On the Local Linear Rate of Consensus on the Stiefel Manifold

Shixiang Chen[1], Alfredo Garcia[2], Mingyi Hong[3] and Shahin Shahrampour[4]

***Abstract*—Coordinated group behavior arising from purely local interactions has been successfully modeled with distributed consensus-seeking dynamics, where the local behavior is aimed at minimizing the disagreement with neighboring peers. However, it has been recently shown that when constrained by a manifold geometry, distributed consensus-seeking dynamics may ultimately fail to converge to a global consensus state. In this paper, we study discrete-time consensus-seeking dynamics on the Stiefel manifold and identify conditions on the network topology to ensure convergence to a global consensus state. We further prove a (local) linear convergence rate to the consensus state that is on par with the well-known rate in the Euclidean space. These results have implications for consensus applications constrained by manifold geometry, such as synchronization and collective motion, and they can be used for convergence analysis of decentralized Riemannian optimization on the Stiefel Manifold.**

I. Introduction

A common question related to the study of biological, socio-economic or engineering multi-agent systems pertains to the ways in which coordinated group behavior can be explained from purely local interactions (e.g., flocking of birds, schooling of fish and other forms of synchronized behavior). Consensus is a generic approach that has been successfully used to this end. The main premise is that local behavior is expected to reduce disagreement with "neighbors", i.e., peers whose state can be observed or readily exchanged according to a communication network. A measure of local disagreement can be formally expressed as $\varphi_i(\mathbf{x}) := \frac{1}{2}\sum_{j=1}^N W_{ij}\|x_i - x_j\|_F^2$ where $x_i \in \mathbb{R}^{d\times r}$ denotes the state of agent $i \in \{1,\dots,N\}$, $\mathbf{x}^\top := (x_1^\top \ x_2^\top \ \dots \ x_N^\top)$ and $W_{ij} \geq 0$ is a weight such that $W_{ij} = W_{ji} \in (0,1]$ if and only if agents i and j are neighbors and $W_{ij} = 0$ otherwise. Gradient descent dynamics of the form:

$$x_{i,k+1} = x_{i,k} - \alpha \nabla \varphi_i(\mathbf{x}_k),$$

where $\alpha > 0$ is the step-size are often referred to as consensus dynamics. Assuming the network is connected, all agents asymptotically agree on a state, say x^*, i.e. $\lim_{k\to\infty} x_{i,k} = x^*$ for all $i \in \{1,\dots,N\}$, and the asymptotic agreement state minimizes the consensus potential $\varphi(\mathbf{x}) := \frac{1}{2}\sum_{i=1}^N \varphi_i(\mathbf{x})$, a global measure of disagreement.

This work is supported by NSF-ECCS-2136206 and NSF-CCF-CIF-1910385 awards.

[1]School of Mathematical Sciences, University of Science and Technology of China, Hefei, Anhui, China. Email address: shxchen@ustc.edu.cn

[2]The Wm Michael Barnes '64 Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX 77843. Email address: alfredo.garcia@tamu.edu

[3]The Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455. Email address: mhong@umn.edu

[4]The Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA 02115. Email address: s.shahrampour@northeastern.edu

Consensus on manifolds has recently attracted significant attention [1]–[4] due to its applications to synchronization in planetary scale sensor networks [5], modeling of collective motion in flocks in the Earth's atmosphere [6], synchronization of quantum bits [7], and the Kuramoto models [3], [8]. However, in a manifold geometry, the individual dynamics of consensus gradient descent flows (in continuous-time) do not necessarily converge to an agreement state [1], i.e., $\varphi(\mathbf{x}) \neq 0$ and individual agents may converge to *different* states. Therefore, it is crucial to identify conditions under which consensus can be achieved when the dynamics are constrained to a manifold geometry.

In our recent work [9], we have shown consensus dynamics play a key role in decentralized multi-agent optimization constrained by the Stiefel manifold geometry, which in turn has applications in dictionary learning [10] and training deep neural networks with orthogonal constraints [11], [12].

Let $\varphi_i^t(\mathbf{x}) := \frac{1}{2}\sum_{j=1}^N W_{ij}^t\|x_i - x_j\|_F^2$ be the local consensus potential. In this paper, we study consensus dynamics on the Stiefel manifold described as follows:

$$x_{i,k+1} = \mathrm{Retr}_{x_{i,k}}(-\alpha \cdot \mathrm{grad}\varphi_i^t(\mathbf{x}_k)), \tag{I.1}$$

where an update along a negative Riemannian gradient direction $-\mathrm{grad}\varphi_i^t(\mathbf{x}_k)$ on the tangent space is followed by a retraction operation $\mathrm{Retr}_{x_{i,k}}(\cdot)$ in order to ensure feasibility. We show that the consensus dynamics (I.1) converges (locally) Q-linearly[1] to the solutions of the following *non-convex* optimization problem

$$\begin{aligned} \min_{\mathbf{x}} \ & \left\{ \varphi^t(\mathbf{x}) := \frac{1}{2}\sum_{i=1}^N \varphi_i^t(\mathbf{x}) \right\} \\ \text{s.t.} \quad & x_i \in \mathrm{St}(d,r), \ i = 1,\dots,N. \end{aligned} \tag{C-St}$$

where $\mathrm{St}(d,r) = \{x \in \mathbb{R}^{d\times r} : x^\top x = I_r\}$ is the Stiefel manifold and I_r is the $r \times r$ identity matrix. Furthermore, we prove that the linear rate depends on the connectivity of the graph and more specifically on the magnitude of the second largest singular value of W.

The paper is organized as follows. In Section I-A we discuss the relevant literature of consensus dynamics on Riemannian manifolds, and in Section I-B we briefly summarize our technical contributions that enable the Q-linear characterization of convergence. In Section II we lay out the mathematical setting and provide a brief overview of Riemannian optimization. In Section III we formally state the decentralized consensus dynamics and describe in a series of remarks the technical challenges associated with establishing the convergence. In Section IV we present the global convergence of the dynamics,

[1]A sequence $\{a_k\}$ is said to converge Q-linearly to a if there exists $\rho \in (0,1)$, such that $\lim_{k\to\infty} \frac{|a_{k+1}-a|}{|a_k - a|} = \rho$.

and in Section V we develop the Riemannian restricted secant inequality (RSI) and formally establish a local linear rate of convergence. Finally, in Section VI we illustrate the convergence characterization with numerical experiments, including an application to decentralized manifold optimization by means of a nested-loop algorithm, where agents alternate between performing Riemannian gradient updates (outer loop) and consensus updates (inner loop). Section VII concludes, and Section VIII includes proofs of all technical results.

A. Literature Review

The literature on optimization on Riemannian manifolds can be broadly classified into *intrinsic* and *extrinsic* methods. Intrinsic optimization algorithms are defined in terms of the inherent manifold geometry, such as geodesic distances, Riemannian gradient, and exponential and logarithm maps. In contrast, the extrinsic algorithms are based on a specific embedding of the manifolds in Euclidean space.

Intrinsic methods for optimization on manifolds with bounded curvature include the discrete-time Riemannian gradient method (RGM) [13]. This work shows that when applied to minimizing the consensus potential on Grassmannian manifold and special orthogonal group $\mathrm{SO}(d)$, the RGM method has a sub-linear rate of convergence. The dynamics (I.1) differ from the Riemannian consensus algorithm in [13] in that the retraction map is used instead of the exponential map to guarantee feasibility.

The authors of [14] consider stochastic RGM and examine its application for minimizing the consensus potential on the manifold of symmetric positive definite matrices. In [15], RGM is also studied for minimizing the consensus potential on the Grassmannian manifold and special orthogonal groups. However, it is only shown that RGM converges to a critical point. To achieve the global consensus, a synchronization algorithm on the tangent space, requiring communicating an extra variable, is presented in [15, Section 7]. Other results showing global consensus are graph dependent. For example, the authors of [15] show that global consensus is achievable on equally weighted complete graphs for $\mathrm{SO}(d)$ and Grassmannian. For general connected undirected graphs, the survey paper [16] summarizes three solutions to achieve almost global consensus on the circle (i.e., $d = 2$ and $r = 1$): potential reshaping [8], the gossip algorithm [17] and dynamic consensus [15]. However, such procedures could suffer from slow convergence.

Previous work analyzing consensus dynamics on the Stiefel manifold mostly focused on *local* convergence. Recently, the authors of [2], [3] have shown almost global consensus for problem (C-St) whenever $r \leq \frac{2}{3}d - 1$. More specifically, all second-order critical points are global optima, and thus, the measure of stable manifold of saddle points is zero. This can be proved by showing that the Riemannian Hessian at all saddle points has negative curvature, i.e., the strict saddle property in [18] holds true. Therefore, if we randomly initialize RGM, it will almost always converge to the global optimal point [3], [18]. Additionally, [3] also conjectures that the strict saddle property holds for $d \geq 3$ and $r \leq d - 2$. The scenarios $r = d - 1$ and $r = d$ correspond to the multiply connected ($\mathrm{St}(d, d-1) \cong \mathrm{SO}(d)$) and not connected case ($\mathrm{St}(d, d) \cong \mathrm{O}(d)$, where $\mathrm{O}(d)$ is the orthogonal group.), respectively, which yields multi-stable systems [1].

However, none of the aforementioned works discusses the local linear rate of RGM on $\mathrm{St}(d, r)$ with $r > 1$. We could prove the linear rate if the Riemannian Hessian were positive definite [19] near a consensus point, but the Riemannian Hessian is degenerate at all consensus points (see Section V). The linear rate of consensus can be established by reparameterization on the circle [8] or computing the generalized Lyapunov-type numbers on the sphere [20], but it is not known how to generalize them to $r > 1$. In this paper we study the convergence of consensus dynamics (I.1) using the recent advances in non-convex optimization [18] as well as optimization over Stiefel manifold [19], [21], [22].

B. Summary of Technical Contributions

The characterization of convergence for consensus dynamics on the Stiefel manifold is enabled by a number of technical contributions, which we summarize as follows:

1) We identify a sufficient condition on the stepsize $\alpha > 0$ in order to guarantee global convergence for the consensus dynamics (I.1) in Theorem 2 for the case that $r \leq \frac{2}{3}d-1$. This result is based on Theorem 1 in the form of a new descent lemma, which enables us to obtain a better bound on the algorithm step size compared to the existing work.
2) In Theorem 3, we establish a sufficient and necessary condition for a first-order critical point to be global optimum. We also show via an example that the *box* region characterized in Theorem 3 has a tight upper bound. This helps with identifying suitable local neighborhoods wherein the convergence of dynamics (I.1) is linear.
3) We identify a surrogate for local strong convexity for problem (C-St). It is called the *Restricted Secant Inequality* (RSI) and is derived in Proposition 2. This inequality facilitates the proof by allowing us to disregard the second-order information in the analysis.
4) We prove the local Q-linear rate of convergence of $\mathrm{dist}(\mathbf{x}_k, \mathcal{X}^*)$, i.e. the Euclidean distance between $\mathbf{x}_k$ and $\mathcal{X}^*$, which is the optimal solution set for the problem (C-St) in the following form:

$$\mathcal{X}^* := \{\mathbf{x} \in \mathrm{St}(d, r)^N : x_1 = x_2 = \ldots = x_N\}. \quad \text{(I.2)}$$

We show that the convergence rate asymptotically scales with the second largest singular value of W, thereby tending to its counterpart in the Euclidean space. We characterize two local regions for such convergence in Theorem 4.

II. Preliminaries

A. Multi-agent Systems

To represent the network, we use a graph $\mathcal{G}$ in which connected nodes can communicate with each other. We assume that $\mathcal{G}$ satisfies the following assumption.

Assumption 1. *We assume that the undirected graph $\mathcal{G}$ is connected and the corresponding weight matrix W is doubly stochastic and symmetric, i.e.,*

- $W = W^\top$.
- $1 > W_{ij} \geq 0$; $1 > W_{ii} > 0$; $\sum_{i=1}^N W_{ij} = \sum_{i=1}^N W_{ji} = 1$.

Note that a doubly stochastic matrix on an undirected connected network can be constructed easily following Laplacian-based constant edge weight matrix [23], the Metropolis rule and the Maximum-Degree rule [24], to name a few. Under Assumption 1, any power of the matrix W, i.e., $W^t := W^{t-1}W$ (t an integer greater than one) is also doubly stochastic and symmetric. Moreover, the second largest singular value of W^t, denoted by σ_2^t, lies in $[0, 1)$. The consensus potential with weight matrix W^t is defined as $\varphi^t(\mathbf{x}) := \frac{1}{2}\sum_{i=1}^N \varphi_i^t(\mathbf{x})$ where $\varphi_i^t(\mathbf{x}) := \frac{1}{2}\sum_{j=1}^N W_{ij}^t \|x_i - x_j\|_F^2$.

In what follows, to simplify the notation, we denote the Stiefel manifold $\mathrm{St}(d, r)$ by $\mathcal{M}$. We have the following notations:

- $\mathcal{G} = (\mathcal{V}, \mathcal{E})$: the undirected graph with $|\mathcal{V}| = N$ nodes.
- $\mathbf{x}$: the collection of all local variables x_i by stacking them, i.e., $\mathbf{x}^\top = (x_1^\top \ x_2^\top \ \dots \ x_N^\top)$.
- $\mathcal{M}^N = \mathcal{M} \times \dots \times \mathcal{M}$: the $N-$fold Cartesian product.
- $[N] := \{1, 2, \dots, N\}$. For any $\mathbf{x} \in (\mathbb{R}^{d\times r})^N$, the i-th block is captured by $[\mathbf{x}]_i = x_i$.
- $\nabla\varphi(\mathbf{x})$: Euclidean gradient; $\nabla\varphi_i(\mathbf{x}) := [\nabla\varphi(\mathbf{x})]_i$: the i-th block of $\nabla\varphi(\mathbf{x})$.
- $\mathrm{T}_x\mathcal{M}$: the tangent space of $\mathcal{M}$ at point x.
- $N_x\mathcal{M}$: the normal space of $\mathcal{M}$ at point x.
- $\mathrm{Tr}(\cdot)$: trace operator; $\langle x, y\rangle = \mathrm{Tr}(x^\top y)$: inner product on $\mathrm{T}_x\mathcal{M}$ is induced from the Euclidean inner product.
- $\mathrm{grad}\varphi(\mathbf{x})$: Riemannian gradient; $\mathrm{grad}\varphi_i(\mathbf{x}) := [\mathrm{grad}\varphi(\mathbf{x})]_i$: the i-th block of $\mathrm{grad}\varphi(\mathbf{x})$. $\mathrm{Hess} f(\mathbf{x})$ denotes the Riemannian Hessian operator.
- D captures the differential of f and $Df(x)[\xi]$ denotes the directional derivative along ξ.
- $\|\cdot\|_F$: the Frobenius norm; $\|\cdot\|_2$: the Euclidean operator norm.
- $\mathcal{P}_C$: the orthogonal projection onto a closed set C.
- $\mathbf{1}_N \in \mathbb{R}^N$: the vector of all ones; $J := \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top$.

Definition 1 (Consensus)**.** *Consensus is the configuration where $x_i = x_j \in \mathcal{M}$ for all $i, j \in [N]$.*

B. Optimality Condition

We introduce some preliminaries about optimization on a Riemannian manifold. Let us consider the following optimization problem over a product matrix manifold $\mathcal{M}^N$

$$\min f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{M}^N. \tag{II.1}$$

Under the Euclidean metric, the Riemannian gradient $\mathrm{grad} f(\mathbf{x})$ on $\mathcal{M}^N$ is given by $\mathrm{grad} f(\mathbf{x}) = \mathcal{P}_{\mathrm{T}_\mathbf{x}\mathcal{M}^N}(\nabla f(\mathbf{x}))$, where $\mathcal{P}_{\mathrm{T}_\mathbf{x}\mathcal{M}^N}$ is the orthogonal projection onto $\mathrm{T}_\mathbf{x}\mathcal{M}^N$. More specifically, the projection onto the $i-$th block of tangent space, $\mathrm{T}_{x_i}\mathcal{M}$, is given by

$$\mathcal{P}_{\mathrm{T}_{x_i}\mathcal{M}}(y) = y - \frac{1}{2}x_i(x_i^\top y + y^\top x_i),$$

for any $y \in \mathbb{R}^{d\times r}$ (see [21]), and

$$\mathcal{P}_{N_{x_i}\mathcal{M}}(y) = \frac{1}{2}x_i(x_i^\top y + y^\top x_i).$$

According to our notation in Section II.A, we have

$$\mathrm{grad}\varphi_i^t(\mathbf{x}) = \nabla\varphi_i^t(\mathbf{x}) - \frac{1}{2}x_i(x_i^\top\nabla\varphi_i^t(\mathbf{x}) + \nabla\varphi_i^t(\mathbf{x})^\top x_i).$$

The Riemannian Hessian $\mathrm{Hess} f(\mathbf{x})$ is given by $\mathrm{Hess} f(\mathbf{x})[\xi] = \mathcal{P}_{\mathrm{T}_\mathbf{x}\mathcal{M}^N}(D(\mathbf{x} \mapsto \mathcal{P}_{\mathrm{T}_\mathbf{x}\mathcal{M}^N}\nabla f(\mathbf{x}))[\xi])$ for any $\xi \in \mathrm{T}_\mathbf{x}\mathcal{M}^N$, i.e., the projection of differential of the Riemannian gradient [19], [22]. A point $\mathbf{x}$ is a first-order critical point (or critical point) if $\mathrm{grad} f(\mathbf{x}) = \mathbf{0}$. $\mathbf{x}$ is called a second-order critical point if $\mathrm{grad} f(\mathbf{x}) = \mathbf{0}$ and $\mathrm{Hess} f(\mathbf{x}) \succcurlyeq \mathbf{0}$.

