Journal of Machine Learning Research 26 (2025) 1-37 Submitted 11/24; Revised 4/25; Published 5/25

A Decentralized Proximal Gradient Tracking Algorithm for Composite Optimization on Riemannian Manifolds

Lei Wang[*] WLKINGS@LSEC.CC.AC.CN
Department of Applied Mathematics
The Hong Kong Polytechnic University
Hong Kong, China

Le Bao LEBAO@PSU.EDU
Department of Statistics
The Pennsylvania State University
University Park, PA, USA

Xin Liu LIUXIN@LSEC.CC.AC.CN
State Key Laboratory of Mathematical Sciences
Academy of Mathematics and Systems Science, Chinese Academy of Sciences
and University of Chinese Academy of Sciences
Beijing, China

Editor: Shiqian Ma

Abstract

This paper focuses on minimizing a smooth function combined with a nonsmooth regularization term on a compact Riemannian submanifold embedded in the Euclidean space under a decentralized setting. Typically, there are two types of approaches at present for tackling such composite optimization problems. The first, subgradient-based approaches, rely on subgradient information of the objective function to update variables, achieving an iteration complexity of $O(\epsilon^{-4}\log^2(\epsilon^{-2}))$. The second, smoothing approaches, involve constructing a smooth approximation of the nonsmooth regularization term, resulting in an iteration complexity of $O(\epsilon^{-4})$. This paper proposes a proximal gradient type algorithm that fully exploits the composite structure. The global convergence to a stationary point is established with a significantly improved iteration complexity of $O(\epsilon^{-2})$. To validate the effectiveness and efficiency of our proposed method, we present numerical results from real-world applications, showcasing its superior performance compared to existing approaches.

Keywords: decentralized optimization, gradient tracking, manifold optimization, proximal gradient method, Riemannian manifold

1. Introduction

Many practical applications in the realms of machine learning and data science can be characterized as optimization problems on Riemannian manifolds, such as sparse principal component analysis (PCA) (Jolliffe et al., 2003; d'Aspremont et al., 2008; Journée et al., 2010; Wang et al., 2025), subspace learning (Mishra et al., 2019; Ye and Zhang, 2021), matrix completion (Boumal and Absil, 2015), sparse canonical correlation analysis (Gao

[*] Corresponding author.

and Ma, 2023), orthogonal dictionary learning (Zhu et al., 2019; Zhai et al., 2020; Li et al., 2021), deep neural networks with batch normalization (Cho and Lee, 2017). In today's era of big data, information is frequently gathered and processed by distributed systems located in various places, which are typically interconnected through communication networks. It is noteworthy that centralized algorithms often prove to be inefficient or even infeasible due to constraints related to storage requirement, communication bandwidth, and data privacy. Consequently, there is a pressing need to design and develop efficient decentralized algorithms.

Given a set of $d \in \mathbb{N}$ agents connected by a communication network, our focus is on the composite optimization problems on the Riemannian manifold of the following form:

$$\min_{X \in \mathcal{M}} \ \frac{1}{d} \sum_{i=1}^{d} f_i(X) + r(X), \tag{1.1}$$

where $f_i : \mathbb{R}^{n\times p} \to \mathbb{R}$ is a smooth local function privately owned by agent i, $r : \mathbb{R}^{n\times p} \to \mathbb{R}$ is a convex function known to all the agents, and $\mathcal{M}$ is a compact Riemannian submanifold embedded in $\mathbb{R}^{n\times p}$ (Absil et al., 2008). Employing a regularizer allows us to use prior knowledge about the problem structure explicitly. Throughout this paper, we make the following assumptions on the objective functions.

Assumption 1 *The following statements hold in the problem* (1.1)*.*

(i) *f_i is smooth and its gradient ∇f_i is Lipschitz continuous over the convex hull of $\mathcal{M}$, denoted by $\mathrm{conv}(\mathcal{M})$, with the corresponding Lipschitz constant $L_f > 0$.*

(ii) *r is convex and Lipschitz continuous with the corresponding Lipschitz constant $L_r > 0$. Moreover, the proximal mapping $\mathrm{Prox}_{\lambda r}(X)$ of r, which is defined by*

$$\mathrm{Prox}_{\lambda r}(X) := \operatorname*{arg\,min}_{Y \in \mathbb{R}^{n\times p}} \ r(Y) + \frac{1}{2\lambda} \left\| Y - X \right\|_{\mathrm{F}}^2 ,$$

is easy-to-compute for any $\lambda > 0$ and $X \in \mathbb{R}^{n\times p}$.

Under the decentralized setting, designing efficient algorithms to solve the problem (1.1) becomes particularly challenging. The complexity arises primarily from the combination of nonsmooth nature of objective functions and nonconvexity of manifold constraints.

1.1 Network Setting

We consider a scenario in which the agents can only exchange information with their immediate neighbors. The network $\mathtt{G} = (\mathtt{V}, \mathtt{E})$ captures the communication links diffusing information among the agents. Here, $\mathtt{V} = [d] := \{1, 2, \ldots, d\}$ is composed of all the agents and $\mathtt{E} = \{(i, j) \mid i \text{ and } j \text{ are connected}\}$ represents the set of communication links. Throughout this paper, we make the following assumptions on the network.

Assumption 2 *The communication network $\mathtt{G} = (\mathtt{V}, \mathtt{E})$ is connected. Furthermore, there exists a mixing matrix $W = [W(i,j)] \in \mathbb{R}^{d\times d}$ associated with $\mathtt{G}$ satisfying the following conditions.*

(i) W *is symmetric and nonnegative.*

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

(iii) $W(i,j) = 0$ *if* $i \neq j$ *and* $(i,j) \notin \mathtt{E}$, *and* $W(i,j) > 0$ *otherwise.*

The assumptions about the mixing matrix are standard in the literature (Nedić et al., 2018), under which W is primitive and doubly stochastic and conforms to the underlying network topology. Then invoking the Perron-Frobenius Theorem (Pillai et al., 2005), we know that the eigenvalues of W lie in $(-1, 1]$ and

$$\sigma := \left\| W - \mathbf{1}_d \mathbf{1}_d^\top / d \right\|_2 < 1. \tag{1.2}$$

The parameter σ characterizes the connectivity of the network $\mathtt{G}$ and plays a prominent part in the analysis of decentralized methods. In fact, $\sigma = 0$ indicates the full connectivity of $\mathtt{G}$, while σ approaches 1 as the connectivity of $\mathtt{G}$ worsens.

Finally, it is noteworthy that the mixing matrix W in Assumption 2, which always exists, can be efficiently constructed via the exchange of local degree information of $\mathtt{G}$ on a neighbor-to-neighbor basis. Interested readers can refer to (Xiao and Boyd, 2004; Gharesifard and Cortés, 2012; Shi et al., 2015a; Nedić et al., 2018) for more details.

1.2 Literature Survey

Recent years have witnessed the repaid development of decentralized optimization in the Euclidean space, marked by the emergence of various algorithms for different types of problems, such as decentralized gradient descent algorithms (Nedić and Ozdaglar, 2009; Yuan et al., 2016; Zeng and Yin, 2018), gradient tracking methods (Xu et al., 2015; Qu and Li, 2017; Nedić et al., 2017; Sun et al., 2022; Song et al., 2024), primal-dual frameworks (Shi et al., 2015a; Ling et al., 2015; Chang et al., 2015; Hajinezhad and Hong, 2019), proximal gradient approaches (Shi et al., 2015b; Scutari and Sun, 2019; Li et al., 2019; Xin et al., 2021; Yan et al., 2023), decentralized Newton methods (Bajovic et al., 2017; Zhang et al., 2021; Daneshmand et al., 2021), and so on. Interested readers can refer to some recent surveys (Nedić et al., 2018; Xin et al., 2020; Chang et al., 2020) and references therein for a complete review of the decentralized algorithms in the Euclidean space.

The investigation of decentralized algorithms for optimization problems on Riemannian manifolds is still in its infancy. Existing studies, predominantly centered on the particular case of Stiefel manifolds, can be categorized into two main groups, which will be briefly introduced below.

The first group of algorithms extends classical decentralized methods in the Euclidean space by leveraging the geometric tools derived from Riemannian optimization (Absil et al., 2008). For instance, Chen et al. (2021) propose a decentralized Riemannian stochastic gradient descent (DRSGD) method along with its gradient-tracking variant DRGTA. Subsequently, these two methods are generalized to compact Riemannian submanifolds in Deng and Hu (2023), resulting in algorithms named DPRGD and DPRGT, respectively. Moreover, the algorithm DRCGD, developed in Chen et al. (2024), incorporates the Riemannian conjugate gradient method into the framework of Deng and Hu (2023) to enhance convergence rates. Additionally, Hu and Deng (2024) propose a decentralized algorithm DPRGC to improve the communication efficiency, which achieves the global consensus on manifolds through single-step communication. The aforementioned approaches are tailored for smooth objective functions and, therefore, may not be directly applicable to the problem (1.1). Recently, Wang et al. (2024) have introduced a decentralized Riemannian subgradient method (DRSM) aimed at tackling nonsmooth optimization problems on Riemannian manifolds. This method operates under the assumption that each local function is weakly convex. However, it is worth noting that DRSM suffers from a slow convergence rate, attributed to its reliance solely on subgradient information.

The second group of algorithms employs penalty functions to handle nonconvex manifold constraints, including DESTINY (Wang and Liu, 2022), VRSGT (Wang and Liu, 2023b), and THANOS (Wang and Liu, 2023a). These algorithms require only a single round of communication per iteration to guarantee the convergence. Specifically, DESTINY incorporates a gradient tracking scheme into the minimization of an approximate augment Lagrangian function. VRSGT goes a step further by integrating the variance reduction technique to simultaneously reduce the communication and sampling complexities. To deal with nonsmooth regularizers, THANOS constructs an approximation of the objective function based on the Moreau envelope. It is crucial to emphasize that these algorithms heavily rely on the specific structure of the Stiefel manifold, posing challenges for their extension to more general Riemannian manifolds.

1.3 Our Contributions

In response to the growing demand for processing large-scale datasets in practical applications, we propose a novel decentralized algorithm called DR-ProxGT for solving the problem (1.1). Our approach involves solving a proximal gradient subproblem on the tangent space, where we leverage the gradient tracking technique to estimate the Euclidean gradient across the whole network. The convergence analysis reveals that DR-ProxGT converges globally to a stationary point of (1.1) and exhibits an iteration complexity of $O(\epsilon^{-2})$. To the best of our knowledge, this achieves the best result in the literature. Please refer to Table 1 for a comparative overview of existing complexity results.

Remarkably, DR-ProxGT is the first decentralized algorithm to solve the nonsmooth optimization problem (1.1) under the single-step consensus with guaranteed convergence. It is important to note that the convergence results presented in existing works (Chen et al., 2021; Wang et al., 2024; Hu and Deng, 2024) can not be straightforwardly extended to deal with the nonsmooth regularization term. Distinct from their methods, we eliminate the dependence on moving average techniques, thereby avoiding the introduction of an extra sufficiently small constant in the consensus steps that would otherwise slow down the convergence rate. Moreover, we establish a descent inequality for the nonsmooth composite function in (1.1) and derive a refined uniform bound of gradient trackers. These technical novelties require meticulous handling of the proximal mapping. Our theoretical analysis paves the way for developing new algorithms to solve the problem (1.1) under the decentralized setting.

Finally, we conduct comprehensive numerical experiments to compare DR-ProxGT with two competing algorithms. The test results are strongly in favor of our algorithm in practical applications.

Table 1: A summary of the iteration complexity for existing algorithms to find an ϵ-stationary point of the problem (1.1).

Algorithm	Manifold	Iteration Complexity
DRSM (Wang et al., 2024)	Stiefel manifold	$O(\epsilon^{-4}\log^2(\epsilon^{-2}))$
THANOS (Wang and Liu, 2023a)	Stiefel manifold	$O(\epsilon^{-4})$
DR-ProxGT (this work)	compact manifold	$O(\epsilon^{-2})$

1.4 Notations

The following notations are adopted throughout this paper. The Euclidean inner product of two matrices Y_1, Y_2 with the same size is defined as $\langle Y_1, Y_2\rangle = \mathrm{tr}(Y_1^\top Y_2)$, where $\mathrm{tr}(B)$ stands for the trace of a square matrix B. And the notation $I_p \in \mathbb{R}^{p\times p}$ represents the $p\times p$ identity matrix. The Frobenius and spectral norm of a matrix C are denoted by $\|C\|_{\mathrm{F}}$ and $\|C\|_2$, respectively. The (i,j)-th entry of a matrix C is represented by $C(i,j)$. The notation $\mathbf{1}_d \in \mathbb{R}^d$ stands for the d-dimensional vector of all ones. We define the distance and the projection of a point $X \in \mathbb{R}^{n\times p}$ onto a set $\mathcal{C} \subset \mathbb{R}^{n\times p}$ by $\mathrm{dist}(X,\mathcal{C}) := \inf\{\|Y - X\|_{\mathrm{F}} \mid Y \in \mathcal{C}\}$ and $\mathrm{Proj}_{\mathcal{C}}(X) := \arg\min_{Y\in\mathcal{C}} \|Y - X\|_{\mathrm{F}}$, respectively. The Kronecker product is denoted by $\otimes$. Given a differentiable function $g(X) : \mathbb{R}^{n\times p} \to \mathbb{R}$, the Euclidean gradient of g with respect to X is represented by $\nabla g(X)$. Further notations will be introduced wherever they occur.

1.5 Outline

The remainder of this paper is organized as follows. Section 2 draws into some preliminaries of Riemannian optimization. In Section 3, we devise a proximal gradient type algorithm for solving the problem (1.1). The convergence properties of the proposed algorithm are investigated in Section 4. Numerical results are presented in Section 5 to evaluate the performance of our algorithm. Finally, this paper concludes with concluding remarks and key insights in Section 6.

2. Preliminaries

This section introduces and reviews some basic notions and concepts regarding Riemannian manifolds that are closely related to the present work in this paper.

