \nabla\mathbf{f}(\mathbf{x}^*)\rangle.
\end{aligned} \tag{21}$$

Apply the inequality $\tfrac{1}{4}\|\mathbf{x} - \mathbf{x}^+\|_{\mathbf{G}}^2 \leq \tfrac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{G}}^2 + \tfrac{1}{2}\|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{G}}^2$, then the key inequality (18) is obtained. □

In the following theorem, we assume $\mathbf{G} \neq \mathbf{0}$ (i.e., $\widetilde{\mathbf{W}} \neq (\mathbf{I} + \mathbf{W})/2$). It is easy to amend the proof to show the result for this special case.

Theorem 4.1 (Q-linear convergence of EXTRA) *Under Assumptions 1–3, we define*

$$r_1 = \frac{4\theta - 3}{4(1-\theta)^2\lambda_{\max}(\overline{\mathbf{W}}^{-1}(\mathbf{I}-\widetilde{\mathbf{W}}))} > 0, \tag{22}$$

$$r_2 = \frac{1}{2\lambda_{\max}(\mathbf{G}\overline{\mathbf{W}}^{-1})} > 0, \tag{23}$$

$$r_3 = \frac{r_1 r_2}{r_1 + r_2 + r_1 r_2} \in (0, 1) \tag{24}$$

and choose two small parameters ξ and η such that

$$\xi \in \left(0, \min\left\{\frac{r_3}{4\lambda_{\max}(\overline{\mathbf{W}}\mathbf{M})}, 1\right\}\right), \tag{25}$$

$$\eta \in \left(0, \frac{\lambda_{\min}(\overline{\mathbf{W}})\xi}{4\alpha\lambda_{\min}(\overline{\mathbf{W}}) - 2\alpha^2 L}\right). \tag{26}$$

In addition, we define

$$\mathbf{P} := \mathbf{H} + \tfrac{\xi}{2}(\mathbf{I} - \mathbf{W}) \succ \mathbf{0},$$
$$\mathbf{Q} := \mathbf{M} + (r_3 - 2\xi\lambda_{\max}(\overline{\mathbf{W}}\mathbf{M}))\overline{\mathbf{W}}^{-1} \succ \mathbf{0}.$$

Then, for any stepsize $\alpha \in (0, \frac{2\lambda_{\min}(\overline{\mathbf{W}})}{L})$, we have

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{\mathbf{P}}^2 + \|\mathbf{y}^{k+1} - \mathbf{y}^*\|_{\mathbf{Q}}^2 \leq \rho(\|\mathbf{x}^k - \mathbf{x}^*\|_{\mathbf{P}}^2 + \|\mathbf{y}^k - \mathbf{y}^*\|_{\mathbf{Q}}^2), \tag{27}$$

where

$$\rho := \max\Big\{ 1 - \Big(2\alpha - \frac{\alpha^2 L}{\lambda_{\min}(\overline{\mathbf{W}})}\Big)\mu_{\mathbf{g}}, \Big(4\alpha - \frac{2\alpha^2 L}{\lambda_{\min}(\overline{\mathbf{W}})}\Big)\frac{\eta}{\xi},$$
$$1 - \frac{r_3 - 4\xi\lambda_{\max}(\overline{\mathbf{W}}\mathbf{M})}{r_3 + (1 - 2\xi)\lambda_{\max}(\overline{\mathbf{W}}\mathbf{M})}\Big\} < 1. \tag{28}$$

Proof From (18) in Lemma 4.3, we have

$$\|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{H}}^2 + \|\mathbf{y}^+ - \mathbf{y}^*\|_{\mathbf{M}}^2$$
$$\leq \|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{H}}^2 + \|\mathbf{y} - \mathbf{y}^*\|_{\mathbf{M}}^2 - \|\mathbf{x} - \mathbf{x}^+\|_{(\theta - \frac{3}{4})(\mathbf{I} - \widetilde{\mathbf{W}})}^2 - \|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{G}}^2$$
$$- \|\mathbf{x} - \mathbf{x}^+\|_{\overline{\mathbf{W}}}^2 - 2\alpha\langle \mathbf{x}^+ - \mathbf{x}^*, \nabla\mathbf{f}(\mathbf{x}) - \nabla\mathbf{f}(\mathbf{x}^*)\rangle. \tag{29}$$

Then, we find an upper bound of $-\|\mathbf{x} - \mathbf{x}^+\|_{\overline{\mathbf{W}}}^2 - 2\alpha\langle \mathbf{x}^+ - \mathbf{x}^*, \nabla\mathbf{f}(\mathbf{x}) - \nabla\mathbf{f}(\mathbf{x}^*)\rangle$.

$$- \|\mathbf{x} - \mathbf{x}^+\|_{\overline{\mathbf{W}}}^2 - 2\alpha\langle \mathbf{x}^+ - \mathbf{x}^*, \nabla\mathbf{f}(\mathbf{x}) - \nabla\mathbf{f}(\mathbf{x}^*)\rangle$$
$$= \alpha^2\|\nabla\mathbf{f}(\mathbf{x}) - \nabla\mathbf{f}(\mathbf{x}^*)\|_{\overline{\mathbf{W}}^{-1}}^2 - 2\alpha\langle \mathbf{x} - \mathbf{x}^*, \nabla\mathbf{f}(\mathbf{x}) - \nabla\mathbf{f}(\mathbf{x}^*)\rangle$$
$$- \|\overline{\mathbf{W}}(\mathbf{x} - \mathbf{x}^+) - \alpha(\nabla\mathbf{f}(\mathbf{x}) - \nabla\mathbf{f}(\mathbf{x}^*))\|_{\overline{\mathbf{W}}^{-1}}^2$$
$$\leq -\Big(2\alpha - \frac{\alpha^2 L}{\lambda_{\min}(\overline{\mathbf{W}})}\Big)\langle \mathbf{x} - \mathbf{x}^*, \nabla\mathbf{f}(\mathbf{x}) - \nabla\mathbf{f}(\mathbf{x}^*)\rangle$$
$$- \|\overline{\mathbf{W}}(\mathbf{x} - \mathbf{x}^+) - \alpha(\nabla\mathbf{f}(\mathbf{x}) - \nabla\mathbf{f}(\mathbf{x}^*))\|_{\overline{\mathbf{W}}^{-1}}^2,$$

where the inequality comes from (8). Combining it with (29), we have

$$\|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{H}}^2 + \|\mathbf{y}^+ - \mathbf{y}^*\|_{\mathbf{M}}^2 - \|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{H}}^2 - \|\mathbf{y} - \mathbf{y}^*\|_{\mathbf{M}}^2$$
$$\leq -\Big(2\alpha - \frac{\alpha^2 L}{\lambda_{\min}(\overline{\mathbf{W}})}\Big)\langle \mathbf{x} - \mathbf{x}^*, \nabla\mathbf{f}(\mathbf{x}) - \nabla\mathbf{f}(\mathbf{x}^*)\rangle$$
$$- \|\overline{\mathbf{W}}(\mathbf{x} - \mathbf{x}^+) - \alpha(\nabla\mathbf{f}(\mathbf{x}) - \nabla\mathbf{f}(\mathbf{x}^*))\|_{\overline{\mathbf{W}}^{-1}}^2$$
$$- \|\mathbf{x} - \mathbf{x}^+\|_{(\theta - \frac{3}{4})(\mathbf{I} - \widetilde{\mathbf{W}})}^2 - \|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{G}}^2. \tag{30}$$

