Decentralized Non-Smooth Optimization Over the Stiefel Manifold

Jinxin Wang
Department of Syst. Eng. and Eng. Manag.
CUHK, Hong Kong SAR
jxwang@se.cuhk.edu.hk

Jiang Hu
Mass General Hospital and Harvard Medical School
Harvard University, USA
hujiangopt@gmail.com

Shixiang Chen
School of Mathematical Sciences
USTC, China
shxchen@ustc.edu.cn

Zengde Deng
ByteDance
China
dengzengde@gmail.com

Anthony Man-Cho So
Department of Syst. Eng. and Eng. Manag.
CUHK, Hong Kong SAR
manchoso@se.cuhk.edu.hk

***Abstract*—We focus on a class of non-smooth optimization problems over the Stiefel manifold in the decentralized setting, where a connected network of many agents cooperatively minimize a finite-sum objective function with each component being weakly convex in the ambient Euclidean space. Such optimization problems, albeit frequently encountered in applications, are quite challenging due to their non-smoothness and non-convexity. To tackle them, we propose an iterative method called the decentralized Riemannian subgradient method (DRSM). When the problem at hand possesses a sharpness property, we show the local linear convergence of DRSM using geometrically diminishing stepsizes. Numerical experiments are conducted to demonstrate the superior performance of DRSM in different applications.**

***Index Terms*—decentralized non-smooth optimization, Stiefel manifold, Riemannian subgradient method, sharpness**

I. Introduction

Decentralized optimization has gained more and more attention during the past decades in various fields ranging from machine learning to control [1]–[3]. The decentralized network operates differently from a centralized network as it does not require a central server, offering several advantages. First, eliminating the server as an intermediate step results in significant savings in communication resources. With no unified coordination and configuration of servers, the communication network among clients becomes more diverse, allowing for more flexible and efficient communication structures—an essential advantage over centralized networks. Second, by removing the central server, the decentralized network eradicates a single point of failure, potentially increasing the robustness and reliability of the learning system. Despite the advantages of the decentralized setting, algorithms for non-smooth optimization problems with non-convex manifold constraints remain largely unexplored, which motivates us to delve into and address this gap in our research.

In this paper, we consider the following problem of weakly convex (possibly non-smooth) optimization over the Stiefel manifold in a decentralized (i.e., multi-agent) manner:

$$\min\ f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \qquad \text{s.t.} \quad x \in \mathcal{M}. \tag{1}$$

Here, each local component $f_i : \mathbb{R}^{d\times r} \to \mathbb{R}$ ($i \in [n] := \{1,\ldots,n\}$) is assumed to be ρ-weakly convex in the ambient Euclidean space $\mathbb{R}^{d\times r}$ (recall that $g(\cdot)$ is ρ-weakly convex if $g(\cdot) + \frac{\rho}{2}\|\cdot\|_F^2$ is convex for some constant $\rho \geq 0$) and $\mathcal{M} := \mathrm{St}(d,r) = \left\{x \in \mathbb{R}^{d\times r} : d \geq r, x^\top x = I_r\right\}$ is the Stiefel manifold.

In the multi-agent setting, there is a connected undirected communication network represented by the graph $\mathcal{G}$. Each node of $\mathcal{G}$ corresponds to an agent, and the network of n agents aim to collectively solve (1). The i-th agent holds a local copy x_i of the variable x in (1). Let $\mathcal{N}_i$ be the neighborhood of i including itself. For any $i \in [n]$ and $j \in \mathcal{N}_i$, the equality constraint $x_i = x_j$ is required. As $\mathcal{G}$ is connected, we have the consensus constraint $x_1 = x_2 = \cdots = x_n$. Then, an equivalent reformulation of (1) is

$$\begin{aligned} &\min\ f(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(x_i) \\ &\text{s.t. } x_1 = x_2 = \cdots = x_n,\ x_i \in \mathcal{M}, \quad \forall i \in [n], \end{aligned} \tag{2}$$

where the variable $\boldsymbol{x}^\top := (x_1^\top, x_2^\top, \ldots, x_n^\top) \in \mathbb{R}^{r\times nd}$. It is worth noting that various machine learning applications, such as decentralized robust subspace recovery and decentralized dictionary learning, can be captured by the formulation (2) [4]–[7].

