WEAKLY CONVEX OPTIMIZATION OVER STIEFEL MANIFOLD USING RIEMANNIAN SUBGRADIENT-TYPE METHODS*

XIAO LI†, SHIXIANG CHEN‡, ZENGDE DENG§, QING QU¶, ZHIHUI ZHU‖, AND ANTHONY MAN-CHO SO#

Abstract. We consider a class of nonsmooth optimization problems over the Stiefel manifold, in which the objective function is weakly convex in the ambient Euclidean space. Such problems are ubiquitous in engineering applications but still largely unexplored. We present a family of Riemannian subgradient-type methods—namely Riemannian subgradient, incremental subgradient, and stochastic subgradient methods—to solve these problems and show that they all have an iteration complexity of $\mathcal{O}(\varepsilon^{-4})$ for driving a natural stationarity measure below ε. In addition, we establish the local linear convergence of the Riemannian subgradient and incremental subgradient methods when the problem at hand further satisfies a sharpness property and the algorithms are properly initialized and use geometrically diminishing stepsizes. To the best of our knowledge, these are the first convergence guarantees for using Riemannian subgradient-type methods to optimize a class of nonconvex nonsmooth functions over the Stiefel manifold. The fundamental ingredient in the proof of the aforementioned convergence results is a new *Riemannian subgradient inequality* for restrictions of weakly convex functions on the Stiefel manifold, which could be of independent interest. We also show that our convergence results can be extended to handle a class of compact embedded submanifolds of the Euclidean space. Finally, we discuss the sharpness properties of various formulations of the robust subspace recovery and orthogonal dictionary learning problems and demonstrate the convergence performance of the algorithms on both problems via numerical simulations.

Key words. manifold optimization, nonconvex optimization, orthogonality constraint, iteration complexity, linear convergence, robust subspace recovery, dictionary learning

AMS subject classifications. 68Q25, 65K10, 90C90, 90C26, 90C06.

1. Introduction.

In this paper, we consider the problem of optimizing a function with finite-sum structure over the Stiefel manifold—i.e.,

$$\begin{aligned} \underset{\boldsymbol{X}\in\mathbb{R}^{n\times r}}{\text{minimize}} \;\; & f(\boldsymbol{X}) := \frac{1}{m}\sum_{i=1}^{m} f_i(\boldsymbol{X}) \\ \text{subject to} \;\; & \boldsymbol{X} \in \mathrm{St}(n,r) \end{aligned} \tag{1.1}$$

with $\mathrm{St}(n,r) := \{\boldsymbol{X} \in \mathbb{R}^{n\times r} : \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_r\}$ and $\boldsymbol{I}_r$ being the $r \times r$ identity matrix—where each component $f_i : \mathbb{R}^{n\times r} \to \mathbb{R}$ $(i = 1, \ldots, m)$ is assumed to be weakly convex

*Submitted to the editors March 24, 2021. The first and second authors contributed equally to this paper. Most of the work of the first author was done when he was affiliated with the Department of Electronic Engineering, The Chinese University of Hong Kong.

Funding: X. Li was partially supported by the University Development Fund UDF01001808 of CUHK (SZ). Q. Qu was partially supported by the Moore-Sloan fellowship. Z. Zhu was partially supported by NSF Grant 1704458 and NSF Grant CCF-2008460. A. M.-C. So was partially supported by the Hong Kong Research Grants Council (RGC) General Research Fund (GRF) Project CUHK 14208117 and the CUHK Research Sustainability of Major RGC Funding Schemes Project 3133236.

†Corresponding author. School of Data Science, The Chinese University of Hong Kong, Shenzhen. (lixiao@cuhk.edu.cn, https://sites.google.com/view/xli).

‡Department of Industrial and Systems Engineering, Texas A&M University. (sxchen@tamu.edu).

§Cainiao Network, Hangzhou, China. (dengzengde@gmail.com).

¶Department of Electrical Engineering and Computer Science, University of Michigan. (qingqu@umich.edu, https://qingqu.engin.umich.edu/).

‖Department of Electrical and Computer Engineering, University of Denver. (zhihui.zhu@du.edu, http://mysite.du.edu/~zzhu61/).

#Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong. (manchoso@se.cuhk.edu.hk, http://www.se.cuhk.edu.hk/~manchoso).

in the ambient Euclidean space $\mathbb{R}^{n\times r}$. Recall that a function h is said to be *weakly convex* if $h(\cdot)+\frac{\tau}{2}\|\cdot\|_2^2$ is convex for some constant $\tau\geq 0$ [55]. In particular, the objective function in (1.1) can be nonconvex and nonsmooth. Our interest in (1.1) stems from the fact that it arises in many applications from different engineering fields such as representation learning and imaging science. As an illustration, let us present two motivating applications, in which nonsmooth formulations have clear advantages over smooth ones.

1.1. Motivating applications.

Application 1: Robust subspace recovery. Fitting a linear subspace to a dataset corrupted by outliers is a fundamental problem in machine learning and statistics, primarily known as robust principal component analysis (RPCA) [56] or robust subspace recovery (RSR) [33]. In this problem, one is given measurements $\widetilde{\boldsymbol{Y}}$ of the form $\widetilde{\boldsymbol{Y}}=\begin{bmatrix}\boldsymbol{Y} & \boldsymbol{O}\end{bmatrix}\boldsymbol{\Gamma}\in\mathbb{R}^{n\times m}$, where the columns of $\boldsymbol{Y}\in\mathbb{R}^{n\times m_1}$ form inlier points spanning a d-dimensional subspace $\mathcal{S}$; the columns of $\boldsymbol{O}\in\mathbb{R}^{n\times m_2}$ form outlier points with no linear structure; $\boldsymbol{\Gamma}\in\mathbb{R}^{m\times m}$ is an unknown permutation, and the goal is to recover the subspace $\mathcal{S}$. It is well-known that the presence of outliers can severely affect the quality of the solutions obtained by the classic PCA approach, which involves minimizing a smooth least-squares loss [56]. In order to obtain solutions that are more robust against outliers, the recent works [33,34,40] propose to minimize the nonsmooth least absolute deviation (LAD) loss. This leads to the formulation

$$\begin{aligned}&\underset{\boldsymbol{X}\in\mathbb{R}^{n\times d}}{\text{minimize}}\ \ f(\boldsymbol{X}):=\frac{1}{m}\sum_{i=1}^{m}\big\|(\boldsymbol{I}_n-\boldsymbol{X}\boldsymbol{X}^\top)\widetilde{\boldsymbol{y}}_i\big\|_2\\&\text{subject to}\ \ \boldsymbol{X}\in\mathrm{St}(n,d),\end{aligned}\tag{1.2}$$

where $\widetilde{\boldsymbol{y}}_i\in\mathbb{R}^n$ $(i=1,\ldots,m)$ denotes the i-th column of $\widetilde{\boldsymbol{Y}}$ and the columns of a global minimizer of (1.2) are expected to form an orthonormal basis of the subspace $\mathcal{S}$. The weak convexity of the components of the objective function in (1.2) can be verified by following the arguments in the proof of [37, Proposition 6]. Thus, the formulation (1.2) is an instance of problem (1.1). On another front, the works [54,69,70] consider a dual form of the problem, which leads to the so-called dual principal component pursuit (DPCP) formulation:

$$\begin{aligned}&\underset{\boldsymbol{X}\in\mathbb{R}^{n\times r}}{\text{minimize}}\ \ f(\boldsymbol{X}):=\frac{1}{m}\sum_{i=1}^{m}\big\|\widetilde{\boldsymbol{y}}_i^\top\boldsymbol{X}\big\|_2\\&\text{subject to}\ \ \boldsymbol{X}\in\mathrm{St}(n,r).\end{aligned}\tag{1.3}$$

In contrast to the primal formulation (1.2), the dual formulation (1.3) aims to find an orthogonal basis of $\mathcal{S}^\perp$ (the orthogonal complement to $\mathcal{S}$) with dimension $r=n-d$. It is clear that the components of the objective function in (1.3) are convex, thus showing that the formulation (1.3) is also an instance of problem (1.1).

Application 2: Learning sparsely-used dictionaries. A problem that arises in many machine learning and computer vision applications is dictionary learning (DL), whose goal is to find a suitable compact representation of certain input data $\boldsymbol{Y}=\begin{bmatrix}\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m\end{bmatrix}\in\mathbb{R}^{n\times m}$ [39,47,61]. Informally, this entails factorizing the data $\boldsymbol{Y}$ into a dictionary $\boldsymbol{A}$ and a sparse code matrix $\boldsymbol{S}=\begin{bmatrix}\boldsymbol{s}_1,\ldots,\boldsymbol{s}_m\end{bmatrix}$; i.e., $\boldsymbol{Y}\approx\boldsymbol{A}\boldsymbol{S}$. When the dictionary $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ is orthogonal and the code matrix $\boldsymbol{S}\in\mathbb{R}^{n\times m}$ is

sufficiently sparse, the product $\boldsymbol{A}^\top \boldsymbol{Y} \approx \boldsymbol{S}$ should be sparse. Thus, one may approach the problem by finding the sparsest vectors in the row space of $\boldsymbol{Y}$ [45, 50, 52]. This motivates the following formulation [3]:

$$\begin{aligned} &\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \;\; f(\boldsymbol{x}) := \frac{1}{m}\left\|\boldsymbol{Y}^\top \boldsymbol{x}\right\|_1 = \frac{1}{m}\sum_{i=1}^m \left|\boldsymbol{y}_i^\top \boldsymbol{x}\right| \\ &\text{subject to} \;\; \boldsymbol{x} \in \mathrm{St}(n,1). \end{aligned} \tag{1.4}$$

Note that the solution to (1.4) only returns one column of $\boldsymbol{A}$. Thus, some extra refinement technique, such as deflation [53] or repetitive independent trials [3], is needed to fully solve the DL problem. It has been shown in [3] that under a suitable statistical model, the formulation (1.4) requires fewer samples for exact recovery of the dictionary $\boldsymbol{A}$ than the smooth variant considered in [52, 53]. Still, since the approach based on (1.4) recovers the columns of $\boldsymbol{A}$ one at a time, it can be rather sensitive to noise. To circumvent this difficulty, one possibility is to directly recover the orthogonal dictionary $\boldsymbol{A}$ by

$$\begin{aligned} &\underset{\boldsymbol{X}\in\mathbb{R}^{n\times n}}{\text{minimize}} \;\; f(\boldsymbol{X}) := \frac{1}{m}\left\|\boldsymbol{Y}^\top \boldsymbol{X}\right\|_1 = \frac{1}{m}\sum_{i=1}^m \left\|\boldsymbol{y}_i^\top \boldsymbol{X}\right\|_1 \\ &\text{subject to} \;\; \boldsymbol{X} \in \mathrm{St}(n,n); \end{aligned} \tag{1.5}$$

cf. [59, 65]. This approach can be easily extended to handle any complete (i.e., square and invertible) dictionaries via preconditioning [52, 65]. Clearly, both (1.4) and (1.5) are instances of (1.1).

1.2. Main contributions. We study three *Riemannian subgradient-type methods* for solving problem (1.1), namely Riemannian subgradient method, Riemannian incremental subgradient method, and Riemannian stochastic subgradient method (see Subsection 2.2). To analyze the convergence behavior of these methods, we first extend the surrogate stationarity measure developed in [12, 16] for weakly convex minimization in the Euclidean space to one for weakly convex minimization over the Stiefel manifold (see Subsection 4.1). Then, we show that the iterates generated by the aforementioned Riemannian subgradient-type methods will drive the surrogate stationarity measure to zero at a rate of $\mathcal{O}(k^{-\frac{1}{4}})$, where k is the iteration index (see Subsections 4.2 and 4.3). Such a complexity guarantee matches that established in [12] for a host of algorithms that solve weakly convex minimization problems in the *Euclidean space*. Next, we show that if problem (1.1) further satisfies the *sharpness* property (see Definition 1), then the Riemannian subgradient and incremental subgradient methods with properly designed geometrically diminishing stepsizes and a good initialization will converge to the set of local minima associated with the sharpness property at a *linear* rate (see Section 5). To the best of our knowledge, our work is the *first* to establish the iteration complexities and convergence rates of Riemannian subgradient-type methods for optimizing a class of nonconvex nonsmooth functions over the Stiefel manifold. We also extend the above convergence results to the setting where the constraint is a compact embedded submanifold of the Euclidean space (see Section 6). Lastly, we show that under certain conditions on the inlier and outlier distributions, the LAD (1.2) and DPCP (1.3) formulations of the RSR problem satisfy the sharpness property (see Subsection 7.1). Consequently, we are able to obtain recovery guarantees for the so-called Haystack model of the input data that are competitive with state-of-the-art results.

The key to establishing the aforementioned convergence results is an algorithm-independent property that we discovered for restrictions of weakly convex functions on the Stiefel manifold, which we term the *Riemannian subgradient inequality* (see Section 3). This is one of the main contributions of this work and could be of independent interest for other Riemannian optimization problems. We believe that our results will have broad implications on understanding the convergence behavior of algorithms for solving more general manifold optimization problems with nonsmooth objectives.

1.3. Connections with prior arts.

Nonsmooth optimization in Euclidean space. The problem of minimizing a weakly convex function over a convex constraint set is well studied in the literature. The main algorithms for this task include subgradient-type methods [12, 13, 36] and proximal point-type methods [15]. The convergence analyses of these algorithms rely on a certain *weakly convex inequality*. We extend this line of work by considering a nonconvex constraint set—i.e., the Stiefel manifold—and develop an analog of the weakly convex inequality on the Stiefel manifold called the *Riemannian subgradient inequality*. Such an inequality allows us to resort to the analysis techniques for weakly convex minimization in the Euclidean space and prove new convergence results for our Riemannian subgradient-type methods when solving the problem of weakly convex minimization over the Stiefel manifold (1.1).

Smooth optimization over Riemannian manifold. Riemannian smooth optimization has been extensively studied over the years; see, e.g., [2, 7, 24, 26, 38] and the references therein. Recently, global sublinear convergence results for Riemannian gradient descent and Riemannian trust region have been presented in [7]. The analysis relies on the assumption that the pullback of the objective function f to the tangent spaces of the manifold has a Lipschitz continuous gradient, which allows one to follow the analyses of the corresponding methods for unconstrained smooth optimization. However, such an approach breaks down when f is nonsmooth, as the gradient of the pullback of f may not exist.

Nonsmooth optimization over Riemannian manifold. In contrast to Riemannian smooth optimization, Riemannian nonsmooth optimization is relatively less explored [1]. In the following, we briefly review some state-of-the-art results in this area and explain their limitations and connections to our results.

Riemannian nonsmooth optimization with geodesic convexity. Recently, the works [4, 18, 19, 66] study the convergence behavior of Riemannian subgradient-type methods when the objective function is geodesically convex over a Riemannian manifold. Thanks to the availability of a geodesic version of the convex subgradient inequality, the conventional analysis for convex optimization in the Euclidean space can be carried over to geodesically convex optimization over a Riemannian manifold. In particular, an asymptotic convergence result is first established in [19], while a global convergence rate of $\mathcal{O}(k^{-\frac{1}{2}})$ is established in [4, 18], for the Riemannian subgradient method. The work [66] considers the setting where the objective function is geodesically strongly convex over the Riemannian manifold and shows that the rate can be improved to $\mathcal{O}(k^{-1})$ for Riemannian projected subgradient methods. Unfortunately, these results are not useful for understanding problem (1.1). This is because the constraint in (1.1) is a *compact* manifold, and every continuous function that is geodesically convex on a compact Riemannian manifold can only be a constant; see, e.g., [5, Proposition 2.2] and [64] .

Riemannian gradient sampling algorithms. For general Riemannian nonsmooth optimization, the recent works [22, 23] propose Riemannian gradient sampling algorithms, which are motivated by the gradient sampling algorithms for nonconvex nonsmooth optimization in the Euclidean space [9]. As introduced in [22, 23], given the current iterate $\boldsymbol{X}_k$, a typical Riemannian gradient sampling algorithm first samples some points $\{\boldsymbol{X}_k^j\}_{j=1}^J$ in the neighborhood of $\boldsymbol{X}_k$ at which the objective function f is differentiable, where the number of sampled points J usually needs to be larger than the dimension of the manifold $\mathcal{M}$. Then, to obtain a descent direction, it solves the quadratic program

$$\boldsymbol{\xi}_k = -\operatorname*{argmin}_{\boldsymbol{G}\in\operatorname{conv}(\mathcal{W})} \|\boldsymbol{G}\|^2, \tag{1.6}$$

where $\operatorname{conv}(\mathcal{W})$ denotes the convex hull of $\mathcal{W} := \{\operatorname{grad} f(\boldsymbol{X}_k^1), \ldots, \operatorname{grad} f(\boldsymbol{X}_k^J)\}$ and $\operatorname{grad} f$ is the Riemannian gradient of f on $\mathcal{M}$. The update can then be performed via classical retractions on $\mathcal{M}$ using the descent direction $\boldsymbol{\xi}_k$. This type of algorithms can potentially be utilized to solve a large class of Riemannian nonsmooth optimization problems. However, they are only known to converge asymptotically without any rate guarantee [22, 23]. Moreover, in order to tackle problem (1.1) with large n and r using a Riemannian gradient sampling algorithm, one has to sample a large number of Riemannian gradients in each iteration, which makes the subproblem (1.6) very expensive to solve. By contrast, although we assume that the objective function in (1.1) has weakly convex components, we can establish the convergence of various Riemannian subgradient-type methods with explicit rate guarantees. In addition, each iteration of those methods involves only the computation of a Riemannian subgradient, which can potentially be much cheaper.

Two types of proximal point methods. Another classic approach to tackling Riemannian nonsmooth optimization is to apply proximal point-type methods. The idea is to iteratively compute the proximal mapping of the objective function over the Riemannian manifold [14, 20]. These methods are shown to converge globally at a sublinear rate, based on the so-called sufficient decrease property. However, the main issue with this type of methods is that each subproblem is as difficult as the original problem, which renders them not practical. When specialized to the Stiefel manifold, such a difficulty has been alleviated by some recent advances [10, 11, 25]. Specifically, they propose to compute the proximal mapping over the tangent space instead of over the Stiefel manifold, which results in a linearly constrained convex subproblem that is much easier to solve than the original problem. They also prove that the new algorithms converge globally at a sublinear rate. Nonetheless, the subproblem still needs to be solved by an iterative algorithm. By contrast, the methods considered in this paper do not need to solve expensive subproblems except for the computation of one Riemannian subgradient. As such, our overall computational complexities are much lower.

Splitting-type methods. There are also splitting-type methods for solving Riemannian nonsmooth optimization problems, such as the manifold ADMM-type algorithms in [30, 31]. In this approach, the problem at hand is typically split into two subproblems—one involves optimizing a smooth function over the Riemannian manifold, the other involves optimizing a nonsmooth function without any constraint. These subproblems are then solved in an alternating manner. Despite their simplicity, these methods often do not have any convergence guarantee.

Nonsmooth optimization over Stiefel manifold for specific problems. Finally, we close this subsection by mentioning several problem-specific results. The recent works [3] and [69, 70] propose to use the Riemannian subgradient method to solve the orthogonal DL problem (1.4) and RSR problem (1.3), respectively, and establish its local linear convergence when solving these problems. The proofs are based on a certain regularity condition instead of the sharpness property studied in this work. We will give a detailed comparison between the said regularity condition and the sharpness property in Section 5. For now, it is worth noting that the analyses in [3, 69, 70] critically depend on the specific model structure of the problem at hand and cannot be easily generalized. By contrast, we develop a more general framework for analyzing Riemannian subgradient-type methods when applied to a family of nonsmooth nonconvex optimization problems over certain compact Riemannian submanifolds, which can yield both global and local convergence guarantees.