The inequality (30) shows that $\{(\mathbf{x}^k, \mathbf{y}^k)\}_{k=1}^{\infty}$ is a Cauchy sequence converging to the fixed point $(\mathbf{x}^*, \mathbf{y}^*)$ of (4). From (9), we can bound the first term on the right-hand side of (30) as

$$-\Big(2\alpha - \frac{\alpha^2 L}{\lambda_{\min}(\overline{\mathbf{W}})}\Big)\langle \mathbf{x} - \mathbf{x}^*, \nabla\mathbf{f}(\mathbf{x}) - \nabla\mathbf{f}(\mathbf{x}^*)\rangle$$
$$\leq \Big(2\alpha - \frac{\alpha^2 L}{\lambda_{\min}(\overline{\mathbf{W}})}\Big)\eta\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{I} - \mathbf{W}}^2 - \Big(2\alpha - \frac{\alpha^2 L}{\lambda_{\min}(\overline{\mathbf{W}})}\Big)\mu_{\mathbf{g}}\|\mathbf{x} - \mathbf{x}^*\|^2. \tag{31}$$

Next, we bound the two terms involving successive iterated points, i.e., $-\|\overline{\mathbf{W}}(\mathbf{x} - \mathbf{x}^+) - \alpha(\nabla\mathbf{f}(\mathbf{x}) - \nabla\mathbf{f}(\mathbf{x}^*))\|_{\overline{\mathbf{W}}^{-1}}^2 - \|\mathbf{x} - \mathbf{x}^+\|_{(\theta - \frac{3}{4})(\mathbf{I} - \widetilde{\mathbf{W}})}^2$. Note that

$$\overline{\mathbf{W}}(\mathbf{x} - \mathbf{x}^+) - \alpha(\nabla\mathbf{f}(\mathbf{x}) - \nabla\mathbf{f}(\mathbf{x}^*))$$

$$= \mathbf{G}(\mathbf{x}^+ - \mathbf{x}^*) - (\mathbf{y}^+ - \mathbf{y}^*) + (1 - \theta)(\mathbf{I} - \widetilde{\mathbf{W}})(\mathbf{x} - \mathbf{x}^+). \tag{32}$$

We use T_1, T_2, and T_3 to denote the three terms on the right-hand side of (32), respectively. Using the definition of r_1 in (22), we have

$$
\begin{aligned}
&- \|T_1 + T_2 + T_3\|_{\overline{\mathbf{W}}^{-1}}^2 - \|\mathbf{x} - \mathbf{x}^+\|_{(\theta - \frac{3}{4})(\mathbf{I} - \widetilde{\mathbf{W}})}^2 \\
&= - \|T_1 + T_2\|_{\overline{\mathbf{W}}^{-1}}^2 - 2\langle \overline{\mathbf{W}}^{-\frac{1}{2}}(T_1 + T_2), \overline{\mathbf{W}}^{-\frac{1}{2}} T_3 \rangle - \|T_3\|_{\overline{\mathbf{W}}^{-1}}^2 \\
&\quad - \tfrac{4\theta - 3}{4(1-\theta)} \|\mathbf{x} - \mathbf{x}^+\|_{(1-\theta)(\mathbf{I} - \widetilde{\mathbf{W}})}^2 \\
&\leq - \|T_1 + T_2\|_{\overline{\mathbf{W}}^{-1}}^2 - 2\langle \overline{\mathbf{W}}^{-\frac{1}{2}}(T_1 + T_2), \overline{\mathbf{W}}^{-\frac{1}{2}} T_3 \rangle - (1 + r_1)\|T_3\|_{\overline{\mathbf{W}}^{-1}}^2 \\
&\leq - \tfrac{r_1}{1+r_1} \|T_1 + T_2\|_{\overline{\mathbf{W}}^{-1}}^2,
\end{aligned}
$$

where the last inequality comes from the Cauchy inequality

$$-2\langle a, b \rangle \leq \tfrac{1}{1+r_1} \|a\|^2 + (1 + r_1)\|b\|^2.$$

Combining it with the last term $-\|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{G}}^2$ on the right-hand side of (30), we have

$$
\begin{aligned}
&- \tfrac{r_1}{1+r_1} \|T_1 + T_2\|_{\overline{\mathbf{W}}^{-1}}^2 - \|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{G}}^2 \\
&\leq - \tfrac{r_1}{1+r_1} \|T_2\|_{\overline{\mathbf{W}}^{-1}}^2 - \tfrac{2r_1}{1+r_1} \langle \overline{\mathbf{W}}^{-\frac{1}{2}} T_1, \overline{\mathbf{W}}^{-\frac{1}{2}} T_2 \rangle - \tfrac{r_1}{1+r_1} \|T_1\|_{\overline{\mathbf{W}}^{-1}}^2 \\
&\quad - r_2 \|T_1\|_{\overline{\mathbf{W}}^{-1}}^2 - \tfrac{1}{2} \|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{G}}^2 \\
&\leq - r_3 \|\mathbf{y}^+ - \mathbf{y}^*\|_{\overline{\mathbf{W}}^{-1}}^2 - \tfrac{\xi}{2} \|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{G}}^2, \tag{33}
\end{aligned}
$$

where $\xi < 1$ is a small positive parameter, and r_2 and r_3 are defined as (23) and (24), respectively. Since $\mathbf{G} = (\mathbf{I} - \mathbf{W}) - 2(\widetilde{\mathbf{W}} - \mathbf{W})$, we have

$$\|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{G}}^2 = \|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{I} - \mathbf{W}}^2 - 2\|\mathbf{y} - \mathbf{y}^+\|_{\mathbf{M}}^2. \tag{34}$$