To obtain the consensual optimal solution to problem (2), each node of $\mathcal{G}$ needs to mix its local decision variable with its immediate neighbors according to predefined weights. We introduce a mixing matrix $W \in \mathbb{R}^{n\times n}$ to model the mixing process, whose (i,j)-th entry $W_{ij} \geq 0$ represents the weight assigned to node j by node i. The following assumption on W is commonly used in decentralized learning.

Assumption 1.1: The mixing matrix $W \in \mathbb{R}^{n\times n}$ is symmetric and doubly stochastic, that is, $W = W^\top, W \geq 0, \sum_{j\in[n]} W_{ij} = 1$ for all $i \in [n]$. Moreover, we have $W_{ij} = 0$ if and only if $j \notin \mathcal{N}_i$.

The first and second authors contributed equally to this paper.

An immediate consequence of the Perron-Frobenius theorem [8] is that the eigenvalues of W lie in $(-1, 1]$. In addition, the second-largest singular value σ_2 of W satisfies $\sigma_2 \in [0, 1)$.

A. Related work

Problem (2), a weakly convex optimization problem over the Stiefel manifold, can be non-smooth and non-convex, rendering it quite challenging to solve. If the Stiefel manifold constraint is absent in problem (2), decentralized (sub)gradient methods were studied in [1], [9]–[12] and a distributed dual averaging subgradient method was proposed in [13], [14]. During the past few years, there have been significant efforts in designing decentralized algorithms for smooth optimization over the Stiefel manifold [15]–[17]. Specifically, the work [15] developed a decentralized version of the Riemannian gradient method, the work [16] studied a decentralized power method for solving the distributed principal component analysis problem, and the work [17] proposed a decentralized augmented Lagrangian method. The related works are summarized in Table I. By sharp contrast, the study of the general decentralized non-smooth non-convex problem (2) is still in its infancy. The work closest to ours is [18], which established a convergence guarantee for the Riemannian subgradient method when solving the *centralized* counterpart (1).

TABLE I
COMPARISON WITH RELATED WORKS. "S.T." MEANS "SUBJECT TO", WHICH INDICATES THE FEASIBLE REGION.

f_i s.t.	L-smooth, non-convex	Non-smooth, weakly convex
$\mathbb{R}^{d\times r}$	DGD [11]	DPSM [12]
$\mathcal{M}$	DRSGD/DRGTA [15] DESTINY [17]	**DRSM** (this paper)

B. Our contribution

In this paper, we propose the *decentralized* Riemannian subgradient method (DRSM) for solving *non-smooth* weakly convex optimization problems over the Stiefel manifold of the form (2). To the best of our knowledge, this is *the first* work to propose a decentralized method for solving problem (2). We show the *local linear* convergence rate of DRSM using geometrically diminishing stepsizes under the regularity condition of sharpness. Numerical experiments are conducted to demonstrate the superior performance of DRSM in different applications.

II. PRELIMINARIES

A. Notation

We use $\otimes$ to denote the Kronecker product. Given a vector x, we use $\|x\|_2$ and $\|x\|_1$ to denote its Euclidean norm and ℓ_1-norm, respectively. Given a matrix x, we use $\|x\|_F$ to denote its Frobenius norm and $\|\boldsymbol{x}\|_{F,\infty}$ with $\boldsymbol{x}^\top := (x_1^\top, x_2^\top, \dots, x_n^\top) \in \mathbb{R}^{r\times nd}$ to denote the norm $\max_{i\in[n]} \|x_i\|_F$. We denote by 1_n the n-dimensional all-one vector. Given a symmetric matrix $W \in \mathbb{R}^{n\times n}$, we use $\lambda_2(W)$ and $\lambda_n(W)$ to denote its second-largest eigenvalue and smallest eigenvalue, respectively. For a nonempty closed set $\mathcal{X}$, we use $\operatorname{dist}(x, \mathcal{X}) := \inf_{y\in\mathcal{X}} \|x - y\|_F$ to denote the distance between a point x and $\mathcal{X}$. We denote the Euclidean average of the points $x_1, \dots, x_n \in \mathbb{R}^{d\times r}$ as $\hat{x} := \frac{1}{n}\sum_{i=1}^n x_i$.