Proposition 1. *([22]) Let $\mathbf{x} \in \mathcal{M}^N$ be a local optimum for (II.1). If f is differentiable at $\mathbf{x}$, then $\mathrm{grad} f(\mathbf{x}) = \mathbf{0}$. Moreover, if f is twice differentiable at $\mathbf{x}$, then $\mathrm{Hess} f(\mathbf{x}) \succcurlyeq \mathbf{0}$.*

C. The Retraction Operator

The second-order retraction [19, Definition 4.1.1] is the approximation of the exponential mapping, more suitable for the computation purpose. In this paper, we only use the polar-decomposition based retraction to present a simple proof. The polar retraction is given by

$$\mathrm{Retr}_x(\xi) = (x + \xi)(I_r + \xi^\top\xi)^{-1/2}, \tag{II.2}$$

which is also the orthogonal projection of $x+\xi$ onto $\mathcal{M}$. The first important property of the retraction (see [22], [25]) is:

$$\begin{aligned} \|\mathrm{Retr}_x(\xi) - (x + \xi)\|_F \leq M\|\xi\|_F^2, \\ \forall x \in \mathcal{M}, \quad \forall \xi \in \mathrm{T}_x\mathcal{M}, \end{aligned} \tag{P1}$$

where $M > 0$ is a constant given in [22], [25]. More details on M can be found in Appendix. This property implies that $\mathrm{Retr}_x(\xi)$ is locally a good approximation to $x+\xi$. Secondly, for all $x \in \mathcal{M}$ and $\xi \in \mathrm{T}_x\mathcal{M}$, the following inequality holds for any $y \in \mathcal{M}$ [26, Lemma 1]:

$$\|\mathrm{Retr}_x(\xi) - y\|_F \leq \|x + \xi - y\|_F. \tag{II.3}$$

III. Distributed Riemannian Consensus

The discrete-time RGM applied to solve problem (C-St) is described in Algorithm 1. Since it can be implemented in a distributed fashion, we name it Distributed Riemannian Consensus algorithm on Stiefel manifold (DRCS). For large values of t, $W^t \approx \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top$ and thus the corresponding graph is approximately completely connected. Hence, in the analysis below, we identify sufficient conditions on how large t must be (which can be interpreted as a requirement on the graph connectivity) to ensure convergence to a global consensus state.

For the convergence analysis below, it will also be convenient to link the algorithm with an equivalent formulation of the optimization problem (C-St). Namely, given that $\|x\|_F^2 = r$ holds true for any $x \in \mathcal{M}$, (C-St) is equivalent to

$$\begin{aligned} \max_\mathbf{x} \left\{ h^t(\mathbf{x}) := \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N W_{ij}^t \langle x_i, x_j\rangle \right\} \\ \text{s.t.} \quad x_i \in \mathcal{M}, \ \forall i \in [N]. \end{aligned} \tag{III.2}$$

Algorithm 1 Distributed Riemannian Consensus on Stiefel manifold (DRCS)

1: **Input**: random initial point $\mathbf{x}_0 \in \mathcal{M}^N$, stepsize $0 < \alpha < 2/L_t$ and an integer $t \geq 1$, where $L_t := 1 - \lambda_N(W^t)$ and $\lambda_N(W^t)$ is the smallest eigenvalue of W^t.
2: **for** $k = 0, 1, \ldots$ **do** ▷ For each node $i \in [N]$, in parallel
3: Compute $\nabla \varphi_i^t(\mathbf{x}_k)$
4: Update

$$x_{i,k+1} = \mathrm{Retr}_{x_{i,k}}\left(-\alpha \mathcal{P}_{\mathrm{T}_{x_{i,k}}\mathcal{M}}\left(\nabla \varphi_i^t(\mathbf{x}_k)\right)\right) \tag{III.1}$$

5: **end for**

Hence, DRCS can be seen as applying Riemannian gradient ascent to solve (III.2). That is, (III.1) is equivalent to

$$x_{i,k+1} = \mathrm{Retr}_{x_{i,k}}\left(\alpha \mathcal{P}_{\mathrm{T}_{x_{i,k}}\mathcal{M}}\Big(\sum_{j=1}^{N} W_{ij}^t x_{j,k}\Big)\right). \tag{III.3}$$

A. Consensus in Euclidean Space: A Revisit

Let us briefly review the consensus with convex constraints in the Euclidean space (C-E) [27], which will give us some insights to study the convergence rate of DRCS. Euclidean consensus can be achieved by solving the optimization problem below

$$\min \varphi(\mathbf{x}) \quad \text{s.t.} \quad x_i \in \mathcal{C},\ i = 1, \ldots, N, \tag{C-E}$$

where $\mathcal{C}$ is a closed convex set in the Euclidean space. Then, the iteration is given by [28]

$$x_{i,k+1} = \mathcal{P}_{\mathcal{C}}\left(\sum_{j=1}^{N} W_{ij} x_{j,k}\right) \quad \forall i \in [N]. \tag{EuC}$$

Let us denote the Euclidean mean via

$$\hat{x} := \frac{1}{N}\sum_{i=1}^{N} x_i \text{ and } \hat{\mathbf{x}} := \mathbf{1}_N \otimes \hat{x}. \tag{III.4}$$

One can easily verify that

$$\|\mathbf{x}_k - \hat{\mathbf{x}}_k\|_{\mathrm{F}} \leq \sigma_2 \|\mathbf{x}_{k-1} - \hat{\mathbf{x}}_{k-1}\|_{\mathrm{F}}. \tag{III.5}$$

Therefore, the Q-linear rate of (EuC) is equal to σ_2. On the other hand, the iteration (EuC) is the same as applying projected gradient descent (PGD) method to solve the problem (C-E). That is, we have

$$\mathbf{x}_{k+1} = \mathcal{P}_{\mathcal{C}^N}\left((W \otimes I_d)\mathbf{x}_k\right) = \mathcal{P}_{\mathcal{C}^N}\left(\mathbf{x}_k - \alpha_e \nabla \varphi(\mathbf{x}_k)\right), \tag{III.6}$$

with stepsize $\alpha_e = 1$. Following the proof of linear rate for strongly convex functions [29, Theorem 2.1.15], one needs the inequality in [29, Theorem 2.1.12], specialized to our problem as follows

$$\langle \mathbf{x} - \hat{\mathbf{x}}, \nabla \varphi(\mathbf{x}) \rangle \geq \frac{\mu L}{\mu + L}\|\mathbf{x} - \hat{\mathbf{x}}\|_{\mathrm{F}}^2 + \frac{1}{\mu + L}\|\nabla \varphi(\mathbf{x})\|_{\mathrm{F}}^2. \tag{III.7}$$

The constants are given by

$$\mu := 1 - \lambda_2(W) \quad \text{and} \quad L := 1 - \lambda_N(W),$$

where $\lambda_2(W)$ is the second largest eigenvalue of W, and $\lambda_N(W)$ is the smallest eigenvalue of W, respectively. This inequality can be obtained using the eigenvalue decomposition of $I_N - W$. We provide the proof in the Appendix, and we call (III.7) "restricted secant inequality". With this, if $\alpha_e = \frac{2}{\mu + L}$, we get

$$\|\mathbf{x}_k - \hat{\mathbf{x}}_k\|_{\mathrm{F}} \leq \left(\frac{L - \mu}{L + \mu}\right)^k \|\mathbf{x}_0 - \hat{\mathbf{x}}_0\|_{\mathrm{F}}.$$

It can be shown by simple calculations that $\frac{L-\mu}{L+\mu} \leq \sigma_2$. This suggests that the PGD can achieve faster convergence rate with $\alpha_e = \frac{2}{\mu + L}$. When $\alpha_e = 1$, the rate of σ_2 can be shown via combining (III.7) with $L\|\mathbf{x} - \hat{\mathbf{x}}\|_{\mathrm{F}} \geq \|\nabla \varphi(\mathbf{x})\|_{\mathrm{F}} \geq \mu \|\mathbf{x} - \hat{\mathbf{x}}\|_{\mathrm{F}}$. The proof is provided in the Appendix.

B. Consensus on Stiefel Manifold: Challenges and Insights

The DRCS iteration (III.1) is an extension of Euclidean consensus with convex constraint [28], where the projection onto convex set is replaced with a retraction operator, and the Euclidean gradient is substituted by the Riemannian gradient. The standard results [19], [22] on RGM already show global sub-linear rate of DRCS. However, to obtain the *local Q-linear* rate, we need to exploit the specific problem structure. To analyze DRCS, there are two main challenges.

First, due to the non-linearity of $\mathcal{M}$, the Euclidean mean $\hat{x} := \frac{1}{N}\sum_{i=1}^{N} x_i$ is infeasible. We need to use the average point defined on the manifold. The second challenge comes from the non-convexity of $\mathcal{M}$. Previous work (e.g., [28]) usually discusses the convex constraint in the Euclidean space, which depends on the non-expansive property of the projection operator onto convex constraint. We cannot use this property due to non-convexity of the Stiefel manifold.

To solve these issues, we use the so-called induced arithmetic mean (IAM) [15] of $x_1, \ldots, x_N$ over $\mathcal{M}$, defined by

$$\begin{aligned} \bar{x} &\in \operatorname*{argmin}_{y \in \mathcal{M}} \sum_{i=1}^{N} \|y - x_i\|_{\mathrm{F}}^2 \\ &= \operatorname*{argmax}_{y \in \mathcal{M}} \langle y, \sum_{i=1}^{N} x_i \rangle = \mathcal{P}_{\mathcal{M}}(\hat{x}), \end{aligned} \tag{IAM}$$

where $\mathcal{P}_{\mathcal{M}}(\cdot)$ is the orthogonal projection onto $\mathcal{M}$. Note that when $\hat{x}$ does not have full column rank, then $\mathcal{P}_{\mathcal{M}}(\hat{x})$ has multiple solutions. In this scenario, we can let $\bar{x}$ be any element of $\mathcal{P}_{\mathcal{M}}(\hat{x})$. We define

$$\bar{x}_k \in \mathcal{P}_{\mathcal{M}}(\hat{x}_k) \quad \text{and} \quad \bar{\mathbf{x}}_k = \mathbf{1}_N \otimes \bar{x}_k, \tag{III.8}$$

to denote IAM of $x_{1,k}, \ldots, x_{N,k}$. The IAM $\bar{\mathbf{x}}$ is also the projection of $\mathbf{x}$ onto the consensus set. The distance between $\mathbf{x}$ and $\mathcal{X}^*$ is given by

$$\mathrm{dist}^2(\mathbf{x}, \mathcal{X}^*) = \min_{y \in \mathcal{M}} \frac{1}{N}\sum_{i=1}^{N} \|y - x_i\|_{\mathrm{F}}^2 = \frac{1}{N}\|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}}^2.$$

The terminology IAM is derived from [30], where the IAM on $\mathrm{SO}(3)$ is called the projected arithmetic mean. The IAM is different from the Fréchet mean [13], [31], [32] (or the Karcher mean [33], [34]). We use IAM since it is computationally convenient and easier to adapt to the Euclidean linear structure.

The following lemma suggests that $\mathrm{dist}(\mathbf{x},\mathcal{X}^*)<\sqrt{2}$ implies that $\mathcal{P}_{\mathcal{M}}(\hat{x})=\bar{x}$ is unique.

Lemma 1. *For $\mathbf{x}\in\mathcal{M}^N$, if $\mathrm{dist}(\mathbf{x},\mathcal{X}^*)<\sqrt{2}$, then $\mathcal{P}_{\mathcal{M}}(\hat{x})$ is unique.*

Therefore, when $\mathrm{dist}(\mathbf{x},\mathcal{X}^*)<\sqrt{2}$, we can define the $l_{F,\infty}$ distance between $\mathbf{x}$ and $\mathcal{X}^*$ as

$$\mathrm{dist}_{F,\infty}(\mathbf{x},\mathcal{X}^*)=\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F},\infty}:=\max_{i\in[N]}\|x_i-\bar{x}\|_{\mathrm{F}}. \quad (l_{F,\infty})$$

Throughout the remainder of this paper, whenever we use the notation $\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F},\infty}$, we implicitly assume the condition $\mathrm{dist}(\mathbf{x},\mathcal{X}^*)<\sqrt{2}$, which ensures that the notation is well-defined.

Then, we build the connection between the Euclidean mean and IAM in the following lemma, which will be key to convergence analysis of Algorithm 1.

Lemma 2. *For any $\mathbf{x}\in\mathcal{M}^N$, let $\hat{\mathbf{x}}=\mathbf{1}_N\otimes\hat{x}$ as in (III.4). Similarly, let $\bar{\mathbf{x}}=\mathbf{1}_N\otimes\bar{x}$, where $\bar{x}$ is defined in (IAM). We have*

$$\frac{1}{2}\|\mathbf{x}-\bar{\mathbf{x}}\|_F^2\le\|\mathbf{x}-\hat{\mathbf{x}}\|_F^2\le\|\mathbf{x}-\bar{\mathbf{x}}\|_F^2. \quad (\text{III.9})$$

Moreover, if $\|\mathbf{x}-\bar{\mathbf{x}}\|_F^2\le N/2$, one has

$$\|\bar{x}-\hat{x}\|_F\le\frac{2\sqrt{r}\|\mathbf{x}-\bar{\mathbf{x}}\|_F^2}{N}, \quad (\text{P2})$$

and

$$\|\mathbf{x}-\hat{\mathbf{x}}\|_F^2\ge\|\mathbf{x}-\bar{\mathbf{x}}\|_F^2-\frac{4r\|\mathbf{x}-\bar{\mathbf{x}}\|_F^4}{N}. \quad (\text{III.10})$$

The inequality (III.9) is tight, since we have $\frac{1}{2}\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2=\|\mathbf{x}-\hat{\mathbf{x}}\|_{\mathrm{F}}^2=Nr$ when $\sum_{i=1}^N x_i=0$ and $\|\mathbf{x}-\hat{\mathbf{x}}\|_{\mathrm{F}}^2=\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2$ when $x_1=x_2=\ldots=x_N$. The inequality (P2) suggests that the Euclidean mean will converge to IAM quadratically if $\mathbf{x}$ is close to $\bar{\mathbf{x}}$.

We now show the relation between $\nabla\varphi^t(\mathbf{x})$ and $\mathrm{grad}\varphi^t(\mathbf{x})$. Denoting $\mathcal{P}_{N_x\mathcal{M}}$ as the orthogonal projection onto the normal space $N_x\mathcal{M}$, a useful property of the projection $\mathcal{P}_{\mathrm{T}_x\mathcal{M}}(y-x),\forall y\in\mathcal{M}$ [26, Section 6] is that

$$\begin{aligned}&\mathcal{P}_{\mathrm{T}_x\mathcal{M}}(x-y)=x-y-\mathcal{P}_{N_x\mathcal{M}}(x-y)\\&=x-y-\frac{1}{2}x\left((x-y)^\top x+x^\top(x-y)\right)\\&=x-y-\frac{1}{2}x(x-y)^\top(x-y),\end{aligned} \quad (\text{P3})$$

where we used $x^\top x=y^\top y=I_r$. This property implies that

$$\mathcal{P}_{\mathrm{T}_x\mathcal{M}}(x-y)=x-y+\mathcal{O}(\|y-x\|_{\mathrm{F}}^2).$$

We will use (P3) to derive a descent lemma on the Stiefel manifold similar to the Euclidean-type inequality [29], which is helpful to identify the stepsize for global convergence. The stepsize α will be determined by the constant L_t in Theorem 1 and the constant M in equation (P1).

Theorem 1 (Descent lemma)**.** *For the function $\varphi^t(\mathbf{x})$ defined in (C-St), we have*

$$\begin{aligned}&\varphi^t(\mathbf{y})-\left[\varphi^t(\mathbf{x})+\left\langle\mathrm{grad}\varphi^t(\mathbf{x}),\mathbf{y}-\mathbf{x}\right\rangle\right]\\&\le\frac{L_t}{2}\|\mathbf{y}-\mathbf{x}\|_F^2,\quad\forall\mathbf{x},\mathbf{y}\in\mathcal{M}^N,\end{aligned} \quad (\text{III.11})$$

where $L_t=1-\lambda_N(W^t)$ and $\lambda_N(W^t)$ is the smallest eigenvalue of W^t.

By now, we have obtained three second-order properties (P1), (P2) and (P3). This will help us to solve the non-linearity issue and get a similar Riemannian restricted secant inequality. Before that, in the next section, we show the global convergence of Algorithm 1 with a tight estimation of the stepsize α; see discussions on L_t in Appendix.