Therefore,

$$
\begin{aligned}
&- \tfrac{r_1}{1+r_1} \|T_1 + T_2\|_{\overline{\mathbf{W}}^{-1}}^2 - \|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{G}}^2 \\
&\leq - r_3 \|\mathbf{y}^+ - \mathbf{y}^*\|_{\overline{\mathbf{W}}^{-1}}^2 - \tfrac{\xi}{2} \|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{I} - \mathbf{W}}^2 - \xi \|\mathbf{y} - \mathbf{y}^+\|_{\mathbf{M}}^2 \\
&\leq - r_3 \|\mathbf{y}^+ - \mathbf{y}^*\|_{\overline{\mathbf{W}}^{-1}}^2 - \tfrac{\xi}{2} \|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{I} - \mathbf{W}}^2 + 2\xi \|\mathbf{y}^+ - \mathbf{y}^*\|_{\mathbf{M}}^2 + 2\xi \|\mathbf{y} - \mathbf{y}^*\|_{\mathbf{M}}^2 \\
&\leq - (r_3/\lambda_{\max}(\overline{\mathbf{W}}\mathbf{M}) - 2\xi)\|\mathbf{y}^+ - \mathbf{y}^*\|_{\mathbf{M}}^2 - \tfrac{\xi}{2} \|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{I} - \mathbf{W}}^2 + 2\xi \|\mathbf{y} - \mathbf{y}^*\|_{\mathbf{M}}^2. \tag{35}
\end{aligned}
$$

Let $\xi < r_3/(4\lambda_{\max}(\overline{\mathbf{W}}\mathbf{M}))$, then we have $r_3/\lambda_{\max}(\overline{\mathbf{W}}\mathbf{M}) - 2\xi > 2\xi$. Putting (31) and (35) together onto (30), we have

$$\|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{H}}^2 + \tfrac{\xi}{2} \|\mathbf{x}^+ - \mathbf{x}^*\|_{\mathbf{I} - \mathbf{W}}^2 + (1 + (r_3/\lambda_{\max}(\overline{\mathbf{W}}\mathbf{M}) - 2\xi))\|\mathbf{y}^+ - \mathbf{y}^*\|_{\mathbf{M}}^2$$

$$\leq \Big(1 - \Big(2\alpha - \tfrac{\alpha^2 L}{\lambda_{\min}(\overline{\mathbf{W}})}\Big)\mu_{\mathbf{g}}\Big)\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{H}}^2 + \Big(2\alpha - \tfrac{\alpha^2 L}{\lambda_{\min}(\overline{\mathbf{W}})}\Big)\eta\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{I} - \mathbf{W}}^2$$
$$+ (1 + 2\xi)\|\mathbf{y} - \mathbf{y}^*\|_{\mathbf{M}}^2.$$

Let ρ be defined as (28), we get (27). Note that the choice of ξ and η affects the definition of $\mathbf{P}$ and $\mathbf{Q}$, but not the algorithm. Hence, for any $\alpha \in (0, \tfrac{2\lambda_{\min}(\overline{\mathbf{W}})}{L})$, Q-linear convergence is guaranteed for $(\mathbf{x}^k - \mathbf{x}^*, \mathbf{y}^k - \mathbf{y}^*)$.

Because $\|\mathbf{x}^k - \mathbf{x}^*\|_{\mathbf{P}}^2 \leq \|\mathbf{x}^k - \mathbf{x}^*\|_{\mathbf{P}}^2 + \|\mathbf{y}^k - \mathbf{y}^*\|_{\mathbf{Q}}^2$, the sequence $\{\|\mathbf{x}^k - \mathbf{x}^*\|_{\mathbf{P}}^2\}_{k=1}^{\infty}$ converges R-linearly to 0 at the rate of ρ. □

Two special cases are not covered by the theorem: $\theta = 1$ and $\widetilde{\mathbf{W}} = \tfrac{\mathbf{I} + \mathbf{W}}{2}$. When $\theta = 1$, we have $r_1 = \infty$ and $r_3 = \tfrac{r_2}{1 + r_2}$. When $\widetilde{\mathbf{W}} = \tfrac{\mathbf{I} + \mathbf{W}}{2}$, i.e., $\mathbf{G} = \mathbf{0}$, we have $r_2 = \infty$ and $r_3 = \tfrac{r_1}{1 + r_1}$. In both cases, the linear convergence rate is

$$\rho = \max\Big\{\, 1 - \Big(2\alpha - \tfrac{2\alpha^2 L}{2 - \theta + \theta\lambda_{\min}(\mathbf{W})}\Big)\mu_{\mathbf{g}}, \Big(4\alpha - \tfrac{4\alpha^2 L}{2 - \theta + \theta\lambda_{\min}(\mathbf{W})}\Big)\tfrac{\eta}{\xi}, \\ 1 - \tfrac{\beta r_3 - 4\xi(2 - \theta\beta)}{\beta r_3 + (1 - 2\xi)(2 - \theta\beta)} \Big\}, \tag{36}$$

where $\beta = 1 - \lambda_2(\mathbf{W})$ is the spectral gap. It is exactly the limit of ρ in (28) with r_1 or r_2 approaching infinity.

Remark 4.1 The upper bound for the stepsize α, $2(1 - \theta + \theta\lambda_{\min}(\widetilde{\mathbf{W}}))/L$, is much larger than that in [26] for ensuring linear convergence, $2\mu_{\mathbf{g}}\lambda_{\min}(\widetilde{\mathbf{W}})/L^2$, when $\widetilde{\mathbf{W}}$ is positive definite. In the special case $\widetilde{\mathbf{W}} = (\mathbf{I} + \mathbf{W})/2$, we have $\alpha < (2 - \theta + \theta\lambda_{\min}(\mathbf{W}))/L$. Since we can choose θ as close as possible to 3/4, the upper bound of α attains $(3\lambda_{\min}(\mathbf{W}) + 5)/(4L)$, which coincides the optimal bound given in [9] for general convex functions. In [9], the linear convergence was shown under the strong convexity of all functions $\{f_i\}_{i=1}^n$.

4.2 NIDS Without Non-smooth Term

We consider NIDS next. In the smooth case, NIDS can be recovered by PAPC applied to the primal form of the decentralized consensus problem

$$\text{minimize}_{\mathbf{x}} \ \mathbf{f}(\mathbf{x}), \quad \text{s.t.} \sqrt{\mathbf{I} - \mathbf{W}}\mathbf{x} = \mathbf{0}.$$

It motivates us to show the inequality in Lemma 4.5.