B. Consensus on the Stiefel manifold

The consensus problem over the Stiefel manifold $\mathcal{M}$ is to minimize the weighted squared distance among all local variables, which can be formulated as

$$\begin{aligned} \min\ & \varphi^t(\boldsymbol{x}) := \frac{1}{4}\sum_{i=1}^n \sum_{j=1}^n W_{ij}^t \|x_i - x_j\|_F^2 \\ \text{s.t. } & x_i \in \mathcal{M}, \quad \forall i \in [n]. \end{aligned} \tag{C-St}$$

Here, W^t denotes the t-th power of the doubly stochastic matrix W with $t \geq 1$ being an integer. In the k-th ($k = 0, 1, \dots$) iteration, the Riemannian gradient method DRCS proposed in [19] for problem (C-St) is given by

$$\begin{aligned} x_{i,k+1} &= \mathcal{R}_{x_{i,k}}(-\alpha \operatorname{grad} \varphi_i^t(\boldsymbol{x}_k)) \\ &= \mathcal{R}_{x_{i,k}}(\alpha \mathcal{P}_{\mathrm{T}_{x_{i,k}}\mathcal{M}}(\sum_{j=1}^n W_{ij}^t x_{j,k})). \end{aligned} \tag{DRCS}$$

Here, $\operatorname{grad} \varphi_i^t(\boldsymbol{x}_k) \in \mathbb{R}^{d\times r}$ represents the Riemannian gradient of φ^t with respect to $x_{i,k}$, $\mathrm{T}_x\mathcal{M}$ is the tangent space to $\mathcal{M}$ at x, $\mathcal{P}_{\mathrm{T}_x\mathcal{M}}(y)$ is the projection of $y \in \mathbb{R}^{d\times r}$ onto $\mathrm{T}_x\mathcal{M}$, and $\mathcal{R}_{x_{i,k}}(\cdot)$ is a retraction operator. We refer the reader to [20]–[22] for an introduction to manifold optimization. In the sequel, we only use the polar decomposition-based retraction to simplify theoretical analysis. Although problem (C-St) is non-convex, it has been shown in [19] that DRCS converges Q-linearly in a local region. Specifically, define

$$\begin{aligned} &\mathcal{N} := \mathcal{N}_1 \cap \mathcal{N}_2,\ \mathcal{N}_1 := \left\{\boldsymbol{x} : \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_{\mathrm{F}}^2 \leq n\delta_1^2\right\}, \\ &\mathcal{N}_2 := \{\boldsymbol{x} : \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_{\mathrm{F},\infty} \leq \delta_2\}, \end{aligned} \tag{3}$$

where $\bar{x} := \mathcal{P}_{\mathcal{M}}(\hat{x}) \in \arg\min_{y\in\mathcal{M}} \sum_{i=1}^n \|y - x_i\|_F^2$ is the induced arithmetic mean (IMA) on the Stiefel manifold, $\mathcal{P}_{\mathcal{M}}(\hat{x})$ represents any point that is the projection of $\hat{x}$ onto $\mathcal{M}$, $\bar{\boldsymbol{x}} := 1_n \otimes \bar{x}$, and δ_1, δ_2 satisfy $\delta_1 \leq \frac{1}{5\sqrt{r}}\delta_2$ and $\delta_2 \leq \frac{1}{6}$. We have the following local linear convergence result [19]:

Fact 2.1: Suppose that Assumption 1.1 holds. Let the stepsize α satisfy $0 < \alpha \leq \bar{\alpha} := \min\{\nu\frac{\Phi}{L_t}, 1, \frac{1}{M}\}$ and $t \geq \lceil \log_{\sigma_2}(\frac{1}{2\sqrt{n}})\rceil$, where $\nu \in [0, 1]$, $\Phi = 2 - \delta_2^2$, $L_t = 1 - \lambda_n(W^t) \in (0, 2]$, and M is a finite constant depending on the specific choice of the retraction. The sequence of iterates $\{\boldsymbol{x}_k\}$ generated by (DRCS) achieves consensus at a linear rate if the initialization satisfies $\boldsymbol{x}_0 \in \mathcal{N}$. That is, we have $\boldsymbol{x}_k \in \mathcal{N}$ for all $k \geq 0$ and $\|\boldsymbol{x}_{k+1} - \bar{\boldsymbol{x}}_{k+1}\|_F \leq \rho_t \|\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k\|_F$, where $\bar{\boldsymbol{x}}_k := 1_n \otimes \left(\mathcal{P}_{\mathcal{M}}(\frac{1}{n}\sum_{i=1}^n x_{i,k})\right)$, $\rho_t := \sqrt{1 - 2(1-\nu)\alpha\gamma_t}$, $\mu_t = 1 - \lambda_2(W^t)$, and $\gamma_t = (1 - 4r\delta_1^2)(1 - \frac{\delta_2^2}{2})\mu_t \geq \frac{\mu_t}{2} \geq \frac{1-\sigma_2^t}{2}$.