IV. Global Convergence Analysis

We study the global convergence of sequence $\{\mathbf{x}_k\}$ generated by (III.1). We note that the almost sure convergence to consensus when $r\le\frac{2}{3}d-1$ for the *continuous-time* gradient flow was established in [3]. In contrast, we consider *discrete-time* dynamics, which in turn requires the use of the retraction operator to ensure feasibility. In this section, we characterize bounds on the stepsize α to ensure convergence under the constraint $r\le\frac{2}{3}d-1$ (Theorem 2). While we cannot prove global convergence for general r,d, we can still build on the results of [25], [35], [36] to provide a necessary and sufficient condition for the optimality of critical points (Theorem 3), which characterizes the landscape of the problem and implies that the convergence to optimal set can only be established in local regions. Our main results on the local linear rate are presented in Section V, and they hold for any r,d.

Definition 2 (Łojasiewicz inequality)**.** *We say that $\mathbf{x}\in\mathcal{M}^N$ satisfies the Łojasiewicz inequality for gradient $\mathrm{grad}f(\mathbf{x})$ if there exist $\tilde{\Delta}>0$, $\Lambda>0$ and $\theta\in(0,1/2]$ such that for all $\mathbf{y}\in\mathcal{M}^N$ with $\|\mathbf{y}-\mathbf{x}\|_F<\tilde{\Delta}$, it holds that*

$$|f(\mathbf{y})-f(\mathbf{x})|^{1-\theta}\le\Lambda\|\mathrm{grad}f(\mathbf{x})\|_F. \quad (\text{Ł})$$

Since $\varphi^t(\mathbf{x})$ is a real analytic function, and the Stiefel manifold is a compact real-analytic submanifold, it is well known that a Łojasiewicz inequality holds at each critical point of problem (C-St) [36]. Therefore, we know that the sequence $\{\mathbf{x}_k\}$ converges to a single critical point with properly chosen α, which is a stronger convergence result than the subsequence convergence in [22]. The exponent θ decides the local convergence rate.

Lemma 3. *Let $G:=\max_{\mathbf{x}\in\mathcal{M}^N}\|\mathrm{grad}\varphi^t(\mathbf{x})\|_F$. Given any $t\ge1$ and $\alpha\in(0,\frac{2}{2MG+L_t})$, where M is the constant in (P1), the sequence $\{\mathbf{x}_k\}$ generated by Algorithm 1 converges to a critical point of problem (C-St) sub-linearly. Furthermore, if some critical point is a limit point of $\{\mathbf{x}_k\}$ and has exponent $\theta=1/2$ in (Ł), $\{\varphi^t(\mathbf{x}_k)\}$ converges to 0 Q-linearly and the sequence $\{\mathbf{x}_k\}$ converges to the critical point R-linearly*[2].

The proof can be found in [37], which follows [36, Section 2.3] and [22], but here we use the descent lemma (Theorem 1). Lemma 3 shows the convergence to a critical point. However, we are more interested in the convergence to consensus states (see Definition 1), i.e., global optima. In Theorem 2 we prove that DRCS almost always converges to the optimal point set $\mathcal{X}^*$ [18], which is a discrete-time version of [3, Theorem 4].

[2]A sequence $\{a_k\}$ is said to converge R-linearly to a if there exists a sequence $\{\varepsilon_k\}$ such that $|a_k-a|\le\varepsilon_k$ and $\{\varepsilon_k\}$ converges Q-linearly to 0.

Theorem 2. *When $r \leq \frac{2}{3}d - 1$, let $\alpha \in (0, C_{\mathcal{M},\varphi^t})$, where $C_{\mathcal{M},\varphi^t} := \min\{\frac{\hat{r}}{G}, \frac{1}{\hat{B}}, \frac{2}{2MG+L_t}\}$, $\hat{r}$ and $\hat{B}$ are two constants related to the retraction (defined in [18, Prop. 9][3]). Let $\mathbf{x}_0$ be a random initial point of Algorithm 1. Then, the set $\{\mathbf{x}_0 \in \mathcal{M}^N : \{\mathbf{x}_k\}$ converges to a point of $\mathcal{X}^*\}$ has measure 1.*

Proof. It is shown in [3] that all second-order critical points of problem (C-St) are global optima whenever $r \leq \frac{2}{3}d - 1$. Combining this and the Łojasiewicz inequality, we can use [18, Theorem 2, Corollary 6] to complete the proof. □

Theorem 2 needs the condition $r \leq \frac{2}{3}d-1$. For general d, r, we cannot prove global convergence, but we can characterize the landscape of the problem. The next theorem shows that when the states of neighboring agents are close enough to each other, any first-order critical point is global optimum.

Theorem 3. *Suppose that $\mathbf{x}$ is a first-order critical point of problem (C-St). Then, $\mathbf{x}$ is a global optimal point if and only if there exists some $y \in \mathbb{R}^{d\times r}$ (with $\|y\|_2 \leq 1$) such that $\langle x_i, y\rangle > r-1$ for all $i \in [N]$. Moreover, a first-order critical point $\mathbf{x}$ is a global optimal point if and only if*

$$\mathbf{x} \in \mathcal{L} := \{\mathbf{x} : \|\mathbf{x} - \bar{\mathbf{x}}\|_{F,\infty} < \sqrt{2}\}.$$

Proof. Let $B := W^t \otimes I_d$. The necessity is trivial by letting $y = [B\mathbf{x}]_i$ if $x_1 = x_2 = \ldots = x_N$. Now, if $\mathbf{x}$ is a first-order critical point, then it follows from Proposition 1 that

$$\begin{aligned}\mathrm{grad}\varphi_i^t(\mathbf{x}) &= \nabla\varphi_i^t(\mathbf{x}) - \frac{1}{2}x_i(x_i^\top\nabla\varphi_i^t(\mathbf{x}) + \nabla\varphi_i^t(\mathbf{x})^\top x_i)\\ &= (I_d - \frac{1}{2}x_ix_i^\top)(\nabla\varphi_i^t(\mathbf{x}) - x_i\nabla\varphi_i^t(\mathbf{x})^\top x_i) = 0, \quad \forall i \in [N].\end{aligned}$$

By our definition, we have $\forall i \in [N]$

$$\nabla\varphi_i^t(\mathbf{x}) - x_i\nabla\varphi_i^t(\mathbf{x})^\top x_i = -[B\mathbf{x}]_i + x_i([B\mathbf{x}]_i^\top x_i).$$

Note that since $I_d - \frac{1}{2}x_ix_i^\top$ is invertible, one has

$$[B\mathbf{x}]_i - x_i([B\mathbf{x}]_i^\top x_i) = 0, \quad \forall i \in [N]. \tag{IV.1}$$

Multiplying both sides by $x_i^\top$ yields

$$x_i^\top[B\mathbf{x}]_i = [B\mathbf{x}]_i^\top x_i, \quad \forall i \in [N]. \tag{IV.2}$$

For the sufficiency, let $\Gamma_i := \sum_{j=1}^N W_{ij}^t(x_j^\top x_i)$, $i \in [N]$. From (IV.1), we get

$$x_i\Gamma_i = \sum_{j=1}^N W_{ij}^t x_j, \quad \forall i \in [N]. \tag{IV.3}$$

Summing above over $i \in [N]$ yields $\sum_{i=1}^N x_i\Gamma_i = \sum_{i=1}^N x_i$. Taking inner product with y on both sides gives $\sum_{i=1}^N \langle y, x_i(I_r - \Gamma_i)\rangle = 0$. Note that $I_r - \Gamma_i$ is symmetric for all i due to (IV.2) and it is also positive semi-definite. Since $\langle x_i, y\rangle > r-1$ for all i, we get that $\Omega_i := \frac{1}{2}(x_i^\top y + y^\top x_i)$ is positive definite. Then, it follows that

$$\langle y, x_i(I_r - \Gamma_i)\rangle = \mathrm{Tr}(\Omega_i^{1/2}(I_r - \Gamma_i)\Omega_i^{1/2}) \geq 0.$$

The equation $\sum_{i=1}^N \langle y, x_i(I_r - \Gamma_i)\rangle = 0$ suggests that $I_r = \Gamma_i$, which also implies $x_1 = x_2 = \ldots = x_N$ by (IV.3).

[3]Specifically, they are given in Appendix.

Furthermore, suppose $y = \bar{x}$ which is the IAM of $\mathbf{x}$. The condition $\|\mathbf{x} - \bar{\mathbf{x}}\|_{F,\infty} < \sqrt{2}$ means that $\|\bar{x} - x_i\|_F^2 < 2$, or equivalently, $\langle y, x_i\rangle > r - 1$ for all $i \in [N]$. □

Theorem 3 establishes a sufficient and necessary condition for a first-order critical point to be global optimum. In Example 1 (see Appendix), we show that there exists a first-order critical point $\mathbf{x}$ satisfying $\max_{i\in[N]} \|x_i - \bar{x}\|_F = \sqrt{2}$ which is not global optimal. Therefore, the upper bound for the radius of $\mathcal{L}$ is also tight in Theorem 3.

When $r = 1$, the region $\mathcal{L}$ is the same as that of $\mathcal{S}$ defined in [13]. Specifically, on the sphere S^{d-1}, $\mathcal{S}$ corresponds to the hemisphere, which is *the largest convex set on* S^{d-1}. Geometrically, it means that x_i cannot be the antipode of any x_j, which is known as the *cut locus* [31]. However, the region $\mathcal{S}$ is unknown for general case $r > 1$. In [8], [13], [20], it was shown that the continuous Riemannian gradient flow starting in $\mathcal{L}$ converges to $\mathcal{X}^*$ on sphere S^{d-1} and the convergence rate is linear [8], [20]. However, it is still unclear whether an algorithm could achieve global consensus initialized in $\mathcal{L}$ when $r > 1$. The main challenge here is that the vanilla gradient method cannot guarantee that the sequence stays in $\|\mathbf{x} - \bar{\mathbf{x}}\|_{F,\infty} < \sqrt{2}$. In Lemma 4, we can also obtain the same result on S^{d-1} ($r = 1$) as that of [20], but we need a different proof since we work with Euclidean distance. The proof is provided in [37] due to the space limitation. The generalization to $r > 1$ is challenging and interesting for future study.

Lemma 4. *Let $r = 1$ and assume that there exists a $y \in \mathrm{St}(d, 1)$ such that the initial point $\mathbf{x}_0$ of Algorithm 1 satisfies $\langle x_{i,0}, y\rangle \geq \delta, \quad \forall i \in [N]$ for some $\delta > 0$. Then, the sequence $\{\mathbf{x}_k\}$ generated by Algorithm 1 with $\alpha \leq 1$ and $t \geq 1$ satisfies*

$$\langle x_{i,k}, y\rangle \geq \delta, \quad \forall i \in [N], \ \forall k \geq 0. \tag{IV.4}$$

Combining Lemma 3, Lemma 4 and Theorem 3, we have the following result. On the sphere, if the initial point $\mathbf{x}_0$ satisfies $\langle x_{i,0}, y\rangle > 0, \quad \forall i \in [N]$ for some $y \in \mathcal{M}$, the sequence $\{\mathbf{x}_k\}$ generated by Algorithm 1 with $t \geq 1$ and $0 < \alpha \leq \min\{1, \frac{1}{MG+L_t/2}\}$, where M, G are defined in Lemma 3, will converge to a point in $\mathcal{X}^*$, i.e., the sequence reaches a consensus state.

V. Local Linear Convergence

In this section, we study the local linear convergence rate of Algorithm 1 for general d, r. Typically, a local linear rate can be obtained if the Riemannian Hessian is non-singular at global optimal points. However, the Riemannian Hessian of $\varphi^t(\mathbf{x})$ is a linear operator. For any tangent vector $\eta^\top = [\eta_1^\top, \ldots, \eta_N^\top]$, we have [38]

$$\begin{aligned}\langle \eta, \mathrm{Hess}\varphi^t(\mathbf{x})[\eta]\rangle = \|\eta\|_F^2 - \sum_{i=1}^N\sum_{j=1}^N W_{ij}^t\langle\eta_i, \eta_j\rangle\\ - \sum_{i=1}^N\langle\eta_i, \eta_i(\frac{1}{2}[\nabla\varphi_i^t(\mathbf{x})^\top x_i + x_i^\top\nabla\varphi_i^t(\mathbf{x})])\rangle.\end{aligned} \tag{V.1}$$

Following [3], if we let $x_1 = \ldots = x_N$ and $\eta_i = \mathcal{P}_{\mathrm{T}_{x_i}\mathcal{M}}\xi$ for any $\xi \in \mathbb{R}^{d\times r}$, (V.1) becomes $0 = \sum_{i=1}^N \langle\eta_i, \mathrm{Hess}\varphi_i^t(\mathbf{x})[\eta_i]\rangle$. Therefore, similar to the Euclidean case, the Riemannian

Hessian at any consensus point has a zero eigenvalue. This motivates us to consider an alternative to the strong convexity. Luckily, there are more relaxed conditions (than strong convexity) for Euclidean problems.

We firstly present our main result, which is the local Q-linear convergence rate of Algorithm 1. Before proceeding, we define two local regions $\mathcal{N}_{R,t}$ and $\mathcal{N}_{l,t}$, where the local linear rate holds. $\mathcal{N}_{R,t}$ is given by

$$\mathcal{N}_{R,t} := \mathcal{N}_{1,t} \cap \mathcal{N}_{2,t}, \tag{V.2}$$

where

$$\mathcal{N}_{1,t} := \{\mathbf{x} : \|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}}^2 \leq N\delta_{1,t}^2\} \tag{V.3}$$
$$\mathcal{N}_{2,t} := \{\mathbf{x} : \|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F},\infty} \leq \delta_{2,t}\}, \tag{V.4}$$

and $\delta_{1,t}, \delta_{2,t}$ satisfy

$$\delta_{1,t} \leq \frac{1}{5\sqrt{r}}\delta_{2,t} \quad \textit{and} \quad \delta_{2,t} \leq \frac{1}{6}. \tag{V.5}$$

Define

$$\mu_t := 1 - \lambda_2(W^t), \tag{V.6}$$

where $\lambda_2(W^t)$ is the second largest eigenvalue of W^t. The region $\mathcal{N}_{l,t}$ is given by

$$\mathcal{N}_{l,t} := \{\mathbf{x} : \varphi^t(\mathbf{x}) \leq \frac{\mu_t}{4}\} \cap \{\mathbf{x} : \|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}}^2 \leq N\delta_{3,t}^2\}, \tag{V.7}$$

where $\delta_{3,t}$ satisfies

$$\delta_{3,t} \leq \min\Big\{\frac{1}{\sqrt{N}}, \frac{1}{4\sqrt{r}}\Big\}. \tag{V.8}$$

The upper bounds for constants $\delta_{1,t}, \delta_{2,t}$ may not be optimal since we use second-order approximation in the development of RSI and we need to guarantee that the DRCS iterates stay in the local region $\mathcal{N}_{R,t}$. However, Theorem 3 implies that the radius of $\mathcal{N}_{2,t}$ cannot be larger than $\sqrt{2}$, the radius of $\mathcal{L}$, which is the manifold property, while $\mathcal{N}_{l,t}$ is decided by the connectivity of the network. If the connectivity is stronger, then the region is larger. More discussions on the constants will be given in Remark 1.

Our main result of this section is presented in Theorem 4. The proof is given in Appendix. To prove it, two main steps will be established in the next two subsections: Section V-A and Section V-B. We use the following two constants in the presentation of the theorem

$$\gamma_t := \begin{cases} \gamma_{R,t} = (1 - 4r\delta_{1,t}^2)(1 - \frac{\delta_{2,t}^2}{2})\mu_t, & \mathbf{x} \in \mathcal{N}_{R,t} \\ \gamma_{l,t} = \mu_t(1 - 4r\delta_{3,t}^2) - \varphi^t(\mathbf{x}), & \mathbf{x} \in \mathcal{N}_{l,t}, \end{cases} \tag{V.9}$$

$$\Phi := \begin{cases} \Phi_R := 2 - \|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F},\infty}^2, & \mathbf{x} \in \mathcal{N}_{R,t} \\ \Phi_l := 2 - \|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}}^2, & \mathbf{x} \in \mathcal{N}_{l,t}. \end{cases} \tag{V.10}$$

Theorem 4. *Let Assumption 1 hold. (1). Let $\nu \in (0,1)$ and the stepsize α satisfy $0 < \alpha \leq \min\{\frac{\nu\Phi}{L_t}, 1, \frac{1}{M}\}$, where Φ is given in (V.10) and $t \geq \lceil \log_{\sigma_2}(\frac{1}{2\sqrt{N}}) \rceil$, and M is given in (P1). The sequence $\{\mathbf{x}_k\}$ in Algorithm 1 achieves consensus linearly if the initialization satisfies $\mathbf{x}_0 \in \mathcal{N}_{R,t}$ defined by (V.2). That is, we have $\mathbf{x}_k \in \mathcal{N}_{R,t}$ for all $k \geq 0$ and*

$$\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_F^2 \leq (1 - 2\alpha(1-\nu)\gamma_t)^k \|\mathbf{x}_0 - \bar{\mathbf{x}}_0\|_F^2, \tag{V.11}$$

where γ_t is defined in (V.9). Moreover, if $\alpha \leq \frac{2}{2MG + L_t}$, $\bar{\mathbf{x}}_k$ also converges to a single point.
(2). If $\mathbf{x}_0 \in \mathcal{N}_{l,t}$ and $\alpha \leq \min\{\frac{2}{L_t + 2MG}, \frac{\Phi}{L_t}\}$, one has (V.11) and $\mathbf{x}_k \in \mathcal{N}_{l,t}$ for all $k \geq 0, t \geq 1$, where Φ is defined in (V.10).

