Journal of Computational Mathematics
Vol.xx, No.x, 2024, 1–29.

http://www.global-sci.org/jcm
doi:10.4208/jcm.2407-m2023-0282

DECENTRALIZED DOUGLAS-RACHFORD SPLITTING METHODS FOR SMOOTH OPTIMIZATION OVER COMPACT SUBMANIFOLDS*

Kangkang Deng
Department of Mathematics, National University of Defense Technology, Changsha 410005, China
Email: freedeng1208@gmail.com
Jiang Hu[1)]
Department of Mathematics, University of California, Berkeley, CA 94720, USA
Email: hujiangopt@gmail.com
Hongxia Wang
Department of Mathematics, National University of Defense Technology, Changsha 410005, China
Email: wanghongxia@nudt.edu.cn

Abstract

We study decentralized smooth optimization problems over compact submanifolds. Recasting it as a composite optimization problem, we propose a decentralized Douglas-Rachford splitting algorithm (DDRS). When the proximal operator of the local loss function does not have a closed-form solution, an inexact version of DDRS (iDDRS), is also presented. Both algorithms rely on careful integration of the nonconvex Douglas-Rachford splitting algorithm with gradient tracking and manifold optimization. We show that our DDRS and iDDRS achieve the convergence rate of $\mathcal{O}(1/k)$. The main challenge in the proof is how to handle the nonconvexity of the manifold constraint. To address this issue, we utilize the concept of proximal smoothness for compact submanifolds. This ensures that the projection onto the submanifold exhibits convexity-like properties, which allows us to control the consensus error across agents. Numerical experiments on the principal component analysis are conducted to demonstrate the effectiveness of our decentralized DRS compared with the state-of-the-art ones.

Mathematics subject classification: 65N06, 65B99.
Key words: Decentralized optimization, Compact submanifold, Douglas-Rachford splitting, Proximal smoothness, Convergence rate.

1. Introduction

Owing to concerns about privacy and robustness, decentralized optimization over manifolds has garnered significant attention in machine learning, optimization control, and signal processing. Examples include principal component analysis [8,36,49], low-rank matrix completion [5,19,29], and low-dimension subspace learning [19,29]. The problem can be mathematically formulated as follows:

* Received December 17, 2023 / Revised version received April 20, 2024 / Accepted July 8, 2024 / Published online September 23, 2024 /
1) Corresponding author

$$\begin{aligned} &\min_{x_1,\cdots,x_n} \quad \sum_{i=1}^{n} f_i(x_i) \\ &\text{s.t.} \quad x_1 = \cdots = x_n, \quad x_i \in \mathcal{M}, \quad \forall\, i = 1, \ldots, n, \end{aligned} \tag{1.1}$$

where $f_i : \mathbb{R}^{d\times r} \to \mathbb{R}$ is a continuously differentiable function held privately by the i-th agent, and $\mathcal{M}$ is a compact submanifold of $\mathbb{R}^{d\times r}$, e.g. Stiefel manifold, Oblique manifold [1, 4, 21].

While numerous algorithms [3, 18, 26, 37, 41, 42, 45, 51] have been explored for decentralized optimization with nonconvex objective functions, there are only a few papers dealing with the nonconvex constraint. This is an important issue because there is a frequent interest in optimizing nonconvex functions over nonconvex sets, especially compact submanifolds, see, e.g. [8, 19, 29, 36, 49]. Such nonconvex constraint introduces additional challenges in the implementation and analysis of decentralized optimization algorithms. These scenarios often require global solutions for a series of nonconvex constrained optimization problems (e.g. projections to the nonconvex manifold), potentially obstructing the use of conventional tools (e.g. the one-Lipschitz continuity of the projection mapping to the closed convex set) for algorithmic complexity analysis.

The Douglas-Rachford splitting (DRS) is recognized as a famous and efficient splitting algorithm in solving convex and nonconvex optimization problems. The DRS is closely related to the more popular alternating direction method of multipliers (ADMM). The authors in [14] first showed that ADMM is an application of the Douglas-Rachford splitting method (DRSM) to the dual problem when the primal problem is convex. A recent study in [43] shows the primal equivalence between DRS and ADMM in the nonconvex case and demonstrates the convergence of both methods using the Douglas-Rachford envelope. This leads to the following question: Can we design provably convergent decentralized DRS methods for solving (1.1)?

1.1. Our contributions

In this paper, we leverage a novel fusion of gradient tracking and DRS, presenting two decentralized DRS algorithms to solve the decentralized manifold optimization problem (1.1).

- **An easy-to-implement paradigm of decentralized DRS.** By utilizing the decentralized communication graph to construct an inexact projection to the consensus set, we develop a decentralized DRS method (DDRS). To mitigate potential consensus distortions caused by the nonconvexity of the manifold constraints, the communication graph needs to be well-connected (which corresponds to a large enough number of communication rounds, i.e. t, in Algorithm 3.1). Moreover, for cases where the proximal operator of the loss function lacks a closed-form solution, we present an inexact decentralized DRS method (iDDRS), where the inexactness of evaluating the proximal operator associated with the loss function gradually decreases. Numerical results on eigenvalue problems demonstrate the superior efficacy of our algorithm compared with state-of-the-art methods. DDRS and iDDRS are the first splitting algorithms for solving decentralized manifold optimization problems.

- **Harnessing convex-like properties for best-known convergence complexity.** Compared to algorithms for convex constraints, the main challenge in the convergence analysis of our algorithms arises from the nonconvexity of manifold constraints. To address this, we employ a powerful property of the compact submanifold from variational analysis,

called proximal smoothness. With a well-connected communication graph, we ensure that all iterations stay within a small neighborhood of the manifold (see Lemma 4.2). Then, by leveraging the convex-like properties of the projection operator within such neighborhood and the Douglas-Rachford envelope, we establish the global convergence of DDRS with a convergence rate of $\mathcal{O}(1/K)$ (see Theorem 4.1). For the iDDRS, we require that the evaluation errors of the proximal operator of the loss function are summable, thus achieving the same convergence rate as its exact counterpart.

1.2. Related works

Over the past decades, decentralized optimization has attracted increasing interest due to its wide applications. For the Euclidean case (i.e. $\mathcal{M} = \mathbb{R}^{d\times r}$), one seminal approach, the decentralized gradient descent (DGD) method, is explored in [31, 44, 50]. Its limitation in achieving convergence using fixed step sizes spurred further research. To address this issue, algorithms using local historic iterative information, such as EXTRA [38,39], gradient-tracking [30,34,48], and the proximal gradient primal-dual algorithm [18] have been proposed. The connections among these algorithms are studied in [7]. Another popular approach is the decentralized ADMM, see, e.g. [15, 25, 27, 28, 40, 52], which empirically converges faster than the former methods. These studies can only tackle the nonconvexity from f_i and fail to converge if the constraint set is nonconvex. Given the equivalence between ADMM and DRS, one of our motivations is to design fast decentralized DRS algorithms for solving (1.1) with nonconvex submanifold constraint.

For the case where $\mathcal{M}$ is the Stiefel manifold, the authors [8, 9] propose a decentralized Riemannian gradient descent method and its gradient-tracking version [8]. When the local objective function is of negative log-probability type, a decentralized Riemannian natural gradient method is proposed in [19]. To use a single step of consensus, augmented Lagrangian methods [46, 47] are proposed. For the general submanifold setting, through a theoretical study on the regularity condition of the consensus on the manifold, the authors [22] establish the linear consensus results of the projected gradient method and Riemannian gradient method. Based on that, a decentralized projected gradient descent method and its gradient-tracking version are presented in [12]. However, there is no splitting-type method available for addressing the decentralized manifold optimization problem. Inspired by the superior performance of the DRS method over the gradient-type methods in the centralized Euclidean setting [16, 17, 24, 43], we aim to develop decentralized DRS methods for solving (1.1).

1.3. Notation

For any positive integer n, let $[n] = 1, 2, \ldots, n$. Define $J = \mathbf{1}_n\mathbf{1}_n^\top/n$, where $\mathbf{1}_n \in \mathbb{R}^n$ is a vector with all entries set to 1. For a real number a, we use $\lceil a \rceil$ to denote the smallest integer greater than a. For a matrix $x \in \mathbb{R}^{n\times d}$, we use $\|x\|$ to denote its Frobenius norm. For a square matrix and an integer t, W^t denotes the t-th power of W. Let $\mathbf{W} = W \otimes I_d \in \mathbb{R}^{(nd)\times(nd)}$, with $\otimes$ representing the Kronecker product. For the submanifold $\mathcal{M} \subset \mathbb{R}^{d\times r}$, we always set the Euclidean metric as the Riemannian metric. We denote the tangent space and the normal space of $\mathcal{M}$ at a point x as $T_x\mathcal{M}$ and $N_x\mathcal{M}$, respectively.

Given n agents $(x_1, \ldots, x_n)$, where $x_i \in \mathbb{R}^{d\times r}, i \in [n]$, we denote

$$\mathbf{x} = \left(x_1^\top, \ldots, x_n^\top\right)^\top, \quad \hat{\mathbf{x}} = \left(\hat{x}^\top, \ldots, \hat{x}^\top\right)^\top, \quad \bar{\mathbf{x}} := \left(\bar{x}^\top, \ldots, \bar{x}^\top\right)^\top,$$

where $\hat{x}, \bar{x}$ are defined in (2.9) and (2.10), respectively. We define $\|\mathbf{x}\|_{F,\infty} := \max_i \|x_i\|$. We also denote $f(\mathbf{x}) = \sum_{i=1}^n f_i(x_i)$. Its Euclidean gradient is given by

$$\nabla f(\mathbf{x}) = \big(\nabla f_1(x_1)^\top, \ldots, \nabla f_n(x_n)^\top\big)^\top,$$

where $\nabla f_i(x_i)$ denotes the Euclidean gradient of f_i at x_i. If $x_i \in \mathcal{M}, i \in [n]$, we denote the Riemannian gradient of f as

$$\mathrm{grad} f(\mathbf{x}) = \big(\mathrm{grad} f_1(x_1)^\top, \ldots, \mathrm{grad} f_n(x_n)^\top\big)^\top,$$

where $\mathrm{grad} f_i(x_i)$ denotes the Riemannian gradient of f_i at x_i defined in (2.2). We denote the n-fold Cartesian product of $\mathcal{M}$ as $\mathcal{M}^n = \underbrace{\mathcal{M} \times \cdots \times \mathcal{M}}_{n}$.

2. Preliminary

In the context of decentralized optimization, the communication accessibility across the agents is modeled by an undirected connected network graph $G = (\mathcal{V}; \mathcal{E})$ with $|\mathcal{V}| = n$. Let W be the adjacency matrix of the graph. We then make the following standard assumption on W [8, 51].

Assumption 2.1. We assume that the mixing matrix W satisfies the following conditions:

(i) $W_{ij} \geq 0$ for any $i, j \in [n]$ and $W_{ij} = 0$ if and only if $(i, j) \notin \mathcal{E}$.

(ii) $W = W^\top$ and $W\mathbf{1}_n = \mathbf{1}_n$.

(iii) The null space of $(I - W)$ is $\mathrm{span}(\mathbf{1}_n) := \{c\mathbf{1}_n : c \in \mathbb{R}\}$.

It follows from [33] that the second largest singular value of W, denoted as $\sigma_2(W)$, is strictly less than 1. To simplify the notation, we use σ_2 to represent $\sigma_2(W)$.

2.1. Manifold optimization

Manifold optimization has attracted much attention in the past few decades, as evident in works such as [1, 4, 21]. The goal of manifold optimization is to minimize a real-valued function over a manifold, i.e.

$$\min_{x \in \mathcal{M}} h(x), \tag{2.1}$$

where $\mathcal{M}$ is a Riemannian manifold and $h : \mathcal{M} \to \mathbb{R}$ is a real-valued function. The Riemannian gradient $\mathrm{grad} h(x) \in T_x\mathcal{M}$ is the unique tangent vector satisfying

$$\langle \mathrm{grad}\, h(x), \xi \rangle = dh(x)[\xi], \quad \forall\, \xi \in T_x\mathcal{M}. \tag{2.2}$$

If $\mathcal{M}$ is a submanifold embedded in $\mathbb{R}^{d\times r}$ and the function h can be extended to $\mathbb{R}^{d\times r}$, then the Riemannian gradient of h at x can be computed as

$$\mathrm{grad}\, h(x) = \mathcal{P}_{T_x\mathcal{M}}\big(\nabla h(x)\big),$$

where $\mathcal{P}_{T_x\mathcal{M}}$ represents the orthogonal projection onto $T_x\mathcal{M}$. We say x^* is a stationary point of (2.1) if $\mathrm{grad}\, h(x^*) = 0$.

2.2. Proximal smoothness

The notion of proximal smoothness, as introduced by [10], refers to the characteristic of a closed set whereby the nearest-point projection becomes a singleton when the point is close enough to the set. Specifically, for any positive real number γ, we define the γ-tube around $\mathcal{M}$ as

$$U_{\mathcal{M}}(\gamma) := \{x : \mathrm{dist}(x, \mathcal{M}) < \gamma\}.$$

We say a closed set $\mathcal{M}$ is γ-proximally smooth if the projection operator $\mathcal{P}_{\mathcal{M}}(x)$ is a singleton whenever $x \in U_{\mathcal{M}}(\gamma)$. It is worth noting that any compact C^2-submanifold of $\mathbb{R}^{d\times r}$ is a proximally smooth set [2, 10, 11]. For instance, the Stiefel manifold is a set that is 1-proximally smooth. Throughout this paper, we assume that $\mathcal{M}$ is 2γ-proximally smooth. By following the proof in [10, Theorem 4.8], a 2γ-proximally smooth set $\mathcal{M}$ satisfies the following property:

$$\|\mathcal{P}_{\mathcal{M}}(x) - \mathcal{P}_{\mathcal{M}}(y)\| \le 2\|x - y\|, \quad \forall\, x, y \in \bar{U}_{\mathcal{M}}(\gamma), \tag{2.3}$$

where

$$\bar{U}_{\mathcal{M}}(\gamma) := \{x : \mathrm{dist}(x, \mathcal{M}) \le \gamma\}$$

is the closure of $U_{\mathcal{M}}(\gamma)$. Moreover, for any point $x \in \mathcal{M}$ and a normal $v \in N_x\mathcal{M}$, it holds that

$$\langle v, y - x\rangle \le \frac{\|v\|}{4\gamma}\|y - x\|^2, \quad \forall\, y \in \mathcal{M}. \tag{2.4}$$

This is often referred to as the normal inequality [10, 11].