[10, Lemma 1] shows that, with the initialization $(\mathbf{d}^0 = \mathbf{0}, \mathbf{x}^0)$, the fixed point $(\mathbf{d}^* \in \mathbf{Range}(\mathbf{I} - \mathbf{W}), \mathbf{x}^*)$ of (5) satisfies

$$\mathbf{d}^* + \nabla\mathbf{f}(\mathbf{x}^*) = \mathbf{0}, \tag{37a}$$

$$(\mathbf{I} - \mathbf{W})\mathbf{x}^* = \mathbf{0}, \tag{37b}$$

and $\mathbf{x}^*$ is the consensual solution to the problem (1). We will use the following important equality, which can be derived from (5)

$$(\mathbf{I} - \tfrac{\mathbf{I}-\mathbf{W}}{2})(\mathbf{d}^{k+1} - \mathbf{d}^k) = \tfrac{\mathbf{I}-\mathbf{W}}{2\alpha}(\mathbf{x}^{k+1} - \mathbf{x}^*). \tag{38}$$

Motivated by the proof of EXTRA, we introduce another matrix to measure the distance to the fixed point. We still pick $\theta \in (\tfrac{3}{4}, 1]$ such that

$$\theta\left(\tfrac{\mathbf{I}+\mathbf{W}}{2}\right) + (1-\theta)\mathbf{I} = \mathbf{I} - \theta\left(\tfrac{\mathbf{I}-\mathbf{W}}{2}\right) \succ \mathbf{0}. \tag{39}$$

Define a new symmetric matrix

$$\widetilde{\mathbf{M}} = 2(\mathbf{I} - \mathbf{W})^{\dagger} - \theta\mathbf{I} = \left(\tfrac{\mathbf{I}-\mathbf{W}}{2}\right)^{\dagger} - \theta\mathbf{I}. \tag{40}$$

Then, $\widetilde{\mathbf{M}}$ is a norm over $\mathbf{Range}(\mathbf{I} - \mathbf{W})$. Note that $\widetilde{\mathbf{M}}$ is invertible because $\widetilde{\mathbf{M}}\mathbf{1} = -\theta\mathbf{1}$. In the following proofs, we use the same simplification $\mathbf{x}$ and $\mathbf{x}^+$.

Lemma 4.4 *(Equality) Let $\{(\mathbf{d}^k, \mathbf{x}^k)\}_{k=1}^{\infty}$ be the sequence generated by (5), we have the following two equalities:*

$$\langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{d}^{k+1} - \mathbf{d}^* \rangle = \alpha\langle \mathbf{d}^{k+1} - \mathbf{d}^k, \mathbf{d}^{k+1} - \mathbf{d}^* \rangle_{\widetilde{\mathbf{M}}-(1-\theta)\mathbf{I}} \tag{41a}$$

$$\langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{d}^{k+1} - \mathbf{d}^k \rangle = \alpha\|\mathbf{d}^{k+1} - \mathbf{d}^k\|^2_{\widetilde{\mathbf{M}}-(1-\theta)\mathbf{I}}. \tag{41b}$$

Proof Since $\mathbf{d}^+ - \mathbf{d}^* \in \mathbf{Range}(\mathbf{I} - \mathbf{W})$, we have

$$\begin{aligned} \langle \mathbf{x}^+ - \mathbf{x}^*, \mathbf{d}^+ - \mathbf{d}^* \rangle &= \langle (\mathbf{I} - \mathbf{W})(\mathbf{x}^+ - \mathbf{x}^*), (\mathbf{I} - \mathbf{W})^{\dagger}(\mathbf{d}^+ - \mathbf{d}^*) \rangle \\ &= \alpha\langle (2\mathbf{I} - (\mathbf{I} - \mathbf{W}))(\mathbf{d}^+ - \mathbf{d}), (\mathbf{I} - \mathbf{W})^{\dagger}(\mathbf{d}^+ - \mathbf{d}^*) \rangle \\ &= \alpha\langle (2(\mathbf{I} - \mathbf{W})^{\dagger} - \mathbf{I})(\mathbf{d}^+ - \mathbf{d}), \mathbf{d}^+ - \mathbf{d}^* \rangle, \end{aligned} \tag{42}$$

where the second equality follows (38). Replacing $\mathbf{d}^*$ with $\mathbf{d}$ in (42), we get (41b) in the same way. $\square$

Lemma 4.5 *(A key inequality for NIDS) Let $\{(\mathbf{d}^k, \mathbf{x}^k)\}_{k=1}^{\infty}$ be the sequence generated by (5). We have, with any $r_4 \in (0, \theta - \tfrac{3}{4})$,*

$$\begin{aligned} &\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 + \alpha^2\|\mathbf{d}^{k+1} - \mathbf{d}^*\|^2_{\widetilde{\mathbf{M}}+(\theta-\frac{1}{2}+2r_4)\mathbf{I}} \\ \leq &\|\mathbf{x}^k - \mathbf{x}^*\|^2 + \alpha^2\|\mathbf{d}^k - \mathbf{d}^*\|^2_{\widetilde{\mathbf{M}}+(\theta-\frac{1}{2}-2r_4)\mathbf{I}} - \alpha^2\|\mathbf{d}^k - \mathbf{d}^{k+1}\|^2_{\widetilde{\mathbf{M}}+(\theta-\frac{3}{4}-r_4)\mathbf{I}} \\ &+ \alpha^2\|\nabla\mathbf{f}(\mathbf{x}^k) - \nabla\mathbf{f}(\mathbf{x}^*)\|^2 - 2\alpha\langle \mathbf{x}^k - \mathbf{x}^*, \nabla\mathbf{f}(\mathbf{x}^k) - \nabla\mathbf{f}(\mathbf{x}^*) \rangle. \end{aligned} \tag{43}$$

Proof The iteration (5) and the definition of $\widetilde{\mathbf{M}}$ in (40) show

$$
\begin{aligned}
&2\alpha \langle \mathbf{x} - \mathbf{x}^*, \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{x}^*) \rangle \\
&= 2\langle \mathbf{x} - \mathbf{x}^*, \mathbf{x} - \mathbf{x}^+ \rangle - 2\alpha \langle \mathbf{x} - \mathbf{x}^*, \mathbf{d}^+ - \mathbf{d}^* \rangle \\
&= 2\langle \mathbf{x} - \mathbf{x}^+, \mathbf{x} - \mathbf{x}^* \rangle - 2\alpha \langle \mathbf{x} - \mathbf{x}^*, \mathbf{d}^+ - \mathbf{d}^* \rangle - 2\alpha \langle \mathbf{x}^+ - \mathbf{x}^*, \mathbf{d}^+ - \mathbf{d}^* \rangle \\
&= 2\langle \mathbf{x} - \mathbf{x}^+, \mathbf{x} - \alpha \mathbf{d}^+ - \mathbf{x}^* + \alpha \mathbf{d}^* \rangle + 2\alpha^2 \langle \mathbf{d} - \mathbf{d}^+, \mathbf{d}^+ - \mathbf{d}^* \rangle_{\widetilde{\mathbf{M}} - (1-\theta)\mathbf{I}} \\
&= 2\langle \mathbf{x} - \mathbf{x}^+, \mathbf{x}^+ - \mathbf{x}^* + \alpha \nabla \mathbf{f}(\mathbf{x}) - \alpha \nabla \mathbf{f}(\mathbf{x}^*) \rangle + 2\alpha^2 \langle \mathbf{d} - \mathbf{d}^+, \mathbf{d}^+ - \mathbf{d}^* \rangle_{\widetilde{\mathbf{M}} - (1-\theta)\mathbf{I}},
\end{aligned}
$$

where the first and the last equalities use (5b) and the third one follows (41a). From (5b), we obtain

$$
\begin{aligned}
&2\alpha \langle \mathbf{x} - \mathbf{x}^+, \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{x}^*) \rangle \\
&= \|\mathbf{x} - \mathbf{x}^+\|^2 + \alpha^2 \|\nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{x}^*)\|^2 - \|\mathbf{x} - \mathbf{x}^+ - \alpha \nabla \mathbf{f}(\mathbf{x}) + \alpha \nabla \mathbf{f}(\mathbf{x}^*)\|^2 \\
&= \|\mathbf{x} - \mathbf{x}^+\|^2 + \alpha^2 \|\nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{x}^*)\|^2 - \alpha^2 \|\mathbf{d}^+ - \mathbf{d}^*\|^2.
\end{aligned}
\tag{44}
$$

