Randomized Submanifold Subgradient Method for Optimization Over the Stiefel Manifold

Andy Yat-ming Cheung* Jinxin Wang† Man-Chung Yue‡ Anthony Man-Cho So§

May 19, 2025

Abstract

Optimization over the Stiefel manifold is a fundamental computational problem in many scientific and engineering applications. Despite considerable research effort, high-dimensional optimization problems over the Stiefel manifold remain challenging, particularly when the objective function is nonsmooth. In this paper, we propose a novel coordinate-type algorithm, named *randomized submanifold subgradient method* (RSSM), for minimizing a possibly nonsmooth weakly convex function over the Stiefel manifold and study its convergence behavior. Similar to coordinate-type algorithms in the Euclidean setting, RSSM exhibits low per-iteration cost and is suitable for high-dimensional problems. We prove that RSSM has an iteration complexity of $\mathcal{O}(\varepsilon^{-4})$ for driving a natural stationarity measure below ε, both in expectation and in almost-sure senses. To the best of our knowledge, this is the first convergence guarantee for coordinate-type algorithms for nonsmooth optimization over the Stiefel manifold. To establish the said guarantee, we develop two new theoretical tools, namely a Riemannian subgradient inequality for weakly convex functions on proximally smooth matrix manifolds and an averaging operator that induces an adaptive metric on the ambient Euclidean space, which could be of independent interest. Lastly, we present numerical results on robust subspace recovery and orthogonal dictionary learning to demonstrate the viability of our proposed method.

1 Introduction

This paper focuses on the optimization problem

$$\begin{array}{ll} \min & f(X) \\ \text{s.t.} & X \in \mathrm{St}(n,p), \end{array} \tag{1}$$

where $n,p \geq 1$ are positive integers with $n \geq p$, $\mathrm{St}(n,p) = \{X \in \mathbb{R}^{n\times p} : X^\top X = I\}$ is the Stiefel manifold, and $f : \mathbb{R}^{n\times p} \to \mathbb{R}$ is a function that is τ-weakly convex on some bounded convex neighborhood $\Omega \subseteq \mathbb{R}^{n\times p}$ containing the Stiefel manifold $\mathrm{St}(n,p)$ (i.e., the function $f(\cdot) + \frac{\tau}{2}\|\cdot\|^2$ is convex on Ω) for some $\tau \in \mathbb{R}$. Note that f is not necessarily smooth. Throughout the paper, unless otherwise specified, we assume that $\tau \geq 0$.

*Department of Applied Mathematics, The Hong Kong Polytechnic University (andy.cheung@polyu.edu.hk); Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong (andycheungym.maths@gmail.com).

†Booth School of Business, The University of Chicago (jinxin.wang@chicagobooth.edu).

‡Department of Data and Systems Engineering and Musketeers Foundation Institute of Data Science, The University of Hong Kong (mcyue@hku.hk).

§Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong (manchoso@se.cuhk.edu.hk).

Problem (1) has attracted much attention from both optimization and machine learning communities due to its broad range of applications, including the Procrustes problem [GD04, SI13], the joint diagonalization problem [TCA09], Kohn-Sham total energy minimization [LWW+15], robust subspace recovery [ZWR+18], fair PCA [VTYN22], and orthogonal dictionary learning [LCD+21]. For recent algorithmic developments and applications of optimization over the Stiefel manifold, we refer the reader to [AMS09, Bou23, CMSZ24] and the references therein.

In modern applications, the dimension n or p of the Stiefel manifold could be very large. For example, as pointed out in [LJW+19], deep neural networks attain an optimal generalization error when the weight matrix has orthonormal columns or rows. Constraining the columns or rows of the weight matrix to be orthonormal in the training of deep neural networks naturally gives rise to high-dimensional optimization problems over the Stiefel manifold. Unfortunately, despite considerable research effort, existing algorithms for optimization over the Stiefel manifold do not scale well and are only suitable for small- to medium-scale problems. When the problem has a nonsmooth objective function, it becomes even more challenging. The main goal of this paper is to address these challenges by developing an efficient algorithm with convergence guarantees for nonsmooth optimization over a high-dimensional Stiefel manifold.

In the Euclidean setting, coordinate-type algorithms constitute a classical approach to tackling high-dimensional optimization problems and have shown promising performance in many applications [Nes12, Wri15]. A natural idea for achieving our goal is therefore to extend coordinate-type algorithms to the manifold setting. This idea has been explored in a number of prior works, though they focus exclusively on smooth objective functions. Roughly speaking, the approaches in existing works can be divided into three groups.

The first applies to manifolds that have a simple separable structure (e.g., a product manifold). Such a structure allows the development of coordinate-type algorithms that use simple operations to maintain the feasibility of the iterates; see, e.g., [HML21a, HML21b, PV23, TKRH21]. However, as the Stiefel manifold does not have an obvious separable structure, it is difficult to directly apply the approaches in this group to tackle optimization problems over the Stiefel manifold.

The second involves performing coordinate decomposition in the tangent spaces to the manifold. In [GHN22], a general framework, called tangent subspace descent, is proposed for developing coordinate-type algorithms for manifold optimization. At each iteration, the tangent space to the manifold at the current iterate is decomposed into low-dimensional subspaces, and the next iterate is obtained by updating along a chosen subspace using the exponential map. The crux of this framework is the choice of the low-dimensional subspaces of the tangent space. As an instantiation of the framework, the work [GHN22] presents a coordinate-type algorithm for optimization over the Stiefel manifold, which modifies at most two columns per iteration. Such a column-wise update scheme can be seen as a generalization of the algorithms in [JZQ+22, MA22, SC14] for optimization over the orthogonal manifold, which are derived based on Givens rotations. To further reduce the per-iteration computational cost, a row-wise update scheme is devised in [HJM24] by another choice of tangent subspaces. The approach in [HJM24] is also suitable for other manifolds, including the Grassmannian manifold. However, all these methods either rely on the exponential map or modify at most two rows or columns. It is unclear whether coordinate-type algorithms that update more columns or rows at each iteration and/or use a low-cost retraction instead of an exponential map can be developed. Using a similar idea as in [GHN22] with a Cholesky factorization-friendly basis for the tangent space, the work [DR23] proposes a coordinate-type algorithm for optimization over the manifold of positive definite matrices. Although each iteration of this algorithm can update more than two columns or rows and avoid matrix exponentiation (which is required for the exponential map of the manifold of positive definite matrices), the efficiency of the update is guaranteed only when the objective function is smooth and takes a certain composite form (see [DR23, Eq.(2)]).

Table 1: Coordinate-type methods for optimization over the Stiefel manifold.

Methods	Coordinate	Objective	Update	Feasible	Retraction
[GHN22] (TSD)	intrinsic	Lipschitz Riemannian gradient	modify at most two columns	✔	exponential map
[HJM24]	intrinsic	Lipschitz Riemannian gradient	modify two rows	✔	exponential map
[GLY19] (PCAL)	ambient	C^2 with bounded Euclidean Hessian	modify multiple columns	✗	N/A
[HPT25] (RSDM)	intrinsic	bounded Euclidean gradient and Hessian	left orthogonal action[1]	✔	retraction
Our work (RSSM)	ambient	weakly convex, nonsmooth	modify multiple columns	✔	polar decomposition (4)

The third consists of penalty approaches, which transform the original manifold optimization problem into an unconstrained one by introducing a penalty term for constraint violation in the objective function. The work [GLY19] studies a penalty approach and develops coordinate-type algorithms for optimization over the Stiefel manifold. Nevertheless, the approach in [GLY19] leads to infeasible-point algorithms, which might not be desirable in some applications.

The above discussion motivates the search for a coordinate-type algorithm that can tackle nonsmooth optimization problems over the Stiefel manifold and possesses more favorable properties than those in the aforementioned works. Our contributions can be summarized as follows.

- We devise a new coordinate-type algorithm, called the *randomized submanifold subgradient method* (RSSM), for minimizing a possibly nonsmooth weakly convex function over the Stiefel manifold (i.e., Problem (1)). A major novelty of RSSM lies in the way the "coordinate blocks" are defined: Instead of performing coordinate updates with respect to the intrinsic coordinates by decomposing the tangent spaces to the manifold as in [GHN22, HJM24], RSSM performs coordinate updates with respect to the ambient Euclidean coordinates by decomposing the Stiefel manifold into submanifold blocks and taking a retracted partial Riemannian subgradient step along a randomly chosen submanifold block.

- We show that the said submanifold blocks are 1-proximally smooth and that the corresponding projections can be computed efficiently in closed form. RSSM therefore exhibits a low per-iteration computational cost and is particularly suitable for high-dimensional instances of Problem (1).

- We prove that RSSM has an iteration complexity of $\mathcal{O}(\varepsilon^{-4})$ for driving a natural stationarity measure of Problem (1) below $\varepsilon > 0$, both in expectation and in almost-sure senses.

[1]Equivalent to the addition of a low rank matrix; see [HPT25, Section 4].

- To facilitate our analysis, we develop two theoretical tools. The first is a new Riemannian subgradient inequality for weakly convex functions on proximally smooth matrix manifolds, which could be of independent interest in other matrix optimization studies. The second is a novel theoretical construct called the *averaging operator*, which is a positive definite operator on $\mathbb{R}^{n\times p}$ and permeates the proofs via the adaptive metric it induces on the ambient Euclidean space. Remarkably, these two tools together allow us to prove the sufficient decrease of our proposed method.

Before leaving this section, let us mention the recent paper [HPT25], which presents another submanifold-based coordinate-type algorithm called RSDM for optimization over the Stiefel manifold and appears around the same time as the preliminary version of this paper [CWYS24]. Unlike RSSM, which operates on submanifolds that are defined in the ambient Euclidean coordinates and can handle nonsmooth objective functions, RSDM operates on submanifolds that are defined in intrinsic coordinates and applies only to smooth objective functions. For a comparison of various coordinate-type algorithms for optimization over the Stiefel manifold, see Table 1.

2 Notation and Preliminaries

Let $\xi, \eta \in \mathbb{R}^{n\times p}$ be arbitrary. The *Frobenius inner product* between ξ and η is denoted by $\langle \xi, \eta \rangle = \operatorname{tr}(\xi^\top \eta)$, while the *Frobenius norm* of ξ is denoted by $\|\xi\| = \sqrt{\langle \xi, \xi \rangle}$. The *operator norm* and *nuclear norm* of ξ are denoted by $\|\xi\|_{\mathsf{op}} = \sup_{v\in\mathbb{R}^p,\, \|v\|=1} \|\xi v\|$ and $\|\xi\|_* = \operatorname{tr}((\xi^\top\xi)^{1/2})$, respectively. Given a self-adjoint positive definite linear operator $\mathcal{D} : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$, the *Mahalanobis inner product* between ξ and η is denoted by $\langle \xi, \eta\rangle_{\mathcal{D}} = \langle \mathcal{D}(\xi), \eta\rangle$, while the *Mahalanobis norm* of ξ is denoted by $\|\xi\|_{\mathcal{D}} = \sqrt{\langle \mathcal{D}(\xi), \xi\rangle}$. When $n = p$, the *symmetric part* and *skew-symmetric part* of ξ are denoted by $\operatorname{Sym}(\xi) = \frac{1}{2}(\xi + \xi^\top)$ and $\operatorname{Skew}(\xi) = \frac{1}{2}(\xi - \xi^\top)$, respectively. The symbols I, J, and 0 represent the identity, all-one, and zero matrices, respectively. For a subset $C \subseteq [p] = \{1, \ldots, p\}$ and a matrix $X \in \mathbb{R}^{n\times p}$, the cardinality of C is denoted by $|C|$, and the submatrix obtained by extracting all columns of X corresponding to the indices in C is denoted by $X_C \in \mathbb{R}^{n\times|C|}$. For $\ell \geq 2$, the collection of 2-sets or unordered pairs in $[\ell]$ is denoted by $\binom{[\ell]}{2} = \{\, \{i,j\} : i,j \in [\ell] \text{ and } i \neq j \,\}$. Given a closed subset $\mathcal{S} \subseteq \mathbb{R}^{n\times p}$ and a point $X \in \mathbb{R}^{n\times p}$, the *distance* from $X \in \mathbb{R}^{n\times p}$ to $\mathcal{S}$ is denoted by $\operatorname{dist}(X, \mathcal{S}) = \inf_{Y\in\mathcal{S}} \|X - Y\|$, while the projection of X onto $\mathcal{S}$ is denoted by $\mathcal{P}_{\mathcal{S}}(X) = \{Y \in \mathcal{S} : \|X - Y\| = \operatorname{dist}(X, \mathcal{S})\}$.

2.1 Weakly Convex Optimization on Stiefel Manifold

Given a τ-weakly convex function $h : \mathbb{R}^{n\times p} \to \mathbb{R}$, which is necessarily Clarke regular [Via83, Proposition 4.5], we have $h(\cdot) = g(\cdot) - \frac{\tau}{2}\|\cdot\|^2$ for some convex function $g : \mathbb{R}^{n\times p} \to \mathbb{R}$. The Clarke subdifferential of h at $X \in \mathbb{R}^{n\times p}$ is given by $\partial h(X) = \partial g(X) - \tau X$, where $\partial g(X)$ is the usual convex subdifferential of g at $X \in \mathbb{R}^{n\times p}$; see [Via83, Proposition 4.6]. Moreover, we have

$$h(Y) \geq h(X) + \langle \widetilde{\nabla} h(X), Y - X \rangle - \tfrac{\tau}{2}\|Y - X\|^2 \tag{2}$$

for all $X, Y \in \mathbb{R}^{n\times p}$ and $\widetilde{\nabla} h(X) \in \partial h(X)$; see [Via83, Proposition 4.8].

Let $\mathcal{M} \subseteq \mathbb{R}^{n\times p}$ be a smooth matrix manifold and $T_X\mathcal{M}$ be the tangent space to $\mathcal{M}$ at a point $X \in \mathcal{M}$. For a weakly convex (and hence Clarke regular) function h in the ambient Euclidean space $\mathbb{R}^{n\times p}$, the Riemannian (Clarke) subdifferential of h at $X \in \mathcal{M}$ is given by

$$\partial_{\mathcal{M}} h(X) = \mathcal{P}_{T_X\mathcal{M}}(\partial h(X)) = \{\, \mathcal{P}_{T_X\mathcal{M}}(\widetilde{\nabla} h(X)) \,:\, \widetilde{\nabla} h(X) \in \partial h(X) \,\}; \tag{3}$$

see [YZS14, Theorem 5.1]. We call an $X \in \mathcal{M}$ that satisfies $0 \in \partial_{\mathcal{M}} h(X)$ a *stationary point* of h over $\mathcal{M}$. Given an Euclidean subgradient $\widetilde{\nabla} h(X) \in \partial h(X)$ of h at $X \in \mathcal{M}$, the corresponding Riemannian subgradient is denoted by $\widetilde{\nabla}_{\mathcal{M}} h(X) = \mathcal{P}_{T_X\mathcal{M}}(\widetilde{\nabla} h(X))$. In particular, for the Stiefel manifold, it follows from [AMS09, Example 3.6.2] that $T_X\mathrm{St}(n,p) = \{\, \xi \in \mathbb{R}^{n\times p} : \xi^\top X + X^\top \xi = 0 \,\}$ and $\mathcal{P}_{T_X\mathrm{St}(n,p)}(\xi) = \xi - X\mathrm{sym}(X^\top \xi)$.

Many iterative algorithms for optimization over the manifold $\mathcal{M}$ update its iterate X by first moving on the tangent space $T_X\mathcal{M}$ along a given direction and then mapping the resulting point back onto the manifold. The latter can be achieved using the exponential map (see, e.g., [AMS09, Section 5.4]), though it can be numerically challenging to compute. This motivates the use of a *retraction*, which is a locally smooth map from the tangent bundle of $\mathcal{M}$ to $\mathcal{M}$ that approximates the exponential map up to the first order; see [AM12, Definition 1]. One important example of a retraction on $\mathcal{M}$ is the *projective retraction*, which is given by $(X,\xi) \mapsto \mathcal{P}_{\mathcal{M}}(X+\xi)$ for $X \in \mathcal{M}$ and $\xi \in T_X\mathcal{M}$ [AM12, Proposition 5]. For the Stiefel manifold $\mathrm{St}(n,p)$, we focus on the polar decomposition-based retraction

$$\mathrm{Retr}_X(\xi) = (X+\xi)(I+\xi^\top\xi)^{-1/2} \tag{4}$$

for $X \in \mathrm{St}(n,p)$ and $\xi \in T_X\mathrm{St}(n,p)$. It is known that the polar decomposition-based retraction (4) coincides with the projective retraction on $\mathrm{St}(n,p)$, and the projection $\mathcal{P}_{\mathrm{St}(n,p)}$ satisfies the Lipschitz-like property

$$\| \,\mathrm{Retr}_X(\xi) - Y \,\| = \| \,\mathcal{P}_{\mathrm{St}(n,p)}(X+\xi) - \mathcal{P}_{\mathrm{St}(n,p)}(Y) \,\| \le \| \,X + \xi - Y \,\| \tag{5}$$

for all $X, Y \in \mathrm{St}(n,p)$ and $\xi \in T_X\mathrm{St}(n,p)$; see [LCD+21, Lemma 1]. Moreover, the map $\xi \mapsto \mathcal{P}_{\mathrm{St}(n,p)}(X+\xi)$ satisfies the second-order boundedness condition

$$\| \,\mathcal{P}_{\mathrm{St}(n,p)}(X+\xi) - X - \xi \,\| \le \|\xi\|^2 \tag{6}$$

for all $X \in \mathrm{St}(n,p)$ and $\xi \in T_X\mathrm{St}(n,p)$ with $\|\xi\| \le 1$; see [LSW19, Appendix E.1].