III. OUR METHOD

Motivated by Fact 2.1, our proposed DRSM proceeds as follows. In the k-th iteration, it performs a consensus step and then updates the local variable using a Riemannian subgradient direction, i.e., for $i \in [n]$,

$$x_{i,k+1} = \mathcal{R}_{x_{i,k}}(\alpha \mathcal{P}_{\mathrm{T}_{x_{i,k}}\mathcal{M}}(\sum_{j=1}^n W_{ij}^t x_{j,k}) - \beta_k \tilde{\nabla}_{\mathcal{R}} f_i(x_{i,k})), \tag{DRSM}$$

Algorithm 1 Decentralized Riemannian Subgradient Method (DRSM) for Solving Problem (2)

```
Input: $\boldsymbol{x}_0 \in \mathcal{N}$, an integer $t \geq \log_{\sigma_2}\left(\frac{1}{2\sqrt{n}}\right)$, $0 < \alpha \leq \bar{\alpha}$ with $\bar{\alpha}$ being given in Fact 2.1.
for $k = 1, 2, \ldots$ {each node $i \in [n]$ in parallel} do
    Choose geometrically diminishing stepsizes $\beta_k$.
    Perform the update according to (DRSM).
end for
```

where $\tilde{\nabla}_{\mathcal{R}} f_i(x_{i,k})$ is a Riemannian subgradient of f_i at the point $x_{i,k}$ (see [23, Section 2.2]), α and β_k are stepsizes to be determined shortly, and $t \geq 1$ is an integer denoting the t-th power of the mixing matrix W (i.e., performing multistep consensus). We summarize the algorithm in Algorithm 1. One can view the update (DRSM) as applying the Riemannian subgradient method to the following penalized version of problem (2): $\min_{x_i \in \mathcal{M}} \beta_k f(\boldsymbol{x}) + \alpha \varphi^t(\boldsymbol{x})$. To gradually approach consensus, we need to increase the effect of φ^t, or equivalently, decrease the effect of f. Therefore, β_k should be diminishing. We will formally specify this requirement later.

IV. Local Linear Convergence Under Sharpness

In this section, we aim at deriving convergence guarantee for DRSM when applied to problem (2) with the sharpness property besides just weak convexity. In the centralized setting, to establish the (local) linear convergence rate of iterative methods for non-convex problems or convex but not strongly-convex problems, certain regularity conditions (e.g., the error bound condition [24], the Kurdyka-Łojasiewicz inequality [25], or the sharpness property [26]) are usually required. In addition, there have been many attempts to establish strong convergence results in decentralized settings under the aforementioned regularity properties; see, e.g., [12], [27]. Motivated by such a line of research, we show that if problem (1) possesses the following sharpness property [18], [28], then with geometrically diminishing stepsizes (i.e., $\beta_k = \mu_0 \gamma^k$ with $\mu_0 > 0$ and $\gamma \in (0,1)$), our proposed DRSM for problem (2) would converge at a linear rate, provided that it is initialized with a suitable point.

Definition 4.1 (Sharpness): A set $\mathcal{X} \subseteq \mathcal{M}$ is called a set of weak sharp minima for the function $f : \mathbb{R}^{d \times r} \to \mathbb{R}$ with parameter $\kappa > 0$ if there exists a constant $B > 0$ such that for every $x \in U_{\mathcal{X}}(B) \cap \mathcal{M}$ and every $y \in \mathcal{X}$, $f(x) - f(y) \geq \kappa \cdot \operatorname{dist}(x, \mathcal{X})$, where $U_{\mathcal{X}}(B)$ is the B-tube around $\mathcal{X}$ defined as $U_{\mathcal{X}}(B) := \{y \in \mathbb{R}^{d \times r} : \operatorname{dist}(y, \mathcal{X}) < B\}$.