Theorem 4 has significant implications for various applications, such as synchronization in planetary-scale sensor networks [5], modeling of collective motion in flocks in the Earth's atmosphere [6], synchronization of quantum bits [7], and the Kuramoto models [3], [8], certifying their rapid convergence. Furthermore, this result sheds light on designing decentralized algorithms for Stiefel manifold optimization [39], as elaborated later in our experiments.

A. Restricted Secant Inequality

To prove Theorem 4, we need to establish a new RSI in the Riemannian form. Notice that Stiefel manifold is embedded in Euclidean space; we start with generalizing (III.7) to its Riemannian version as follows

$$\begin{aligned} &\big\langle \mathbf{x} - \mathcal{P}_{\mathrm{T}_{\mathbf{x}}\mathcal{M}^N}\bar{\mathbf{x}}, \mathrm{grad}\varphi^t(\mathbf{x}) \big\rangle \\ &\geq c_d \|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}}^2 + c_g \|\mathrm{grad}\varphi^t(\mathbf{x})\|_{\mathrm{F}}^2, \end{aligned} \tag{V.12}$$

where $c_d > 0, c_g > 0$ and $\mathbf{x}$ is in some neighborhood of $\mathcal{X}^*$. This is natural as for the Riemannian problem (C-St), we need to substitute the Euclidean gradient with Riemannian gradient. Moreover, the IAM $\bar{\mathbf{x}}$ should be mapped into the tangent space $\mathrm{T}_{\mathbf{x}}\mathcal{M}^N$. However, the map $\mathrm{Exp}_{\mathbf{x}}^{-1}(\bar{\mathbf{x}})$ is difficult to compute. Note that $\mathrm{Exp}_{\mathbf{x}}$ is a local diffeomorphism. By the inverse function theorem, we have $\mathrm{Exp}_{\mathbf{x}}^{-1}(\bar{\mathbf{x}}) = \bar{\mathbf{x}} - \mathbf{x} + \mathcal{O}(\|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}}^2)$. Using the property in (P3), we know that $\mathcal{P}_{\mathrm{T}_{\mathbf{x}}\mathcal{M}^N}(\bar{\mathbf{x}} - \mathbf{x})$ is a second-order approximation to $\mathrm{Exp}_{\mathbf{x}}^{-1}(\bar{\mathbf{x}})$. As such, we directly project $\bar{\mathbf{x}}$ onto the tangent space of $\mathbf{x}$ without recourse to the inverse of any retraction. Then, since

$$\begin{aligned} &\big\langle \mathbf{x} - \mathcal{P}_{\mathrm{T}_{\mathbf{x}}\mathcal{M}^N}\bar{\mathbf{x}}, \mathrm{grad}\varphi^t(\mathbf{x}) \big\rangle \\ &= \big\langle \mathcal{P}_{\mathrm{T}_{\mathbf{x}}\mathcal{M}^N}(\mathbf{x} - \bar{\mathbf{x}}), \mathcal{P}_{\mathrm{T}_{\mathbf{x}}\mathcal{M}^N}\nabla\varphi^t(\mathbf{x}) \big\rangle \\ &= \big\langle \mathbf{x} - \bar{\mathbf{x}}, \mathrm{grad}\varphi^t(\mathbf{x}) \big\rangle, \end{aligned}$$

we will get the following definition of RSI from (V.12)

$$\big\langle \mathbf{x} - \bar{\mathbf{x}}, \mathrm{grad}\varphi^t(\mathbf{x}) \big\rangle \geq c_d \|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}}^2 + c_g \|\mathrm{grad}\varphi^t(\mathbf{x})\|_{\mathrm{F}}^2. \tag{RSI}$$

To establish the (RSI), we first show the quadratic growth (QG) property of $\varphi^t(\mathbf{x})$ (Lemma 5). In the Euclidean space, especially for convex problems, QG condition is equivalent to the RSI as well as the Łojasiewicz inequality with $\theta = 1/2$ [40]. To the best of our knowledge, QG cannot be used directly to establish the linear rate of GD, and it is usually required to show the equivalence to Luo-Tseng [41] error bound inequality (ERB) [42]. However, for nonconvex problems, RSI is strictly stronger than QG. Detailed discussions are provided in Appendix VIII-C.

Lemma 5 (Quadratic growth)**.** *For any $t \geq 1$ and $\mathbf{x} \in \mathcal{M}^N$, we have that*

$$\varphi^t(\mathbf{x}) - \varphi^t(\bar{\mathbf{x}}) \geq \frac{\mu_t}{2}\|\mathbf{x} - \hat{\mathbf{x}}\|_F^2 \geq \frac{\mu_t}{4}\|\mathbf{x} - \bar{\mathbf{x}}\|_F^2. \tag{QG}$$

Moreover, if $\|\mathbf{x}-\bar{\mathbf{x}}\|_F^2 \le \frac{N}{8r}$, *we have*

$$\varphi^t(\mathbf{x}) - \varphi^t(\bar{\mathbf{x}}) \ge \frac{\mu_t}{2}(1-\frac{4r}{N}\|\mathbf{x}-\bar{\mathbf{x}}\|_F^2)\|\mathbf{x}-\bar{\mathbf{x}}\|_F^2. \quad \text{(QG')}$$

Proof. We rewrite the objective $\varphi^t(\mathbf{x})$ as follows

$$\begin{aligned} 2\varphi^t(\mathbf{x}) &= \sum_{i=1}^N \|x_i\|_{\mathrm{F}}^2 - \sum_{i=1,j=1}^N W_{ij}^t \langle x_i, x_j\rangle \\ &= \sum_{i=1}^N \langle x_i, x_i - \sum_{j=1}^N W_{ij}^t x_j\rangle = \left\langle \nabla\varphi^t(\mathbf{x}), \mathbf{x}\right\rangle. \end{aligned}$$

Note that as $\langle \nabla\varphi^t(\mathbf{x}), \hat{\mathbf{x}}\rangle = 0$, we get

$$\begin{aligned} 2\varphi^t(\mathbf{x}) &= \left\langle \nabla\varphi^t(\mathbf{x}), \mathbf{x}-\hat{\mathbf{x}}\right\rangle \\ &\overset{\text{(III.7)}}{\ge} \frac{\mu_t L_t}{\mu_t+L_t}\|\mathbf{x}-\hat{\mathbf{x}}\|_{\mathrm{F}}^2 + \frac{1}{\mu_t+L_t}\|\nabla\varphi^t(\mathbf{x})\|_{\mathrm{F}}^2 \\ &\ge \mu_t\|\mathbf{x}-\hat{\mathbf{x}}\|_{\mathrm{F}}^2, \end{aligned}$$

where the last inequality follows from $\|\nabla\varphi^t(\mathbf{x})\|_{\mathrm{F}} \ge \mu_t\|\mathbf{x}-\hat{\mathbf{x}}\|_{\mathrm{F}}$. Combining above with Lemma 2 and observing $\varphi^t(\bar{\mathbf{x}}) = 0$ completes the proof for both (QG) and (QG'). □

The second inequality (QG') is a local quadratic growth property, which is tighter than (QG).

Next, we discuss how to establish (RSI) based on Lemma 5. We will derive (RSI) in the separate forms

$$\left\langle \mathbf{x}-\bar{\mathbf{x}}, \mathrm{grad}\varphi^t(\mathbf{x})\right\rangle \ge c_d'\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2, \quad c_d' > 0 \qquad \text{(RSI-1)}$$

and

$$\left\langle \mathbf{x}-\bar{\mathbf{x}}, \mathrm{grad}\varphi^t(\mathbf{x})\right\rangle \ge c_g'\|\mathrm{grad}\varphi^t(\mathbf{x})\|_{\mathrm{F}}^2 \quad c_g' > 0. \qquad \text{(RSI-2)}$$

Then, (RSI) can be obtained by any convex combination of (RSI-1) and (RSI-2). To proceed with the analysis, we define for $i \in [N]$

$$p_i := \frac{1}{2}(x_i-\bar{x})^\top(x_i-\bar{x}), \qquad \text{(V.13)}$$

and

$$q_i := \frac{1}{2}\sum_{j=1}^N W_{ij}^t(x_i-x_j)^\top(x_i-x_j). \qquad \text{(V.14)}$$

Let $\mathbf{y} = \bar{\mathbf{x}}$ in (VIII.8)(see Appendix). We get

$$\begin{aligned} \left\langle \mathrm{grad}\varphi^t(\mathbf{x}), \mathbf{x}-\bar{\mathbf{x}}\right\rangle &= \left\langle \nabla\varphi^t(\mathbf{x}), \mathbf{x}-\bar{\mathbf{x}}\right\rangle - \sum_{i=1}^N \langle p_i, q_i\rangle \\ &= 2\varphi^t(\mathbf{x}) - \sum_{i=1}^N \langle p_i, q_i\rangle, \end{aligned} \qquad \text{(V.15)}$$

where in the last equation we used the following two identities $2\varphi^t(\mathbf{x}) = \langle \nabla\varphi^t(\mathbf{x}), \mathbf{x}\rangle$ and $\langle \nabla\varphi^t(\mathbf{x}), \bar{\mathbf{x}}\rangle = 0$. The term $\sum_{i=1}^N \langle p_i, q_i\rangle$ is non-negative, so if we substitute (V.15) into (RSI), we observe that RSI is stronger than QG. Moreover, by Cauchy-Schwarz inequality, we have

$$\sum_{i=1}^N \langle p_i, q_i\rangle \le \max_{i\in[N]}\|p_i\|_{\mathrm{F}} \cdot 2\varphi^t(\mathbf{x}) \le \varphi^t(\mathbf{x})\cdot\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F},\infty}^2. \qquad \text{(V.16)}$$

Hence, we see that if $\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F},\infty} < \sqrt{2}$, we have $\langle \mathrm{grad}\varphi^t(\mathbf{x}), \mathbf{x}-\bar{\mathbf{x}}\rangle > 0$, which implies that the direction $-\mathrm{grad}\varphi^t(\mathbf{x})$ is positively correlated with the direction $\bar{\mathbf{x}}-\mathbf{x}$. However, it is difficult to guarantee $\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|_{\mathrm{F},\infty} < \sqrt{2}$ for every k, since $\bar{\mathbf{x}}_k$ is not fixed. We will see in Lemma 7 that a large enough value for t can help us circumvent this problem in the region $\mathcal{N}_{2,t}$. Moreover, note that

$$\sum_{i=1}^N \langle p_i, q_i\rangle \le \varphi^t(\mathbf{x})\cdot\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2, \qquad \text{(V.17)}$$

so we can also establish (RSI-1) when $\varphi^t(\mathbf{x}) = O(\mu_t)$, as we will see in Lemma 6.

To conclude, the two inequalities (V.16) and (V.17) correspond to two neighborhoods of $\mathcal{X}^*$: $\mathcal{N}_{R,t}$ and $N_{l,t}$, which are defined in (V.2) and (V.7). The (RSI-1) is formally established in the following lemma.

Lemma 6. *Let* μ_t *be the constant given in* (V.6) *and* $t \ge 1$.

1) Suppose $\mathbf{x} \in \mathcal{N}_{R,t}$, *where* $\mathcal{N}_{R,t}$ *is defined by* (V.2). *There exists a constant* $\gamma_{R,t} > 0$ *defined in* (V.9):

$$\gamma_{R,t} := (1-4r\delta_{1,t}^2)(1-\frac{\delta_{2,t}^2}{2})\mu_t \ge \frac{\mu_t}{2},$$

such that the following holds:

$$\left\langle \mathbf{x}-\bar{\mathbf{x}}, \mathrm{grad}\varphi^t(\mathbf{x})\right\rangle \ge \gamma_{R,t}\|\bar{\mathbf{x}}-\mathbf{x}\|_F^2. \qquad \text{(V.18)}$$

2) For $\mathbf{x} \in \mathcal{N}_{l,t}$, *where* $\mathcal{N}_{l,t}$ *is defined by* (V.7), *we also have* (RSI-1), *in which* $c_d' = \gamma_{l,t} = \mu_t(1-4r\delta_{3,t}^2) - \varphi^t(\mathbf{x}) \ge \frac{\mu_t}{2}$.

Proof. (1). Combining (V.15) with (V.16), we get

$$\begin{aligned} \left\langle \mathbf{x}-\bar{\mathbf{x}}, \mathrm{grad}\varphi^t(\mathbf{x})\right\rangle &\overset{\text{(V.15)}}{=} 2\varphi^t(\mathbf{x}) - \sum_{i=1}^N \langle p_i, q_i\rangle \\ &\overset{\text{(V.16)}}{\ge} \varphi^t(\mathbf{x})\cdot(2-\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F},\infty}^2). \end{aligned} \qquad \text{(V.19)}$$

Since $\mathbf{x} \in \mathcal{N}_{R,t}$, invoking (QG') in Lemma 5, we get

$$\left\langle \mathbf{x}-\bar{\mathbf{x}}, \mathrm{grad}\varphi^t(\mathbf{x})\right\rangle \ge (1-4r\delta_{1,t}^2)(1-\frac{\delta_{2,t}^2}{2})\mu_t\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2,$$

where using the conditions (V.5) completes the proof.
(2). For $\mathbf{x} \in \mathcal{N}_{l,t}$, combining (V.15), (V.17) and (QG') yields

$$\begin{aligned} \left\langle \mathbf{x}-\bar{\mathbf{x}}, \mathrm{grad}\varphi^t(\mathbf{x})\right\rangle &\ge [\mu_t(1-4r\delta_{3,t}^2) - \varphi^t(\mathbf{x})]\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2 \\ &\ge \frac{1}{2}\mu_t\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2, \end{aligned} \qquad \text{(V.20)}$$

where we used the conditions in (V.8).

□

Remark 1. *In the proof, we derive* $\gamma_{R,t}$ *and* $\gamma_{l,t}$ *by combining* (QG') *with* (V.16) *and* (V.17), *respectively. For* (V.18), *any* $\delta_{1,t}, \delta_{2,t}$ *satisfying* $(1-4r\delta_{1,t}^2)(1-\frac{\delta_{2,t}^2}{2}) \ge 1/2$ *might seem sufficient. However, we impose the condition on* $\delta_{1,t}, \delta_{2,t}$ *in* (V.5) *in order to guarantee* $\mathbf{x}_k \in \mathcal{N}_{2,t}$ *for all* $k \ge 0$. *Moreover, we find that by combining* (QG) *with* (V.16), *one can also get* (RSI-1) *without the constraint* $\mathcal{N}_{1,t}$. *But the coefficient will be smaller. For simplicity, we only show the results in* $\mathcal{N}_{1,t} \cap \mathcal{N}_{2,t}$. *Similarly, for* $\mathcal{N}_{l,t}$, $\delta_{3,t} \le \frac{1}{4\sqrt{r}}$ *is enough to ensure RSI. We*

impose $\delta_{3,t} \leq 1/\sqrt{N}$ to get Proposition 2, which is useful to ensure $\{\mathbf{x}_k\}_k \in \mathcal{N}_{l,t}$. In fact, $\delta_{3,t} \leq 1/\sqrt{N}$ does not shrink the region since $\varphi^t(\mathbf{x}) \leq \mu_t/4$ implies a small region by Lemma 5. Also, since $\delta_{3,t} \leq 1/\sqrt{N}$, it is clear that $\mathcal{N}_{l,t}$ is smaller than $\mathcal{N}_{R,t}$ when N is large enough.

Next, we are ready to present the (RSI) inequality.

Proposition 2 (Restricted secant inequality)**.** *The following two inequalities hold for $\mathbf{x} \in \mathcal{N}_{R,t}$ and $\mathbf{x} \in \mathcal{N}_{l,t}$*

$$\left\langle \mathbf{x} - \bar{\mathbf{x}}, \operatorname{grad}\varphi^t(\mathbf{x}) \right\rangle \geq \frac{\Phi}{2L_t} \|\operatorname{grad}\varphi^t(\mathbf{x})\|_F^2, \tag{V.21}$$

and

$$\begin{aligned} &\left\langle \mathbf{x} - \bar{\mathbf{x}}, \operatorname{grad}\varphi^t(\mathbf{x}) \right\rangle \\ &\geq \nu \cdot \frac{\Phi}{2L_t} \|\operatorname{grad}\varphi^t(\mathbf{x})\|_F^2 + (1-\nu)\gamma_t \|\mathbf{x} - \bar{\mathbf{x}}\|_F^2, \end{aligned} \tag{RSI-I}$$

for any $\nu \in [0,1]$, where γ_t and $\Phi > 1$ are constants related to $\mathbf{x}$, which are given by (V.9) and (V.10).