Together with the basic equality $2\langle a - b, b - c \rangle = \|a - c\|^2 - \|b - c\|^2 - \|a - b\|^2$, we get

$$
\begin{aligned}
&\|\mathbf{x}^+ - \mathbf{x}^*\|^2 + \alpha^2 \|\mathbf{d}^+ - \mathbf{d}^*\|^2_{\widetilde{\mathbf{M}} - (1-\theta)\mathbf{I}} \\
&= \|\mathbf{x} - \mathbf{x}^*\|^2 + \alpha^2 \|\mathbf{d} - \mathbf{d}^*\|^2_{\widetilde{\mathbf{M}} - (1-\theta)\mathbf{I}} - \alpha^2 \|\mathbf{d} - \mathbf{d}^+\|^2_{\widetilde{\mathbf{M}} - (1-\theta)\mathbf{I}} - \alpha^2 \|\mathbf{d}^+ - \mathbf{d}^*\|^2 \\
&\quad + \alpha^2 \|\nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{x}^*)\|^2 - 2\alpha \langle \mathbf{x} - \mathbf{x}^*, \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{x}^*) \rangle.
\end{aligned}
\tag{45}
$$

Since $r_4 < \theta - \frac{3}{4} \le 1/4$, the following inequality holds,

$$
-\left(\tfrac{1}{2} - 2r_4\right)\|\mathbf{d}^+ - \mathbf{d}^*\|^2 \le \left(\tfrac{1}{2} - 2r_4\right)\|\mathbf{d} - \mathbf{d}^*\|^2 - \left(\tfrac{1}{4} - r_4\right)\|\mathbf{d} - \mathbf{d}^+\|^2.
$$

Adding it onto both sides of (45), we get (43). $\qquad\square$

Theorem 4.2 (Q-linear convergence for NIDS) *Under Assumptions 1-3, we define*

$$
r_5 = \max \left(2, \frac{(\lambda_{\max}(\mathbf{I} - \mathbf{W}) - 2)^2}{2 - (\frac{3}{4} + r_4)\lambda_{\max}(\mathbf{I} - \mathbf{W})} \right).
\tag{46}
$$

For any stepsize $\alpha \in (0, \frac{2}{L})$, we choose $\eta \in (0, \frac{1}{\alpha(2-\alpha L)r_5})$ and define

$$
\rho_3 = \max \left\{ 1 - \alpha(2 - \alpha L)\mu_{\mathbf{g}}, \alpha(2 - \alpha L)\eta r_5, 1 - \frac{4r_4}{2\lambda_{\max}((\mathbf{I}-\mathbf{W})^+) - \frac{1}{2} + 2r_4} \right\} < 1,
\tag{47}
$$

Then, we have

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2_{\mathbf{I}+\frac{\mathbf{I}-\mathbf{W}}{r_5}} + \alpha^2 \|\mathbf{d}^{k+1} - \mathbf{d}^*\|^2_{\mathbf{Q}}$$

$$\leq \rho(\|\mathbf{x}^k - \mathbf{x}^*\|^2_{\mathbf{I}+\frac{\mathbf{I}-\mathbf{W}}{r_5}} + \alpha^2 \|\mathbf{d}^k - \mathbf{d}^*\|^2_{\mathbf{Q}}), \tag{48}$$

where $\mathbf{Q}:=\widetilde{\mathbf{M}} + (\theta - \frac{1}{2} + 2r_4)\mathbf{I} \succ 0$.

Proof Given any $\alpha \in (0, \frac{2}{L})$, we have

$$\alpha^2 \|\nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{x}^*)\|^2 - 2\alpha\langle \mathbf{x} - \mathbf{x}^*, \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{x}^*)\rangle$$

$$\leq -\alpha(2 - \alpha L)\langle \mathbf{x} - \mathbf{x}^*, \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{x}^*)\rangle$$

$$= -\alpha(2 - \alpha L)\langle \mathbf{x} - \mathbf{x}^*, \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{x}^*)\rangle - \alpha(2 - \alpha L)\eta\|\mathbf{x} - \mathbf{x}^*\|^2_{\mathbf{I}-\mathbf{W}}$$

$$+ \alpha(2 - \alpha L)\eta\|\mathbf{x} - \mathbf{x}^*\|^2_{\mathbf{I}-\mathbf{W}}$$

$$\leq -\alpha(2 - \alpha L)\mu_{\mathbf{g}}\|\mathbf{x} - \mathbf{x}^*\|^2 + \alpha(2 - \alpha L)\eta\|\mathbf{x} - \mathbf{x}^*\|^2_{\mathbf{I}-\mathbf{W}},$$

where the first inequality is from (8) and the second one uses (restricted) strong convexity (9). Together with (43), we have

$$\|\mathbf{x}^+ - \mathbf{x}^*\|^2 + \alpha^2\|\mathbf{d}^+ - \mathbf{d}^*\|^2_{\widetilde{\mathbf{M}}+(\theta-\frac{1}{2}+2r_4)\mathbf{I}}$$

$$\leq \|\mathbf{x} - \mathbf{x}^*\|^2 + \alpha^2\|\mathbf{d} - \mathbf{d}^*\|^2_{\widetilde{\mathbf{M}}+(\theta-\frac{1}{2}-2r_4)\mathbf{I}} - \alpha^2\|\mathbf{d} - \mathbf{d}^+\|^2_{\widetilde{\mathbf{M}}+(\theta-\frac{3}{4}-r_4)\mathbf{I}}$$

$$- \alpha(2 - \alpha L)\mu_{\mathbf{g}}\|\mathbf{x} - \mathbf{x}^*\|^2 + \alpha(2 - \alpha L)\eta\|\mathbf{x} - \mathbf{x}^*\|^2_{\mathbf{I}-\mathbf{W}}, \tag{49}$$