2.1 Proximal Smoothness

Our theoretical analysis heavily relies on the concept of proximal smoothness, which tides us over the obstacle incurred by the nonconvexity of manifolds. Following Clarke et al. (1995), we say a closed set $\mathcal{C}$ is δ-proximally smooth for a constant $\delta > 0$ if the projection $\mathrm{Proj}_{\mathcal{C}}(X)$ is a singleton whenever $\mathrm{dist}(X,\mathcal{C}) < \delta$. It should be noted that the operation operator $\mathrm{Proj}_{\mathcal{C}}(X)$ of a δ-proximally smooth set $\mathcal{C}$ is Lipschitz continuous as long as X is not far away from $\mathcal{C}$. Specifically, for any $\gamma \in (0,\delta)$, the following relationship holds whenever $\operatorname{dist}(X, \mathcal{C}) < \gamma$ and $\operatorname{dist}(Y, \mathcal{C}) < \gamma$,

$$\|\mathrm{Proj}_{\mathcal{C}}(X) - \mathrm{Proj}_{\mathcal{C}}(Y)\|_{\mathrm{F}} \leq \frac{\delta}{\delta - \gamma} \|X - Y\|_{\mathrm{F}} .$$

As is shown in Clarke et al. (1995), Davis et al. (2020), and Balashov and Kamalov (2021), any compact C^2-manifold $\mathcal{M}$ embedded in the Euclidean space is proximally smooth. For instance, the Stiefel manifold is 1-proximally smooth and the Grassmann manifold is $1/\sqrt{2}$-proximally smooth. Throughout this paper, we assume that the manifold $\mathcal{M}$ is δ-proximally smooth for a constant $\delta > 0$. Then for any $X, Y \in \mathcal{R}(\gamma) := \{X \in \mathbb{R}^{n \times p} \mid \operatorname{dist}(X, \mathcal{M}) < \gamma\}$ with $\gamma \in (0, \delta)$, we have

$$\|\mathrm{Proj}_{\mathcal{M}}(X) - \mathrm{Proj}_{\mathcal{M}}(Y)\|_{\mathrm{F}} \leq \frac{\delta}{\delta - \gamma} \|X - Y\|_{\mathrm{F}} . \tag{2.1}$$

Below is another crucial inequality regarding the projection operator $\mathrm{Proj}_{\mathcal{M}}$ indicating that $X + \mathrm{Proj}_{\mathcal{T}_X}(V)$ is a second-order approximation of $\mathrm{Proj}_{\mathcal{M}}(X + V)$ for any $X \in \mathcal{M}$ and $V \in \mathbb{R}^{n \times p}$. Specifically, it holds that

$$\left\|\mathrm{Proj}_{\mathcal{M}}(X + V) - X - \mathrm{Proj}_{\mathcal{T}_X}(V)\right\|_{\mathrm{F}} \leq M_{pj} \|V\|_{\mathrm{F}}^2 , \tag{2.2}$$

where $M_{pj} > 0$ is a constant. We refer interested readers to Deng and Hu (2023) for the construction of (2.2).

2.2 Consensus on Riemannian Manifolds

In decentralized networks, where only local communications are permitted, each agent i has to maintain its local copy $X_i \in \mathcal{M}$ of the common variable X in the problem (1.1). Let $\hat{X}$ be the Euclidean average of $\{X_i\}_{i=1}^d$ defined by

$$\hat{X} := \frac{1}{d} \sum_{i=1}^d X_i = \operatorname*{arg\,min}_{Y \in \mathbb{R}^{n \times p}} \sum_{i=1}^d \|Y - X_i\|_{\mathrm{F}}^2 .$$

To access the consensus error, the quantity $\sum_{i=1}^d \|X_i - \hat{X}\|_{\mathrm{F}}^2$ is typically used in the convergence analysis of Euclidean decentralized algorithms.

It should be noted that the average $\hat{X}$ may not necessarily lie in the manifold $\mathcal{M}$ even if $X_i \in \mathcal{M}$ for $i \in [d]$, owing to the nonconvexity of $\mathcal{M}$. This observation motivates the definition of induced arithmetic mean introduced in Sarlette and Sepulchre (2009) as follows,

$$\bar{X} \in \operatorname*{arg\,min}_{Y \in \mathcal{M}} \sum_{i=1}^d \|Y - X_i\|_{\mathrm{F}}^2 = \mathrm{Proj}_{\mathcal{M}}(\hat{X}). \tag{2.3}$$

According to the stationarity condition of the above problem, we have $\bar{X} - \hat{X} \in \mathcal{N}_{\bar{X}}$. In the development of our algorithm, we will guarantee that $\hat{X} \in \mathcal{R}(\gamma)$. The proximal smoothness of $\mathcal{M}$ then results in that $\mathrm{Proj}_{\mathcal{M}}(\hat{X})$ is a singleton.

Moreover, as is discussed in Deng and Hu (2023), the distance between $\hat{X}$ and $\bar{X}$ can be controlled by the consensus error measured by $\bar{X}$ under a certain condition. Specifically, for any $\{X_i \in \mathcal{M}\}_{i=1}^d$ satisfying $\max_{i\in[d]} \|X_i - \bar{X}\|_{\mathrm{F}} \leq \gamma$, there exists a constant $M_{av} > 0$ such that

$$\left\|\hat{X} - \bar{X}\right\|_{\mathrm{F}} \leq \frac{M_{av}}{d} \sum_{i=1}^{d} \left\|X_i - \bar{X}\right\|_{\mathrm{F}}^2, \tag{2.4}$$

which will be used in the subsequent analysis.

2.3 Stationarity Condition

In this subsection, we delve into the stationarity condition of the problem (1.1). Towards this end, we first introduce the definition of Clarke subgradient (Clarke, 1990) for nonsmooth functions.

Definition 3 *Suppose $f : \mathbb{R}^{n\times p} \to \mathbb{R}$ is a Lipschitz continuous function. The generalized directional derivative of f at the point $X \in \mathbb{R}^{n\times p}$ along the direction $H \in \mathbb{R}^{n\times p}$ is defined by:*

$$f^{\circ}(X; H) := \limsup_{Y\to X,\, t\to 0^+} \frac{f(Y + tH) - f(Y)}{t}.$$

Based on generalized directional derivative of f, the (Clarke) subgradient of f is defined by:

$$\partial f(X) := \{G \in \mathbb{R}^{n\times p} \mid \langle G, H\rangle \leq f^{\circ}(X; H)\}.$$

Additionally, we employ geometric concepts of Riemannian manifolds to present the stationarity condition. For each point $X \in \mathcal{M}$, the tangent space to $\mathcal{M}$ at X is referred to as $\mathcal{T}_X$, and the notation $\mathcal{N}_X$ represents the normal space at X. In this paper, we consider the Riemannian metric $\langle\cdot,\cdot\rangle_X$ on $\mathcal{T}_X$ that is induced from the Euclidean inner product $\langle\cdot,\cdot\rangle$, i.e., $\langle V_1, V_2\rangle_X = \langle V_1, V_2\rangle = \mathrm{tr}(V_1^\top V_2)$. Roughly speaking, the tangent space $\mathcal{T}_X$ intuitively contains the possible directions in which one can tangentially pass through X, while the normal space $\mathcal{N}_X$ is the orthogonal complement of $\mathcal{T}_X$ in $\mathbb{R}^{n\times p}$. For precise statements on these notions, interested readers can refer to the monograph of Absil et al. (2008).

For convenience, we denote $f(X) := \sum_{i=1}^d f_i(X)/d$ and $h(X) := f(X) + r(X)$. Now we are prepared to state the stationarity condition of the problem (1.1), which has been thoroughly discussed in Yang et al. (2014) and Chen et al. (2020).

Definition 4 *A point $X \in \mathcal{M}$ is called a stationary point of the problem* (1.1) *if it satisfies the following stationarity condition,*

$$0 \in \mathrm{Proj}_{\mathcal{T}_X}\left(\partial h(X)\right) = \mathrm{Proj}_{\mathcal{T}_X}\left(\nabla f(X) + \partial r(X)\right), \tag{2.5}$$

or equivalently, there exists an element $R(X) \in \partial r(X)$ such that $\nabla f(X) + R(X) \in \mathcal{N}_X$.

As previously noted, the decentralized framework considered in this paper inherently precludes access to global information of the problem (1.1), and each agent i has to independently maintain a local variable X_i. Consequently, it becomes imperative to design an algorithm that achieves a global consensus across all local variables. To capture this fundamental challenge, we introduce the following concept of an approximate stationary point.

Definition 5 *A point* $\mathbf{X} = [X_1^\top, \dots, X_d^\top]^\top$ *with* $X_i \in \mathcal{M}$ *for all* $i \in [d]$ *is called an* ϵ*-stationary point of problem* (1.1) *with* $\epsilon > 0$ *if there exists* $\mathbf{S} = [S_1^\top, \dots, S_d^\top]^\top$ *such that*

$$\begin{cases} \operatorname{dist}\left(0, \operatorname{Proj}_{\mathcal{T}_{X_i}}\left(\nabla f(X_i) + \partial r(X_i + S_i)\right)\right) \leq \epsilon, \\ \left\|X_i - \bar{X}\right\|_{\mathrm{F}} \leq \epsilon, \\ \|S_i\|_{\mathrm{F}} \leq \epsilon, \end{cases}$$

for any $i \in [d]$*. Here,* $\bar{X} \in \mathcal{M}$ *represents the induced arithmetic mean defined in* (2.3)*.*

The above definition is inspired by Chen et al. (2020) and arises from a perturbation of the stationarity condition (2.5). Obviously, when $\epsilon = 0$, each local variable X_i precisely satisfies the condition (2.5), thereby qualifying as a stationary point of the problem (1.1).

3. Decentralized Riemannian Proximal Gradient Tracking

In this section, we devise an efficient decentralized algorithm to solve the problem (1.1) by exploiting the composite structure. Each iteration of our algorithm attempts to tackle a convex subproblem coupled with a linear constraint. The objection function of this subproblem is constructed by linearizing the function f around the current iterate. Simultaneously, the linear constraint in question captures the structure of the tangent space. A key feature of our algorithm is the integration of the gradient tracking technique, which is employed to estimate ∇f across the entire network. With the help of gradient tracking, our algorithm can effectively achieve exact consensus.

3.1 Algorithm Development

We introduce two auxiliary local variables, $D_i \in \mathbb{R}^{n\times p}$ and $S_i \in \mathbb{R}^{n\times p}$, for agent i in our algorithm. Specifically, D_i is designed to track the global gradient ∇f through the exchange of local gradient information, while S_i aims at estimating the search direction on the tangent space based on D_i.

Hereafter, we use the notations $X_i^{(k)}$, $D_i^{(k)}$, and $S_i^{(k)}$ to represent the k-th iterate of X_i, D_i, and S_i, respectively. The key steps of our algorithm from the perspective of each agent are outlined below.

Step 1: S-update. Given the composite structure, a natural approach to tackle the problem (1.1) involves evaluating the following proximal gradient step constrained to the tangent space, as proposed by Chen et al. (2020),

$$\min_{S \in \mathcal{T}_X} \ \langle \nabla f(X), S \rangle + \frac{1}{2\tau} \|S\|_{\mathrm{F}}^2 + r(X + S),$$

where $\tau > 0$ is a constant. The intuition behind this method is to seek a descent direction on the tangent space by replacing the smooth term f with its first-order approximation around the current estimate. Under the decentralized setting, however, the global gradient ∇f is not available. To overcome this challenge, we introduce the auxiliary variable $D_i \in \mathbb{R}^{n\times p}$ at agent i to estimate ∇f across the whole network. Specifically, at iteration k, each agent

i aims to solve the following subproblem to obtain the search direction,

$$S_i^{(k)} := \operatorname*{arg\,min}_{S_i \in \mathcal{T}_{X_i^{(k)}}} \left\langle D_i^{(k)}, S_i \right\rangle + \frac{1}{2\tau} \|S_i\|_{\mathrm{F}}^2 + r(X_i^{(k)} + S_i), \tag{3.1}$$

where $\tau > 0$ remains a constant. The subproblem in (3.1) involves minimizing a strongly convex function subject to linear constraints. In the specific scenario where $\mathcal{M}$ is the Stiefel manifold, Chen et al. (2020) and Liu et al. (2024) have leveraged the semi-smooth Newton method and the fixed-point method, respectively, to address a subproblem with a comparable structure. Notably, these methodologies can be seamlessly extended to encompass the case of compact manifolds embedded in the Euclidean space.

Step 2: X-update. Once the search direction is determined, our algorithm executes a local update along the direction of $S_i^{(k)}$ for agent i at iteration k, combining the consensus step with the projection onto $\mathcal{M}$,

$$X_i^{(k+1)} := \mathrm{Proj}_{\mathcal{M}} \left(\sum_{j=1}^{d} W(i,j) X_j^{(k)} + \eta S_i^{(k)} \right), \tag{3.2}$$

where $\eta > 0$ is a stepsize.

Step 3: D-update. Finally, leveraging the updated information, each agent i computes the new estimate of ∇f by resorting to the gradient tracking technique (Zhu and Martínez, 2010; Daneshmand et al., 2020),

$$D_i^{(k+1)} := \sum_{j=1}^{d} W(i,j) D_j^{(k)} + \nabla f_i(X_i^{(k+1)}) - \nabla f_i(X_i^{(k)}). \tag{3.3}$$

Diverging from existing approaches (Chen et al., 2021; Deng and Hu, 2023), our algorithm directly tracks the Euclidean gradient instead of the Riemannian gradient (Absil et al., 2008). This strategy serves to alleviate the computational burden associated with projecting the Euclidean gradient onto the tangent space.

The whole procedure is summarized in Algorithm 1, named *Decentralized Riemannian proximal gradient tracking* and abbreviated to DR-ProxGT. Our approach merges the Riemannian proximal gradient method with the gradient tracking technique, resulting in a substantial enhancement in iteration complexity. The details will be elaborated upon in the subsequent section.

3.2 Comparison with Existing Methods

In our algorithm, each iteration requires only a single round of communication to guarantee the convergence. Similar results have also been established in the literature (Wang and Liu, 2022, 2023a,b). However, it is crucial to note that these algorithms are built upon the penalty function specifically crafted for the Stiefel manifold, making it difficult to generalize to broader contexts.

Algorithm 1: decentralized Riemannian proximal gradient tracking.

```
Input: $X_{\text{initial}} \in \mathcal{M}$, $\tau > 0$, and $\eta > 0$.
Set $k := 0$.
for $i \in [d]$ do
    Initialize $X_i^{(k)} := X_{\text{initial}}$ and $D_i^{(k)} := \nabla f_i(X_i^{(k)})$.
while "not converged" do
    for $i \in [d]$ do
        Compute $S_i^{(k)}$ by (3.1).
        Update $X_i^{(k+1)}$ by (3.2).
        Update $D_i^{(k+1)}$ by (3.3).
    Set $k := k + 1$.
Output: $\{X_i^{(k)}\}_{i=1}^d$.
```

Very recently, Hu and Deng (2024) have proposed a decentralized algorithm DPRGC with single-step consensus for smooth optimization problems on manifolds. The main iteration of DPRGC leverages the moving average technique and reads as follows,

$$X_i^{(k+1)} := \mathrm{Proj}_{\mathcal{M}}\left((1-\omega)X_i^{(k)} + \omega \sum_{j=1}^{d} W(i,j) X_j^{(k)} + \eta U_i^{(k)} \right),$$

where $\eta > 0$ is a stepsize, $\omega \in (0,1]$ is a constant, and $U_i^{(k)}$ represents a descent direction. Similar techniques have also been employed in Chen et al. (2021) and Wang et al. (2024). It is worth mentioning that, to control the consensus error, DPRGC imposes a stringent condition on the parameter ω, requiring it to be sufficiently small. In practical applications, this condition may hinder the convergence rate of algorithms.

4. Convergence Analysis

The global convergence of our proposed Algorithm 1 is rigorously established under mild conditions in this section. To facilitate the narrative, we define the following notations.