2.2 Proximal Smoothness

The following notion will be crucial as we deal with submanifolds of the Stiefel manifold in our subsequent development.

Definition 2.1 ([CSW95, Theorems 4.1 and 4.11])**.** *A closed set $\mathcal{S} \subseteq \mathbb{R}^{n\times p}$ is R-proximally smooth if the projection $\mathcal{P}_{\mathcal{S}}(X)$ is a singleton whenever* $\mathrm{dist}(X,\mathcal{S}) < R$.

It is known that the Stiefel manifold $\mathrm{St}(n,p)$ is 1-proximally smooth for any integers $n, p \ge 1$ with $n \ge p$; see, e.g., [AM12, Proposition 7]. Moreover, a smooth manifold $\mathcal{M}$ is R-proximally smooth if and only if

$$2R \,\langle\, \zeta, X' - X \,\rangle \le \|\zeta\| \cdot \|X' - X\|^2 \tag{7}$$

for all $X, X' \in \mathcal{M}$ and $\zeta \in N_X\mathcal{M}$, where $N_X\mathcal{M}$ denotes the *normal space* to $\mathcal{M}$ at X. This can be deduced, e.g., by combining the results in [ANT16, Theorem 2.2, Proposition 2.3] and [RW09, Example 6.8].

3 Randomized Submanifold Subgradient Method

In this section, we introduce RSSM, a coordinate-type method for solving Problem (1). Central to the development of RSSM is the notion of a *submanifold block*, which is a submanifold of the Stiefel manifold induced by a collection of orthonormal columns. Each iteration of RSSM performs a projected Riemannian subgradient step on the submanifold block induced by randomly selected columns of the current iterate. As we shall see, submanifold blocks enjoy many desirable properties, which facilitate the easy computation of the projected Riemannian subgradient step.

3.1 Submanifold Block

To develop a coordinate-type algorithm for optimization over the Stiefel manifold, an intuitive approach is to update a subset of columns at each iteration. One fundamental challenge of such an approach is to maintain the feasibility of an iterate after updating the columns. This can be tackled by restricting the update to the subspace that is orthogonal to the span of the unaltered columns. We are thus led to the following definition.

Definition 3.1 (Submanifold Block)**.** *Let $C \subseteq [p]$ be fixed. The* submanifold block *at* $X \in \mathrm{St}(n,p)$ *with respect to* C *is defined as*

$$\mathcal{M}_{X_{[p]\setminus C}} = \{\, Y \in \mathrm{St}(n,|C|) \,:\, X_{[p]\setminus C}^\top Y = 0 \,\}. \tag{8}$$

By definition, the submanifold block $\mathcal{M}_{X_{[p]\setminus C}}$ can be written as the intersection $\mathcal{M}_{X_{[p]\setminus C}} = \mathrm{St}(n,|C|) \cap \{Y \in \mathbb{R}^{n\times|C|} : X_{[p]\setminus C}^\top Y = 0\}$. The intersection is transversal, since $T_Z\mathrm{St}(n,|C|) + T_Z\{Y \in \mathbb{R}^{n\times|C|} : X_{[p]\setminus C}^\top Y = 0\} = \mathbb{R}^{n\times|C|}$ for any $Z \in \mathcal{M}_{X_{[p]\setminus C}}$. Therefore, by [GP10, p.30], the submanifold block $\mathcal{M}_{X_{[p]\setminus C}}$ is itself a smooth submanifold of the Stiefel manifold $\mathrm{St}(n,|C|)$. The tangent space to $\mathcal{M}_{X_{[p]\setminus C}}$ at $X_C \in \mathcal{M}_{X_{[p]\setminus C}}$ is

$$T_{X_C}\mathcal{M}_{X_{[p]\setminus C}} = \{\, \xi \in T_{X_C}\mathrm{St}(n,|C|) \,:\, X_{[p]\setminus C}^\top \xi = 0 \,\}. \tag{9}$$

As mentioned above, our proposed RSSM performs a projected Riemannian subgradient step on a submanifold block at each iteration. The following lemma guarantees that the projection is well-defined, despite the non-convexity of the submanifold block, and provides a closed-form formula for the projection. In particular, when the search direction lies in the tangent space to the submanifold block, the projection reduces to the projection onto a lower-dimensional Stiefel manifold, which is readily computable via (4).

Lemma 3.2 (Proximal Smoothness and Projection onto Submanifold Block)**.** *For any $C \subseteq [p]$ and $X \in \mathrm{St}(n,p)$, the submanifold block $\mathcal{M}_{X_{[p]\setminus C}}$ is 1-proximally smooth. In particular, for any $\Xi \in \mathbb{R}^{n\times|C|}$ with $\mathrm{dist}(\Xi, \mathcal{M}_{X_{[p]\setminus C}}) < 1$, we have*

$$\begin{aligned}\mathcal{P}_{\mathcal{M}_{X_{[p]\setminus C}}}(\Xi) &= \mathcal{P}_{\mathrm{St}(n,|C|)}\left(\left(I - X_{[p]\setminus C}X_{[p]\setminus C}^\top\right)\Xi\right)\\ &= \left(I - X_{[p]\setminus C}X_{[p]\setminus C}^\top\right)\Xi\left[\,\Xi^\top\left(I - X_{[p]\setminus C}X_{[p]\setminus C}^\top\right)\Xi\,\right]^{-\frac{1}{2}}.\end{aligned}$$

If in addition $X_{[p]\setminus C}^\top\Xi = 0$, then $\mathcal{P}_{\mathcal{M}_{X_{[p]\setminus C}}}(\Xi) = \mathcal{P}_{\mathrm{St}(n,|C|)}(\Xi)$.

Proof. Let $\Xi \in \mathbb{R}^{n\times|C|}$ be such that $\mathrm{dist}(\Xi, \mathcal{M}_{X_{[p]\setminus C}}) < 1$. Consider the decomposition $\Xi = X_{[p]\setminus C}X_{[p]\setminus C}^\top\Xi + (I - X_{[p]\setminus C}X_{[p]\setminus C}^\top)\Xi$. For any $U \in \mathcal{M}_{X_{[p]\setminus C}}$, we have

$$\|\Xi - U\|^2 = \left\|X_{[p]\setminus C}X_{[p]\setminus C}^\top\Xi + (I - X_{[p]\setminus C}X_{[p]\setminus C}^\top)\Xi - U\right\|^2$$

$$= \left\| X_{[p]\setminus C} X_{[p]\setminus C}^\top \Xi \right\|^2 + \left\| (I - X_{[p]\setminus C} X_{[p]\setminus C}^\top)\Xi - U \right\|^2.$$

Since $\mathrm{dist}(\Xi, \mathrm{St}(n, |C|)) \le \mathrm{dist}(\Xi, \mathcal{M}_{X_{[p]\setminus C}}) < 1$ and $\mathrm{St}(n, |C|)$ is 1-proximally smooth, we obtain

$$\begin{aligned} &\operatorname*{argmin}_{U \in \mathrm{St}(n,|C|)} \left\| (I - X_{[p]\setminus C} X_{[p]\setminus C}^\top)\Xi - U \right\|^2 = \mathcal{P}_{\mathrm{St}(n,|C|)} \left((I - X_{[p]\setminus C} X_{[p]\setminus C}^\top)\Xi \right) \\ &= (I - X_{[p]\setminus C} X_{[p]\setminus C}^\top)\Xi \left[\Xi^\top (I - X_{[p]\setminus C} X_{[p]\setminus C}^\top)\Xi \right]^{-1/2} \in \mathcal{M}_{X_{[p]\setminus C}} \subseteq \mathrm{St}(n, |C|). \end{aligned}$$

Here, the second equality follows from the fact that $\mathcal{P}_{\mathrm{St}(n,|C|)}(\widetilde{\Xi}) = \widetilde{\Xi}(\widetilde{\Xi}^\top \widetilde{\Xi})^{-1/2}$ when $\mathrm{dist}(\widetilde{\Xi}, \mathrm{St}(n, |C|)) < 1$ and that

$$\mathrm{dist}\left((I - X_{[p]\setminus C} X_{[p]\setminus C}^\top)\Xi, \mathrm{St}(n, |C|) \right) \le \mathrm{dist}\left((I - X_{[p]\setminus C} X_{[p]\setminus C}^\top)\Xi, \mathcal{M}_{X_{[p]\setminus C}} \right) \le \mathrm{dist}(\Xi, \mathcal{M}_{X_{[p]\setminus C}}) < 1.$$

Therefore, the projection $\mathcal{P}_{\mathcal{M}_{X_{[p]\setminus C}}}(\Xi)$ is uniquely given by

$$\begin{aligned} \mathcal{P}_{\mathcal{M}_{X_{[p]\setminus C}}}(\Xi) &= \operatorname*{argmin}_{U \in \mathcal{M}_{X_{[p]\setminus C}}} \|\Xi - U\|^2 = \operatorname*{argmin}_{U \in \mathcal{M}_{X_{[p]\setminus C}}} \left\| (I - X_{[p]\setminus C} X_{[p]\setminus C}^\top)\Xi - U \right\|^2 \\ &= (I - X_{[p]\setminus C} X_{[p]\setminus C}^\top)\Xi \left[\Xi^\top (I - X_{[p]\setminus C} X_{[p]\setminus C}^\top)\Xi \right]^{-1/2}. \end{aligned}$$

It follows that $\mathcal{M}_{X_{[p]\setminus C}}$ is 1-proximally smooth. Furthermore, if $X_{[p]\setminus C}^\top \Xi = 0$, then the above formula gives $\mathcal{P}_{\mathcal{M}_{X_{[p]\setminus C}}}(\Xi) = \Xi(\Xi^\top \Xi)^{-1/2} = \mathcal{P}_{\mathrm{St}(n,|C|)}(\Xi)$. □

3.2 Partial Riemannian Subgradient Oracle

Recall that the objective function f in Problem (1) is τ-weakly convex on some bounded convex neighborhood $\Omega \subseteq \mathbb{R}^{n\times p}$ containing $\mathrm{St}(n, p)$. By [Via83, Proposition 4.4], we know that f is L-Lipschitz continuous on Ω for some $L > 0$. To define the updates in our proposed RSSM, we need the subgradient of the restriction of f to a submanifold block. Towards that end, let $X \in \mathbb{R}^{n\times p}$ and $C \subseteq [p]$ be fixed. Consider the map $\Phi_{X,C} : \mathbb{R}^{n\times|C|} \to \mathbb{R}^{n\times p}$ given by $\Phi_{X,C}(Y) = Y I_C^\top + X_{[p]\setminus C} I_{[p]\setminus C}^\top$. We define the *partial Euclidean (Clarke) subdifferential* of f with respect to C at X as

$$\partial^C f(X) = \partial(f \circ \Phi_{X,C})(X_C) \subseteq \mathbb{R}^{n\times|C|}.$$

A *partial Euclidean (Clarke) subgradient* of f with respect to C at X is denoted by $\widetilde{\nabla}^C f(X) = \widetilde{\nabla}(f \circ \Phi_{X,C})(X_C) \in \partial^C f(X)$. The following proposition reveals the relationship between the partial and full Euclidean (Clarke) subdifferentials.

Proposition 3.3 (Relationship between Partial and Full Euclidean Subdifferentials)**.** *For any $X \in \mathbb{R}^{n\times p}$ and $C \subseteq [p]$, we have $\partial^C f(X) = \partial f(X) I_C$, i.e., for any $\widetilde{\nabla}^C f(X) \in \partial^C f(X)$, there exists $\widetilde{\nabla} f(X) \in \partial f(X)$ such that $\widetilde{\nabla}^C f(X) = \widetilde{\nabla} f(X) I_C$.*

Proof. The function f, being weakly convex, is Clarke regular [Via83, Proposition 4.5]. Hence, by the subdifferential chain rule [RW09, Theorem 10.6] and the fact that $\Phi_{X,C}(X_C) = X$, we have $\partial^C f(X) = \partial(f \circ \Phi_{X,C})(X_C) = \nabla\Phi_{X,C}(X_C)^* \partial f(X) = \partial f(X) I_C$, where $\nabla\Phi_{X,C}(X_C)^*$ is the adjoint of the Jacobian of $\Phi_{X,C}$ at X_C. □

The index set C induces the submanifold block $\mathcal{M}_{X_{[p]\setminus C}}$ in (8). We define the *partial Riemannian (Clarke) subdifferential* of f with respect to C at $X \in \mathrm{St}(n,p)$ as

$$\partial^C_{\mathcal{M}_{X_{[p]\setminus C}}} f(X) = \partial_{\mathcal{M}_{X_{[p]\setminus C}}} (f \circ \Phi_{X,C})(X_C).$$

A *partial Riemannian (Clarke) subgradient* of f with respect to C at X is denoted by

$$\widetilde{\nabla}^C_{\mathcal{M}_{X_{[p]\setminus C}}} f(X) = \widetilde{\nabla}_{\mathcal{M}_{X_{[p]\setminus C}}} (f \circ \Phi_{X,C})(X_C) = \mathcal{P}_{T_{X_C}\mathcal{M}_{X_{[p]\setminus C}}} \left(\widetilde{\nabla}^C f(X) \right) \in \partial^C_{\mathcal{M}_{X_{[p]\setminus C}}} f(X), \tag{10}$$

where $\widetilde{\nabla}^C f(X) \in \partial^C f(X)$. It can be verified that $\mathcal{P}_{T_{X_C}\mathcal{M}_{X_{[p]\setminus C}}}(\xi) = X_C \,\mathrm{Skew}(X_C^\top \xi) + (I - XX^\top)\xi$ for all $\xi \in \mathbb{R}^{n\times|C|}$.

Given an integer $\ell \geq 2$, let $\mathfrak{C} = \{ C_1, \ldots, C_\ell \}$ be a partition of $[p]$, i.e., $C_i \cap C_j = \varnothing$ and $\bigcup_{i=1}^{\ell} C_i = [p]$. For notational simplicity, given a matrix $X \in \mathrm{St}(n,p)$, we denote $p_i = |C_i|$, $X_i = X_{C_i}$, $X_{-i} = X_{[p]\setminus C_i}$ for $i \in [\ell]$; and $C_{ij} = C_i \cup C_j$, $p_{ij} = |C_{ij}|$, $X_{ij} = X_{C_{ij}}$, $X_{-ij} = X_{[p]\setminus C_{ij}}$, $\mathcal{M}_{X_{-ij}} = \{ Y \in \mathrm{St}(n, p_{ij}) : X_{-ij}^\top Y = 0 \} \subseteq \mathrm{St}(n, p_{ij})$ for $\{i,j\} \in \binom{[\ell]}{2}$. We now present our proposed RSSM in Algorithm 1.

Algorithm 1 Randomized Submanifold Subgradient Method for Problem (1)

Input: An initial point $X^0 \in \mathrm{St}(n,p)$, a partition $\mathfrak{C} = \{ C_1, \ldots, C_\ell \}$ of $[p]$ with $\ell \geq 2$, and a sequence of stepsizes $\{\gamma_k\}_k \subseteq (0, \frac{1}{L})$.