2.3. The Douglas-Rachford splitting method

The DRS is recognized as a famous and efficient splitting algorithm in solving convex and nonconvex optimization problems. Recently, its primal equivalence with the more popular ADMM is established in [43]. Consider the following composite optimization problem:

$$\min_{x\in\mathbb{R}^p} \varphi_1(x) + \varphi_2(x), \tag{2.5}$$

where $\varphi_1, \varphi_2 : \mathbb{R}^p \to \mathbb{R}$ are proper, lower semicontinuous, extended real-valued functions. Starting from some $x_k, s_k, z_k \in \mathbb{R}^p$, one iteration of the classical DRS applied to (2.5) with stepsize α amounts to

$$\begin{cases} s_{k+1} = s_k + z_k - x_k, \\ x_{k+1} = \mathrm{prox}_{\alpha\varphi_1}(s_{k+1}), \\ z_{k+1} = \mathrm{prox}_{\alpha\varphi_2}(2x_{k+1} - s_{k+1}), \end{cases} \tag{2.6}$$

where $\mathrm{prox}_{\alpha\varphi_1}$ is a proximal operator of φ_1 defined by

$$\mathrm{prox}_{\alpha\varphi_1}(x) = \arg\min_{y\in\mathbb{R}^p} \varphi_1(y) + \frac{1}{2\alpha}\|y - x\|^2. \tag{2.7}$$

The authors in [24] present the first general analysis of global convergence of the classical DRS for fully nonconvex problems where one function is Lipschitz differentiable. The authors in [43] consider the relaxed DRS and give a tight convergence result. Their convergence analysis is based on the Douglas-Rachford envelope (DRE), first introduced in [32] for convex problems and generalized to nonconvex cases. In particular, the DRE of (2.5) is defined as

$$\varphi_\alpha^{\mathrm{DR}}(x) := \min_{w\in\mathbb{R}^p} \left\{ \varphi_2(w) + \varphi_1(u) + \langle \nabla\varphi_1(u), w - u\rangle + \frac{1}{2\alpha}\|w - u\|^2 \right\}, \tag{2.8}$$

where $u := \mathrm{prox}_{\alpha\varphi_1}(x)$. They show that the DRE serves as an exact and continuously differentiable merit function for the original problem.

2.4. Stationary point

Let $x_1, \ldots, x_n \in \mathcal{M}$ represent the local copies of x at each agent. We denote $\hat{x}$ as their Euclidean average point, given by

$$\hat{x} := \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2.9}$$

Let $\mathcal{P}_{\mathcal{M}}$ be the orthogonal projection to $\mathcal{M}$. We define $\bar{x}$ is an element in $\mathcal{P}_{\mathcal{M}}(\hat{x})$, i.e.

$$\bar{x} \in \mathrm{argmin}_{y\in\mathcal{M}} \sum_{i=1}^{n} \|y - x_i\|^2 = \mathcal{P}_{\mathcal{M}}(\hat{x}). \tag{2.10}$$

Any element $\bar{x}$ in $\mathcal{P}_{\mathcal{M}}(\hat{x})$ is the induced arithmetic mean of $\{x_i\}_{i=1}^n$ on $\mathcal{M}$ [35]. The ϵ-stationary point of problem (1.1) is defined as follows.

Definition 2.1. *The set of points $\{x_1, x_2, \cdots, x_n\} \subset \mathcal{M}$ is called an ϵ-stationary point of* (1.1) *if there exists an $\bar{x} \in \mathcal{P}_{\mathcal{M}}(\hat{x})$ such that*

$$\|\mathbf{x} - \bar{\mathbf{x}}\|^2 \le \epsilon, \quad \|\mathrm{grad} f(\bar{\mathbf{x}})\|^2 \le \epsilon,$$

where

$$\mathbf{x} = \left(x_1^\top, \ldots, x_n^\top\right)^\top, \quad \bar{\mathbf{x}} = (\bar{x}^\top, \ldots, \bar{x}^\top)^\top.$$

In the following development, we always assure that $\hat{x} \in \bar{U}_{\mathcal{M}}(\gamma)$. Consequently, $\mathcal{P}_{\mathcal{M}}(\hat{x})$ is a singleton and we have $\bar{x} = \mathcal{P}_{\mathcal{M}}(\hat{x})$.

3. Decentralized Douglas-Rachford Splitting Methods

In this section, we will present two decentralized DRS methods for solving (1.1). We first give the notations as follows. Let $x_{i,k}$ denote the i-agent in the k-iteration. Denote

$$\hat{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{i,k},$$

and $\bar{x}_k$ be the projection of $\hat{x}$ onto $\mathcal{M}$. We also denote

$$\mathbf{x}_k = \left(x_{1,k}^\top, \cdots, x_{n,k}^\top\right)^\top, \quad \hat{\mathbf{x}}_k = \left(\hat{x}_k^\top, \cdots, \hat{x}_k^\top\right)^\top, \quad \bar{\mathbf{x}}_k = \left(\bar{x}_k^\top, \cdots, \bar{x}_k^\top\right)^\top.$$

3.1. Decentralized DRS

By Assumption 2.1, the constraint $x_1 = \cdots = x_n$ can be reformulated as $(I_{nd} - \mathbf{W})\mathbf{x} = 0$, where I_{nd} is the $nd \times nd$ identity matrix. Let us define

$$\mathcal{C} := \left\{\mathbf{x} \in \mathcal{M}^n : (I_{nd} - \mathbf{W})\mathbf{x} = 0\right\},$$

and denote $\delta_{\mathcal{C}}$ by the indicator function of $\mathcal{C}$. Then, problem (1.1) can be written as

$$\min_{\mathbf{x}\in\mathbb{R}^{nd\times r}} f(\mathbf{x}) + \delta_{\mathcal{C}}(\mathbf{x}). \tag{3.1}$$

Note that the nearest-point projection of a point $\mathbf{x}$ to $\mathcal{C}$ has an explicit formulation, namely,

$$\mathcal{P}_{\mathcal{C}}(\mathbf{x}) = \underset{\{\mathbf{y}\in\mathcal{M}^n:y_1=\cdots=y_n\}}{\arg\min} \|\mathbf{y}-\mathbf{x}\|^2 = \mathcal{P}_{\mathcal{M}^n}(\hat{\mathbf{x}}) =: \bar{\mathbf{x}}. \tag{3.2}$$

Given $\mathbf{s}_0, \mathbf{z}_0, \mathbf{x}_0 \in \mathcal{M}^n$, at the k-th iterate, a direct application of the DRS method for solving (3.1) has the following update scheme:

$$\begin{cases} \mathbf{s}_{k+1} = \mathbf{s}_k + \mathbf{z}_k - \mathbf{x}_k, \\ \mathbf{x}_{k+1} = \mathrm{prox}_{\alpha f}(\mathbf{s}_{k+1}), \\ \mathbf{y}_{k+1} = 2\mathbf{x}_{k+1} - \mathbf{s}_{k+1}, \\ \mathbf{z}_{k+1} = \mathcal{P}_{\mathcal{C}}(\mathbf{y}_{k+1}). \end{cases} \tag{3.3}$$

Let

$$\hat{y}_k = \frac{1}{n}\sum_{i=1}^n y_{i,k}.$$

The agent-wise version of (3.3) can be written as: For any $i \in [n]$,

$$\begin{cases} s_{i,k+1} = s_{i,k} + z_{i,k} - x_{i,k}, \\ x_{i,k+1} = \mathrm{prox}_{\alpha f_i}(s_{i,k+1}), \\ y_{i,k+1} = 2x_{i,k+1} - s_{i,k+1}, \\ z_{i,k+1} = \mathcal{P}_{\mathcal{M}}(\hat{y}_{k+1}). \end{cases} \tag{3.4}$$

Note that in the update of $z_{i,k+1}$, the i-th agent needs to collect $\{y_{i,k+1}\}_{i=1}^n$. This can be easily achieved in the centralized setting, but could be a critical issue for the decentralized setting where only partial communication along the graph is allowed.

To address the above problem in the z-update, a natural way is to investigate the local average instead of the global average. However, due to the existence of the nonconvexity of $\mathcal{M}$, a careful design to control the approximation error is needed. Before introducing our approaches, let us rewrite the z-update in a form with a more explicit dependence on x. It is easily shown that the update of $x_{i,k+1}$ implies that

$$s_{i,k+1} = x_{i,k+1} + \alpha\nabla f_i(x_{i,k+1}).$$

This together with the update of $y_{i,k+1}$ yields

$$y_{i,k+1} = x_{i,k+1} - \alpha\nabla f_i(x_{i,k+1}), \quad i \in [n]. \tag{3.5}$$

Therefore, the last two rows in (3.4) can be simplified as follows:

$$z_{i,k+1} = \mathcal{P}_{\mathcal{M}}\left(\frac{1}{n}\sum_{j=1}^n \left(x_{j,k+1} - \alpha\nabla f_j(x_{j,k+1})\right)\right). \tag{3.6}$$

Based on the above formulation, and utilizing the adjacency matrix W under Assumption 2.1, we approximate $\sum_{j=1}^n x_{j,k+1}/n$ by $\sum_{j=1}^n W_{ij}^t x_{j,k+1}$, where t is an integer, denoting the communication rounds. To obtain a better performance, we adopt the gradient tracking techniques [8, 30, 34] on the gradient

$$-\frac{\alpha}{n}\sum_{j=1}^n \nabla f_j(x_{j,k+1}) = \frac{1}{n}\sum_{j=1}^n (x_{j,k+1} - s_{j,k+1}),$$

i.e.

$$d_{i,k+1} = \sum_{j=1}^{n} W_{ij}^t d_{i,k} + x_{i,k+1} - s_{i,k+1} - (x_{i,k} - s_{i,k}), \tag{3.7}$$

where $d_{i,0} := x_{0,k} - s_{0,k}$. Then, these approximations give a modified and operational update

$$z_{i,k+1} = \mathcal{P}_{\mathcal{M}}\left(\sum_{j=1}^{n} W_{ij}^t x_{j,k+1} + d_{i,k+1} \right). \tag{3.8}$$

With (3.7) and (3.8), our decentralized DRS method performs the following update in the k-th iteration, for $i = 1, \ldots, n$,

$$\begin{cases} s_{i,k+1} = s_{i,k} + z_{i,k} - x_{i,k}, \\ x_{i,k+1} = \mathrm{prox}_{\alpha f_i}(s_{i,k+1}), \\ d_{i,k+1} = \sum_{j=1}^{n} W_{ij}^t d_{j,k} + x_{i,k+1} - s_{i,k+1} - (x_{i,k} - s_{i,k}), \\ z_{i,k+1} = \mathcal{P}_{\mathcal{M}}\left(\sum_{j=1}^{n} W_{ij}^t x_{j,k+1} + d_{i,k+1} \right). \end{cases} \tag{3.9}$$

The detailed description is given in Algorithm 3.1.

Algorithm 3.1: Decentralized DRS Method for Solving (1.1).

Require: Initial point $\mathbf{s}_0, \mathbf{z}_0 \in \mathcal{M}^n$, an integer t, the step size α.

Let $x_{i,0} = s_{i,0}$, $d_{i,0} = x_{i,0} - s_{i,0}$ and $y_{i,0} = 2x_{i,0} - s_{i,0}$ on each node $i \in [n]$.

for $k = 0, \ldots$ (for each node $i \in [n]$, in parallel) **do**

  Update $s_{i,k+1} = s_{i,k} + z_{i,k} - x_{i,k}$.

  Update $x_{i,k+1} = \mathrm{prox}_{\alpha f_i}(s_{i,k+1})$.

  Update $y_{i,k+1} = 2x_{i,k+1} - s_{i,k+1}$.

  Perform gradient tracking

$$d_{i,k+1} = \sum_{j=1}^{n} W_{ij}^t d_{j,k} + x_{i,k+1} - s_{i,k+1} - (x_{i,k} - s_{i,k}).$$

  Update $z_{i,k+1} = \mathcal{P}_{\mathcal{M}}\left(\sum_{j=1}^{n} W_{ij}^t x_{j,k+1} + d_{i,k+1} \right)$.

end

As will be seen in the next section, we need the communication graph to be well-connected (which corresponds to a sufficiently large integer t) to tackle the nonconvexity from the manifold constraint. Basically speaking, a large t will guarantee that the iterates remain in the proximally smooth neighborhood of $\mathcal{M}$, which allows us to utilize the convex-like properties, (2.3) and (2.4). Moreover, $\{y_{i,k}\}$ is an auxiliary sequence, which is useful for the subsequent analysis.

3.2. Inexact decentralized DRS

Note that in (3.9), the computation of the exact proximal operator, denoted as $\mathrm{prox}_{\alpha f}$, is required. This computation is time-consuming in some cases. Therefore, we investigate the

convergence of the algorithm when $\text{prox}_{\alpha f}$ is computed approximately with a tolerance ϵ_k. In particular, one can find $x_{i,k+1}$ satisfying

$$x_{i,k+1} - s_{i,k+1} + \alpha \nabla f_i(x_{i,k+1}) = \mu_{i,k+1}, \quad i \in [n],$$

where

$$\|\mu_{i,k+1}\|^2 \leq \epsilon_{k+1}, \quad i \in [n].$$

This implies that

$$x_{i,k+1} = \text{prox}_{\alpha f_i}(s_{i,k} + \mu_{i,k}).$$

The detailed iterative process is given in Algorithm 3.2. We emphasize that the introductions of iDDRS aim for a complete analysis of the decentralized Douglas-Rachford splitting-type methods when the evaluation of the proximal mapping is costly. For the principal component analysis problem, its proximal operator can be exactly calculated with a lower cost, as shown in Section 5. Therefore, we do not present the numerical result of iDDRS in this paper. However, we will investigate more appropriate applications to test the iDDRS algorithm in the future.

Algorithm 3.2: Inexact Decentralized DRS Method for Solving (1.1).

Require: Initial point $\mathbf{s}_0, \mathbf{z}_0 \in \mathcal{M}^n$, an integer t, the step size α, the sequence $\{\epsilon_k\}$ $\epsilon_0 < \delta_2$.

1 Let $x_{i,0} = s_{i,0}$, $d_{i,0} = x_{i,0} - s_{i,0}$ and $y_{i,0} = 2x_{i,0} - s_{i,0}$ on each node $i \in [n]$.

2 **for** $k = 0, \ldots$ (for each node $i \in [n]$, in parallel) **do**

3 Update $s_{i,k+1} = s_{i,k} + z_{i,k} - x_{i,k}$.

4 Update $x_{i,k+1} = \text{prox}_{\alpha f_i}(s_{i,k+1} + \mu_{i,k+1})$, $\|\mu_{i,k+1}\|^2 \leq \epsilon_k$.

5 Update $y_{i,k+1} = 2x_{i,k+1} - s_{i,k+1}$.

6 Take gradient tracking

$$d_{i,k+1} = \sum_{j=1}^{n} W_{ij}^t d_{j,k} + x_{i,k+1} - s_{i,k+1} - (x_{i,k} - s_{i,k}).$$

7 Update $z_{i,k+1} = \mathcal{P}_{\mathcal{M}}\left(\sum_{j=1}^{n} W_{ij}^t x_{j,k+1} + d_{i,k+1}\right)$.

8 **end**

4. Convergence Analysis

This section will provide the main convergence result for our algorithms DDRS and iDDRS. We first make the following assumptions.