- $J = \mathbf{1}_d \mathbf{1}_d^\top / d$, $\mathbf{J} = J \otimes I_n$, $\mathbf{W} = W \otimes I_n$.
- $\hat{X}^{(k)} = \dfrac{1}{d}\sum\limits_{i=1}^{d} X_i^{(k)}$, $\bar{X}^{(k)} \in \mathrm{Proj}_{\mathcal{M}}(\hat{X}^{(k)})$, $\hat{D}^{(k)} = \dfrac{1}{d}\sum\limits_{i=1}^{d} D_i^{(k)}$, $\hat{G}^{(k)} = \dfrac{1}{d}\sum\limits_{i=1}^{d} \nabla f_i(X_i^{(k)})$.
- $\mathbf{X}^{(k)} = [(X_1^{(k)})^\top, \dots, (X_d^{(k)})^\top]^\top$, $\hat{\mathbf{X}}^{(k)} = (\mathbf{1}_d \otimes I_n)\hat{X}^{(k)} = \mathbf{J}\mathbf{X}^{(k)}$, $\bar{\mathbf{X}}^{(k)} = (\mathbf{1}_d \otimes I_n)\bar{X}^{(k)}$.
- $\mathbf{D}^{(k)} = [(D_1^{(k)})^\top, \dots, (D_d^{(k)})^\top]^\top$, $\hat{\mathbf{D}}^{(k)} = (\mathbf{1}_d \otimes I_n)\hat{D}^{(k)} = \mathbf{J}\mathbf{D}^{(k)}$.
- $\mathbf{G}^{(k)} = [(\nabla f_1(X_1^{(k)}))^\top, \dots, (\nabla f_d(X_d^{(k)}))^\top]^\top$, $\hat{\mathbf{G}}^{(k)} = (\mathbf{1}_d \otimes I_n)\hat{G}^{(k)} = \mathbf{J}\mathbf{G}^{(k)}$.
- $\mathbf{S}^{(k)} = [(S_1^{(k)})^\top, \dots, (S_d^{(k)})^\top]^\top$, $\hat{S}^{(k)} = \dfrac{1}{d}\sum\limits_{i=1}^{d} S_i^{(k)}$.

The following constants will also be used in the subsequent analysis.

$$\begin{cases} M_g := \sup\left\{\|\nabla f_i(X)\|_{\mathrm{F}} \mid X \in \mathcal{M}, i \in [d]\right\}, \\ M_l := \sup\left\{\|X - Y\|_{\mathrm{F}} \mid X, Y \in \mathcal{M}\right\}, \\ \underline{h} := \inf\left\{f(X) + r(X) \mid X \in \operatorname{conv}(\mathcal{M})\right\}. \end{cases} \tag{4.1}$$

Since ∇f_i is Lipschitz continuous and $\mathcal{M}$ is compact, the above constants are well-defined.

4.1 Boundedness of Iterates

The purpose of this subsection is to show the boundedness of the iterate sequence generated by Algorithm 1, which is denoted by $\{(\mathbf{X}^{(k)}, \mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$. It is obvious that the sequence $\{\mathbf{X}^{(k)}\}$ is bounded since $\mathcal{M}$ is a compact manifold. Next, we prove that the sequence $\{\mathbf{D}^{(k)}\}$ is bounded.

Lemma 6 *Suppose that Assumption 1 and Assumption 2 hold. Let $\{(\mathbf{X}^{(k)}, \mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$ be the iterate sequence generated by Algorithm 1. Then for any $k \in \mathbb{N}$, it holds that*

$$\left\|\mathbf{D}^{(k)}\right\|_{\mathrm{F}} \leq C_{gt},$$

where $C_{gt} := 2\sqrt{d}M_g/(1-\sigma) + \sqrt{d}M_g$ is a positive constant with M_g defined in (4.1).

Proof According to the triangular inequality, we have

$$\left\|\mathbf{G}^{(k+1)} - \mathbf{G}^{(k)}\right\|_{\mathrm{F}} \leq \left\|\mathbf{G}^{(k+1)}\right\|_{\mathrm{F}} + \left\|\mathbf{G}^{(k)}\right\|_{\mathrm{F}} \leq 2\sqrt{d}M_g,$$

where the last inequality follows from the definition of M_g. And straightforward calculations give rise to the following relationship,

$$\begin{aligned} \left\|\mathbf{D}^{(k+1)} - \hat{\mathbf{D}}^{(k+1)}\right\|_{\mathrm{F}} &= \left\|(\mathbf{W} - \mathbf{J})\left(\mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)}\right) + (I_{dn} - \mathbf{J})\left(\mathbf{G}^{(k+1)} - \mathbf{G}^{(k)}\right)\right\|_{\mathrm{F}} \\ &\leq \left\|(\mathbf{W} - \mathbf{J})\left(\mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)}\right)\right\|_{\mathrm{F}} + \left\|\mathbf{G}^{(k+1)} - \mathbf{G}^{(k)}\right\|_{\mathrm{F}} \\ &\leq \sigma\left\|\mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)}\right\|_{\mathrm{F}} + 2\sqrt{d}M_g. \end{aligned}$$

Then, by mathematical induction, we can obtain that

$$\left\|\mathbf{D}^{(k+1)} - \hat{\mathbf{D}}^{(k+1)}\right\|_{\mathrm{F}} \leq \sigma^{k+1}\left\|\mathbf{D}^{(0)} - \hat{\mathbf{D}}^{(0)}\right\|_{\mathrm{F}} + \frac{2\sqrt{d}M_g(1-\sigma^{k+1})}{1-\sigma}.$$

which together with $\mathbf{D}^{(0)} = \mathbf{G}^{(0)}$ and $\hat{\mathbf{D}}^{(0)} = \hat{\mathbf{G}}^{(0)}$ yields that

$$\begin{aligned} \left\|\mathbf{D}^{(k+1)} - \hat{\mathbf{D}}^{(k+1)}\right\|_{\mathrm{F}} &\leq \sigma^{k+1}\left\|\mathbf{G}^{(0)} - \hat{\mathbf{G}}^{(0)}\right\|_{\mathrm{F}} + \frac{2\sqrt{d}M_g(1-\sigma^{k+1})}{1-\sigma} \\ &\leq \frac{2\sqrt{d}M_g(1-\sigma^{k+2})}{1-\sigma}. \end{aligned}$$

In addition, it can be straightforwardly verified that

$$\left\| \hat{D}^{(k)} \right\|_{\mathrm{F}} = \left\| \frac{1}{d} \sum_{i=1}^{d} \nabla f_i(X_i^{(k)}) \right\|_{\mathrm{F}} \leq \frac{1}{d} \sum_{i=1}^{d} \left\| \nabla f_i(X_i^{(k)}) \right\|_{\mathrm{F}} \leq M_g. \tag{4.2}$$

Therefore, we can obtain that

$$\begin{aligned} \left\| \mathbf{D}^{(k+1)} \right\|_{\mathrm{F}} &\leq \left\| \mathbf{D}^{(k+1)} - \hat{\mathbf{D}}^{(k+1)} \right\|_{\mathrm{F}} + \left\| \hat{\mathbf{D}}^{(k+1)} \right\|_{\mathrm{F}} \\ &= \left\| \mathbf{D}^{(k+1)} - \hat{\mathbf{D}}^{(k+1)} \right\|_{\mathrm{F}} + \sqrt{d} \left\| \hat{D}^{(k+1)} \right\|_{\mathrm{F}} \\ &\leq \frac{2\sqrt{d} M_g (1 - \sigma^{k+2})}{1 - \sigma} + \sqrt{d} M_g \\ &\leq \frac{2\sqrt{d} M_g}{1 - \sigma} + \sqrt{d} M_g = C_{gt}, \end{aligned}$$

where the last inequality follows from the fact that $\sigma \in [0, 1)$. The proof is completed. ■

Then we prove that the norm of $S_i^{(k)}$ is controlled by that of $D_i^{(k)}$ for any $i \in [d]$ and $k \in \mathbb{N}$ in the following lemma.

Lemma 7 *Suppose that Assumption 1 and Assumption 2 hold. Let* $\{(\mathbf{X}^{(k)}, \mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$ *be the iterate sequence generated by Algorithm 1. Then, for any* $i \in [d]$ *and* $k \in \mathbb{N}$*, it holds that*

$$\left\| S_i^{(k)} \right\|_{\mathrm{F}} \leq \tau \left\| D_i^{(k)} \right\|_{\mathrm{F}} + \tau L_r.$$

Proof To begin with, the assertion of this lemma is obvious if $S_i^{(k)} = 0$. Next, we investigate the case that $S_i^{(k)} \neq 0$. For convenience, we denote the objective function in (3.1) by

$$g_i^{(k)}(S) := \left\langle D_i^{(k)}, S \right\rangle + \frac{1}{2\tau} \left\| S \right\|_{\mathrm{F}}^2 + r(X_i^{(k)} + S).$$

Since $g_i^{(k)}$ is strongly convex with modulus $1/\tau$, we have

$$g_i^{(k)}(S') \geq g_i^{(k)}(S) + \left\langle \partial g_i^{(k)}(S), S' - S \right\rangle + \frac{1}{2\tau} \left\| S' - S \right\|_{\mathrm{F}}^2, \tag{4.3}$$

for any $S, S' \in \mathbb{R}^{n \times p}$. In particular, if $S, S' \in \mathcal{T}_{X_i^{(k)}}$, it holds that

$$\left\langle \partial g_i^{(k)}(S), S' - S \right\rangle = \left\langle \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}} \left(\partial g_i^{(k)}(S) \right), S' - S \right\rangle.$$

Then it follows from the first-order optimality condition of (3.1) that

$$0 \in \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}} \left(\partial g_i^{(k)}(S_i^{(k)}) \right).$$

Upon taking $S = S_i^{(k)}$ and $S' = 0$ in (4.3), we can obtain that

$$g_i^{(k)}(0) - g_i^{(k)}(S_i^{(k)}) \geq \frac{1}{2\tau} \left\| S_i^{(k)} \right\|_{\mathrm{F}}^2,$$

which together with the Lipschitz continuity of r infers that

$$\begin{aligned}\frac{1}{\tau} \left\| S_i^{(k)} \right\|_{\mathrm{F}}^2 &\leq r(X_i^{(k)}) - r(X_i^{(k)} + S_i^{(k)}) - \left\langle D_i^{(k)}, S_i^{(k)} \right\rangle \\ &\leq L_r \left\| S_i^{(k)} \right\|_{\mathrm{F}} + \left\| D_i^{(k)} \right\|_{\mathrm{F}} \left\| S_i^{(k)} \right\|_{\mathrm{F}}.\end{aligned}$$

Hence, we can arrive at the conclusion that

$$\left\| S_i^{(k)} \right\|_{\mathrm{F}} \leq \tau \left\| D_i^{(k)} \right\|_{\mathrm{F}} + \tau L_r,$$

as desired. ■

Now the boundedness of the sequence $\{\mathbf{S}^{(k)}\}$ can be straightforwardly obtained by combining the above two lemmas.

Corollary 8 *Let $\{(\mathbf{X}^{(k)}, \mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$ be the iterate sequence generated by Algorithm 1. Suppose that Assumption 1 and Assumption 2 hold. Then for any $k \in \mathbb{N}$, we have*

$$\left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2 \leq C_{pg}^2 \tau^2,$$

where $C_{pg} := \sqrt{2}(C_{gt}^2 + dL_r^2)^{1/2}$ is a positive constant.

Proof This is a direct consequence of Lemma 6 and Lemma 7, and the proof is omitted here. ■

4.2 Consensus and Tracking Errors

This subsection is devoted to building the upper bound of consensus error $\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\|_{\mathrm{F}}^2$ and tracking error $\|\mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)}\|_{\mathrm{F}}^2$. Towards this end, we first show that, based on the following lemma, $\mathrm{Proj}_{\mathcal{M}}(\hat{X}^{(k)})$ is a singleton and $\bar{X}^{(k)} = \mathrm{Proj}_{\mathcal{M}}(\hat{X}^{(k)})$ for any $k \in \mathcal{N}$.

Lemma 9 *Suppose that Assumption 1 and Assumption 2 hold. Let $\{(\mathbf{X}^{(k)}, \mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$ be the iterate sequence generated by Algorithm 1 with*

$$0 < \eta \leq 1, \ \text{and} \ 0 < \tau \leq \min\left\{ \frac{\delta}{4C_{pg}}, \frac{\gamma(\delta(1-\sigma) - \gamma)}{2(\delta - \gamma)C_{pg}} \right\}, \tag{4.4}$$

where $\gamma < \min\{\delta/4, \delta(1-\sigma)\}$ is a positive constant. Then the following relationship holds for any $k \in \mathbb{N}$,

$$\left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}} \leq \gamma.$$

Proof We will use mathematical induction to prove this lemma. The argument $\|\mathbf{X}^{(0)} - \bar{\mathbf{X}}^{(0)}\|_{\mathrm{F}} \leq \gamma$ directly holds resulting from the initialization. Now, we assume that $\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\|_{\mathrm{F}} \leq \gamma$, and investigate the situation of $\|\mathbf{X}^{(k+1)} - \bar{\mathbf{X}}^{(k+1)}\|_{\mathrm{F}}$.

Our first purpose is to show that

$$\sum_{j=1}^{d} W(i,j) X_j^{(k)} + \eta S_i^{(k)} \in \mathcal{R}(\delta/2), \tag{4.5}$$

for any $i \in [d]$. In fact, since $\|X_j^{(k)} - \bar{X}^{(k)}\|_{\mathrm{F}} \leq \|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\|_{\mathrm{F}} \leq \gamma$ for any $j \in [d]$, we have

$$\left\| \sum_{j=1}^{d} W(i,j) X_j^{(k)} - \bar{X}^{(k)} \right\|_{\mathrm{F}} \leq \sum_{j=1}^{d} W(i,j) \left\| X_j^{(k)} - \bar{X}^{(k)} \right\|_{\mathrm{F}} \leq \gamma,$$

and

$$\left\| \hat{X}^{(k)} - \bar{X}^{(k)} \right\|_{\mathrm{F}} \leq \frac{1}{d} \sum_{j=1}^{d} \left\| X_j^{(k)} - \bar{X}^{(k)} \right\|_{\mathrm{F}} \leq \gamma,$$

which indicate that $\sum_{j=1}^{d} W(i,j) X_j^{(k)} \in \mathcal{R}(\gamma)$ and $\hat{X}^{(k)} \in \mathcal{R}(\gamma)$, respectively. Then straightforward calculations yield that

$$\begin{aligned} \left\| \sum_{j=1}^{d} W(i,j) X_j^{(k)} + \eta S_i^{(k)} - \bar{X}^{(k)} \right\|_{\mathrm{F}} \leq{}& \left\| \sum_{j=1}^{d} W(i,j) X_j^{(k)} - \bar{X}^{(k)} \right\|_{\mathrm{F}} + \eta \left\| S_i^{(k)} \right\|_{\mathrm{F}} \\ \leq{}& \gamma + \tau C_{pg} \leq \frac{\delta}{2}. \end{aligned}$$

Hence, the relationship (4.5) holds due to the fact that $\bar{X}^{(k)} \in \mathcal{M}$.