1.4. Notation. We use $\mathrm{T}_{\boldsymbol{X}}\,\mathrm{St} := \{\boldsymbol{\xi} \in \mathbb{R}^{n\times r} : \boldsymbol{\xi}^\top \boldsymbol{X} + \boldsymbol{X}^\top \boldsymbol{\xi} = 0\}$ to denote the tangent space to the Stiefel manifold $\mathrm{St}(n,r)$ at the point $\boldsymbol{X} \in \mathrm{St}(n,r)$. Let $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \mathrm{trace}(\boldsymbol{A}^\top \boldsymbol{B})$ denote the Euclidean inner product of two matrices $\boldsymbol{A}, \boldsymbol{B}$ of the same dimensions and $\|\boldsymbol{A}\|_F = \sqrt{\langle \boldsymbol{A}, \boldsymbol{A}\rangle}$ denote the Frobenius norm of $\boldsymbol{A}$. We endow the Stiefel manifold $\mathrm{St}(n,r)$ with the Riemannian metric inherited from the Euclidean inner product; i.e., $\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \mathrm{trace}(\boldsymbol{X}^\top \boldsymbol{Y})$ for any $\boldsymbol{X}, \boldsymbol{Y} \in \mathrm{T}_{\boldsymbol{Z}}\,\mathrm{St}$ and $\boldsymbol{Z} \in \mathrm{St}(n,r)$. For a closed set $\mathcal{C} \subseteq \mathbb{R}^{n\times r}$, we use $\mathcal{P}_{\mathcal{C}}$ to denote the orthogonal projector onto $\mathcal{C}$ and $\mathrm{dist}(\boldsymbol{X}, \mathcal{C}) := \inf_{\boldsymbol{Y}\in\mathcal{C}} \|\boldsymbol{X} - \boldsymbol{Y}\|_F$ to denote the distance between $\boldsymbol{X}$ and $\mathcal{C}$. We use $x \lesssim y$ and $x \gtrsim y$ to denote $x \le cy$ and $x \ge cy$ for some universal constant c, respectively.

2. Preliminaries. In this section, we first review some basic notions in Riemannian optimization and then present the Riemannian subgradient-type algorithms for solving problem (1.1).

2.1. Optimization over Stiefel manifold.

Riemannian subgradient and first-order optimality condition. By our assumption, the objective function f in (1.1) is τ-weakly convex for some $\tau \ge 0$; i.e, there exists a convex function $g : \mathbb{R}^{n\times r} \to \mathbb{R}$ such that $f(\boldsymbol{X}) = g(\boldsymbol{X}) - \frac{\tau}{2}\|\boldsymbol{X}\|_F^2$ for any $\boldsymbol{X} \in \mathbb{R}^{n\times r}$ [55, Proposition 4.3]. Although f may not be convex, we may define its (Euclidean) subdifferential ∂f via

$$\partial f(\boldsymbol{X}) = \partial g(\boldsymbol{X}) - \tau \boldsymbol{X}, \quad \forall \boldsymbol{X} \in \mathbb{R}^{n\times r}; \tag{2.1}$$

see [55, Proposition 4.6]. Note that since g is convex, ∂g is simply its usual convex subdifferential. Hence, the subdifferential ∂f in (2.1) is well defined.

Using the properties of weakly convex functions in [55, Proposition 4.5] and the result in [63, Theorem 5.1], the Riemannian subdifferential $\partial_{\mathcal{R}} f$ of f on the Stiefel manifold $\mathrm{St}(n,r)$ is given by

$$\partial_{\mathcal{R}} f(\boldsymbol{X}) = \mathcal{P}_{\mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}}(\partial f(\boldsymbol{X})), \quad \forall \boldsymbol{X} \in \mathrm{St}(n,r). \tag{2.2}$$

In particular, given an Euclidean subgradient $\widetilde{\nabla} f(\boldsymbol{X}) \in \partial f(\boldsymbol{X})$ of f at $\boldsymbol{X} \in \mathrm{St}(n,r)$, we obtain a corresponding Riemannian subgradient $\widetilde{\nabla}_{\mathcal{R}} f(\boldsymbol{X}) \in \partial_{\mathcal{R}} f(\boldsymbol{X})$ through $\widetilde{\nabla}_{\mathcal{R}} f(\boldsymbol{X}) = \mathcal{P}_{\mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}}(\widetilde{\nabla} f(\boldsymbol{X}))$. Recall that for any $\boldsymbol{B} \in \mathbb{R}^{n\times r}$, the projection of $\boldsymbol{B}$ onto $\mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}$ is given by $\mathcal{P}_{\mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}}(\boldsymbol{B}) = \boldsymbol{B} - \frac{1}{2}\boldsymbol{X}\left(\boldsymbol{B}^\top \boldsymbol{X} + \boldsymbol{X}^\top \boldsymbol{B}\right)$ [2, Example 3.6.2].

Using (2.2), we call $\boldsymbol{X} \in \mathrm{St}(n,r)$ a *stationary point* of problem (1.1) if it satisfies the following first-order optimality condition:

$$\boldsymbol{0} \in \partial_{\mathcal{R}} f(\boldsymbol{X}). \tag{2.3}$$

Retractions on Stiefel manifold. To enable search along curves on the Stiefel manifold, we need the notion of a retraction (see [2, Definition 4.1.1] for the definition). There are four commonly used retractions on the Stiefel manifold. These include the exponential map [17] and those based on the QR decomposition, Cayley transformation [60], and polar decomposition. It is mentioned in [11] that among the above four retractions, the polar decomposition-based one is the most efficient in terms of computational complexity. Therefore, we shall focus on polar decomposition-based retraction, which is given by

$$\mathrm{Retr}_{\boldsymbol{X}}(\boldsymbol{\xi}) = (\boldsymbol{X}+\boldsymbol{\xi})(\boldsymbol{I}_r + \boldsymbol{\xi}^\top\boldsymbol{\xi})^{-\frac{1}{2}}. \tag{2.4}$$

However, we remark that our results also apply to the other three retractions; see Section 6 for a detailed discussion.

As the following lemma shows, given any $\boldsymbol{X} \in \mathrm{St}(n,r)$ and $\boldsymbol{\xi} \in \mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}$, the polar decomposition-based retraction at $\boldsymbol{X}$ essentially computes the projection of $\boldsymbol{X}+\boldsymbol{\xi}$ onto $\mathrm{St}(n,r)$. Moreover, this projection has a Lipschitz-like behavior, even though $\mathrm{St}(n,r)$ is nonconvex.

LEMMA 1. *Let $\boldsymbol{X} \in \mathrm{St}(n,r)$ and $\boldsymbol{\xi} \in \mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}$ be given. Consider the point $\boldsymbol{X}^+ = \boldsymbol{X}+\boldsymbol{\xi}$. Then, the polar decomposition-based retraction (2.4) satisfies $\mathrm{Retr}_{\boldsymbol{X}}(\boldsymbol{\xi}) = \boldsymbol{X}^+\left(\boldsymbol{X}^{+\top}\boldsymbol{X}^+\right)^{-\frac{1}{2}} = \mathcal{P}_{\mathrm{St}}(\boldsymbol{X}^+)$ and*

$$\left\|\mathrm{Retr}_{\boldsymbol{X}}(\boldsymbol{\xi}) - \overline{\boldsymbol{X}}\right\|_F \le \left\|\boldsymbol{X}^+ - \overline{\boldsymbol{X}}\right\|_F = \left\|\boldsymbol{X}+\boldsymbol{\xi}-\overline{\boldsymbol{X}}\right\|_F, \quad \forall\ \overline{\boldsymbol{X}} \in \mathrm{St}(n,r).$$

Proof. It is well known that the convex hull of the Stiefel manifold $\mathrm{St}(n,r)$ is given by $H \equiv H(n,r) := \{\boldsymbol{Y} \in \mathbb{R}^{n\times r} : \|\boldsymbol{Y}\|_2 \le 1\}$, where $\|\boldsymbol{Y}\|_2$ denotes the spectral norm (i.e. the largest singular value) of $\boldsymbol{Y}$; see, e.g., [27]. Let us first show that $\mathrm{Retr}_{\boldsymbol{X}}(\boldsymbol{\xi}) = \mathcal{P}_{\mathrm{St}}(\boldsymbol{X}^+) = \mathcal{P}_H(\boldsymbol{X}^+)$. Let $\boldsymbol{X}^+ = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be an SVD of $\boldsymbol{X}^+$. Since $\boldsymbol{\xi} \in \mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}$, we have $\boldsymbol{X}^{+\top}\boldsymbol{X}^+ = \boldsymbol{I}_r + \boldsymbol{\xi}^\top\boldsymbol{\xi}$, which implies that all the singular values of $\boldsymbol{X}^+$ are at least 1. This, together with the Hoffman-Wielandt Theorem for singular values (see, e.g., [51]), implies that $\mathcal{P}_{\mathrm{St}}(\boldsymbol{X}^+) = \mathcal{P}_H(\boldsymbol{X}^+) = \boldsymbol{U}\boldsymbol{V}^\top$, as desired.

Now, observe that $\mathrm{Retr}_{\boldsymbol{X}}(\boldsymbol{\xi}) = \boldsymbol{X}^+\left(\boldsymbol{X}^{+\top}\boldsymbol{X}^+\right)^{-\frac{1}{2}} = \boldsymbol{U}\boldsymbol{V}^\top$ and $\overline{\boldsymbol{X}} \in H(n,r)$. Hence, we have $\left\|\mathrm{Retr}_{\boldsymbol{X}}(\boldsymbol{\xi}) - \overline{\boldsymbol{X}}\right\|_F = \left\|\mathcal{P}_H(\boldsymbol{X}^+) - \mathcal{P}_H\left(\overline{\boldsymbol{X}}\right)\right\|_F$. Upon noting that projections onto closed convex sets are 1-Lipschitz, the proof is complete. ∎

2.2. A family of Riemannian subgradient-type methods.

Riemannian subgradient method. We begin by revisiting the Riemannian gradient method for smooth optimization over the Stiefel manifold. Let $h : \mathbb{R}^{n\times r} \to \mathbb{R}$ be a smooth function and consider

$$\begin{aligned} &\underset{\boldsymbol{X}\in\mathbb{R}^{n\times r}}{\text{minimize}}\ \ h(\boldsymbol{X}) \\ &\text{subject to}\ \ \boldsymbol{X} \in \mathrm{St}(n,r). \end{aligned}$$

A generic Riemannian gradient method for solving the above problem is given by

$$\boldsymbol{X}_{k+1} = \mathrm{Retr}_{\boldsymbol{X}_k}(\boldsymbol{\xi}_k) \quad \text{with} \quad \boldsymbol{\xi}_k = -\gamma_k \,\mathrm{grad}\, h(\boldsymbol{X}_k),$$

where $\mathrm{grad}\, h(\boldsymbol{X}_k)$ is the Riemannian gradient of h at $\boldsymbol{X}_k$, $\gamma_k > 0$ is the stepsize, and Retr is any retraction on the Stiefel manifold; see, e.g., [2, Section 4.2]. Since problem (1.1) involves a possibly nonsmooth objective function, one approach to

tackling it is to apply a natural generalization of the Riemannian gradient method, namely the Riemannian subgradient method:

$$\boxed{\boldsymbol{X}_{k+1} = \mathrm{Retr}_{\boldsymbol{X}_k}(\boldsymbol{\xi}_k) \quad \text{with} \quad \boldsymbol{\xi}_k = -\gamma_k \widetilde{\nabla}_{\mathcal{R}} f(\boldsymbol{X}_k).} \tag{2.5}$$

Here, recall that $\widetilde{\nabla}_{\mathcal{R}} f(\boldsymbol{X}_k) \in \partial_{\mathcal{R}} f(\boldsymbol{X}_k)$ is a Riemannian subgradient of f at $\boldsymbol{X}_k \in \mathrm{St}(n,r)$, which can be obtained by taking $\widetilde{\nabla} f(\boldsymbol{X}) \in \partial f(\boldsymbol{X})$ and setting $\widetilde{\nabla}_{\mathcal{R}} f(\boldsymbol{X}) = \mathcal{P}_{\mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}}(\widetilde{\nabla} f(\boldsymbol{X}))$; see Subsection 2.1.

Riemannian incremental and stochastic subgradient methods. Recall that the objective function in (1.1) has the *finite-sum structure* $f = \frac{1}{m}\sum_{i=1}^m f_i$. In many modern machine learning tasks, the number of components m can be very large. Thus, it is not desirable to evaluate the full Riemannian subgradient of f. This motivates us to introduce two variants of the Riemannian subgradient method (2.5), namely the Riemannian incremental subgradient method and Riemannian stochastic subgradient method, to better exploit the finite-sum structure in (1.1). Let us now give a high-level description of these two methods.

Starting with the current iterate $\boldsymbol{X}_k$, both methods generate a sequence of m inner iterates $\boldsymbol{X}_{k,1}, \ldots, \boldsymbol{X}_{k,i}, \ldots, \boldsymbol{X}_{k,m}$ via

$$\boxed{\boldsymbol{X}_{k,i} = \mathrm{Retr}_{\boldsymbol{X}_{k,i-1}}(\boldsymbol{\xi}_{k,i-1}) \quad \text{with} \quad \boldsymbol{\xi}_{k,i-1} = -\gamma_k \widetilde{\nabla}_{\mathcal{R}} f_{\ell_i}(\boldsymbol{X}_{k,i-1})} \tag{2.6}$$

with $\boldsymbol{X}_{k,0} = \boldsymbol{X}_k$, where f_{ℓ_i} is selected from $\{f_1, \ldots, f_m\}$ according to a certain rule. The next iterate $\boldsymbol{X}_{k+1}$ is then obtained by setting $\boldsymbol{X}_{k+1} = \boldsymbol{X}_{k,m}$. The difference between the incremental and stochastic methods lies in the rule for selecting the component function f_{ℓ_i}. In particular,

- *Riemannian incremental subgradient method* picks the component function f_{ℓ_i} *sequentially* from f_1 to f_m—i.e., $\boldsymbol{\xi}_{k,i-1} = -\gamma_k \widetilde{\nabla}_{\mathcal{R}} f_i(\boldsymbol{X}_{k,i-1})$;
- *Riemannian stochastic subgradient method* picks the component function f_{ℓ_i} *independently* and *uniformly* from $\{f_1, \ldots, f_m\}$ in each inner iteration (2.6)—i.e., $\boldsymbol{\xi}_{k,i-1} = -\gamma_k \widetilde{\nabla}_{\mathcal{R}} f_{\ell_i}(\boldsymbol{X}_{k,i-1})$ with $\ell_i \sim_{i.i.d.} \mathrm{Uniform}\big(\{1, \ldots, m\}\big)$.

3. Riemannian Subgradient Inequality over Stiefel Manifold. Naturally, we are interested in the convergence behavior of the Riemannian subgradient-type methods introduced in Subsection 2.2 when applied to problem (1.1). Towards that end, let us derive a useful inequality, which we call the *Riemannian subgradient inequality*, for restrictions of weakly convex functions on the Stiefel manifold. The main motivation for deriving such an inequality is that an analogous one for weakly convex functions in the Euclidean space, known as the *weakly convex inequality*, plays a fundamental role in the convergence analysis of subgradient-type methods for solving weakly convex minimization problems [12, 13, 36, 37]. To begin, recall that for a τ-weakly convex function $h : \mathbb{R}^{n\times r} \to \mathbb{R}$, the weakly convex inequality states that

$$\begin{aligned} h(\boldsymbol{Y}) \geq h(\boldsymbol{X}) + \Big\langle \widetilde{\nabla} h(\boldsymbol{X}), \boldsymbol{Y} - \boldsymbol{X} \Big\rangle - \frac{\tau}{2}\|\boldsymbol{Y} - \boldsymbol{X}\|_F^2, \\ \forall\ \widetilde{\nabla} h(\boldsymbol{X}) \in \partial h(\boldsymbol{X});\ \boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times r} \end{aligned} \tag{3.1}$$

[55, Proposition 4.8]. The following is our extension of the above inequality to one for weakly convex functions that are restricted on the Stiefel manifold.

THEOREM 1 (Riemannian Subgradient Inequality). *Suppose that $h : \mathbb{R}^{n\times r} \to \mathbb{R}$ is τ-weakly convex for some $\tau \geq 0$. Then, for any bounded open convex set $\mathcal{U}$ that*

contains $\mathrm{St}(n,r)$*, there exists a constant* $L>0$ *such that* h *is* L*-Lipschitz continuous on* $\mathcal{U}$ *and satisfies*

$$h(\boldsymbol{Y}) \geq h(\boldsymbol{X}) + \left\langle \widetilde{\nabla}_{\mathcal{R}} h(\boldsymbol{X}), \boldsymbol{Y}-\boldsymbol{X} \right\rangle - \frac{\tau+L}{2}\|\boldsymbol{Y}-\boldsymbol{X}\|_F^2, \quad \forall\ \widetilde{\nabla}_{\mathcal{R}} h(\boldsymbol{X}) \in \partial_{\mathcal{R}} h(\boldsymbol{X}); \boldsymbol{X}, \boldsymbol{Y} \in \mathrm{St}(n,r). \tag{3.2}$$

Before we proceed to prove Theorem 1, let us highlight the differences between the weakly convex inequality (3.1) and the Riemannian subgradient inequality (3.2). First, the former involves elements in the Euclidean subdifferential ∂h, while the latter involves elements in the Riemannian subdifferential $\partial_{\mathcal{R}} h$. Second, the former holds for all pairs of points in the Euclidean space $\mathbb{R}^{n\times r}$, while the latter only holds for all pairs of points on the Stiefel manifold $\mathrm{St}(n,r)$. Third, the latter involves the extra compensation term $-\frac{L}{2}\|\boldsymbol{Y}-\boldsymbol{X}\|_F^2$, which accounts for the behavior of the restriction of h on the Stiefel manifold $\mathrm{St}(n,r)$.