Assumption 4.1. For any given i, the function f_i is L-smooth, i.e. for any $x, y \in \mathbb{R}^{d\times r}$,

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_f \|x - y\|. \tag{4.1}$$

Using (4.1), we can readily obtain a quadratic upper bound for f_i: For $x, y \in \text{conv}(\mathcal{M})$, it holds that

$$f_i(y) \leq f_i(x) + \langle \nabla f_i(x), y - x\rangle + \frac{L_f}{2}\|y - x\|^2, \quad i \in [n]. \tag{4.2}$$

Moreover, it follows from [12, Lemma 4.2] that there exists $L > L_f$ such that

$$\|\mathrm{grad} f_i(x) - \mathrm{grad} f_i(y)\| \leq L\|x - y\|, \quad i \in [n]. \tag{4.3}$$

Before analyzing our algorithms, we give the following two useful lemmas, which are independent of the algorithm and will be used in subsequent analyses. For an L-smooth h, its proximal operator $\mathrm{prox}_{\alpha h}$ is strongly monotone and cocoercive for small enough α [43, Proposition 2.3].

Proposition 4.1. *Let h be an L-smooth function and $0<\alpha<1/L$. Then, $\mathrm{prox}_{\alpha h}$ is $(1/(1+\alpha L))$-strongly monotone and $(1-\alpha L)$-cocoercive, in the sense that*

$$\begin{aligned} \langle x - x', s - s' \rangle &\geq \frac{1}{1+\alpha L}\|s - s'\|^2, \\ \langle x - x', s - s' \rangle &\geq (1-\alpha L)\|x - x'\|^2 \end{aligned} \tag{4.4}$$

for all s, s', where $x = \mathrm{prox}_{\alpha h}(s)$ and $x' = \mathrm{prox}_{\alpha h}(s')$. In particular,

$$\frac{1}{1+\alpha L}\|s - s'\| \leq \|x - x'\| \leq \frac{1}{1-\alpha L}\|s - s'\|. \tag{4.5}$$

We give the following bound on the distance between $\mathrm{prox}_{\alpha f}(x)$ and x for an L-smooth h and $\alpha < 1/(2L)$.

Lemma 4.1. *Let f be an L-smooth function and $0<\alpha<1/(2L)$. Then, for any $x=\mathrm{prox}_{\alpha f}(s+\mu)$ with $\|\mu\| \leq \epsilon$, it holds that*

$$\|s - x\| \leq \alpha(3\|\nabla f(0)\| + 2L\|s\| + 2\epsilon) + \epsilon. \tag{4.6}$$

When $\epsilon = 0$, it reduced to

$$\|s - x\| \leq \alpha(3\|\nabla f(0)\| + 2L\|s\|). \tag{4.7}$$

Proof. We only prove (4.6). It follows from the definition of $\mathrm{prox}_{\alpha f}$ that

$$x - s + \alpha\nabla f(x) = \mu.$$

Then, by using the L-smoothness of f and the triangle inequality, we have

$$\|x\| = \|s - \alpha\nabla f(x) + \mu\| \leq \|s\| + \epsilon + \alpha(\|\nabla f(0)\| + L\|x\|).$$

This gives

$$\|x\| \leq \frac{\|s\| + \epsilon + \alpha\|\nabla f(0)\|}{1 - \alpha L}.$$

Therefore,

$$\begin{aligned} \|s - x\| = \alpha\|\nabla f(x)\| + \epsilon &\leq \alpha(\|\nabla f(0)\| + L\|x\|) + \epsilon \\ &\leq \alpha\left(\|\nabla f(0)\| + \frac{L\|s\| + \epsilon + \|\nabla f(0)\|}{1 - \alpha L}\right) + \epsilon \\ &\leq \alpha\left(3\|\nabla f(0)\| + 2L\|s\| + 2\epsilon\right) + \epsilon. \end{aligned}$$

The proof is complete. □

4.1. Convergence of DDRS

This subsection will provide the main convergence result of our DDRS (Algorithm 3.1). We first show that the update of $z_{i,k+1}$ is well-defined by ensuring that its projection onto $\mathcal{M}$ results in a singleton. Subsequently, we introduce a descent lemma for our DDRS, drawing from the DRE function as defined in (2.8). Lastly, we establish the convergence rate of $\mathcal{O}(1/k)$ to reach a stationary point.

Due to the proximal smoothness of the manifold constraint, the update of the variable $z_{i,k+1}$ is well-defined only when the term $\sum_{j=1}^n W_{ij}^t x_{j,k+1}+d_{i,k+1}$ lies within the neighborhood $\bar{U}_{\mathcal{M}}(\gamma)$, i.e. for all $k \geq 0$,

$$\sum_{j=1}^n W_{ij}^t x_{j,k+1} + d_{i,k+1} \in \bar{U}_{\mathcal{M}}(\gamma). \tag{4.8}$$

Therefore, before demonstrating the main convergence result of Algorithm 3.1, we will prove that (4.8) holds under some mild conditions. Let us define several constants that will be used in the next analysis, namely,

$$\delta_1 := \frac{\gamma}{4}, \quad \delta_2 := \frac{\delta_1}{12}, \quad \delta_3 := 2\delta_2 + \zeta, \tag{4.9}$$

where γ occurs in (2.3). Let $\mathcal{N}_1, \mathcal{N}_2$ be two neighborhoods defined by

$$\begin{aligned} \mathcal{N}_1 &:= \left\{\mathbf{x} \in \mathbb{R}^{nd\times r} : \|\hat{x} - \bar{x}\| \leq \delta_1\right\}, \\ \mathcal{N}_2 &:= \left\{\mathbf{x} \in \mathbb{R}^{nd\times r} : \|\hat{x} - \bar{x}\| \leq 10\delta_2\right\}, \end{aligned} \tag{4.10}$$

The next lemma demonstrates that under certain conditions on α, t, if $\mathbf{x}_0 \in \mathcal{N}_1$ and $\mathbf{s}_0 \in \mathcal{N}_2$, then for all k, it holds that $\mathbf{x}_k \in \mathcal{N}_1$ and $\sum_{j=1}^n W_{ij}^t x_{i,k} + d_{i,k}$ remains within the neighborhood $\bar{U}_{\mathcal{M}}(\gamma)$. This latter result allows us to invoke the Lipschitz continuity (2.3) of $\mathcal{P}_{\mathcal{M}}$ over $\bar{U}_{\mathcal{M}}(\gamma)$ in the subsequent analysis. We note that for a sufficiently large t, the used communication graph is well-connected, i.e. the second-largest singular value of W^t is small enough. We provide the proof in Section 4.1.1.

Lemma 4.2. *Suppose that Assumptions* 2.1 *and* 4.1 *hold. Let* $\{\mathbf{s}_k, \mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k\}$ *be generated by Algorithm* 3.1 *with*

$$\begin{aligned} 0 < \alpha &\leq \min\left\{\frac{1}{2L}, \frac{\delta_2}{3\|\nabla f(0)\| + 2L\left(\zeta + \delta_2\right)}\right\}, \\ t &\geq \left\lceil \max\left\{\log_{\sigma_2}\left(\frac{1}{4\sqrt{n}}\right), \log_{\sigma_2}\left(\frac{\delta_3}{\delta_2\sqrt{n}}\right)\right\}\right\rceil. \end{aligned}$$

If

$$\|\mathbf{d}_0\|_{F,\infty} \leq 4\delta_2, \quad \|\mathbf{s}_0\|_{F,\infty} \leq \zeta + \delta_2, \quad \mathbf{x}_0 \in \mathcal{N}_1, \quad \mathbf{z}_0 \in \mathcal{N}_2,$$

then it holds that for any integer $k > 0$*,*

$$\mathbf{x}_k \in \mathcal{N}_1, \quad \mathbf{z}_k \in \mathcal{N}_2, \tag{4.11}$$

where $\delta_1, \delta_2, \delta_3$ *are defined in* (4.9) *and* $\mathcal{N}_1, \mathcal{N}_2$ *are defined in* (4.10)*. Moreover, we have that for any integer* $k > 0$*,*

$$\sum_{j=1}^n W_{ij}^t x_{j,k} + d_{i,k} \in \bar{U}_{\mathcal{M}}(\gamma), \quad i \in [n]. \tag{4.12}$$

As shown in [32], the DRE defined in (2.8) can serve as the potential function to analyze the convergence of the DRS method. In particular, given any $\mathbf{s}$, we define $\mathbf{x} = \mathrm{prox}_{\alpha f}(\mathbf{s})$, $\mathbf{y} = \mathbf{x} - \alpha\nabla f(\mathbf{x})$ and $\bar{\mathbf{y}} = \mathcal{P}_{\mathcal{C}}(\mathbf{y})$. The DRE of (3.1) is defined as follows:

$$\begin{aligned}\varphi_\alpha^{\mathrm{DR}}(\mathbf{s}) :&= f(\mathbf{x}) + \min_{\mathbf{w}\in\mathcal{C}}\left\{\langle\nabla f(\mathbf{x}), \mathbf{w}-\mathbf{x}\rangle + \frac{1}{2\alpha}\|\mathbf{w}-\mathbf{x}\|^2\right\}\\ &= f(\mathbf{x}) + \langle\nabla f(\mathbf{x}), \bar{\mathbf{y}}-\mathbf{x}\rangle + \frac{1}{2\alpha}\|\bar{\mathbf{y}}-\mathbf{x}\|^2. \end{aligned} \tag{4.13}$$

We then have the following descent lemma on φ_α.

Lemma 4.3. *Suppose that Assumption* 4.1 *holds. Let* $\{\mathbf{s}_k, \mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k\}$ *be generated by Algorithm* 3.1. *Then, it holds that*

$$\begin{aligned}\varphi_\alpha^{\mathrm{DR}}(\mathbf{s}_0) - \varphi_\alpha^{\mathrm{DR}}(\mathbf{s}_{k+1}) \geq &\sum_{\ell=0}^{k}\frac{1-\alpha L-2\alpha^2L^2}{2\alpha}\|\mathbf{x}_{\ell+1}-\mathbf{x}_\ell\|^2\\ &-\frac{1+\alpha L}{2\alpha}\sum_{\ell=0}^{k}\left(\alpha\|\mathbf{x}_{\ell+1}-\mathbf{x}_\ell\|^2+\frac{1}{\alpha}\|\mathbf{z}_\ell-\bar{\mathbf{y}}_\ell\|^2\right).\end{aligned} \tag{4.14}$$

Proof. It follows from the definition of φ_α in (4.13) and $\bar{\mathbf{y}}_k \in \mathcal{C}$ that

$$\begin{aligned}\varphi_\alpha^{\mathrm{DR}}(\mathbf{s}_k) &\leq f(\mathbf{x}_{k+1}) + \langle\nabla f(\mathbf{x}_{k+1}), \bar{\mathbf{y}}_k-\mathbf{x}_{k+1}\rangle + \frac{1}{2\alpha}\|\bar{\mathbf{y}}_k-\mathbf{x}_{k+1}\|^2\\ &= f(\mathbf{x}_{k+1}) + \langle\nabla f(\mathbf{x}_{k+1}), \mathbf{x}_k-\mathbf{x}_{k+1}\rangle + \langle\nabla f(\mathbf{x}_{k+1}), \bar{\mathbf{y}}_k-\mathbf{x}_k\rangle + \frac{1}{2\alpha}\|\bar{\mathbf{y}}_k-\mathbf{x}_{k+1}\|^2\\ &\leq f(\mathbf{x}_k) + \frac{L}{2}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 + \langle\nabla f(\mathbf{x}_{k+1}), \bar{\mathbf{y}}_k-\mathbf{x}_k\rangle + \frac{1}{2\alpha}\|\bar{\mathbf{y}}_k-\mathbf{x}_{k+1}\|^2,\end{aligned} \tag{4.15}$$

where the second inequality is due to the L_f-smoothness of f_i's and $L > L_f$. Then we have

$$\begin{aligned}\varphi_\alpha^{\mathrm{DR}}(\mathbf{s}_k) \leq\ & f(\mathbf{x}_k) + \frac{L}{2}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 + \langle\nabla f(\mathbf{x}_k), \bar{\mathbf{y}}_k-\mathbf{x}_k\rangle + \frac{1}{2\alpha}\|\bar{\mathbf{y}}_k-\mathbf{x}_{k+1}\|^2\\ &+\langle\nabla f(\mathbf{x}_{k+1})-\nabla f(\mathbf{x}_k), \bar{\mathbf{y}}_k-\mathbf{x}_k\rangle\\ =\ &\varphi_\alpha^{\mathrm{DR}}(\mathbf{s}_{k-1}) + \langle\nabla f(\mathbf{x}_{k+1})-\nabla f(\mathbf{x}_k), \bar{\mathbf{y}}_k-\mathbf{x}_k\rangle + \frac{1+\alpha L}{2\alpha}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2\\ &+\frac{1}{\alpha}\langle\mathbf{x}_{k+1}-\mathbf{x}_k, \mathbf{x}_k-\bar{\mathbf{y}}_k\rangle.\end{aligned} \tag{4.16}$$

By (3.9), we have

$$\begin{aligned}\mathbf{x}_k-\bar{\mathbf{y}}_k &= \mathbf{x}_k-\mathbf{z}_k+\mathbf{z}_k-\bar{\mathbf{y}}_k = \mathbf{s}_k-\mathbf{s}_{k+1}+\mathbf{z}_k-\bar{\mathbf{y}}_k\\ &= \mathbf{x}_k-\mathbf{x}_{k+1}+\alpha\big(\nabla f(\mathbf{x}_k)-\nabla f(\mathbf{x}_{k+1})\big)+\mathbf{z}_k-\bar{\mathbf{y}}_k.\end{aligned} \tag{4.17}$$

Plugging (4.17) into (4.16) gives

$$\begin{aligned}&\varphi_\alpha^{\mathrm{DR}}(\mathbf{s}_{k-1}) - \varphi_\alpha^{\mathrm{DR}}(\mathbf{s}_k)\\ \geq\ &-\alpha\|\nabla f(\mathbf{x}_{k+1})-\nabla f(\mathbf{x}_k)\|^2 + \frac{1}{2\alpha}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2\\ &-\frac{L}{2}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 + \langle\nabla f(\mathbf{x}_{k+1})-\nabla f(\mathbf{x}_k), \mathbf{z}_k-\bar{\mathbf{y}}_k\rangle - \frac{1}{\alpha}\langle\mathbf{x}_{k+1}-\mathbf{x}_k, \mathbf{z}_k-\bar{\mathbf{y}}_k\rangle\\ \geq\ &\frac{1-\alpha L-2\alpha^2L^2}{2\alpha}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 - \left(L+\frac{1}{\alpha}\right)\|\mathbf{x}_{k+1}-\mathbf{x}_k\|\|\mathbf{z}_k-\bar{\mathbf{y}}_k\|.\end{aligned} \tag{4.18}$$

Summing (4.18) in k gives

$$\varphi_\alpha^{\mathrm{DR}}(\mathbf{s}_0) - \varphi_\alpha^{\mathrm{DR}}(\mathbf{s}_{k+1}) \geq \sum_{\ell=0}^{k} \frac{1-\alpha L - 2\alpha^2 L^2}{2\alpha} \|\mathbf{x}_{\ell+1} - \mathbf{x}_\ell\|^2 - \frac{1+\alpha L}{2\alpha} \sum_{\ell=0}^{k} \left(\alpha \|\mathbf{x}_{\ell+1} - \mathbf{x}_\ell\|^2 + \frac{1}{\alpha} \|\mathbf{z}_\ell - \bar{\mathbf{y}}_\ell\|^2 \right),$$

where the inequality is from the Lipschitz continuity of ∇f_i and the Cauchy-Schwarz inequality. The proof is complete. □

Putting the above results together, we establish the following convergence rate of $\mathcal{O}(1/k)$ to reach a stationary point, which matches the best-known result of the decentralized gradient-type methods [8, 12, 46]. We provide the proof in Section 4.1.2.