Next, we proceed to prove that $\|\mathbf{X}^{(k+1)} - \bar{\mathbf{X}}^{(k+1)}\|_{\mathrm{F}} \leq \gamma$. For convenience, we denote $Y_i^{(k)} = \mathrm{Proj}_{\mathcal{M}}(\sum_{j=1}^{d} W(i,j) X_j^{(k)})$ and $\mathbf{Y}^{(k)} = [(Y_1^{(k)})^\top, \dots, (Y_d^{(k)})^\top]^\top$ for $i \in [d]$ and $k \in \mathbb{N}$. Since the inclusion (4.5) holds and $\sum_{j=1}^{d} W(i,j) X_j^{(k)} \in \mathcal{R}(\gamma) \subset \mathcal{R}(\delta/2)$, we can deduce from the inequality (2.1) that

$$\begin{aligned} & \left\| \mathbf{X}^{(k+1)} - \mathbf{Y}^{(k)} \right\|_{\mathrm{F}}^2 \\ ={}& \sum_{i=1}^{d} \left\| X_i^{(k+1)} - Y_i^{(k)} \right\|_{\mathrm{F}}^2 \\ ={}& \sum_{i=1}^{d} \left\| \mathrm{Proj}_{\mathcal{M}} \left(\sum_{j=1}^{d} W(i,j) X_j^{(k)} + \eta S_i^{(k)} \right) - \mathrm{Proj}_{\mathcal{M}} \left(\sum_{j=1}^{d} W(i,j) X_j^{(k)} \right) \right\|_{\mathrm{F}}^2 \\ \leq{}& 4\eta^2 \sum_{i=1}^{d} \left\| S_i^{(k)} \right\|_{\mathrm{F}}^2 = 4\eta^2 \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2. \end{aligned} \tag{4.6}$$

Then it can be readily verified that

$$\left\| \mathbf{W}\mathbf{X}^{(k)} - \hat{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 = \left\| (\mathbf{W} - \mathbf{J}) \left(\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right) \right\|_{\mathrm{F}}^2 \leq \sigma^2 \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2.$$

Invoking the inequality (2.1) again with the inclusions $\hat{X}^{(k)} \in \mathcal{R}(\gamma)$ and $\sum_{j=1}^{d} W(i,j)X_j^{(k)} \in \mathcal{R}(\gamma)$, we can further obtain that

$$\begin{aligned}\left\|\mathbf{Y}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 &= \sum_{i=1}^{d}\left\|\mathrm{Proj}_{\mathcal{M}}\left(\sum_{j=1}^{d} W(i,j)X_j^{(k)}\right) - \mathrm{Proj}_{\mathcal{M}}\left(\hat{X}^{(k)}\right)\right\|_{\mathrm{F}}^2 \\ &\leq \left(\frac{\delta}{\delta-\gamma}\right)^2 \sum_{i=1}^{d}\left\|\sum_{j=1}^{d} W(i,j)X_j^{(k)} - \hat{X}^{(k)}\right\|_{\mathrm{F}}^2 \\ &= \left(\frac{\delta}{\delta-\gamma}\right)^2 \left\|\mathbf{W}\mathbf{X}^{(k)} - \hat{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 \leq \frac{\delta^2\sigma^2}{(\delta-\gamma)^2}\left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 .\end{aligned} \tag{4.7}$$

Furthermore, it follows from the inequalities (4.6) and (4.7) that

$$\begin{aligned}\left\|\mathbf{X}^{(k+1)} - \bar{\mathbf{X}}^{(k+1)}\right\|_{\mathrm{F}} &\leq \left\|\mathbf{X}^{(k+1)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}} \\ &\leq \left\|\mathbf{X}^{(k+1)} - \mathbf{Y}^{(k)}\right\|_{\mathrm{F}} + \left\|\mathbf{Y}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}} \\ &\leq 2\eta\left\|\mathbf{S}^{(k)}\right\|_{\mathrm{F}} + \frac{\delta\sigma}{\delta-\gamma}\left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}} \\ &\leq 2\tau C_{pg} + \frac{\gamma\delta\sigma}{\delta-\gamma} \leq \gamma,\end{aligned}$$

where the last inequality holds due to the conditions of τ and γ. The proof is completed. ■

According to the proof of Lemma 9, it follows that the relationship (4.5) holds for any $k \in \mathbb{N}$ in Algorithm 1 under the condition (4.4). Moreover, the relationship $\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\|_{\mathrm{F}} \leq \gamma$ directly implies that the inclusion $\hat{X}^{(k)} \in \mathcal{R}(\gamma)$ is valid. As a consequence, $\mathrm{Proj}_{\mathcal{M}}(\hat{X}^{(k)})$ is a singleton and $\bar{X}^{(k)} = \mathrm{Proj}_{\mathcal{M}}(\hat{X}^{(k)})$. Next, we establish the upper bound of consensus errors.

Lemma 10 *Suppose that Assumption 1 and Assumption 2 hold. Let $\{(\mathbf{X}^{(k)}, \mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$ be the iterate sequence generated by Algorithm 1 with the algorithmic parameters satisfying the condition (4.4). Then for any $k \in \mathbb{N}$, it holds that*

$$\left\|\mathbf{X}^{(k+1)} - \bar{\mathbf{X}}^{(k+1)}\right\|_{\mathrm{F}}^2 \leq \frac{1+\zeta^2}{2}\left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 + \frac{4\eta^2(1+\zeta^2)}{1-\zeta^2}\left\|\mathbf{S}^{(k)}\right\|_{\mathrm{F}}^2,$$

where $\zeta = \delta\sigma/(\delta-\gamma) \in (0,1)$ is a constant.

Proof According to the definition of $\bar{X}^{(k+1)}$ and Young's inequality, it follows that

$$\begin{aligned}\left\|\mathbf{X}^{(k+1)} - \bar{\mathbf{X}}^{(k+1)}\right\|_{\mathrm{F}}^2 \leq \left\|\mathbf{X}^{(k+1)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 &= \left\|\mathbf{X}^{(k+1)} - \mathbf{Y}^{(k)} + \mathbf{Y}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 \\ &\leq \frac{1+\zeta^2}{1-\zeta^2}\left\|\mathbf{X}^{(k+1)} - \mathbf{Y}^{(k)}\right\|_{\mathrm{F}}^2 + \frac{1+\zeta^2}{2\zeta^2}\left\|\mathbf{Y}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 .\end{aligned}$$

Then by virtue of the relationships (4.6) and (4.7), we can attain that

$$\left\|\mathbf{X}^{(k+1)} - \bar{\mathbf{X}}^{(k+1)}\right\|_{\mathrm{F}}^2 \leq \frac{1+\zeta^2}{2} \left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 + \frac{4\eta^2(1+\zeta^2)}{1-\zeta^2} \left\|\mathbf{S}^{(k)}\right\|_{\mathrm{F}}^2,$$

which completes the proof. ■

Eventually, we conclude this subsection by bounding the tracking errors.

Lemma 11 *Suppose that Assumption 1 and Assumption 2 hold. Let* $\{(\mathbf{X}^{(k)}, \mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$ *be the iterate sequence generated by Algorithm 1 with the algorithmic parameters satisfying the condition* (4.4)*. Then for any* $k \in \mathbb{N}$*, it holds that*

$$\begin{aligned}\left\|\mathbf{D}^{(k+1)} - \hat{\mathbf{D}}^{(k+1)}\right\|_{\mathrm{F}}^2 \leq{} & \frac{1+\sigma^2}{2} \left\|\mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)}\right\|_{\mathrm{F}}^2 + \frac{6L_f^2(1+\sigma^2)}{1-\sigma^2} \left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 \\ & + \frac{16\eta^2 L_f^2(1+\sigma^2)}{1-\sigma^2} \left\|\mathbf{S}^{(k)}\right\|_{\mathrm{F}}^2.\end{aligned}$$

Proof By straightforward calculations, we can attain that

$$\begin{aligned}\left\|\mathbf{D}^{(k+1)} - \hat{\mathbf{D}}^{(k+1)}\right\|_{\mathrm{F}}^2 ={} & \left\|(\mathbf{W}-\mathbf{J})\left(\mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)}\right) + (I_{dn}-\mathbf{J})\left(\mathbf{G}^{(k+1)} - \mathbf{G}^{(k)}\right)\right\|_{\mathrm{F}}^2 \\ \leq{} & \frac{1+\sigma^2}{2\sigma^2} \left\|(\mathbf{W}-\mathbf{J})\left(\mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)}\right)\right\|_{\mathrm{F}}^2 \\ & + \frac{1+\sigma^2}{1-\sigma^2} \left\|(I_{dn}-\mathbf{J})\left(\mathbf{G}^{(k+1)} - \mathbf{G}^{(k)}\right)\right\|_{\mathrm{F}}^2 \\ \leq{} & \frac{1+\sigma^2}{2} \left\|\mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)}\right\|_{\mathrm{F}}^2 + \frac{1+\sigma^2}{1-\sigma^2} \left\|\mathbf{G}^{(k+1)} - \mathbf{G}^{(k)}\right\|_{\mathrm{F}}^2.\end{aligned}$$

In light of the Lipschitz continuity of ∇f_i, we have

$$\left\|\mathbf{G}^{(k+1)} - \mathbf{G}^{(k)}\right\|_{\mathrm{F}} \leq L_f \left\|\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}\right\|_{\mathrm{F}}.$$

Moreover, it follows from the relationships (4.6) and (4.7) that

$$\begin{aligned}\left\|\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}\right\|_{\mathrm{F}}^2 \leq{} & 2\left\|\mathbf{X}^{(k+1)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 + 2\left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 \\ \leq{} & 4\left\|\mathbf{X}^{(k+1)} - \mathbf{Y}^{(k)}\right\|_{\mathrm{F}}^2 + 4\left\|\mathbf{Y}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 + 2\left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 \\ \leq{} & 6\left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 + 16\eta^2\left\|\mathbf{S}^{(k)}\right\|_{\mathrm{F}}^2.\end{aligned}$$

The proof is completed by collecting the above three inequalities. ■

4.3 Sufficient Descent Property

In this subsection, we construct a merit function to monitor the progress of Algorithm 1, which equips the function value with consensus and tracking errors. While none of these three terms is guaranteed to decrease along the iterates, a suitable combination of them does satisfy a sufficient descent property. We start from the following technical lemma.

Lemma 12 *Suppose that Assumption 2 holds. Let $\{(\mathbf{X}^{(k)}, \mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$ be the iterate sequence generated by Algorithm 1. Then for any $k \in \mathbb{N}$, we have*

$$\left\| \hat{X}^{(k+1)} - \hat{X}^{(k)} - \eta \hat{S}^{(k)} \right\|_{\mathrm{F}} \leq \frac{8M_{pj} + \sqrt{d} M_{tg}}{d} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 + \frac{2\eta^2 M_{pj}}{d} \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2 .$$

Proof To begin with, straightforward manipulations lead to that

$$\begin{aligned}
& \left\| \hat{X}^{(k+1)} - \hat{X}^{(k)} - \eta \hat{S}^{(k)} \right\|_{\mathrm{F}} \\
\leq {} & \frac{1}{d} \sum_{i=1}^{d} \left\| X_i^{(k+1)} - X_i^{(k)} - \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}} \left(\sum_{j=1}^{d} W(i,j) X_j^{(k)} + \eta S_i^{(k)} - X_i^{(k)} \right) \right\|_{\mathrm{F}} \\
& + \frac{1}{d} \left\| \sum_{i=1}^{d} \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}} \left(\sum_{j=1}^{d} W(i,j) X_j^{(k)} - X_i^{(k)} \right) \right\|_{\mathrm{F}} .
\end{aligned}$$

As a direct consequence of the relationship (2.2), we can proceed to show that

$$\begin{aligned}
& \frac{1}{d} \sum_{i=1}^{d} \left\| X_i^{(k+1)} - X_i^{(k)} - \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}} \left(\sum_{j=1}^{d} W(i,j) X_j^{(k)} + \eta S_i^{(k)} - X_i^{(k)} \right) \right\|_{\mathrm{F}} \\
\leq {} & \frac{M_{pj}}{d} \sum_{i=1}^{d} \left\| \sum_{j=1}^{d} W(i,j) X_j^{(k)} + \eta S_i^{(k)} - X_i^{(k)} \right\|_{\mathrm{F}}^2 \\
\leq {} & \frac{2M_{pj}}{d} \sum_{i=1}^{d} \left\| \sum_{j=1}^{d} W(i,j) X_j^{(k)} - X_i^{(k)} \right\|_{\mathrm{F}}^2 + \frac{2\eta^2 M_{pj}}{d} \sum_{i=1}^{d} \left\| S_i^{(k)} \right\|_{\mathrm{F}}^2 \\
= {} & \frac{2M_{pj}}{d} \left\| \left(\mathbf{W} - I_{dn} \right) \mathbf{X}^{(k)} \right\|_{\mathrm{F}}^2 + \frac{2\eta^2 M_{pj}}{d} \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2 .
\end{aligned}$$

Moreover, it can be readily verified that

$$\left\| \left(\mathbf{W} - I_{dn} \right) \mathbf{X}^{(k)} \right\|_{\mathrm{F}} = \left\| \left(\mathbf{W} - I_{dn} \right) \left(\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right) \right\|_{\mathrm{F}} \leq 2 \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}} .$$

According to Lemma 5.3 in Deng and Hu (2023), we have

$$\frac{1}{d} \left\| \sum_{i=1}^{d} \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}} \left(\sum_{j=1}^{d} W(i,j) X_j^{(k)} - X_i^{(k)} \right) \right\|_{\mathrm{F}} \leq \frac{\sqrt{d} M_{tg}}{d} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 ,$$

where $M_{tg} > 0$ is a constant. Collecting the above four relationships, we can obtain the assertion of this lemma. ■

Corollary 13 *Let $\{(\mathbf{X}^{(k)}, \mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$ be the iterate sequence generated by Algorithm 1 with*

$$0 < \eta \leq 1, \ \text{and} \ 0 < \tau \leq \frac{\sqrt{2d}}{4M_{pj}C_{pg}}. \tag{4.8}$$

Suppose that Assumption 1 and Assumption 2 hold. Then for any $k \in \mathbb{N}$, we have

$$\left\| \hat{X}^{(k+1)} - \hat{X}^{(k)} \right\|_{\mathrm{F}}^2 \leq \frac{4M_l^2(8M_{pj} + \sqrt{d}M_{tg})^2}{d} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 + \frac{4\eta^2}{d} \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2,$$

where M_l is a positive constant defined in (4.1)*.*

Proof Combining Corollary 8 with the condition (4.8) gives rise to that

$$\eta^2 \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2 \leq \frac{d}{8M_{pj}^2}. \tag{4.9}$$

In light of Lemma 12, we have

$$\begin{aligned}
&\left\| \hat{X}^{(k+1)} - \hat{X}^{(k)} \right\|_{\mathrm{F}}^2 \\
\leq{}& 2\left\| \hat{X}^{(k+1)} - \hat{X}^{(k)} - \eta\hat{S}^{(k)} \right\|_{\mathrm{F}}^2 + 2\eta^2 \left\| \hat{S}^{(k)} \right\|_{\mathrm{F}}^2 \\
\leq{}& 2\left(\frac{8M_{pj} + \sqrt{d}M_{tg}}{d} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 + \frac{2\eta^2 M_{pj}}{d} \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2 \right)^2 + \frac{2\eta^2}{d} \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2 \\
\leq{}& \frac{4(8M_{pj} + \sqrt{d}M_{tg})^2}{d^2} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^4 + \frac{16\eta^4 M_{pj}^2}{d^2} \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^4 + \frac{2\eta^2}{d} \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2 \\
\leq{}& \frac{4M_l^2(8M_{pj} + \sqrt{d}M_{tg})^2}{d} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 + \frac{4\eta^2}{d} \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2,
\end{aligned}$$

where the last inequality is valid due to the fact that $\left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 \leq dM_l^2$ and the relationship (4.9). We complete the proof. ■

The following lemma indicates that $\hat{D}^{(k)}$ is an estimate of $\nabla f(\hat{X}^{(k)})$ with the approximation error controlled by the consensus error.