The equality (38) gives

$$\|\mathbf{x}^+ - \mathbf{x}^*\|^2_{\mathbf{I}-\mathbf{W}}$$

$$= \|(\mathbf{I} - \mathbf{W})(\mathbf{x}^+ - \mathbf{x}^*)\|^2_{(\mathbf{I}-\mathbf{W})^\dagger} = \alpha^2\|(2\mathbf{I} - (\mathbf{I} - \mathbf{W}))(\mathbf{d}^+ - \mathbf{d})\|^2_{(\mathbf{I}-\mathbf{W})^\dagger}$$

$$= \alpha^2\|\mathbf{d} - \mathbf{d}^+\|^2_{(2\mathbf{I}-(\mathbf{I}-\mathbf{W}))(\mathbf{I}-\mathbf{W})^\dagger(2\mathbf{I}-(\mathbf{I}-\mathbf{W}))} = \alpha^2\|\mathbf{d} - \mathbf{d}^+\|^2_{4(\mathbf{I}-\mathbf{W})^\dagger-4\mathbf{I}+(\mathbf{I}-\mathbf{W})}$$

$$\leq \alpha^2 r_5\|\mathbf{d} - \mathbf{d}^+\|^2_{\widetilde{\mathbf{M}}+(\theta-\frac{3}{4}-r_4)\mathbf{I}}, \tag{50}$$

where the second equality follows (38), the fourth equality comes from $\mathbf{d} - \mathbf{d}^+ \in$ **Range**$(\mathbf{I} - \mathbf{W})$, and the inequality holds with the definition of r_5 in (46). Combing (49) and (50), we derive

$$\|\mathbf{x}^+ - \mathbf{x}^*\|^2 + \frac{1}{r_5}\|\mathbf{x}^+ - \mathbf{x}^*\|^2_{\mathbf{I}-\mathbf{W}} + \alpha^2\|\mathbf{d}^+ - \mathbf{d}^*\|^2_{\widetilde{\mathbf{M}}+(\theta-\frac{1}{2}+2r_4)\mathbf{I}}$$

$$\leq (1 - \alpha(2 - \alpha L)\mu_{\mathbf{g}})\|\mathbf{x} - \mathbf{x}^*\|^2 + \alpha(2 - \alpha L)\eta\|\mathbf{x} - \mathbf{x}^*\|^2_{\mathbf{I}-\mathbf{W}}$$

$$+ \alpha^2\|\mathbf{d} - \mathbf{d}^*\|^2_{\widetilde{\mathbf{M}}+(\theta-\frac{1}{2}-2r_4)\mathbf{I}}. \tag{51}$$

Let ρ_3 be defined as (47), and we show (48). Meanwhile, the Q-linear convergence of $(\mathbf{d}^k, \mathbf{x}^k)$ implies the R-linear convergence of $\mathbf{x}^k$. □

This theorem shows that NIDS is still linearly convergent over a relaxed $\mathbf{W}$ and keeps the network-independent stepsize, which attains $\frac{2}{L}$ practically.

5 Numerical Experiments

In this section, we compare the performance of EXTRA and NIDS over the relaxed mixing matrices in the following two scenarios:[1]

- Comparison of decentralized gradient descent (DGD), EXTRA, and NIDS with different stepsizes for a doubly stochastic matrix $\mathbf{W}$.
- Comparison of EXTRA and NIDS with different stepsizes for a relaxed matrix $\mathbf{W}$.

We consider the following decentralized sensing problem. Each agent $i \in \{1, \ldots, n\}$ has its own private measured data $M_i \in \mathbb{R}^{m_i \times p}$ and $y_i \in \mathbb{R}^{m_i}$ based on the unknown common variable $x \in \mathbb{R}^p$. Suppose that $y_i = M_i x + e_i$ with independently identically distributed random noise $e_i \in \mathbb{R}^{m_i}$. The goal is to estimate x cooperatively over the network, and the problem is

$$\text{minimize}_x \ \ \bar{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \| M_i x - y_i \|_2^2.$$

The data $\{M_i\}_{i=1}^n$ and x are generated from Gaussian distribution. We normalize each M_i such that $\| M_i^\top M_i \| = 10$, i.e., $L = 10$. In both scenarios, we set $n = 10$, $p = 5$, $\mathbf{x}^0 = \mathbf{0}$, and $\widetilde{\mathbf{W}} = \frac{\mathbf{I}+\mathbf{W}}{2}$ for EXTRA.

For the first scenario, we construct the matrix $\mathbf{W}$ based on the Metropolis constant edge weight matrix in [26, §2.4]. In this case, $\widetilde{\mathbf{W}}$ is positive definite, and we can set $\theta = \frac{3}{4}$. Then, $\overline{\mathbf{W}} = \frac{5\mathbf{I}+3\mathbf{W}}{8}$. We implement EXTRA with three different stepsizes: $\alpha_1 = \frac{(1+\lambda_{\min}(\mathbf{W}))\mu_{\bar{f}}}{100}$ (the stepsize for linear convergence in [26]), $\alpha_2 = \frac{1+\lambda_{\min}(\mathbf{W})}{10}$ (the stepsize for convergence only in [26]), and $\alpha_3 = \frac{5+3\lambda_{\min}(\mathbf{W})}{40}$ (our largest stepsize). For NIDS, the stepsize is set to $\alpha_4 = \frac{1}{5}$ although it is the upper bound of the stepsize which is not attainable in our proof theoretically.

The result with $m_i = 1$ is illustrated in Fig. 1. Because we have $n > p$, the function $\bar{f}(x)$ is strongly convex with probability one. NIDS requires the least number of iteration to attain the expected tolerance. Meanwhile, EXTRA with our proposed stepsize has better performance than that given in [26].

Then, we set $m_i = 10$ in Fig. 2. In this case, individual functions $f_i(x)$ and $\mathbf{f}(\mathbf{x})$ are strongly convex. NIDS and EXTRA with the largest stepsize lead the performance. Here, two results of EXTRA are the same as that of NIDS although they are set with different stepsizes. The observation may indicate that there is an optimal choice of stepsize between α_2 and α_3 for both EXTRA and NIDS. By setting $\alpha_5 = \frac{5+3\lambda_{\min}(\mathbf{W})}{40+\mu_{\bar{f}}}$ for EXTRA and $\alpha_6 = \frac{2}{10+\mu_{\bar{f}}}$, we compare these algorithms in Fig. 3. This figure

---

[1] The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.
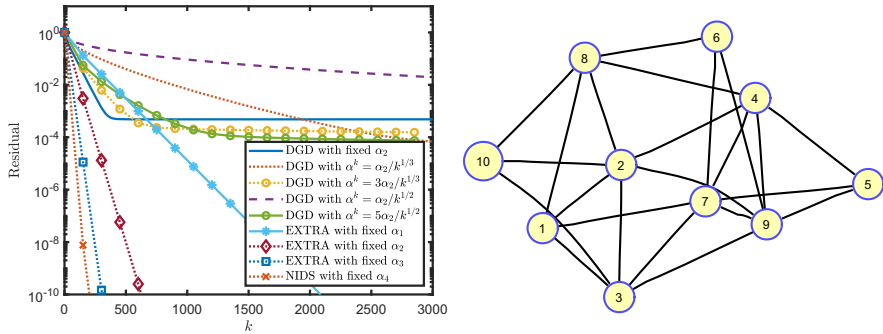
Fig. 1 LEFT: the error $\frac{\|\mathbf{x}^k-\mathbf{x}^*\|_F}{\|\mathbf{x}^0-\mathbf{x}^*\|_F}$ vs iterations for DGD with different stepsizes, EXTRA with three stepsizes, and NIDS. RIGHT: The random network with 10 nodes