Theorem 4.1. *Suppose that Assumptions* 2.1 *and* 4.1 *hold. Let* $\{\mathbf{s}_k, \mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k\}$ *be generated by Algorithm* 3.1 *with*

$$0 < \alpha \leq \min\left\{ \frac{1}{2(1+2L+\mathcal{C}_1 L^2)}, \frac{\delta_2}{3\|\nabla f(0)\| + 2L(\zeta+\delta_2)} \right\},$$
$$t \geq \left\lceil \max\left\{ \log_{\sigma_2}\left(\frac{1}{4\sqrt{n}}\right), \log_{\sigma_2}\left(\frac{\delta_3}{\delta_2\sqrt{n}}\right), \log_{\sigma_2}\frac{1}{12\sqrt{n}} \right\} \right\rceil.$$

Let f^* *be the optimal value of* (3.1). *If* $\|\mathbf{d}_0\| \leq 4\delta_2, \|\mathbf{s}_0\|_{F,\infty} \leq \zeta + \delta_2, \mathbf{x}_0 \in \mathcal{N}_1$ *and* $\mathbf{z}_0 \in \mathcal{N}_2$ *for any* $k \in \mathbb{N}$, *it holds that*

$$\min_{0\leq\ell\leq k} \|\mathbf{x}_\ell - \bar{\mathbf{x}}_\ell\| \leq \frac{8\alpha\left(\mathcal{C}_1\alpha^2L^2+4\right)}{k+1}\left(\varphi_\alpha^{\mathrm{DR}}(\mathbf{x}_0,\bar{\mathbf{y}}_0) - f^* + \frac{\mathcal{C}_2}{\alpha^2}\right) + \frac{2\mathcal{C}_2}{k+1}, \tag{4.19}$$

$$\min_{0\leq\ell\leq k} \|\mathrm{grad} f(\bar{\mathbf{x}}_\ell)\| \leq \frac{72\left(\mathcal{C}_1\alpha^2L^2+4\right)}{(k+1)\alpha}\left(\varphi_\alpha^{\mathrm{DR}}(\mathbf{x}_0,\bar{\mathbf{y}}_0) - f^* + \frac{\mathcal{C}_2}{\alpha^2}\right) + \frac{18\mathcal{C}_2}{(k+1)\alpha^2}, \tag{4.20}$$

where

$$\begin{aligned} \mathcal{C}_1 &:= \frac{32}{\left(1-4\sigma_2^t\right)^2}\left(4\sigma_2^t + \frac{4}{\left(1-\sigma_2^t\right)^2}\right), \\ \mathcal{C}_2 &:= \frac{4}{1-16\sigma_2^{2t}}\|\mathbf{z}_0 - \bar{\mathbf{y}}_0\|^2 + \frac{128}{\left(1-4\sigma_2^t\right)^2\left(1-\sigma_2^{2t}\right)}\|\mathbf{d}_0 - (\hat{\mathbf{x}}_0 - \hat{\mathbf{s}}_0)\|^2. \end{aligned} \tag{4.21}$$

Remark 4.1. Compared to the convergence analysis of projected gradient-type methods [12], DDRS relies on the one-step decrease of the Douglas-Rachford envelope instead of the original objective function. As there are three iterative sequences, $\{x_k\}, \{s_k\}$, and $\{z_k\}$, we need a more careful investigation on the consensus analysis, especially in maintaining the neighborhood of the proximal smoothness.

4.1.1. Proof of Lemma 4.2

Before proving Lemma 4.2, we need the following two lemmas, i.e. Lemmas 4.4 and 4.5. Building upon Lemma 4.1, we can establish that in Algorithm 3.1, when $\|\mathbf{s}_0\|_{F,\infty}$ is bounded by a specific constant, it follows that both $\|\mathbf{s}_k\|_{F,\infty}$ and $\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{F,\infty}$ are also bounded, for all $k > 0$. Furthermore, one can shows that $\|\nabla f(x_{i,k})\|$ and $\|d_{i,k}\|$ are bounded.

Lemma 4.4. *Suppose that Assumptions* 2.1 *and* 4.1 *hold. Let* $\{\mathbf{s}_k, \mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k\}$ *be generated by Algorithm* 3.1. *Let* δ_2 *be defined in* (4.9). *If*

$$0 < \alpha \leq \min\left\{\frac{1}{2L}, \frac{\delta_2}{3\|\nabla f(0)\| + 2L\left(\zeta + \delta_2\right)}\right\}, \quad t \geq \left\lceil \log_{\sigma_2}\left(\frac{1}{4\sqrt{n}}\right)\right\rceil,$$

$\|\mathbf{d}_0\|_{F,\infty} \leq 4\delta_2$ *and* $\|\mathbf{s}_0\|_{F,\infty} \leq \zeta + \delta_2$, *then it holds that for any* k,

$$\|\mathbf{s}_k\|_{F,\infty} \leq \zeta + \delta_2, \qquad \|\mathbf{x}_k - \mathbf{s}_k\|_{F,\infty} \leq \delta_2, \tag{4.22}$$

$$\|\nabla f(x_{i,k})\| \leq \delta_2/\alpha, \quad i \in [n], \tag{4.23}$$

$$\|d_{i,k}\| \leq 4\delta_2, \qquad i \in [n]. \tag{4.24}$$

Proof. Firstly, we prove (4.22) by induction. Note that $\|\mathbf{s}_0\|_{F,\infty} \leq \zeta + \delta_2$ and by Lemma 4.1,

$$\begin{aligned}\|x_{i,0} - s_{i,0}\| &\leq \alpha\left(\|\nabla f(0)\| + \frac{L\|s_{i,0}\| + \|\nabla f(0)\|}{1 - \alpha L}\right)\\ &\leq \alpha\big(3\|\nabla f(0)\| + 2L(\zeta + \delta_2)\big) \leq \delta_2,\end{aligned} \tag{4.25}$$

which implies that $\|\mathbf{x}_0 - \mathbf{s}_0\|_{F,\infty} \leq \delta_2$. Suppose for some $k \geq 0$ that $\|\mathbf{s}_k\|_{F,\infty} \leq \zeta + \delta_2$ and $\|\mathbf{x}_k - \mathbf{s}_k\|_{F,\infty} \leq \delta_2$. Then we have

$$\|\mathbf{s}_{k+1}\|_{F,\infty} \leq \|\mathbf{z}_k\|_{F,\infty} + \|\mathbf{x}_k - \mathbf{s}_k\|_{F,\infty} \leq \zeta + \delta_2.$$

Similar with (4.25), we further have that $\|x_{i,k+1} - s_{i,k+1}\| \leq \delta_2$, for any $i \in [n]$, which implies $\|\mathbf{x}_k - \mathbf{s}_k\|_{F,\infty} \leq \delta_2$. Secondly, (4.23) follows from the fact that $x_{i,k} = s_{i,k} - \alpha\nabla f(x_{i,k})$. Finally, we prove (4.24) by induction. Suppose that (4.24) holds for some $k \geq 0$. It follows from (3.7) and $d_{i,0} = x_{i,0} - s_{i,0}$ that

$$\hat{d}_k = \hat{x}_k - \hat{s}_k = \alpha\hat{g}_k, \tag{4.26}$$

where

$$\hat{g}_k := \frac{1}{n}\sum_{i=1}^{n}\nabla f(x_{i,k}).$$

Then we have that

$$\begin{aligned}\|d_{i,k+1} - \alpha\hat{g}_k\| &= \left\|\sum_{j=1}^{n} W_{ij}^t d_{j,k} - \alpha\hat{g}_k + \alpha\nabla f_i(x_{i,k+1}) - \alpha\nabla f_i(x_{i,k})\right\|\\ &\overset{(4.26)}{\leq} \left\|\sum_{j=1}^{n}\left(W_{ij}^t - \frac{1}{n}\right)d_{j,k}\right\| + \alpha\|\nabla f_i(x_{i,k+1}) - \nabla f_i(x_{i,k})\|\\ &\leq \sigma_2^t\sqrt{n}\max_i\|d_{i,k}\| + 2\delta_2\\ &\leq \frac{1}{4}\max_i\|d_{i,k}\| + 2\delta_2 \leq 3\delta_2,\end{aligned} \tag{4.27}$$

where the second inequality follows from (4.23) and the bound on the total variation distance between any row of W^t and $\mathbf{1}/n$ [6, 13], i.e.

$$\max_i\sum_{j=1}^{n}\left|W_{i,j}^t - \frac{1}{n}\right| \leq \sqrt{n}\sigma_2^t. \tag{4.28}$$

Hence,

$$\begin{aligned}\|d_{i,k+1}\| &\leq \|d_{i,k+1} - \alpha\hat{g}_k\| + \|\alpha\hat{g}_k\| \\ &\leq 3\delta_2 + \alpha \max_i \|\nabla f(x_{i,k})\| \overset{(4.23)}{\leq} 3\delta_2 + \delta_2 \leq 4\delta_2. \end{aligned} \tag{4.29}$$

The proof is complete. □

The following lemma shows that when $\mathbf{x}_k \in \mathcal{N}_1$, the term $\sum_{j=1}^n W_{ij}^t x_{j,k} + d_{i,k}$ will lie in the neighborhood $\bar{U}_{\mathcal{M}}(\gamma)$.

Lemma 4.5. *Suppose that Assumptions* 2.1 *and* 4.1 *hold. Let* $\{\mathbf{s}_k, \mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k\}$ *be generated by Algorithm* 3.1*. Under the same condition on* $\alpha, \mathbf{d}_0$ *and* $\mathbf{s}_0$ *as in Lemma* 4.4*, if*

$$t \geq \left\lceil \max\left\{ \log_{\sigma_2}\left(\frac{1}{4\sqrt{n}}\right), \log_{\sigma_2}\left(\frac{\delta_3}{\delta_2\sqrt{n}}\right)\right\}\right\rceil,$$

and $\mathbf{x}_k \in \mathcal{N}_1$*, then it holds that*

$$\sum_{j=1}^n W_{ij}^t x_{j,k} + d_{i,k} \in \bar{U}_{\mathcal{M}}(\gamma), \quad i \in [n], \tag{4.30}$$

where $\delta_1, \delta_2, \delta_3$ *are defined in* (4.9) *and* $\mathcal{N}_1$ *is defined in* (4.10)*.*

Proof. It follows from the updated rule of $x_{i,k+1}$ and $s_{i,k+1}$ in Algorithm 3.1 that

$$\begin{aligned} x_{i,k+1} &= x_{i,k+1} - s_{i,k+1} + s_{i,k+1} \\ &= x_{i,k+1} - s_{i,k+1} + s_{i,k} - x_{i,k} + z_{i,k} \\ &= z_{i,k} + \alpha\nabla f(x_{i,k}) - \alpha\nabla f(x_{i,k+1}). \end{aligned} \tag{4.31}$$

Then we have that

$$\begin{aligned}\|x_{i,k+1}\| &\leq \|x_{i,k+1} - z_{i,k}\| + \|z_{i,k}\| \\ &\leq \alpha\|\nabla f_i(x_{i,k+1}) - \nabla f_i(x_{i,k})\| + \|z_{i,k}\| \\ &\leq 2\delta_2 + \zeta = \delta_3, \end{aligned} \tag{4.32}$$

where the last inequality follows from (4.23) and Assumption 4.1. It follows that for any $i \in [n]$,

$$\begin{aligned}\left\|\sum_{j=1}^n W_{ij}^t x_{j,k} + d_{i,k} - \bar{x}_k\right\| &\leq \left\|\sum_{j=1}^n W_{ij}^t x_{j,k} + d_{i,k} - \hat{x}_k\right\| + \|\hat{x}_k - \bar{x}_k\| \\ &\leq \left\|\sum_{j=1}^n (W_{ij}^t - \frac{1}{n})x_{j,k}\right\| + \|d_{i,k}\| + \|\hat{x}_k - \bar{x}_k\| \\ &\leq \sum_{j=1}^n \left|W_{ij}^t - \frac{1}{n}\right| \max_j \|x_{j,k}\| + 4\delta_2 + \delta_1 \\ &\leq \sqrt{n}\sigma_2^t\delta_3 + 4\delta_2 + \delta_1 \\ &\leq 5\delta_2 + \delta_1 \leq \frac{17}{12}\delta_1 \leq \gamma, \end{aligned}$$

where the fourth inequality follows from (4.28). Combining the fact that $\bar{x}_k \in \mathcal{M}$, we obtain (4.30). The proof is complete. □

Proof of Lemma 4.2. We prove it by induction on both $\|\hat{z}_k - \bar{z}_k\|$ and $\|\hat{x}_k - \bar{x}_k\|$. Suppose for some $k \geq 0$ such that (4.11) holds. Then we have

$$\begin{aligned}\|\hat{x}_{k+1} - \bar{x}_{k+1}\| &\leq \|\hat{x}_{k+1} - \bar{z}_k\| \leq \|\hat{x}_{k+1} - \hat{z}_k\| + \|\hat{z}_k - \bar{z}_k\| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\|x_{i,k+1} - z_{i,k}\| + 10\delta_2 \\ &\overset{(4.31)}{\leq} \max_i \alpha\|\nabla f_i(x_{i,k+1}) - \nabla f_i(x_{i,k})\| + 10\delta_2 \\ &\overset{(4.23)}{\leq} 2\delta_2 + 10\delta_2 \leq \delta_1, \end{aligned} \tag{4.33}$$

where the first inequality use $\bar{x}_k = \mathcal{P}_{\mathcal{M}}(\hat{x}_k)$ and $\bar{z}_k \in \mathcal{M}$. In addition,

$$\begin{aligned}\|\hat{z}_{k+1} - \bar{z}_{k+1}\| &\leq \|\hat{z}_{k+1} - \bar{x}\| \leq \frac{1}{n}\sum_{i=1}^{n}\|z_{i,k+1} - \bar{x}\| \\ &\leq \frac{2}{n}\sum_{i=1}^{n}\left\|\sum_{j=1}^{n}W_{ij}^t x_{j,k} + d_{i,k} - \hat{x}\right\| \\ &\leq 2\big(\sqrt{n}\sigma_2^t\delta_3 + 4\delta_2\big) \leq 10\delta_2. \end{aligned} \tag{4.34}$$

Finally, (4.12) follows from Lemma 4.5. We completed the proof. □

4.1.2. Proof of Theorem 4.1

Before proving Lemma 4.2, we need the following two lemmas, i.e. Lemmas 4.10 and 4.11. The following lemma shows the discrepancy between $\mathbf{z}_k$ and $\bar{\mathbf{y}}_k$.