Proof of Theorem 1. The Lipschitz continuity of h on $\mathcal{U}$ follows directly from [55, Proposition 4.4] and the boundedness of $\mathcal{U}$. Since h is τ-weakly convex on $\mathbb{R}^{n\times r}$, for any $\boldsymbol{X}, \boldsymbol{Y} \in \mathrm{St}(n,r) \subseteq \mathbb{R}^{n\times r}$, the inequality (3.1) implies that

$$\begin{aligned} h(\boldsymbol{Y}) &\geq h(\boldsymbol{X}) + \left\langle \widetilde{\nabla} h(\boldsymbol{X}), \boldsymbol{Y}-\boldsymbol{X} \right\rangle - \frac{\tau}{2}\|\boldsymbol{Y}-\boldsymbol{X}\|_F^2 \\ &= h(\boldsymbol{X}) + \left\langle \mathcal{P}_{\mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}}(\widetilde{\nabla} h(\boldsymbol{X})) + \mathcal{P}^{\perp}_{\mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}}(\widetilde{\nabla} h(\boldsymbol{X})), \boldsymbol{Y}-\boldsymbol{X} \right\rangle - \frac{\tau}{2}\|\boldsymbol{Y}-\boldsymbol{X}\|_F^2, \end{aligned} \tag{3.3}$$

where

$$\mathcal{P}^{\perp}_{\mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}}(\boldsymbol{B}) = \frac{1}{2}\boldsymbol{X}\left(\boldsymbol{B}^\top\boldsymbol{X} + \boldsymbol{X}^\top\boldsymbol{B}\right), \quad \forall\ \boldsymbol{B} \in \mathbb{R}^{n\times r} \tag{3.4}$$

[2, Example 3.6.2]. Now, we compute

$$\begin{aligned} \left\langle \mathcal{P}^{\perp}_{\mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}}(\widetilde{\nabla} h(\boldsymbol{X})), \boldsymbol{Y}-\boldsymbol{X} \right\rangle &= \left\langle \mathcal{P}^{\perp}_{\mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}}(\widetilde{\nabla} h(\boldsymbol{X})), \mathcal{P}^{\perp}_{\mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}}(\boldsymbol{Y}-\boldsymbol{X}) + \mathcal{P}_{\mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}}(\boldsymbol{Y}-\boldsymbol{X}) \right\rangle \\ &= \left\langle \mathcal{P}^{\perp}_{\mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}}(\widetilde{\nabla} h(\boldsymbol{X})), \mathcal{P}^{\perp}_{\mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}}(\boldsymbol{Y}-\boldsymbol{X}) \right\rangle \\ &= \left\langle \widetilde{\nabla} h(\boldsymbol{X}), \mathcal{P}^{\perp}_{\mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}}(\boldsymbol{Y}-\boldsymbol{X}) \right\rangle \\ &\overset{(i)}{=} \frac{1}{2}\left\langle \widetilde{\nabla} h(\boldsymbol{X}), \boldsymbol{X}\left(\boldsymbol{Y}^\top\boldsymbol{X} + \boldsymbol{X}^\top\boldsymbol{Y} - 2\boldsymbol{I}_r\right) \right\rangle \\ &\overset{(ii)}{\geq} -\frac{1}{2}\|\widetilde{\nabla} h(\boldsymbol{X})\|_F \left\|\boldsymbol{Y}^\top\boldsymbol{X} + \boldsymbol{X}^\top\boldsymbol{Y} - 2\boldsymbol{I}_r\right\|_F \\ &\overset{(iii)}{\geq} -\frac{L}{2}\left\|\boldsymbol{Y}^\top\boldsymbol{X} + \boldsymbol{X}^\top\boldsymbol{Y} - 2\boldsymbol{I}_r\right\|_F, \end{aligned} \tag{3.5}$$

where (i) comes from (3.4), (ii) is due to the fact that $\boldsymbol{X} \in \mathrm{St}(n,r)$, and (iii) follows from [46, Theorem 9.13] and the L-Lipschitz continuity of h on $\mathcal{U}$. Note that

$$\left\|\boldsymbol{Y}^\top\boldsymbol{X} + \boldsymbol{X}^\top\boldsymbol{Y} - 2\boldsymbol{I}_r\right\|_F = \left\|(\boldsymbol{X}-\boldsymbol{Y})^\top(\boldsymbol{X}-\boldsymbol{Y})\right\|_F \leq \|\boldsymbol{X}-\boldsymbol{Y}\|_F^2 \tag{3.6}$$

since $\boldsymbol{X}, \boldsymbol{Y} \in \mathrm{St}(n,r)$. Combining (3.5) and (3.6) and recalling (3.3), we get

$$h(\boldsymbol{Y}) \geq h(\boldsymbol{X}) + \left\langle \mathcal{P}_{\mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}}(\widetilde{\nabla} h(\boldsymbol{X})), \boldsymbol{Y}-\boldsymbol{X} \right\rangle - \frac{\tau+L}{2}\|\boldsymbol{Y}-\boldsymbol{X}\|_F^2.$$

Since $\boldsymbol{X}, \boldsymbol{Y} \in \mathrm{St}(n,r)$, $\widetilde{\nabla} h(\boldsymbol{X}) \in \partial h(\boldsymbol{X})$ are arbitrary and $\partial_{\mathcal{R}} h(\boldsymbol{X}) = \mathcal{P}_{\mathrm{T}_{\boldsymbol{X}}\,\mathrm{St}}(\partial h(\boldsymbol{X}))$ (see (2.2)), the proof is complete. ∎

As we shall see in subsequent sections, the Riemannian subgradient inequality plays a similar role to the weakly convex inequality and allows us to connect the analysis of Riemannian subgradient-type methods with that of their Euclidean counterparts. In particular, equipped with Theorem 1, we can obtain the iteration complexities of the Riemannian subgradient-type methods introduced in Subsection 2.2 for addressing problem (1.1). Moreover, if problem (1.1) further possesses certain sharpness property (see Definition 1), then the aforementioned methods with geometrically diminishing stepsizes and a proper initialization will achieve local linear convergence to the set of so-called weak sharp minima (again, see Definition 1).

Although the Riemannian subgradient inequality in Theorem 1 focuses on the Stiefel manifold, it can be extended to a class of compact embedded submanifolds of the Euclidean space. We shall present such an extension in Section 6.

4. Global Convergence.

In this section, we study the iteration complexities of Riemannian subgradient-type methods for solving problem (1.1). Our analysis relies on the Riemannian subgradient inequality in Theorem 1.

4.1. Surrogate stationarity measure.

In classical Euclidean nonsmooth convex optimization, the iteration complexities of subgradient-type methods are typically presented in terms of the suboptimality gap $f(\boldsymbol{X}_k) - \min f$; see, e.g., [44, Theorem 3.2.2], [42, Proposition 2.3]. On the other hand, in Riemannian smooth optimization, which typically involves nonconvex constraints, the iteration complexities of various methods can be expressed in terms of the continuous stationarity measure $\|\mathrm{grad} f(\boldsymbol{X}_k)\|_F$ [7]. However, for the Riemannian nonsmooth optimization problem (1.1), neither the suboptimality gap $f(\boldsymbol{X}_k) - \min f$ (due to nonconvexity) nor the minimum-norm Riemannian subgradient $\mathrm{dist}\,(0, \partial_{\mathcal{R}} f(\boldsymbol{X}_k))$ (due to nonsmoothness) is an appropriate stationarity measure. Therefore, in order to establish the iteration complexities of Riemannian subgradient-type methods, we need to find a surrogate stationarity measure that can track the progress of those methods.

Towards that end, we borrow ideas from the recent works [12,16] on weakly convex minimization in the Euclidean space, which propose to use the gradient of the Moreau envelope of the weakly convex function at hand as a surrogate stationarity measure. To begin, let us define, for any $\lambda > 0$, the following analogs of the Moreau envelope and proximal mapping for problem (1.1), which take into account the effect of the Stiefel manifold constraint on the problem:

$$\begin{cases} f_\lambda(\boldsymbol{X}) = \min\limits_{\boldsymbol{Y}\in \mathrm{St}(n,r)} \left\{ f(\boldsymbol{Y}) + \dfrac{1}{2\lambda}\|\boldsymbol{Y}-\boldsymbol{X}\|_F^2 \right\}, & \boldsymbol{X} \in \mathrm{St}(n,r), \\ P_{\lambda f}(\boldsymbol{X}) \in \operatorname{argmin}\limits_{\boldsymbol{Y}\in \mathrm{St}(n,r)} \left\{ f(\boldsymbol{Y}) + \dfrac{1}{2\lambda}\|\boldsymbol{Y}-\boldsymbol{X}\|_F^2 \right\}, & \boldsymbol{X} \in \mathrm{St}(n,r). \end{cases} \tag{4.1}$$

We remark that the Moreau envelope and proximal mapping defined above differ from those in [20] in that the proximal term $\boldsymbol{Y} \mapsto \frac{1}{2\lambda}\|\boldsymbol{Y}-\boldsymbol{X}\|_F^2$ is based on the Euclidean distance rather than the geodesic distance. This will facilitate our later analysis.

By (2.2) and (2.3), the point $P_{\lambda f}(\boldsymbol{X})$ satisfies the first-order optimality condition $\boldsymbol{0} \in \partial_{\mathcal{R}} f\left(P_{\lambda f}(\boldsymbol{X})\right) + \frac{1}{\lambda}\mathcal{P}_{\mathrm{T}_{P_{\lambda f}(\boldsymbol{X})}\mathrm{St}}\left(P_{\lambda f}(\boldsymbol{X}) - \boldsymbol{X}\right)$. It follows that

$$\boxed{\begin{aligned} \mathrm{dist}\left(\boldsymbol{0}, \partial_{\mathcal{R}} f\left(P_{\lambda f}(\boldsymbol{X})\right)\right) &\le \lambda^{-1} \cdot \left\| \mathcal{P}_{\mathrm{T}_{P_{\lambda f}(\boldsymbol{X})}\mathrm{St}}\left(P_{\lambda f}(\boldsymbol{X}) - \boldsymbol{X}\right) \right\|_F \\ &\le \lambda^{-1} \cdot \left\| P_{\lambda f}(\boldsymbol{X}) - \boldsymbol{X} \right\|_F =: \Theta(\boldsymbol{X}). \end{aligned}} \tag{4.2}$$

In particular, we see from (2.3) that $\boldsymbol{X} \in \mathrm{St}(n,r)$ is a stationary point of problem (1.1) when $\Theta(\boldsymbol{X}) = 0$. This motivates us to use $\boldsymbol{X} \mapsto \Theta(\boldsymbol{X})$ as a surrogate stationarity measure of problem (1.1) and call $\boldsymbol{X} \in \mathrm{St}(n,r)$ an *ε-nearly stationary point* of problem (1.1) if it satisfies $\Theta(\boldsymbol{X}) \le \varepsilon$.

The careful reader may note that the proximal mapping $P_{\lambda f}$ in (4.1) needs not yield a unique point at a given $\boldsymbol{X} \in \mathrm{St}(n,r)$. Nevertheless, for the purpose of defining the surrogate stationarity measure, we can choose any point returned by $P_{\lambda f}$ at $\boldsymbol{X}$, as each of them plays exactly the same role in our analysis and will satisfy the convergence rate bounds in Theorem 2.

4.2. Riemannian subgradient and incremental subgradient methods. Using the surrogate stationarity measure Θ, we are now ready to establish the iteration complexities of the Riemannian subgradient and incremental subgradient methods. We will focus on analyzing the Riemannian incremental subgradient method, as the Riemannian subgradient method can be regarded as its special case where there is only one (i.e., $m = 1$) component function.

To begin, let us establish a relationship between the surrogate stationarity measure Θ and the sufficient decrease of the Moreau envelope f_λ.

PROPOSITION 1. *Suppose that each component function $f_i : \mathbb{R}^{n\times r} \to \mathbb{R}$ $(i = 1,\dots,m)$ in problem (1.1) is τ-weakly convex on $\mathbb{R}^{n\times r}$ for some $\tau \ge 0$. Let $\mathcal{U}$ be any bounded open convex set that contains $\mathrm{St}(n,r)$. Furthermore, let $\{\boldsymbol{X}_k\}$ be the sequence generated by the Riemannian incremental subgradient method (2.6) with an arbitrary initialization for solving problem (1.1). Then, for any $\lambda < \frac{1}{2(L+\tau)}$ in (4.1), we have for any $k \ge 0$*

$$m\gamma_k \Theta^2(\boldsymbol{X}_k) \le \frac{2\big(f_\lambda(\boldsymbol{X}_k) - f_\lambda(\boldsymbol{X}_{k+1})\big) + \frac{\gamma_k^2 L^2}{\lambda} m^2 + \frac{\gamma_k^3 L^2 (L+\tau)}{\lambda} C(m)}{2\lambda\left(\frac{1}{2\lambda} - (L+\tau)\right)},$$

where $L > 0$ is an upper bound on the Lipschitz constants of $f_1,\dots,f_m$ on $\mathcal{U}$ and $C(m) = \frac{1}{3}m(m-1)(2m-1)$.

Proof. According to (4.1), we have

$$\begin{aligned} f_\lambda(\boldsymbol{X}_{k+1}) &= f\left(P_{\lambda f}(\boldsymbol{X}_{k+1})\right) + \frac{1}{2\lambda}\left\|P_{\lambda f}(\boldsymbol{X}_{k+1}) - \boldsymbol{X}_{k+1}\right\|_F^2 \\ &\le f\left(P_{\lambda f}(\boldsymbol{X}_k)\right) + \frac{1}{2\lambda}\left\|P_{\lambda f}(\boldsymbol{X}_k) - \boldsymbol{X}_{k+1}\right\|_F^2, \end{aligned} \tag{4.3}$$

where the last inequality follows from the optimality of $P_{\lambda f}(\boldsymbol{X}_{k+1})$ and the fact that $P_{\lambda f}(\boldsymbol{X}_k) \in \mathrm{St}(n,r)$. We claim that for $l = 1,\dots,m$,

$$\begin{aligned} \left\|P_{\lambda f}(\boldsymbol{X}_k) - \boldsymbol{X}_{k,l}\right\|_F^2 \le \left\|\boldsymbol{X}_k - P_{\lambda f}(\boldsymbol{X}_k)\right\|_F^2 - 2\gamma_k \sum_{i=1}^{l}\left(f_i(\boldsymbol{X}_{k,i-1}) - f_i\left(P_{\lambda f}(\boldsymbol{X}_k)\right)\right) \\ + \gamma_k(L+\tau)\sum_{i=1}^{l}\left\|\boldsymbol{X}_{k,i-1} - P_{\lambda f}(\boldsymbol{X}_k)\right\|_F^2 + l\gamma_k^2 L^2. \end{aligned} \tag{4.4}$$

The proof is by induction on l. For $l = 1$, recalling that $\boldsymbol{X}_{k,0} = \boldsymbol{X}_k$, we compute

$$\begin{aligned} \left\|P_{\lambda f}(\boldsymbol{X}_k) - \boldsymbol{X}_{k,1}\right\|_F^2 &\le \left\|\boldsymbol{X}_k + \boldsymbol{\xi}_{k,0} - P_{\lambda f}(\boldsymbol{X}_k)\right\|_F^2 \\ &\le \left\|\boldsymbol{X}_k - P_{\lambda f}(\boldsymbol{X}_k)\right\|_F^2 - 2\gamma_k\left(f_1(\boldsymbol{X}_k) - f_1\left(P_{\lambda f}(\boldsymbol{X}_k)\right)\right) \\ &\quad + \gamma_k(L+\tau)\left\|\boldsymbol{X}_k - P_{\lambda f}(\boldsymbol{X}_k)\right\|_F^2 + \gamma_k^2 L^2, \end{aligned} \tag{4.5}$$

where we used (2.6) and Lemma 1 in the first inequality and Theorem 1 and the fact that $\left\|\widetilde{\nabla}_{\mathcal{R}} f_i(\boldsymbol{X}_{k,i-1})\right\|_F \le \left\|\widetilde{\nabla} f_i(\boldsymbol{X}_{k,i-1})\right\|_F \le L$ in the second inequality. The inductive step can be completed by following the same derivations as in (4.5). Thus, the claim (4.4) is established. Setting $l = m$ in (4.4) and plugging it into (4.3), we obtain

$$\begin{aligned} f_\lambda(\boldsymbol{X}_{k+1}) \le f_\lambda(\boldsymbol{X}_k) + \frac{\gamma_k}{\lambda}\sum_{i=1}^m \left(f_i\left(P_{\lambda f}(\boldsymbol{X}_k)\right) - f_i(\boldsymbol{X}_{k,i-1})\right) \\ + \frac{\gamma_k(L+\tau)}{2\lambda}\sum_{i=1}^m \|\boldsymbol{X}_{k,i-1} - P_{\lambda f}(\boldsymbol{X}_k)\|_F^2 + \frac{m\gamma_k^2 L^2}{2\lambda}, \end{aligned} \tag{4.6}$$

where we used the relation $f_\lambda(\boldsymbol{X}_k) = f\left(P_{\lambda f}(\boldsymbol{X}_k)\right) + \frac{1}{2\lambda}\|\boldsymbol{X}_k - P_{\lambda f}(\boldsymbol{X}_k)\|_F^2$ (since $\boldsymbol{X}_k \in \mathrm{St}(n,r)$).

Next, we claim that for $i = 1, \ldots, m$,

$$\|\boldsymbol{X}_{k,i-1} - \boldsymbol{X}_k\|_F \le (i-1)\gamma_k L. \tag{4.7}$$

The proof is again by induction on i. The claim trivially holds when $i = 1$. Suppose that (4.7) holds for $i = j$. For $i = j+1$, we compute $\|\boldsymbol{X}_{k,j} - \boldsymbol{X}_k\|_F \le \|\boldsymbol{X}_{k,j-1} + \boldsymbol{\xi}_{k,j-1} - \boldsymbol{X}_k\|_F \le j\gamma_k L$, where we used (2.6) and Lemma 1 in the first inequality. This completes the inductive step and the proof of the claim.

With (4.7), we have

$$\begin{aligned} f_i\left(P_{\lambda f}(\boldsymbol{X}_k)\right) - f_i(\boldsymbol{X}_{k,i-1}) &= f_i\left(P_{\lambda f}(\boldsymbol{X}_k)\right) - f_i(\boldsymbol{X}_k) + f_i(\boldsymbol{X}_k) - f_i(\boldsymbol{X}_{k,i-1}) \\ &\le (i-1)\gamma_k L^2 + f_i\left(P_{\lambda f}(\boldsymbol{X}_k)\right) - f_i(\boldsymbol{X}_k) \end{aligned} \tag{4.8}$$

and

$$\begin{aligned} \|\boldsymbol{X}_{k,i-1} - P_{\lambda f}(\boldsymbol{X}_k)\|_F^2 &= \|\boldsymbol{X}_{k,i-1} - \boldsymbol{X}_k + \boldsymbol{X}_k - P_{\lambda f}(\boldsymbol{X}_k)\|_F^2 \\ &\le 2(i-1)^2\gamma_k^2 L^2 + 2\|\boldsymbol{X}_k - P_{\lambda f}(\boldsymbol{X}_k)\|_F^2. \end{aligned} \tag{4.9}$$

Plugging (4.8) and (4.9) into (4.6) yields

$$\begin{aligned} f_\lambda(\boldsymbol{X}_{k+1}) \le f_\lambda(\boldsymbol{X}_k) + \frac{m^2\gamma_k^2 L^2 + \frac{1}{3}m(m-1)(2m-1)\gamma_k^3 L^2(L+\tau)}{2\lambda} \\ + \frac{m\gamma_k}{\lambda}\left(f\left(P_{\lambda f}(\boldsymbol{X}_k)\right) - f(\boldsymbol{X}_k)\right) + \frac{m\gamma_k(L+\tau)}{\lambda}\|\boldsymbol{X}_k - P_{\lambda f}(\boldsymbol{X}_k)\|_F^2. \end{aligned} \tag{4.10}$$

By definition of the Moreau envelope and proximal mapping in (4.1), we have

$$\begin{aligned} &-\left[f(\boldsymbol{X}_k) - f\left(P_{\lambda f}(\boldsymbol{X}_k)\right) - (L+\tau)\|\boldsymbol{X}_k - P_{\lambda f}(\boldsymbol{X}_k)\|_F^2\right] \\ &= -\left[f(\boldsymbol{X}_k) - \left(f\left(P_{\lambda f}(\boldsymbol{X}_k)\right) + \frac{1}{2\lambda}\|\boldsymbol{X}_k - P_{\lambda f}(\boldsymbol{X}_k)\|_F^2\right)\right. \\ &\qquad \left. + \left(\frac{1}{2\lambda} - (L+\tau)\right)\|\boldsymbol{X}_k - P_{\lambda f}(\boldsymbol{X}_k)\|_F^2\right] \\ &\le -\left(\frac{1}{2\lambda} - (L+\tau)\right)\|\boldsymbol{X}_k - P_{\lambda f}(\boldsymbol{X}_k)\|_F^2, \end{aligned} \tag{4.11}$$

where the last inequality is due to $f_\lambda(\boldsymbol{X}_k) = f\left(P_{\lambda f}(\boldsymbol{X}_k)\right) + \frac{1}{2\lambda}\left\|\boldsymbol{X}_k - P_{\lambda f}(\boldsymbol{X}_k)\right\|_F^2$ (since $\boldsymbol{X}_k \in \mathrm{St}(n,r)$) and $f_\lambda(\boldsymbol{X}_k) \le f(\boldsymbol{X}_k)$. Since $\lambda < \frac{1}{2(L+\tau)}$ by assumption, the desired result then follows by substituting (4.11) into (4.10) and recognizing that $\Theta(\boldsymbol{X}_k) = \lambda^{-1}\left\|P_{\lambda f}(\boldsymbol{X}_k) - \boldsymbol{X}_k\right\|_F$ (see (4.2)). ∎

Using Proposition 1, we obtain our iteration complexity result for the Riemannian subgradient and incremental subgradient methods.