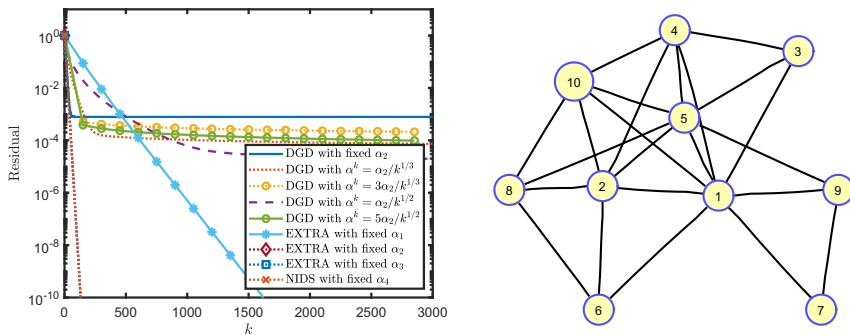**Fig. 2** LEFT: the error $\frac{\|\mathbf{x}^k-\mathbf{x}^*\|_F}{\|\mathbf{x}^0-\mathbf{x}^*\|_F}$ vs iterations for DGD with different stepsizes, EXTRA with three stepsizes, and NIDS. RIGHT: The random network with 10 nodes

suggests that the optimal stepsize may depend on the problem/functions. How to find the optimal stepsize is an important research topic and beyond the scope of this paper.

Next, we turn to the relaxed mixing matrices. Based on the previous created $\mathbf{W}$, we replace it by $\mathbf{W}_{\mathrm{new}} = \frac{4\mathbf{W}-\mathbf{I}}{3}$ to scale the range of eigenvalues to $(-\frac{5}{3}, 1]$. In this case, some diagonal entries of $\mathbf{W}_{\mathrm{new}}$ may be negative. We consider the worst topology of network, line topology, i.e., each agent has at most two neighbors. In this experiment, we solve the same problem using EXTRA and NIDS on the unrelaxed and relaxed mixing matrices, respectively, over the line. For NIDS, since the stepsize is network-independent, we relax the mixing matrix $\mathbf{W}$ to $\mathbf{W}_{\mathrm{new}}$ more aggressively so that $\lambda_{\min}(\mathbf{W}_{\mathrm{new}})$ approaches $-\frac{5}{3}$ and compare the performance with the unrelaxed case of NIDS under $\alpha = \frac{1}{5}$. For EXTRA, we set the stepsize to $\alpha = \frac{5+3\lambda_{\min}(\mathbf{W})}{40}$ and compare the performance with the relaxed one under the stepsize $\alpha = \frac{5+3\lambda_{\min}(\mathbf{W}_{\mathrm{new}})}{40}$ where we only perturb $\mathbf{W}$ mildly so that $\lambda_{\min}(\mathbf{W}_{\mathrm{new}})$ approaches -1. The result is shown in Fig. 4.
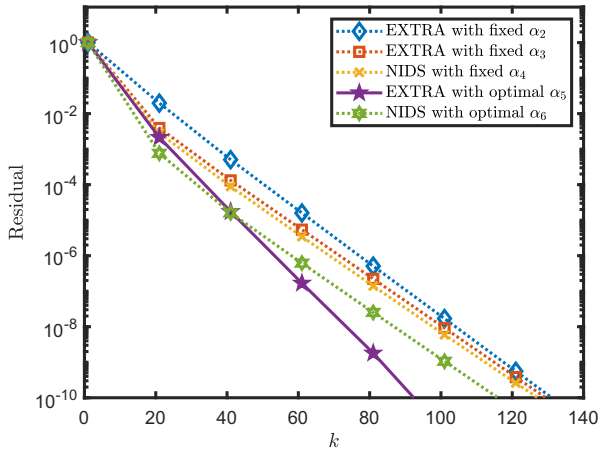
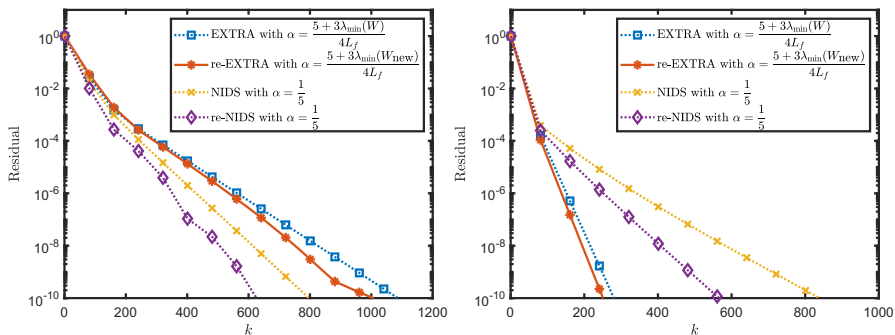**Fig. 3** The comparison of proved stepsizes for EXTRA and NIDS with the optimal choice

Fig. 4 The figure of residuals $\frac{\|\mathbf{x}^k - \mathbf{x}^*\|_F}{\|\mathbf{x}^0 - \mathbf{x}^*\|_F}$ with respect to iteration. The left graph is for strongly convex $\bar{f}(x)$, and the right one is for strongly convex $\mathbf{f}(\mathbf{x})$. re-EXTRA and re-NIDS stand for implementing EXTRA and NIDS over relaxed $\mathbf{W}_{\text{new}}$

From Fig. 4, if the topology of network is weak, switching to the relaxed mixing matrix may offer better performance when using NIDS and EXTRA to solve the problem. The improvement for NIDS is more distinguished.

6 Conclusion

In this paper, we relax the mixing matrices and prove the linear convergence of EXTRA and NIDS under the (restricted) strongly convexity assumption on $\bar{f}$. A larger upper bound of the stepsize is derived for EXTRA compared with that given in [26] and [1]. NIDS can choose a network-independent stepsize, and this stepsize can be chosen as

the same as that of centralized ones. We relax the conditions for the mixing matrices and the functions, while keeping the same stepsize.

In numerical experiments on linear regression, EXTRA with the larger stepsize converges faster than using the $\mu_{\bar{f}}$-dependent stepsize in [26]. Over the unrelaxed mixing matrix, NIDS leads the performance in most cases and is the easiest to implement. If the topology of network is weak, using the relaxed mixing matrix can accelerate NIDS. For EXTRA, in general, we may not choose the mixing matrices to be relaxed due to the tiny improvement, but the larger stepsize derived in the relaxed case is competent to be considered.