1: **for** $k = 0, 1, 2, \ldots$ **do**

2: Generate $\{i,j\} \sim \mathrm{Uniform}\binom{[\ell]}{2}$ and compute the partial Riemannian subgradient

$$\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k) = \mathcal{P}_{T_{X^k_{ij}}\mathcal{M}_{X^k_{-ij}}} \left(\widetilde{\nabla}^{C_{ij}} f(X^k) \right) \in T_{X^k_{ij}}\mathcal{M}_{X^k_{-ij}}. \tag{11}$$

3: Perform the updates

$$X^{k+1}_{ij} = \mathcal{P}_{\mathcal{M}_{X^k_{-ij}}} \left(X^k_{ij} - \gamma_k \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k) \right) = \mathcal{P}_{\mathrm{St}(n,p_{ij})} \left(X^k_{ij} - \gamma_k \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k) \right), \tag{12}$$

$$X^{k+1}_{-ij} = X^k_{-ij}.$$

4: **end for**

At each iteration k, RSSM first randomly selects an unordered column block index pair $\{i,j\} \in \binom{[\ell]}{2}$ uniformly. Then, it computes the partial Riemannian subgradient $\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k)$ according to (11). Lastly, it updates the block X_{ij} via the retracted partial Riemannian subgradient step in (12) and keeps the block X_{-ij} unchanged. We note that line 3 of RSSM is well defined, in the sense that the two projections in (12) are equal and yield a singleton. This follows from Lemma 3.2, because we have $\left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k) \right\| \leq L$ by the L-Lipschitz continuity of f and

$$\mathrm{dist}\left(X^k_{ij} - \gamma_k \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k), \mathcal{M}_{X^k_{-ij}} \right) \leq \gamma_k \left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k) \right\| < 1$$

by the fact that $\gamma_k \in (0, \frac{1}{L})$.

Remark 3.4. *Besides selecting the index pair $\{i,j\}$ randomly according to the uniform distribution, there are other randomized or deterministic selection rules, such as cyclic, Gauss-Southwell, etc. The* *uniformly random selection rule is theoretically advantageous due to its simplicity, especially in the Riemannian setting. Specifically, for deterministic selection rules, in order to analyze the aggregated effect of a full epoch of iterations that cover all the columns, one needs to relate the partial Riemannian subgradients at consecutive iterates, which typically requires the notion of parallel transport.*

3.3 Per-Iteration Complexity

The proposition below establishes the per-iteration complexity of our proposed RSSM in terms of number of floating point operations when we choose a *uniform* partition of the column indices.

Proposition 3.5 (Per-iteration Complexity)**.** *Let $\mathfrak{C} = \{ C_1, \ldots, C_\ell \}$ be a partition of $[p]$ with $\ell = |\mathfrak{C}| \geq 2$ and $|C_i| \leq \lceil \frac{p}{\ell} \rceil$ for all $i \in [\ell]$. Given the partial Euclidean subgradient $\widetilde{\nabla}^{C_{ij}} f(X) \in \mathbb{R}^{n \times p_{ij}}$, an iteration of RSSM requires $\mathcal{O}\left(\frac{np^2}{\ell}\right)$ floating point operations.*

Proof. Given a matrix $\xi_{ij} \in \mathbb{R}^{n \times p_{ij}}$, we can implement (11) as follows:

1. Find $Y = X^\top \xi_{ij} \in \mathbb{R}^{p \times p_{ij}}$, which involves $\mathcal{O}(npp_{ij})$ floating point operations.
2. Update $Y(C_{ij}, :) \in \mathbb{R}^{p_{ij} \times p_{ij}}$ by $\mathrm{Sym}(Y(C_{ij}, :))$, which involves $\mathcal{O}(p_{ij}^2)$ floating point operations.
3. Perform the multiplication $XY \in \mathbb{R}^{n \times p_{ij}}$, which involves $\mathcal{O}(npp_{ij})$ floating point operations.
4. Compute $\mathcal{P}_{T_{X_{ij}} \mathcal{M}_{X_{-ij}}}(\xi_{ij}) = \xi_{ij} - XY \in \mathbb{R}^{n \times p_{ij}}$, which involves np_{ij} floating point operations.

Hence, given the partial Euclidean subgradient $\widetilde{\nabla}^{C_{ij}} f(X) \in \mathbb{R}^{n \times p_{ij}}$, computing the partial Riemannian subgradient $\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) \in \mathbb{R}^{n \times p_{ij}}$ via (11) requires $\mathcal{O}(npp_{ij} + np_{ij} + p_{ij}^2) = \mathcal{O}(npp_{ij})$ (since $p_{ij} < p$) floating point operations. Moreover, given a matrix $\Xi \in \mathbb{R}^{n \times p_{ij}}$, performing the polar decomposition $\mathcal{P}_{\mathrm{St}(n,p_{ij})}(\Xi)$ requires $\mathcal{O}(np_{ij}^2)$ floating point operations; see, e.g., [GVL13, Chapter 8.6.4]. This implies that the updates in (12) require $\mathcal{O}(np_{ij} + np_{ij}^2) = \mathcal{O}(np_{ij}^2)$ floating point operations. Putting the above together and noting that $p_{ij} \leq \lceil \frac{2p}{\ell} \rceil = \mathcal{O}(\frac{p}{\ell})$, we conclude that an iteration of RSSM requires $\mathcal{O}(npp_{ij} + np_{ij}^2) = \mathcal{O}(npp_{ij}) = \mathcal{O}(\frac{np^2}{\ell})$ floating point operations. □

4 Convergence Analysis

In this section, we study the convergence behavior of the proposed RSSM. We first prove a generalization of [LCD+21, Theorem 1], which is instrumental to the analysis of subgradient-type methods for weakly convex optimization over a compact proximally smooth manifold. Recall that the Stiefel manifold and the submanifold block introduced in Definition 3.1 are both compact proximally smooth; see Section 2.2 and Lemma 3.2.

Lemma 4.1 (Riemannian Subgradient Inequality)**.** *Let $\mathcal{M} \subseteq \mathbb{R}^{n \times p}$ be a compact R-proximally smooth manifold. Suppose that $h : \mathbb{R}^{n \times p} \to \mathbb{R}$ is τ-weakly convex for some $\tau \in \mathbb{R}$. Then, for any bounded convex neighborhood $\mathcal{U}$ that contains $\mathcal{M}$, there exists a constant $L > 0$ such that h is L-Lipschitz continuous on $\mathcal{U}$. Moreover, for any $X, Y \in \mathcal{M}$, $\widetilde{\nabla} h(X) \in \partial h(X)$, and $\widetilde{\nabla}_{\mathcal{M}} h(X) = \mathcal{P}_{T_X \mathcal{M}} \widetilde{\nabla} h(X) \in \partial_{\mathcal{M}} h(X)$, we have*

$$h(Y) \geq h(X) + \langle \widetilde{\nabla}_{\mathcal{M}} h(X), Y - X \rangle - \tfrac{\tau + \|\widetilde{\nabla} h(X)\|/R}{2} \|Y - X\|^2 \tag{13}$$

$$\geq h(X) + \langle \widetilde{\nabla}_{\mathcal{M}} h(X), Y - X \rangle - \tfrac{\tau + L/R}{2} \|Y - X\|^2. \tag{14}$$

We remark that Lemma 4.1 applies even when h is strongly convex, i.e., $\tau < 0$.

Proof of Lemma 4.1. The Lipschitz continuity of h on any bounded convex neighborhood that contains $\mathcal{M}$ follows directly from [Via83, Proposition 4.4]. Now, for any $X, Y \in \mathcal{M}$ and $\widetilde{\nabla} h(X) \in \partial h(X)$, we use (2) to get

$$
\begin{aligned}
h(Y) &\geq h(X) + \langle \widetilde{\nabla} h(X), Y - X \rangle - \tfrac{\tau}{2} \| Y - X \|^2 \\
&= h(X) + \langle \mathcal{P}_{T_X \mathcal{M}} \widetilde{\nabla} h(X) + \mathcal{P}_{N_X \mathcal{M}} \widetilde{\nabla} h(X), Y - X \rangle - \tfrac{\tau}{2} \| Y - X \|^2.
\end{aligned}
$$

By (7), we have

$$
\langle \mathcal{P}_{N_X \mathcal{M}} \widetilde{\nabla} h(X), Y - X \rangle \geq - \frac{\| \mathcal{P}_{N_X \mathcal{M}} \widetilde{\nabla} h(X) \|}{2R} \| Y - X \|^2 \geq - \frac{\| \widetilde{\nabla} h(X) \|}{2R} \| Y - X \|^2.
$$

Putting the above together and using the expression for $\partial_{\mathcal{M}} h(X)$ in (3), we get (13). Since the L-Lipschitz continuity of h on $\mathcal{U}$ implies that $\| \widetilde{\nabla} h(X) \| \leq L$, we get (14). $\square$

In what follows, we shall continue to work under the setting where the objective function f of Problem (1) is τ-weakly convex and L-Lipschitz continuous on some bounded convex neighborhood $\Omega \subseteq \mathbb{R}^{n \times p}$ containing $\mathrm{St}(n,p)$ for some $\tau \geq 0$ and $L > 0$, and the set $\mathfrak{C} = \{ C_1, \ldots, C_\ell \}$ forms a partition of $[p]$ with $\ell = |\mathfrak{C}| \geq 2$.

4.1 Averaging Operator

Motivated by our uniformly random selection rule for the submanifold blocks, we now study the average of the *lifted* partial Riemannian subgradients $\left\{ \mathcal{P}_{T_{X_{ij}} \mathcal{M}_{X_{-ij}}}(\xi_{ij}) I_{ij}^\top : \{i,j\} \in \binom{[\ell]}{2} \right\}$ on $T_X \mathrm{St}(n,p)$. Note that given the partition $\mathfrak{C}$ of $[p]$, we have $X_{-ij}^\top (\mathcal{P}_{T_{X_{ij}} \mathcal{M}_{X_{-ij}}}(\xi_{ij})) = 0$ for all $\{i,j\} \in \binom{[\ell]}{2}$. Consider the linear operator $\mathcal{A}_X : \mathbb{R}^{n \times p} \to \mathbb{R}^{n \times p}$, called the *averaging operator*, defined as

$$
\mathcal{A}_X(\xi) = \frac{1}{\binom{\ell}{2}} \sum_{\{i,j\} \in \binom{[\ell]}{2}} (I - X_{-ij} X_{-ij}^\top) \xi_{ij} I_{ij}^\top. \tag{15}
$$

Basically, the averaging operator $\mathcal{A}_X$ computes a simple average of the projections of partial Euclidean subgradients onto the orthogonal complement of the subspace spanned by the unaltered columns. Upon letting $X_\perp \in \mathrm{St}(n, n-p)$ be such that $[X \ X_\perp] \in \mathrm{St}(n,n)$, we can express $\mathcal{A}_X$ as

$$
\begin{aligned}
\mathcal{A}_X(\xi) &= X \left(\binom{\ell}{2}^{-1} \sum_{\{i,j\} \in \binom{[\ell]}{2}} I_{ij} X_{ij}^\top \xi_{ij} I_{ij}^\top \right) + X_\perp \left(X_\perp^\top \left(\binom{\ell}{2}^{-1} \sum_{\{i,j\} \in \binom{[\ell]}{2}} \xi_{ij} I_{ij}^\top \right) \right) \\
&= X \left(\binom{\ell}{2}^{-1} Q \boxdot (X^\top \xi) \right) + X_\perp \left(\tfrac{2}{\ell} X_\perp^\top \xi \right),
\end{aligned} \tag{16}
$$

where $Q = J + (\ell - 2) I$ and $\boxdot$ denotes the *block Hadamard product* with respect to $\mathfrak{C}$, i.e., for $A = (a_{ij}) \in \mathbb{R}^{\ell \times \ell}$ and $B = (B_{ij}) \in \mathbb{R}^{p \times p}$ with $B_{ij} \in \mathbb{R}^{p_i \times p_j}$, we define $A \boxdot B = (a_{ij} B_{ij}) \in \mathbb{R}^{p \times p}$. In particular, we have

$$
\mathcal{A}_X(\xi) = X \begin{bmatrix} \frac{2}{\ell} X_1^\top \xi_1 & \binom{\ell}{2}^{-1} X_1^\top \xi_2 & \cdots & \binom{\ell}{2}^{-1} X_1^\top \xi_\ell \\ \binom{\ell}{2}^{-1} X_2^\top \xi_1 & \frac{2}{\ell} X_2^\top \xi_2 & \cdots & \binom{\ell}{2}^{-1} X_2^\top \xi_\ell \\ \vdots & \vdots & \ddots & \vdots \\ \binom{\ell}{2}^{-1} X_\ell^\top \xi_1 & \binom{\ell}{2}^{-1} X_\ell^\top \xi_2 & \cdots & \frac{2}{\ell} X_\ell^\top \xi_\ell \end{bmatrix} + X_\perp \left(\tfrac{2}{\ell} X_\perp^\top \xi \right).
$$

We show in Lemma A.1 that $\mathcal{A}_X$ and $\mathcal{P}_{T_X \mathrm{St}(n,p)}$ commute, and

$$\mathcal{A}_X \circ \mathcal{P}_{T_X \mathrm{St}(n,p)}(\xi) = \binom{\ell}{2}^{-1} \sum_{\{i,j\} \in \binom{[\ell]}{2}} \mathcal{P}_{T_{X_{ij}} \mathcal{M}_{X_{-ij}}}(\xi_{ij}) I_{ij}^\top.$$

Furthermore, we show in Lemma A.2 that $\mathcal{A}_X$ is positive definite, which implies that the Mahalanobis inner product $\langle \xi, \eta \rangle_{\mathcal{A}_X^{-1}} = \langle \mathcal{A}_X^{-1}(\xi), \eta \rangle$ and the Mahalanobis norm $\|\xi\|_{\mathcal{A}_X^{-1}} = \sqrt{\langle \mathcal{A}_X^{-1}(\xi), \xi \rangle}$ are well defined.

The following lemma highlights the importance of $\langle \cdot, \cdot \rangle_{\mathcal{A}_X^{-1}}$.

Lemma 4.2 (Average of Partial Riemannian Subgradients)**.** *Suppose that $\{i,j\} \sim \mathrm{Uniform}\binom{[\ell]}{2}$. For any $X \in \mathrm{St}(n,p)$, $\widetilde{\nabla} f(X) \in \partial f(X)$, and $\eta \in \mathbb{R}^{n \times p}$, we have*

$$\mathbb{E}\left[\left\langle \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) I_{ij}^\top, \eta \right\rangle_{\mathcal{A}_X^{-1}} \right] = \langle \widetilde{\nabla}_{\mathrm{St}(n,p)} f(X), \eta \rangle, \tag{17}$$

$$\mathbb{E}\left[\left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) I_{ij}^\top \right\|^2_{\mathcal{A}_X^{-1}} \right] = \|\widetilde{\nabla}_{\mathrm{St}(n,p)} f(X)\|^2, \tag{18}$$

where $\widetilde{\nabla}_{\mathrm{St}(n,p)} f(X) = \mathcal{P}_{T_X \mathrm{St}(n,p)}(\widetilde{\nabla} f(X))$.

We remark that in (17) and (18), both the full and partial Riemannian subgradients, $\widetilde{\nabla}_{\mathrm{St}(n,p)} f(X) = \mathcal{P}_{T_X \mathrm{St}(n,p)}(\widetilde{\nabla} f(X))$ and $\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) = \mathcal{P}_{T_{X_{ij}} \mathcal{M}_{X_{-ij}}}(\widetilde{\nabla} f(X) I_{ij})$ (the latter expression follows from (10) and Proposition 3.3), depend on the choice of $\widetilde{\nabla} f(X)$.