Note that if $\mathcal{X}$ is a set of weak sharp minima for f, then it is the set of minimizers of f over $U_{\mathcal{X}}(B) \cap \mathcal{M}$. In addition, when f is continuous (e.g., if f is weakly convex; see [23, Section 2.2]), then $\mathcal{X}$ can be chosen as a closed set.

We first estimate the consensus error $\|\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k\|_F$ when using geometrically diminishing stepsizes.

Lemma 4.2: Let the stepsizes in DRSM be chosen as $\beta_k = \mu_0 \gamma^k, k \geq 0$, where $0 < \mu_0 \leq \min\left\{\frac{1-\rho_t}{L}\delta_1, \frac{\alpha\delta_1}{5L}\right\}$, L satisfies $\|\tilde{\nabla}_{\mathcal{R}} f_i(x)\|_F \leq L$, and $\rho_t^\delta \leq \gamma < 1, \delta \in (0,1)$. If Assumption 1.1 holds and $\boldsymbol{x}_0 \in \mathcal{N}$, then we have $\|\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k\|_F = \mathcal{O}(\beta_k)$. The following two assumptions are required to prove the local linear convergence of DRSM under the sharpness condition.

Assumption 4.3: There exists a weak sharp minimum $x^* \in \mathcal{M}$ of problem (1) that is isolated.

Assumption 4.4: Let $\boldsymbol{x}_0^\top = (x_{1,0}^\top, \ldots, x_{n,0}^\top)$ be the initial point of DRSM. Let $\Gamma \geq 3$ be a constant. Define

$$e_0 := \min\left\{\max\left\{\frac{\kappa}{(\rho+L)\Gamma}, \sqrt{\sum_{i=1}^{n} \frac{\|x_{i,0} - x^*\|_F^2}{n}}\right\}, \frac{B}{\Gamma}\right\},$$

$$a := 2(L + \kappa + \alpha L L_t)L, \quad b := (4\alpha\sqrt{r} + 2\alpha^2 r)L_t L^2,$$

$$q := \frac{2\kappa e_0}{\Gamma} - (\rho+L)e_0^2.$$

We assume that the constant $\mu_0 > 0$ is strictly less than

$$\min\left\{\frac{e_0}{2\kappa - (\rho+L)e_0}, \frac{q}{L^2 e_0^2 + \frac{4(a+b)}{(1-\rho_t)^2}}, \frac{(1-\rho_t)e_0}{2\sqrt{n}L}\right\}. \quad (4)$$

With the above setup, we establish the local linear convergence result in the following theorem. Its proof can be found in [23].

Theorem 4.5: Suppose that the conditions in Lemma 4.2, Assumption 4.3, and Assumption 4.4 hold. Suppose further that the initial point $\boldsymbol{x}_0$ satisfies the following two conditions:

$$\sum_{i=1}^{n} \|x_{i,0} - x^*\|_F^2 < \frac{n}{\Gamma^2} \min\left\{\left(\frac{2\kappa}{\rho+L}\right)^2, B^2\right\}, \quad (5)$$

$$\|\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0\|_F = 0. \quad (6)$$

Then, there exists a sufficiently small constant $\delta > 0$ such that for $\gamma = \rho_t^\delta$, we have for every $i \in [n]$,

$$\sum_{i=1}^{n} \|x_{i,k} - x^*\|_F^2 \leq n\gamma^{2k} e_0^2, \ \|x_{i,k} - x^*\|_F^2 \leq \Gamma^2 \gamma^{2k} e_0^2 \quad (7)$$

for the sequence $\{x_{i,k}\}_{k \geq 0}$ generated by DRSM.

Some comments on Theorem 4.5 are in order.

i) The condition (5) requires that the initial points $x_{i,0}, i \in [n]$ should be all close to x^*. One can simply initialize all agents with the same value, i.e., $x_{1,0} = x_{2,0} = \cdots = x_{n,0}$, to satisfy (6).