B. Staying in the Local Region

Since the RSI condition holds in the local region $\mathcal{N}_{R,t}$, the main difficulty now is to show that $\mathbf{x}_k \in \mathcal{N}_{2,t}$. We can show that $\mathbf{x}_k$ always stays in $\mathcal{N}_{R,t} = \mathcal{N}_{1,t} \cap \mathcal{N}_{2,t}$ if the stepsize α satisfies $0 \leq \alpha \leq \min\{\frac{\Phi}{L_t}, 1, \frac{1}{M}\}$ and $t \geq \lceil \log_{\sigma_2}(\frac{1}{2\sqrt{N}}) \rceil$. The upper bounds $\frac{1}{M}$ and 1 are due to $\mathbf{x}_k \in \mathcal{N}_{2,t}$.

Lemma 7 (Stay in $\mathcal{N}_{R,t}$)**.** *Let $\mathbf{x}_k \in \mathcal{N}_{R,t}$, $0 \leq \alpha \leq \min\{\frac{\Phi}{L_t}, 1, \frac{1}{M}\}$ and $t \geq \lceil \log_{\sigma_2}(\frac{1}{2\sqrt{N}}) \rceil$, where the radius of $\mathcal{N}_{R,t}$ is given by (V.5) and M is given in (P1). We then have that $\mathbf{x}_{k+1} \in \mathcal{N}_{R,t}$.*

The lower bound $\lceil \log_{\sigma_2}(\frac{1}{2\sqrt{N}}) \rceil$ may not be a small number. For example, when W is the lazy Metropolis matrix of a regular connected graph, σ_2 usually scales as $1 - \mathcal{O}(\frac{1}{N^2})$ [43, Remark 2] and $\log_{\sigma_2}(\frac{1}{2\sqrt{N}}) = \mathcal{O}(N^2 \log N)$. However, for example, for a star graph this can be $\mathcal{O}(\log N)$. It will be interesting to investigate (as a future work) under what conditions Lemma 7 holds for $t = 1$. Here, we require this condition to ensure the algorithm is in a proper local neighborhood. From the above result, we see that the stepsize is upper bounded by $\frac{\Phi}{L_t}$ and $\frac{1}{M}$, and they show the role of the network and the manifold. The condition $\alpha \leq \Phi/L_t$ guarantees that $\mathbf{x}_k \in \mathcal{N}_{1,t}$ and $\alpha \leq \min\{1, 1/M\}$ ensures that $\mathbf{x}_k \in \mathcal{N}_{2,t}$. For the simplicity, we discuss the constant M in Section VIII-A in Appendix.

Combining Theorem 4 with Lemma 3 and Theorem 2, we conclude the following result.

Theorem 5. *When $\alpha < \min\{C_{\mathcal{M},\varphi^t}, \frac{\nu\Phi}{L_t}, 1\}$ and $r \leq \frac{2}{3}d - 1$, with random initialization, $\{\mathbf{x}_k\}$ firstly converges sub-linearly and then linearly for any $t \geq 1$, almost surely.*

The condition on the stepsize α depends on the function, network, and manifold properties, which is expected based on distributed optimization techniques in the Euclidean space. For the global convergence purpose, one could use the method in [44] to estimate L_t, μ_t in a distributed fashion, but the estimate of $C_{\mathcal{M},\varphi^t}$ is difficult to obtain. Therefore, similar to the Euclidean distributed optimization methods in practice, setting $\alpha = 1$ is a good starting point to apply the algorithm to real-world datasets, as demonstrated in Section VI. Otherwise, if $\alpha = 1$ does not converge, a non-exhaustive grid search can find a smaller stepsize ensuring convergence.

C. Asymptotic Rate

To get the rate of σ_2^t, we need to ensure $c_d = \frac{\mu_t L_t}{\mu_t + L_t}$ and $c_g = \frac{1}{\mu_t + L_t}$ in (RSI). We show this asymptotically for any $\mathbf{x} \in \mathcal{N}_{l,t}$. Firstly, by (V.15) we have

$$\left\langle \operatorname{grad}\varphi^t(\mathbf{x}), \mathbf{x} - \bar{\mathbf{x}} \right\rangle = \left\langle \nabla\varphi^t(\mathbf{x}), \mathbf{x} - \hat{\mathbf{x}} \right\rangle - \sum_{i=1}^{N} \langle p_i, q_i \rangle, \tag{V.22}$$

where p_i and q_i are given in (V.13)-(V.14). Using (III.7) and (III.10) yields

$$\begin{aligned} &\left\langle \nabla\varphi^t(\mathbf{x}), \mathbf{x} - \hat{\mathbf{x}} \right\rangle \\ &\geq \frac{\mu_t L_t}{\mu_t + L_t} \|\mathbf{x} - \hat{\mathbf{x}}\|_{\mathrm{F}}^2 + \frac{1}{\mu_t + L_t} \|\nabla\varphi^t(\mathbf{x})\|_{\mathrm{F}}^2 \\ &\geq \frac{\mu_t L_t}{\mu_t + L_t} (1 - \frac{4r}{N} \|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}}^2) \|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}}^2 \\ &\quad + \frac{1}{\mu_t + L_t} \|\operatorname{grad}\varphi^t(\mathbf{x})\|_{\mathrm{F}}^2, \end{aligned} \tag{V.23}$$

where we also used $\|\operatorname{grad}\varphi^t(\mathbf{x})\|_{\mathrm{F}} \leq \|\nabla\varphi^t(\mathbf{x})\|_{\mathrm{F}}$ by the non-expansiveness of $\mathcal{P}_{\mathrm{T}_{\mathbf{x}}\mathcal{M}^N}$. Substituting (V.23) into (V.22) and noting (V.17), we get

$$\begin{aligned} &\left\langle \operatorname{grad}\varphi^t(\mathbf{x}), \mathbf{x} - \bar{\mathbf{x}} \right\rangle \\ &\geq \frac{\mu_t L_t}{\mu_t + L_t} (1 - \frac{4r}{N} \|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}}^2 - \frac{\mu_t + L_t}{\mu_t L_t} \varphi^t(\mathbf{x})) \|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}}^2 \\ &\quad + \frac{1}{\mu_t + L_t} \|\operatorname{grad}\varphi^t(\mathbf{x})\|_{\mathrm{F}}^2. \end{aligned}$$

When $\|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}} \to 0$, we have $\varphi^t(\mathbf{x}) \to 0$ by Theorem 1. Thus, we get

$$c_d = \frac{\mu_t L_t}{\mu_t + L_t} (1 - \frac{4r}{N} \|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}}^2 - \frac{\mu_t + L_t}{\mu_t L_t} \varphi^t(\mathbf{x})) \to \frac{\mu_t L_t}{\mu_t + L_t}.$$

By the same arguments as of Theorem 4, we get the asymptotic rate being $\frac{L_t - \mu_t}{L_t + \mu_t}$ with $\alpha = \frac{2}{L_t + \mu_t}$, and $\frac{L_t - \mu_t}{L_t + \mu_t} \leq \sigma_2^t$. Also, using similar arguments as (VIII.3) in Appendix, we can get the rate of σ_2^t with $\alpha = 1$ as the Euclidean case by noting that the error bound inequality (ERB) in Appendix is asymptotically $\mu_t \|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}} \leq \|\operatorname{grad}\varphi^t(\mathbf{x})\|_{\mathrm{F}}$.

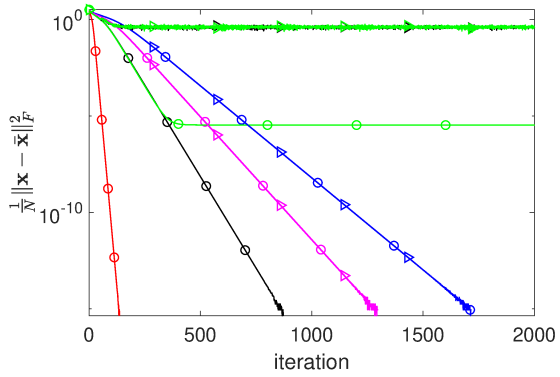
VI. Numerical Experiments

A. Consensus Simulation

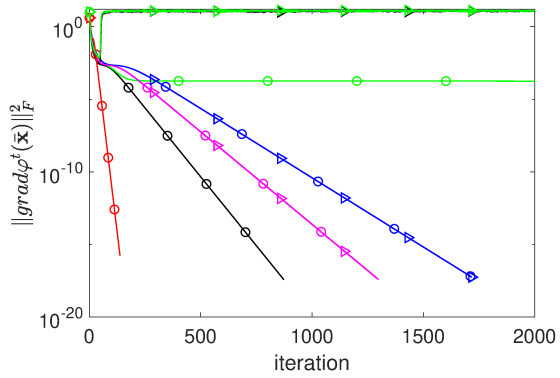
We now provide the numerical experiments by evaluating our method on a ring graph with $N = 30$ nodes. The matrix W is given as follows: $W_{ii} = 1/3$ for all $i \in \{1, ..., 30\}$; $W_{ij} = 1/3$ if i and j are neighbors and $W_{ij} = 0$ otherwise.

We compare the polar retraction and the exponential map using different stepsizes and $t \geq 1$. For $t = 1$, we run Algorithm 1 with four choices of stepsize: $1/L_t, 2/(L_t + \mu_t), 2/L_t, 1$. For $t = 10$, we only use $\alpha = 1$ for simplicity. The algorithms are stopped when we reach the target accuracy of $\frac{1}{N}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 \leq 2 \times 10^{-16}$. The dimension of the variable is

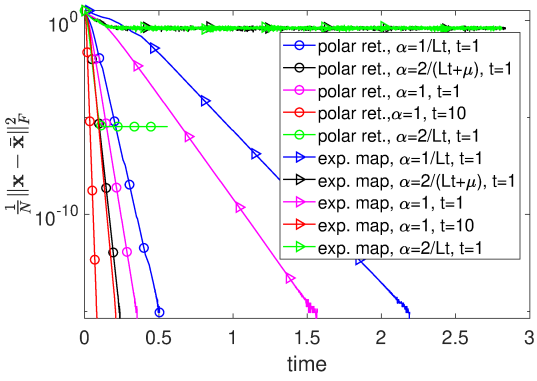
This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2023.3330735



(a) distance vs iteration



(b) gradient vs iteration



(c) distance vs time

Fig. 1: Numerical results for $N = 30, d = 5, r = 2$. All three sub-figures use the same legends as that in Figure (c).

$d = 5, r = 2$. The initial points are sampled independently from an identical uniform distribution. In Fig. 1, we have $L_t = 1 - \lambda_{\min}(W^t) = \frac{4}{3}$ when $t = 1$.

Fig. 1 (a) presents the convergence of log-scale distance $\frac{1}{N}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2$, and Fig 1 (b) shows the log-scale $\|\mathrm{grad}\varphi^t(\mathbf{x})\|_{\mathrm{F}}^2$ versus the iteration number. We see that Algorithm 1 with $\alpha = 2/L_t$ does not converge to a critical point for the polar retraction and the exponential map, which is consistent with the stepsize range in Lemma 3. We also observe that $\alpha = 2/(\mu_t + L_t)$ produces the fastest convergence for the polar retraction (when $t = 1$), but the exponential map does not converge for this value of stepsize. This is because they have different constants M, which changes the permissible stepsize range, and for the exponential maps, this range is smaller. When the chosen stepsize can guarantee the convergence, the polar retraction and the exponential map perform similarly in terms of the iteration numbers required to reach a certain accuracy level. Moreover, the convergence rate of $\alpha = 1, t = 10$ (red lines) is about 10 times of that of $\alpha = 1, t = 1$ (pink lines). In the Fig. 1 (c), we demonstrate the convergence versus the CPU time. We see that the polar retraction is always faster than the exponential map. Finally, we remark that although Algorithm 1 with $\alpha = \frac{2}{\mu_t + L_t}$ converges fast, it takes additional effort to obtain L_t, μ_t. Therefore, we recommend using $\alpha = 1$ as a default starting point in practice.

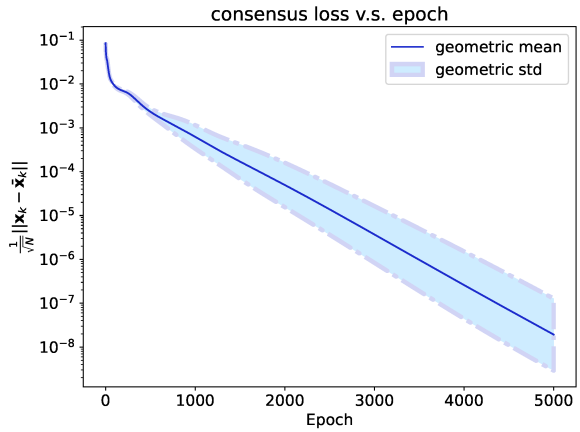


Fig. 2: DRGTA algorithm for PCA problem: consensus error v.s. algorithm epoch number.

B. Application to Decentralized Optimization

We next illustrate the importance of linear rate for solving decentralized optimization problems on the Stiefel manifold. In a follow-up work [39], we proposed two decentralized Riemannian gradient algorithms, where consensus plays a key role in their convergence. We focus on one of them here, namely decentralized Riemannian gradient tracking algorithm (DRGTA), and we apply that to solve the decentralized principle component analysis (PCA) problem. Specifically, we solve the following PCA problem using DRGTA:

$$\min_{\mathbf{x}\in\mathcal{M}^N} -\frac{1}{2N}\sum_{i=1}^{N}\left\langle x_i, A_i^\top A_i x_i\right\rangle, \quad \text{s.t.} \quad x_1 = \ldots = x_N, \tag{VI.1}$$

where for agent $i \in [N]$, $A_i \in \mathbb{R}^{m_i \times d}$ denotes the local data matrix and m_i is the sample size. Denote the global data matrix by $A := [A_1^\top \ A_2^\top \ \ldots \ A_N^\top]^\top$. The data matrix A is given by the MNIST dataset [45], with $\sum_{i=1}^N m_i = 60000$ samples and $d = 784$. The experiment is conducted in Python with mpi4py, on a single Intel I9 13900KF CPU with 24 cores.

DRGTA is a manifold optimization algorithm in which agents alternate between performing Riemannian gradient updates (outer loop) and DRCS algorithm (inner loop). For a

ring graph with $N = 20$ nodes, in Fig. 2, we show the consensus error $\frac{1}{\sqrt{N}}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}$ with respect to the epoch number (iteration of the outer loop) of DRGTA for the case of $r = 5, t = 1$. The results represent the average of error for 20 random initializations. In each run, we randomly initialize all agents variables $x_1, x_2, \ldots, x_N$, such that they satisfy the condition of linear convergence in Theorem 4. Note that $x_1 = x_2 = \ldots = x_N = x_0$ also satisfies the condition. Fig. 2 depicts the geometric mean and the geometric standard deviation of the consensus errors. Given this empirical linear rate observed in Fig. 2, if the convergence of DRCS was slower than linear, it would have exacerbated the overall performance of DRGTA, making the rate sublinear. Therefore, the linear rate achieved by DRCS could play a key role in the design of consensus-based decentralized manifold optimization techniques. Theorem 4 facilitates the convergence rate analysis of the two decentralized Riemannian gradient methods in [39]. We also present the computation time cost of each part of DRGTA for 5000 epochs in Table I, including the computation of consensus gradient $\nabla\varphi^t(\mathbf{x}_k)$ in the inner loop, the gradient of PCA in the outer loop, the retraction operation in the outer loop, and the projection onto tangent space for both loops. These results are the arithmetic mean and standard deviation of 20 experiment runs. The time cost of the consensus gradient $\nabla\varphi^t(\mathbf{x}_k)$ is negligible in DRGTA, certifying the low computational overhead of DRCS.

TABLE I. COMPUTATION TIME OF DRGTA, IN SECONDS.

	$\nabla\varphi^t(\mathbf{x}_k)$	gradient of PCA	retraction	$\mathcal{P}_{\mathrm{T}_{\mathcal{M}^N}}$
mean(std)	0.29(0.32)	43.56(4.02)	60.79(7.51)	1.02(0.30)

VII. Conclusion

In this paper, we provided the global and local convergence analysis of DRCS, a distributed method for consensus on the Stiefel manifold. We showed that the convergence rate asymptotically matches the Euclidean counterpart, which scales with the second largest singular value of the communication matrix. The main technical contribution is to generalize the Euclidean restricted secant inequality to the Riemannian version. In the future work, we would like to study the preservation of iteration in the region $\mathcal{N}_{2,t}$ (with $t = 1$) and to estimate the constant $C_{\mathcal{M},\varphi^t}$ for stepsize.

VIII. Appendix

A. More Discussions on Constants

– Lipschitz Constant L_t in Theorem 1: We remark that a closely related inequality is the restricted Lipschitz-type gradient presented in [22, Lemma 4], which is defined by the pull back function $g(\xi) := \varphi^t(\mathrm{Retr}_{\mathbf{x}}(\xi))$, whose Lipschitz constant $\tilde{L}$ relies on the retraction and the Lipschitz constant of Euclidean gradient. Also, the stepsize of RGM in [22] depends on the norm of Euclidean gradient. Therefore, the range of our stepsize is larger than that in [22, Theorem 5]. Our inequality does not rely on the retraction, which could be of independent interest.