Proof of Lemma 4.2. Using Lemma A.1 and the expression for $\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X)$ mentioned above, we have

$$\begin{aligned}
\mathbb{E}\left[\left\langle \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) I_{ij}^\top, \eta \right\rangle_{\mathcal{A}_X^{-1}} \right] &= \left\langle \frac{1}{\binom{\ell}{2}} \sum_{\{i,j\} \in \binom{[\ell]}{2}} \mathcal{P}_{T_{X_{ij}} \mathcal{M}_{X_{-ij}}}(\widetilde{\nabla} f(X) I_{ij}) I_{ij}^\top, \eta \right\rangle_{\mathcal{A}_X^{-1}} \\
&= \left\langle \mathcal{A}_X \circ \mathcal{P}_{T_X \mathrm{St}(n,p)}(\widetilde{\nabla} f(X)), \eta \right\rangle_{\mathcal{A}_X^{-1}} = \langle \widetilde{\nabla}_{\mathrm{St}(n,p)} f(X), \eta \rangle,
\end{aligned}$$

which shows that (17) holds. Next, for $\{i,j\} \in \binom{[\ell]}{2}$, we define $\mathcal{Q}_{ij} : \mathbb{R}^{n \times p} \to \mathbb{R}^{n \times p}$ as $\mathcal{Q}_{ij}(\xi) = \mathcal{P}_{T_{X_{ij}} \mathcal{M}_{X_{-ij}}}(\xi I_{ij}) I_{ij}^\top$. Since $\mathcal{P}_{T_{X_{ij}} \mathcal{M}_{X_{-ij}}} \circ \mathcal{P}_{T_{X_{ij}} \mathcal{M}_{X_{-ij}}} = \mathcal{P}_{T_{X_{ij}} \mathcal{M}_{X_{-ij}}}$, we have $\mathcal{Q}_{ij} \circ \mathcal{Q}_{ij} = \mathcal{Q}_{ij}$. Moreover, by (31), we have $\mathcal{A}_X \circ \mathcal{Q}_{ij} = \mathcal{Q}_{ij} \circ \mathcal{A}_X$. This, together with Lemma A.2, implies that $\mathcal{A}_X^{-1} \circ \mathcal{Q}_{ij} = \mathcal{Q}_{ij} \circ \mathcal{A}_X^{-1}$. It follows that

$$\begin{aligned}
&\left\langle \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) I_{ij}^\top, \widetilde{\nabla} f(X) \right\rangle_{\mathcal{A}_X^{-1}} = \left\langle \mathcal{A}_X^{-1} \left(\mathcal{P}_{T_{X_{ij}} \mathcal{M}_{X_{-ij}}}(\widetilde{\nabla} f(X) I_{ij}) I_{ij}^\top \right), \widetilde{\nabla} f(X) \right\rangle \\
&= \left\langle \mathcal{A}_X^{-1} \left(\mathcal{Q}_{ij} \circ \mathcal{Q}_{ij}(\widetilde{\nabla} f(X)) \right), \widetilde{\nabla} f(X) \right\rangle = \left\langle \mathcal{Q}_{ij} \left(\mathcal{A}_X^{-1} \circ \mathcal{Q}_{ij}(\widetilde{\nabla} f(X)) \right), \widetilde{\nabla} f(X) \right\rangle \\
&= \left\langle \mathcal{P}_{T_{X_{ij}} \mathcal{M}_{X_{-ij}}} \left(\mathcal{A}_X^{-1} \left(\mathcal{P}_{T_{X_{ij}} \mathcal{M}_{X_{-ij}}}(\widetilde{\nabla} f(X) I_{ij}) I_{ij}^\top \right) I_{ij} \right) I_{ij}^\top, \widetilde{\nabla} f(X) \right\rangle \\
&= \left\langle \mathcal{A}_X^{-1} \left(\mathcal{P}_{T_{X_{ij}} \mathcal{M}_{X_{-ij}}}(\widetilde{\nabla} f(X) I_{ij}) I_{ij}^\top \right), \mathcal{P}_{T_{X_{ij}} \mathcal{M}_{X_{-ij}}}(\widetilde{\nabla} f(X) I_{ij}) I_{ij}^\top \right\rangle = \left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) I_{ij}^\top \right\|^2_{\mathcal{A}_X^{-1}}.
\end{aligned}$$

By taking $\eta = \widetilde{\nabla} f(X)$ in (17), we obtain (18). □

Our next lemma provides an estimate of the change in the local metrics at successive iterates generated by RSSM.

Lemma 4.3 (Metric Comparison)**.** *Suppose that* $\{i,j\} \sim \mathrm{Uniform}\binom{[\ell]}{2}$*. Given* $X \in \mathrm{St}(n,p)$*,* $\widetilde{\nabla} f(X) \in \partial f(X)$*, and* $\gamma \in (0, \frac{1}{L})$*, let* $X^+ \in \mathrm{St}(n,p)$ *be the (unique) matrix satisfying*

$$X_{ij}^+ = \mathcal{P}_{\mathcal{M}_{X_{-ij}}}\left(X_{ij} - \gamma \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X)\right)$$

and $X_{-ij}^+ = X_{-ij}$*. Then, for any* $Y \in \mathrm{St}(n,p)$*, we have*

$$\begin{aligned}\mathbb{E}\left[\|Y - X^+\|^2_{\mathcal{A}^{-1}_{X^+}}\right] &\le \mathbb{E}\left[\|Y - X^+\|^2_{\mathcal{A}^{-1}_{X}}\right] + 2(\ell-2)\gamma L \|Y - X\|^2 \\ &\quad + (\ell-2)\gamma^2 L^2\left(\|Y-X\|^2 + 2\|Y-X\|\right) + (\ell-2)\gamma^3 L^3 (\|Y-X\|^2 + 1). \end{aligned} \tag{19}$$

Before proving Lemma 4.3, we present a lemma that characterizes the operator norm of the difference between two orthogonal projections, whose proof can be found in [GVL13, Theorem 2.5.1].

Lemma 4.4. *Let* $X, Y \in \mathrm{St}(n,p)$ *and* $X_\perp, Y_\perp \in \mathrm{St}(n, n-p)$ *be such that* $[X\ X_\perp]$*,* $[Y\ Y_\perp] \in \mathrm{St}(n,n)$*. Then, we have* $\|XX^\top - YY^\top\|_{\mathsf{op}} = \|X_\perp^\top Y\|_{\mathsf{op}} = \|Y_\perp^\top X\|_{\mathsf{op}}$*.*

Proof of Lemma 4.3. Consider the linear operator $\Psi : \mathbb{R}^{n\times p} \to \mathbb{R}^{n \times p}$ defined as

$$\Psi_X(\xi) = X\left[(J - I) \boxdot (X^\top \xi)\right].$$

It is easily seen that Ψ_X is self-adjoint. By (32), for any $Y \in \mathrm{St}(n,p)$, we have

$$\begin{aligned}&\|Y - X^+\|^2_{\mathcal{A}^{-1}_{X^+}} - \|Y - X^+\|^2_{\mathcal{A}^{-1}_{X}} \\ &= \langle \mathcal{A}^{-1}_{X^+}(Y - X^+), Y - X^+ \rangle - \langle \mathcal{A}^{-1}_{X}(Y - X^+), Y - X^+ \rangle \\ &= \tfrac{\ell(\ell-2)}{2} \langle (\Psi_{X^+} - \Psi_X)(Y - X^+), Y - X^+ \rangle \\ &= \tfrac{\ell(\ell-2)}{2} \big(\langle (\Psi_{X^+} - \Psi_X)(Y - X), Y - X \rangle \\ &\qquad + 2\langle (\Psi_{X^+} - \Psi_X)(X - X^+), Y - X \rangle + \langle (\Psi_{X^+} - \Psi_X)(X - X^+), X - X^+ \rangle \big).\end{aligned}$$

Now, for any $\xi \in \mathbb{R}^{n\times p}$, we have

$$(\Psi_{X^+} - \Psi_X)(\xi) = (X_i^+ X_i^{+\top} - X_i X_i^\top)\xi_{-i} I_{-i}^\top + (X_j^+ X_j^{+\top} - X_j X_j^\top)\xi_{-j} I_{-j}^\top \tag{20}$$

and, using (5),

$$\begin{aligned}\langle (\Psi_{X^+} - \Psi_X)(\xi), \xi \rangle &= \left(\|X_i^{+\top}\xi_{-i}\|^2 - \|X_i^\top \xi_{-i}\|^2\right) + \left(\|X_j^{+\top}\xi_{-j}\|^2 - \|X_j^\top \xi_{-j}\|^2\right) \\ &= \langle (X_i^+ + X_i)^\top \xi_{-i}, (X_i^+ - X_i)^\top \xi_{-i} \rangle + \langle (X_j^+ + X_j)^\top \xi_{-j}, (X_j^+ - X_j)^\top \xi_{-j} \rangle \\ &\le \|(X_i^+ + X_i)^\top \xi_{-i}\| \cdot \|X_i^+ - X_i\| \cdot \|\xi_{-i}\| + \|(X_j^+ + X_j)^\top \xi_{-j}\| \cdot \|X_j^+ - X_j\| \cdot \|\xi_{-j}\| \\ &\le \|(X_i^+ + X_i)^\top \xi\| \cdot \|X_i^+ - X_i\| \cdot \|\xi\| + \|(X_j^+ + X_j)^\top \xi\| \cdot \|X_j^+ - X_j\| \cdot \|\xi\| \\ &\le \|(X_{ij}^+ + X_{ij})^\top \xi\| \cdot \|X_{ij}^+ - X_{ij}\| \cdot \|\xi\| \le \gamma \left\|\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X)\right\| \cdot \|(X_{ij}^+ + X_{ij})^\top \xi\| \cdot \|\xi\|.\end{aligned}$$

Since $\gamma < \frac{1}{L}$, the second-order boundedness condition (6) implies that

$$\begin{aligned}\|(X_{ij}^+ + X_{ij})^\top \xi\| &= \left\|\left(2X_{ij} - \gamma \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) + X_{ij}^+ - X_{ij} + \gamma \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X)\right)^\top \xi\right\| \\ &\le 2\|X_{ij}^\top \xi\| + \gamma\left\|\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X)\right\| \cdot \|\xi\| + \gamma^2 \left\|\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X)\right\|^2 \cdot \|\xi\|.\end{aligned}$$

It follows that

$$\langle (\Psi_{X^+} - \Psi_X)(\xi), \xi \rangle \le 2\gamma \left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) \right\| \cdot \|X_{ij}^\top \xi\| \cdot \|\xi\| + \gamma^2 \left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) \right\|^2 \cdot \|\xi\|^2 + \gamma^3 \left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) \right\|^3 \cdot \|\xi\|^2. \tag{21}$$

By applying (20) with $\xi = X - X^+$ and using Lemma 4.4, we get

$$\begin{aligned}
&\|(\Psi_{X^+} - \Psi_X)(X - X^+)\| \le \|X_{ij}^+ X_{ij}^{+\top} - X_{ij} X_{ij}^\top\|_{\mathsf{op}} \cdot \|X - X^+\| \\
&\le \|(I - X_{ij}X_{ij}^\top) X_{ij}^+\|_{\mathsf{op}} \cdot \|X_{ij} - X_{ij}^+\| = \|(I - X_{ij}X_{ij}^\top)(X_{ij}^+ - X_{ij})\|_{\mathsf{op}} \cdot \|X_{ij} - X_{ij}^+\| \\
&\le \|X_{ij} - X_{ij}^+\|^2 \le \gamma^2 \left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) \right\|^2.
\end{aligned} \tag{22}$$

By taking $\xi = Y - X$ in (21) and combining the result with (22), we obtain

$$\begin{aligned}
&\|Y - X^+\|^2_{\mathcal{A}^{-1}_{X^+}} - \|Y - X^+\|^2_{\mathcal{A}^{-1}_X} \\
&\le \tfrac{\ell(\ell-2)}{2} \big(\langle (\Psi_{X^+} - \Psi_X)(Y - X), Y - X \rangle \\
&\quad + 2\,\|(\Psi_{X^+} - \Psi_X)(X - X^+)\| \cdot \|Y - X\| + \|(\Psi_{X^+} - \Psi_X)(X - X^+)\| \cdot \|X - X^+\|\big) \\
&\le \tfrac{\ell(\ell-2)}{2} \Big(2\gamma \left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) \right\| \cdot \|X_{ij}^\top (Y - X)\| \cdot \|Y - X\| \\
&\qquad + \gamma^2 \left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) \right\|^2 \cdot \|Y - X\|^2 + \gamma^3 \left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) \right\|^3 \cdot \|Y - X\|^2 \\
&\qquad + 2\gamma^2 \left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) \right\|^2 \cdot \|Y - X\| + \gamma^3 \left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) \right\|^3 \Big) \\
&\le (\ell - 2) \Big(2\gamma \left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) I_{ij}^\top \right\|_{\mathcal{A}^{-1}_X} \cdot \sqrt{\tfrac{\ell}{2}} \|X_{ij}^\top (Y - X)\| \cdot \|Y - X\| \\
&\qquad + \gamma^2 (1 + \gamma L) \left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) I_{ij}^\top \right\|^2_{\mathcal{A}^{-1}_X} \cdot \|Y - X\|^2 \\
&\qquad + 2\gamma^2 \left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) I_{ij}^\top \right\|^2_{\mathcal{A}^{-1}_X} \cdot \|Y - X\| + \gamma^3 L \left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) I_{ij}^\top \right\|^2_{\mathcal{A}^{-1}_X} \Big),
\end{aligned} \tag{23}$$

where the last inequality follows from Lemma A.2 and the fact that $\left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) \right\| \le L$. Now, we take expectation on both sides of the above inequality with respect to $\{i, j\} \sim \mathrm{Uniform}\binom{[\ell]}{2}$. Observe that

$$\mathbb{E}\left[\|X_{ij}^\top (Y - X)\|^2\right] = \frac{1}{\binom{\ell}{2}} \sum_{\{i,j\} \in \binom{[\ell]}{2}} \left(\|X_i^\top (Y - X)\|^2 + \|X_j^\top (Y - X)\|^2\right) = \tfrac{2}{\ell} \|X^\top (Y - X)\|^2.$$

Moreover, by the Cauchy–Schwarz inequality, we get

$$\begin{aligned}
&\mathbb{E}\left[\left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) I_{ij}^\top \right\|_{\mathcal{A}^{-1}_X} \cdot \sqrt{\tfrac{\ell}{2}} \|X_{ij}^\top (Y - X)\|\right] \\
&\le \sqrt{\mathbb{E}\left[\left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X_{-ij}}} f(X) I_{ij}^\top \right\|^2_{\mathcal{A}^{-1}_X}\right]} \cdot \sqrt{\tfrac{\ell}{2} \mathbb{E}\left[\|X_{ij}^\top (Y - X)\|^2\right]}.
\end{aligned}$$

This, together with (18), (23), and the fact that $\|\widetilde{\nabla}_{\mathrm{St}(n,p)} f(X)\| \le L$, yields

$$\mathbb{E}\left[\|Y - X^+\|^2_{\mathcal{A}^{-1}_{X^+}}\right] \le \mathbb{E}\left[\|Y - X^+\|^2_{\mathcal{A}^{-1}_X}\right] + 2(\ell - 2)\gamma \|\widetilde{\nabla}_{\mathrm{St}(n,p)} f(X)\| \cdot \|X^\top (Y - X)\| \cdot \|Y - X\|$$

$$
\begin{aligned}
&+ (\ell-2)\gamma^2(1+\gamma L)\|\widetilde{\nabla}_{\mathrm{St}(n,p)} f(X)\|^2 \cdot \|Y-X\|^2 \\
&+ 2(\ell-2)\gamma^2\|\widetilde{\nabla}_{\mathrm{St}(n,p)} f(X)\|^2 \cdot \|Y-X\| + (\ell-2)\gamma^3 L\|\widetilde{\nabla}_{\mathrm{St}(n,p)} f(X)\|^2 \\
\le\ & \mathbb{E}\left[\|Y-X^+\|^2_{\mathcal{A}_X^{-1}}\right] + 2(\ell-2)\gamma L\|Y-X\|^2 \\
&+ (\ell-2)\gamma^2 L^2\left(\|Y-X\|^2 + 2\|Y-X\|\right) + (\ell-2)\gamma^3 L^3(\|Y-X\|^2+1),
\end{aligned}
$$

as desired. □

4.2 Adaptive Moreau Envelope, Adaptive Proximal Map, and Surrogate Stationarity Measure

For nonsmooth optimization over a compact embedded submanifold, the work [LCD+21] defines Riemannian analogues of the Moreau envelope and proximal map to facilitate the convergence analysis of Riemannian subgradient-type methods. However, these constructs are not well suited for capturing the effects of the random submanifold block updates in our proposed RSSM. To address this difficulty, we introduce adaptive versions of the Moreau envelope and proximal map in [LCD+21], in which we use the norm induced by the inverse of the averaging operator to measure proximity. Specifically, for Problem (1), given the partition $\mathfrak{C}$ of $[p]$, we define the *adaptive Moreau envelope* and *adaptive proximal map* at $X \in \mathrm{St}(n,p)$ as

$$
f_\lambda^{\mathfrak{C}}(X) = \min_{Y \in \mathrm{St}(n,p)} \left\{ f(Y) + \tfrac{1}{2\lambda}\|Y-X\|^2_{\mathcal{A}_X^{-1}} \right\}
$$

and

$$
P_{\lambda f}^{\mathfrak{C}}(X) = \operatorname*{argmin}_{Y \in \mathrm{St}(n,p)} \left\{ f(Y) + \tfrac{1}{2\lambda}\|Y-X\|^2_{\mathcal{A}_X^{-1}} \right\}, \tag{24}
$$

respectively. As we shall see, the above constructs allow us to use Lemma 4.2 to study the average effects of the random submanifold block updates in RSSM. We first show that for small $\lambda > 0$, the adaptive proximal map $P_{\lambda f}^{\mathfrak{C}}$ is single-valued and satisfies a Lipschitz-like property; cf. [WHC+23, Lemma 4.2].