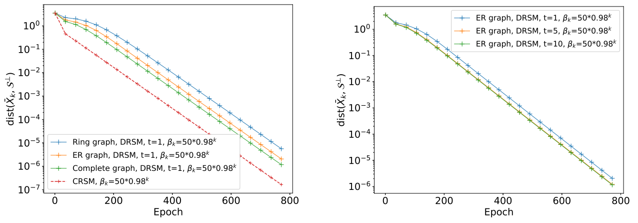
ii) An immediate conclusion is that

$$\|\bar{x}_k - x^*\|_F^2 \leq 4\|\hat{x}_k - x^*\|_F^2$$
$$\leq \frac{4}{n} \sum_{i=1}^{n} \|x_{i,k} - x^*\|_F^2 \leq 4\gamma^{2k} e_0^2, \quad (8)$$

where the first inequality comes from [23, Lemma 2.3] and the last inequality is from (7).

V. Numerical Experiments

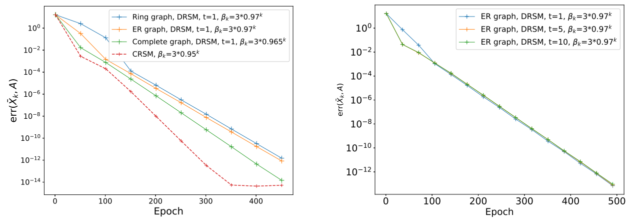
We conduct numerical experiments on the decentralized dual principal component pursuit (DPCP) problem as well as the decentralized orthogonal dictionary learning (ODL) problem to compare our DRSM algorithm with its centralized counterpart (CRSM) [18]. Throughout the experiments, for the network topology, we consider three different choices: A complete graph, a ring graph, and an Erdös-Rényi (ER) random graph where each possible edge is generated independently with probability 0.3.



(a) Geometrically diminishing stepsizes

(b) Geometrically diminishing stepsizes and different t

Fig. 1. Convergence performance of Riemannian subgradient-type methods for solving the DPCP problem.



(a) Geometrically diminishing stepsizes

(b) Geometrically diminishing stepsizes and different t

Fig. 2. Convergence performance of Riemannian subgradient-type methods for solving the ODL problem.

A. Dual principal component pursuit (DPCP)

In the DPCP problem, one is given some measurements $\tilde{Y} = [Y \ O]\Gamma \in \mathbb{R}^{d\times m}$, where the columns of $Y \in \mathbb{R}^{d\times m_1}$ form inlier points spanning a $(d-r)$-dimensional subspace $\mathcal{S}$, the columns of $O \in \mathbb{R}^{d\times m_2}$ form outlier points with no linear structure, and $\Gamma \in \mathbb{R}^{m\times m}$ with $m = m_1 + m_2$ is an unknown permutation. To recover the subspace $\mathcal{S}$ (or $\mathcal{S}^\perp$), one aims to solve

$$\min_{X\in\mathbb{R}^{nd\times r}} f(X) := \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{N}\sum_{j=1}^{N}\left\|(\tilde{y}_{i,j})^\top X_i\right\|_2\Big) \tag{9}$$
$$\text{s.t. } X_1 = X_2 = \cdots = X_n,\ X_i \in \mathrm{St}(d,r),$$

where $X^\top := (X_1^\top, X_2^\top, \ldots, X_n^\top)$, $m = n \times N$, and $\tilde{y}_{i,j} \in \mathbb{R}^d$ is the j-th column vector of the data in the i-th local node. We generate the measurements $\tilde{Y}$ following [18] with $d = 100$ and $r = 10$. After that, we randomly allocate the m column vectors of $\tilde{Y}$ to $n = 10$ local nodes such that each node has $N = 500$ column vectors. The initialization is set to satisfy $X_1 = \cdots = X_n$ and we randomly generate X_1 on $\mathrm{St}(d,r)$.

The DPCP problem is weakly convex and possesses the sharpness property with high probability under suitable conditions [18]. In our experiments, suppose that the underlying subspace $\mathcal{S}^\perp$ is the column space of a matrix $X_{\text{true}} \in \mathrm{St}(d,r)$. We measure the performance by the distance between the IMA in the k-th iteration and the low-dimensional subspace $\mathcal{S}^\perp$, that is $\mathrm{dist}(\bar{X}_k, \mathcal{S}^\perp) = \min_{Q\in O(r)} \|\bar{X}_k Q - X_{\text{true}}\|_F$, where $O(r)$ represents the set of $r \times r$ orthogonal matrices.