– Constants $\hat{r}$ and $\hat{B}$ in Lemma 2: The two constants $\hat{r}$ and $\hat{B}$ in Lemma 2 are directly obtained from the proof of [18, Prop. 9]. For the completeness, we introduce them here. Firstly, since $\mathcal{M}$ is Stiefel manifold, the polar retraction is unique and smooth in a neighborhood of radius $\hat{r} = 1$ of the manifold [46]. Secondly, define $h_{\mathbf{x}}(\alpha) = \det(D\mathrm{Retr}_{\mathbf{x}}(-\alpha\mathrm{grad}\varphi(\mathbf{x}))(I_d - \alpha D(\mathcal{P}_{\mathrm{T}_{\mathbf{x}}\mathcal{M}^N})(\mathbf{x})))$. Since $\mathcal{M}^N$ is a compact smooth manifold, letting $\alpha < \frac{\hat{r}}{\max_{\mathbf{x}\in\mathcal{M}^N}\|\nabla\varphi(\mathbf{x})\|_{\mathrm{F}}}$, $\mathrm{Retr}_{\mathbf{x}}(-\alpha\mathrm{grad}\varphi(\mathbf{x}))$ and its derivatives exist. Then,

$$\hat{B} := \max_{\mathbf{x}\in\mathcal{M}^N,\alpha} |\frac{dh_{\mathbf{x}}}{d\alpha}(\alpha)| < \infty, \text{s.t. } \alpha < \frac{\hat{r}}{\max_{\mathbf{x}\in\mathcal{M}^N}\|\nabla\varphi(\mathbf{x})\|_{\mathrm{F}}}.$$

– Constant M in (P1): We have $M = 1$ in (P1) for the polar retraction if $\alpha\|\mathrm{grad}\varphi^t(x_{i,k})\|_{\mathrm{F}} \leq 1$ according to [25, Append. E]. By our choice of $\alpha \leq 1$ and $\mathbf{x}_k \in \mathcal{N}_{R,t}$, we indeed have $\alpha\|\mathrm{grad}\varphi^t(x_{i,k})\|_{\mathrm{F}} \leq 2\delta_{2,t} \leq 1$ according to Lemma 8. However, we do not plan to remove the term $\frac{1}{M}$.

B. Proofs

Proof of inequality (III.7). Without loss of generality, we assume $d = r = 1$. Let $U_1, U_2, \ldots, U_N$ be the orthonormal eigenvectors of $I_N - W$, corresponding to the eigenvalues $0 = \lambda_1 < \lambda_2 \leq \ldots \leq \lambda_N$. Then, we have that $\mathbf{x} - \hat{\mathbf{x}} = \sum_{i=1}^N c_i U_i$. Since $\mathbf{x} - \hat{\mathbf{x}}$ is orthogonal to $\mathrm{span}\{U_1\}$, we have $c_1 = 0$. Note that $\nabla\varphi(\mathbf{x}) = (I_N - W)\mathbf{x} = (I_N - W)(\mathbf{x} - \hat{\mathbf{x}})$. We get

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_{\mathrm{F}}^2 = \sum_{i=2}^N c_i^2 \quad \text{and} \quad \|\nabla\varphi(\mathbf{x})\|_{\mathrm{F}}^2 = \sum_{i=2}^N c_i^2\lambda_i^2. \quad \text{(VIII.1)}$$

Then, to prove (III.7), we have that

$$\begin{aligned}
&\langle \mathbf{x} - \hat{\mathbf{x}}, \nabla\varphi(\mathbf{x})\rangle = \langle \mathbf{x} - \hat{\mathbf{x}}, (I_N - W)(\mathbf{x} - \hat{\mathbf{x}})\rangle \\
&= \left\langle \sum_{i=2}^N c_i U_i, \sum_{i=2}^N c_i\lambda_i U_i \right\rangle = \sum_{i=2}^N c_i^2\lambda_i \geq \frac{1}{L+\mu}\sum_{i=2}^N (\mu L c_i^2 + c_i^2\lambda_i^2) \\
&= \frac{\mu L}{\mu + L}\|\mathbf{x} - \hat{\mathbf{x}}\|_{\mathrm{F}}^2 + \frac{1}{\mu + L}\|\nabla\varphi(\mathbf{x})\|_{\mathrm{F}}^2, \qquad \text{(VIII.2)}
\end{aligned}$$

where the inequality follows due to $\mu = \lambda_2$ and $L = \lambda_N$. □

Proof of linear rate of PGD with $\alpha_e = 1$. Firstly, one can easily verify $L\|\mathbf{x} - \hat{\mathbf{x}}\|_{\mathrm{F}} \geq \|\nabla\varphi(\mathbf{x})\|_{\mathrm{F}} \geq \mu\|\mathbf{x} - \hat{\mathbf{x}}\|_{\mathrm{F}}$ using (VIII.1). We then have

$$\begin{aligned}
&\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}_{k+1}\|_{\mathrm{F}}^2 \leq \|\mathbf{x}_{k+1} - \hat{\mathbf{x}}_k\|_{\mathrm{F}}^2 \\
&\leq \|\mathbf{x}_k - \hat{\mathbf{x}}_k\|_{\mathrm{F}}^2 + \|\nabla\varphi(\mathbf{x}_k)\|_{\mathrm{F}}^2 - 2\langle\nabla\varphi(\mathbf{x}_k), \mathbf{x}_k - \hat{\mathbf{x}}_k\rangle \\
&\overset{\text{(III.7)}}{\leq} (1 - \frac{2\mu L}{\mu + L})\|\mathbf{x}_k - \hat{\mathbf{x}}_k\|_{\mathrm{F}}^2 + (1 - \frac{2}{\mu + L})\|\nabla\varphi(\mathbf{x}_k)\|_{\mathrm{F}}^2. \qquad \text{(VIII.3)}
\end{aligned}$$

If $\frac{2}{\mu+L} \geq 1$, i.e., $\lambda_2(W) + \lambda_N(W) \geq 0$, this implies $\sigma_2 = \lambda_2(W)$. Combining $\|\nabla\varphi(\mathbf{x})\|_{\mathrm{F}} \geq \mu\|\mathbf{x} - \hat{\mathbf{x}}\|_{\mathrm{F}}$ with (VIII.3) yields

$$\begin{aligned}
&\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}_k\|_{\mathrm{F}}^2 \\
&\leq (1 - \frac{2\mu L}{\mu + L} + \mu^2 - \frac{2\mu^2}{L + \mu})\|\mathbf{x}_k - \hat{\mathbf{x}}_k\|_{\mathrm{F}}^2 \\
&= (1 - \mu)^2\|\mathbf{x}_k - \hat{\mathbf{x}}_k\|_{\mathrm{F}}^2 = \sigma_2^2\|\mathbf{x}_k - \hat{\mathbf{x}}_k\|_{\mathrm{F}}^2.
\end{aligned}$$

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2023.3330735

If $\frac{2}{\mu+L} < 1$, then $\lambda_2(W) + \lambda_N(W) < 0$, this implies $\sigma_2 = -\lambda_N(W)$. Combining $\|\nabla\varphi(\mathbf{x})\|_{\mathrm{F}} \le L\|\mathbf{x}-\hat{\mathbf{x}}\|_{\mathrm{F}}$ with (VIII.3) implies

$$\begin{aligned}\|\mathbf{x}_{k+1}-\hat{\mathbf{x}}_k\|_{\mathrm{F}}^2 &\le (1-\frac{2\mu L}{\mu+L}+L^2-\frac{2L^2}{L+\mu})\|\mathbf{x}_k-\hat{\mathbf{x}}_k\|_{\mathrm{F}}^2\\ &= (1-L)^2\|\mathbf{x}_k-\hat{\mathbf{x}}_k\|_{\mathrm{F}}^2 = \sigma_2^2\|\mathbf{x}_k-\hat{\mathbf{x}}_k\|_{\mathrm{F}}^2.\end{aligned}$$

□

Proof of Lemma 1. Let $usv^\top = \hat{x}$ be the singular value decomposition and $\hat{\sigma}_1 \ge \ldots \ge \hat{\sigma}_r \ge 0$ be the singular values of $\hat{x}$. Since $\bar{x} = \mathcal{P}_{\mathcal{M}}(\hat{x}) = uv^\top$, we get

$$\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2 = \sum_{i=1}^{N}(2r - 2\langle x_i, \bar{x}\rangle) = 2N(r - \langle \hat{x}, \bar{x}\rangle) = 2N(r - \|\hat{x}\|_*), \tag{VIII.4}$$

where $\|\cdot\|_*$ is the trace norm. Hence, by assumption $\mathrm{dist}^2(\mathbf{x}, \mathcal{X}^*) = \frac{1}{N}\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2 < 2$ and (VIII.4), we get $\|\hat{x}\|_* > r-1$.

Noticing that $\hat{\sigma}_i \in [0,1]$ for all $i \in [r]$, we get the smallest singular value $\hat{\sigma}_r > 0$. Therefore, $\hat{x}$ has full rank and $\mathcal{P}_{\mathcal{M}}(\hat{x})$ is unique.

□

Proof of Lemma 2. Note that

$$\begin{aligned}\|\mathbf{x}-\hat{\mathbf{x}}\|_{\mathrm{F}}^2 &= \sum_{i=1}^{N}\|x_i-\hat{x}\|_{\mathrm{F}}^2 = N(r-\|\hat{x}\|_{\mathrm{F}}^2)\\ &= N(\sqrt{r}+\|\hat{x}\|_{\mathrm{F}})(\sqrt{r}-\|\hat{x}\|_{\mathrm{F}}) \le 2N(r-\sqrt{r}\|\hat{x}\|_{\mathrm{F}}),\end{aligned} \tag{VIII.5}$$

where the inequality is due to $\|\hat{x}\|_{\mathrm{F}} \le \sqrt{r}$.

Let $\hat{\sigma}_1 \ge \ldots \ge \hat{\sigma}_r \ge 0$ be the singular values of $\hat{x}$. It is clear that $\hat{\sigma}_1 \le 1$ since $\|\hat{x}\|_2 \le \frac{1}{N}\sum_{i=1}^{N}\|x_i\|_2 \le 1$. The inequality $\|\hat{x}\|_* = \sum_{i=1}^{r}\hat{\sigma}_i \le \sqrt{r}\sqrt{\sum_{i=1}^{r}\hat{\sigma}_i^2} = \sqrt{r}\|\hat{x}\|_{\mathrm{F}}$, together with (VIII.5) and (VIII.4) imply that $\|\mathbf{x}-\hat{\mathbf{x}}\|_{\mathrm{F}}^2 \le \|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2$. Next, we have $\|\hat{x}\|_* = \sum_{i=1}^{r}\hat{\sigma}_i \ge \sum_{i=1}^{r}\hat{\sigma}_i^2 = \|\hat{x}\|_{\mathrm{F}}^2$. This yields

$$\frac{1}{2}\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2 = N(r-\|\hat{x}\|_*) \le N(r-\|\hat{x}\|_{\mathrm{F}}^2) = \|\mathbf{x}-\hat{\mathbf{x}}\|_{\mathrm{F}}^2,$$

which proves (III.9).

By utilizing the fact $\|\mathbf{x}-\hat{\mathbf{x}}\|_{\mathrm{F}} \le \|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}$ in (III.9), we have

$$\sqrt{r\sum_{i=1}^{r}\hat{\sigma}_i^2} = \sqrt{r}\|\hat{x}\|_{\mathrm{F}} \ge \|\hat{x}\|_{\mathrm{F}}^2 = r - \frac{1}{N}\|\hat{\mathbf{x}}-\mathbf{x}\|_{\mathrm{F}}^2 \ge r - \frac{1}{N}\|\bar{\mathbf{x}}-\mathbf{x}\|_{\mathrm{F}}^2, \tag{VIII.6}$$

where we used $\|\hat{x}\|_{\mathrm{F}} = \|\frac{1}{N}\sum_{i=1}^{N}x_i\|_{\mathrm{F}} \le \sqrt{r}$. If $\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2 \le N/2$ (by assumption), we can square both sides of above and note $\hat{\sigma}_i^2 \le 1$ for $i \in [r-1]$ to get

$$\hat{\sigma}_r^2 \ge 1 - 2\frac{\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2}{N} + \frac{\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^4}{N^2 r} \ge 1 - 2\frac{\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2}{N}.$$

Then, we have

$$\hat{\sigma}_r \ge \sqrt{1-2\frac{\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2}{N}} \ge 1 - 2\frac{\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2}{N}, \tag{VIII.7}$$

where we used $\sqrt{1-s} \ge 1-s$ for any $1 \ge s \ge 0$. Recall that $\bar{x} = \mathcal{P}_{\mathcal{M}}(\hat{x}) = uv^\top$. Hence, it follows that

$$\|\hat{x}-\bar{x}\|_{\mathrm{F}}^2 = \sum_{i=1}^{r}(1-\hat{\sigma}_i)^2 \le \frac{4r\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^4}{N^2}.$$

Hence, we have proved (P2). Finally,

$$\begin{aligned}\|\mathbf{x}-\hat{\mathbf{x}}\|_{\mathrm{F}}^2 &= \sum_{i=1}^{N}\langle x_i-\hat{x}, x_i-\hat{x}\rangle\\ &= \sum_{i=1}^{N}\langle x_i-\hat{x}, x_i-\bar{x}\rangle + \sum_{i=1}^{N}\langle x_i-\hat{x}, \bar{x}-\hat{x}\rangle\\ &= \|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2 + \sum_{i=1}^{N}\langle \bar{x}-\hat{x}, x_i-\bar{x}\rangle\\ &= \|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2 - N\|\bar{x}-\hat{x}\|_{\mathrm{F}}^2\\ &\overset{\text{(P2)}}{\ge} \|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^2 - \frac{4r\|\mathbf{x}-\bar{\mathbf{x}}\|_{\mathrm{F}}^4}{N},\end{aligned}$$

where we used $\sum_{i=1}^{N}\langle x_i-\hat{x}, \bar{x}-\hat{x}\rangle = 0$ in the third line. □

Proof of Theorem 1. We firstly show that for any $\mathbf{x}, \mathbf{y} \in \mathcal{M}^N$, we have

$$\begin{aligned}&\langle \mathrm{grad}\varphi^t(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle = \langle \nabla\varphi^t(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle +\\ &\frac{1}{4}\sum_{i=1}^{N}\langle\sum_{j=1}^{N}W_{ij}^t(x_i-x_j)^\top(x_i-x_j), (y_i-x_i)^\top(y_i-x_i)\rangle\\ &\ge \langle \nabla\varphi^t(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle.\end{aligned} \tag{VIII.8}$$

It follows from the relationship (P3) that

$$\begin{aligned}&\langle \mathrm{grad}\varphi^t(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle = \langle \nabla\varphi^t(\mathbf{x}), \mathcal{P}_{\mathrm{T}_{\mathbf{x}}\mathcal{M}^N}(\mathbf{y}-\mathbf{x})\rangle\\ &= \langle \nabla\varphi^t(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle - \sum_{i=1}^{N}\langle \nabla\varphi_i^t(\mathbf{x}), \mathcal{P}_{N_{x_i}\mathcal{M}}(y_i-x_i)\rangle\\ &= \langle \nabla\varphi^t(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle\\ &\quad + \frac{1}{4}\sum_{i=1}^{N}\langle \nabla\varphi_i^t(\mathbf{x})^\top x_i + x_i^\top\nabla\varphi_i^t(\mathbf{x}), (y_i-x_i)^\top(y_i-x_i)\rangle.\end{aligned}$$

Since

$$\frac{1}{2}[\nabla\varphi_i^t(\mathbf{x})^\top x_i + x_i^\top\nabla\varphi_i^t(\mathbf{x})] = \frac{1}{2}\sum_{j=1}^{N}W_{ij}^t(x_i-x_j)^\top(x_i-x_j),$$

is positive semi-definite, we get

$$\sum_{i=1}^{N}\left\langle \nabla\varphi_i^t(\mathbf{x}), \frac{1}{2}x_i(y_i-x_i)^\top(y_i-x_i)\right\rangle \ge 0. \tag{VIII.9}$$

Therefore, we get (VIII.8). Note that the largest eigenvalue of $\nabla^2\varphi^t(\mathbf{x}) = (I_N - W^t)\otimes I_d$ is $L_t = 1-\lambda_N(W^t)$ in Euclidean space, where $\lambda_N(W^t)$ denotes the smallest eigenvalue of W^t. For any $\mathbf{x}, \mathbf{y} \in \mathcal{M}^N$, it follows that [29]

$$\varphi^t(\mathbf{y}) - \left[\varphi^t(\mathbf{x}) + \langle \nabla\varphi^t(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle\right] \le \frac{L_t}{2}\|\mathbf{y}-\mathbf{x}\|_{\mathrm{F}}^2. \tag{VIII.10}$$

Together with (VIII.8), this implies that

$$\varphi^t(\mathbf{y}) - \left[\varphi^t(\mathbf{x}) + \langle \mathrm{grad}\varphi^t(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle\right] \le \frac{L_t}{2}\|\mathbf{x}-\mathbf{y}\|_{\mathrm{F}}^2. \tag{VIII.11}$$