Lemma 4.5. *Consider the setting of Lemma 4.3. If $\lambda \in (0, \frac{\ell}{2(\tau+(2\ell-1)L)})$, then the adaptive proximal map $P_{\lambda f}^{\mathfrak{C}}$ is single-valued and satisfies*

$$
\|P_{\lambda f}^{\mathfrak{C}}(X) - P_{\lambda f}^{\mathfrak{C}}(X^+)\| \le \frac{\binom{\ell}{2}}{\frac{\ell}{2} - \lambda(\tau+(2\ell-1)L)}\|X-X^+\|. \tag{25}
$$

Proof. Let $h_X : \mathbb{R}^{n\times p} \to \mathbb{R}$ be the function defined as $h_X(Z) = f(Z) + \frac{1}{2\lambda}\|Z-X\|^2_{\mathcal{A}_X^{-1}}$. By Lemma A.2, when $\lambda < \frac{\ell}{2\tau}$, the function h_X is $\left(\frac{\ell}{2\lambda} - \tau\right)$-strongly convex. Moreover, for any $Z \in \mathbb{R}^{n\times p}$, we have $\partial h_X(Z) = \partial f(Z) + \frac{1}{\lambda}\mathcal{A}_X^{-1}(Z-X)$.

If $Z \in P_{\lambda f}^{\mathfrak{C}}(X) = \operatorname{argmin}_{Z' \in \mathrm{St}(n,p)} h_X(Z')$, then $h_X(Z) = f(Z) + \frac{1}{2\lambda}\|Z-X\|^2_{\mathcal{A}_X^{-1}} \le f(X)$. This, together with Lemma A.2, implies that $\frac{\ell}{2}\cdot\frac{1}{2\lambda}\|Z-X\|^2 \le \frac{1}{2\lambda}\|Z-X\|^2_{\mathcal{A}_X^{-1}} \le f(X) - f(Z) \le L\,\|Z-X\|$, or equivalently, $\|Z-X\| \le \frac{4\lambda L}{\ell}$. Hence, we have

$$
\|\widetilde{\nabla} h_X(Z)\| \le \|\widetilde{\nabla} f(Z)\| + \tfrac{1}{\lambda}\|\mathcal{A}_X^{-1}\|_{\mathsf{op}} \cdot \|Z-X\| \le L + \tfrac{1}{\lambda}\tbinom{\ell}{2}\tfrac{4\lambda L}{\ell} = (2\ell-1)L \tag{26}
$$

for all $\widetilde{\nabla} f(Z) \in \partial f(Z)$ and $\widetilde{\nabla} h_X(Z) = \widetilde{\nabla} f(Z) + \frac{1}{\lambda}\mathcal{A}_X^{-1}(Z-X) \in \partial h_X(Z)$.

Now, let $Y, Z \in P^{\mathfrak{C}}_{\lambda f}(X)$ be arbitrary. The first-order optimality condition of (24) implies that $0 \in \partial_{\mathrm{St}(n,p)} h_X(Y)$ and $0 \in \partial_{\mathrm{St}(n,p)} h_X(Z)$. This, together with the Riemannian subgradient inequality (13), the subgradient bound (26), and the 1-proximal smoothness of $\mathrm{St}(n,p)$, yields

$$
\begin{aligned}
h_X(Y) &\geq h_X(Z) + \frac{\frac{\ell}{2\lambda} - \tau - (2\ell-1)L}{2}\|Y - Z\|^2, \\
h_X(Z) &\geq h_X(Y) + \frac{\frac{\ell}{2\lambda} - \tau - (2\ell-1)L}{2}\|Y - Z\|^2.
\end{aligned}
$$

It follows that when $\lambda < \frac{\ell}{2(\tau + (2\ell-1)L)}$, we have $\|Y - Z\|^2 \leq 0$, i.e., $Y = Z$. This shows that $P^{\mathfrak{C}}_{\lambda f}$ is single-valued.

Using the Riemannian subgradient inequality (13) again, we have

$$
\begin{aligned}
h_X(P^{\mathfrak{C}}_{\lambda f}(X^+)) &\geq h_X(P^{\mathfrak{C}}_{\lambda f}(X)) + \frac{\frac{\ell}{2\lambda} - \tau - (2\ell-1)L}{2}\|P^{\mathfrak{C}}_{\lambda f}(X^+) - P^{\mathfrak{C}}_{\lambda f}(X)\|^2, \\
h_{X^+}(P^{\mathfrak{C}}_{\lambda f}(X)) &\geq h_{X^+}(P^{\mathfrak{C}}_{\lambda f}(X^+)) + \frac{\frac{\ell}{2\lambda} - \tau - (2\ell-1)L}{2}\|P^{\mathfrak{C}}_{\lambda f}(X) - P^{\mathfrak{C}}_{\lambda f}(X^+)\|^2.
\end{aligned}
$$

Summing the above two inequalities gives

$$
\begin{aligned}
&\left(\tfrac{\ell}{2\lambda} - \tau - (2\ell-1)L\right)\|P^{\mathfrak{C}}_{\lambda f}(X) - P^{\mathfrak{C}}_{\lambda f}(X^+)\|^2 \\
&\leq h_X(P^{\mathfrak{C}}_{\lambda f}(X^+)) - h_X(P^{\mathfrak{C}}_{\lambda f}(X)) + h_{X^+}(P^{\mathfrak{C}}_{\lambda f}(X)) - h_{X^+}(P^{\mathfrak{C}}_{\lambda f}(X^+)) \\
&= \frac{1}{2\lambda}\Bigg(\|P^{\mathfrak{C}}_{\lambda f}(X^+) - X\|^2_{\mathcal{A}_X^{-1}} - \|P^{\mathfrak{C}}_{\lambda f}(X) - X\|^2_{\mathcal{A}_X^{-1}} \\
&\qquad + \|P^{\mathfrak{C}}_{\lambda f}(X) - X^+\|^2_{\mathcal{A}_{X^+}^{-1}} - \|P^{\mathfrak{C}}_{\lambda f}(X^+) - X^+\|^2_{\mathcal{A}_{X^+}^{-1}} \Bigg) \\
&= \frac{1}{2\lambda}\Bigg(\|P^{\mathfrak{C}}_{\lambda f}(X^+)\|^2_{\mathcal{A}_X^{-1}} - \|P^{\mathfrak{C}}_{\lambda f}(X^+)\|^2_{\mathcal{A}_{X^+}^{-1}} + \|P^{\mathfrak{C}}_{\lambda f}(X)\|^2_{\mathcal{A}_{X^+}^{-1}} - \|P^{\mathfrak{C}}_{\lambda f}(X)\|^2_{\mathcal{A}_X^{-1}} \\
&\qquad - \ell\,\langle P^{\mathfrak{C}}_{\lambda f}(X^+), X\rangle + \ell\,\langle P^{\mathfrak{C}}_{\lambda f}(X), X\rangle - \ell\,\langle P^{\mathfrak{C}}_{\lambda f}(X), X^+\rangle + \ell\,\langle P^{\mathfrak{C}}_{\lambda f}(X^+), X^+\rangle \Bigg) \\
&\leq \frac{\ell(\ell-2)}{2\lambda}\|X^+ - X\| \cdot \|P^{\mathfrak{C}}_{\lambda f}(X^+) - P^{\mathfrak{C}}_{\lambda f}(X)\| + \frac{\ell}{2\lambda}\langle P^{\mathfrak{C}}_{\lambda f}(X^+) - P^{\mathfrak{C}}_{\lambda f}(X), X^+ - X\rangle \\
&\leq \frac{\ell(\ell-1)}{2\lambda}\|X^+ - X\| \cdot \|P^{\mathfrak{C}}_{\lambda f}(X^+) - P^{\mathfrak{C}}_{\lambda f}(X)\| = \frac{1}{\lambda}\binom{\ell}{2}\|X^+ - X\| \cdot \|P^{\mathfrak{C}}_{\lambda f}(X^+) - P^{\mathfrak{C}}_{\lambda f}(X)\|,
\end{aligned}
$$

where the second equality follows from the fact that $\mathcal{A}_X^{-1}(X) = \frac{\ell}{2}X$ and $\mathcal{A}_{X^+}^{-1}(X^+) = \frac{\ell}{2}X^+$, and the second inequality follows from (33) and the Cauchy–Schwarz inequality. This establishes the Lipschitz-like property (25). $\square$

Now, motivated by the definition of the adaptive proximal map $P^{\mathfrak{C}}_{\lambda f}$ in (24), let us consider the map $\Theta^{\mathfrak{C}}_{\lambda f} : \mathbb{R}^{n\times p} \to \mathbb{R}_+$ defined as $\Theta^{\mathfrak{C}}_{\lambda f}(X) = \frac{1}{\lambda}\|P^{\mathfrak{C}}_{\lambda f}(X) - X\|_{\mathcal{A}_X^{-1}}$. The following proposition shows that $\Theta^{\mathfrak{C}}_{\lambda f}$ can be viewed as a *surrogate stationarity measure* of Problem (1).

Proposition 4.6 (Surrogate Stationarity Measure)**.** *Let $\lambda \in (0, \frac{\ell}{2(\tau+(2\ell-1)L)})$ be fixed, so that $P^{\mathfrak{C}}_{\lambda f}$ is single-valued. Then, for any $X \in \mathrm{St}(n,p)$, the following assertions hold:*

(a) $\mathrm{dist}(0, \partial_{\mathrm{St}(n,p)} f(P^{\mathfrak{C}}_{\lambda f}(X))) \leq \sqrt{\binom{\ell}{2}}\frac{1}{\lambda}\|P^{\mathfrak{C}}_{\lambda f}(X) - X\|_{\mathcal{A}_X^{-1}} = \sqrt{\binom{\ell}{2}}\,\Theta^{\mathfrak{C}}_{\lambda f}(X)$ *and* $\frac{\frac{\ell}{2} - \lambda(\tau+L)}{\sqrt{2\ell}}\Theta^{\mathfrak{C}}_{\lambda f}(X) = \frac{\frac{\ell}{2} - \lambda(\tau+L)}{\sqrt{2\ell}\lambda}\|P^{\mathfrak{C}}_{\lambda f}(X) - X\|_{\mathcal{A}_X^{-1}} \leq \mathrm{dist}(0, \partial_{\mathrm{St}(n,p)} f(X))$.

(b) $\Theta^{\mathfrak{C}}_{\lambda f}(X) = 0 \iff X = P^{\mathfrak{C}}_{\lambda f}(X) \iff 0 \in \partial_{\mathrm{St}(n,p)} f(X)$.

Proof. We first prove assertion (a). Let $Z = P^{\mathfrak{C}}_{\lambda f}(X)$. By the first-order optimality condition of (24), we have $\mathcal{P}_{T_Z \mathrm{St}(n,p)}\big(\frac{1}{\lambda}\mathcal{A}_X^{-1}(X - Z)\big) \in \partial_{\mathrm{St}(n,p)} f(Z)$. This, together with Lemma A.2, implies that $\operatorname{dist}(0, \partial_{\mathrm{St}(n,p)} f(Z)) \le \frac{1}{\lambda}\|\mathcal{A}_X^{-1}(X - Z)\| \le \frac{1}{\lambda}\sqrt{\binom{\ell}{2}}\|X - Z\|_{\mathcal{A}_X^{-1}}$. Now, let $V \in \partial_{\mathrm{St}(n,p)} f(X)$ be arbitrary. By the Riemannian subgradient inequality (14) and Lemma A.2, we have

$$\begin{aligned}\langle V, X - Z\rangle &\ge f(X) - f(Z) - \tfrac{\tau+L}{2}\|Z - X\|^2 \ge \tfrac{1}{2\lambda}\|Z - X\|^2_{\mathcal{A}_X^{-1}} - \tfrac{\tau+L}{2}\|Z - X\|^2 \\ &\ge \left(\tfrac{1}{2\lambda} - \tfrac{\tau+L}{2}\cdot\tfrac{2}{\ell}\right)\|Z - X\|_{\mathcal{A}_X^{-1}}\cdot\sqrt{\tfrac{\ell}{2}}\|Z - X\|.\end{aligned}$$

This implies that $\|V\| \ge \sqrt{\frac{\ell}{2}}\cdot\frac{1}{\ell}\left(\frac{\ell}{2\lambda} - \tau - L\right)\|Z - X\|_{\mathcal{A}_X^{-1}}$, where $\frac{\ell}{2\lambda} - \tau - L > (2\ell - 1)L - L = 2(\ell - 1)L > 0$.

Assertion (b) follows directly from assertion (a). □

4.3 Convergence Rate of RSSM

Our final goal is to establish the convergence rate of RSSM. We first prove a recursion lemma.

Lemma 4.7 (Recursion Lemma)**.** *Let $\{X^k\}_k$ be the iterates generated by Algorithm 1 with stepsizes $\gamma_k \in (0, \frac{1}{L})$ for all $k \ge 0$. Then, for any $k \ge 0$ and $Y \in \mathrm{St}(n,p)$, we have*

$$\begin{aligned}\mathbb{E}\left[\|Y - X^{k+1}\|^2_{\mathcal{A}^{-1}_{X^{k+1}}}\,\middle|\, X^k\right] &\le \|Y - X^k\|^2_{\mathcal{A}^{-1}_{X^k}} + 2\gamma_k\left[f(Y) - f(X^k) + \tfrac{\tau+(2\ell-3)L}{\ell}\|Y - X^k\|^2_{\mathcal{A}^{-1}_{X^k}}\right] \\ &\quad + \gamma_k^2 L^2\left(1 + (\ell-2)[4\|Y - X^k\| + \|Y - X^k\|^2]\right) + (\ell-2)\gamma_k^3 L^3(\|Y - X^k\|^2 + 3) + (\ell-2)\gamma_k^4 L^4.\end{aligned}$$

Proof. Let $\{i,j\} \sim \mathrm{Uniform}\binom{[\ell]}{2}$ be drawn. Given $X^k \in \mathrm{St}(n,p)$, we pick an arbitrary $\widetilde{\nabla} f(X^k) \in \partial f(X^k)$ and set $\widetilde{\nabla}_{\mathrm{St}(n,p)} f(X) = \mathcal{P}_{T_X \mathrm{St}(n,p)}(\widetilde{\nabla} f(X))$. Since $\gamma_k < \frac{1}{L}$, we have $\left\|X^{k+1}_{ij} - X^k_{ij} + \gamma_k \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k)\right\| \le \gamma_k^2\left\|\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k)\right\|^2$ by the second-order boundedness condition (6). Denoting the identity operator on $\mathbb{R}^{n\times p}$ by $\mathcal{I}$, we bound

$$\begin{aligned}&\|Y - X^{k+1}\|^2_{\mathcal{A}^{-1}_{X^k}} = \tfrac{\ell}{2}\|Y - X^{k+1}\|^2 + \|Y - X^{k+1}\|^2_{\mathcal{A}^{-1}_{X^k} - \frac{\ell}{2}\mathcal{I}} \\ &\le \tfrac{\ell}{2}\left\|Y - X^k + \gamma_k\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k) I_{ij}^\top\right\|^2 + \Bigg(\left\|Y - X^k + \gamma_k\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k) I_{ij}^\top\right\|_{\mathcal{A}^{-1}_{X^k} - \frac{\ell}{2}\mathcal{I}} \\ &\quad + \|\mathcal{A}^{-1}_{X^k} - \tfrac{\ell}{2}\mathcal{I}\|^{1/2}_{\mathsf{op}}\cdot\left\|X^{k+1} - X^k + \gamma_k\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k) I_{ij}^\top\right\|\Bigg)^2 \\ &\le \left\|Y - X^k + \gamma_k\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k) I_{ij}^\top\right\|^2_{\mathcal{A}^{-1}_{X^k}} \\ &\quad + \ell(\ell-2)\gamma_k^2\left\|\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k)\right\|^2\cdot\left\|Y - X^k + \gamma_k\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k) I_{ij}^\top\right\| + \tfrac{\ell(\ell-2)}{2}\gamma_k^4\left\|\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k)\right\|^4 \\ &\le \left\|Y - X^k + \gamma_k\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k) I_{ij}^\top\right\|^2_{\mathcal{A}^{-1}_{X^k}} \\ &\quad + 2(\ell-2)\gamma_k^2\left\|\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k) I_{ij}^\top\right\|^2_{\mathcal{A}^{-1}_{X^k}}\cdot\left(\|Y - X^k\| + \gamma_k L\right) + (\ell-2)\gamma_k^4 L^2\left\|\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k) I_{ij}^\top\right\|^2_{\mathcal{A}^{-1}_{X^k}}\end{aligned}$$