We present the linear convergence rate of DRSM with geometrically decaying stepsizes. In each epoch k, we set the stepsize for DRSM and CRSM as $\beta_k = 50 \times 0.98^k$ and set $t = 1, \alpha = 1$ for the multistep consensus in our DRSM. The convergence results are shown in Figure 1(a). From Figure 1(a), our proposed DRSM converges linearly in all three graphs, which is in line with our theoretical analysis. In Figure 1(b), we show the convergence performance of DRSM with geometrically diminishing stepsizes and varying t. It can be observed that the convergence behavior of DRSM with different t is similar.

B. Orthogonal dictionary learning (ODL)

For the ODL problem, the goal is to obtain a suitable compact representation of the observed data $Y \in \mathbb{R}^{d\times m}$. Assuming that the observation Y can be approximated by $Y \approx AS$, where $A \in \mathrm{St}(d,d)$ represents the underlying orthogonal dictionary to be estimated and each column of $S \in \mathbb{R}^{d\times m}$ is sparse, we try to recover the entire dictionary A by considering the formulation

$$\min_{X\in\mathbb{R}^{nd\times d}} f(X) := \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{N}\sum_{j=1}^{N}\left\|(y_{i,j})^\top X_i\right\|_1\Big) \tag{10}$$
$$\text{s.t. } X_1 = X_2 = \cdots = X_n,\ X_i \in \mathrm{St}(d,d),$$

where $X^\top = (X_1^\top, X_2^\top, \ldots, X_n^\top)$, $m = n \times N$, and $y_{i,j} \in \mathbb{R}^d$ is the j-th column vector of the data in the i-th local node. We generate the data A, S, and Y following [18] with $d = 30$ and $m = 1650$. Then, we randomly allocate the m columns of Y to $n = 10$ local nodes with $N = 165$ column vectors on each node. We also use random Gaussian initialization to generate $X_1 \in \mathrm{St}(d,d)$ and set $X_1 = \cdots = X_n$. The performance measure is defined as the error between $\bar{X}$ and A; i.e., $\mathrm{err}(\bar{X}, A) = \sum_{i=1}^{d} |\max_{1\le j\le d} |[\bar{X}_i^\top A]_j| - 1|$.

Figure 2(a) shows the linear convergence of DRSM and CRSM when geometrically diminishing stepsizes of the form $\beta_k = \mu_0\gamma^k$ are used. As shown in Theorem 4.5, smaller γ can be used for DRSM on graphs with better connectivity. Hence, we set $\mu_0 = 3, \gamma = 0.97$ for the ring and ER graphs, $\mu_0 = 3, \gamma = 0.965$ for the complete graph, and $\mu_0 = 3, \gamma = 0.95$ for CRSM. We also show the performance of DRSM with geometrically diminishing stepsizes and varying t in Figure 2(b). It can be observed that DRSM with $t = 5$ or 10 converges faster in the initial process than DRSM with $t = 1$ and they behave similarly later.

VI. Concluding Remarks

We proposed the decentralized Riemannian subgradient method (DRSM) for solving decentralized weakly convex (possibly non-smooth) optimization problems over the Stiefel manifold and showed that it enjoys a local linear convergence rate if the problem at hand exhibits the sharpness property. Future directions include exploring practical optimization problems over other embedded manifolds and provably alleviating the communication burden since multiple rounds of communications are required per iteration in DRSM.

References

[1] A. Nedic, A. Ozdaglar, and P. A. Parrilo, “Constrained consensus and optimization in multi-agent networks,” *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, 2010.

[2] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, “Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[3] B. Ying, K. Yuan, Y. Chen, H. Hu, P. Pan, and W. Yin, “Exponential graph is provably efficient for decentralized deep training,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 975–13 987, 2021.

[4] V. Huroyan and G. Lerman, “Distributed robust subspace recovery,” *SIAM Journal on Scientific Computing*, vol. 40, no. 5, pp. A3067–A3090, 2018.

[5] A. Daneshmand, Y. Sun, G. Scutari, F. Facchinei, and B. Sadler, “Decentralized dictionary learning over time-varying digraphs,” *Journal of Machine Learning Research*, vol. 20, 2019.