Lemma 4.6. *Suppose that Assumptions* 2.1 *and* 4.1 *hold. Under the conditions as same as in Lemma* 4.2*, it holds that*

$$\|\mathbf{d}_{k+1} - (\hat{\mathbf{x}}_{k+1} - \hat{\mathbf{s}}_{k+1})\| \leq \sigma_2^t\|\mathbf{d}_k - (\hat{\mathbf{x}}_k - \hat{\mathbf{s}}_k)\| + \alpha L\|\mathbf{x}_{k+1} - \mathbf{x}_k\|, \tag{4.35}$$

$$\|\mathbf{z}_{k+1} - \bar{\mathbf{y}}_{k+1}\| \leq 4\sigma_2^t(\|\mathbf{z}_k - \bar{\mathbf{y}}_k\| + \alpha L\|\mathbf{x}_{k+1} - \mathbf{x}_k\|) + 2\|\mathbf{d}_{k+1} - (\hat{\mathbf{x}}_{k+1} - \hat{\mathbf{s}}_{k+1})\|. \tag{4.36}$$

Proof. Denote $q_{i,k} := x_{i,k} - s_{i,k} = \alpha\nabla f(x_{i,k})$. Since $\hat{\mathbf{d}}_{k+1} = \hat{\mathbf{x}}_{k+1} - \hat{\mathbf{s}}_{k+1}$, we have

$$\begin{aligned}\|\mathbf{d}_{k+1} - (\hat{\mathbf{x}}_{k+1} - \hat{\mathbf{s}}_{k+1})\| &\leq \|\mathbf{d}_{k+1} - (\hat{\mathbf{x}}_k - \hat{\mathbf{s}}_k)\| \\ &\leq \|\mathbf{W}^t\mathbf{d}_k - (\hat{\mathbf{x}}_k - \hat{\mathbf{s}}_k)\| + \|\mathbf{q}_{k+1} - \mathbf{q}_k\| \\ &\leq \sigma_2^t\|\mathbf{d}_k - (\hat{\mathbf{x}}_k - \hat{\mathbf{s}}_k)\| + \alpha L\|\mathbf{x}_{k+1} - \mathbf{x}_k\|. \end{aligned}$$

Note that

$$\begin{aligned}\|\hat{y}_{k+1} - \bar{y}_{k+1}\| &\leq \|\hat{y}_{k+1} - \bar{z}_k\| \leq \|\hat{y}_{k+1} - \hat{z}_k\| + \|\hat{z}_k - \bar{z}_k\| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\|x_{i,k+1} - s_{i,k+1} + x_{i,k+1} - z_{i,k}\| + \|\hat{z}_k - \bar{z}_k\| \\ &\leq \delta_2 + 2\delta_2 + 10\delta_2 = 13\delta_2, \end{aligned} \tag{4.37}$$

which means $\|\hat{y}_{k+1} - \bar{y}_{k+1}\| \leq \gamma$. Note that

$$\|\hat{\mathbf{z}}_k - \bar{\mathbf{y}}_k\|^2 = n\left\|\frac{1}{n}\sum_{i=1}^{n}(z_{i,k} - \bar{y}_k)\right\|^2 \leq n \cdot \frac{1}{n^2} \cdot n\|\mathbf{z}_k - \bar{\mathbf{y}}_k\|^2 \leq \|\mathbf{z}_k - \bar{\mathbf{y}}_k\|^2. \tag{4.38}$$

Then, we have

$$\begin{aligned}\|\mathbf{z}_{k+1}-\bar{\mathbf{y}}_{k+1}\| \leq\ & 2\|\mathbf{W}^t\mathbf{x}_{k+1}+\mathbf{d}_{k+1}-\hat{\mathbf{y}}_{k+1}\| \\ \leq\ & 2\sigma_2^t\|\mathbf{x}_{k+1}-\hat{\mathbf{x}}_{k+1}\|+2\|\mathbf{d}_{k+1}-(\hat{\mathbf{x}}_{k+1}-\hat{\mathbf{s}}_{k+1})\| \\ \leq\ & 2\sigma_2^t(\|\mathbf{z}_k-\hat{\mathbf{z}}_k\|+\|\mathbf{q}_{k+1}-\mathbf{q}_k\|+\|\hat{\mathbf{q}}_{k+1}-\hat{\mathbf{q}}_k\|) \\ & +2\|\mathbf{d}_{k+1}-(\hat{\mathbf{x}}_{k+1}-\hat{\mathbf{s}}_{k+1})\| \\ \leq\ & 4\sigma_2^t(\|\mathbf{z}_k-\bar{\mathbf{y}}_k\|+\alpha L\|\mathbf{x}_{k+1}-\mathbf{x}_k\|) \\ & +2\|\mathbf{d}_{k+1}-(\hat{\mathbf{x}}_{k+1}-\hat{\mathbf{s}}_{k+1})\|,\end{aligned}$$

where the first inequality is from 2-Lipschitz of $\mathcal{P}_{\mathcal{M}}$ over $\bar{U}_{\mathcal{M}}(\gamma)$, the third inequality is due to $\mathbf{x}_{k+1}=\mathbf{z}_k+\mathbf{q}_{k+1}-\mathbf{q}_k$ from (4.31), and the last inequality follows from (4.38) and the Lipschitz continuity of f_i. We complete the proof. □

With the recursion inequalities (4.35) and (4.36), we can bound $\sum_{\ell=0}^k\|\mathbf{z}_\ell-\bar{\mathbf{y}}_\ell\|^2$ by $\sum_{\ell=0}^k\|\mathbf{x}_{\ell+1}-\mathbf{x}_\ell\|^2$.

Lemma 4.7. *Suppose that Assumption* 2.1, 4.1 *and the conditions in Lemma* 4.2 *hold. Then it holds that*

$$\sum_{\ell=0}^k\|\mathbf{z}_\ell-\bar{\mathbf{y}}_\ell\|^2\leq\mathcal{C}_1\alpha^2L^2\sum_{\ell=0}^k\|\mathbf{x}_{\ell+1}-\mathbf{x}_\ell\|^2+\mathcal{C}_2, \tag{4.39}$$

where $\mathcal{C}_1,\mathcal{C}_2$ *are defined by Theorem* 4.1.

Proof. Note that $\sigma_2^t<4\sigma_2^t\leq 1/\sqrt{n}\leq 1$. Applying [48, Lemma 2] to (4.35) and (4.36) gives

$$\sum_{\ell=0}^k\|\mathbf{d}_k-(\hat{\mathbf{x}}_k-\hat{\mathbf{s}}_k)\|^2\leq\frac{4\alpha^2L^2}{\left(1-\sigma_2^t\right)^2}\sum_{\ell=0}^k\|\mathbf{x}_{\ell+1}-\mathbf{x}_\ell\|^2+\frac{4}{1-\sigma_2^{2t}}\|\mathbf{d}_0-(\hat{\mathbf{x}}_0-\hat{\mathbf{s}}_0)\|^2. \tag{4.40}$$

Similarly, we have that

$$\begin{aligned}\sum_{\ell=0}^k\|\mathbf{z}_\ell-\bar{\mathbf{y}}_\ell\|^2\leq\ & \frac{32}{\left(1-4\sigma_2^t\right)^2}\sum_{\ell=0}^k\left(4\sigma_2^{2t}\alpha^2L^2\|\mathbf{x}_{\ell+1}-\mathbf{x}_\ell\|^2+\|\mathbf{d}_\ell-(\hat{\mathbf{x}}_\ell-\hat{\mathbf{s}}_\ell)\|^2\right) \\ & +\frac{4}{1-16\sigma_2^{2t}}\|\mathbf{z}_0-\bar{\mathbf{y}}_0\|^2 \\ \leq\ & \frac{32}{\left(1-4\sigma_2^t\right)^2}\sum_{\ell=0}^k\left(4\sigma_2^{2t}\alpha^2L^2+\frac{4\alpha^2L^2}{\left(1-\sigma_2^t\right)^2}\right)\|\mathbf{x}_{\ell+1}-\mathbf{x}_\ell\|^2 \\ & +\frac{128}{\left(1-4\sigma_2^t\right)^2\left(1-\sigma_2^{2t}\right)}\|\mathbf{d}_0-(\hat{\mathbf{x}}_0-\hat{\mathbf{s}}_0)\|^2 \\ & +\frac{4}{1-16\sigma_2^{2t}}\|\mathbf{z}_0-\bar{\mathbf{y}}_0\|^2.\end{aligned} \tag{4.41}$$

The proof is complete. □

Now we give the proof of Theorem 4.1.

Proof of Theorem 4.1. It follows from (4.14) and (4.39) that

$$\begin{aligned}\varphi_\alpha^{\mathrm{DR}}(\mathbf{s}_0)-\varphi_\alpha^{\mathrm{DR}}(\mathbf{s}_{k+1})\geq\ & \frac{(1+\alpha L)\left(1-\alpha-2\alpha L-\mathcal{C}_1\alpha L^2\right)}{2\alpha}\sum_{\ell=0}^k\|\mathbf{x}_{\ell+1}-\mathbf{x}_\ell\|^2-\frac{\mathcal{C}_2(1+\alpha L)}{2\alpha^2} \\ \geq\ & \frac{1}{4\alpha}\sum_{\ell=0}^k\|\mathbf{x}_{\ell+1}-\mathbf{x}_\ell\|^2-\frac{\mathcal{C}_2(1+\alpha L)}{2\alpha^2},\end{aligned} \tag{4.42}$$

where the first inequality comes from Lemma 4.7, and the second inequality is due to the assumption on α. By Assumption 4.1, we have from [43, Theorem 3.4] that

$$\varphi_\alpha^{\mathrm{DR}}(\mathbf{x}_k, \bar{\mathbf{y}}_k) \geq \inf_{\mathbf{x}} \{f(\mathbf{x}) + \delta_{\mathcal{C}}(\mathbf{x})\} = f^* > -\infty.$$

Then, it follows from (4.42) that

$$\frac{1}{k+1}\sum_{\ell=0}^{k} \|\mathbf{x}_{\ell+1} - \mathbf{x}_\ell\|^2 \leq \frac{4\alpha}{k+1}\left(\varphi_\alpha^{\mathrm{DR}}(\mathbf{x}_0, \bar{\mathbf{y}}_0) - f^* + \frac{\mathcal{C}_2}{\alpha^2}\right). \tag{4.43}$$

It follows from the definition of $\bar{\mathbf{y}}_k$ that

$$\bar{\mathbf{y}}_k - \big(\hat{\mathbf{x}}_k - \alpha\nabla f(\hat{\mathbf{x}}_k)\big) \in N_{\bar{\mathbf{y}}_k}\mathcal{M}^n.$$

This further implies that

$$\begin{aligned}
\|\mathrm{grad} f(\bar{\mathbf{y}}_k)\| &= \mathrm{dist}\big(\nabla f(\bar{\mathbf{y}}_k), N_{\bar{\mathbf{y}}_k}\mathcal{M}^n\big) \\
&\leq \left\|\nabla f(\bar{\mathbf{y}}_k) - \frac{\bar{\mathbf{y}}_k - \hat{\mathbf{x}}_k}{\alpha} + \nabla f(\hat{\mathbf{x}}_k)\right\| \\
&\leq \frac{2}{\alpha}\|\bar{\mathbf{y}}_k - \hat{\mathbf{x}}_k\| \leq \frac{2}{\alpha}\|\bar{\mathbf{y}}_k - \mathbf{x}_k\|.
\end{aligned} \tag{4.44}$$

By combining with the fact that $\|\bar{\mathbf{y}}_k - \hat{\mathbf{x}}_k\| \leq \|\bar{\mathbf{y}}_k - \mathbf{x}_k\|$, we have that

$$\begin{aligned}
\|\mathrm{grad} f(\bar{\mathbf{x}}_k)\| &\leq \|\mathrm{grad} f(\bar{\mathbf{y}}_k)\| + \|\mathrm{grad} f(\bar{\mathbf{y}}_k) - \mathrm{grad} f(\bar{\mathbf{x}}_k)\| \\
&\leq \frac{2}{\alpha}\|\bar{\mathbf{y}}_k - \mathbf{x}_k\| + L\|\bar{\mathbf{y}}_k - \bar{\mathbf{x}}_k\| \\
&\leq \frac{2}{\alpha}\|\bar{\mathbf{y}}_k - \mathbf{x}_k\| + 2L\|\bar{\mathbf{y}}_k - \hat{\mathbf{x}}_k\| \\
&\leq \frac{2+2\alpha L}{\alpha}\|\bar{\mathbf{y}}_k - \mathbf{x}_k\| \leq \frac{3}{\alpha}\|\bar{\mathbf{y}}_k - \mathbf{x}_k\|,
\end{aligned} \tag{4.45}$$

where the third inequality utilizes that

$$\begin{aligned}
\|\bar{\mathbf{y}}_k - \bar{\mathbf{x}}_k\| &\leq \|\bar{\mathbf{y}}_k - \hat{\mathbf{x}}_k\| + \|\hat{\mathbf{x}}_k - \bar{\mathbf{x}}_k\| \\
&\leq \|\bar{\mathbf{y}}_k - \hat{\mathbf{x}}_k\| + \|\bar{\mathbf{y}}_k - \hat{\mathbf{x}}_k\| \leq 2\|\bar{\mathbf{y}}_k - \hat{\mathbf{x}}_k\|.
\end{aligned}$$

By the triangle inequality and Proposition 4.1, we have

$$\begin{aligned}
\|\bar{\mathbf{x}}_k - \mathbf{x}_k\| &\leq \|\bar{\mathbf{y}}_k - \mathbf{x}_k\| \\
&\leq \|\bar{\mathbf{y}}_k - \mathbf{z}_k\| + \|\mathbf{z}_k - \mathbf{x}_k\| \\
&= \|\bar{\mathbf{y}}_k - \mathbf{z}_k\| + \|\mathbf{s}_{k+1} - \mathbf{s}_k\| \\
&\leq \|\bar{\mathbf{y}}_k - \mathbf{z}_k\| + 2\|\mathbf{x}_{k+1} - \mathbf{x}_k\|,
\end{aligned} \tag{4.46}$$

where the first inequality uses the fact that $\bar{\mathbf{x}}_k$ belongs to the projection of $\mathbf{x}_k$ onto $\mathcal{M}^n$, the last inequality follows from (4.5) and $\alpha < 1/(2L)$. Then, it holds that

$$\begin{aligned}
\frac{1}{k+1}\sum_{\ell=0}^{k}\|\bar{\mathbf{x}}_\ell - \mathbf{x}_\ell\|^2 &\leq \frac{1}{k+1}\sum_{\ell=0}^{k}(\|\bar{\mathbf{y}}_\ell - \mathbf{z}_\ell\| + 2\|\mathbf{x}_{\ell+1} - \mathbf{x}_\ell\|)^2 \\
&\leq \frac{2}{k+1}\cdot\left[\big(\mathcal{C}_1\alpha^2L^2 + 4\big)\sum_{\ell=0}^{k}\|\mathbf{x}_{\ell+1} - \mathbf{x}_\ell\|^2 + \mathcal{C}_2\right].
\end{aligned} \tag{4.47}$$

Combining (4.45) and (4.47) gives

$$\frac{1}{k+1}\sum_{\ell=0}^{k}\|\mathrm{grad} f(\bar{\mathbf{x}}_\ell)\|^2 \le \frac{9}{\alpha^2}\frac{1}{k+1}\sum_{\ell=0}^{k}\|\bar{\mathbf{y}}_\ell-\mathbf{x}_\ell\|^2 \\ \overset{(4.39)}{\le} \frac{18\mathcal{C}_1\alpha^2L^2+72}{(k+1)\alpha^2}\sum_{\ell=0}^{k}\|\mathbf{x}_{\ell+1}-\mathbf{x}_\ell\|^2+\frac{18\mathcal{C}_2}{(k+1)\alpha^2}. \tag{4.48}$$

Putting (4.43), (4.47), and (4.48) together gives (4.19) and (4.20). □

4.2. Convergence of iDDRS

This subsection will provide the main convergence result of our iDDRS (Algorithm 3.2). Here, we shall consider the following assumption on the inexactness of evaluating the proximal mapping of f_i.