$$
\begin{aligned}
&\leq \|Y - X^k\|^2_{\mathcal{A}^{-1}_{X^k}} + 2\gamma_k \left\langle \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k) I_{ij}^\top, Y - X^k \right\rangle_{\mathcal{A}^{-1}_{X^k}} \\
&\quad + \gamma_k^2 \left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k) I_{ij}^\top \right\|^2_{\mathcal{A}^{-1}_{X^k}} \big(1 + 2(\ell-2)\|Y - X^k\|\big) \\
&\quad + 2(\ell-2)\gamma_k^3 L \left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k) I_{ij}^\top \right\|^2_{\mathcal{A}^{-1}_{X^k}} + (\ell-2)\gamma_k^4 L^2 \left\| \widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k) I_{ij}^\top \right\|^2_{\mathcal{A}^{-1}_{X^k}},
\end{aligned}
$$

where the first inequality follows from (5); the second and third inequalities are due to Lemma A.2 and the fact that $\left\|\widetilde{\nabla}^{C_{ij}}_{\mathcal{M}_{X^k_{-ij}}} f(X^k)\right\| \leq L$. Conditioning on X^k and taking expectation on both sides of the above inequality with respect to $\{i,j\} \sim \text{Uniform}\binom{[\ell]}{2}$, we get

$$
\begin{aligned}
\mathbb{E}\left[\|Y - X^{k+1}\|^2_{\mathcal{A}^{-1}_{X^k}} \,\middle|\, X^k \right] &\leq \|Y - X^k\|^2_{\mathcal{A}^{-1}_{X^k}} + 2\gamma_k \langle \widetilde{\nabla}_{\mathrm{St}(n,p)} f(X^k), Y - X^k \rangle \\
&\quad + \gamma_k^2 L^2 (1 + 2(\ell-2)\|Y - X^k\|) + 2(\ell-2)\gamma_k^3 L^3 + (\ell-2)\gamma_k^4 L^4 \\
&\leq \|Y - X^k\|^2_{\mathcal{A}^{-1}_{X^k}} + 2\gamma_k \left[f(Y) - f(X^k) + \tfrac{\tau+L}{2}\|Y - X^k\|^2 \right] \\
&\quad + \gamma_k^2 L^2 (1 + 2(\ell-2)\|Y - X^k\|) + 2(\ell-2)\gamma_k^3 L^3 + (\ell-2)\gamma_k^4 L^4,
\end{aligned}
$$

where the first inequality is due to Lemma 4.2 and the fact that $\|\widetilde{\nabla}_{\mathrm{St}(n,p)} f(X^k)\| \leq L$, and the second inequality is due to Lemma 4.1. It then follows from (19) that

$$
\begin{aligned}
\mathbb{E}\left[\|Y - X^{k+1}\|^2_{\mathcal{A}^{-1}_{X^{k+1}}} \,\middle|\, X^k \right] &\leq \|Y - X^k\|^2_{\mathcal{A}^{-1}_{X^k}} + 2\gamma_k \left[f(Y) - f(X^k) + \tfrac{\tau+L}{2}\|Y - X^k\|^2 \right] \\
&\quad + \gamma_k^2 L^2 (1 + 2(\ell-2)\|Y - X^k\|) + 2(\ell-2)\gamma_k^3 L^3 + (\ell-2)\gamma_k^4 L^4 \\
&\quad + 2(\ell-2)\gamma_k L \|Y - X^k\|^2 + (\ell-2)\gamma_k^2 L^2 \left(\|Y - X^k\|^2 + 2\|Y - X^k\| \right) \\
&\quad + (\ell-2)\gamma_k^3 L^3 (\|Y - X^k\|^2 + 1) \\
&\leq \|Y - X^k\|^2_{\mathcal{A}^{-1}_{X^k}} + 2\gamma_k \left[f(Y) - f(X^k) + \tfrac{\tau+(2\ell-3)L}{2} \cdot \tfrac{2}{\ell} \|Y - X^k\|^2_{\mathcal{A}^{-1}_{X^k}} \right] \\
&\quad + \gamma_k^2 L^2 \left(1 + (\ell-2)[4\|Y - X^k\| + \|Y - X^k\|^2] \right) \\
&\quad + (\ell-2)\gamma_k^3 L^3 (\|Y - X^k\|^2 + 3) + (\ell-2)\gamma_k^4 L^4.
\end{aligned}
$$

Note that the bound above is independent of the choice of $\widetilde{\nabla} f(X^k) \in \partial f(X^k)$. This completes the proof. $\square$

Lemma 4.7 allows us to prove the following inequality, which, upon rearranging, essentially establishes the sufficient decrease of the adaptive Moreau envelopes.

Proposition 4.8 (Sufficient Decrease of $f^{\mathfrak{C}}_\lambda$). *Let $\{X^k\}_k$ be the iterates generated by Algorithm 1 with stepsizes $\gamma_k \in (0, \frac{1}{L})$ for all $k \geq 0$. Then, for any $k \geq 0$ and $\lambda \in (0, \frac{\ell}{2(\tau+(2\ell-1)L)})$, conditioning on X^k, we have*

$$
\begin{aligned}
\gamma_k \Theta^{\mathfrak{C}}_{\lambda f}(X^k)^2 &= \frac{\gamma_k}{\lambda^2} \|P^{\mathfrak{C}}_{\lambda f}(X^k) - X^k\|^2_{\mathcal{A}^{-1}_{X^k}} \\
&\leq \frac{f^{\mathfrak{C}}_\lambda(X^k) - \mathbb{E}[\, f^{\mathfrak{C}}_\lambda(X^{k+1}) \mid X^k \,] + \frac{9\gamma_k^2 L^2}{2\lambda} + \frac{2(\ell-2)\gamma_k^3 L^3}{\lambda} + \frac{(\ell-2)\gamma_k^4 L^4}{2\lambda}}{\frac{\lambda}{\ell}\left[\frac{\ell}{2\lambda} - (\tau + (2\ell-3)L) \right]}. \qquad (27)
\end{aligned}
$$

Proof. Applying Lemma 4.7 with $Y = P^{\mathfrak{C}}_{\lambda f}(X^k) \in \mathrm{St}(n,p)$, we have

$$
\begin{aligned}
&\mathbb{E}[\, f^{\mathfrak{C}}_{\lambda}(X^{k+1}) \,|\, X^k \,] \le \mathbb{E}\left[f(P^{\mathfrak{C}}_{\lambda f}(X^k)) + \tfrac{1}{2\lambda}\|P^{\mathfrak{C}}_{\lambda f}(X^k) - X^{k+1}\|^2_{\mathcal{A}^{-1}_{X^{k+1}}} \,\middle|\, X^k \right] \\
&\le f^{\mathfrak{C}}_{\lambda}(X^k) + \tfrac{\gamma_k}{\lambda}\left[f(P^{\mathfrak{C}}_{\lambda f}(X^k)) - f(X^k) + \tfrac{\tau+(2\ell-3)L}{\ell}\|P^{\mathfrak{C}}_{\lambda f}(X^k) - X^k\|^2_{\mathcal{A}^{-1}_{X^k}} \right] \\
&\quad + \tfrac{\gamma_k^2 L^2}{2\lambda}\left(1 + (\ell-2)\left[4\left(\tfrac{4\lambda L}{\ell}\right) + \left(\tfrac{4\lambda L}{\ell}\right)^2\right]\right) + \tfrac{(\ell-2)\gamma_k^3 L^3}{2\lambda}\left(\left(\tfrac{4\lambda L}{\ell}\right)^2 + 3\right) + \tfrac{(\ell-2)\gamma_k^4 L^4}{2\lambda} \\
&\le f^{\mathfrak{C}}_{\lambda}(X^k) + \tfrac{\lambda}{\ell}\left(\tau + (2\ell-3)L - \tfrac{\ell}{2\lambda}\right)\cdot \gamma_k \Theta^{\mathfrak{C}}_{\lambda f}(X^k)^2 \\
&\quad + \tfrac{\gamma_k^2 L^2}{2\lambda}\left(1 + (\ell-2)\tfrac{16\lambda L(\ell+\lambda L)}{\ell^2}\right) + \tfrac{(\ell-2)\gamma_k^3 L^3}{2\lambda}\cdot\tfrac{16(\lambda L)^2 + 3\ell^2}{\ell^2} + \tfrac{(\ell-2)\gamma_k^4 L^4}{2\lambda} \\
&< f^{\mathfrak{C}}_{\lambda}(X^k) + \tfrac{\lambda}{\ell}\left(\tau + (2\ell-3)L - \tfrac{\ell}{2\lambda}\right)\cdot \gamma_k \Theta^{\mathfrak{C}}_{\lambda f}(X^k)^2 \\
&\quad + \tfrac{\gamma_k^2 L^2}{2\lambda}\left(1 + (\ell-2)\tfrac{4(2\ell+1)}{\ell^2}\right) + \tfrac{(\ell-2)\gamma_k^3 L^3}{2\lambda}\cdot\tfrac{4+3\ell^2}{\ell^2} + \tfrac{(\ell-2)\gamma_k^4 L^4}{2\lambda} \\
&\le f^{\mathfrak{C}}_{\lambda}(X^k) + \tfrac{\lambda}{\ell}\left(\tau + (2\ell-3)L - \tfrac{\ell}{2\lambda}\right)\cdot \gamma_k \Theta^{\mathfrak{C}}_{\lambda f}(X^k)^2 + \tfrac{9\gamma_k^2 L^2}{2\lambda} + \tfrac{2(\ell-2)\gamma_k^3 L^3}{\lambda} + \tfrac{(\ell-2)\gamma_k^4 L^4}{2\lambda},
\end{aligned}
$$

where the second inequality holds because $\lambda \in (0, \frac{\ell}{2(\tau+(2\ell-1)L)})$ and $\|P^{\mathfrak{C}}_{\lambda f}(X^k) - X^k\| \le \frac{4\lambda L}{\ell}$ from the proof of Lemma 4.5, the third inequality follows from $f(P^{\mathfrak{C}}_{\lambda f}(X^k)) + \frac{1}{2\lambda}\|P^{\mathfrak{C}}_{\lambda f}(X^k) - X^k\|^2_{\mathcal{A}^{-1}_{X^k}} \le f(X^k)$, the fourth inequality is due to $\lambda < \frac{\ell}{2(\tau+(2\ell-1)L)} < \frac{\ell}{2(2\ell-1)L} < \frac{1}{2L}$, and the last inequality uses the fact that $\ell \ge 2$. Upon rearranging and noting that $\frac{\ell}{2\lambda} - (\tau + (2\ell-3)L) > 0$, we obtain (27), as desired. ∎

By choosing suitable stepsizes $\{\gamma_k\}_k$ in Proposition 4.8 and using a telescoping argument, we obtain the following convergence rate of RSSM.

Theorem 4.9 (Non-Asymptotic Convergence Rate)**.** *Let $\lambda \in (0, \frac{\ell}{2(\tau+(2\ell-1)L)})$, $T \ge 1$ be fixed and $\{X^k\}_{k=0}^T$ be the iterates generated by Algorithm 1 with the following stepsizes:*

(a) If $\gamma_k = \frac{\Delta}{\ell^{1/3}(T+1)^{1/2}}$ for $k = 0, 1, \ldots, T$, where $\Delta \in \left(0, \frac{1}{L}\right]$ is any constant, then

$$
\begin{aligned}
&\min_{0\le k\le T} \mathbb{E}[\Theta^{\mathfrak{C}}_{\lambda f}(X^k)^2] \\
&\le \frac{f^{\mathfrak{C}}_{\lambda}(X^0) - \min_{X\in\mathrm{St}(n,p)} f^{\mathfrak{C}}_{\lambda}(X) + \frac{9L^2\Delta^2}{2\lambda\ell^{2/3}} + \frac{2L^3\Delta^3}{\lambda(T+1)^{1/2}} + \frac{L^4\Delta^4}{2\lambda\ell^{1/3}(T+1)}}{\frac{\lambda}{\ell}\left(\frac{\ell}{2\lambda} - (\tau + (2\ell-3)L)\right)\cdot\frac{\Delta}{\ell^{1/3}}(T+1)^{1/2}}.
\end{aligned}
$$

(b) If $\gamma_k = \frac{\Delta_k}{\ell^{1/3}(k+1)^{1/2}}$ with $\Delta_k \in [\Delta_{\mathsf{min}}, \frac{1}{L}]$ for $k = 0, 1, \ldots, T$, where $\Delta_{\mathsf{min}} > 0$ is any constant, then

$$
\min_{0\le k\le T} \mathbb{E}[\Theta^{\mathfrak{C}}_{\lambda f}(X^k)^2] \le \frac{f^{\mathfrak{C}}_{\lambda}(X^0) - \min_{X\in\mathrm{St}(n,p)} f^{\mathfrak{C}}_{\lambda}(X) + \frac{9(1+\log T)+6\ell^{1/3}+2}{2\lambda\ell^{1/3}}}{\frac{\lambda}{\ell}\left(\frac{\ell}{2\lambda} - (\tau + (2\ell-3)L)\right)\cdot\frac{\Delta_{\mathsf{min}}}{\ell^{1/3}}(T+1)^{1/2}}.
$$

Proof. Taking expectation on both sides of (27), we have

$$
\gamma_k\, \mathbb{E}[\Theta^{\mathfrak{C}}_{\lambda f}(X^k)^2] \le \frac{\mathbb{E}\left[f^{\mathfrak{C}}_{\lambda}(X^k)\right] - \mathbb{E}\left[f^{\mathfrak{C}}_{\lambda}(X^{k+1})\right] + \frac{9\gamma_k^2 L^2}{2\lambda} + \frac{2(\ell-2)\gamma_k^3 L^3}{\lambda} + \frac{(\ell-2)\gamma_k^4 L^4}{2\lambda}}{\frac{\lambda}{\ell}\left[\frac{\ell}{2\lambda} - (\tau + (2\ell-3)L)\right]}.
$$

Summing the above inequality over $k = 0, 1, \ldots, T$ gives

$$\min_{0\le k\le T} \mathbb{E}[\,\Theta^{\mathfrak{C}}_{\lambda f}(X^k)^2\,] \le \frac{f^{\mathfrak{C}}_\lambda(X^0) - \mathbb{E}[\,f^{\mathfrak{C}}_\lambda(X^{T+1})] + \frac{9L^2}{2\lambda}\sum_{k=0}^{T}\gamma_k^2 + \frac{2(\ell-2)L^3}{\lambda}\sum_{k=0}^{T}\gamma_k^3 + \frac{(\ell-2)L^4}{2\lambda}\sum_{k=0}^{T}\gamma_k^4}{\frac{\lambda}{\ell}\left(\frac{\ell}{2\lambda} - (\tau + (2\ell-3)L)\right)\sum_{k=0}^{T}\gamma_k}. \tag{28}$$

If $\gamma_k = \frac{\Delta}{\sqrt{\ell(T+1)}}$ for $k = 0, 1, \ldots, T$, then $\sum_{k=0}^T \gamma_k = \frac{\Delta}{\sqrt{\ell}}\sqrt{T+1}$, $\sum_{k=0}^T \gamma_k^2 = \frac{\Delta^2}{\ell}$, $\sum_{k=0}^T \gamma_k^3 = \frac{\Delta^3}{\ell^{3/2}\sqrt{T+1}}$, and $\sum_{k=0}^T \gamma_k^4 = \frac{\Delta^4}{\ell^2(T+1)}$. Substituting these into (28) gives (a).

On the other hand, if $\gamma_k = \frac{\Delta_k}{\sqrt{\ell(k+1)}}$ with $\Delta_k \in [\Delta_{\min}, \frac{1}{L}]$ for $k = 0, 1, \ldots, T$, then $\sum_{k=0}^T \gamma_k^2 \le \frac{1}{\ell L^2}\left(1 + \int_1^T \frac{du}{u}\right) = \frac{1}{\ell L^2}(1+\log T)$, $\sum_{k=0}^T \gamma_k^3 \le \frac{1}{\ell^{3/2}L^3}\left(1 + \int_1^\infty \frac{du}{u^{3/2}}\right) = \frac{3}{2\ell^{3/2}L^3}$, $\sum_{k=0}^T \gamma_k^4 \le \frac{1}{\ell^2 L^4}\left(1 + \int_1^\infty \frac{du}{u^2}\right) = \frac{2}{\ell^2 L^4}$, and $\sum_{k=0}^T \gamma_k \ge (T+1)\gamma_T = \frac{\Delta_{\min}}{\sqrt{\ell}}\sqrt{T+1}$. Substituting these into (28) gives (b). □

Choosing $\lambda = \frac{\ell}{4(\tau+(2\ell-1)L)}$ and $T \ge \ell^{2/3} - 1$ in Theorem 4.9(a) yields

$$\min_{0\le k\le T} \mathbb{E}[\,\Theta^{\mathfrak{C}}_{\lambda f}(X^k)\,] \le \sqrt{\frac{f^{\mathfrak{C}}_\lambda(X^0) - \min_{X\in \mathrm{St}(n,p)} f^{\mathfrak{C}}_\lambda(X) + \frac{7}{\lambda \ell^{1/3}}}{\frac{\lambda}{\ell}\left(\frac{\ell}{2\lambda} - \frac{\ell}{4\lambda}\right)\cdot \frac{\Delta}{\ell^{1/3}}(T+1)^{1/2}}} = \mathcal{O}\left(\sqrt[4]{\frac{\ell^{2/3}}{T}}\right),$$

which implies that the iteration complexity (*i.e.*, the number of iterations required to compute a point with the expected surrogate stationarity measure bounded by ε) of RSSM is $\mathcal{O}(\ell^{2/3}\varepsilon^{-4})$.