[6] J. Chen, Z. J. Towfic, and A. H. Sayed, “Dictionary learning over distributed models,” *IEEE Transactions on Signal Processing*, vol. 63, no. 4, pp. 1001–1016, 2014.

[7] H. Raja and W. U. Bajwa, “Cloud k-SVD: A collaborative dictionary learning algorithm for big, distributed data,” *IEEE Transactions on Signal Processing*, vol. 64, no. 1, pp. 173–188, 2015.

[8] S. U. Pillai, T. Suel, and S. Cha, “The Perron-Frobenius theorem: Some of its applications,” *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp. 62–75, 2005.

[9] J. Tsitsiklis, D. Bertsekas, and M. Athans, “Distributed asynchronous deterministic and stochastic gradient optimization algorithms,” *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 803–812, 1986.

[10] K. Yuan, Q. Ling, and W. Yin, “On the convergence of decentralized gradient descent,” *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1835–1854, 2016.

[11] J. Zeng and W. Yin, “On nonconvex decentralized gradient descent,” *IEEE Transactions on Signal Processing*, vol. 66, no. 11, pp. 2834–2848, 2018.

[12] S. Chen, A. Garcia, and S. Shahrampour, “On distributed non-convex optimization: Projected subgradient method for weakly convex problems in networks,” *IEEE Transactions on Automatic Control*, 2021.

[13] J. C. Duchi, A. Agarwal, and M. J. Wainwright, “Dual averaging for distributed optimization: Convergence analysis and network scaling,” *IEEE Transactions on Automatic Control*, vol. 57, no. 3, pp. 592–606, 2011.

[14] C. Liu, Z. Zhou, J. Pei, Y. Zhang, and Y. Shi, “Decentralized composite optimization in stochastic networks: A dual averaging approach with linear convergence,” *IEEE Transactions on Automatic Control*, 2022.

[15] S. Chen, A. Garcia, M. Hong, and S. Shahrampour, “Decentralized Riemannian gradient descent on the Stiefel manifold,” *arXiv preprint arXiv:2102.07091*, 2021.

[16] H. Ye and T. Zhang, “DeEPCA: Decentralized exact PCA with linear convergence rate.” *Journal of Machine Learning Research*, vol. 22, no. 238, pp. 1–27, 2021.

[17] L. Wang and X. Liu, “Decentralized optimization over the Stiefel manifold by an approximate augmented Lagrangian function,” *IEEE Transactions on Signal Processing*, vol. 70, pp. 3029–3041, 2022.

[18] X. Li, S. Chen, Z. Deng, Q. Qu, Z. Zhu, and A. Man-Cho So, “Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods,” *SIAM Journal on Optimization*, vol. 31, no. 3, pp. 1605–1634, 2021.

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

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

[21] J. Hu, X. Liu, Z.-W. Wen, and Y.-X. Yuan, “A brief introduction to manifold optimization,” *Journal of the Operations Research Society of China*, vol. 8, no. 2, pp. 199–248, 2020.

[22] N. Boumal, *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

[23] J. Wang, J. Hu, S. Chen, Z. Deng, and A. M.-C. So, “Decentralized weakly convex optimization over the stiefel manifold,” *arXiv preprint arXiv:2303.17779*, 2023.

[24] Z. Zhou and A. M.-C. So, “A unified approach to error bounds for structured convex optimization problems,” *Mathematical Programming*, vol. 165, no. 2, pp. 689–728, 2017.

[25] G. Li and T. K. Pong, “Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods,” *Foundations of Computational Mathematics*, vol. 18, no. 5, pp. 1199–1232, 2018.

[26] D. Davis, D. Drusvyatskiy, K. J. MacPhee, and C. Paquette, “Subgradient methods for sharp weakly convex functions,” *Journal of Optimization Theory and Applications*, vol. 179, no. 3, pp. 962–982, 2018.

[27] Y. Tian, Y. Sun, and G. Scutari, “Asynchronous decentralized successive convex approximation,” *arXiv preprint arXiv:1909.10144*, 2019.

[28] J. V. Burke and M. C. Ferris, “Weak sharp minima in mathematical programming,” *SIAM Journal on Control and Optimization*, vol. 31, no. 5, pp. 1340–1359, 1993.