Assumption 4.2. $\{\epsilon_k\}_{k\in\mathbb{N}}$ is summable, i.e. $\sum_k \epsilon_k \le \mathcal{D} < \infty$ for some constant $\mathcal{D}$.

For the ease of analysis, we assume $\epsilon_0 < \delta_2$, where δ_2 is defined by (4.9). Similar to Lemma 4.2, we have the following result. We provide the proof in Section 4.2.1.

Lemma 4.8. *Suppose that Assumptions* 2.1, 4.1, 4.2 *hold. Let* $\{\mathbf{s}_k,\mathbf{x}_k,\mathbf{y}_k,\mathbf{z}_k\}$ *be generated by Algorithm* 3.2 *with*

$$0<\alpha\le\min\left\{\frac{1}{2L},\frac{\delta_2-\epsilon_0}{3\|\nabla f(0)\|+2L\left(\zeta+\delta_2\right)+2\epsilon_0}\right\},$$
$$t\ge\left\lceil\max\left\{\log_{\sigma_2}\left(\frac{1}{4\sqrt{n}}\right),\log_{\sigma_2}\left(\frac{\delta_3}{\delta_2\sqrt{n}}\right)\right\}\right\rceil.$$

If $\mathbf{x}_0\in\mathcal{N}_1$ *and* $\mathbf{z}_0\in\mathcal{N}_2$, *then it holds that for any integer* $k>0$,

$$\mathbf{x}_k\in\mathcal{N}_1,\quad \mathbf{z}_k\in\mathcal{N}_2, \tag{4.49}$$

where $\delta_1,\delta_2,\delta_3$ *are defined in* (4.9) *and* $\mathcal{N}_1,\mathcal{N}_2$ *are defined in* (4.10). *Moreover, we have that for any integer* $k>0$,

$$\sum_{j=1}^{n}W_{ij}^t x_{j,k}+d_{i,k}\in\bar{U}_{\mathcal{M}}(\gamma),\quad i\in[n]. \tag{4.50}$$

Denote $\mu_k=(\mu_{1,k}^\top,\cdots,\mu_{n,k}^\top)^\top$. The following lemma is similar to Lemma 4.3, we omit the proof.

Lemma 4.9. *Suppose that Assumption* 4.1 *holds. Let* $\{\mathbf{s}_k,\mathbf{x}_k,\mathbf{y}_k,\mathbf{z}_k\}$ *be generated by Algorithm* 3.2. *Then,*

$$\varphi_\alpha^{\mathrm{DR}}(\mathbf{s}_0)-\varphi_\alpha^{\mathrm{DR}}(\mathbf{s}_{k+1})\ge\sum_{\ell=0}^{k}\frac{1-\alpha L-2\alpha^2L^2}{2\alpha}\|\mathbf{x}_{\ell+1}-\mathbf{x}_\ell\|^2 \\ -\frac{1+\alpha L}{2\alpha}\sum_{\ell=0}^{k}\left(\alpha\|\mathbf{x}_{\ell+1}-\mathbf{x}_\ell\|^2+\frac{2}{\alpha}\left(\|\mathbf{z}_l-\bar{\mathbf{y}}_l\|^2+2\sqrt{n}\epsilon_\ell\right)\right). \tag{4.51}$$

The following theorem shows that Algorithm 3.2 achieves the convergence rate of $\mathcal{O}(1/K)$. We provide the proof in Section 4.2.2.

Theorem 4.2. *Suppose that Assumption* 2.1, 4.1, 4.2 *hold. Let* $\{\mathbf{s}_k, \mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k\}$ *be generated by Algorithm* 3.2 *with*

$$\alpha \le \min\left\{\frac{1}{2\big(1+2L+2\mathcal{C}_3L^2\big)}, \frac{\delta_2-\epsilon_0}{3\|\nabla f(0)\|+2L\left(\zeta+\delta_2\right)+2\epsilon_0}\right\},$$
$$t \ge \left\lceil \max\left\{2\log_{\sigma_2}\left(\frac{1}{n}\right), \log_{\sigma_2}\left(\frac{\delta_3}{\delta_2\sqrt{n}}\right), \log_{\sigma_2}\frac{1}{12\sqrt{n}}\right\}\right\rceil.$$

Let f^* *be the optimal value of* (3.1). *If* $\|\mathbf{d}_0\| \le 4\delta_2$, $\|\mathbf{s}_0\|_{F,\infty} \le \zeta+\delta_2$, $\mathbf{x}_0 \in \mathcal{N}_1$ *and* $\mathbf{z}_0 \in \mathcal{N}_2$, *it holds that for any* $k \in \mathbb{N}$,

$$\min_{0\le\ell\le k}\|\mathbf{x}_\ell - \bar{\mathbf{x}}_\ell\| \le \frac{8\alpha\big(\mathcal{C}_1\alpha^2L^2+4\big)}{k+1}\left(\varphi_\alpha^{\mathrm{DR}}(\mathbf{x}_0,\bar{\mathbf{y}}_0)-f^*+\frac{\mathcal{C}_3}{\alpha^2}\right)+\frac{2\mathcal{C}_5}{k+1}, \tag{4.52}$$
$$\min_{0\le\ell\le k}\|\mathrm{grad} f(\bar{\mathbf{x}}_\ell)\| \le \frac{72\big(\mathcal{C}_3\alpha^2L^2+4\big)}{(k+1)\alpha}\left(\varphi_\alpha^{\mathrm{DR}}(\mathbf{x}_0,\bar{\mathbf{y}}_0)-f^*+\frac{\mathcal{C}_5}{\alpha^2}\right)+\frac{18\mathcal{C}_5}{(k+1)\alpha^2}, \tag{4.53}$$

where

$$\begin{aligned}
\mathcal{C}_3 &:= \frac{128}{\big(1-4\sigma_2^t\big)^2}\left(\sigma_2^{2t}\alpha^2L^2+\frac{\alpha^2L^2}{\big(1-\sigma_2^t\big)^2}\right)\mathcal{D},\\
\mathcal{C}_4 &:= \frac{512n+128n\big(1-\sigma_2^t\big)^2}{\big(1-4\sigma_2^t\big)^2\big(1-\sigma_2^t\big)^2}\sum_{\ell=0}^{k}\epsilon_\ell\\
&\quad+\frac{128}{\big(1-4\sigma_2^t\big)^2\big(1-\sigma_2^{2t}\big)}\|\mathbf{d}_0-(\hat{\mathbf{x}}_0-\hat{\mathbf{s}}_0)\|^2+\frac{4}{1-16\sigma_2^{2t}}\|\mathbf{z}_0-\bar{\mathbf{y}}_0\|^2,\\
\mathcal{C}_5 &:= \mathcal{C}_4+2\sqrt{n}\mathcal{D}.
\end{aligned} \tag{4.54}$$

4.2.1. Proof of Lemma 4.8

Before proving Lemma 4.8, we first give the following two lemmas.

Lemma 4.10. *Suppose that Assumptions* 2.1, 4.1, 4.2 *hold. Let* $\{\mathbf{s}_k, \mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k\}$ *be generated by Algorithm* 3.2. *Given any* $\delta_2 > 0$, *if*

$$\epsilon_0 < \delta_2, \quad 0 < \alpha \le \min\left\{\frac{1}{2L}, \frac{\delta_2-\epsilon_0}{3\|\nabla f(0)\|+2L\left(\zeta+\delta_2\right)+2\epsilon_0}\right\},$$
$$t \ge \left\lceil \log_{\sigma_2}\left(\frac{1}{4\sqrt{n}}\right)\right\rceil, \quad \|\mathbf{d}_0\|_{F,\infty} \le 4\delta_2, \quad \|\mathbf{s}_0\|_{F,\infty} \le \zeta+\delta_2,$$

then it holds that for any k,

$$\|\mathbf{s}_k\|_{F,\infty} \le \zeta+\delta_2, \quad \|\mathbf{x}_k-\mathbf{s}_k\|_{F,\infty} \le \delta_2, \tag{4.55}$$
$$\|\alpha\nabla f(x_{i,k})+\mu_{i,k}\| \le \delta_2, \quad i \in [n], \tag{4.56}$$
$$\|d_{i,k}\| \le 4\delta_2, \qquad\qquad i \in [n]. \tag{4.57}$$

Proof. We prove it by induction on both $\|\mathbf{s}_k\|_{F,\infty}$ and $\|\mathbf{x}_k-\mathbf{s}_k\|_{F,\infty}$. Note that $\|\mathbf{s}_0\|_{F,\infty} \le \zeta+\delta_2$ and by Lemma 4.1,

$$\|x_{i,0}-s_{i,0}\| \le \alpha\big(3\|\nabla f(0)\|+2L(\zeta+\delta_2)+2\epsilon_0\big)+\epsilon_0 \le \delta_2, \tag{4.58}$$

which implies that $\|\mathbf{x}_0 - \mathbf{s}_0\|_{F,\infty} \leq \delta_2$. Suppose for some $k \geq 0$ that $\|\mathbf{s}_k\|_{F,\infty} \leq \zeta + \delta_2$ and $\|\mathbf{x}_k - \mathbf{s}_k\|_{F,\infty} \leq \delta_2$. Then we have

$$\|\mathbf{s}_{k+1}\|_{F,\infty} \leq \|\mathbf{z}_k\|_{F,\infty} + \|\mathbf{x}_k - \mathbf{s}_k\|_{F,\infty} \leq \zeta + \delta_2.$$

Similar with (4.58), we further have that

$$\|x_{i,k+1} - s_{i,k+1}\| \leq \alpha\big(3\|\nabla f(0)\| + 2L(\zeta + \delta_2) + 2\epsilon_{k+1}\big) + \epsilon_{k+1} \leq \delta_2,$$

where we use $\epsilon_{k+1} \leq \epsilon_0$. This implies that $\|\mathbf{x}_{k+1} - \mathbf{s}_{k+1}\| \leq \delta_2$. Moreover, (4.56) follows from the fact that

$$x_{i,k} = s_{i,k} - \alpha\nabla f(x_{i,k}) + \mu_{i,k}.$$

Finally, we prove (4.57) by induction. Suppose for some $k \geq 0$ such that (4.57) holds. It follows from $d_{i,0} = x_{i,0} - s_{i,0}$ that

$$\hat{d}_k = \hat{x}_k - \hat{s}_k = \alpha\hat{g}_k + \hat{\mu}_k, \tag{4.59}$$

where

$$\hat{g}_k := \frac{1}{n}\sum_{i=1}^{n} \nabla f(x_{i,k}).$$

Then we have that

$$\begin{aligned}
&\|d_{i,k+1} - (\hat{x}_k - \hat{s}_k)\| \\
&= \left\| \sum_{j=1}^{n} W_{ij}^t d_{j,k} - (\hat{x}_k - \hat{s}_k) \right\| + \|\alpha\nabla f_i(x_{i,k+1}) + \mu_{i,k+1}\| + \|\alpha\nabla f_i(x_{i,k}) + \mu_{i,k}\| \\
&\overset{(4.59)}{\leq} \left\| \sum_{j=1}^{n} \Big(W_{ij}^t - \frac{1}{n}\Big) d_{j,k} \right\| + \|\alpha\nabla f_i(x_{i,k+1}) + \mu_{i,k+1}\| + \|\alpha\nabla f_i(x_{i,k}) + \mu_{i,k}\| \\
&\overset{(4.28)(4.56)}{\leq} \sigma_2^t\sqrt{n}\max_i \|d_{i,k}\| + 2\delta_2 \\
&\leq \frac{1}{4}\max_i \|d_{i,k}\| + 2\delta_2 \leq 3\delta_2.
\end{aligned} \tag{4.60}$$

Hence,

$$\begin{aligned}
\|d_{i,k+1}\| &\leq \|d_{i,k+1} - (\hat{x}_k - \hat{s}_k)\| + \|\hat{x}_k - \hat{s}_k\| \\
&\leq 3\delta_2 + \max_i \|\alpha\nabla f(x_{i,k}) + \mu_{i,k}\| \overset{(4.56)}{\leq} 4\delta_2.
\end{aligned} \tag{4.61}$$

The proof is completed. □

The following lemma is similar to Lemma 4.5, we omit the proof.

Lemma 4.11. *Suppose that Assumptions* 2.1, 4.1, 4.2 *hold. Let* $\{\mathbf{s}_k, \mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k\}$ *be generated by Algorithm* 3.2. *Under the same condition on* $\alpha, \mathbf{d}_0$ *and* $\mathbf{s}_0$ *as in Lemma* 4.10, *if*

$$t \geq \left\lceil \max\left\{ \log_{\sigma_2}\left(\frac{1}{4\sqrt{n}}\right), \log_{\sigma_2}\left(\frac{\delta_3}{\delta_2\sqrt{n}}\right) \right\} \right\rceil$$

and $\mathbf{x}_k \in \mathcal{N}_1$, *then it holds that*

$$\sum_{j=1}^{n} W_{ij}^t x_{j,k} + d_{i,k} \in \bar{U}_{\mathcal{M}}(\gamma), \quad i \in [n], \tag{4.62}$$

where $\delta_1, \delta_2, \delta_3$ *are defined in* (4.9) *and* $\mathcal{N}_1$ *is defined in* (4.10).

Now we give the proof of Lemma 4.8.