Theorem 4.9 shows that with sufficiently many iterations, our proposed RSSM will produce an iterate $\bar{X}$ such that the expectation $\mathbb{E}[\,\Theta^{\mathfrak{C}}_{\lambda f}(\bar{X})^2\,]$ is smaller than any given threshold. However, it does not imply that the random quantity $\Theta^{\mathfrak{C}}_{\lambda f}(\bar{X})$ will be small with high probability or that the sequence $\{\mathbb{E}[\,\Theta^{\mathfrak{C}}_{\lambda f}(X^k)^2\,]\}_k$ converges. To complement Theorem 4.9, we now demonstrate the almost-sure convergence behavior of RSSM. To begin, let $\Lambda = \left\{(\omega_1, \omega_2, \ldots) : \omega_i \in \binom{[\ell]}{2} \text{ for } i = 1, 2, \ldots\right\}$ denote the set of all possible sequences of pairs in $\binom{[\ell]}{2}$ generated by Algorithm 1. It is possible to construct a σ-algebra $\mathscr{F}$ on Λ and a probability measure $\mathbb{P}$ on $\mathscr{F}$ based on the uniform distribution on $\binom{[\ell]}{2}$ such that the triple $(\Lambda, \mathscr{F}, \mathbb{P})$ is a probability space; see, e.g., the subsection "Sequence Space" in Chapter 1, Section 2 of [Bil95]. This allows us to study various events associated with the (random) sequence of iterates generated by Algorithm 1.

Theorem 4.10 (Almost-Sure Convergence)**.** *Let $\{X^k\}_k$ be the iterates generated by Algorithm 1 with stepsizes $\{\gamma_k\}_k$ satisfying $\gamma_k \in (0, \frac{1}{L})$ for all $k \ge 0$, $\sum_{k\ge 0}\gamma_k = \infty$, and $\sum_{k\ge 0}\gamma_k^2 < \infty$. Then, for any $\lambda \in (0, \frac{\ell}{2(\tau+(2\ell-1)L)})$, we will have $\lim_{k\to\infty} \Theta^{\mathfrak{C}}_{\lambda f}(X^k) = 0$ $\mathbb{P}$-almost surely. Hence, every accumulation point of $\{X^k\}_k$ will be a stationary point of Problem (1) $\mathbb{P}$-almost surely.*

Proof. The proof involves verifying the conditions in Lemma A.5 with $Y^k = X^k$, $\mu_k = \gamma_k$ for $k \ge 0$ and $\Phi = \Theta^{\mathfrak{C}}_{\lambda f}$. First, using Lemma 4.5 with $X = X^k$ and $X^+ = X^{k+1}$ in (25) and the definition of $\Theta^{\mathfrak{C}}_{\lambda f}$, we have $|\Theta^{\mathfrak{C}}_{\lambda f}(X^{k+1}) - \Theta^{\mathfrak{C}}_{\lambda f}(X^k)| \le L_\Theta \|X^{k+1} - X^k\|$ for some $L_\Theta > 0$. Next, we have $\|X^k - X^{k+1}\| \le L\gamma_k$ for all $k \ge 0$ and $\sum_{k\ge 0}\gamma_k = \infty$. Lastly, let $f^* = \min_{X\in\mathrm{St}(n,p)} f^{\mathfrak{C}}_\lambda(X) = \min_{X\in\mathrm{St}(n,p)} f(X)$. Since $\gamma_k < \frac{1}{L}$ for all $k \ge 0$, by Proposition 4.8, we have

$$\mathbb{E}\left[\,f^{\mathfrak{C}}_\lambda(X^{k+1}) - f^* \,\middle|\, X^k\right]$$

$$\le \left(f_\lambda^{\mathfrak{C}}(X^k) - f^*\right) - \tfrac{\lambda}{\ell}\left(\tfrac{\ell}{2\lambda} - (\tau + (2\ell-3)L)\right)\cdot \gamma_k \Theta_{\lambda f}^{\mathfrak{C}}(X^k)^2 + \tfrac{9\gamma_k^2 L^2}{2\lambda} + \tfrac{2(\ell-2)\gamma_k^3 L^3}{\lambda} + \tfrac{(\ell-2)\gamma_k^4 L^4}{2\lambda}$$
$$\le \left(f_\lambda^{\mathfrak{C}}(X^k) - f^*\right) - \tfrac{\lambda}{\ell}\left(\tfrac{\ell}{2\lambda} - (\tau + (2\ell-3)L)\right)\cdot \gamma_k \Theta_{\lambda f}^{\mathfrak{C}}(X^k)^2 + \tfrac{\gamma_k^2 L^2}{\lambda}\left(\tfrac{9}{2} + \tfrac{5}{2}(\ell-2)\right).$$

Since $\sum_{k\ge 0}\gamma_k^2 < \infty$, by the supermartingale convergence theorem [RS71, Theorem 1], we will have $\sum_{k\ge 0}\gamma_k \Theta_{\lambda f}^{\mathfrak{C}}(X^k)^2 < \infty$ $\mathbb{P}$-almost surely. Thus, by Lemma A.5, we will have $\lim_{k\to\infty}\Theta_{\lambda f}^{\mathfrak{C}}(X^k) = 0$ $\mathbb{P}$-almost surely. In particular, every accumulation point of $\{X^k\}_k$ will be a stationary point of Problem (1) $\mathbb{P}$-almost surely. □

With a slightly more conservative choice of stepsizes, we obtain the following almost-sure asymptotic convergence rate result, which complements the in-expectation non-asymptotic convergence rate results in Theorem 4.9.

Theorem 4.11 (Almost-Sure Asymptotic Convergence Rate). *Let $\Delta_{\mathsf{min}} > 0$ be a given constant and $\{\Delta_k\}_k$ be a sequence in $(\Delta_{\mathsf{min}}, \frac{\sqrt{2}\log 2}{L})$. Let $\{X^k\}_k$ be the sequence of iterates generated by Algorithm 1 with stepsizes $\gamma_k = \frac{\Delta_k}{\sqrt{k+2}\log(k+2)}$ for $k \ge 0$. Then, for any $\lambda \in (0, \frac{\ell}{2(\tau+(2\ell-1)L)})$, we will have $\liminf_{k\to\infty}\sqrt[4]{k+2}\cdot\Theta_{\lambda f}^{\mathfrak{C}}(X^k) = 0$ $\mathbb{P}$-almost surely.*

Proof. Observe that $\sum_{k=0}^K \gamma_k^2 \le \gamma_0^2 + \int_0^K \frac{(\sqrt{2}\log 2/L)^2}{(u+2)\log(u+2)^2}\,du \le \gamma_0^2 + \frac{(\sqrt{2}\log 2/L)^2}{\log 2} < \frac{4}{L^2}$ and $\sum_{k=0}^K \gamma_k \ge \sum_{k=0}^K \frac{\Delta_{\mathsf{min}}}{\sqrt{K+2}\log(K+2)} = \frac{\Delta_{\mathsf{min}}(K+1)}{\sqrt{K+2}\log(K+2)} \to +\infty$ as $K \to +\infty$. Let $\Lambda_\delta = \{\omega \in \Lambda : \liminf_{k\to\infty}\sqrt{k+2}\cdot\Theta_{\lambda f}^{\mathfrak{C}}(X^k(\omega))^2 \ge \delta\}$ and assume to the contrary that $\mathbb{P}(\Lambda_\delta) > 0$ for some $\delta > 0$. Then, there exists a $\bar{k} \ge 0$ such that $\mathbb{P}(\bigcap_{k\ge\bar{k}}\Lambda_{\delta,k}) > 0$, where $\Lambda_{\delta,k} = \{\omega \in \Lambda : \sqrt{k+2}\cdot\Theta_{\lambda f}^{\mathfrak{C}}(X^k(\omega))^2 \ge \delta\}$. This implies that

$$\mathbb{P}\left(\left\{\omega \in \Lambda : \sum_{k\ge\bar{k}} \frac{\Theta_{\lambda f}^{\mathfrak{C}}(X^k(\omega))^2}{\sqrt{k+2}\,\log(k+2)} \ge \sum_{k\ge\bar{k}}\frac{\delta}{(k+2)\log(k+2)}\right\}\right) \ge \mathbb{P}\left(\bigcap_{k\ge\bar{k}}\Lambda_{\delta,k}\right) > 0.$$

Note that $\sum_{k\ge\bar{k}}\frac{1}{(k+2)\log(k+2)} \ge \int_{\bar{k}+3}^{\infty}\frac{du}{u\log u} = \infty$. However, as shown in the proof of Theorem 4.10, we will have $\sum_{k\ge 0}\gamma_k\Theta_{\lambda f}^{\mathfrak{C}}(X^k(\omega))^2 < \infty$ $\mathbb{P}$-almost surely. Thus, we obtain a contradiction. □

5 Numerical Results

In this section, we present numerical results to demonstrate the performance of our proposed RSSM on the dual principal component pursuit (DPCP) formulation of robust subspace recovery (RSR) and on orthogonal dictionary learning (ODL). We also compare the performance of RSSM with that of the Riemannian subgradient method (RSM) [LCD+21].

5.1 DPCP Formulation of Robust Subspace Recovery

The DPCP formulation of robust subspace recovery aims to identify a low-dimensional linear subspace for a dataset corrupted by outliers, which plays a fundamental role in machine learning, signal processing, and statistics. In robust subspace recovery, one is given some measurements $\widetilde{Y} = [\,Y\ O\,]\,\Gamma \in \mathbb{R}^{n\times m}$, where the columns of $Y \in \mathbb{R}^{n\times m_1}$ form inlier points spanning a d-dimensional subspace $\mathcal{S}$, the columns of $O \in \mathbb{R}^{n\times m_2}$ form outlier points with no linear structure, $\Gamma \in \mathbb{R}^{m\times m}$ is an unknown permutation, and $m = m_1 + m_2$. To recover the subspace $\mathcal{S}$ via obtaining an orthonormal basis for $\mathcal{S}^\perp$, one could adopt the *holistic approach* [DZVR21] and solve

$$\min_{X \in \mathbb{R}^{n \times (n-d)}} \ f(X) = \frac{1}{m} \sum_{j=1}^{m} \|\tilde{y}_j^\top X\| \quad \text{s.t.} \quad X \in \mathrm{St}(n, n-d). \tag{29}$$

In our experiments, we first generate a random matrix $S \in \mathrm{St}(n, d)$ with $p = n - d = 90$ and $n = 100$. We also generate $m_1 = 1500$ inlier points by $Y = SR$ with $R \in \mathrm{St}(d, m_1)$ and $m_2 = 3500$ outlier points $O \in \mathrm{St}(n, m_2)$ at random. For both RSM and RSSM, we use the same initialization $X^0 = \mathrm{argmin}_{X \in \mathrm{St}(n,p)} \|\widetilde{Y}^\top X\|^2$, where $\widetilde{Y} = [Y \ O]\Gamma$ for some random permutation matrix Γ. We partition the columns into 3, 5, or 10 blocks (i.e., $\ell = 3, 5, 10$) and adopt the stepsizes $\gamma_k = \frac{\Delta_k}{\sqrt{k+2}\log(k+2)}$ with $\Delta_k = 0.9$ for RSM and $\Delta_k = 0.9\binom{\ell}{2}^{2(0.991)^k - 1}$ for RSSM for $k \geq 0$.

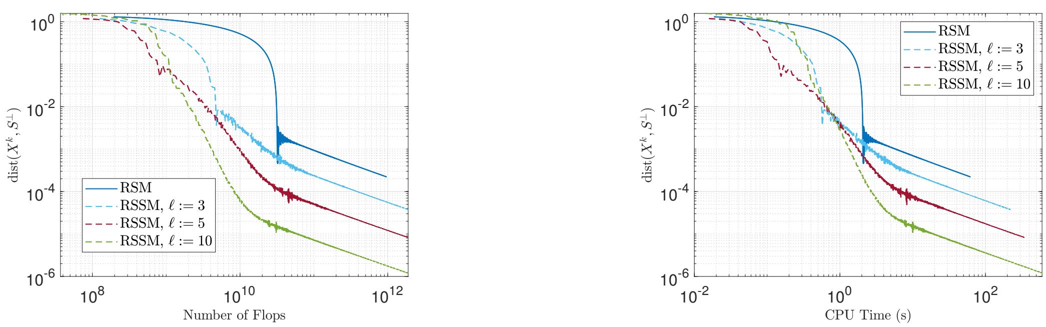

Figure 1: Robust subspace recovery: RSM vs RSSM ($\Delta_k = 0.9\binom{\ell}{2}^{2(0.991)^k - 1} \in [0.02, 40.5]$).

We measure the performance by the *estimation error*

$$\mathrm{dist}(X, S^\perp) = \min_{R \in \mathrm{O}(p)} \|X - S^\perp R\| = \sqrt{2(p - \|X^\top S^\perp\|_*)} = \sqrt{2(p - \|(I - SS^\top)X\|_*)},$$

where $\mathrm{O}(p) = \mathrm{St}(p, p)$. Figure 1 shows the log-log plots of estimation error against flops and CPU time. We observe that RSSM outperforms RSM in terms of both speed and accuracy.

5.2 Orthogonal Dictionary Learning

ODL aims to decompose a data matrix $Y \in \mathbb{R}^{n \times m}$ into the multiplication of an orthonormal dictionary and a sparse matrix, i.e., $Y = XS$ with $X \in \mathrm{St}(n, n)$ and each column of $S \in \mathbb{R}^{n \times m}$ being sparse. To recover the dictionary X, a natural optimization formulation is $\min_{X \in \mathrm{St}(n,n)} \|Y^\top X\|_1$.

We generate synthetic data as in [BJS19, LCD+21]. First, the true orthogonal dictionary $X^\star \in \mathrm{O}(n)$ with $n = 60$, is generated randomly according to the uniform distribution. The sample size is $m = 4648 \approx 10 \times n^{1.5}$. Then we generate a sparse matrix $S \in \mathbb{R}^{n \times m}$ with i.i.d. entries $S_{ij} \sim \mathcal{N}(0, 1)$ with probability 0.3 and $S_{ij} = 0$ otherwise. The observation is thus $Y = X^\star S$. Next, we generate a uniformly random orthogonal matrix X^0 on $\mathrm{St}(n, n)$ to initialize the algorithms. We partition the columns into 3, 5 or 10 blocks, i.e., $\ell = 3, 5, 10$, and adopt the stepsize $\gamma_k = \frac{\Delta_k}{\sqrt{k+2}\log(k+2)}$ with $\Delta_k = 10^{-3}$ for RSM and $\Delta_k = 10^{-3}\binom{\ell}{2}^{4(0.995)^k - 1}$ for RSSM.

To evaluate the performance of the algorithms, we define the error between X and $X^\star$ as $\mathrm{err}(X, X^\star) = \sum_{i=1}^{n} \left| \max_{1 \leq j \leq n} |x_i^\top x_j^\star| - 1 \right|$, where x_i and $x_j^\star$ denote the i-th and j-th columns of X and $X^\star$ respectively. Figure 2 shows the log-log plots of error against flops and CPU time. Once again, we observe that RSSM outperforms RSM in terms of both speed and accuracy.

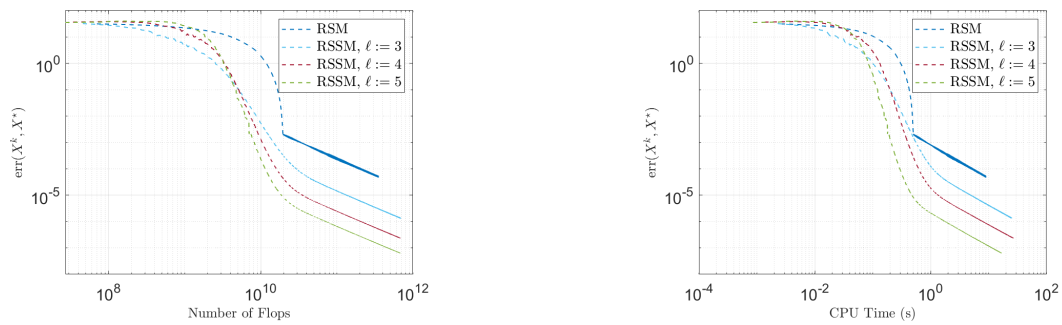

Figure 2: Orthogonal dictionary learning: RSM vs RSSM ($\Delta_k = 10^{-3}\binom{\ell}{2}^{4(0.995)^k-1}$).