References

1. Alghunaim, S., Yuan, K., Sayed, A.H.: A linearly convergent proximal gradient algorithm for decentralized optimization. In: Advances in Neural Information Processing Systems, pp. 2848–2858 (2019)
2. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Mach. Learn. **3**(1), 1–122 (2011)
3. Chang, T.H., Hong, M., Wang, X.: Multi-agent distributed optimization via inexact consensus ADMM. IEEE Trans. Signal Process. **63**(2), 482–497 (2015)
4. Chen, A., Ozdaglar, A.: A fast distributed proximal-gradient method. In: the 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 601–608 (2012)
5. Di Lorenzo, P., Scutari, G.: NEXT: in-network nonconvex optimization. IEEE Trans. Signal Inf. Process. Netw. **2**(2), 120–136 (2016)
6. Glowinski, R., Marroco, A.: Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique **9**(R2), 41–76 (1975)
7. Hong, M., Chang, T.H.: Stochastic proximal gradient consensus over random networks. IEEE Trans. Signal Process. **65**(11), 2933–2948 (2017)
8. Jakovetic, D., Xavier, J., Moura, J.: Fast distributed gradient methods. IEEE Trans. Autom. Control **59**, 1131–1146 (2014)
9. Li, Z., Yan, M.: A primal-dual algorithm with optimal stepsizes and its application in decentralized consensus optimization. arXiv preprint arXiv:1711.06785 (2017)
10. Li, Z., Shi, W., Yan, M.: A decentralized proximal-gradient method with network independent stepsizes and separated convergence rates. IEEE Trans. Signal Process. **67**(17), 4494–4506 (2019)
11. Ling, Q., Ribeiro, A.: Decentralized dynamic optimization through the alternating direction method of multipliers. In: 2013 IEEE 14th Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pp. 170–174. IEEE (2013)
12. Mokhtari, A., Shi, W., Ling, Q., Ribeiro, A.: A decentralized second-order method for dynamic optimization. In: IEEE 55th Conference on Decision and Control (CDC), pp. 6036–6043. IEEE (2016)
13. Nedić, A., Olshevsky, A., Shi, W., Uribe, C.A.: Geometrically convergent distributed optimization with uncoordinated step-sizes. In: American Control Conference (ACC), pp. 3950–3955. IEEE (2017)
14. Nedić, A., Olshevsky, A.: Distributed optimization over time-varying directed graphs. In: The 52nd IEEE Annual Conference on Decision and Control, pp. 6855–6860 (2013)
15. Nedić, A.: Distributed optimization over networks. In: Multi-agent Optimization, pp. 1–84. Springer (2018)
16. Nedic, A.: Asynchronous broadcast-based convex optimization over a network. IEEE Trans. Autom. Control **56**(6), 1337–1351 (2010)
17. Nedić, A., Olshevsky, A.: Stochastic gradient-push for strongly convex functions on time-varying directed graphs. IEEE Trans. Autom. Control **61**(12), 3936–3947 (2016)
18. Nedić, A., Ozdaglar, A.: Distributed subgradient methods for multi-agent optimization. IEEE Trans. Autom. Control **54**, 48–61 (2009)

19. Nedić, A., Olshevsky, A., Shi, W.: Achieving geometric convergence for distributed optimization over time-varying graphs. SIAM J. Optim. **27**(4), 2597–2633 (2017)
20. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Springer, New York (2013)
21. Pu, S., Shi, W., Xu, J., Nedić, A.: A push-pull gradient method for distributed optimization in networks. In: 2018 IEEE Conference on Decision and Control (CDC), pp. 3385–3390. IEEE (2018)
22. Qu, G., Li, N.: Harnessing smoothness to accelerate distributed optimization. In: IEEE 55th Conference on Decision and Control (CDC), pp. 159–166. IEEE (2016)
23. Ram, S.S., Nedić, A., Veeravalli, V.V.: Distributed stochastic subgradient projection algorithms for convex optimization. J. Optim. Theory Appl. **147**(3), 516–545 (2010)
24. Seaman, K., Bach, F., Bubeck, S., Lee, Y.T., Massoulié, L.: Optimal algorithms for smooth and strongly convex distributed optimization in networks. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 3027–3036. JMLR. org (2017)
25. Shi, W., Ling, Q., Yuan, K., Wu, G., Yin, W.: On the linear convergence of the ADMM in decentralized consensus optimization. IEEE Trans. Signal Process. **62**(7), 1750–1761 (2014)
26. Shi, W., Ling, Q., Wu, G., Yin, W.: EXTRA: an exact first-order algorithm for decentralized consensus optimization. SIAM J. Optim. **25**(2), 944–966 (2015)
27. Shi, W., Ling, Q., Wu, G., Yin, W.: A proximal gradient algorithm for decentralized composite optimization. IEEE Trans. Signal Process. **63**(22), 6013–6023 (2015)
28. Sun, Y., Scutari, G., Palomar, D.: Distributed nonconvex multiagent optimization over time-varying networks. In: 2016 50th Asilomar Conference on Signals, Systems and Computers, pp. 788–794. IEEE (2016)
29. Uribe, C.A., Lee, S., Gasnikov, A., Nedić, A.: A dual approach for optimal algorithms in distributed optimization over networks. arXiv preprint arXiv:1809.00710 (2018)
30. Wei, E., Ozdaglar, A.: On the $O(1/k)$ convergence of asynchronous distributed alternating direction method of multipliers. In: Global Conference on Signal and Information Processing (GlobalSIP), pp. 551–554. IEEE (2013)
31. Xi, C., Khan, U.: On the linear convergence of distributed optimization over directed graphs. arXiv preprint arXiv:1510.02149 (2015)
32. Xu, J., Zhu, S., Soh, Y., Xie, L.: Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant step sizes. In: Proceedings of the 54th IEEE Conference on Decision and Control (CDC), pp. 2055–2060 (2015)
33. Yuan, K., Ling, Q., Yin, W.: On the convergence of decentralized gradient descent. SIAM J. Optim. **26**(3), 1835–1854 (2016)
34. Yuan, K., Ying, B., Zhao, X., Sayed, A.H.: Exact diffusion for distributed optimization and learning-part I: algorithm development. IEEE Trans. Signal Process. **67**(3), 708–723 (2018)
35. Yuan, K., Ying, B., Zhao, X., Sayed, A.H.: Exact diffusion for distributed optimization and learning-part II: convergence analysis. IEEE Trans. Signal Process. **67**(3), 724–739 (2018)
36. Zeng, J., Yin, W.: ExtraPush for convex smooth decentralized optimization over directed networks. J. Comput. Math. **35**(4), 381–394 (2017)
37. Zhu, M., Martínez, S.: Discrete-time dynamic average consensus. Automatica **46**(2), 322–329 (2010)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.