THEOREM 2. *Under the setting of Proposition 1, the following hold:*

(a) If we choose the constant stepsize $\gamma_k = \frac{1}{m\sqrt{T+1}}$, $k = 0, 1, \ldots$ *with* T *being the total number of iterations, then*

$$\min_{0\le k\le T} \Theta^2(\boldsymbol{X}_k) \le \frac{2\big(f_\lambda\left(\boldsymbol{X}_0\right) - \min f_\lambda\big) + \frac{L^2}{\lambda} + \frac{L^2(L+\tau)}{\lambda m^3}C(m)}{2\lambda\left(\frac{1}{2\lambda} - (L+\tau)\right)\sqrt{T+1}}.$$

(b) If we choose the diminishing stepsizes $\gamma_k = \frac{1}{m\sqrt{k+1}}$, $k = 0, 1, \ldots$, *then*

$$\min_{0\le k\le T} \Theta^2(\boldsymbol{X}_k) \le \frac{2\big(f_\lambda\left(\boldsymbol{X}_0\right) - \min f_\lambda\big) + \left(\frac{L^2}{\lambda} + \frac{L^2(L+\tau)}{\lambda m^3}C(m)\right)\left(\ln(T+1)+1\right)}{2\lambda\left(\frac{1}{2\lambda} - (L+\tau)\right)\sqrt{T+1}}.$$

Proof. By summing both sides of the relation in Proposition 1 over $k = 0, 1, \ldots, T$, we deduce that

$$\min_{0\le k\le T} \Theta^2(\boldsymbol{X}_k) \le \frac{2\big(f_\lambda\left(\boldsymbol{X}_0\right) - \min f_\lambda\big) + \frac{L^2}{\lambda}m^2\sum_{k=0}^T \gamma_k^2 + \frac{L^2(L+\tau)}{\lambda}C(m)\sum_{k=0}^T\gamma_k^3}{2\lambda\left(\frac{1}{2\lambda} - (L+\tau)\right)m\sum_{k=0}^T\gamma_k}.$$

The result in (a) follows immediately by substituting $\gamma_k = \frac{1}{m\sqrt{T+1}}$ into the above inequality, while that in (b) follows by substituting $\gamma_k = \frac{1}{m\sqrt{k+1}}$ into the above inequality and noting that $\sum_{k=0}^T \frac{1}{\sqrt{k+1}} > \sqrt{T+1}$ and $\sum_{k=0}^T \frac{1}{k+1} < \ln(T+1)+1$. ∎

By taking $\lambda = \frac{1}{4(L+\tau)}$ and using the constant stepsize $\gamma_k = \frac{1}{m\sqrt{T+1}}$, $k = 0, 1, \ldots$, we see from Theorem 2 that

$$\min_{0\le k\le T} \Theta(\boldsymbol{X}_k) \le \frac{2\sqrt{\big(f_\lambda\left(\boldsymbol{X}_0\right) - \min f_\lambda\big) + 2L^2(L+\tau)(1+L+\tau)}}{(T+1)^{1/4}}.$$

In particular, the iteration complexity of the Riemannian (incremental) subgradient method for computing an ε-nearly stationary point of problem (1.1) is $\mathcal{O}(\varepsilon^{-4})$. It is worth noting that this matches the iteration complexity of a host of methods for solving weakly convex minimization problems in the Euclidean space [12].

4.3. Riemannian stochastic subgradient method. Now, let us turn to analyze the Riemannian stochastic subgradient method. Instead of focusing on objective functions with a finite-sum structure as in (1.1), we consider the following more general stochastic optimization problem over the Stiefel manifold:

$$\begin{aligned} &\underset{\boldsymbol{X}\in\mathbb{R}^{n\times r}}{\text{minimize}} \;\; f(\boldsymbol{X}) := \mathrm{E}_{\zeta\sim D}[g(\boldsymbol{X},\zeta)] \\ &\text{subject to} \;\; \boldsymbol{X} \in \mathrm{St}(n,r). \end{aligned} \tag{4.12}$$

Here, we assume that the function $\boldsymbol{X} \mapsto g(\boldsymbol{X},\zeta)$ is τ-weakly convex ($\tau \ge 0$) for each realization ζ and the function f is finite-valued on $\mathbb{R}^{n\times r}$. Furthermore, we assume the

existence of a bounded open convex set $\mathcal{U}$ containing $\mathrm{St}(n,r)$ such that $\boldsymbol{X} \mapsto g(\boldsymbol{X},\zeta)$ is Lipschitz continuous on $\mathcal{U}$ with some constant $L(\zeta) > 0$ and $L^2 = \mathrm{E}_{\zeta\sim D}[L(\zeta)^2] < +\infty$. This would then imply that for any $\boldsymbol{X} \in \mathrm{St}(n,r)$, we have

$$\mathrm{E}_{\zeta\sim D}\left[\left\|\widetilde{\nabla} g(\boldsymbol{X},\zeta)\right\|_F^2\right] \le L^2, \tag{4.13}$$

where $\widetilde{\nabla} g(\boldsymbol{X},\zeta) \in \partial g(\boldsymbol{X},\zeta)$. Moreover, the function f is L-Lipschitz continuous on $\mathcal{U}$. When D is the empirical distribution on m data samples, problem (4.12) reduces to our original finite-sum optimization problem (1.1). If all the component functions are finite-valued and weakly convex, then the above two assumptions hold.

Now, suppose that the Riemannian stochastic subgradient method is equipped with a *Riemannian stochastic subgradient oracle*, which has the following properties:

(a) The oracle can generate i.i.d. samples according to the distribution D.
(b) Given a point $\boldsymbol{X} \in \mathrm{St}(n,r)$, the oracle generates a sample $\zeta \sim D$ and returns a stochastic subgradient $\widetilde{\nabla} g(\boldsymbol{X},\zeta) \in \partial g(\boldsymbol{X},\zeta)$ with $\mathrm{E}_{\zeta\sim D}[\widetilde{\nabla} g(\boldsymbol{X},\zeta)] \in \partial f(\boldsymbol{X})$, from which one can obtain a Riemannian stochastic subgradient $\widetilde{\nabla}_{\mathcal{R}} g(\boldsymbol{X},\zeta) \in \partial_{\mathcal{R}} g(\boldsymbol{X},\zeta)$ with $\mathrm{E}_{\zeta\sim D}[\widetilde{\nabla}_{\mathcal{R}} g(\boldsymbol{X},\zeta)] \in \partial_{\mathcal{R}} f(\boldsymbol{X})$.

We remark that the above properties mirror those of the stochastic subgradient oracle for stochastic optimization in the Euclidean space; see, e.g., Assumptions (A1) and (A2) in [43].

At the current iterate $\boldsymbol{X}_k$, the Riemannian stochastic subgradient oracle generates a sample $\zeta_k \sim D$ that is independent of $\{\zeta_0,\ldots,\zeta_{k-1}\}$ and returns a Riemannian stochastic subgradient $\widetilde{\nabla}_{\mathcal{R}} g(\boldsymbol{X}_k,\zeta_k)$. Then, the Riemannian stochastic subgradient method generates the next iterate $\boldsymbol{X}_{k+1}$ via

$$\boldsymbol{X}_{k+1} = \mathrm{Retr}_{\boldsymbol{X}_k}(\boldsymbol{\xi}_k) \quad \text{with} \quad \boldsymbol{\xi}_k = -\gamma_k \widetilde{\nabla}_{\mathcal{R}} g(\boldsymbol{X}_k,\zeta_k). \tag{4.14}$$

This generalizes the update (2.6) introduced in Subsection 2.2 for the case where D is the empirical distribution on m data samples.

Similar to the analysis of the Riemannian subgradient and incremental subgradient methods, we begin by establishing the following result; cf. Proposition 1:

PROPOSITION 2. *Suppose that the aforementioned assumptions on problem (4.12) hold, and that a Riemannian stochastic subgradient oracle having properties (a)–(b) above is available. Let $\{\boldsymbol{X}_k\}$ be the sequence generated by the Riemannian stochastic subgradient method (4.14) with arbitrary initialization for solving problem (4.12). Then, for any $\lambda < \frac{1}{L+\tau}$ in (4.1), we have*

$$\gamma_k \,\mathrm{E}\left[\Theta^2(\boldsymbol{X}_k)\right] \le \frac{2\big(\mathrm{E}\left[f_\lambda(\boldsymbol{X}_k)\right] - \mathrm{E}\left[f_\lambda(\boldsymbol{X}_{k+1})\right]\big) + \frac{\gamma_k^2 L^2}{\lambda}}{\lambda\left(\frac{1}{\lambda} - (L+\tau)\right)}, \quad \forall\ k \ge 0.$$

Proof. Using (4.1), the optimality of $P_{\lambda f}(\boldsymbol{X}_{k+1})$, Lemma 1, and the fact that $P_{\lambda f}(\boldsymbol{X}_k) \in \mathrm{St}(n,r)$, we obtain

$$\begin{aligned}
\mathrm{E}_{\zeta_k\sim D}\left[f_\lambda(\boldsymbol{X}_{k+1})\right] &\le f\left(P_{\lambda f}(\boldsymbol{X}_k)\right) + \frac{1}{2\lambda}\,\mathrm{E}_{\zeta_k\sim D}\left[\left\|P_{\lambda f}(\boldsymbol{X}_k) - \boldsymbol{X}_{k+1}\right\|_F^2\right] \\
&\le f\left(P_{\lambda f}(\boldsymbol{X}_k)\right) + \frac{1}{2\lambda}\,\mathrm{E}_{\zeta_k\sim D}\left[\left\|\boldsymbol{X}_k - \gamma_k\widetilde{\nabla}_{\mathcal{R}} g(\boldsymbol{X}_k,\zeta_k) - P_{\lambda f}(\boldsymbol{X}_k)\right\|_F^2\right] \\
&\le f_\lambda(\boldsymbol{X}_k) + \frac{\gamma_k}{\lambda}\,\mathrm{E}_{\zeta_k\sim D}\left[\left\langle\widetilde{\nabla}_{\mathcal{R}} g(\boldsymbol{X}_k,\zeta_k), P_{\lambda f}(\boldsymbol{X}_k) - \boldsymbol{X}_k\right\rangle\right] + \frac{\gamma_k^2 L^2}{2\lambda},
\end{aligned}$$

where the first inequality is due to the optimality of $P_{\lambda f}(\boldsymbol{X}_{k+1})$, the second inequality comes from Lemma 1 and the fact that $P_{\lambda f}(\boldsymbol{X}_k) \in \mathrm{St}(n,r)$, and the third inequality is due to (4.13) and the fact that $\left\|\widetilde{\nabla}_{\mathcal{R}} g(\boldsymbol{X},\zeta)\right\|_F \le \left\|\widetilde{\nabla} g(\boldsymbol{X},\zeta)\right\|_F$. Since we have $\mathrm{E}_{\zeta_k\sim D}\left[\widetilde{\nabla}_{\mathcal{R}} g(\boldsymbol{X}_k,\zeta_k)\right] \in \partial_{\mathcal{R}} f(\boldsymbol{X}_k)$, the L-Lipschitz continuity of f on $\mathcal{U}$, Theorem 1, and (4.11) imply that

$$\mathrm{E}_{\zeta_k\sim D}\left[f_\lambda(\boldsymbol{X}_{k+1})\right] \le f_\lambda(\boldsymbol{X}_k) - \frac{\gamma_k}{2\lambda}\left(\frac{1}{\lambda} - (L+\tau)\right)\left\|\boldsymbol{X}_k - P_{\lambda f}(\boldsymbol{X}_k)\right\|_F^2 + \frac{\gamma_k^2 L^2}{2\lambda}.$$

Upon taking expectation with respect to all the previous realizations $\zeta_0,\ldots,\zeta_{k-1}$ on both sides, we get

$$\mathrm{E}\left[f_\lambda(\boldsymbol{X}_{k+1})\right] \le \mathrm{E}\left[f_\lambda(\boldsymbol{X}_k)\right] - \frac{\gamma_k}{2\lambda}\left(\frac{1}{\lambda} - (L+\tau)\right)\mathrm{E}\left[\left\|P_{\lambda f}(\boldsymbol{X}_k) - \boldsymbol{X}_k\right\|_F^2\right] + \frac{\gamma_k^2 L^2}{2\lambda}.$$

The desired result then follows by rearranging the above inequality and recognizing that $\Theta(\boldsymbol{X}_k) = \lambda^{-1}\left\|P_{\lambda f}(\boldsymbol{X}_k) - \boldsymbol{X}_k\right\|_F$ (see (4.2)). ∎

Now, we can bound the iteration complexity of the Riemannian stochastic subgradient method using Proposition 2.

THEOREM 3. *Under the setting of Proposition 2, suppose that we choose the constant stepsize $\gamma_k = \frac{1}{\sqrt{T+1}}$, $k = 0,1,\ldots$ with T being the total number of iterations and the algorithm returns $\boldsymbol{X}_{\overline{k}}$ with $\overline{k}$ sampled from $\{1,\ldots,T\}$ uniformly at random. Then, we have*

$$\mathrm{E}\left[\Theta^2\left(\boldsymbol{X}_{\overline{k}}\right)\right] \le \frac{1}{\lambda\left(\frac{1}{\lambda} - (L+\tau)\right)}\frac{2\left(f_\lambda(\boldsymbol{X}_0) - \min f_\lambda\right) + \frac{L^2}{\lambda}}{\sqrt{T+1}},$$

where the expectation is taken over all random choices by the algorithm.

Proof. By summing both sides of the relation in Proposition 2 over $k = 0,1,\ldots,T$, we have

$$\sum_{k=0}^{T}\gamma_k\,\mathrm{E}\left[\Theta^2(\boldsymbol{X}_k)\right] \le \frac{2\left(f_\lambda(\boldsymbol{X}_0) - \min f_\lambda\right) + \frac{L^2}{\lambda}\sum_{k=0}^{T}\gamma_k^2}{\lambda\left(\frac{1}{\lambda} - (L+\tau)\right)}.$$

It follows that

$$\sum_{k=0}^{T}\frac{\gamma_k}{\sum_{k=0}^{T}\gamma_k}\,\mathrm{E}\left[\Theta^2(\boldsymbol{X}_k)\right] \le \frac{1}{\lambda\left(\frac{1}{\lambda} - (L+\tau)\right)}\frac{2\left(f_\lambda(\boldsymbol{X}_0) - \min f_\lambda\right) + \frac{L^2}{\lambda}\sum_{k=0}^{T}\gamma_k^2}{\sum_{k=0}^{T}\gamma_k}.$$

To complete the proof, it remains to substitute $\gamma_k = \frac{1}{\sqrt{T+1}}$ into the above inequality and note that the resulting LHS is exactly $\mathrm{E}\left[\Theta^2\left(\boldsymbol{X}_{\overline{k}}\right)\right]$ with the expectation being taken with respect to $\zeta_0,\ldots,\zeta_{T-1},\overline{k}$. ∎

5. Local Linear Convergence for Sharp Instances. So far our discussion on problem (1.1) does not assume any structure on the objective function f besides weak convexity. However, many applications, such as those discussed in Subsection 1.1, give rise to weakly convex objective functions that are not arbitrary but have rather concrete structure. It is thus natural to ask whether the methods we considered can exploit this structure and provably achieve faster convergence rates than those established in Section 4. In this section, we introduce a regularity property of problem (1.1)

called *sharpness* and show that the Riemannian subgradient and incremental subgradient methods will achieve a local linear convergence rate when applied to instances of (1.1) that possess the sharpness property. Then, we will discuss in Section 7 how the notion of sharpness captures, in a unified manner, the structure of both the dual principal component pursuit (DPCP) formulation (1.3) of the robust subspace recovery (RSR) problem and the single-column formulation (1.4) of the orthogonal dictionary learning (DL) problem.

5.1. Sharpness: Weak sharp minima. To begin, let us introduce the notion of a weak sharp minima set.

DEFINITION 1 (Sharpness; cf. [8, 28, 35]). *We say that $\mathcal{X} \subseteq \mathrm{St}(n,r)$ is a set of* weak sharp minima *for the function $h : \mathbb{R}^{n\times r} \to \mathbb{R}$ with parameter $\alpha > 0$ if there exists a constant $\rho > 0$ such that for any $\boldsymbol{X} \in \mathcal{B} := \{\boldsymbol{X} \in \mathbb{R}^{n\times r} : \mathrm{dist}(\boldsymbol{X}, \mathcal{X}) < \rho\} \cap \mathrm{St}(n,r)$, we have*

$$h(\boldsymbol{X}) - h(\boldsymbol{Y}) \geq \alpha\, \mathrm{dist}(\boldsymbol{X}, \mathcal{X})$$

for all $\boldsymbol{Y} \in \mathcal{X}$, where $\mathrm{dist}(\boldsymbol{X}, \mathcal{X}) := \inf_{\boldsymbol{Y}\in\mathcal{X}} \|\boldsymbol{Y} - \boldsymbol{X}\|_F$.

From the definition, it is immediate that if $\mathcal{X}$ is a set of weak sharp minima for h, then it is the set of minimizers of h over $\mathcal{B}$, and the function value grows linearly with the distance to $\mathcal{X}$. Moreover, if h is continuous (e.g., when h is weakly convex), then $\mathcal{X}$ can be taken as closed.

Similar notions of sharpness play a fundamental role in establishing the linear convergence of a host of methods for weakly convex minimization in the Euclidean space. For instance, it is shown in [21] that the subgradient method with geometrically diminishing stepsizes will converge linearly to the optimal solution set when applied to minimize a sharp convex function. Later, the work [13] establishes a similar linear convergence result for sharp weakly convex minimization. In the recent work [36], it is shown that the incremental subgradient, proximal point, and prox-linear methods will converge linearly when applied to minimize a sharp weakly convex function. In this paper, we extend, for the first time, the above results to the manifold setting by establishing the linear convergence of Riemannian subgradient-type methods for minimizing a weakly convex function over the Stiefel manifold under the sharpness property in Definition 1.

5.2. Riemannian subgradient and incremental subgradient methods. Again, we will focus on analyzing the Riemannian incremental subgradient method. The analysis of the Riemannian subgradient method will follow as a special case. We first present the following result, which is crucial for our subsequent development.