The proof is completed. □

Example 1. *For any number $N > 1$ and any $1 \le r \le d-2$, consider the fully connected network with the weight matrix*

$$W = \begin{pmatrix} 1/N & 1/N & \dots & 1/N \\ 1/N & 1/N & \dots & 1/N \\ \vdots & \vdots & \vdots & \vdots \\ 1/N & 1/N & \dots & 1/N \end{pmatrix}.$$

Let the initial point $\mathbf{x} \in \mathcal{M}^N$ satisfy the following conditions:

- *The first $r-1$ columns of $x_1, \dots, x_N$ are identical and can be represented as a matrix denoted by $b \in \mathbb{R}^{d\times(r-1)}$.*
- *The last column of $x_1, \dots, x_N$ is denoted by $a_1, \dots, a_N$, respectively. That is, we have*

$$x_i = [b \ a_i] \quad i = 1, \dots, N.$$

- *Assume that $\sum_{i=1}^N a_i = 0$ and that $a_1, a_2, \dots, a_N$ span a two-dimensional subspace $\mathcal{A}$.*

Then, we have $\hat{x} = [b \ 0]$. Note that $\mathcal{A}$ lies in $\mathbb{R}^{d-(r-1)}$ and $d-(r-1) \ge 3$. By definition of IAM, we know that the first $r-1$ columns of $\bar{x}$ form a matrix equal to b. Let z denote the last column of $\bar{x}$, where z is a unit vector that is orthogonal to b and $\mathcal{A}$. It follows that $\|\bar{x} - x_i\|_F = \sqrt{2}$ for all $i = 1, ..., N$. Note that

$$\begin{aligned}\mathrm{grad}\varphi_i^t(\mathbf{x}) &= \mathcal{P}_{\mathrm{T}_{x_i}\mathcal{M}}(x_i - \sum_{j=1}^N W_{ij}^t x_j) = -\mathcal{P}_{\mathrm{T}_{x_i}\mathcal{M}} \sum_{j=1}^N W_{ij}^t x_j \\ &= -\frac{1}{N}\mathcal{P}_{\mathrm{T}_{x_i}\mathcal{M}} \sum_{j=1}^N x_j = -\mathcal{P}_{\mathrm{T}_{x_i}\mathcal{M}}\hat{x} = 0.\end{aligned}$$

Hence, $\mathbf{x}$ is a first-order critical point, $\max_{i\in[N]} \|x_i - \bar{x}\|_F = \sqrt{2}$, but it is not global optimum since $x_i \ne x_j$ for all i, j.

Proofs for Section V. To prove Proposition 2, we need the following bounds for $\mathrm{grad}\varphi^t(\mathbf{x})$ by noting that $\varphi^t(\mathbf{x})$ is Lipschitz smooth as shown in Theorem 1. The following lemma will be helpful to show (RSI-2).

Lemma 8. *For any $\mathbf{x} \in \mathcal{M}^N$, it follows that*

$$\|\sum_{i=1}^N \mathrm{grad}\varphi_i^t(\mathbf{x})\|_F \le L_t \|\mathbf{x} - \bar{\mathbf{x}}\|_F^2, \tag{VIII.12}$$

and

$$\|\mathrm{grad}\varphi^t(\mathbf{x})\|_F^2 \le 2L_t \cdot \varphi^t(\mathbf{x}), \tag{VIII.13}$$

where L_t is the Lipschitz constant given in Theorem 1. Moreover, suppose $\mathbf{x} \in \mathcal{N}_{2,t}$, where $\mathcal{N}_{2,t}$ is defined in (V.4). We then have

$$\max_{i\in[N]} \|\mathrm{grad}\varphi_i^t(\mathbf{x})\|_F \le 2\delta_{2,t}. \tag{VIII.14}$$

Proof of Lemma 8. First, using (P3) we have

$$\mathrm{grad}\varphi_i^t(\mathbf{x}) = x_i - \sum_{j=1}^N W_{ij}^t x_j - \frac{1}{2}x_i \sum_{j=1}^N W_{ij}^t (x_i - x_j)^\top (x_i - x_j). \tag{VIII.15}$$

Since $\sum_{i=1}^N \nabla\varphi_i^t(\mathbf{x}) = \sum_{i=1}^N (x_i - \sum_{j=1}^N W_{ij}^t x_j) = 0$, we have

$$\begin{aligned}\|\sum_{i=1}^N \mathrm{grad}\varphi_i^t(\mathbf{x})\|_{\mathrm{F}} &= \frac{1}{2}\|\sum_{i=1}^N x_i \sum_{j=1}^N W_{ij}^t (x_i - x_j)^\top (x_i - x_j)\|_{\mathrm{F}} \\ &\le \frac{1}{2}\sum_{i=1}^N \|\sum_{j=1}^N W_{ij}^t (x_i - x_j)^\top (x_i - x_j)\|_{\mathrm{F}} \\ &\le \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N W_{ij}^t \|x_i - x_j\|_{\mathrm{F}}^2 = 2\varphi^t(\mathbf{x}) \le L_t \|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}}^2,\end{aligned}$$

where the last inequality follows from (III.11). Moreover, it is clear that we have

$$\begin{aligned}0 &\le \varphi^t(\mathbf{x} - \frac{1}{L_t}\nabla\varphi^t(\mathbf{x})) \\ &\overset{\text{(VIII.10)}}{\le} \varphi^t(\mathbf{x}) + \langle \nabla\varphi^t(\mathbf{x}), -\frac{1}{L_t}\nabla\varphi^t(\mathbf{x})\rangle + \frac{1}{2L_t}\|\nabla\varphi^t(\mathbf{x})\|_{\mathrm{F}}^2 \\ &= \varphi^t(\mathbf{x}) - \frac{1}{2L_t}\|\nabla\varphi^t(\mathbf{x})\|_{\mathrm{F}}^2.\end{aligned}$$

Since $\mathrm{grad}\varphi_i^t(\mathbf{x}) = \mathcal{P}_{\mathrm{T}_{x_i}\mathcal{M}}(\nabla\varphi_i^t(\mathbf{x}))$, we get

$$\|\mathrm{grad}\varphi^t(\mathbf{x})\|_{\mathrm{F}}^2 \le \|\nabla\varphi^t(\mathbf{x})\|_{\mathrm{F}}^2 \le 2L_t \cdot \varphi^t(\mathbf{x}).$$

Finally, it follows from $\mathbf{x} \in \mathcal{N}_{2,t}$ that

$$\|\mathrm{grad}\varphi_i^t(\mathbf{x})\|_{\mathrm{F}} \le \|\sum_{j=1}^N W_{ij}^t (x_j - x_i)\|_{\mathrm{F}} \le 2\delta_{2,t}.$$

□

Proof of Proposition 2. First, we prove it for $\mathbf{x} \in \mathcal{N}_{R,t}$. It follows from (V.15) and (V.16) that

$$\langle \mathbf{x} - \bar{\mathbf{x}}, \mathrm{grad}\varphi^t(\mathbf{x})\rangle \ge \Phi_R \cdot \varphi^t(\mathbf{x}).$$

Combining with (VIII.13), we get $\langle \mathbf{x} - \bar{\mathbf{x}}, \mathrm{grad}\varphi^t(\mathbf{x})\rangle \ge \frac{\Phi_R}{2L_t}\|\mathrm{grad}\varphi^t(\mathbf{x})\|_{\mathrm{F}}^2$.
Secondly, for $\mathbf{x} \in \mathcal{N}_{l,t}$, we have the similar arguments by combining (V.15) with (V.17). Furthermore, if $\mathbf{x} \in \mathcal{N}_{R,t}$ or $\mathbf{x} \in \mathcal{N}_{l,t}$, we notice that (RSI-I) is the convex combination of (V.21) and (V.18). □

We have the following bound in (VIII.16) for the total variation distance between any row of W^t and the uniform distribution.

Lemma 9. *Given any $\mathbf{x} \in \mathcal{N}_{2,t}$, where $\mathcal{N}_{2,t}$ is defined in (V.4), if $t \ge \lceil \log_{\sigma_2}(\frac{1}{2\sqrt{N}})\rceil$, we have*

$$\max_{i\in[N]} \|\sum_{j=1}^N (W_{ij}^t - 1/N)x_j\|_F \le \frac{\delta_{2,t}}{2}. \tag{VIII.16}$$

Proof. Note that W^t is doubly stochastic with σ_2^t as the second largest singular value. As $\mathbf{x} \in \mathcal{N}_2$, it follows that $\|x_i - \bar{x}\|_{\mathrm{F}} \le \delta_{2,t}$ for all $i \in [N]$. We then have

$$\begin{aligned}&\max_{i\in[N]} \|\sum_{j=1}^N (W_{ij}^t - 1/N)x_j\|_{\mathrm{F}} \\ &= \max_{i\in[N]} \|\sum_{j=1}^N (W_{ij}^t - 1/N)(x_j - \bar{x})\|_{\mathrm{F}} \\ &\le \max_{i\in[N]} \sum_{j=1}^N |W_{ij}^t - 1/N|\delta_{2,t} \le \sqrt{N}\sigma_2^t\delta_{2,t},\end{aligned}$$

where the last inequality follows from the bound on the total variation distance between any row of W^t and $\frac{1}{N}\mathbf{1}_N^\top$ [47,

Prop.3] [24, Sec 1.1.2]. The conclusion is obtained by setting $t \geq \lceil \log_{\sigma_2}(\frac{1}{2\sqrt{N}}) \rceil$. □

Following a perturbation lemma of the polar decomposition [48, Theorem 2.4], we get the following technical lemma, which will be useful to bound the Euclidean distance between two consecutive points $\bar{x}_k$ and $\bar{x}_{k+1}$.

Lemma 10. *Suppose $\mathbf{x}, \mathbf{y} \in \mathcal{N}_{1,t}$, we have*

$$\|\bar{x} - \bar{y}\|_F \leq \frac{1}{1-2\delta_{1,t}^2}\|\hat{x} - \hat{y}\|_F,$$

where $\bar{x}$ and $\bar{y}$ are the IAM of $x_1, \ldots, x_N$ and $y_1, \ldots, y_N$, respectively.

Proof. Let $\hat{x} = \frac{1}{N}\sum_{i=1}^N x_i$ and $\hat{y} = \frac{1}{N}\sum_{i=1}^N y_i$ be the Euclidean average points of $\mathbf{x}$ and $\mathbf{y}$. Then, $\bar{x}$ and $\bar{y}$ are the (generalized) polar factor [48] of $\hat{x}$ and $\hat{y}$, respectively. We have

$$\sigma_r(\hat{x}) \overset{\text{(VIII.7)}}{\geq} 1 - 2\frac{\|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}}^2}{N} \overset{(i)}{\geq} 1 - 2\delta_{1,t}^2 > 0,$$

where (i) follows from $\mathbf{x} \in \mathcal{N}_{1,t}$. Similarly, we have $\sigma_r(\hat{y}) \geq 1 - 2\delta_{1,t}^2$ since $\mathbf{y} \in \mathcal{N}_{1,t}$.
Then, it follows from [48, Theorem 2.4] that

$$\|\bar{y} - \bar{x}\|_{\mathrm{F}} \leq \frac{2}{\sigma_r(\hat{x}) + \sigma_r(\hat{y})}\|\hat{y} - \hat{x}\|_{\mathrm{F}} \leq \frac{1}{1-2\delta_{1,t}^2}\|\hat{x} - \hat{y}\|_{\mathrm{F}}.$$

The proof is completed. □

We use Lemma 10 for the following lemma.

Lemma 11. *If $\mathbf{x}_k \in \mathcal{N}_{R,t}, \mathbf{x}_{k+1} \in \mathcal{N}_{1,t}$ and $x_{i,k+1} = \mathrm{Retr}_{x_{i,k}}(-\alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k))$, where $\delta_{1,t}$ and $\delta_{2,t}$ are given by (V.5), it follows that*

$$\|\bar{x}_k - \bar{x}_{k+1}\|_F \leq \frac{L_t}{1-2\delta_{1,t}^2}\frac{\alpha + 2M\alpha^2 L_t}{N}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_F^2.$$

Proof. We have that

$$\begin{aligned}
&\|\hat{x}_k - \hat{x}_{k+1}\|_{\mathrm{F}} \leq \|\hat{x}_k - \frac{\alpha}{N}\sum_{i=1}^N \mathrm{grad}\varphi_i^t(\mathbf{x}_k) - \hat{x}_{k+1}\|_{\mathrm{F}} \\
&+ \|\frac{\alpha}{N}\sum_{i=1}^N \mathrm{grad}\varphi_i^t(\mathbf{x}_k)\|_{\mathrm{F}} \\
&\overset{\text{(P1)}}{\leq} \frac{M}{N}\sum_{i=1}^N \|\alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k)\|_{\mathrm{F}}^2 + \alpha\|\frac{1}{N}\sum_{i=1}^N \mathrm{grad}\varphi_i^t(\mathbf{x}_k)\|_{\mathrm{F}} \\
&\overset{\text{(VIII.12)}}{\leq} \frac{2L_t^2 M\alpha^2 + L_t\alpha}{N}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2.
\end{aligned}$$

Therefore, it follows from Lemma 10 that

$$\begin{aligned}
\|\bar{x}_k - \bar{x}_{k+1}\|_{\mathrm{F}} &\leq \frac{1}{1-2\delta_{1,t}^2}\cdot\|\hat{x}_k - \hat{x}_{k+1}\|_{\mathrm{F}} \\
&\leq \frac{L_t}{1-2\delta_{1,t}^2}\frac{\alpha + 2M\alpha^2 L_t}{N}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2.
\end{aligned}$$

□

Using Lemma 9 and Lemma 11, we can prove Lemma 7.

Proof of Lemma 7. First, we verify that $\mathbf{x}_{k+1} \in \mathcal{N}_{1,t}$. Since $\mathbf{x}_k \in \mathcal{N}_{R,t}$, it follows from Lemma 6 that

$$\begin{aligned}
&\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|_{\mathrm{F}}^2 \leq \|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 \\
&\overset{\text{(II.3)}}{\leq} \sum_{i=1}^N \|x_{i,k} - \alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k) - \bar{x}_k\|_{\mathrm{F}}^2 \\
&= \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 - 2\alpha\left\langle \mathrm{grad}\varphi^t(\mathbf{x}_k), \mathbf{x}_k - \bar{\mathbf{x}}_k\right\rangle + \|\alpha \mathrm{grad}\varphi^t(\mathbf{x}_k)\|_{\mathrm{F}}^2 \\
&\overset{\text{(RSI-I)}}{\leq} (1 - 2\alpha(1-\nu)\gamma_{R,t})\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 \\
&\quad + \left(\alpha^2 - \frac{\alpha\nu\Phi_R}{L_t}\right)\|\mathrm{grad}\varphi^t(\mathbf{x}_k)\|_{\mathrm{F}}^2,
\end{aligned} \tag{VIII.17}$$

for any $\nu \in [0,1]$, where the last inequality holds by noting $\Phi_R \geq 1$ for $\mathbf{x} \in \mathcal{N}_{R,t}$. By letting $\nu = 1$ and $\alpha \leq \frac{\Phi_R}{L_t}$, we get

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|_{\mathrm{F}}^2 \leq \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2. \tag{VIII.18}$$

and thus $\mathbf{x}_{k+1} \in \mathcal{N}_{1,t}$.
Next, let us verify $\mathbf{x}_{k+1} \in \mathcal{N}_{2,t}$. For each $i \in [N]$, one has

$$\begin{aligned}
&\|x_{i,k+1} - \bar{x}_k\|_{\mathrm{F}} \\
&\overset{\text{(II.3)}}{\leq} \|x_{i,k} - \alpha \mathrm{grad}\varphi_i^t(\mathbf{x}_k) - \bar{x}_k\|_{\mathrm{F}} \\
&\overset{\text{(VIII.15)}}{=} \|(1-\alpha)(x_{i,k} - \bar{x}_k) + \alpha(\hat{x}_k - \bar{x}_k) + \alpha\sum_{j=1}^N W_{ij}^t(x_{j,k} - \hat{x}_k) \\
&\quad + \frac{\alpha}{2}x_{i,k}\sum_{j=1}^N W_{ij}^t(x_{i,k} - x_{j,k})^\top(x_{i,k} - x_{j,k})\|_{\mathrm{F}} \\
&\leq (1-\alpha)\delta_{2,t} + \alpha\|\hat{x}_k - \bar{x}_k\|_{\mathrm{F}} + \alpha\|\sum_{j=1}^N (W_{ij}^t - \frac{1}{N})x_{j,k}\|_{\mathrm{F}} \\
&\quad + \frac{1}{2}\|\alpha\sum_{j=1}^N W_{ij}^t(x_{i,k} - x_{j,k})^\top(x_{i,k} - x_{j,k})\|_{\mathrm{F}} \\
&\overset{\text{(P2)}}{\leq} (1-\alpha)\delta_{2,t} + 2\alpha\delta_{1,t}^2\sqrt{r} + \alpha\|\sum_{j=1}^N (W_{ij}^t - \frac{1}{N})x_{j,k}\|_{\mathrm{F}} \\
&\quad + 2\alpha\delta_{2,t}^2 \\
&\overset{\text{(VIII.16)}}{\leq} (1 - \frac{\alpha}{2})\delta_{2,t} + 2\alpha\delta_{1,t}^2\sqrt{r} + 2\alpha\delta_{2,t}^2.
\end{aligned}$$