Lemma 14 *Suppose that Assumption 1 and Assumption 2 hold. Let $\{(\mathbf{X}^{(k)}, \mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$ be the iterate sequence generated by Algorithm 1. Then it holds that*

$$\left\| \nabla f(\hat{X}^{(k)}) - \hat{D}^{(k)} \right\|_{\mathrm{F}}^2 \leq \frac{L_f^2}{d} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2.$$

Proof It follows from the definition of $\hat{X}^{(k)}$ that

$$\sum_{i=1}^{d} \left\| X_i^{(k)} - \hat{X}^{(k)} \right\|_{\mathrm{F}}^2 \leq \sum_{i=1}^{d} \left\| X_i^{(k)} - \bar{X}^{(k)} \right\|_{\mathrm{F}}^2 = \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 .$$

Since ∇f_i is Lipschitz continuous with the corresponding Lipschitz constant L_f, we have

$$\begin{aligned} \left\| \nabla f(\hat{X}^{(k)}) - \hat{D}^{(k)} \right\|_{\mathrm{F}}^2 & = \left\| \frac{1}{d} \sum_{i=1}^{d} \left(\nabla f_i(\hat{X}^{(k)}) - \nabla f_i(X_i^{(k)}) \right) \right\|_{\mathrm{F}}^2 \\ & \leq \frac{1}{d} \sum_{i=1}^{d} \left\| \nabla f_i(\hat{X}^{(k)}) - \nabla f_i(X_i^{(k)}) \right\|_{\mathrm{F}}^2 \\ & \leq \frac{L_f^2}{d} \sum_{i=1}^{d} \left\| X_i^{(k)} - \hat{X}^{(k)} \right\|_{\mathrm{F}}^2 \leq \frac{L_f^2}{d} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 . \end{aligned}$$

The proof is completed. ■

Now we can prove a descent inequality for the function f based on the Lipschitz continuity of ∇f.

Proposition 15 *Let $\{(\mathbf{X}^{(k)}, \mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$ be the iterate sequence generated by Algorithm 1 with the algorithmic parameters satisfying the condition* (4.8)*. Suppose that Assumption 1 and Assumption 2 hold. Then, for any $k \in \mathbb{N}$, it holds that*

$$f(\hat{X}^{(k+1)}) \leq f(\hat{X}^{(k)}) + \eta \left\langle \hat{D}^{(k)}, \hat{S}^{(k)} \right\rangle + \frac{C_{fx}}{d} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 + \frac{\eta^2 C_{fs}}{d} \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2 ,$$

where C_{fx} and C_{fs} are two positive constants defined by

$$C_{fx} = 8 M_g M_{pj} + \sqrt{d} M_g M_{tg} + L_f + 3 M_l^2 L_f (8 M_{pj} + \sqrt{d} M_{tg})^2,$$

and

$$C_{fs} = 2 M_g M_{pj} + 3 L_f,$$

respectively.

Proof In view of the Lipschitz continuity of ∇f, we have

$$\begin{aligned} f(\hat{X}^{(k+1)}) & \leq f(\hat{X}^{(k)}) + \left\langle \nabla f(\hat{X}^{(k)}), \hat{X}^{(k+1)} - \hat{X}^{(k)} \right\rangle + \frac{L_f}{2} \left\| \hat{X}^{(k+1)} - \hat{X}^{(k)} \right\|_{\mathrm{F}}^2 \\ & = f(\hat{X}^{(k)}) + \left\langle \nabla f(\hat{X}^{(k)}) - \hat{D}^{(k)}, \hat{X}^{(k+1)} - \hat{X}^{(k)} \right\rangle + \left\langle \hat{D}^{(k)}, \hat{X}^{(k+1)} - \hat{X}^{(k)} \right\rangle \\ & \quad + \frac{L_f}{2} \left\| \hat{X}^{(k+1)} - \hat{X}^{(k)} \right\|_{\mathrm{F}}^2 . \end{aligned}$$

It follows from Young's inequality that

$$\left\langle \nabla f(\hat{X}^{(k)}) - \hat{D}^{(k)}, \hat{X}^{(k+1)} - \hat{X}^{(k)} \right\rangle \leq \frac{1}{L_f} \left\| \nabla f(\hat{X}^{(k)}) - \hat{D}^{(k)} \right\|_{\mathrm{F}}^2 + \frac{L_f}{4} \left\| \hat{X}^{(k+1)} - \hat{X}^{(k)} \right\|_{\mathrm{F}}^2 .$$

As a direct consequence of Lemma 14, we can proceed to show that

$$\left\langle \nabla f(\hat{X}^{(k)}) - \hat{D}^{(k)}, \hat{X}^{(k+1)} - \hat{X}^{(k)} \right\rangle \leq \frac{L_f}{d} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 + \frac{L_f}{4} \left\| \hat{X}^{(k+1)} - \hat{X}^{(k)} \right\|_{\mathrm{F}}^2,$$

which is followed by

$$\begin{aligned} f(\hat{X}^{(k+1)}) \leq f(\hat{X}^{(k)}) &+ \left\langle \hat{D}^{(k)}, \hat{X}^{(k+1)} - \hat{X}^{(k)} \right\rangle + \frac{L_f}{d} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 \\ &+ \frac{3L_f}{4} \left\| \hat{X}^{(k+1)} - \hat{X}^{(k)} \right\|_{\mathrm{F}}^2. \end{aligned}$$

Moreover, it can be readily verified that

$$\begin{aligned} \left\langle \hat{D}^{(k)}, \hat{X}^{(k+1)} - \hat{X}^{(k)} \right\rangle = {} & \left\langle \hat{D}^{(k)}, \hat{X}^{(k+1)} - \hat{X}^{(k)} - \eta \hat{S}^{(k)} \right\rangle + \eta \left\langle \hat{D}^{(k)}, \hat{S}^{(k)} \right\rangle \\ \leq {} & \left\| \hat{D}^{(k)} \right\|_{\mathrm{F}} \left\| \hat{X}^{(k+1)} - \hat{X}^{(k)} - \eta \hat{S}^{(k)} \right\|_{\mathrm{F}} + \eta \left\langle \hat{D}^{(k)}, \hat{S}^{(k)} \right\rangle \\ \leq {} & \frac{8M_g M_{pj} + \sqrt{d} M_g M_{tg}}{d} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 + \frac{2\eta^2 M_g M_{pj}}{d} \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2 \\ & + \eta \left\langle \hat{D}^{(k)}, \hat{S}^{(k)} \right\rangle, \end{aligned}$$

where the last inequality results from Lemma 12 and the relationship (4.2). Combining the above two relationships, we can obtain that

$$\begin{aligned} f(\hat{X}^{(k+1)}) \leq f(\hat{X}^{(k)}) + \eta \left\langle \hat{D}^{(k)}, \hat{S}^{(k)} \right\rangle + \frac{8M_g M_{pj} + \sqrt{d} M_g M_{tg} + L_f}{d} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 \\ + \frac{2\eta^2 M_g M_{pj}}{d} \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2 + \frac{3L_f}{4} \left\| \hat{X}^{(k+1)} - \hat{X}^{(k)} \right\|_{\mathrm{F}}^2, \end{aligned}$$

which together with Corollary 13 yields that

$$f(\hat{X}^{(k+1)}) \leq f(\hat{X}^{(k)}) + \eta \left\langle \hat{D}^{(k)}, \hat{S}^{(k)} \right\rangle + \frac{C_{fx}}{d} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 + \frac{\eta^2 C_{fs}}{d} \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2,$$

as desired. ■

In order to show a similar descent inequality for the function r, we need the following two technical lemmas.

Lemma 16 *Suppose that Assumption 1 holds. Let* $\{(\mathbf{X}^{(k)}, \mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$ *be the iterate sequence generated by Algorithm 1. Then, for any* $k \in \mathbb{N}$*, it holds that*

$$\begin{aligned} \frac{1}{d} \sum_{i=1}^{d} \left\langle D_i^{(k)}, \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}} (\bar{X}^{(k)} - X_i^{(k)}) \right\rangle \leq {} & \frac{1}{4dL_r} \left\| \mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)} \right\|_{\mathrm{F}}^2 \\ & + \frac{L_r + M_g M_{tg}/2}{d} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2. \end{aligned}$$

Proof To begin with, the Young's inequality leads to that

$$\begin{aligned}
&\frac{1}{d}\sum_{i=1}^{d}\left\langle D_i^{(k)} - \hat{D}^{(k)}, \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)})\right\rangle \\
\leq{}& \frac{1}{4dL_r}\sum_{i=1}^{d}\left\|D_i^{(k)} - \hat{D}^{(k)}\right\|_{\mathrm{F}}^2 + \frac{L_r}{d}\sum_{i=1}^{d}\left\|X_i^{(k)} - \bar{X}^{(k)}\right\|_{\mathrm{F}}^2 \\
={}& \frac{1}{4dL_r}\left\|\mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)}\right\|_{\mathrm{F}}^2 + \frac{L_r}{d}\left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2.
\end{aligned}$$

Since $\mathrm{Proj}_{\mathcal{T}_{\bar{X}^{(k)}}}(\cdot)$ is a linear operator, we can obtain that

$$\frac{1}{d}\sum_{i=1}^{d}\left\langle \hat{D}^{(k)}, \mathrm{Proj}_{\mathcal{T}_{\bar{X}^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)})\right\rangle = \left\langle \hat{D}^{(k)}, \mathrm{Proj}_{\mathcal{T}_{\bar{X}^{(k)}}}(\bar{X}^{(k)} - \hat{X}^{(k)})\right\rangle = 0,$$

where the last equality follows from the fact that $\bar{X}^{(k)} - \hat{X}^{(k)} \in \mathcal{N}_{\bar{X}^{(k)}}$. By virtue of the Lipschitz continuity of $\mathrm{Proj}_{\mathcal{T}_X}(\cdot)$ with respect to X, we have

$$\left\|\mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)}) - \mathrm{Proj}_{\mathcal{T}_{\bar{X}^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)})\right\|_{\mathrm{F}} \leq \frac{M_{tg}}{2}\left\|X_i^{(k)} - \bar{X}^{(k)}\right\|_{\mathrm{F}}^2,$$

which further implies that

$$\begin{aligned}
&\left\langle \hat{D}^{(k)}, \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)})\right\rangle \\
={}& \left\langle \hat{D}^{(k)}, \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)}) - \mathrm{Proj}_{\mathcal{T}_{\bar{X}^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)})\right\rangle \\
\leq{}& \left\|\hat{D}^{(k)}\right\|_{\mathrm{F}}\left\|\mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)}) - \mathrm{Proj}_{\mathcal{T}_{\bar{X}^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)})\right\|_{\mathrm{F}} \\
\leq{}& \frac{M_g M_{tg}}{2}\left\|X_i^{(k)} - \bar{X}^{(k)}\right\|_{\mathrm{F}}^2.
\end{aligned}$$

Finally, we can arrive at the conclusion that

$$\begin{aligned}
&\frac{1}{d}\sum_{i=1}^{d}\left\langle D_i^{(k)}, \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)})\right\rangle \\
={}& \frac{1}{d}\sum_{i=1}^{d}\left\langle D_i^{(k)} - \hat{D}^{(k)}, \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)})\right\rangle \\
&+ \frac{1}{d}\sum_{i=1}^{d}\left\langle \hat{D}^{(k)}, \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)})\right\rangle \\
\leq{}& \frac{1}{4dL_r}\left\|\mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)}\right\|_{\mathrm{F}}^2 + \frac{L_r + M_g M_{tg}/2}{d}\left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2,
\end{aligned}$$

which completes the proof. ■

Lemma 17 *Suppose that Assumption 1 holds. Let* $\{(\mathbf{X}^{(k)}, \mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$ *be the iterate sequence generated by Algorithm 1 with* $0 < \eta \leq 1$*. Then, for any* $k \in \mathbb{N}$*, it holds that*

$$\begin{aligned} r(\hat{X}^{(k+1)}) \leq{}& (1-\eta) r(\hat{X}^{(k)}) + \frac{\eta}{d} \sum_{i=1}^{d} r(X_i^{(k)} + S_i^{(k)}) \\ &+ \frac{L_r(8M_{pj} + \sqrt{d} M_{tg})}{d} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 + \frac{2\eta^2 L_r M_{pj}}{d} \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2 . \end{aligned}$$

Proof It follows from the convexity of r and Jensen's inequality that

$$r(\hat{X}^{(k)} + \hat{S}^{(k)}) \leq \frac{1}{d} \sum_{i=1}^{d} r(X_i^{(k)} + S_i^{(k)}),$$

and

$$\begin{aligned} r(\hat{X}^{(k)} + \eta \hat{S}^{(k)}) &= r(\eta(\hat{X}^{(k)} + \hat{S}^{(k)}) + (1-\eta)\hat{X}^{(k)}) \\ &\leq \eta r(\hat{X}^{(k)} + \hat{S}^{(k)}) + (1-\eta) r(\hat{X}^{(k)}). \end{aligned}$$

Then, in light of the Lipschitz continuity of r, we have

$$\begin{aligned} & r(\hat{X}^{(k+1)}) - r(\hat{X}^{(k)} + \eta \hat{S}^{(k)}) \\ \leq{}& L_r \left\| \hat{X}^{(k+1)} - \hat{X}^{(k)} - \eta \hat{S}^{(k)} \right\|_{\mathrm{F}} \\ \leq{}& \frac{L_r(8M_{pj} + \sqrt{d} M_{tg})}{d} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 + \frac{2\eta^2 L_r M_{pj}}{d} \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2 , \end{aligned}$$

where the last inequality follows from Lemma 12. Combining the above three inequalities, we complete the proof. ■

Then the following proposition establishes a descent inequality for the function r.