PROPOSITION 3. *Under the setting of Proposition 1, for any $\overline{\boldsymbol{X}} \in \mathrm{St}(n,r)$, we have*

$$\begin{aligned}\left\|\boldsymbol{X}_{k+1} - \overline{\boldsymbol{X}}\right\|_F^2 \leq (1 + 2m\gamma_k(L+\tau)) \left\|\boldsymbol{X}_k - \overline{\boldsymbol{X}}\right\|_F^2 - 2m\gamma_k \left(f(\boldsymbol{X}_k) - f(\overline{\boldsymbol{X}})\right) \\ + m^2\gamma_k^2 L^2 + C(m)\gamma_k^3 L^2(L+\tau), \quad \forall\ k \geq 0,\end{aligned}$$

where $C(m) = \frac{1}{3}m(m-1)(2m-1)$.

Proof. According to Lemma 1, for any $\overline{\boldsymbol{X}} \in \mathrm{St}(n,r)$, we have

$$\begin{aligned}\left\|\boldsymbol{X}_{k,i}-\overline{\boldsymbol{X}}\right\|_F^2 &\le \left\|\boldsymbol{X}_{k,i-1}+\boldsymbol{\xi}_{k,i-1}-\overline{\boldsymbol{X}}\right\|_F^2\\ &\overset{(i)}{\le} \left\|\boldsymbol{X}_{k,i-1}-\overline{\boldsymbol{X}}\right\|_F^2 - 2\gamma_k\left\langle\widetilde{\nabla}_{\mathcal{R}} f_i(\boldsymbol{X}_{k,i-1}),\boldsymbol{X}_{k,i-1}-\overline{\boldsymbol{X}}\right\rangle+\gamma_k^2L^2\\ &\overset{(ii)}{\le} \left\|\boldsymbol{X}_{k,i-1}-\overline{\boldsymbol{X}}\right\|_F^2 - 2\gamma_k\left(f_i(\boldsymbol{X}_{k,i-1})-f_i(\overline{\boldsymbol{X}})\right)\\ &\quad+\gamma_k(L+\tau)\left\|\boldsymbol{X}_{k,i-1}-\overline{\boldsymbol{X}}\right\|_F^2+\gamma_k^2L^2,\end{aligned}\tag{5.1}$$

where (i) follows from the fact that $\left\|\widetilde{\nabla}_{\mathcal{R}} f_i(\boldsymbol{X}_{k,i-1})\right\| \le \left\|\widetilde{\nabla} f_i(\boldsymbol{X}_{k,i-1})\right\|_F \le L$ and (ii) is from Theorem 1. Following the derivations of (4.7)–(4.9), we get

$$\begin{aligned}f_i(\overline{\boldsymbol{X}})-f_i(\boldsymbol{X}_{k,i-1}) &\le (i-1)\gamma_kL^2-(f_i(\boldsymbol{X}_k)-f_i(\overline{\boldsymbol{X}})),\\ \left\|\boldsymbol{X}_{k,i-1}-\overline{\boldsymbol{X}}\right\|_F^2 &\le 2(i-1)^2\gamma_k^2L^2+2\left\|\boldsymbol{X}_k-\overline{\boldsymbol{X}}\right\|_F^2.\end{aligned}$$

Substituting the above two upper bounds into (5.1) gives

$$\begin{aligned}\left\|\boldsymbol{X}_{k,i}-\overline{\boldsymbol{X}}\right\|_F^2 \le{}& \left\|\boldsymbol{X}_{k,i-1}-\overline{\boldsymbol{X}}\right\|_F^2 - 2\gamma_k\left(f_i(\boldsymbol{X}_k)-f_i(\overline{\boldsymbol{X}})\right)+2\gamma_k(L+\tau)\left\|\boldsymbol{X}_k-\overline{\boldsymbol{X}}\right\|_F^2\\ &+(2i-1)\gamma_k^2L^2+2(i-1)^2\gamma_k^3L^2(L+\tau).\end{aligned}$$

Upon summing both sides of the above inequality over $i=1,\ldots,m$, we obtain

$$\begin{aligned}\left\|\boldsymbol{X}_{k+1}-\overline{\boldsymbol{X}}\right\|_F^2 \le{}& (1+2m\gamma_k(L+\tau))\left\|\boldsymbol{X}_k-\overline{\boldsymbol{X}}\right\|_F^2-2m\gamma_k\left(f(\boldsymbol{X}_k)-f(\overline{\boldsymbol{X}})\right)\\ &+m^2\gamma_k^2L^2+\frac{1}{3}m(m-1)(2m-1)\gamma_k^3L^2(L+\tau),\end{aligned}$$

which completes the proof. ∎

In order for Riemannian subgradient-type methods to achieve linear convergence when solving sharp instances of problem (1.1), we need to choose the stepsizes appropriately. Motivated by previous works [13, 21, 36, 42, 49] on sharp weakly convex minimization in the Euclidean space, let us consider using geometrically diminishing stepsizes of the form $\gamma_k=\beta^k\gamma_0$, $k=0,1,\ldots$. Then, by applying Proposition 3, we can establish the following local linear convergence result:

THEOREM 4. *Consider the setting of Proposition 1. Suppose further that $\mathcal{X}$ is a set of weak sharp minima for the objective function f in (1.1) with parameter $\alpha>0$ over the set $\mathcal{B}$ defined in Definition 1. Let $\{\boldsymbol{X}_k\}$ be the sequence generated by Riemannian incremental subgradient method (2.6) for solving problem (1.1), in which the initial point $\boldsymbol{X}_0$ satisfies $\mathrm{dist}(\boldsymbol{X}_0,\mathcal{X})<\min\left\{\frac{\alpha}{L+\tau},\rho\right\}$ (so that $\boldsymbol{X}_0\in\mathcal{B}$) and the stepsizes satisfy $\gamma_k=\beta^k\gamma_0$, $k=0,1,\ldots$, where*

$$\gamma_0<\min\left\{\frac{2me_0(\alpha-(L+\tau)e_0)}{d(m)L^2},\frac{e_0}{2m(\alpha-(L+\tau)e_0)}\right\},$$

$$\beta\in[\beta_{\min},1)\quad\text{with}\quad\beta_{\min}:=\sqrt{1+2m\left(L+\tau-\frac{\alpha}{e_0}\right)\gamma_0+\frac{d(m)L^2}{e_0^2}\gamma_0^2},$$

$$d(m)=\frac{5}{3}m^2-m+\frac{1}{3},\quad\text{and}\quad e_0=\min\left\{\max\left\{\mathrm{dist}(\boldsymbol{X}_0,\mathcal{X}),\frac{\alpha}{2(L+\tau)}\right\},\rho\right\}.$$

Then, we have

$$\mathrm{dist}(\boldsymbol{X}_k,\mathcal{X})\le\beta^k\cdot e_0,\quad\forall\ k\ge0.$$

Proof. We first show that $\beta_{\min} \in (0,1)$ and $\gamma_0 > 0$ are well defined. Towards that end, note that $\beta_{\min} = \sqrt{1+v(\gamma_0)}$ with $v(\gamma) = 2m\left(L+\tau-\frac{\alpha}{e_0}\right)\gamma + \frac{d(m)L^2}{e_0^2}\gamma^2$ being quadratic in γ. By definition of γ_0, we immediately have $v(\gamma_0) < 0$. Moreover, the function $\gamma \mapsto v(\gamma)$ attains its minimum at $\overline{\gamma} = \frac{me_0(\alpha-(L+\tau)e_0)}{d(m)L^2}$ with value $v(\overline{\gamma}) = -\frac{m^2(\alpha-(L+\tau)e_0)^2}{d(m)L^2} > -\frac{\alpha^2}{L^2} \geq -1$, where the first inequality is due to $\frac{m^2}{d(m)} \leq 1$ for $m \geq 1$ and $e_0 < \frac{\alpha}{L+\tau}$, and the second inequality is implied by the sharpness assumption because $\alpha\|\boldsymbol{X}-\overline{\boldsymbol{X}}\|_F \leq f(\boldsymbol{X})-f(\overline{\boldsymbol{X}}) \leq L\|\boldsymbol{X}-\overline{\boldsymbol{X}}\|_F$ for any $\boldsymbol{X} \in \mathcal{B}$ and $\overline{\boldsymbol{X}} \in \mathcal{P}_{\mathcal{X}}(\boldsymbol{X})$. Hence, we have $v(\gamma_0) \in (-1,0)$, which implies that $\beta_{\min} \in (0,1)$. On the other hand, since $e_0 < \frac{\alpha}{L+\tau}$, the upper bound on the initial stepsize γ_0 is positive. It follows that γ_0 is well defined.

We now prove the theorem by induction on k. The base case $k = 0$ follows directly from the definition of e_0. For the inductive step, suppose that $\operatorname{dist}(\boldsymbol{X}_k, \mathcal{X}) \leq \beta^k \cdot e_0$ for some $k \geq 0$. Note that this implies $\boldsymbol{X}_k \in \mathcal{B}$. Let $\overline{\boldsymbol{X}} \in \mathcal{P}_{\mathcal{X}}(\boldsymbol{X}_k)$. Clearly, we have $\operatorname{dist}(\boldsymbol{X}_k, \mathcal{X}) = \left\|\boldsymbol{X}_k - \overline{\boldsymbol{X}}\right\|_F$ and $\operatorname{dist}(\boldsymbol{X}_{k+1}, \mathcal{X}) \leq \left\|\boldsymbol{X}_{k+1} - \overline{\boldsymbol{X}}\right\|_F$. Hence, by Proposition 3, the sharpness assumption, and the fact that $\gamma_k \leq \gamma_0$ for $k = 0, 1, \ldots$, we get

$$\begin{aligned}\operatorname{dist}^2(\boldsymbol{X}_{k+1}, \mathcal{X}) \leq (1+2m\gamma_0(L+\tau))\operatorname{dist}^2(\boldsymbol{X}_k, \mathcal{X}) - 2m\gamma_k\alpha\operatorname{dist}(\boldsymbol{X}_k, \mathcal{X}) \\ + m^2\gamma_k^2L^2 + C(m)\gamma_k^3L^2(L+\tau).\end{aligned} \tag{5.2}$$

Observe that the RHS of the above recursion is quadratic in $\operatorname{dist}(\boldsymbol{X}_k, \mathcal{X})$. By definition of γ_0, we have $\gamma_0 < \frac{e_0}{2m(\alpha-(L+\tau)e_0)}$ and hence $\frac{2m\gamma_0\alpha}{1+2m\gamma_0(L+\tau)} < e_0$. This implies that the RHS of (5.2) achieves its maximum when $\operatorname{dist}(\boldsymbol{X}_k, \mathcal{X}) = \beta^k \cdot e_0$. Since $\operatorname{dist}(\boldsymbol{X}_k, \mathcal{X}) \leq \beta^k \cdot e_0$ by the inductive hypothesis, plugging $\gamma_k = \beta^k\gamma_0$ and $\operatorname{dist}(\boldsymbol{X}_k, \mathcal{X}) = \beta^k \cdot e_0$ into (5.2) yields

$$\begin{aligned}&\operatorname{dist}^2(\boldsymbol{X}_{k+1}, \mathcal{X}) \\ &\leq \beta^{2k}e_0^2\left[1+2m\left(L+\tau-\frac{\alpha}{e_0}\right)\gamma_0 + L^2\left(\frac{m^2+C(m)\gamma_0(L+\tau)}{e_0^2}\right)\gamma_0^2\right].\end{aligned} \tag{5.3}$$

Note that $\gamma_0 < \frac{2m(\alpha e_0-(L+\tau)e_0^2)}{d(m)L^2} \leq \frac{m\alpha^2}{2d(m)L^2(L+\tau)} < \frac{1}{m(L+\tau)}$. It then follows from (5.3) that

$$\operatorname{dist}^2(\boldsymbol{X}_{k+1}, \mathcal{X}) \leq \beta^{2k}e_0^2\left[1+2m\left(L+\tau-\frac{\alpha}{e_0}\right)\gamma_0 + \frac{d(m)L^2}{e_0^2}\gamma_0^2\right] \leq \beta^{2(k+1)}e_0^2.$$

This completes the inductive step and hence the proof of Theorem 4. ∎

From Theorem 4, we see that in order to achieve a fast linear convergence rate, one should choose an appropriate γ_0 so that the minimum decay factor $\beta_{\min}$ is as small as possible. By minimizing $\beta_{\min}$ with respect to γ_0, we see that the theoretical minimum value of $\beta_{\min}$ is $\sqrt{1-\frac{m^2(\alpha-(L+\tau)e_0)^2}{d(m)L^2}}$, which is attained at $\gamma_0 = \overline{\gamma}_0 = \frac{me_0(\alpha-(L+\tau)e_0)}{d(m)L^2}$. This suggests that subject to the requirement in Theorem 4, the initial stepsize γ_0 should be set as close to $\overline{\gamma}_0$ as possible. As an illustration, consider the case where the sharpness property holds globally over the Stiefel manifold (i.e., $\mathcal{B} = \operatorname{St}(n,r)$ in Definition 1). Then, the parameter ρ can be set as large as possible. In this case, we have $e_0 = \max\left\{\operatorname{dist}(\boldsymbol{X}_0, \mathcal{X}), \frac{\alpha}{2(L+\tau)}\right\}$, and the condition on γ_0 in Theorem 4 becomes $\gamma_0 < \frac{2me_0(\alpha-(L+\tau)e_0)}{d(m)L^2}$. This implies that we can choose $\gamma_0 = \overline{\gamma}_0$ to obtain the smallest

possible $\beta_{\min}$. Note, however, that the larger the initialization error $\operatorname{dist}(\boldsymbol{X}_0, \mathcal{X})$, the larger the minimum decay factor $\beta_{\min}$. In particular, from the expression for $\beta_{\min}$ above, we see that $\beta_{\min}$ approaches 1 as $\operatorname{dist}(\boldsymbol{X}_0, \mathcal{X})$ approaches its maximum $\frac{\alpha}{L^+\tau}$.

We end this section by comparing the sharpness property with the *Riemannian regularity condition* used in [3] and [69] for orthogonal DL and RSR, respectively. For a target solution set $\mathcal{X}$, the Riemannian regularity condition stipulates the existence of a constant $\kappa > 0$ such that $\left\langle \widetilde{\nabla}_{\mathcal{R}} f(\boldsymbol{X}), \boldsymbol{X} - \boldsymbol{Y} \right\rangle \geq \kappa \operatorname{dist}(\boldsymbol{X}, \mathcal{X})$ for all $\boldsymbol{X}$ in a *small neighborhood* of $\mathcal{X}$ and $\boldsymbol{Y} \in \mathcal{P}_{\mathcal{X}}(\boldsymbol{X})$. This condition is motivated by the need to bound the inner product term on the LHS in the convergence analysis of the Riemannian subgradient method; see (5.1) with $f_i = f$ and $\boldsymbol{X}_{k,i-1} = \boldsymbol{X}_k$. Informally, the Riemannian regularity condition is a combination of the Riemannian subgradient inequality in Theorem 1 and the sharpness property in Definition 1. However, the tangling of these two elements potentially restricts the applicability of the Riemannian regularity condition. In particular, since the Riemannian regularity condition can only hold locally, it cannot be used to establish global convergence and iteration complexity results for the Riemannian subgradient method.

6. Extension to Optimization over a Compact Embedded Submanifold. There is of course no conceptual difficulty in adapting the Riemannian subgradient-type methods in Subsection 2.2 to minimize weakly convex functions over more general manifolds. All that is needed is an efficiently computable retraction on the manifold of interest. In this section, let us briefly demonstrate how the machinery developed in the previous sections can be extended to study the convergence behavior of Riemannian subgradient-type methods when the manifold in question is compact and defined by a certain smooth mapping.

Riemannian subgradient inequality. Our starting point is the following generalization of the Riemannian subgradient inequality in Theorem 1, which applies to restrictions of weakly convex functions on a class of compact embedded submanifolds of the Euclidean space. Some examples of manifolds in this class include the generalized Stiefel manifold, oblique manifold, and symplectic manifold; see, e.g., [2].

COROLLARY 1. *Let $\mathcal{M}$ be a compact submanifold of $\mathbb{R}^p$ given by $\mathcal{M} = \{\boldsymbol{X} \in \mathbb{R}^p : F(\boldsymbol{X}) = \boldsymbol{0}\}$, where $F : \mathbb{R}^p \to \mathbb{R}^q$ is a smooth mapping whose derivative $DF(\boldsymbol{X})$ at $\boldsymbol{X}$ has full row rank for all $\boldsymbol{X} \in \mathcal{M}$. Then, for any weakly convex function $h : \mathbb{R}^p \to \mathbb{R}$, there exists a constant $c > 0$ such that*

$$h(\boldsymbol{Y}) \geq h(\boldsymbol{X}) + \left\langle \widetilde{\nabla}_{\mathcal{R}} h(\boldsymbol{X}), \boldsymbol{Y} - \boldsymbol{X} \right\rangle - c\|\boldsymbol{Y} - \boldsymbol{X}\|_F^2$$

for all $\boldsymbol{X}, \boldsymbol{Y} \in \mathcal{M}$ and $\widetilde{\nabla}_{\mathcal{R}} h(\boldsymbol{X}) \in \partial_{\mathcal{R}} h(\boldsymbol{X})$.