Since $\alpha \geq 0$, by invoking Lemma 11 we get

$$\|\bar{x}_k - \bar{x}_{k+1}\|_{\mathrm{F}} \leq L_t \cdot \frac{2M\alpha^2 L_t + \alpha}{N(1-2\delta_{1,t}^2)}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2 \leq \frac{10\alpha\delta_{1,t}^2}{1-2\delta_{1,t}^2},$$

where the last inequality follows from $\alpha \leq \frac{1}{M}$ and $L_t \leq 2$. Therefore, using the conditions on $\delta_{1,t}$ and $\delta_{2,t}$ in (V.5) gives

$$\begin{aligned}
&\|x_{i,k+1} - \bar{x}_{k+1}\|_{\mathrm{F}} \leq \|x_{i,k+1} - \bar{x}_k\|_{\mathrm{F}} + \|\bar{x}_k - \bar{x}_{k+1}\|_{\mathrm{F}} \\
&\leq (1 - \frac{\alpha}{2})\delta_{2,t} + 2\alpha\delta_{1,t}^2\sqrt{r} + 2\alpha\delta_{2,t}^2 + \frac{10}{1-2\delta_{1,t}^2}\alpha\delta_{1,t}^2 \leq \delta_{2,t}.
\end{aligned}$$

The proof is completed. □

Proof of Theorem 4. Now, we are ready to prove Theorem 4. (1). Since $0 < \alpha \leq \min\{1, \frac{\Phi}{L_t}, \frac{1}{M}\}$. By Lemma 7, we have $\mathbf{x}_k \in \mathcal{N}_{R,t}$ for all $k \geq 0$. By choosing any $\nu \in (0,1)$ and $\alpha \leq \frac{\nu\Phi}{L_t}$, we get from (VIII.17) that

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|_{\mathrm{F}}^2 \leq (1 - 2\alpha(1-\nu)\gamma_{R,t})\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_{\mathrm{F}}^2. \tag{VIII.19}$$

We know that $\mathbf{x}_k$ converges to the optimal set $\mathcal{X}^*$ Q-linearly. Furthermore, if $\alpha \leq \frac{2}{2MG + L_t}$, it follows from Lemma 3 that the limit point of $\mathbf{x}_k$ is unique. Hence, $\bar{\mathbf{x}}_k$ also converges to

a single point.

(2). If $\mathbf{x}_k \in \mathcal{N}_{l,t}$, we have the constant $\Phi = 2 - \frac{1}{2}\|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}}^2 > 1$ in Proposition 2. Since $\alpha \leq \min\{\frac{2}{L_t+2MG}, \frac{\Phi}{L_t}\}$, we have $\mathbf{x}_{k+1} \in \mathcal{N}_{l,t}$ by using the sufficient decrease condition in [37][Lemma 5 (A1)]. The remaining proof follows the same argument of (1). □

C. Discussion on RSI, Quadratic Growth, Error bound and Łojasiewicz Inequality

Lemma 6 implies that the following error bound inequality holds for $\mathbf{x} \in \mathcal{N}_{R,t}$ and $\mathbf{x} \in \mathcal{N}_{l,t}$

$$\|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}} \leq \frac{2}{\mu_t}\|\mathrm{grad}\varphi^t(\mathbf{x})\|_{\mathrm{F}}. \tag{ERB}$$

This inequality is a generalization of the Luo-Tseng error bound [41] for problems in Euclidean space. In [40], the following holds for smooth non-convex problems

$$\text{RSI} \Rightarrow \text{ERB} \Leftrightarrow \text{Łojasiewicz inequality with } \theta = 1/2 \Rightarrow \text{QG}.$$

However, in Euclidean space and for convex problems, they are all equivalent. RSI can be used to show the Q-linear rate of $\mathrm{dist}(\mathbf{x}, \mathcal{X}^*)$, and ERB can be used to establish the Q-linear rate of the objective value and the R-linear rate of $\mathrm{dist}(\mathbf{x}, \mathcal{X}^*)$. Moreover, under mild assumptions QG and ERB are shown to be equivalent for second-order critical points for Euclidean nonconvex problems [49]. Some other error bound inequalities are also obtained over the Stiefel manifold or oblique manifold. For example, Liu et al. [25] established the error bound inequality of any first-order critical point for the eigenvector problem. Our proof of Lemma 6 relies mainly on the doubly stochasticity of W^t and the properties of IAM, and it is fundamentally different from previous works. Another similar form of RSI is the Riemannian regularity condition proposed in [50] for minimizing the nonsmooth problems over the Stiefel manifold.

Following the same argument as [25], the error bound inequality (ERB) implies a growth inequality similar to the Łojasiewicz inequality. However, the neighborhoods $\mathcal{N}_{R,t}$ and $\mathcal{N}_{l,t}$ are relative to the set $\mathcal{X}^*$, which is different from the Definition 2. It can be used to show the Q-linear rate of $\{\varphi^t(\mathbf{x}_k)\}$ only if $\mathbf{x}_k \in \mathcal{N}_{R,t}$ or $\mathbf{x}_k \in \mathcal{N}_{l,t}$ can be guaranteed.

Proposition 3. *For any $\mathbf{x} \in \mathcal{N}_{R,t}$ or $\mathbf{x} \in \mathcal{N}_{l,t}$ it holds that*

$$\varphi^t(\mathbf{x}) \leq \frac{3}{2\mu_t}\|\mathrm{grad}\varphi^t(\mathbf{x})\|_F^2. \tag{VIII.20}$$

Proof. By (V.15), we get

$$\begin{aligned} 2\varphi^t(\mathbf{x}) &= \left\langle \mathrm{grad}\varphi^t(\mathbf{x}), \mathbf{x} - \bar{\mathbf{x}} \right\rangle + \sum_{i=1}^{N} \langle p_i, q_i \rangle \\ &\overset{\text{(ERB)}}{\leq} \frac{2}{\mu_t}\|\mathrm{grad}\varphi^t(\mathbf{x})\|_{\mathrm{F}}^2 + \sum_{i=1}^{N} \langle p_i, q_i \rangle . \end{aligned} \tag{VIII.21}$$

If $\mathbf{x} \in \mathcal{N}_{R,t}$, we use (V.16) to get $(2 - \delta_{2,t}^2)\varphi^t(\mathbf{x}) \leq \frac{2}{\mu_t}\|\mathrm{grad}\varphi^t(\mathbf{x})\|_{\mathrm{F}}^2$. If $\mathbf{x} \in \mathcal{N}_{l,t}$, we use (V.17) to get

$$2\varphi^t(\mathbf{x}) \leq \frac{2}{\mu_t}\|\mathrm{grad}\varphi^t(\mathbf{x})\|_{\mathrm{F}}^2 + \frac{\mu_t}{4}\|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathrm{F}}^2 \overset{\text{(ERB)}}{\leq} \frac{3}{\mu_t}\|\mathrm{grad}\varphi^t(\mathbf{x})\|_{\mathrm{F}}^2.$$

We conclude the proof by noting $\delta_{2,t} \leq 1/6$. □

REFERENCES

[1] J. Markdahl, "Synchronization on Riemannian manifolds: Multiply connected implies multistable," *IEEE Transactions on Automatic Control*, vol. 66, no. 9, pp. 4311–4318, 2021.

[2] J. Markdahl, J. Thunberg, and J. Goncalves, "Almost global consensus on the n-sphere," *IEEE Transactions on Automatic Control*, vol. 63, no. 6, pp. 1664–1675, 2017.

[3] J. Markdahl, J. Thunberg, and J. Goncalves, "High-dimensional Kuramoto models on Stiefel manifolds synchronize complex networks almost globally," *Automatica*, vol. 113, p. 108736, 2020.

[4] J. Markdahl, "A geometric obstruction to almost global synchronization on Riemannian manifolds," *arXiv preprint arXiv:1808.00862*, 2018.

[5] D. A. Paley, "Stabilization of collective motion on a sphere," *Automatica*, vol. 45, no. 1, pp. 212–216, 2009.

[6] S. Al-Abri, W. Wu, and F. Zhang, "A gradient-free three-dimensional source seeking strategy with robustness analysis," *IEEE Transactions on Automatic Control*, vol. 64, no. 8, pp. 3439–3446, 2018.

[7] M. Lohe, "Quantum synchronization over quantum networks," *Journal of Physics A: Mathematical and Theoretical*, vol. 43, no. 46, p. 465301, 2010.

[8] A. Sarlette and R. Sepulchre, "Synchronization on the circle," *The complexity of dynamical systems: a multi-disciplinary perspective*, 2011.

[9] 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 (ICML)*, pp. 1594–1605, 2021.

[10] 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.

[11] M. Arjovsky, A. Shah, and Y. Bengio, "Unitary evolution recurrent neural networks," in *International Conference on Machine Learning*, pp. 1120–1128, PMLR, 2016.

[12] E. Vorontsov, C. Trabelsi, S. Kadoury, and C. Pal, "On orthogonality and learning recurrent networks with long term dependencies," in *International Conference on Machine Learning*, pp. 3570–3578, PMLR, 2017.

[13] R. Tron, B. Afsari, and R. Vidal, "Riemannian consensus for manifolds with bounded curvature," *IEEE Transactions on Automatic Control*, vol. 58, no. 4, pp. 921–934, 2012.

[14] S. Bonnabel, "Stochastic gradient descent on Riemannian manifolds," *IEEE Transactions on Automatic Control*, vol. 58, no. 9, pp. 2217–2229, 2013.

[15] A. Sarlette and R. Sepulchre, "Consensus optimization on manifolds," *SIAM Journal on Control and Optimization*, vol. 48, no. 1, pp. 56–76, 2009.

[16] R. Sepulchre, "Consensus on nonlinear spaces," *Annual reviews in control*, vol. 35, no. 1, pp. 56–64, 2011.

[17] A. Sarlette, S. E. Tuna, V. D. Blondel, and R. Sepulchre, "Global synchronization on the circle," *IFAC Proceedings Volumes*, vol. 41, no. 2, pp. 9045–9050, 2008.

[18] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, "First-order methods almost always avoid strict saddle points," *Math. Program.*, vol. 176, p. 311–337, 2019.

[19] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[20] C. Lageman and Z. Sun, "Consensus on spheres: Convergence analysis and perturbation theory," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 19–24, IEEE, 2016.

[21] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM journal on Matrix Analysis and Applications*, vol. 20, no. 2, pp. 303–353, 1998.

[22] N. Boumal, P.-A. Absil, and C. Cartis, "Global rates of convergence for nonconvex optimization on manifolds," *IMA Journal of Numerical Analysis*, vol. 39, no. 1, pp. 1–33, 2019.

[23] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[24] S. Boyd, P. Diaconis, and L. Xiao, "Fastest mixing Markov chain on a graph," *SIAM review*, vol. 46, no. 4, pp. 667–689, 2004.

[25] H. Liu, A. M.-C. So, and W. Wu, "Quadratic optimization with orthogonality constraint: Explicit Łojasiewicz exponent and linear convergence of retraction-based line-search and stochastic variance-reduced gradient methods," *Mathematical Programming Series A*, vol. 178, no. 1-2, pp. 215–262, 2019.

[26] 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.

[27] J. Tsitsiklis, *Problems in decentralized decision making and computation*. PhD thesis, MIT, 1984.

[28] 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.

[29] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87. Springer Science & Business Media, 2013.

[30] M. Moakher, "Means and averaging in the group of rotations," *SIAM journal on matrix analysis and applications*, vol. 24, no. 1, pp. 1–16, 2002.

[31] B. Afsari, "Riemannian l^p center of mass: existence, uniqueness, and convexity," *Proceedings of the American Mathematical Society*, vol. 139, no. 2, pp. 655–673, 2011.

[32] B. Afsari, R. Tron, and R. Vidal, "On the convergence of gradient descent for finding the Riemannian center of mass," *SIAM Journal on Control and Optimization*, vol. 51, no. 3, pp. 2230–2260, 2013.

[33] K. Grove and H. Karcher, "How to conjugatec 1-close group actions," *Mathematische Zeitschrift*, vol. 132, no. 1, pp. 11–20, 1973.

[34] H. Karcher, "Riemannian center of mass and mollifier smoothing," *Communications on pure and applied mathematics*, vol. 30, no. 5, pp. 509–541, 1977.

[35] P.-A. Absil, R. Mahony, and B. Andrews, "Convergence of the iterates of descent methods for analytic cost functions," *SIAM Journal on Optimization*, vol. 16, no. 2, pp. 531–547, 2005.

[36] R. Schneider and A. Uschmajew, "Convergence results for projected line-search methods on varieties of low-rank matrices via Łojasiewicz inequality," *SIAM Journal on Optimization*, vol. 25, no. 1, pp. 622–646, 2015.

[37] 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.

[38] P.-A. Absil, R. Mahony, and J. Trumpf, "An extrinsic look at the Riemannian hessian," in *International Conference on Geometric Science of Information*, pp. 361–368, Springer, 2013.

[39] S. Chen, A. Garcia, M. Hong, and S. Shahrampour, "Decentralized Riemannian gradient descent on the Stiefel manifold," *International Conference on Machine Learning*, 2021.

[40] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811, Springer, 2016.

[41] Z.-Q. Luo and P. Tseng, "Error bounds and convergence analysis of feasible descent methods: a general approach," *Annals of Operations Research*, vol. 46, no. 1, pp. 157–178, 1993.

[42] D. Drusvyatskiy and A. S. Lewis, "Error bounds, quadratic growth, and linear convergence of proximal methods," *Mathematics of Operations Research*, vol. 43, no. 3, pp. 919–948, 2018.

[43] A. Nedic, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2597–2633, 2017.

[44] M. Yang and C. Y. Tang, "Distributed estimation of graph spectrum," in *2015 American Control Conference (ACC)*, pp. 2703–2708, IEEE, 2015.

[45] Y. LeCun, "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*.

[46] P.-A. Absil and J. Malick, "Projection-like retractions on matrix manifolds," *SIAM Journal on Optimization*, vol. 22, no. 1, pp. 135–158, 2012.

[47] P. Diaconis and D. Stroock, "Geometric bounds for eigenvalues of Markov chains," *The Annals of Applied Probability*, pp. 36–61, 1991.

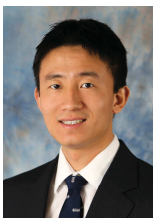
[48] W. Li and W. Sun, "Perturbation bounds of unitary and subunitary polar factors," *SIAM journal on matrix analysis and applications*, vol. 23, no. 4, pp. 1183–1193, 2002.

[49] M.-C. Yue, Z. Zhou, and A. Man-Cho So, "On the quadratic convergence of the cubic regularization method under a local error bound condition," *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 904–932, 2019.

[50] 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," in *Advances in Neural Information Processing Systems*, pp. 9442–9452, 2019.

Shixiang Chen received the Ph.D. degree in Systems Engineering and Engineering Management from The Chinese University of Hong Kong in July, 2019. He is an assistant professor in the School of Mathematical Sciences, University of Science and Technology of China. He was a postdoctoral associate in the Department of Industrial & Systems Engineering at Texas A&M University. His current research interests include design and analysis of optimization algorithms, and their applications in machine learning and signal processing.

Alfredo Garcia received the Degree in electrical engineering from the Universidad de los Andes, Bogotá,Colombia,in 1991,the Diplôme d'Etudes Approfondies in automatic control from the Université Paul Sabatier, Toulouse, France, in 1992, and the Ph.D. degree in industrial and operations engineering from the University of Michigan, Ann Arbor, MI, USA, in 1997. From 1997 to 2001, he was a consultant to government agencies and private utilities in the electric power industry. From 2001 to 2015, he was a Faculty with the Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, USA. From 2015 to 2017, he was a Professor with the Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, USA. In 2018, he joined the Department of Industrial and System Engineering, Texas A&M University, College Station, TX, USA. His research interests include game theory and dynamic optimization with applications in communications and energy networks.



Mingyi Hong received his Ph.D. degree from the University of Virginia, Charlottesville, in 2011. He is an assoicate professor in the Department of Electrical and Computer Engineering at the University of Minnesota, Minneapolis. He serves on the IEEE Signal Processing for Communications and Networking and Machine Learning for Signal Processing Technical Committees. His research interests include optimization theory and applications in signal processing and machine learning. He is a Member of the IEEE.

Shahin Shahrampour received the Ph.D. degree in Electrical and Systems Engineering, the M.A. degree in Statistics (The Wharton School), and the M.S.E. degree in Electrical Engineering, all from the University of Pennsylvania, in 2015, 2014, and 2012, respectively. He is currently an Assistant Professor in the Department of Mechanical and Industrial Engineering at Northeastern University. His research interests include machine learning, optimization, sequential decision-making, and distributed learning, with a focus on developing computationally efficient methods for data analytics. He is a Senior Member of the IEEE.