Proof of Lemma 4.8. We prove it by induction on both $\|\hat{z}_k - \bar{z}_k\|$ and $\|\hat{x}_k - \bar{x}_k\|$. Suppose for some $k \geq 0$ such that (4.11) holds. It follows from the updated rule of $\mathbf{x}_k$ and $\mathbf{s}_k$ in Algorithm 3.1 that

$$x_{i,k+1} = z_{i,k} + \alpha\nabla f(x_{i,k}) - \alpha\nabla f(x_{i,k+1}) + \mu_{i,k} - \mu_{i,k+1}. \tag{4.63}$$

Then we have that

$$\begin{aligned}\|\hat{x}_{k+1} - \bar{x}_{k+1}\| &\leq \|\hat{x}_{k+1} - \bar{z}_k\| \\ &\leq \|\hat{x}_{k+1} - \hat{z}_k\| + \|\hat{z}_k - \bar{z}_k\| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\|x_{i,k+1} - z_{i,k}\| + 12\delta_2 \\ &\overset{(4.56)}{\leq} 2\delta_2 + 10\delta_2 \leq \delta_1,\end{aligned} \tag{4.64}$$

where the first inequality use $\bar{x}_k = \mathcal{P}_{\mathcal{M}}(\hat{x}_k)$ and $\bar{z}_k \in \mathcal{M}$. In addition,

$$\begin{aligned}\|\hat{z}_{k+1} - \bar{z}_{k+1}\| &\leq \|\hat{z}_{k+1} - \bar{x}_{k+1}\| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\|z_{i,k+1} - \bar{x}_{k+1}\| \\ &\leq \frac{2}{n}\sum_{i=1}^{n}\left\|\sum_{j=1}^{n}W_{ij}^t x_{j,k} + d_{i,k} - \hat{x}_{k+1}\right\| \\ &\leq 2\big(\sqrt{n}\sigma_2^t\delta_3 + 4\delta_2\big) \leq 10\delta_2.\end{aligned} \tag{4.65}$$

Finally, (4.50) follows from Lemma 4.11. We completed the proof. □

4.2.2. Proof of Theorem 4.2

Before proving Theorem 4.2, we need the following two lemmas, which are similar to Lemmas 4.6 and 4.7. Here, for the sake of brevity, we will omit the proof of these lemmas.

Lemma 4.12. *Suppose that the conditions in Lemma* 4.8 *hold. Then, it holds that*

$$\|\mathbf{d}_{k+1} - (\hat{\mathbf{x}}_{k+1} - \hat{\mathbf{s}}_{k+1})\| \leq \sigma_2^t\|\mathbf{d}_k - (\hat{\mathbf{x}}_k - \hat{\mathbf{s}}_k)\| + \alpha L\|\mathbf{x}_{k+1} - \mathbf{x}_k\| + 2\sqrt{n}\epsilon_k, \tag{4.66}$$

$$\begin{aligned}\|\mathbf{z}_{k+1} - \bar{\mathbf{y}}_{k+1}\| &\leq 4\sigma_2^t\big(\|\mathbf{z}_k - \bar{\mathbf{y}}_k\| + \alpha L\|\mathbf{x}_{k+1} - \mathbf{x}_k\| + 2\sqrt{n}\epsilon_k\big) \\ &\quad + 2\|\mathbf{d}_{k+1} - (\hat{\mathbf{x}}_{k+1} - \hat{\mathbf{s}}_{k+1})\|.\end{aligned} \tag{4.67}$$

Lemma 4.13. *Suppose that Assumptions* 2.1, 4.1, 4.2 *and the conditions in Lemma* 4.8 *hold. Then it holds that*

$$\sum_{\ell=0}^{k}\|\mathbf{z}_\ell - \bar{\mathbf{y}}_\ell\|^2 \leq \mathcal{C}_3\alpha^2L^2\sum_{\ell=0}^{k}\|\mathbf{x}_{\ell+1} - \mathbf{x}_\ell\|^2 + \mathcal{C}_4, \tag{4.68}$$

where $\mathcal{C}_3, \mathcal{C}_4$ *and* $\mathcal{C}_5$ *are defined in Theorem* 4.2.

Proof of Theorem 4.2. It follows from (4.51) and (4.68) that

$$\begin{aligned}\varphi_\alpha^{\mathrm{DR}}(\mathbf{s}_0)-\varphi_\alpha^{\mathrm{DR}}(\mathbf{s}_{k+1}) &\geq \frac{(1+\alpha L)\big(1-\alpha-2\alpha L-2\mathcal{C}_3\alpha L^2\big)}{2\alpha}\sum_{\ell=0}^{k}\|\mathbf{x}_{\ell+1}-\mathbf{x}_\ell\|^2\\ &\quad-\frac{(1+\alpha L)}{2\alpha^2}\left(\mathcal{C}_4+2\sqrt{n}\sum_{l=0}^{k}\epsilon_l\right)\\ &\geq \frac{1}{4\alpha}\sum_{\ell=0}^{k}\|\mathbf{x}_{\ell+1}-\mathbf{x}_\ell\|^2-\frac{(1+\alpha L)}{2\alpha^2}\mathcal{C}_5, \end{aligned}\tag{4.69}$$

where the first inequality comes from Lemma 4.13, and the second inequality is due to the assumption on α. By Assumption 4.1, we have from [43, Theorem 3.4] that

$$\varphi_\alpha^{\mathrm{DR}}(\mathbf{x}_k,\bar{\mathbf{y}}_k)\geq \inf_{\mathbf{x}}\{f(\mathbf{x})+\delta_{\mathcal{C}}(\mathbf{x})\}=f^*>-\infty.$$

Then, it follows from (4.42) that

$$\frac{1}{k+1}\sum_{\ell=0}^{k}\|\mathbf{x}_{\ell+1}-\mathbf{x}_\ell\|^2\leq\frac{4\alpha}{k+1}\left(\varphi_\alpha^{\mathrm{DR}}(\mathbf{x}_0,\bar{\mathbf{y}}_0)-f^*+\frac{\mathcal{C}_5}{\alpha^2}\right).\tag{4.70}$$

Similar with (4.44), we have that

$$\|\mathrm{grad}f(\bar{\mathbf{x}}_k)\|\leq\frac{3}{\alpha}\|\bar{\mathbf{y}}_k-\mathbf{x}_k\|.\tag{4.71}$$

By the triangle inequality and Proposition 4.1, we have

$$\begin{aligned}\|\mathbf{s}_{k+1}-\mathbf{s}_k\| &\leq \|\mathbf{s}_{k+1}+\mu_{k+1}-\mathbf{s}_k-\mu_k\|+\|\mu_{k+1}-\mu_k\|\\ &\leq 2\|\mathbf{x}_{k+1}-\mathbf{x}_k\|+2\sqrt{n}\epsilon_k.\end{aligned}\tag{4.72}$$

Then we have that

$$\begin{aligned}\|\bar{\mathbf{x}}_k-\mathbf{x}_k\| &\leq \|\bar{\mathbf{y}}_k-\mathbf{x}_k\|\\ &\leq \|\bar{\mathbf{y}}_k-\mathbf{z}_k\|+\|\mathbf{z}_k-\mathbf{x}_k\|\\ &= \|\bar{\mathbf{y}}_k-\mathbf{z}_k\|+\|\mathbf{s}_{k+1}-\mathbf{s}_k\|\\ &\leq \|\bar{\mathbf{y}}_k-\mathbf{z}_k\|+2\|\mathbf{x}_{k+1}-\mathbf{x}_k\|,\\ &\leq \|\bar{\mathbf{y}}_k-\mathbf{z}_k\|+2\|\mathbf{x}_{k+1}-\mathbf{x}_k\|+2\sqrt{n}\epsilon_k.\end{aligned}\tag{4.73}$$

Then, it holds that

$$\begin{aligned}\frac{1}{k+1}\sum_{\ell=0}^{k}\|\bar{\mathbf{x}}_\ell-\mathbf{x}_\ell\|^2 &\leq \frac{1}{k+1}\sum_{\ell=0}^{k}\big(\|\bar{\mathbf{y}}_\ell-\mathbf{z}_\ell\|+2\|\mathbf{x}_{\ell+1}-\mathbf{x}_\ell\|+2\sqrt{n}\epsilon_l\big)^2\\ &\leq \frac{3}{k+1}\cdot\left[\big(\mathcal{C}_3\alpha^2L^2+4\big)\sum_{\ell=0}^{k}\|\mathbf{x}_{\ell+1}-\mathbf{x}_\ell\|^2+\mathcal{C}_5\right].\end{aligned}\tag{4.74}$$

Combining (4.71) and (4.74) gives

$$\begin{aligned}\frac{1}{k+1}\sum_{\ell=0}^{k}\|\mathrm{grad}f(\bar{\mathbf{y}}_\ell)\|^2 &\leq \frac{9}{\alpha^2}\frac{1}{k+1}\sum_{\ell=0}^{k}\|\bar{\mathbf{y}}_\ell-\mathbf{x}_\ell\|^2\\ &\leq \frac{27\mathcal{C}_1\alpha^2L^2+108}{(k+1)\alpha^2}\sum_{\ell=0}^{k}\|\mathbf{x}_{\ell+1}-\mathbf{x}_\ell\|^2+\frac{27\mathcal{C}_5}{(k+1)\alpha^2}.\end{aligned}\tag{4.75}$$

Putting (4.70), (4.74), and (4.75) together gives (4.52) and (4.53). □

5. Numerical Experiments

In this section, we present numerical comparisons of our proposed methods with the existing decentralized manifold optimization algorithms, DRGTA [8] and DPGTA [12], on the decentralized principal component analysis (DPCA).

The DPCA problem can be mathematically formulated as

$$\begin{aligned} &\min_{\mathbf{x}\in\mathcal{M}^n} \quad -\frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}\big(x_i^\top A_i^\top A_i x_i\big), \\ &\text{s.t.} \quad x_1 = \cdots = x_n, \end{aligned} \tag{5.1}$$

where

$$\mathcal{M}^n := \underbrace{\mathrm{St}(d,r)\times\cdots\times\mathrm{St}(d,r)}_{n},$$

$A_i \in \mathbb{R}^{m_i\times d}$ is the local data matrix in i-th agent with m_i samples. Note that for any solution x^* of (5.1), x^*Q with an orthogonal matrix $Q \in \mathbb{R}^{r\times r}$ is also a solution. We use the function

$$d_s(x,x^*) := \min_{Q\in\mathbb{R}^{r\times r},\ Q^\top Q = QQ^\top = I_r} \|xQ - x^*\|$$

to compute the distance between two points x and x^*. As x is constrained on the Stiefel manifold, it always holds that $\mathrm{tr}(x^\top x) = r$. Then, we can define

$$f_i = -\frac{1}{2}\mathrm{tr}\big(x^\top\big(A_i^\top A_i - \|A_i\|_2^2 I\big)x\big)$$

for (5.1). Consequently, the proximal operator $\mathrm{prox}_{\alpha f_i}$ can be exactly calculated via solving linear equations, i.e.

$$\mathrm{prox}_{\alpha f_i}(x) = \big(I + \alpha\big(\|A_i\|_2^2 I - A_i^\top A_i\big)\big)^{-1}x.$$

Moreover, we can save $(I + \alpha(\|A_i\|_2^2 I - A_i^\top A_i))^{-1}$ to significantly reduce the computational costs. Given these, we only present the numerical results of DDRS.

5.1. Synthetic dataset

We set $m_1 = \cdots = m_n = 1000, d = 10$, and $r = 5$. Then, a matrix $B \in \mathbb{R}^{1000n\times d}$ is generated from the singular value decomposition

$$B = U\Sigma V^\top,$$

where $U \in \mathbb{R}^{1000n\times d}$ and $V \in \mathbb{R}^{d\times d}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{d\times d}$ is a diagonal matrix. To control the distributions of the singular values, we set $\tilde{\Sigma} = \mathrm{diag}(\xi^j)$ with $\xi \in (0,1)$. Furthermore, A is set as

$$A = U\tilde{\Sigma}V^\top \in \mathbb{R}^{1000n\times d}.$$

A_i is obtained by randomly splitting the rows of A into n subsets with equal cardinalities. It is easy to check the first r columns of V form the solution of (5.1). In the experiments, we set ξ and n to 0.8 and 8, respectively.

Each algorithm employs a fixed step size $\alpha = \hat{\beta}n/\sum_{i=1}^{n} m_i$, and the grid search is carried out to determine the optimal $\hat{\beta}$. Both DRPGT and DRGTA utilize the polar decomposition for their retraction operations. The connectivity between agents is simulated using various graph

matrices like the Erdos-Renyi (ER) network with probability settings $p = 0.3, 0.6$ and the Ring network. In addition, we opted for the Metropolis constant edge weight matrix [38] as our choice of the mixing matrix W.

The results are presented in Figs. 5.1 and 5.2. It can be seen from Fig. 5.1 that DDRS with the multiple-step consensus (i.e. $t = 10$) allows using a larger step size and converges faster

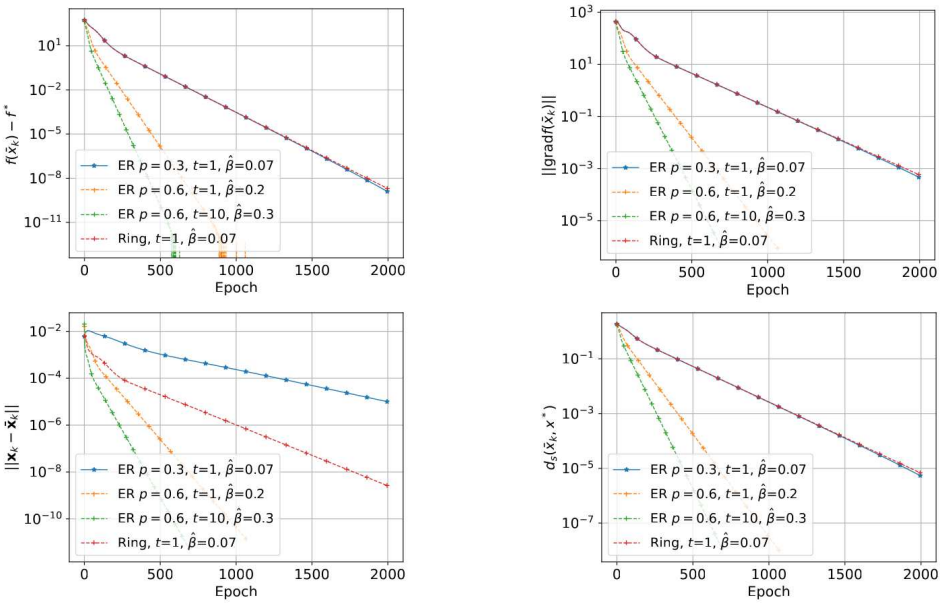

Fig. 5.1. Numerical results of DDRS for solving DPCA on the synthetic dataset with different network graphs and different t.