Proof. By our assumptions on F and [2, Equation (3.19)], we have $\mathrm{T}_{\boldsymbol{X}}\mathcal{M} = \ker(DF(\boldsymbol{X}))$, where $\ker(T)$ denotes the kernel of the operator T. Thus, the projector $\mathcal{P}^{\perp}_{\mathrm{T}_{\boldsymbol{X}}\mathcal{M}}$ is given by $DF(\boldsymbol{X})^\top (DF(\boldsymbol{X})DF(\boldsymbol{X})^\top)^{-1} DF(\boldsymbol{X})$. Following the proof of Theorem 1, we need to bound

$$\begin{aligned}
&\left\langle \widetilde{\nabla} h(\boldsymbol{X}), \mathcal{P}^{\perp}_{\mathrm{T}_{\boldsymbol{X}}\mathcal{M}}(\boldsymbol{Y} - \boldsymbol{X}) \right\rangle \geq -\|\widetilde{\nabla} h(\boldsymbol{X})\|_F \cdot \|\mathcal{P}^{\perp}_{\mathrm{T}_{\boldsymbol{X}}\mathcal{M}}(\boldsymbol{Y} - \boldsymbol{X})\|_F \\
&= -\|\widetilde{\nabla} h(\boldsymbol{X})\|_F \cdot \|DF(\boldsymbol{X})^\top (DF(\boldsymbol{X})DF(\boldsymbol{X})^\top)^{-1} DF(\boldsymbol{X})(\boldsymbol{Y} - \boldsymbol{X})\|_F \\
&\geq -\|\widetilde{\nabla} h(\boldsymbol{X})\|_F \cdot \max_{\boldsymbol{X} \in \mathcal{M}} \|DF(\boldsymbol{X})^\top (DF(\boldsymbol{X})DF(\boldsymbol{X})^\top)^{-1}\|_F \cdot \|DF(\boldsymbol{X})(\boldsymbol{Y} - \boldsymbol{X})\|_F.
\end{aligned}$$

Since h is weakly convex on $\mathbb{R}^p$, it is Lipschitz continuous on any bounded open convex set $\mathcal{U}$ that contains $\mathcal{M}$. Thus, the term $\|\widetilde{\nabla} h(\boldsymbol{X})\|_F$ is bounded above. Moreover, the

compactness of $\mathcal{M}$ implies that the term $\max_{\boldsymbol{X}\in\mathcal{M}} \|DF(\boldsymbol{X})^\top (DF(\boldsymbol{X})DF(\boldsymbol{X})^\top)^{-1}\|_F$ is also bounded above. Lastly, observe that $F(\boldsymbol{Y}) = F(\boldsymbol{X}) + DF(\boldsymbol{X})(\boldsymbol{Y}-\boldsymbol{X}) + \mathcal{O}(\|\boldsymbol{Y}-\boldsymbol{X}\|_F^2)$ by Taylor's theorem and $F(\boldsymbol{X}) = F(\boldsymbol{Y}) = \boldsymbol{0}$ whenever $\boldsymbol{X}, \boldsymbol{Y} \in \mathcal{M}$. Hence, we have $\|DF(\boldsymbol{X})(\boldsymbol{Y}-\boldsymbol{X})\|_F = \mathcal{O}(\|\boldsymbol{Y}-\boldsymbol{X}\|_F^2)$. Putting these together, we conclude that $\left|\left\langle \widetilde{\nabla} h(\boldsymbol{X}), \mathcal{P}_{\mathrm{T}_{\boldsymbol{X}}\mathcal{M}}^{\perp}(\boldsymbol{Y}-\boldsymbol{X})\right\rangle\right| = \mathcal{O}(\|\boldsymbol{Y}-\boldsymbol{X}\|_F^2)$. The rest of the argument is similar to that in the proof of Theorem 1. ∎

General retractions. The notion of retraction introduced in Subsection 2.1 for the Stiefel manifold can be easily adapted to that for general manifolds. Specifically, a retraction on the manifold $\mathcal{M}$ is a smooth map $\mathrm{Retr} : \mathrm{T}\mathcal{M} \to \mathcal{M}$ from the tangent bundle $\mathrm{T}\mathcal{M}$ onto the manifold $\mathcal{M}$ that satisfies $\mathrm{Retr}_{\boldsymbol{X}}(\boldsymbol{0}) = \boldsymbol{X}$ and $D\,\mathrm{Retr}_{\boldsymbol{X}}(\boldsymbol{0}) = \boldsymbol{Id}$ for all $\boldsymbol{X} \in \mathcal{M}$. Unlike the polar decomposition-based retraction on the Stiefel manifold, a general retraction may not have the Lipschitz-like property in Lemma 1. Nevertheless, a retraction on a compact submanifold $\mathcal{M}$ satisfies a *second-order boundedness* property [7]; i.e., there exists a constant $b \geq 0$ such that for all $\boldsymbol{X} \in \mathcal{M}$ and $\boldsymbol{\xi} \in \mathrm{T}_{\boldsymbol{X}}\mathcal{M}$,

$$\|\,\mathrm{Retr}_{\boldsymbol{X}}(\boldsymbol{\xi}) - \boldsymbol{X} - \boldsymbol{\xi}\|_F \leq b\|\boldsymbol{\xi}\|_F^2.$$

This allows us to replace the result in Lemma 1 by

$$\begin{aligned}\left\|\mathrm{Retr}_{\boldsymbol{X}}(\boldsymbol{\xi}) - \overline{\boldsymbol{X}}\right\|_F &= \|(\boldsymbol{X}+\boldsymbol{\xi}) - \overline{\boldsymbol{X}} + \mathrm{Retr}_{\boldsymbol{X}}(\boldsymbol{\xi}) - (\boldsymbol{X}+\boldsymbol{\xi})\|_F \\ &\leq \|\boldsymbol{X}+\boldsymbol{\xi}-\overline{\boldsymbol{X}}\|_F + b\|\boldsymbol{\xi}\|_F^2,\end{aligned}$$

which holds for any $\boldsymbol{X}, \overline{\boldsymbol{X}} \in \mathcal{M}$ and $\boldsymbol{\xi} \in \mathrm{T}_{\boldsymbol{X}}\mathcal{M}$. Although the above inequality has the extra term $b\|\boldsymbol{\xi}\|_F^2$, it can still be used to establish convergence guarantees (with slightly worse constants) for the Riemannian subgradient-type methods considered in Subsection 2.2. Specifically, by following the analyses in Sections 4 and 5, we can show that for problem (1.1) with the Stiefel manifold $\mathrm{St}(n,r)$ being replaced by a manifold of the type considered in Corollary 1, the iteration complexity of Riemannian subgradient-type methods for computing an ε-nearly stationary point is $\mathcal{O}(\varepsilon^{-4})$, and the Riemannian subgradient and incremental subgradient methods will achieve a local linear convergence rate if the instance satisfies the sharpness property in Definition 1.

7. Applications and Numerical Results. In this section, we apply the Riemannian subgradient-type methods in Subsection 2.2 to solve the RSR and orthogonal DL problems. As described in Section 1, the objective functions of both problems are weakly convex. Thus, Theorem 2 and Theorem 3 ensure that the Riemannian subgradient-type methods with arbitrary initialization will have a global convergence rate of $\mathcal{O}(k^{-1/4})$ when utilized to solve those problems. We also discuss the sharpness properties of the RSR and orthogonal DL problems. For reproducible research, our code for generating the numerical results can be found at

https://github.com/lixiao0982/Riemannian-subgradient-methods

7.1. Robust subspace recovery (RSR). We begin with the DPCP formulation (1.3) of the RSR problem, which has a relatively simpler form than the least absolute deviation (LAD) formulation (1.2). Recall that the objective function in (1.3) takes the form $\mathrm{St}(n,r) \ni \boldsymbol{X} \mapsto f(\boldsymbol{X}) = \frac{1}{m}\sum_{i=1}^m \left\|\widetilde{\boldsymbol{y}}_i^\top \boldsymbol{X}\right\|_2$, where $\widetilde{\boldsymbol{y}}_i \in \mathbb{R}^n$ $(i = 1,\ldots,m)$ denotes the i-th column of $\widetilde{\boldsymbol{Y}} = \begin{bmatrix}\boldsymbol{Y} & \boldsymbol{O}\end{bmatrix}\boldsymbol{\Gamma} \in \mathbb{R}^{n\times m}$, the columns $\boldsymbol{y}_i$ of $\boldsymbol{Y} \in \mathbb{R}^{n\times m_1}$ form inlier points spanning a d-dimensional subspace $\mathcal{S}$ with $r = n-d$, the columns $\boldsymbol{o}_i$ of $\boldsymbol{O} \in \mathbb{R}^{n\times m_2}$ form outlier points, and $\boldsymbol{\Gamma} \in \mathbb{R}^{m\times m}$ is an unknown

permutation. Note that f is rotationally invariant; i.e., $f(\boldsymbol{X}) = f(\boldsymbol{X}\boldsymbol{R})$ for any $\boldsymbol{X} \in \mathrm{St}(n,r)$ and $\boldsymbol{R} \in \mathrm{St}(r,r)$.

Sharpness. Let $\boldsymbol{S}^\perp \in \mathrm{St}(n,r)$ be an orthonormal basis of $\mathcal{S}^\perp$. Since the goal of DPCP is to find an orthonormal basis (but not necessary $\boldsymbol{S}^\perp$) for $\mathcal{S}^\perp$, we are interested in the elements in the set $\mathcal{X} = \{\boldsymbol{S}^\perp \boldsymbol{R} \in \mathbb{R}^{n\times r} : \boldsymbol{R} \in \mathrm{St}(r,r)\}$. Due to rotation invariance, f is constant on $\mathcal{X}$. To study the sharpness property of problem (1.3), let us introduce two quantities that reflect how well distributed the inliers and outliers are:

$$c_{\boldsymbol{Y},\min} := \frac{1}{m_1} \inf_{\substack{\boldsymbol{D}\in\mathbb{R}^{n\times \ell},\\ \|\boldsymbol{D}\|_F=1,\, \mathrm{col}(\boldsymbol{D})\subseteq\mathcal{S}}} \sum_{i=1}^{m_1} \|\boldsymbol{y}_i^\top \boldsymbol{D}\|_2, \tag{7.1}$$

$$c_{\boldsymbol{O},\max} := \frac{1}{m_2} \sup_{\substack{\boldsymbol{B}\in\mathbb{R}^{n\times r},\\ \|\boldsymbol{B}\|_F=1}} \sum_{i=1}^{m_2} \|\boldsymbol{o}_i^\top \boldsymbol{B}\|_2. \tag{7.2}$$

Here, $\ell = \min\{d,r\}$ and $\mathrm{col}(\boldsymbol{D})$ denotes the column space of $\boldsymbol{D}$. In a nutshell, larger values of $c_{\boldsymbol{Y},\min}$ (respectively, smaller values of $c_{\boldsymbol{O},\max}$) correspond to a more uniform distribution of inliers (respectively, outliers). As the following proposition shows, the quantities $c_{\boldsymbol{Y},\min}$ and $c_{\boldsymbol{O},\max}$ can be used to capture the sharpness property of the DPCP formulation (1.3).

PROPOSITION 4. *Suppose that $m_2 c_{\boldsymbol{O},\max} \le \frac{1}{2} m_1 c_{\boldsymbol{Y},\min}$. Then, the DPCP formulation (1.3) has $\mathcal{X}$ as a set of weak sharp minima with parameter $\alpha = \frac{1}{m}\big(\frac{1}{2}m_1 c_{\boldsymbol{Y},\min} - m_2 c_{\boldsymbol{O},\max}\big) > 0$ over the set $\mathcal{B} = \mathrm{St}(n,r)$; i.e.,*

$$f(\boldsymbol{X}) - f(\boldsymbol{S}^\perp) \ge \alpha \,\mathrm{dist}(\boldsymbol{X},\mathcal{X}), \quad \forall\ \boldsymbol{X} \in \mathrm{St}(n,r).$$

Proof. Let $\boldsymbol{X} \in \mathrm{St}(n,r)$ be arbitrary. For any $\boldsymbol{X}^\star \in \mathcal{P}_{\mathcal{X}}(\boldsymbol{X})$, we have $f(\boldsymbol{X}^\star) = f(\boldsymbol{S}^\perp)$ and

$$\begin{aligned} f(\boldsymbol{X}) - f(\boldsymbol{S}^\perp) &= \frac{1}{m}\sum_{i=1}^{m} \big\|\widetilde{\boldsymbol{y}}_i^\top \boldsymbol{X}\big\|_2 - \frac{1}{m}\sum_{i=1}^{m} \big\|\widetilde{\boldsymbol{y}}_i^\top \boldsymbol{S}^\perp\big\|_2 \\ &= \frac{1}{m}\sum_{i=1}^{m_1} \big\|\boldsymbol{y}_i^\top \boldsymbol{X}\big\|_2 + \frac{1}{m}\left(\sum_{i=1}^{m_2} \big\|\boldsymbol{o}_i^\top \boldsymbol{X}\big\|_2 - \sum_{i=1}^{m_2} \big\|\boldsymbol{o}_i^\top \boldsymbol{S}^\perp\big\|_2\right), \end{aligned} \tag{7.3}$$

where $\boldsymbol{y}_i$ (respectively, $\boldsymbol{o}_i$) is the i-th column of $\boldsymbol{Y}$ (respectively, $\boldsymbol{O}$), and the second line follows because the inliers $\{\boldsymbol{y}_i\}_{i=1}^{m_1}$ are orthogonal to $\boldsymbol{S}^\perp$. Now, let us derive lower bounds for the two terms on the right-hand side separately.

For the first term, let $\boldsymbol{S} \in \mathbb{R}^{n\times d}$ be an orthonormal basis of the subspace $\mathcal{S}$. By projecting $\boldsymbol{X}$ onto the orthogonal subspaces $\mathcal{S}$ and $\mathcal{S}^\perp$, we have

$$\boldsymbol{X} = \boldsymbol{S}\boldsymbol{S}^\top \boldsymbol{X} + \boldsymbol{S}^\perp (\boldsymbol{S}^\perp)^\top \boldsymbol{X}. \tag{7.4}$$

For $i = 1,\dots,r$, let $\phi_i = \arccos(\sigma_i((\boldsymbol{S}^\perp)^\top \boldsymbol{X}))$ be the i-th smallest principal angle between the subspaces spanned by $\boldsymbol{X}$ and $\mathcal{S}^\perp$, where $\sigma_i(\cdot)$ denotes the i-th largest singular value [51]. Then, we can write $(\boldsymbol{S}^\perp)^\top \boldsymbol{X} = \boldsymbol{U}\cos(\boldsymbol{\Phi})\boldsymbol{W}^\top$, where $\cos(\boldsymbol{\Phi}) \in \mathbb{R}^{r\times r}$ is the diagonal matrix with $\cos(\phi_1) \ge \cdots \ge \cos(\phi_r)$ on its diagonal and $\boldsymbol{U} \in \mathbb{R}^{r\times r}$, $\boldsymbol{W} \in \mathbb{R}^{r\times r}$ are orthogonal matrices. On the other hand, according to [29, Theorem 2.7], the i-th smallest principal angle between the subspaces spanned by $\boldsymbol{X}$ and $\mathcal{S}$ is $\widetilde{\phi}_i = \frac{\pi}{2} - \phi_{r-i+1}$, where $i = 1,\dots,\ell$ with $\ell = \min\{d,r\}$. Hence, we

can write $\boldsymbol{S}^\top \boldsymbol{X} = \boldsymbol{V}\sin(\widetilde{\boldsymbol{\Phi}})\boldsymbol{H}^\top$, where $\sin(\widetilde{\boldsymbol{\Phi}}) \in \mathbb{R}^{\ell\times\ell}$ is the diagonal matrix with $\sin(\phi_r) \geq \cdots \geq \sin(\phi_{r-\ell+1})$ on its diagonal and $\boldsymbol{V} \in \mathbb{R}^{d\times\ell}$, $\boldsymbol{H} \in \mathbb{R}^{r\times\ell}$ are orthogonal matrices. These, together with (7.4), yield $\boldsymbol{X} = \boldsymbol{S}\boldsymbol{V}\sin(\widetilde{\boldsymbol{\Phi}})\boldsymbol{H}^\top + \boldsymbol{S}^\perp\boldsymbol{U}\cos(\boldsymbol{\Phi})\boldsymbol{W}^\top$. Hence, we can bound

$$\begin{aligned}\sum_{i=1}^{m_1}\|\boldsymbol{y}_i^\top\boldsymbol{X}\|_2 &= \sum_{i=1}^{m_1}\|\boldsymbol{y}_i^\top\boldsymbol{S}\boldsymbol{V}\sin(\widetilde{\boldsymbol{\Phi}})\|_2 \\ &= \|\boldsymbol{S}\boldsymbol{V}\sin(\widetilde{\boldsymbol{\Phi}})\|_F\sum_{i=1}^{m_1}\left\|\boldsymbol{y}_i^\top\frac{\boldsymbol{S}\boldsymbol{V}\sin(\widetilde{\boldsymbol{\Phi}})}{\|\boldsymbol{S}\boldsymbol{V}\sin(\widetilde{\boldsymbol{\Phi}})\|_F}\right\|_2 \geq m_1 c_{\boldsymbol{Y},\min}\|\sin(\widetilde{\boldsymbol{\Phi}})\|_F,\end{aligned} \tag{7.5}$$

where $c_{\boldsymbol{Y},\min}$ is defined in (7.1). On the other hand, observe that

$$\begin{aligned}\operatorname{dist}^2(\boldsymbol{X},\mathcal{X}) &= \underset{\boldsymbol{R}\in\mathrm{St}(r,r)}{\text{minimize}}\|\boldsymbol{X}-\boldsymbol{S}^\perp\boldsymbol{R}\|_F^2 = \|\boldsymbol{X}-\boldsymbol{S}^\perp\boldsymbol{U}\boldsymbol{W}^\top\|_F^2 \\ &= 2r - 2\operatorname{trace}(\cos(\boldsymbol{\Phi})) = 2\sum_{i=1}^{\ell}(1-\cos(\phi_i)) = 4\sum_{i=1}^{\ell}\sin^2(\phi_i/2) \leq 4\|\sin(\widetilde{\boldsymbol{\Phi}})\|_F^2,\end{aligned} \tag{7.6}$$

where the second equality follows from the solution to the orthogonal Procrustes problem [48] and the fourth equality utilizes the fact that the number of nonzero principal angles in $\boldsymbol{\Phi}$ is at most $\ell = \min\{d, r\}$ [29, Theorem 2.7]. Combining (7.5) and (7.6) gives

$$\sum_{i=1}^{m_1}\|\boldsymbol{y}_i^\top\boldsymbol{X}\|_2 \geq \frac{1}{2}m_1 c_{\boldsymbol{Y},\min}\operatorname{dist}(\boldsymbol{X},\mathcal{X}). \tag{7.7}$$

Now, let us consider the second term on the right-hand side of (7.3). Let $\boldsymbol{R}^\star = \operatorname{argmin}_{\boldsymbol{R}\in\mathrm{St}(r,r)}\|\boldsymbol{X}-\boldsymbol{S}^\perp\boldsymbol{R}\|_F$. Then, we have

$$\begin{aligned}&\left|\sum_{i=1}^{m_2}\|\boldsymbol{o}_i^\top\boldsymbol{X}\|_2 - \|\boldsymbol{o}_i^\top\boldsymbol{S}^\perp\|_2\right| \leq \sum_{i=1}^{m_2}\left|\|\boldsymbol{o}_i^\top\boldsymbol{X}\|_2 - \|\boldsymbol{o}_i^\top\boldsymbol{S}^\perp\boldsymbol{R}^\star\|_2\right| \\ &\leq \sum_{i=1}^{m_2}\|\boldsymbol{o}_i^\top(\boldsymbol{X}-\boldsymbol{S}^\perp\boldsymbol{R}^\star)\|_2 = \|\boldsymbol{X}-\boldsymbol{S}^\perp\boldsymbol{R}^\star\|_F\sum_{i=1}^{m_2}\left\|\boldsymbol{o}_i^\top\frac{\boldsymbol{X}-\boldsymbol{S}^\perp\boldsymbol{R}^\star}{\|\boldsymbol{X}-\boldsymbol{S}^\perp\boldsymbol{R}^\star\|_F}\right\|_2 \\ &\leq m_2 c_{\boldsymbol{O},\max}\operatorname{dist}(\boldsymbol{X},\mathcal{X}),\end{aligned} \tag{7.8}$$

where $c_{\boldsymbol{O},\max}$ is defined in (7.2).

By plugging (7.7) and (7.8) into (7.3), the desired result follows. ∎

The requirement $m_2 c_{\boldsymbol{O},\max} \leq \frac{1}{2}m_1 c_{\boldsymbol{Y},\min}$ in Proposition 4 determines the number of outliers that can be tolerated. Now, let us give probabilistic estimates of the quantities $c_{\boldsymbol{Y},\min}$ and $c_{\boldsymbol{O},\max}$ under the popular *Haystack model* (see, e.g., [34,40,68]) of the input data. The model stipulates that the inliers $\{\boldsymbol{y}_i\}_{i=1}^{m_1}$ are i.i.d. according to the Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \frac{1}{d}\mathcal{P}_{\mathcal{S}})$ with $\mathcal{P}_{\mathcal{S}}$ being the orthogonal projector onto the d-dimensional subspace $\mathcal{S}$, while the outliers $\{\boldsymbol{o}_i\}_{i=1}^{m_2}$ are i.i.d. according to the Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \frac{1}{n}\boldsymbol{I}_n)$.

LEMMA 2. *Under the Haystack model, the event*

$$c_{\boldsymbol{Y},\min} \geq \sqrt{\frac{2}{d\pi}} - \sqrt{\frac{8\ell}{m_1}} - \frac{c_1}{\sqrt{m_1}}$$

will hold with probability at least $1-2\exp(-\frac{c_1^2 d}{2})$ *for some constant* $c_1>0$, *where* $\ell=\min\{d,r\}$. *Moreover, the event*

$$c_{\boldsymbol{O},\max} \le \frac{1}{\sqrt{n}} + \sqrt{\frac{8r}{m_2}} + \frac{c_2}{\sqrt{m_2}}$$

will hold with probability at least $1-2\exp(-\frac{c_2^2 n}{2})$ *for some constant* $c_2>0$.