Proposition 18 *Let* $\{(\mathbf{X}^{(k)}, \mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$ *be the iterate sequence generated by Algorithm 1 with*

$$0 < \eta \leq \min\{1, 2\tau\}. \tag{4.10}$$

Suppose that Assumption 1 and Assumption 2 hold. Then, for any $k \in \mathbb{N}$*, it holds that*

$$\begin{aligned} r(\hat{X}^{(k+1)}) \leq{}& r(\hat{X}^{(k)}) - \eta \left\langle \hat{D}^{(k)}, \hat{S}^{(k)} \right\rangle + \frac{1}{2dL_r} \left\| \mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)} \right\|_{\mathrm{F}}^2 \\ &+ \frac{C_{rx}}{d} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 + \left(\frac{\eta^2 C_{rs}}{d} - \frac{\eta(1 - 2L_r\tau)}{2d\tau} \right) \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2 , \end{aligned}$$

where C_{rx} *and* C_{rs} *are two positive constants defined by*

$$C_{rx} = L_r(9M_{pj} + \sqrt{d} M_{tg} + M_{av} + 1) + M_g M_{tg}/2 + 1,$$

and

$$C_{rs} = 2L_r M_{pj},$$

respectively.

Proof To begin with, taking $S = S_i^{(k)}$ and $S' = \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)})$ in (4.3) yields that

$$g_i^{(k)}(\mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)})) - g_i^{(k)}(S_i^{(k)}) \geq \frac{1}{2\tau}\left\|\mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)}) - S_i^{(k)}\right\|_{\mathrm{F}}^2 \geq 0,$$

which, after a suitable rearrangement, can be equivalently written as

$$\begin{aligned} r(X_i^{(k)} + S_i^{(k)}) \leq{}& r(X_i^{(k)} + \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)})) - \left\langle D_i^{(k)}, S_i^{(k)}\right\rangle - \frac{1}{2\tau}\left\|S_i^{(k)}\right\|_{\mathrm{F}}^2 \\ &+ \left\langle D_i^{(k)}, \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)})\right\rangle + \frac{1}{2\tau}\left\|\mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)})\right\|_{\mathrm{F}}^2. \end{aligned}$$

As a direct consequence of Lemma 16, we can proceed to show that

$$\begin{aligned} \frac{1}{d}\sum_{i=1}^d r(X_i^{(k)} + S_i^{(k)}) \leq{}& \frac{1}{d}\sum_{i=1}^d r(X_i^{(k)} + \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)})) - \frac{1}{d}\sum_{i=1}^d \left\langle D_i^{(k)}, S_i^{(k)}\right\rangle \\ &- \frac{1}{2d\tau}\left\|\mathbf{S}^{(k)}\right\|_{\mathrm{F}}^2 + \frac{1}{4dL_r}\left\|\mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)}\right\|_{\mathrm{F}}^2 \\ &+ \left(\frac{L_r + M_g M_{tg}/2}{d} + \frac{1}{2d\tau}\right)\left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2. \end{aligned}$$

Moreover, the Lipschitz continuity of r gives rise to that

$$\begin{aligned} & r(X_i^{(k)} + \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)})) \\ ={}& r(X_i^{(k)} + \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)})) - r(\bar{X}^{(k)}) \\ &+ r(\bar{X}^{(k)}) - r(\hat{X}^{(k)}) + r(\hat{X}^{(k)}) \\ \leq{}& L_r\left\|\bar{X}^{(k)} - X_i^{(k)} - \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}(\bar{X}^{(k)} - X_i^{(k)})\right\|_{\mathrm{F}} \\ &+ L_r\left\|\bar{X}^{(k)} - \hat{X}^{(k)}\right\|_{\mathrm{F}} + r(\hat{X}^{(k)}) \\ \leq{}& L_r M_{pj}\left\|X_i^{(k)} - \bar{X}^{(k)}\right\|_{\mathrm{F}}^2 + \frac{L_r M_{av}}{d}\left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 + r(\hat{X}^{(k)}), \end{aligned}$$

where the last inequality results from the relationships (2.4) and (2.2). Hence, we can attain that

$$\begin{aligned} & \frac{1}{d}\sum_{i=1}^d r(X_i^{(k)} + S_i^{(k)}) \\ \leq{}& r(\hat{X}^{(k)}) + \left(\frac{L_r(M_{pj} + M_{av} + 1) + M_g M_{tg}/2}{d} + \frac{1}{2d\tau}\right)\left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 \\ &+ \frac{1}{4dL_r}\left\|\mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)}\right\|_{\mathrm{F}}^2 - \frac{1}{d}\sum_{i=1}^d \left\langle D_i^{(k)}, S_i^{(k)}\right\rangle - \frac{1}{2d\tau}\left\|\mathbf{S}^{(k)}\right\|_{\mathrm{F}}^2. \end{aligned}$$

In addition, it can be straightforwardly verified that

$$\begin{aligned}\left\langle \hat{D}^{(k)}, \hat{S}^{(k)} \right\rangle &= \frac{1}{d}\sum_{i=1}^{d}\left\langle \hat{D}^{(k)}, S_i^{(k)} \right\rangle \\ &= \frac{1}{d}\sum_{i=1}^{d}\left\langle \hat{D}^{(k)} - D_i^{(k)}, S_i^{(k)} \right\rangle + \frac{1}{d}\sum_{i=1}^{d}\left\langle D_i^{(k)}, S_i^{(k)} \right\rangle \\ &\leq \frac{1}{4dL_r}\left\| \mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)} \right\|_{\mathrm{F}}^2 + \frac{L_r}{d}\left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2 + \frac{1}{d}\sum_{i=1}^{d}\left\langle D_i^{(k)}, S_i^{(k)} \right\rangle,\end{aligned}$$

which together with the condition (4.10) further results in that

$$\begin{aligned}&\frac{1}{d}\sum_{i=1}^{d} r(X_i^{(k)} + S_i^{(k)}) \\ &\leq r(\hat{X}^{(k)}) + \frac{L_r(M_{pj} + M_{av} + 1) + M_g M_{tg}/2 + 1}{\eta d}\left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 \\ &\quad + \frac{1}{2\eta d L_r}\left\| \mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)} \right\|_{\mathrm{F}}^2 - \left\langle \hat{D}^{(k)}, \hat{S}^{(k)} \right\rangle + \left(\frac{L_r}{d} - \frac{1}{2d\tau} \right)\left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2.\end{aligned}$$

The last thing to do in the proof is to combine the above relationship with Lemma 17. Finally, we can obtain the assertion of this proposition. ■

To facilitate the narrative, we define the following positive constants.

$$\begin{cases} C_x = C_{fx} + C_{rx} + \dfrac{6L_f^2(1+\sigma^2)(2-\sigma^2)}{L_r(1-\sigma^2)^2}, \quad \rho = \dfrac{C_x(3-\zeta^2)}{1-\zeta^2}, \\ \kappa = \dfrac{2-\sigma^2}{L_r(1-\sigma^2)}, \quad C_s = C_{fs} + C_{rs} + \dfrac{4\rho(1+\zeta^2)}{1-\zeta^2}. \end{cases} \tag{4.11}$$

Now we are in the position to introduce the following quantity,

$$\hbar^{(k)} := f(\hat{X}^{(k)}) + r(\hat{X}^{(k)}) + \frac{\rho}{d}\left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 + \frac{\kappa}{d}\left\| \mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)} \right\|_{\mathrm{F}}^2.$$

The sequence $\{\hbar^{(k)}\}$ satisfies a sufficient descent property.

Corollary 19 *Suppose that Assumption 1 and Assumption 2 hold. Let $\{(\mathbf{X}^{(k)}, \mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$ be the iterate sequence generated by Algorithm 1 with*

$$\begin{cases} 0 < \eta < \min\left\{ 1,\, 2\tau,\, \dfrac{1-2L_r\tau}{2\tau C_s} \right\}, \\ 0 < \tau < \min\left\{ \dfrac{1}{2L_r}, \dfrac{\delta}{4C_{pg}}, \dfrac{\gamma(\delta(1-\sigma)-\gamma)}{2(\delta-\gamma)C_{pg}}, \dfrac{\sqrt{2d}}{4M_{pj}C_{pg}} \right\}. \end{cases} \tag{4.12}$$

Then for any $k \in \mathbb{N}$, the following sufficient descent condition holds, which implies that the sequence $\{\hbar^{(k)}\}$ is monotonically non-increasing.

$$\begin{aligned}\hbar^{(k+1)} \leq \hbar^{(k)} - \frac{C_x(1-\zeta^2)}{2d} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 - \frac{1-\sigma^2}{2dL_r} \left\| \mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)} \right\|_{\mathrm{F}}^2 \\ - \left(\frac{\eta(1-2L_r\tau)}{2d\tau} - \frac{\eta^2 C_s}{d} \right) \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2,\end{aligned}$$

where C_x and C_s are two positive constants defined in (4.11).

Proof Collecting Lemma 10, Lemma 11, Proposition 15, Proposition 18 together, we can obtain the assertion of this corollary. ■

The small-stepsize condition prescribed by (4.12) enables the single-step consensus, albeit at the expense of slower progress per iteration. Conversely, the framework of multi-step consensus adopted in Chen et al. (2021) permits a larger stepsize, though it incurs higher communication overheads. According to the discussions in Nedić et al. (2018), in large-scale networks typical of real-world applications, communication latency often becomes the primary bottleneck that dominates the overall process. Under such circumstances, decentralized algorithms based on single-step consensus emerge as the more practical and efficient choice.

4.4 Global Convergence

Based on the sufficient descent property of $\{\hbar^{(k)}\}$, we can finally establish the global convergence guarantee of Algorithm 1 to a stationary point.

Theorem 20 *Let $\{(\mathbf{X}^{(k)}, \mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$ be the iterate sequence generated by Algorithm 1 with the algorithmic parameters satisfying the condition* (4.12). *Suppose that Assumption 1 and Assumption 2 hold. Then $\{\mathbf{X}^{(k)}\}$ has at least one accumulation point. Moreover, for any accumulation point $\mathbf{X}^*$ of $\{\mathbf{X}^{(k)}\}$, there exists a stationary point $\bar{X}^* \in \mathcal{M}$ of the problem* (1.1) *such that $\mathbf{X}^* = (\mathbf{1}_d \otimes I_n)\bar{X}^*$.*

Proof To begin with, it is obvious that $\hat{X}^{(k)} \in \mathrm{conv}(\mathcal{M})$. Then according to the definition of the constant $\underline{h}$, we know that

$$\hbar^{(k)} \geq f(\hat{X}^{(k)}) + r(\hat{X}^{(k)}) \geq \underline{h}.$$

Hence, it follows from Corollary 19 that the sequence $\{\hbar^{(k)}\}$ is convergent and the following relationships hold.

$$\lim_{k\to\infty} \left\| \mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)} \right\|_{\mathrm{F}}^2 = 0, \quad \lim_{k\to\infty} \left\| \mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)} \right\|_{\mathrm{F}}^2 = 0, \quad \lim_{k\to\infty} \left\| \mathbf{S}^{(k)} \right\|_{\mathrm{F}}^2 = 0. \tag{4.13}$$

In addition, in light of Lemma 6 and Corollary 8, we know that the sequence $\{(\mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$ is bounded. And owing to the compactness of $\mathcal{M}$, the sequence $\{\mathbf{X}^{(k)}\}$ is also bounded. Then from the Bolzano-Weierstrass theorem, it follows that the sequence $\{(\mathbf{X}^{(k)}, \mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$

exists an accumulation point, say $(\mathbf{X}^*, \mathbf{D}^*, \mathbf{S}^*)$. The relationships in (4.13) infers that $\mathbf{X}^* = (\mathbf{1}_d \otimes I_n)\bar{X}^*$ for some $\bar{X}^* \in \mathcal{M}$, $\mathbf{D}^* = (\mathbf{1}_d \otimes I_n)\nabla f(\bar{X}^*)$, and $\mathbf{S}^* = 0$.

Now we fix an arbitrary $i \in [d]$. By the optimality condition of the subproblem (3.1), we have

$$0 \in \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}} \left(D_i^{(k)} + \frac{1}{\tau} S_i^{(k)} + \partial r(X_i^{(k)} + S_i^{(k)}) \right), \tag{4.14}$$

which further yields that

$$\mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}} \left(\nabla f(X_i^{(k)} + S_i^{(k)}) - D_i^{(k)} \right) - \frac{1}{\tau} S_i^{(k)} \in \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}} \left(\partial h(X_i^{(k)} + S_i^{(k)}) \right).$$

Consequently, there exists $E_i^{(k)} \in \mathcal{N}_{X_i^{(k)}}$ such that

$$\mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}} \left(\nabla f(X_i^{(k)} + S_i^{(k)}) - D_i^{(k)} \right) - \frac{1}{\tau} S_i^{(k)} + E_i^{(k)} \in \partial h(X_i^{(k)} + S_i^{(k)}).$$

Let $\{(\mathbf{X}^{(k_l)}, \mathbf{D}^{(k_l)}, \mathbf{S}^{(k_l)})\}$ be the subsequence of $\{(\mathbf{X}^{(k)}, \mathbf{D}^{(k)}, \mathbf{S}^{(k)})\}$ converging to the accumulation point $(\mathbf{X}^*, \mathbf{D}^*, \mathbf{S}^*)$. Then we can obtain that

$$\mathrm{Proj}_{\mathcal{T}_{X_i^{(k_l)}}} \left(\nabla f(X_i^{(k_l)} + S_i^{(k_l)}) - D_i^{(k_l)} \right) - \frac{1}{\tau} S_i^{(k_l)} + E_i^{(k_l)} \in \partial h(X_i^{(k_l)} + S_i^{(k_l)}).$$

Without loss of generality, we can assume that $X_i^{(k_l)} + S_i^{(k_l)} \in \mathcal{R}(\beta)$ for any $l \in \mathbb{N}$, where $\beta > 0$ is a constant. According to Proposition 2.1.2 in Clarke (1990), it follows that the set $\{H \mid H \in \partial h(X), X \in \mathcal{R}(\beta)\}$ is bounded, which results in the boundedness of $E_i^{(k_l)}$. Without loss of generality, we can assume that the sequence $\{E_i^{(k_l)}\}$ is convergent. Let E_i^* denote its limiting point. By virtue of the relationships in (4.13), we can attain that

$$\mathrm{Proj}_{\mathcal{T}_{X_i^{(k_l)}}} \left(\nabla f(X_i^{(k_l)} + S_i^{(k_l)}) - D_i^{(k_l)} \right) - \frac{1}{\tau} S_i^{(k_l)} + E_i^{(k_l)} \to E_i^*,$$

and

$$X_i^{(k_l)} + S_i^{(k_l)} \to \bar{X}^*,$$

as $l \to \infty$. Furthermore, since h is Lipschitz continuous, we also have $h(X_i^{(k_l)} + S_i^{(k_l)}) \to h(\bar{X}^*)$ as $l \to \infty$. Then by Remark 1(ii) in Bolte et al. (2014), it holds that

$$E_i^* \in \partial h(\bar{X}^*).$$

It follows from the smoothness of the projection operator $\mathrm{Proj}_{\mathcal{N}_X}$ with respect to X that

$$E_i^{(k_l)} = \mathrm{Proj}_{\mathcal{N}_{X_i^{(k_l)}}} \left(E_i^{(k_l)} \right) \to \mathrm{Proj}_{\mathcal{N}_{\bar{X}^*}} \left(E_i^* \right),$$

as $l \to \infty$. Hence, we have $E_i^* = \mathrm{Proj}_{\mathcal{N}_{\bar{X}^*}} (E_i^*)$ and

$$0 = \mathrm{Proj}_{\mathcal{T}_{\bar{X}^*}} (E_i^*) \in \mathrm{Proj}_{\mathcal{T}_{\bar{X}^*}} \left(\partial h(\bar{X}^*) \right),$$

which indicates that $\bar{X}^*$ is a stationary point of the problem (1.1) and completes the proof. ■

The final task is to derive the iteration complexity of Algorithm 1 to reach an ϵ-stationary point defined in Definition 5, which is accomplished in the following theorem.