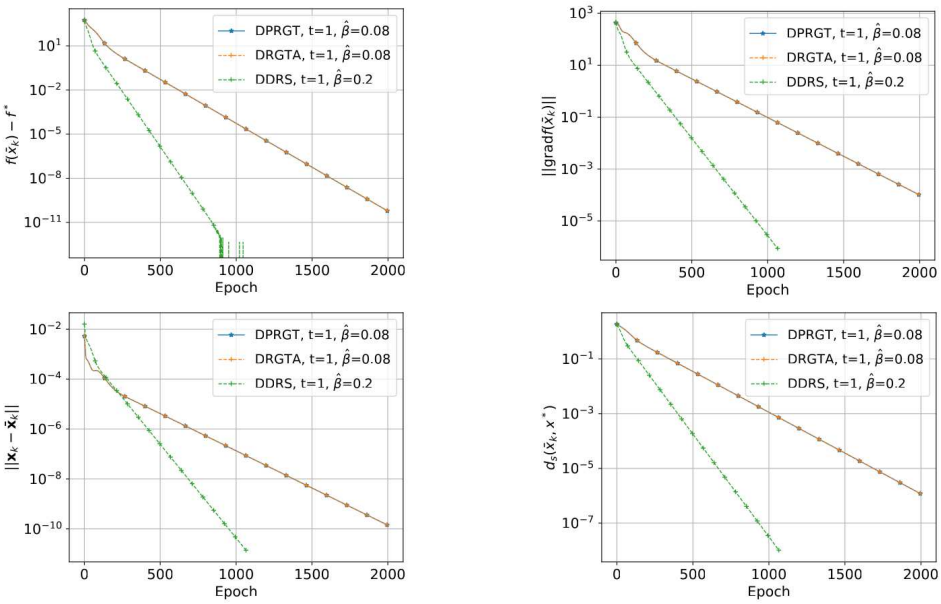

Fig. 5.2. Numerical results of different algorithms for solving DPCA on the synthetic dataset.

than those with the single-step consensus (i.e. $t = 1$). Besides, a denser graph (e.g. ER $p = 0.6$) will give better solutions. For Fig. 5.2, we see that DDRS converges much faster than DPRGT and DRGTA, and DPRGT and DRGTA have very close trajectories on the consensus error, the objective function, the gradient norm, and the distance to the global optimum.

5.2. Mnist dataset

To evaluate the efficiency of our proposed method, we also conduct numerical tests on the Mnist dataset [23]. The testing set, consisting of 60000 handwritten images of size 32×32, is used to generate A_i's. We first normalize the data matrix by dividing 255 and randomly split the data into $n = 8$ agents with equal cardinality. Then, each agent holds a local matrix A_i of dimension $60000/n \times 784$. We compute the first 5 principal components, i.e. $d = 784, r = 5$.

For all algorithms, we use the fixed step sizes $\alpha = \hat{\beta}/60000$ with a best-chosen $\hat{\beta}$. Echoing the findings from the synthetic dataset, Fig. 5.3 reveals that DDRS consistently outperforms both DPRGT and DRGTA, with the latter two showcasing very similar behaviors.

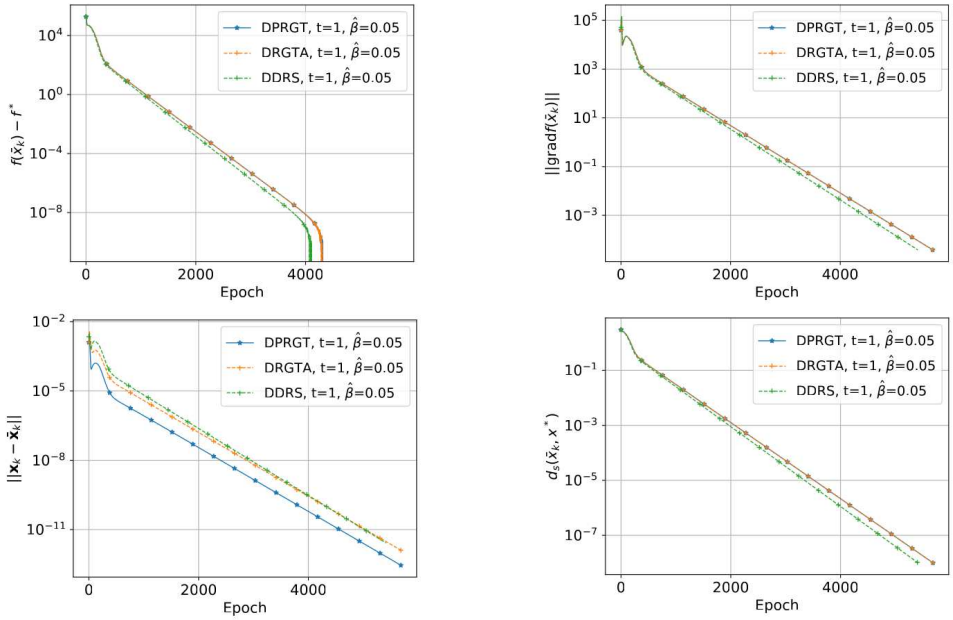

Fig. 5.3. Numerical results of different algorithms for solving DPCA on the Mnist dataset.

6. Conclusion

Through a novel fusion of gradient tracking and DRS, we propose two efficient decentralized Douglas-Rachford splitting algorithms, DDRS and iDDRS, for solving decentralized smooth optimization problems on compact submanifolds. To address the nonconvexity challenge of the manifold constraint, we employ the fundamental concept of proximal smoothness of the compact submanifold and establish the best-known convergence rate $\mathcal{O}(1/K)$ for both algorithms. Numerical experiments validate the superior performance of DDRS. Our work can also be readily extended to solve more general decentralized nonsmooth optimization problems, e.g. the strong prox-regular function [20].

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, 2009.

[2] M.V. Balashov and R.A. Kamalov, The gradient projection method with Armijo's step size on manifolds, *Comput. Math. Math. Phys.*, **61** (2021), 1776–1786.

[3] P. Bianchi and J. Jakubowicz, Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization, *IEEE Trans. Automat.*, **58**:2 (2012), 391–405.

[4] N. Boumal, *An Introduction to Optimization on Smooth Manifolds*, Cambridge University Press, 2023.

[5] N. Boumal and P.-A. Absil, Low-rank matrix completion via preconditioned optimization on the Grassmann manifold, *Linear Algebra Appl.*, **475** (2015), 200–239.

[6] S. Boyd, P. Diaconis, and L. Xiao, Fastest mixing Markov chain on a graph, *SIAM Rev.*, **46**:4 (2004), 667–689.

[7] T.-H. Chang, M. Hong, H.-T. Wai, X. Zhang, and S. Lu, Distributed learning in the nonconvex world: From batch data to streaming and beyond, *IEEE Signal Process. Mag.*, **37**:3 (2020), 26–38.

[8] S. Chen, A. Garcia, M. Hong, and S. Shahrampour, Decentralized Riemannian gradient descent on the Stiefel manifold, in: *International Conference on Machine Learning*, PMLR, (2021), 1594–1605.

[9] S. Chen, A. Garcia, M. Hong, and S. Shahrampour, On the local linear rate of consensus on the Stiefel manifold, *arXiv:2101.09346*, 2021.

[10] F.H. Clarke, R.J. Stern, and P.R. Wolenski, Proximal smoothness and the lower-C^2 property, *J. Convex Anal.*, **2**:1-2 (1995), 117–144.

[11] D. Davis, D. Drusvyatskiy, and Z. Shi, Stochastic optimization over proximally smooth sets, *arXiv: 2002.06309*, 2020.

[12] K. Deng and J. Hu, Decentralized projected Riemannian gradient method for smooth optimization on compact submanifolds, *arXiv:2304.08241*, 2023.

[13] P. Diaconis and D. Stroock, Geometric bounds for eigenvalues of Markov chains, *Ann. Appl. Probab.*, **1**:1 (1991), 36–61.

[14] J. Eckstein and D.P. Bertsekas, On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators, *Math. Program.*, **55** (1992), 293–318.

[15] T. Erseghe, D. Zennaro, E. Dall'Anese, and L. Vangelista, Fast consensus by the alternating direction multipliers method, *IEEE Trans. Signal Process.*, **59**:11 (2011), 5523–5537.

[16] D.R. Han, H.J. He, H. Yang, and X.M. Yuan, A customized Douglas-Rachford splitting algorithm for separable convex minimization with linear constraints, *Numer. Math.*, **127**:1 (2014), 167–200.

[17] H.J. He and D.R. Han, A distributed Douglas-Rachford splitting method for multi-block convex minimization problems, *Adv. Comput. Math.*, **42** (2016), 27–53.

[18] M.Y. Hong, D. Hajinezhad, and M.M. Zhao, Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks, in: *International Conference on Machine Learning*, PMLR, (2017), 1529–1538.

[19] J. Hu, K.K. Deng, N. Li, and Q.Z. Li, Decentralized Riemannian natural gradient methods with Kronecker-product approximations, *arXiv:2303.09611*, 2023.

[20] J. Hu, K.K. Deng, J.Y. Wu, and Q.Z. Li, A projected semismooth Newton method for a class of nonconvex composite programs with strong prox-regularity, *arXiv:2303.05410*, 2023.

[21] J. Hu, X. Liu, Z.W. Wen, and Y.Y. Yuan, A brief introduction to manifold optimization, *J. Oper. Res. Soc. China*, **8** (2020), 199–248.

[22] J. Hu, J.J. Zhang, and K.K. Deng, Achieving consensus over compact submanifolds, *arXiv: 2306.04769*, 2023.

[23] Y. LeCun, C. Cortes, and C.J.C. Burges, The MNIST database of handwritten digits, `http://yann.lecun.com/exdb/mnist/`, 1998.

[24] G.Y. Li and T.K. Pong, Douglas-Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems, *Math. Program.*, **159** (2016), 371–401.
[25] Q. Ling, W. Shi, G. Wu, and A. Ribeiro, DLM: Decentralized linearized alternating direction method of multipliers, *IEEE Trans. Signal Process.*, **63**:15 (2015), 4051–4064.
[26] P.D. Lorenzo and G. Scutari, NEXT: In-network nonconvex optimization, *IEEE Trans. Signal Inform. Process. Netw.*, **2**:2 (2016), 120–136.
[27] M. Maros and J. Jaldn, On the Q-linear convergence of distributed generalized ADMM under non-strongly convex function components, *IEEE Trans. Signal Inform. Process. Netw.*, **5**:3 (2018), 442–453.
[28] G. Mateos, J.A. Bazerque, and G.B. Giannakis, Distributed sparse linear regression, *IEEE Trans. Signal Process.*, **58**:10 (2010), 5262–5276.
[29] B. Mishra, H. Kasai, P. Jawanpuria, and A. Saroop, A Riemannian gossip approach to subspace learning on Grassmann manifold, *Mach. Learn.*, **108**:10 (2019), 1783–1803.
[30] A. Nedic, A. Olshevsky, and W. Shi, Achieving geometric convergence for distributed optimization over time-varying graphs, *SIAM J. Optim.*, **27**:4 (2017), 2597–2633.
[31] A. Nedic and A. Ozdaglar, Distributed subgradient methods for multi-agent optimization, *IEEE Trans. Automat.*, **54**:1 (2009), 48.
[32] P. Patrinos, L. Stella, and A. Bemporad, Douglas-Rachford splitting: Complexity estimates and accelerated variants, in: *53rd IEEE Conference on Decision and Control*, IEEE, (2014), 4234–4239.
[33] S.U. Pillai, T. Suel, and Seunghun Cha, The Perron-Frobenius theorem: Some of its applications, *IEEE Signal Process. Mag.*, **22**:2 (2005), 62–75.
[34] G.N. Qu and N. Li, Harnessing smoothness to accelerate distributed optimization, *IEEE Trans. Control Netw. Syst.*, **5**:3 (2017), 1245–1260.
[35] A. Sarlette and R. Sepulchre, Consensus optimization on manifolds, *SIAM J. Control Optim.*, **48**:1 (2009), 56–76.
[36] A. Scaglione, R. Pagliari, and H. Krim, The decentralized estimation of the sample covariance, in: *2008 42nd Asilomar Conference on Signals, Systems and Computers*, IEEE, (2008), 1722–1726.
[37] G. Scutari and Y. Sun, Distributed nonconvex constrained optimization over time-varying digraphs, *Math. Program.*, **176**:1-2 (2019), 497–544.
[38] W. Shi, Q. Ling, G. Wu, and W.T. Yin, EXTRA: An exact first-order algorithm for decentralized consensus optimization, *SIAM J. Optim.*, **25**:2 (2015), 944–966.
[39] W. Shi, Q. Ling, G. Wu, and W.T. Yin, A proximal gradient algorithm for decentralized composite optimization, *IEEE Trans. Signal Process.*, **63**:22 (2015), 6013–6023.
[40] W. Shi, Q. Ling, K. Yuan, G. Wu, and W.T. Yin, On the linear convergence of the ADMM in decentralized consensus optimization, *IEEE Trans. Signal Process.*, **62**:7 (2014), 1750–1761.
[41] H.R. Sun, S.T. Lu, and M.Y. Hong, Improving the sample and communication complexity for decentralized non-convex optimization: Joint gradient estimation and tracking, in: *International Conference on Machine Learning*, PMLR, (2020), 9217–9228.
[42] T. Tatarenko and B. Touri, Non-convex distributed optimization, *IEEE Trans. Automat.*, **62**:8 (2017), 3744–3757.
[43] A. Themelis and P. Patrinos, Douglas-Rachford splitting and ADMM for nonconvex optimization: Tight convergence results, *SIAM J. Optim.*, **30**:1 (2020), 149–181.
[44] J. Tsitsiklis, D. Bertsekas, and M. Athans, Distributed asynchronous deterministic and stochastic gradient optimization algorithms, *IEEE Trans. Automat.*, **31**:9 (1986), 803–812.
[45] H.-T. Wai, J. Lafond, A. Scaglione, and E. Moulines, Decentralized Frank-Wolfe algorithm for convex and nonconvex problems, *IEEE Trans. Automat.*, **62**:11 (2017), 5522–5537.
[46] L. Wang and X. Liu, Decentralized optimization over the Stiefel manifold by an approximate augmented Lagrangian function, *IEEE Trans. Signal Process.*, **70** (2022), 3029–3041.
[47] L. Wang and X. Liu, A variance-reduced stochastic gradient tracking algorithm for decentralized

optimization with orthogonality constraints, *J. Ind. Manag. Optim.*, **19**:10 (2023), 7753–7776.

[48] J.M. Xu, S.Y. Zhu, Y.C. Soh, and L.H. Xie, Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes, in: *Proceedings of the 54th IEEE Conference on Decision and Control*, (2015), 2055–2060.

[49] H.S. Ye and T. Zhang, DeEPCA: Decentralized exact PCA with linear convergence rate, *J. Mach. Learn. Res.*, **22**:1 (2021), 10777–10803.

[50] K. Yuan, Q. Ling, and W.T. Yin, On the convergence of decentralized gradient descent, *SIAM J. Optim.*, **26**:3 (2016), 1835–1854.

[51] J.S. Zeng and W.T. Yin, On nonconvex decentralized gradient descent, *IEEE Trans. Signal Process.*, **66**:11 (2018), 2834–2848.

[52] J. Zhang, H. Liu, A.M.-C. So, and Q. Ling, A penalty alternating direction method of multipliers for convex composite optimization over decentralized networks, *IEEE Trans. Signal Process.*, **69** (2021), 4282–4295.