The proof of Lemma 2 can be found in Appendix A. Lemma 2 implies that under the Haystack model, if the numbers of inliers m_1 and outliers m_2 satisfy $m_1 \gtrsim d\ell$ and $m_2 \gtrsim nr$, then we will have $c_{\boldsymbol{Y},\min} \gtrsim \frac{1}{\sqrt{d}}$ and $c_{\boldsymbol{O},\max} \lesssim \frac{1}{\sqrt{n}}$ with high probability. Combining Theorem 4, Proposition 4, and Lemma 2, we see that as long as

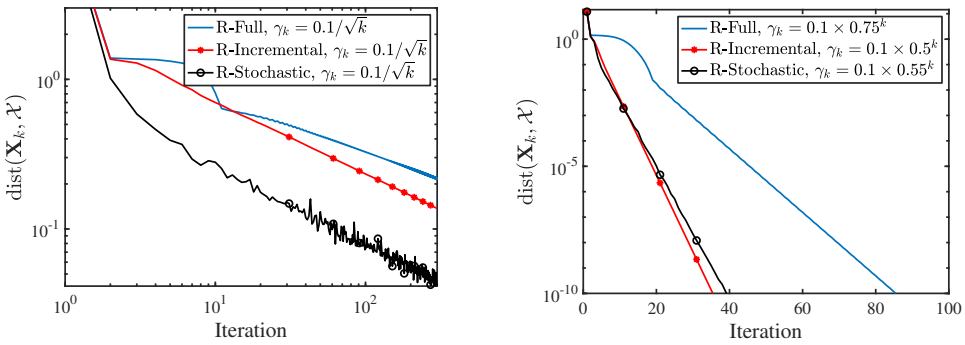
$$m_2 \lesssim \sqrt{\frac{n}{d}} m_1 \tag{7.9}$$

so that $m_2 c_{\boldsymbol{O},\max} \le \frac{1}{2} m_1 c_{\boldsymbol{Y},\min}$, the Riemannian subgradient and incremental subgradient methods with geometrically diminishing stepsizes and a proper initialization will converge linearly to an orthonormal basis of $\mathcal{S}^\perp$. One initialization strategy is to take the bottom eigenvectors of $\widetilde{\boldsymbol{Y}}\widetilde{\boldsymbol{Y}}^\top$ [40, 70].

It is instructive to compare the bound (7.9) with those in the literature. When $d=\mathcal{O}(1)$ or when both d and r are on the order of n, our bound (7.9) holds in the regime $m \gtrsim n^2$. In this regime, the algorithms proposed in [40, 68] can recover $\mathcal{S}$ as long as $m_2 \lesssim \frac{\sqrt{n(n-d)}}{d} m_1$; see [40, Section 5.5.2]. Such a bound is superior to ours when $d=\mathcal{O}(1)$ but is comparable when both d and r are on the order of n. When $r=\mathcal{O}(1)$, our bound (7.9) holds in the regime $m \gtrsim n$, which is superior to the bound $m_2 \lesssim \frac{n-d}{d} m_1$ established in [34, 67] for the same regime. We remark that there are other works [32, 62, 70] studying the RSR problem. However, they differ from our work in that they either assume different data models, require additional data structures, or consider the asymptotic setting $m \to \infty$.

To further demonstrate the power of Proposition 4, let us use it to establish the sharpness property of the LAD formulation (1.2). To begin, let g be the objective function in (1.2) and $\boldsymbol{S} \in \mathrm{St}(n,d)$ be an orthonormal basis of $\mathcal{S}$. We are interested in the set $\mathcal{D}=\{\boldsymbol{S}\boldsymbol{R} : \boldsymbol{R} \in \mathrm{St}(d,d)\}$, whose elements are different orthonormal bases of $\mathcal{S}$. Now, observe that for any $\boldsymbol{X} \in \mathrm{St}(n,r)$, we can find an orthonormal basis $\boldsymbol{Z} \in \mathrm{St}(n,d)$ of $\mathrm{col}(\boldsymbol{X})^\perp$, and vice versa, such that $f(\boldsymbol{X})=g(\boldsymbol{Z})$ (recall that f is the objective function in (1.3)). Hence, Proposition 4 asserts that $g(\boldsymbol{Z})-g(\boldsymbol{S}) \ge \alpha \,\mathrm{dist}(\boldsymbol{X},\mathcal{X})$. By invoking [29, Theorem 2.7], we obtain $\mathrm{dist}(\boldsymbol{X},\mathcal{X})=\mathrm{dist}(\boldsymbol{Z},\mathcal{D})$, which shows that $\mathcal{D}$ is a set of weak sharp minima with parameter α over the set $\mathcal{B}=\mathrm{St}(n,d)$.

Experiments. We first randomly sample a subspace $\mathcal{S}$ with co-dimension $r=10$ in ambient dimension $n=100$. We then generate $m_1=1500$ inliers uniformly at random from the unit sphere in $\mathcal{S}$ and $m_2=3500$ outliers uniformly at random from the unit sphere in $\mathbb{R}^n$. We generate a standard Gaussian random vector and use it to initialize all the algorithms, as such an initialization provides comparable performance with the carefully designed initialization in [40, 70]. The numerical results are displayed in Figure 1. Sublinear convergence can be observed from the log-log plot in Figure 1a, where we use the diminishing stepsizes suggested in Theorem 2 and Theorem 3. In Figure 1b, we use geometrically diminishing stepsizes of the form $\gamma_k=\beta^k\gamma_0$. We fix $\gamma_0=0.1$ and tune the best decay factor β for each algorithm. A linear rate of convergence can be observed, which corroborates our theoretical results.



(a) DPCP: Diminishing stepsizes of the form $\gamma_k = 0.1/\sqrt{k}$, $k = 1, 2, \ldots$

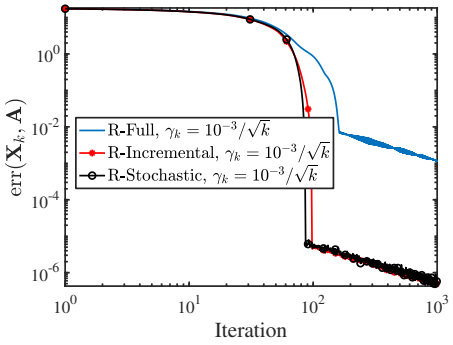
(b) DPCP: Geometrically diminishing stepsizes of the form $\gamma_k = 0.1 \times \beta^k$, $k = 0, 1, \ldots$

Fig. 1: Convergence performance of Riemannian subgradient-type methods for the DPCP formulation (1.3).

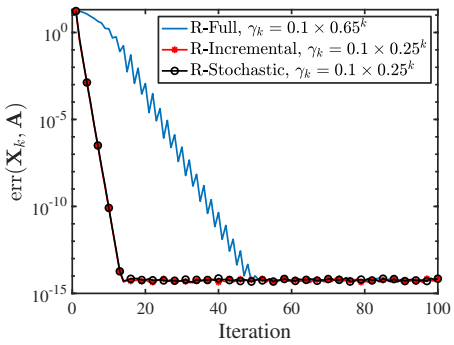
7.2. Orthogonal dictionary learning (ODL). We now turn to the orthogonal DL problem. Given $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{S} \in \mathbb{R}^{n\times m}$, where $\boldsymbol{A} \in \mathrm{St}(n, n)$ is an unknown orthonormal dictionary and each column of $\boldsymbol{S} \in \mathbb{R}^{n\times m}$ is sparse, we can try to recover the columns of $\boldsymbol{A}$ one at a time by considering the formulation (1.4), whose objective function takes the form $\mathbb{S}^{n-1} \ni \boldsymbol{x} \mapsto f(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^m \left|\boldsymbol{y}_i^\top \boldsymbol{x}\right|$, or to recover the entire dictionary by considering the formulation (1.5), whose objective function takes the form $\mathrm{St}(n, n) \ni \boldsymbol{X} \mapsto f(\boldsymbol{X}) = \frac{1}{m}\sum_{i=1}^m \left\|\boldsymbol{y}_i^\top \boldsymbol{X}\right\|_1$.

Sharpness. The sharpness property of the formulation (1.4) has been studied in [3], while that of (1.5) has been studied in [59] only in the *asymptotic* regime; i.e., when the number of samples m tends to infinity. Although we do not yet know how to establish the sharpness property of (1.5) in the *finite-sample* regime, the following numerical results suggest that problem (1.5) likely possesses such a property, as the Riemannian subgradient-type methods with geometrically diminishing stepsizes exhibit linear convergence behavior, even with a random initialization. We leave the study of the sharpness property of (1.5) in the finite-sample regime as a future work.

Experiments. For the orthogonal DL application, we generate synthetic data in the same way as [3]. Specifically, we first generate the underlying orthogonal dictionary $\boldsymbol{A} \in \mathrm{St}(n, n)$ with $n = 30$ randomly and set the number of samples m to be $m = 1643 \approx 10 \times n^{1.5}$. We then generate a sparse coefficient matrix $\boldsymbol{S} \in \mathbb{R}^{n\times m}$, in which each entry follows the Bernoulli-Gaussian distribution with parameter 0.3 (sparsity)—i.e., each entry $\boldsymbol{S}_{i,j}$ is drawn independently from the standard Gaussian distribution with probability 0.3 and is set to zero otherwise. Lastly, we obtain the observation $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{S}$. As before, we generate a standard Gaussian random vector and use it to initialize all the algorithms. To evaluate the performance of the algorithms, we define the error between $\boldsymbol{X}$ and $\boldsymbol{A}$ as $\mathrm{err}(\boldsymbol{X}, \boldsymbol{A}) = \sum_{i=1}^n \left|\max_{1\le j\le n} \left|[\boldsymbol{x}_i^\top \boldsymbol{A}]_j\right| - 1\right|$, where $\boldsymbol{x}_i$ is the i-th column of $\boldsymbol{X}$. Clearly, $\mathrm{err}(\boldsymbol{X}, \boldsymbol{A}) = 0$ when $\boldsymbol{X}$ and $\boldsymbol{A}$ are equal up to permutation and sign ambiguities. The numerical results are shown in Figure 2. The log-log plot in Figure 2a shows the sublinear convergence of Riemannian subgradient-type methods when the diminishing stepsizes suggested in Theorem 2 and



(a) ODL: Diminishing stepsizes of the form $\gamma_k = 10^{-3}/\sqrt{k}$, $k = 1, 2, \ldots$



(b) ODL: Geometrically diminishing stepsizes of the form $\gamma_k = 0.1 \times \beta^k$, $k = 0, 1, \ldots$

Fig. 2: Convergence performance of Riemannian subgradient-type methods for the orthogonal DL problem (1.5).

Theorem 3 are used. Figure 2b shows the linear convergence of those methods when geometrically diminishing stepsizes of the form $\gamma_k = \beta^k \gamma_0$ are used. Here, $\gamma_0 = 0.1$ and the best decay factor β is chosen for each algorithm.

8. Conclusion. In this work, we introduced a family of Riemannian subgradient-type methods for minimizing weakly convex functions over the Stiefel manifold. We proved, for the first time, iteration complexity and local convergence rate results for these methods. Specifically, we showed that all these methods have a global sublinear convergence rate, and that if the problem at hand further possesses the sharpness property, then the Riemannian subgradient and incremental subgradient methods with geometrically diminishing stepsizes and a proper initialization will converge linearly to the set of weak sharp minima of the problem. The key to establishing these results is a new Riemannian subgradient inequality for restrictions of weakly convex functions on the Stiefel manifold, which could be of independent interest. Our results can be extended to cover weakly convex minimization over a class of compact embedded submanifolds of the Euclidean space. Lastly, we showed that certain formulations of the RSR and orthogonal DL problems possess the sharpness property and verified the convergence performance of the Riemannian subgradient-type methods on these problems via numerical simulations.

Our work has opened up several interesting directions for future investigation. First, one can readily generalize our results to weakly convex minimization over a Cartesian product of Stiefel manifolds, which has applications in ℓ_1-PCA [33, 58] and robust phase synchronization [57]. Next, since our results are specific to weakly convex minimization over the Stiefel manifold, it would be interesting to see if they can be extended to handle more general nonconvex nonsmooth functions over a broader class of Riemannian manifolds. We believe that this should be possible based on the analytic framework developed here. Finally, we suspect that the global convergence rate $\mathcal{O}(k^{-\frac{1}{4}})$ we established for the Riemannian subgradient-type methods is not tight. This is because the Riemannian proximal point method for solving problem (1.1) has a global convergence rate of $\mathcal{O}(k^{-\frac{1}{2}})$ [10], and in smooth optimization the gradient

descent method has the same global convergence rate as the proximal point method. Hence, it would be interesting to see if the global convergence rate established in this paper can be improved.

Acknowledgments. We would like to thank Dr. Huikang Liu for fruitful discussions. We also thank the Associate Editor and two anonymous reviewers for their detailed and helpful comments.

REFERENCES

[1] P.-A. Absil and S. Hosseini, *A collection of nonsmooth Riemannian optimization problems*, in Nonsmooth Optimization and Its Applications, Springer, 2019, pp. 1–15.
[2] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, 2009.
[3] Y. Bai, Q. Jiang, and J. Sun, *Subgradient descent learns orthogonal dictionaries*, in International Conference on Learning Representations, 2019.
[4] G. C. Bento, O. P. Ferreira, and J. G. Melo, *Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds*, Journal of Optimization Theory and Applications, 173 (2017), pp. 548–562.
[5] R. L. Bishop and B. O'Neill, *Manifolds of negative curvature*, Transactions of the American Mathematical Society, 145 (1969), pp. 1–49.
[6] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, 2013.
[7] N. Boumal, P.-A. Absil, and C. Cartis, *Global rates of convergence for nonconvex optimization on manifolds*, IMA Journal of Numerical Analysis, 39 (2019), pp. 1–33.
[8] J. V. Burke and M. C. Ferris, *Weak sharp minima in mathematical programming*, SIAM Journal on Control and Optimization, 31 (1993), pp. 1340–1359.
[9] J. V. Burke, A. S. Lewis, and M. L. Overton, *A robust gradient sampling algorithm for nonsmooth, nonconvex optimization*, SIAM Journal on Optimization, 15 (2005), pp. 751–779.
[10] S. Chen, Z. Deng, S. Ma, and A. M.-C. So, *Manifold proximal point algorithms for dual principal component pursuit and orthogonal dictionary learning*, arXiv preprint arXiv:2005.02356, (2020).
[11] S. Chen, S. Ma, A. M.-C. So, and T. Zhang, *Proximal gradient method for nonsmooth optimization over the Stiefel manifold*, SIAM Journal on Optimization, 30 (2020), pp. 210–239.
[12] D. Davis and D. Drusvyatskiy, *Stochastic model-based minimization of weakly convex functions*, SIAM Journal on Optimization, 29 (2019), pp. 207–239.
[13] D. Davis, D. Drusvyatskiy, K. J. MacPhee, and C. Paquette, *Subgradient methods for sharp weakly convex functions*, Journal of Optimization Theory and Applications, 179 (2018), pp. 962–982.
[14] G. de Carvalho Bento, J. X. da Cruz Neto, and P. R. Oliveira, *A new approach to the proximal point method: Convergence on general Riemannian manifolds*, Journal of Optimization Theory and Applications, 168 (2016), pp. 743–755.
[15] D. Drusvyatskiy, *The proximal point method revisited*, SIAG/OPT Views and News, 26 (2018), pp. 1–7.
[16] D. Drusvyatskiy and C. Paquette, *Efficiency of minimizing compositions of convex functions and smooth maps*, Mathematical Programming, 178 (2019), pp. 503–558.
[17] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM Journal on Matrix Analysis and Applications, 20 (1998), pp. 303–353.
[18] O. Ferreira, M. Louzeiro, and L. Prudente, *Iteration-complexity of the subgradient method on Riemannian manifolds with lower bounded curvature*, Optimization, 68 (2019), pp. 713–729.
[19] O. P. Ferreira and P. R. Oliveira, *Subgradient algorithm on Riemannian manifolds*, Journal of Optimization Theory and Applications, 97 (1998), pp. 93–104.
[20] O. P. Ferreira and P. R. Oliveira, *Proximal point algorithm on Riemannian manifold*, Optimization, 51 (2002), pp. 257–270.
[21] J.-L. Goffin, *On convergence rates of subgradient optimization methods*, Mathematical Programming, 13 (1977), pp. 329–347.
[22] S. Hosseini, W. Huang, and R. Yousefpour, *Line search algorithms for locally Lipschitz functions on Riemannian manifolds*, SIAM Journal on Optimization, 28 (2018), pp. 596–