Theorem 21 *Under the setting of Theorem 20, Algorithm 1 will reach an ϵ-stationary point of the problem* (1.1) *after at most $O(\epsilon^{-2})$ iterations.*

Proof According to Lemma 14, we can obtain that

$$\begin{aligned}\left\|\nabla f(X_i^{(k)}) - D_i^{(k)}\right\|_{\mathrm{F}}^2 \leq{}& 3\left\|\nabla f(X_i^{(k)}) - \nabla f(\hat{X}^{(k)})\right\|_{\mathrm{F}}^2 + 3\left\|\nabla f(\hat{X}^{(k)}) - \hat{D}^{(k)}\right\|_{\mathrm{F}}^2 \\ &+ 3\left\|\hat{D}^{(k)} - D_i^{(k)}\right\|_{\mathrm{F}}^2 \\ \leq{}& 3L_f^2\left\|X_i^{(k)} - \hat{X}^{(k)}\right\|_{\mathrm{F}}^2 + \frac{3L_f^2}{d}\left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 \\ &+ 3\left\|\mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)}\right\|_{\mathrm{F}}^2 \\ \leq{}& 6L_f^2\left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 + 3\left\|\mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)}\right\|_{\mathrm{F}}^2.\end{aligned}$$

Moreover, it follows from the relationship (4.14) that

$$\mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}\left(\nabla f(X_i^{(k)}) - D_i^{(k)} - \frac{1}{\tau}S_i^{(k)}\right) \in \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}\left(\nabla f(X_i^{(k)}) + \partial r(X_i^{(k)} + S_i^{(k)})\right).$$

For convenience, we define the following quantity,

$$\Delta_i^{(k)} := \mathrm{dist}^2\left(0, \mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}\left(\nabla f(X_i^{(k)}) + \partial r(X_i^{(k)} + S_i^{(k)})\right)\right),$$

for any $i \in [d]$ and $k \in \mathbb{N}$. Then a straightforward verification reveals that

$$\begin{aligned}\Delta_i^{(k)} \leq{}& \left\|\mathrm{Proj}_{\mathcal{T}_{X_i^{(k)}}}\left(\nabla f(X_i^{(k)}) - D_i^{(k)} - \frac{1}{\tau}S_i^{(k)}\right)\right\|_{\mathrm{F}}^2 \\ \leq{}& 2\left\|\nabla f(X_i^{(k)}) - D_i^{(k)}\right\|_{\mathrm{F}}^2 + \frac{2}{\tau^2}\left\|S_i^{(k)}\right\|_{\mathrm{F}}^2 \\ \leq{}& C_{\max}\left(\left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 + \left\|\mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)}\right\|_{\mathrm{F}}^2 + \left\|\mathbf{S}^{(k)}\right\|_{\mathrm{F}}^2\right),\end{aligned}$$

where $C_{\max} = \max\{12L_f^2, 6, 2\tau^{-2}\} > 1$ is a constant. Let $C_{\min} > 0$ be a constant defined as

$$C_{\min} := \min\left\{\frac{C_x(1-\zeta^2)}{2d}, \frac{1-\sigma^2}{2dL_r}, \frac{\eta(1-2\tau L_r - 2\tau\eta C_s)}{2d\tau}\right\}.$$

As a direct consequence of Corollary 19, we can proceed to show that

$$\hbar^{(k)} - \hbar^{(k+1)} \geq C_{\min}\left(\left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}^2 + \left\|\mathbf{D}^{(k)} - \hat{\mathbf{D}}^{(k)}\right\|_{\mathrm{F}}^2 + \left\|\mathbf{S}^{(k)}\right\|_{\mathrm{F}}^2\right) \geq \frac{C_{\min}}{C_{\max}}\Delta_i^{(k)},$$

which further implies that

$$\begin{aligned}\min_{k=0,1,\dots,K-1}\Delta_i^{(k)} \leq{}& \frac{1}{K}\sum_{k=0}^{K-1}\Delta_i^{(k)} \leq \frac{C_{\max}}{C_{\min}K}\sum_{k=0}^{K-1}\left(\hbar^{(k)} - \hbar^{(k+1)}\right) \\ \leq{}& \frac{C_{\max}(\hbar^{(0)} - \hbar^{(K)})}{C_{\min}K} \leq \frac{C_{\max}(\hbar^{(0)} - \underline{h})}{C_{\min}K}.\end{aligned}$$

Here, $\underline{h}$ is a constant defined in (4.1). Similarly, it holds that

$$\min_{k=0,1,\dots,K-1} \max\left\{ \left\| X_i^{(k)} - \bar{X}^{(k)} \right\|_{\mathrm{F}}^2, \left\| S_i^{(k)} \right\|_{\mathrm{F}}^2 \right\} \leq \frac{\hbar^{(0)} - \underline{h}}{C_{\min} K}.$$

Therefore, we can conclude that Algorithm 1 will reach an ϵ-stationary point of the problem (1.1) after at most

$$K = \frac{C_{\max}(\hbar^{(0)} - \underline{h})}{C_{\min}\epsilon^2} = O\left(\frac{1}{\epsilon^2}\right)$$

iterations. The proof is completed. ■

We conclude this section by discussing the dependence of the iteration complexity on the parameter σ, which approaches 1 as the connectivity of communication networks deteriorates. From the condition (4.12), it can be observed that both τ and η are of order $O((1-\sigma)^2)$, which results in that $C_{\max} = O((1-\sigma)^{-4})$ and $C_{\min} = O(1-\sigma)$. As a result, the proof of Theorem 21 reveals that the iteration complexity of Algorithm 1 is $O((1-\sigma)^{-5}\epsilon^{-2})$.

5. Numerical Experiments

In this section, we conduct a comparative evaluation of the numerical performance for DR-ProxGT against state-of-the-art solvers, specifically focusing on the coordinate-independent sparse estimation (CISE) problem (Chen et al., 2010), the sparse PCA problem (Wang et al., 2023), and the sparse oblique-manifold nonnegative matrix factorization (SOMNMF) problem (Guo et al., 2022). The corresponding experiments are performed on a workstation with dual Intel Xeon Gold 6242R CPU processors (at 3.10 GHz×20×2) and 510 GB of RAM under Ubuntu 20.04. For a fair and consistent comparison, all the algorithms under test are implemented in the `Python` language with the communication realized by the `mpi4py` package.

5.1 Comparison on CISE Problems

We first engage in a numerical comparison between DR-ProxGT and the other two benchmark methods, DRSM (Wang et al., 2024) and THANOS (Wang and Liu, 2023a), on the CISE problem (Chen et al., 2010). The CISE model is adept at achieving sparse sufficient dimension reduction while efficiently screening out irrelevant and redundant variables, which is accomplished by solving the following optimization problem on the Stiefel manifold $\mathcal{S}_{n,p} := \{X \in \mathbb{R}^{n\times p} \mid X^\top X = I_p\}$,

$$\min_{X\in\mathcal{S}_{n,p}} \; -\frac{1}{2}\sum_{i=1}^{d} \mathrm{tr}\left(X^\top A_i A_i^\top X\right) + \mu \left\|X\right\|_{2,1}. \tag{5.1}$$

Here, $A_i \in \mathbb{R}^{n\times m_i}$ represents the local data matrix privately owned by agent i, consisting of m_i samples with n features. This model incorporates an $\ell_{2,1}$-norm regularizer, defined as $\left\|X\right\|_{2,1} := \sum_{i=1}^{n} \left\|X(i)\right\|_{\mathrm{F}}$ with $X(i)$ being the i-th row of X, to shrink the corresponding row vectors of irrelevant variables to zero. The parameter $\mu > 0$ modulates the row sparsity level within the model for variable selection. For convenience, we denote $A = [A_1 \; A_2 \; \cdots \; A_d] \in \mathbb{R}^{n\times m}$ as the global data matrix, where $m = m_1 + m_2 + \cdots + m_d$.

We set the algorithmic parameters $\tau = 10$ and $\eta = 1$ in DR-ProxGT. Moreover, THANOS is equipped with the BB stepsizes proposed in Wang and Liu (2022). For each iteration k, the stepsize for DRSM is set to be $0.5/\sqrt{k}$. In our numerical experiments, all the algorithms are started from the same initial points. Given the nonconvex nature of the optimization problem, different solvers may still occasionally return different solutions when starting from a common initial point at random. As suggested in Huang and Wei (2022), to increase the chance that all solvers find the same solution, we construct the initial point from the leading p left singular vectors of A, which can be computed efficiently by DESTINY (Wang and Liu, 2022) under the decentralized setting.

Our investigation encompasses a widely recognized image dataset MNIST[1] in the realm of machine learning research, which contains $m = 60000$ samples with $n = 784$ features. For our testing, the samples of MNIST are uniformly distributed into $d = 16$ agents. The CISE problem (5.1) is tested with $p = 10$ and $\mu = 0.001$. The corresponding numerical experiments are performed across three different networks associated with Metropolis constant edge weight matrices (Shi et al., 2015a), including ring network, grid network, and Erdös-Rényi network.

For our simulation in this case, we collect and record the following two quantities at each iteration as performance metrics.

- Stationarity violation: $\dfrac{1}{d}\left\|\mathbf{S}^{(k)}\right\|_{\mathrm{F}}$.
- Consensus error: $\dfrac{1}{d}\left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_{\mathrm{F}}$.

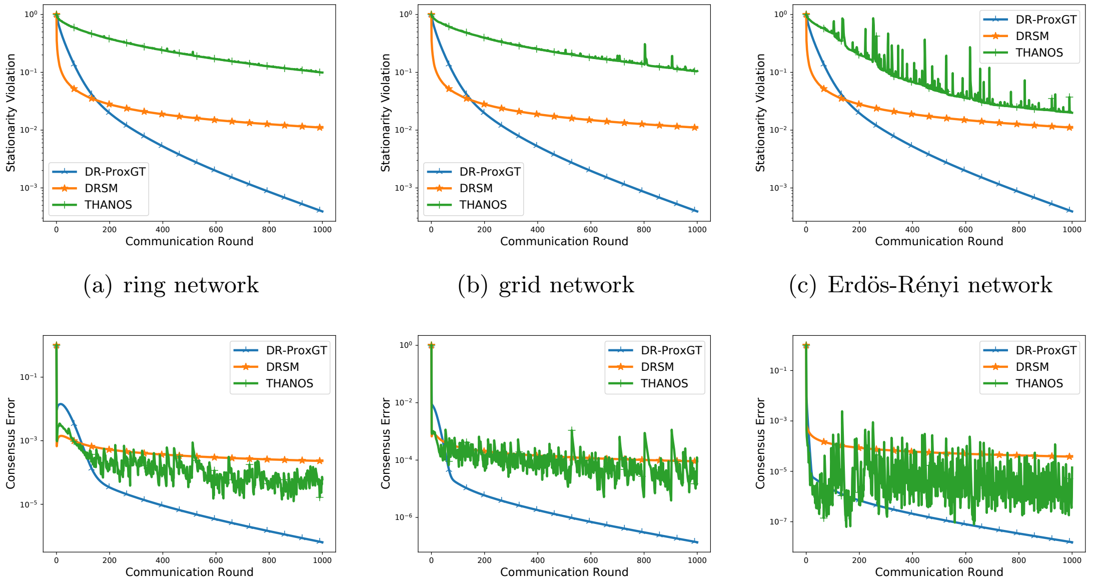
The numerical results of this test are illustrated in Figure 1. Each subfigure, representing a certain structure of networks, depicts the diminishing trend of stationarity violations (top panel) and consensus errors (bottom panel) against the communication rounds on a logarithmic scale. We observe that DR-ProxGT exhibits a superior performance over DRSM and THANOS across all three networks structures. It is worth mentioning that DR-ProxGT appears to converge asymptotically at a linear rate. These observations underscore the effectiveness of DR-ProxGT in achieving both stationarity and consensus rapidly, highlighting its advantage in handling the CISE problem.

5.2 Comparison on Sparse PCA Problems

The next experiment is to evaluate the numerical performance of DR-ProxGT, DRSM, and THANOS on the Global Health Estimates, which are available online[2]. The data were compiled from various sources, including national vital statistics, WHO technical programs, UN partners, the Global Burden of Disease, and scientific research. Our illustrative dataset contains mortality rates across $n = 19$ different age groups for $t = 20$ years (2000–2019) from $d = 50$ European countries, which is structured into a matrix $B \in \mathbb{R}^{n\times dt}$. This situation naturally lends itself to decentralized computations, as it involves data that is privately

[1] http://yann.lecun.com/exdb/mnist/

[2] https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates



(a) ring network (b) grid network (c) Erdös-Rényi network

(d) ring network (e) grid network (f) Erdös-Rényi network

Figure 1: Numerical comparison of DR-ProxGT, DRSM, and THANOS on CISE problems across different structures of networks. The subfigures in the first and second rows depict stationarity violations and consensus errors, respectively.

held within each country. Such an approach is crucial for preserving the privacy of sensitive health information.

A key aspect of this study involves forecasting future mortality rates based on the Lee–Carter model (Lee and Carter, 1992; Basellini et al., 2023), where extracting the principal components of matrix B is a crucial step. In this context, we focus on solving the sparse PCA problem (Wang et al., 2023) with the mortality dataset, aiming to enhance the interpretability and remediate the inconsistency issue. The sparse PCA problem can be formulated as the following optimization model,

$$\min_{X\in\mathcal{S}_{n,p}} \quad -\frac{1}{2}\sum_{i=1}^{d} \operatorname{tr}\left(X^\top B_i B_i^\top X\right) + \lambda \left\|X\right\|_1 . \tag{5.2}$$

Here, $B_i \in \mathbb{R}^{n\times t}$ contains the mortality rates collected from the i-th country. The ℓ_1-norm regularizer $\left\|X\right\|_1 := \sum_{i=1}^{n}\sum_{j=1}^{p} \left|X(i,j)\right|$ is imposed to promote sparsity in X. The parameter $\lambda > 0$ is used to control the amount of sparseness.

In the subsequent experiment, we adopt the same stepsize strategies that are detailed in Subsection 5.1 for the tested algorithms. The initial guesses are also constructed by singular value vectors. We fix the number of principal components to be extracted at $p = 5$ and vary the parameter λ among the values $\{0.05, 0.1, 0.15\}$ in the problem (5.2). The three algorithms under test are configured to perform 600 rounds of communication on an Erdös-Rényi network. Numerical results from this experiment are presented in Table

2 with function values, sparsity levels, and consensus errors recorded. When determining the sparsity level of a solution matrix (i.e., the percentage of zero elements), we consider a matrix element to be zero if its absolute value is less than 1e-5. These results clearly demonstrate that DR-ProxGT outperforms DRSM and THANOS across all performance metrics by substantial margins. This finding is particularly noteworthy as it indicates that the superior performance of DR-ProxGT is not confined to simulated cases, but also extends to real-world applications.

Table 2: Numerical comparison of DR-ProxGT, DRSM, and THANOS on sparse PCA problems for different values of λ.

	Algorithm	Function Value	Sparsity Level	Consensus Error
$\lambda = 0.05$	DR-ProxGT	-7.76	26.32%	2.26e-04
	DRSM	-7.52	21.05%	2.35e-04
	THANOS	-7.65	24.21%	4.96e-04
$\lambda = 0.1$	DR-ProxGT	-7.15	50.53%	3.41e-05
	DRSM	-6.98	30.53%	2.68e-04
	THANOS	-7.02	31.58%	9.26e-05
$\lambda = 0.15$	DR-ProxGT	-6.54	64.21%	4.12e-05
	DRSM	-6.32	32.63%	2.51e-04
	THANOS	-6.47	34.74%	8.25e-05

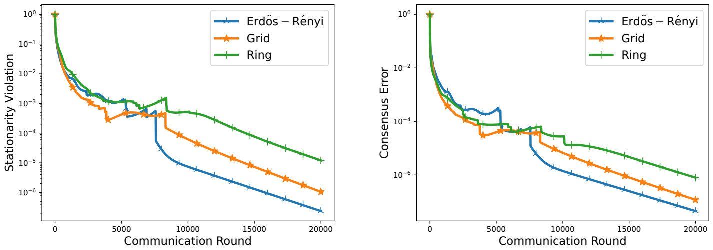
5.3 Performance on SOMNMF Problems

Let diag(Y) be the diagonal matrix whose diagonal elements are those of a square matrix Y. Based on the Hadamard parametrization, Guo et al. (2022) propose the following SOMNMF model on the oblique manifold $\mathcal{O}_{n,p} := \{X \in \mathbb{R}^{n\times p} \mid \mathrm{diag}(X^\top X) = I_p\}$,

$$\min_{X\in\mathcal{O}_{n,p}} \quad \frac{1}{4}\sum_{i=1}^{d} \|P_i - Q_i\,(X \odot X)\|_{\mathrm{F}}^2 + \theta\,\|X\|_1 . \tag{5.3}$$

Here, $P_i \in \mathbb{R}^{m_i\times p}$ and $Q_i \in \mathbb{R}^{m_i\times n}$ contain the local data privately owned by agent i, the notation $\odot$ represents the Hadamard product, and $\theta > 0$ is a constant. Since both DRSM and THANOS are only applicable to optimization problems on the Stiefel manifold, we restrict our evaluation to the performance of DR-ProxGT on the problem (5.3) in this subsection.

For our numerical experiments, the local data matrices P_i and Q_i are randomly generated from the normal distribution. And the SOMNMF problem (5.3) is evaluated with $d = 8$, $m_i = 100$, $n = 5$, $p = 100$ and $\theta = 0.01$. We set the algorithmic parameters $\tau = 0.05$ and $\eta = 1$ in DR-ProxGT, which is terminated after 20000 iterations. Figure 2 depicts the numerical performance of DR-ProxGT across three distinct networks, including ring network, grid network, and Erdös-Rényi network. The numerical results demonstrate that our algorithm manifests its efficiency in solving the SOMNMF problem on the oblique manifold. Moreover, as network connectivity decreases, a degradation in performance occurs, which is consistent with our theoretical analysis.



(a) stationarity violation

(b) consensus error

Figure 2: Numerical performance of DR-ProxGT on SOMNMF problems across different structures of networks.

6. Concluding Remarks

The composite optimization problems on Riemannian manifolds, albeit frequently encountered in many applications, present considerable challenges when addressed under the decentralized setting, primarily due to their intrinsic nonsmoothness and nonconvexity. Current algorithms typically resort to subgradient methods or construct a smooth approximation of the objective function, which often fall short in terms of communication-efficiency. To address this issue, we propose a novel algorithm DR-ProxGT that incorporates the gradient tracking technique into the framework of Riemannian proximal gradient methods. We have rigorously demonstrated that DR-ProxGT achieves global convergence to a stationary point of the problem (1.1) under mild conditions. Furthermore, we have established the iteration complexity of $O(\epsilon^{-2})$, a notable improvement over existing results in the literature. The numerical experiments illustrate that DR-ProxGT significantly surpasses existing algorithms, as evidenced by tests on CISE, sparse PCA, and SOMNMF problems.

Although the potential of DR-ProxGT has already emerged, several intriguing issues remain worth exploring. For instance, we could extend our algorithm to the stochastic and online settings (Wang et al., 2022) to accommodate more general scenarios. Additionally, investigating its generalization to asynchronous environments could further enhance the communication efficiency.

Acknowledgments

The work of Xin Liu was supported in part by the National Key R&D Program of China (2023YFA1009300), National Natural Science Foundation of China (12125108, 12021001, 12288201), and RGC grant JLFS/P-501/24 for the CAS AMSS-PolyU Joint Laboratory in Applied Mathematics. The work of Lei Wang and Le Bao was supported in part by the

World Health Organization and NIH/NIAID R01AI136664 and R01AI170249. This research was conducted while Lei Wang was employed at the Pennsylvania State University.

References

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

D. Bajovic, D. Jakovetic, N. Krejic, and N. K. Jerinkic. Newton-like method with diagonal correction for distributed optimization. *SIAM Journal on Optimization*, 27(2):1171–1203, 2017.

M. V. Balashov and R. A. Kamalov. The gradient projection method with Armijo's step size on manifolds. *Computational Mathematics and Mathematical Physics*, 61:1776–1786, 2021.

U. Basellini, C. G. Camarda, and H. Booth. Thirty years on: A review of the Lee-Carter method for forecasting mortality. *International Journal of Forecasting*, 39(3):1033–1049, 2023.

J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.

N. Boumal and P.-A. Absil. Low-rank matrix completion via preconditioned optimization on the Grassmann manifold. *Linear Algebra and its Applications*, 475:200–239, 2015.

T.-H. Chang, M. Hong, and X. Wang. Multi-agent distributed optimization via inexact consensus ADMM. *IEEE Transactions on Signal Processing*, 63(2):482–497, 2015.

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 Processing Magazine*, 37(3):26–38, 2020.

J. Chen, H. Ye, M. Wang, T. Huang, G. Dai, I. Tsang, and Y. Liu. Decentralized Riemannian conjugate gradient method on the Stiefel manifold. In *Proceedings of the Twelfth International Conference on Learning Representations*, pages 1–12, 2024.

S. Chen, S. Ma, A. M.-C. So, and T. Zhang. Proximal gradient method for nonsmooth optimization over the Stiefel manifold. *SIAM Journal on Optimization*, 30(1):210–239, 2020.

S. Chen, A. Garcia, M. Hong, and S. Shahrampour. Decentralized Riemannian gradient descent on the Stiefel manifold. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 1594–1605. PMLR, 2021.

X. Chen, C. Zou, and R. D. Cook. Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics*, 38(6):3696–3723, 2010.

M. Cho and J. Lee. Riemannian approach to batch normalization. *Advances in Neural Information Processing Systems*, 30, 2017.

F. H. Clarke. *Optimization and Nonsmooth Analysis.* Society for Industrial and Applied Mathematics, Philadelphia, 1990.

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

A. Daneshmand, G. Scutari, and V. Kungurtsev. Second-order guarantees of distributed gradient algorithms. *SIAM Journal on Optimization*, 30(4):3029–3068, 2020.

A. Daneshmand, G. Scutari, P. Dvurechensky, and A. Gasnikov. Newton method over networks is fast up to the statistical precision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 2398–2409. PMLR, 2021.

A. d'Aspremont, F. Bach, and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9(42):1269–1294, 2008.

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

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

S. Gao and Z. Ma. Sparse GCA and thresholded gradient descent. *Journal of Machine Learning Research*, 24(135):1–61, 2023.

B. Gharesifard and J. Cortés. Distributed strategies for generating weight-balanced and doubly stochastic digraphs. *European Journal of Control*, 18(6):539–557, 2012.

Z. Guo, A. Min, B. Yang, J. Chen, H. Li, and J. Gao. A sparse oblique-manifold nonnegative matrix factorization for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

D. Hajinezhad and M. Hong. Perturbed proximal primal-dual algorithm for nonconvex nonsmooth optimization. *Mathematical Programming*, 176(1):207–245, 2019.

J. Hu and K. Deng. Improving the communication in decentralized manifold optimization through single-step consensus and compression. *arXiv:2407.08904*, 2024.

W. Huang and K. Wei. Riemannian proximal gradient methods. *Mathematical Programming*, 194(1-2):371–413, 2022.

I. T. Jolliffe, N. T. Trendafilov, and M. Uddin. A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003.

M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(15): 517–553, 2010.

R. D. Lee and L. R. Carter. Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87(419):659–671, 1992.

X. Li, S. Chen, Z. Deng, Q. Qu, Z. Zhu, and A. M.-C. So. Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods. *SIAM Journal on Optimization*, 31(3):1605–1634, 2021.

Z. Li, W. Shi, and M. Yan. A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. *IEEE Transactions on Signal Processing*, 67(17):4494–4506, 2019.

Q. Ling, W. Shi, G. Wu, and A. Ribeiro. DLM: Decentralized linearized alternating direction method of multipliers. *IEEE Transactions on Signal Processing*, 63(15):4051–4064, 2015.

X. Liu, N. Xiao, and Y.-X. Yuan. A penalty-free infeasible approach for a class of nonsmooth optimization problems over the Stiefel manifold. *Journal of Scientific Computing*, 99(2): 30, 2024.

B. Mishra, H. Kasai, P. Jawanpuria, and A. Saroop. A Riemannian gossip approach to subspace learning on Grassmann manifold. *Machine Learning*, 108:1783–1803, 2019.

A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

A. Nedić, A. Olshevsky, and W. Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

A. Nedić, A. Olshevsky, and M. G. Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5): 953–976, 2018.

S. U. Pillai, T. Suel, and S. Cha. The Perron-Frobenius theorem: Some of its applications. *IEEE Signal Processing Magazine*, 22(2):62–75, 2005.

G. Qu and N. Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

A. Sarlette and R. Sepulchre. Consensus optimization on manifolds. *SIAM Journal on Control and Optimization*, 48(1):56–76, 2009.

G. Scutari and Y. Sun. Distributed nonconvex constrained optimization over time-varying digraphs. *Mathematical Programming*, 176(1):497–544, 2019.

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

W. Shi, Q. Ling, G. Wu, and W. Yin. A proximal gradient algorithm for decentralized composite optimization. *IEEE Transactions on Signal Processing*, 63(22):6013–6023, 2015b.

Z. Song, L. Shi, S. Pu, and M. Yan. Optimal gradient tracking for decentralized optimization. *Mathematical Programming*, 207(1):1–53, 2024.

Y. Sun, G. Scutari, and A. Daneshmand. Distributed optimization based on gradient tracking revisited: Enhancing convergence rate via surrogation. *SIAM Journal on Optimization*, 32(2):354–385, 2022.

B. Wang, S. Ma, and L. Xue. Riemannian stochastic proximal gradient methods for nonsmooth optimization over the Stiefel manifold. *Journal of Machine Learning Research*, 23 (106):1–33, 2022.

J. Wang, J. Hu, S. Chen, Z. Deng, and A. M.-C. So. Decentralized non-smooth optimization over the Stiefel manifold. In *Proceedings of the IEEE 13rd Sensor Array and Multichannel Signal Processing Workshop*, pages 1–5, 2024.

L. Wang and X. Liu. Decentralized optimization over the Stiefel manifold by an approximate augmented Lagrangian function. *IEEE Transactions on Signal Processing*, 70:3029–3041, 2022.

L. Wang and X. Liu. Smoothing gradient tracking for decentralized optimization over the Stiefel manifold with non-smooth regularizers. In *Proceedings of the 62nd IEEE Conference on Decision and Control*, pages 126–132. IEEE, 2023a.

L. Wang and X. Liu. A variance-reduced stochastic gradient tracking algorithm for decentralized optimization with orthogonality constraints. *Journal of Industrial and Management Optimization*, 19(10):7753–7776, 2023b.

L. Wang, X. Liu, and Y. Zhang. A communication-efficient and privacy-aware distributed algorithm for sparse PCA. *Computational Optimization and Applications*, 85(3):1033–1072, 2023.

L. Wang, X. Liu, and X. Chen. The distributionally robust optimization model of sparse principal component analysis. *arXiv:2503.02494*, 2025.

L. Xiao and S. Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.

R. Xin, S. Pu, A. Nedić, and U. A. Khan. A general framework for decentralized optimization with first-order methods. *Proceedings of the IEEE*, 108(11):1869–1889, 2020.

R. Xin, S. Das, U. A. Khan, and S. Kar. A stochastic proximal gradient framework for decentralized non-convex composite optimization: Topology-independent sample complexity and communication efficiency. *arXiv:2110.01594*, 2021.

J. Xu, S. Zhu, Y. C. Soh, and L. Xie. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *Proceedings of the 54th IEEE Conference on Decision and Control*, pages 2055–2060. IEEE, 2015.

Y. Yan, J. Chen, P.-Y. Chen, X. Cui, S. Lu, and Y. Xu. Compressed decentralized proximal stochastic gradient method for nonconvex composite problems with heterogeneous data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 39035–39061. PMLR, 2023.

W. H. Yang, L.-H. Zhang, and R. Song. Optimality conditions for the nonlinear programming problems on Riemannian manifolds. *Pacific Journal of Optimization*, 10(2):415–434, 2014.

H. Ye and T. Zhang. DeEPCA: Decentralized exact PCA with linear convergence rate. *Journal of Machine Learning Research*, 22(238):1–27, 2021.

K. Yuan, Q. Ling, and W. Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

J. Zeng and W. Yin. On nonconvex decentralized gradient descent. *IEEE Transactions on Signal Processing*, 66(11):2834–2848, 2018.

Y. Zhai, Z. Yang, Z. Liao, J. Wright, and Y. Ma. Complete dictionary learning via ℓ_4-norm maximization over the orthogonal group. *Journal of Machine Learning Research*, 21(165): 1–68, 2020.

J. Zhang, Q. Ling, and A. M.-C. So. A Newton tracking algorithm with exact linear convergence for decentralized consensus optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 7:346–358, 2021.

M. Zhu and S. Martínez. Discrete-time dynamic average consensus. *Automatica*, 46(2): 322–329, 2010.

Z. Zhu, T. Ding, D. Robinson, M. Tsakiris, and R. Vidal. A linearly convergent method for non-smooth non-convex optimization on the Grassmannian with applications to robust subspace and dictionary learning. *Advances in Neural Information Processing Systems*, 32, 2019.