619.
[23] S. Hosseini and A. Uschmajew, *A Riemannian gradient sampling algorithm for nonsmooth optimization on manifolds*, SIAM Journal on Optimization., 27 (2017), pp. 173–189.
[24] J. Hu, X. Liu, Z.-W. Wen, and Y.-X. Yuan, *A brief introduction to manifold optimization*, Journal of the Operations Research Society of China, 8 (2020), pp. 199–248.
[25] W. Huang and K. Wei, *Riemannian proximal gradient methods*, arXiv preprint arXiv:1909.06065, (2019).
[26] B. Jiang, S. Ma, A. M.-C. So, and S. Zhang, *Vector transport-free SVRG with general retraction for Riemannian optimization: Complexity analysis and practical implementation*, arXiv preprint arXiv:1705.09059, (2017).
[27] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, *Generalized power method for sparse principal component analysis*, Journal of Machine Learning Research, 11 (2010), pp. 517–553.
[28] M. M. Karkhaneei and N. Mahdavi-Amiri, *Nonconvex weak sharp minima on Riemannian manifolds*, Journal of Optimization Theory and Applications, 183 (2019), pp. 85–104.
[29] A. V. Knyazev and M. E. Argentati, *Majorization for changes in angles between subspaces, Ritz values, and graph Laplacian spectra*, SIAM Journal on Matrix Analysis and Applications, 29 (2007), pp. 15–32.
[30] A. Kovnatsky, K. Glashoff, and M. M. Bronstein, *MADMM: a generic algorithm for non-smooth optimization on manifolds*, in European Conference on Computer Vision, Springer, 2016, pp. 680–696.
[31] R. Lai and S. Osher, *A splitting method for orthogonality constrained problems*, Journal of Scientific Computing, 58 (2014), pp. 431–449.
[32] G. Lerman and T. Maunu, *Fast, robust and non-convex subspace recovery*, Information and Inference: A Journal of the IMA, 7 (2018), pp. 277–336.
[33] G. Lerman and T. Maunu, *An overview of robust subspace recovery*, Proceedings of the IEEE, 106 (2018), pp. 1380–1410.
[34] G. Lerman, M. B. McCoy, J. A. Tropp, and T. Zhang, *Robust computation of linear models by convex relaxation*, Foundations of Computational Mathematics, 15 (2015), pp. 363–410.
[35] C. Li, B. S. Mordukhovich, J. Wang, and J.-C. Yao, *Weak sharp minima on Riemannian manifolds*, SIAM Journal on Optimization, 21 (2011), pp. 1523–1560.
[36] X. Li, Z. Zhu, A. M.-C. So, and J. D. Lee, *Incremental methods for weakly convex optimization*, arXiv preprint arXiv:1907.11687, (2019).
[37] X. Li, Z. Zhu, A. M.-C. So, and R. Vidal, *Nonconvex robust low-rank matrix recovery*, SIAM Journal on Optimization, 30 (2020), pp. 660–686.
[38] 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, 178 (2019), pp. 215–262.
[39] J. Mairal, F. Bach, and J. Ponce, *Sparse modeling for image and vision processing*, Foundations and Trends® in Computer Graphics and Vision, 8 (2014), pp. 85–283.
[40] T. Maunu, T. Zhang, and G. Lerman, *A well-tempered landscape for non-convex robust subspace recovery.*, Journal of Machine Learning Research, 20 (2019), pp. 1–59.
[41] A. Maurer, *A vector-contraction inequality for Rademacher complexities*, in Proceedings of the 27th International Conference on Algorithmic Learning Theory (ALT 2016), R. Ortner, H. U. Simon, and S. Zilles, eds., vol. 9925 of Lecture Notes in Artificial Intelligence, 2016, pp. 3–17.
[42] A. Nedić and D. Bertsekas, *Convergence rate of incremental subgradient algorithms*, in Stochastic Optimization: Algorithms and Applications, S. Uryasev and P. M. Pardalos, eds., vol. 54 of Applied Optimization, Springer Science+Business Media, Dordrecht, 2001, pp. 223–264.
[43] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal on Optimization, 19 (2009), pp. 1574–1609.
[44] Yu. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publishers, Boston, 2004.
[45] Q. Qu, J. Sun, and J. Wright, *Finding a sparse vector in a subspace: Linear sparsity using alternating directions*, IEEE Transactions on Information Theory, 62 (2016), pp. 5855–5880.
[46] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, vol. 317 of Grundlehren der mathematischen Wissenschaften, Springer Science & Business Media, second ed., 2009.
[47] R. Rubinstein, A. M. Bruckstein, and M. Elad, *Dictionaries for sparse representation*

modeling, Proceedings of the IEEE, 98 (2010), pp. 1045–1057.
[48] P. H. Schönemann, *A generalized solution of the orthogonal procrustes problem*, Psychometrika, 31 (1966), pp. 1–10.
[49] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, vol. 3 of Springer Series in Computational Mathematics, Springer–Verlag, Berlin Heidelberg, 1985.
[50] D. A. Spielman, H. Wang, and J. Wright, *Exact recovery of sparsely-used dictionaries*, in Proceedings of the 25th Annual Conference on Learning Theory, 2012, pp. 37.1–37.18.
[51] G. W. Stewart and J. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.
[52] J. Sun, Q. Qu, and J. Wright, *Complete dictionary recovery over the sphere I: Overview and the geometric picture*, IEEE Transactions on Information Theory, 63 (2016), pp. 853–884.
[53] J. Sun, Q. Qu, and J. Wright, *Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method*, IEEE Transactions on Information Theory, 63 (2016), pp. 885–914.
[54] M. C. Tsakiris and R. Vidal, *Dual principal component pursuit*, Journal of Machine Learning Research, 19 (2018), pp. 1–49.
[55] J.-P. Vial, *Strong and weak convexity of sets and functions*, Mathematics of Operations Research, 8 (1983), pp. 231–259.
[56] R. Vidal, Y. Ma, and S. S. Sastry, *Generalized Principal Component Analysis*, vol. 40 of Interdisciplinary Applied Mathematics, Springer-Verlag, New York, 2016.
[57] L. Wang and A. Singer, *Exact and stable recovery of rotations for robust synchronization*, Information and Inference: A Journal of the IMA, 2 (2013), pp. 145–193.
[58] P. Wang, H. Liu, and A. M.-C. So, *Globally convergent accelerated proximal alternating maximization method for L1–principal component analysis*, in Proceedings of the 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2019), 2019, pp. 8147–8151.
[59] Y. Wang, S. Wu, and B. Yu, *Unique sharp local minimum in ℓ_1-minimization complete dictionary learning*, Journal of Machine Learning Research, 21 (2020), pp. 1–52.
[60] Z. Wen and W. Yin, *A feasible method for optimization with orthogonality constraints*, Mathematical Programming, 142 (2013), pp. 397–434.
[61] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, *Sparse representation for computer vision and pattern recognition*, Proceedings of the IEEE, 98 (2010), pp. 1031–1044.
[62] H. Xu, C. Caramanis, and S. Mannor, *Outlier-robust PCA: The high-dimensional case*, IEEE Transactions on Information Theory, 59 (2012), pp. 546–572.
[63] W. H. Yang, L.-H. Zhang, and R. Song, *Optimality conditions for the nonlinear programming problems on Riemannian manifolds*, Pacific Journal of Optimization, 10 (2014), pp. 415–434.
[64] S.-T. Yau, *Non-existence of continuous convex functions on certain Riemannian manifolds*, Mathematische Annalen, 207 (1974), pp. 269–270.
[65] Y. Zhai, Z. Yang, Z. Liao, J. Wright, and Y. Ma, *Complete dictionary learning via ℓ^4-norm maximization over the orthogonal group*, Journal of Machine Learning Research, 21 (2020), pp. 1–68.
[66] H. Zhang and S. Sra, *First-order methods for geodesically convex optimization*, in Proceedings of the 29th Annual Conference on Learning Theory, 2016, pp. 1617–1638.
[67] T. Zhang, *Robust subspace recovery by Tyler's M-estimator*, Information and Inference: A Journal of the IMA, 5 (2016), pp. 1–21.
[68] T. Zhang and G. Lerman, *A novel M-estimator for robust PCA*, Journal of Machine Learning Research, 15 (2014), pp. 749–808.
[69] Z. Zhu, T. Ding, M. Tsakiris, D. Robinson, and R. Vidal, *A linearly convergent method for non-smooth non-convex optimization on Grassmannian with applications to robust subspace and dictionary learning*, in Advances in Neural Information Processing Systems, 2019, pp. 9437–9447.
[70] Z. Zhu, Y. Wang, D. Robinson, D. Naiman, R. Vidal, and M. Tsakiris, *Dual principal component pursuit: Improved analysis and efficient algorithms*, in Advances in Neural Information Processing Systems, 2018, pp. 2171–2181.

Appendix A. Proof of Lemma 2.

The proof follows the framework in [34, Section 8.1.1] with nontrivial modifications in order to handle our matrix-based definitions of $c_{\boldsymbol{Y},\min}$ and $c_{\boldsymbol{O},\max}$.

Part I. We first derive an upper bound on $c_{\boldsymbol{O},\max}$. Recall that under the Haystack model, the outliers $\boldsymbol{o}_1, \ldots, \boldsymbol{o}_{m_2} \in \mathbb{R}^n$ are i.i.d. according to the Gaussian distribution

$\mathcal{N}(\mathbf{0}, \frac{1}{n}\boldsymbol{I}_n)$. Let $\boldsymbol{o} \sim \mathcal{N}(\mathbf{0}, \frac{1}{n}\boldsymbol{I}_n)$ denote an i.i.d. copy of $\boldsymbol{o}_i$. Then, we have

$$\begin{aligned} &\sup_{\|\boldsymbol{B}\|_F=1} \sum_{i=1}^{m_2} \|\boldsymbol{o}_i^\top \boldsymbol{B}\|_2 \\ &\leq \sup_{\|\boldsymbol{B}\|_F=1} \sum_{i=1}^{m_2} \left(\|\boldsymbol{o}_i^\top \boldsymbol{B}\|_2 - \mathrm{E}\left[\|\boldsymbol{o}^\top \boldsymbol{B}\|_2\right]\right) + \sup_{\|\boldsymbol{B}\|_F=1} \sum_{i=1}^{m_2} \mathrm{E}\left[\|\boldsymbol{o}^\top \boldsymbol{B}\|_2\right]. \end{aligned} \tag{A.1}$$

Using Jensen's inequality, we bound the second term as follows:

$$\begin{aligned} &\sup_{\|\boldsymbol{B}\|_F=1} \sum_{i=1}^{m_2} \mathrm{E}\left[\|\boldsymbol{o}^\top \boldsymbol{B}\|_2\right] \leq \sup_{\|\boldsymbol{B}\|_F=1} \sum_{i=1}^{m_2} \sqrt{\mathrm{E}\left[\|\boldsymbol{o}^\top \boldsymbol{B}\|_2^2\right]} \\ &= \sup_{\|\boldsymbol{B}\|_F=1} \sum_{i=1}^{m_2} \sqrt{\sum_{j=1}^{r}\sum_{k=1}^{n} \mathrm{E}\left[o_k^2 b_{kj}^2\right]} = \frac{m_2}{\sqrt{n}}. \end{aligned} \tag{A.2}$$

To estimate the first term in (A.1), let $\{\epsilon_i : i = 1, \ldots, m_2\}$ be independent Rademacher random variables (i.e., $\Pr[\epsilon_i = +1] = \Pr[\epsilon_i = -1] = 1/2$ for $i = 1, \ldots, m_2$) that are independent of $\{\boldsymbol{o}_i : i = 1, \ldots, m_2\}$. By a standard symmetrization argument (see, e.g., [6, Lemma 11.4]), we have

$$\mathrm{E}\left[\sup_{\|\boldsymbol{B}\|_F=1} \sum_{i=1}^{m_2} \left(\|\boldsymbol{o}_i^\top \boldsymbol{B}\|_2 - \mathrm{E}\left[\|\boldsymbol{o}^\top \boldsymbol{B}\|_2\right]\right)\right] \leq 2\,\mathrm{E}\left[\sup_{\|\boldsymbol{B}\|_F=1} \sum_{i=1}^{m_2} \epsilon_i \|\boldsymbol{o}_i^\top \boldsymbol{B}\|_2\right]. \tag{A.3}$$

Furthermore, let $\{\epsilon_{ij} : i = 1, \ldots, m_2;\, j = 1, \ldots, r\}$ be independent Rademacher random variables that are independent of $\{\boldsymbol{o}_i : i = 1, \ldots, m_2\}$ and $\boldsymbol{V} \in \mathbb{R}^{n \times r}$ be the matrix whose j-th column ($j = 1, \ldots, r$) is $\sum_{i=1}^{m_2} \epsilon_{ij} \boldsymbol{o}_i$. Then, by the vector contraction inequality in [41, Corollary 1] and Jensen's inequality, we have

$$\begin{aligned} &\mathrm{E}\left[\sup_{\|\boldsymbol{B}\|_F=1} \sum_{i=1}^{m_2} \epsilon_i \|\boldsymbol{o}_i^\top \boldsymbol{B}\|_2\right] \leq \sqrt{2}\,\mathrm{E}\left[\sup_{\|\boldsymbol{B}\|_F=1} \sum_{i=1}^{m_2}\sum_{j=1}^{r} \epsilon_{ij} \boldsymbol{o}_i^\top \boldsymbol{b}_j\right] \\ &= \sqrt{2}\,\mathrm{E}\left[\sup_{\|\boldsymbol{B}\|_F=1} \sum_{j=1}^{r} \left(\sum_{i=1}^{m_2} \epsilon_{ij} \boldsymbol{o}_i\right)^\top \boldsymbol{b}_j\right] = \sqrt{2}\,\mathrm{E}\left[\sup_{\|\boldsymbol{B}\|_F=1} \langle \boldsymbol{V}, \boldsymbol{B}\rangle\right] = \sqrt{2}\,\mathrm{E}\left[\|\boldsymbol{V}\|_F\right] \\ &\leq \sqrt{2}\sqrt{\sum_{j=1}^{r} \mathrm{E}\left[\left\|\sum_{i=1}^{m_2} \epsilon_{ij} \boldsymbol{o}_i\right\|_2^2\right]} \leq \sqrt{2m_2 r}. \end{aligned}$$

This, together with (A.3), yields

$$\mathrm{E}\left[\sup_{\|\boldsymbol{B}\|_F=1} \sum_{i=1}^{m_2} \left(\|\boldsymbol{o}_i^\top \boldsymbol{B}\|_2 - \mathrm{E}\left[\|\boldsymbol{o}^\top \boldsymbol{B}\|_2\right]\right)\right] \leq 2\sqrt{2m_2 r}. \tag{A.4}$$

Now, observe that the function

$$(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_{m_2}) \mapsto h(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_{m_2}) := \sup_{\|\boldsymbol{B}\|_F=1} \sum_{i=1}^{m_2} \left(\|\boldsymbol{o}_i^\top \boldsymbol{B}\|_2 - \mathrm{E}\left[\|\boldsymbol{o}^\top \boldsymbol{B}\|_2\right]\right)$$

is Lipschitz continuous with constant at most $\sqrt{m_2}$. Hence, using the Gaussian concentration inequality for Lipschitz functions [6, Theorem 5.6] and (A.4), we get

$$\Pr\left[\sup_{\|\boldsymbol{B}\|_F=1}\sum_{i=1}^{m_2}\left(\|\boldsymbol{o}_i^\top\boldsymbol{B}\|_2-\mathrm{E}\left[\|\boldsymbol{o}^\top\boldsymbol{B}\|_2\right]\right)\le 2\sqrt{2m_2 r}+t\right]\ge 1-2\exp\left(-\frac{nt^2}{2m_2}\right). \tag{A.5}$$

Upon substituting (A.5) and (A.2) into (A.1) and letting $t=c_2\sqrt{m_2}$, the desired result follows.

Part II. We now derive a lower bound on $c_{\boldsymbol{Y},\min}$. Again, recall that under the Haystack model, the inliers $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{m_1}\in\mathbb{R}^n$ are i.i.d. according to the Gaussian distribution $\mathcal{N}(\boldsymbol{0},\frac{1}{d}\mathcal{P}_{\mathcal{S}})$. Thus, for $i=1,\ldots,m_1$, we have $\mathrm{E}[\boldsymbol{y}_i\boldsymbol{y}_i^\top]=\frac{1}{d}\mathcal{P}_{\mathcal{S}}=\frac{1}{d}\boldsymbol{S}\boldsymbol{S}^\top$ for some orthonormal basis $\boldsymbol{S}\in\mathrm{St}(n,d)$ of $\mathcal{S}$ and $\boldsymbol{y}_i=\boldsymbol{S}\widetilde{\boldsymbol{y}}_i$ for some $\widetilde{\boldsymbol{y}}_i\in\mathbb{R}^d$. Now, let $\boldsymbol{D}\in\mathbb{R}^{n\times\ell}$ be such that $\|\boldsymbol{D}\|_F=1$ and $\mathrm{col}(\boldsymbol{D})\subseteq\mathcal{S}$; see (7.1). Then, there exists a $\widetilde{\boldsymbol{D}}\in\mathbb{R}^{d\times\ell}$ such that $\boldsymbol{D}=\boldsymbol{S}\widetilde{\boldsymbol{D}}$ and $\|\widetilde{\boldsymbol{D}}\|_F=1$. In particular, we have $\boldsymbol{y}_i^\top\boldsymbol{D}=\widetilde{\boldsymbol{y}}_i^\top\widetilde{\boldsymbol{D}}$, and by the rotational invariance of the Gaussian distribution, the vector $\widetilde{\boldsymbol{y}}_i$ follows the Gaussian distribution $\mathcal{N}(\boldsymbol{0},\frac{1}{d}\boldsymbol{I}_d)$ in $\mathbb{R}^d$. Consequently, we may assume without loss of generality that $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{m_1}\in\mathbb{R}^d$ are i.i.d. according to the Gaussian distribution $\mathcal{N}(\boldsymbol{0},\frac{1}{d}\boldsymbol{I}_d)$ and $\boldsymbol{D}\in\mathbb{R}^{d\times\ell}$ satisfies $\|\boldsymbol{D}\|_F=1$. The rest of the proof will be similar to that of Part I.

Let $\boldsymbol{y}\sim\mathcal{N}(\boldsymbol{0},\frac{1}{d}\boldsymbol{I}_d)$ denote an i.i.d. copy of $\boldsymbol{y}_i$. Then, we have

$$\begin{aligned}&\inf_{\|\boldsymbol{D}\|_F=1}\sum_{i=1}^{m_1}\|\boldsymbol{y}_i^\top\boldsymbol{D}\|_2\\ &\ge\inf_{\|\boldsymbol{D}\|_F=1}\sum_{i=1}^{m_1}\left(\|\boldsymbol{y}_i^\top\boldsymbol{D}\|_2-\mathrm{E}\left[\|\boldsymbol{y}^\top\boldsymbol{D}\|_2\right]\right)+\inf_{\|\boldsymbol{D}\|_F=1}\sum_{i=1}^{m_1}\mathrm{E}\left[\|\boldsymbol{y}^\top\boldsymbol{D}\|_2\right].\end{aligned} \tag{A.6}$$

The first term can be written as $-\sup_{\|\boldsymbol{D}\|_F=1}\sum_{i=1}^{m_1}\left(\mathrm{E}\left[\|\boldsymbol{y}^\top\boldsymbol{D}\|_2\right]-\|\boldsymbol{y}_i^\top\boldsymbol{D}\|_2\right)$. By following the same arguments as in Part I, we obtain

$$\Pr\left[\sup_{\|\boldsymbol{D}\|_F=1}\sum_{i=1}^{m_1}\left(\mathrm{E}\left[\|\boldsymbol{y}^\top\boldsymbol{D}\|_2\right]-\|\boldsymbol{y}_i^\top\boldsymbol{D}\|_2\right)\le 2\sqrt{2m_1\ell}+t\right]\ge 1-2\exp\left(-\frac{dt^2}{2m_1}\right). \tag{A.7}$$

It remains to estimate the second term in (A.6). Let $\boldsymbol{d}_i$ be the i-th column of $\boldsymbol{D}$, where $i=1,\ldots,\ell$. By the Cauchy-Schwarz inequality and the fact that $\|\boldsymbol{D}\|_F=1$, we have

$$\begin{aligned}&\underbrace{\left(\|\boldsymbol{d}_1\|_2^2+\cdots+\|\boldsymbol{d}_\ell\|_2^2\right)}_{=\|\boldsymbol{D}\|_F^2=1}\underbrace{\left[(\boldsymbol{y}^\top\boldsymbol{d}_1)^2+\cdots+(\boldsymbol{y}^\top\boldsymbol{d}_\ell)^2\right]}_{=\|\boldsymbol{y}^\top\boldsymbol{D}\|_2^2}\\ &\ge\left[\|\boldsymbol{d}_1\|_2|\boldsymbol{y}^\top\boldsymbol{d}_1|+\cdots+\|\boldsymbol{d}_\ell\|_2|\boldsymbol{y}^\top\boldsymbol{d}_\ell|\right]^2.\end{aligned}$$

Since $\boldsymbol{y}^\top\boldsymbol{d}_i\sim\mathcal{N}(0,\frac{1}{d}\|\boldsymbol{d}_i\|_2^2)$, we obtain

$$\mathrm{E}\left[\|\boldsymbol{y}^\top\boldsymbol{D}\|_2\right]\ge\mathrm{E}\left[\|\boldsymbol{d}_1\|_2|\boldsymbol{y}^\top\boldsymbol{d}_1|+\cdots+\|\boldsymbol{d}_\ell\|_2|\boldsymbol{y}^\top\boldsymbol{d}_\ell|\right]=\sqrt{\frac{2}{d\pi}}.$$

Note that the above inequality holds as equality when $\boldsymbol{d}_1=\cdots=\boldsymbol{d}_\ell$. This implies that

$$\inf_{\|\boldsymbol{D}\|_F=1}\sum_{i=1}^{m_1}\mathrm{E}\left[\|\boldsymbol{y}^\top\boldsymbol{D}\|_2\right]=m_1\sqrt{\frac{2}{d\pi}}. \tag{A.8}$$

By substituting (A.7) with $t=c_1\sqrt{m_1}$ and (A.8) into (A.6), we complete the proof.