6 Conclusion

In this paper, we proposed a new coordinate-type algorithm RSSM for solving nonsmooth weakly convex optimization problems over high dimensional Stiefel manifolds. The main idea of RSSM is that at each iteration, it decomposes the Stiefel manifold into submanifold blocks and performs a retracted partial Riemannian subgradient step with respect to a randomly selected submanifold block. RSSM enjoys a low per-iteration cost and is especially suitable for high-dimensional applications. Furthermore, we showed that RSSM converges to the set of stationary points at a sublinear rate. To the best of our knowledge, RSSM is the first feasible, coordinate-type algorithm for nonsmooth weakly convex optimization over Stiefel manifolds.

References

[AM12] P.-A. Absil and Jérôme Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012.

[AMS09] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. Optimization algorithms on matrix manifolds. In *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

[ANT16] Samir Adly, Florent Nacry, and Lionel Thibault. Preservation of prox-regularity of sets with applications to constrained optimization. *SIAM Journal on Optimization*, 26(1):448–473, 2016.

[Bil95] Patrick Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., third edition, 1995.

[BJS19] Yu Bai, Qijia Jiang, and Ju Sun. Subgradient descent learns orthogonal dictionaries. In *The 7th International Conference on Learning Representations*, 2019.

[Bou23] Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

[CMSZ24] Shixiang Chen, Shiqian Ma, Anthony Man-Cho So, and Tong Zhang. Nonsmooth optimization over the Stiefel manifold and beyond: Proximal gradient method and recent variants. *SIAM Review*, 66(2):319–352, 2024.

[CSW95] Francis H Clarke, Ronald J Stern, and Peter R Wolenski. Proximal smoothness and the lower-C^2 property. *Journal of Convex Analysis*, 2(1-2):117–144, 1995.

[CWYS24] Andy Yat-Ming Cheung, Jinxin Wang, Man-Chung Yue, and Anthony Man-Cho So. Randomized submanifold subgradient method for optimization over Stiefel manifolds. *arXiv preprint arXiv:2409.01770*, 2024.

[DR23] Yogesh Darmwal and Ketan Rajawat. Low-complexity subspace-descent over symmetric positive definite manifold. *arXiv preprint arXiv:2305.02041*, 2023.

[DZVR21] Tianyu Ding, Zhihui Zhu, René Vidal, and Daniel P Robinson. Dual principal component pursuit for robust subspace learning: Theory and algorithms for a holistic approach. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2739–2748, 2021.

[GD04] John C Gower and Garmt B Dijksterhuis. *Procrustes Problems*, volume 30 of *Oxford Statistical Science Series*. Oxford University Press, 2004.

[GHN22] David H Gutman and Nam Ho-Nguyen. Coordinate descent without coordinates: Tangent subspace descent on Riemannian manifolds. *Mathematics of Operations Research*, 48(1):127–159, 2022.

[GLY19] Bin Gao, Xin Liu, and Yaxiang Yuan. Parallelizable algorithms for optimization problems with orthogonality constraints. *SIAM Journal on Scientific Computing*, 41(3):A1949–A1983, 2019.

[GP10] Victor Guillemin and Alan Pollack. *Differential Topology*, volume 370. American Mathematical Society, 2010.

[GVL13] Gene H Golub and Charles F Van Loan. *Matrix Computations*. The Johns Hopkins University Press, fourth edition, 2013.

[HJM24] Andi Han, Pratik Jawanpuria, and Bamdev Mishra. Riemannian coordinate descent algorithms on matrix manifolds. In *Proceedings of the 41st International Conference on Machine Learning*, pages 17393–17415, 2024.

[HML21a] Minhui Huang, Shiqian Ma, and Lifeng Lai. Projection robust Wasserstein barycenters. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4456–4465, 2021.

[HML21b] Minhui Huang, Shiqian Ma, and Lifeng Lai. A Riemannian block coordinate descent method for computing the projection robust Wasserstein distance. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4446–4455, 2021.

[HPT25] Andi Han, Pierre-Louis Poirion, and Akiko Takeda. Efficient optimization with orthogonality constraint: a randomized Riemannian submanifold method. 2025.

[JZQ+22] Yunjiang Jiang, Han Zhang, Yiming Qiu, Yun Xiao, Bo Long, and Wen-Yun Yang. Givens coordinate descent methods for rotation matrix learning in trainable embedding indexes. In *The 10th International Conference on Learning Representations*, 2022.

[LCD+21] Xiao Li, Shixiang Chen, Zengde Deng, Qing Qu, Zhihui Zhu, and Anthony Man-Cho So. Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods. *SIAM Journal on Optimization*, 31(3):1605–1634, 2021.

[LJW+19] Shuai Li, Kui Jia, Yuxin Wen, Tongliang Liu, and Dacheng Tao. Orthogonal deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1352–1368, 2019.

[LSW19] Huikang Liu, Anthony Man-Cho So, and Weijie 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*, 178:215–262, 2019.

[LWW+15] Xin Liu, Zaiwen Wen, Xiao Wang, Michael Ulbrich, and Yaxiang Yuan. On the analysis of the discretized Kohn–Sham density functional theory. *SIAM Journal on Numerical Analysis*, 53(4):1758–1785, 2015.

[MA22] Estelle Massart and Vinayak Abrol. Coordinate descent on the orthogonal group for recurrent neural network training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7744–7751, 2022.

[Nes12] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[PV23] Liangzu Peng and René Vidal. Block coordinate descent on smooth manifolds. *arXiv preprint arXiv:2305.14744*, 2023.

[RS71] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In Jagdish S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233–257. Academic Press, 1971.

[RW09] R Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

[SC14] Uri Shalit and Gal Chechik. Coordinate-descent for learning orthogonal matrices through Givens rotations. In *Proceedings of the 31st International Conference on Machine Learning*, pages 548–556, 2014.

[SI13] Hiroyuki Sato and Toshihiro Iwai. A Riemannian optimization approach to the matrix singular value decomposition. *SIAM Journal on Optimization*, 23(1):188–212, 2013.

[TCA09] Fabian J Theis, Thomas P Cason, and P A Absil. Soft dimension reduction for ICA by joint diagonalization on the Stiefel manifold. In *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation*, pages 354–361, 2009.

[TKRH21] Yulun Tian, Kasra Khosoussi, David M Rosen, and Jonathan P How. Distributed certifiably correct pose-graph optimization. *IEEE Transactions on Robotics*, 37(6):2137–2156, 2021.

[Via83] Jean-Philippe Vial. Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 8(2):231–259, 1983.

[VTYN22] Hieu Vu, Toan Tran, Man-Chung Yue, and Viet Anh Nguyen. Distributionally robust fair principal components via geodesic descents. In *The 10th International Conference on Learning Representations*, 2022.

[WHC+23] Jinxin Wang, Jiang Hu, Shixiang Chen, Zengde Deng, and Anthony Man-Cho So. Decentralized weakly convex optimization over the Stiefel manifold. *arXiv preprint arXiv:2303.17779*, 2023.

[Wri15] Stephen J Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

[YZS14] Wei Hong Yang, Lei-Hong Zhang, and Ruyi Song. Optimality conditions for the nonlinear programming problems on Riemannian manifolds. *Pacific Journal of Optimization*, 10(2):415–434, 2014.

[ZCZL22] Lei Zhao, Ding Chen, Daoli Zhu, and Xiao Li. Randomized coordinate subgradient method for nonsmooth optimization. *arXiv preprint arXiv:2206.14981*, 2022.

[ZWR+18] Zhihui Zhu, Yifan Wang, Daniel Robinson, Daniel Naiman, Rene Vidal, and Manolis Tsakiris. Dual principal component pursuit: Improved analysis and efficient algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.

A Technical Lemmas

Lemma A.1. *Let $X \in \mathrm{St}(n,p)$ be fixed and $\mathfrak{C}$ be a partition of $[p]$ with $\ell = |\mathfrak{C}| \geq 2$. Then, for any $\xi \in \mathbb{R}^{n\times p}$, we have*

$$\mathcal{P}_{T_X \mathrm{St}(n,p)} \circ \mathcal{A}_X(\xi) = \mathcal{A}_X \circ \mathcal{P}_{T_X \mathrm{St}(n,p)}(\xi) = \frac{1}{\binom{\ell}{2}} \sum_{\{i,j\}\in\binom{[\ell]}{2}} \mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}}(\xi_{ij}) I_{ij}^\top, \tag{30}$$

$$\mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}}\left(\mathcal{A}_X(\xi) I_{ij}\right) I_{ij}^\top = \mathcal{A}_X\left(\mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}}(\xi_{ij}) I_{ij}^\top\right). \tag{31}$$

Proof. Let $\xi \in \mathbb{R}^{n\times p}$ be fixed. On one hand, a direct computation gives

$$\begin{aligned}
\mathcal{A}_X\left(\mathcal{P}_{T_X \mathrm{St}(n,p)}(\xi)\right) &= \frac{1}{\binom{\ell}{2}} \sum_{\{i,j\}\in\binom{[\ell]}{2}} (I - X_{-ij}X_{-ij}^\top)\left(X\,\mathrm{Skew}(X^\top\xi) + (I - XX^\top)\xi\right) I_{ij} I_{ij}^\top \\
&= \frac{1}{\binom{\ell}{2}} \sum_{\{i,j\}\in\binom{[\ell]}{2}} \left(X_{ij} I_{ij}^\top\,\mathrm{Skew}(X^\top\xi) I_{ij} + (I - XX^\top)\xi_{ij}\right) I_{ij}^\top \\
&= \frac{1}{\binom{\ell}{2}} \sum_{\{i,j\}\in\binom{[\ell]}{2}} \mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}}(\xi_{ij}) I_{ij}^\top.
\end{aligned}$$

On the other hand, using (16) and the fact that $XX^\top + X_\perp X_\perp^\top = I$, we have

$$\mathcal{A}_X(\xi) = X\left(\binom{\ell}{2}^{-1} Q \boxdot (X^\top\xi)\right) + (I - XX^\top)(\tfrac{2}{\ell}\,\xi).$$

This gives

$$\begin{aligned}
\mathcal{P}_{T_X \mathrm{St}(n,p)}\left(\mathcal{A}_X(\xi)\right) &= X\,\mathrm{Skew}(\binom{\ell}{2}^{-1} Q \boxdot (X^\top\xi)) + (I - XX^\top)(\tfrac{2}{\ell}\,\xi) \\
&= X\left(\binom{\ell}{2}^{-1} Q \boxdot \mathrm{Skew}(X^\top\xi)\right) + (I - XX^\top)(\tfrac{2}{\ell}\,\xi) = \mathcal{A}_X\left(\mathcal{P}_{T_X \mathrm{St}(n,p)}(\xi)\right).
\end{aligned}$$

Hence, (30) holds. Next, we compute

$$\mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}}\left(\mathcal{A}_X(\xi) I_{ij}\right) I_{ij}^\top = \mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}}\left(\left(X\left(\binom{\ell}{2}^{-1} Q \boxdot (X^\top\xi)\right) + X_\perp(\tfrac{2}{\ell} X_\perp^\top \xi)\right) I_{ij}\right) I_{ij}^\top$$

$$= X\left(\binom{\ell}{2}^{-1} Q \boxdot (I_{ij}\,\mathrm{Skew}(X_{ij}^\top \xi_{ij}) I_{ij}^\top)\right) + (I - XX^\top)(\tfrac{2}{\ell}\,\xi_{ij}) I_{ij}^\top,$$

$$\begin{aligned}\mathcal{A}_X\left(\mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}}(\xi_{ij}) I_{ij}^\top\right) &= \mathcal{A}_X\left(X(I_{ij}\,\mathrm{Skew}(X_{ij}^\top \xi_{ij}) I_{ij}^\top) + X_\perp(X_\perp^\top \xi_{ij} I_{ij}^\top)\right) \\ &= X\left(\binom{\ell}{2}^{-1} Q \boxdot (I_{ij}\,\mathrm{Skew}(X_{ij}^\top \xi_{ij}) I_{ij}^\top)\right) + (I - XX^\top)(\tfrac{2}{\ell}\,\xi_{ij}) I_{ij}^\top.\end{aligned}$$

It follows that (31) holds. □

Lemma A.2. *Let $X \in \mathrm{St}(n,p)$ be fixed and $\mathfrak{C}$ be a partition of $[p]$ with $\ell = |\mathfrak{C}| \geq 2$. The averaging operator $\mathcal{A}_X$ is self-adjoint and has eigenvalues $\frac{2}{\ell}$ and $\binom{\ell}{2}^{-1}$ with multiplicities $(n-p)p + \sum_{i=1}^{\ell} p_i^2$ and $p^2 - \sum_{i=1}^{\ell} p_i^2$, respectively.*

Proof. Fix any $X_\perp \in \mathrm{St}(n, n-p)$ such that $[X\ X_\perp] \in \mathrm{St}(n,n)$. Then, for any $\xi, \eta \in \mathbb{R}^{n\times p}$, we have

$$\begin{aligned}\langle \mathcal{A}_X(\xi), \eta \rangle &= \left\langle X\left(\binom{\ell}{2}^{-1} Q \boxdot (X^\top \xi)\right) + X_\perp(\tfrac{2}{\ell}\, X_\perp^\top \xi), X(X^\top \eta) + X_\perp(X_\perp^\top \eta) \right\rangle \\ &= \left\langle \binom{\ell}{2}^{-1} Q \boxdot (X^\top \xi), X^\top \eta \right\rangle + \left\langle \tfrac{2}{\ell}\, X_\perp^\top \xi, X_\perp^\top \eta \right\rangle \\ &= \left\langle X^\top \xi, \binom{\ell}{2}^{-1} Q \boxdot (X^\top \eta) \right\rangle + \left\langle X_\perp^\top \xi, \tfrac{2}{\ell}\, X_\perp^\top \eta \right\rangle = \langle \xi, \mathcal{A}_X(\eta) \rangle,\end{aligned}$$

i.e., $\mathcal{A}_X$ is self-adjoint. In addition, the following subspaces

$$\left\{ X \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & A_{\ell\ell} \end{bmatrix} + X_\perp B \,:\, A_{ii} \in \mathbb{R}^{p_i \times p_i}, B \in \mathbb{R}^{(n-p)\times p} \right\},$$

$$\left\{ X \begin{bmatrix} 0 & A_{12} & \cdots & A_{1\ell} \\ A_{21} & 0 & & A_{2\ell} \\ \vdots & & \ddots & \vdots \\ A_{\ell 1} & A_{\ell 2} & \cdots & 0 \end{bmatrix} \,:\, A_{ij} \in \mathbb{R}^{p_i \times p_j} \right\}$$

are the eigenspaces of $\mathcal{A}_X$ corresponding to the eigenvalues $\frac{2}{\ell}$ and $\binom{\ell}{2}^{-1}$, respectively. □

Lemma A.3. *Let $X, Y \in \mathrm{St}(n,p)$ be fixed. For any $\xi \in \mathbb{R}^{n\times p}$, we have*

$$\mathcal{A}_Y^{-1}(\xi) - \mathcal{A}_X^{-1}(\xi) = \tfrac{\ell(\ell-2)}{2}\left(Y\left[(J-I)\boxdot(Y^\top \xi)\right] - X\left[(J-I)\boxdot(X^\top \xi)\right]\right). \tag{32}$$

Proof. Let $Q' = J - \frac{\ell-2}{\ell-1} I$ denote the entrywise reciprocal of $Q = J + (\ell-2)I$. It can be verified that

$$\mathcal{A}_X^{-1}(\xi) = X\left(\binom{\ell}{2} Q' \boxdot (X^\top \xi)\right) - (I - XX^\top)(\tfrac{\ell}{2}\,\xi)$$

for all $\xi \in \mathbb{R}^{n\times p}$. It follows that

$$\begin{aligned}\mathcal{A}_Y^{-1}(\xi) - \mathcal{A}_X^{-1}(\xi) &= Y\left(\binom{\ell}{2} Q' \boxdot (Y^\top \xi)\right) + (I - YY^\top)(\tfrac{\ell}{2}\,\xi) - X\left(\binom{\ell}{2} Q' \boxdot (X^\top \xi)\right) - (I - XX^\top)(\tfrac{\ell}{2}\,\xi) \\ &= Y\left(\binom{\ell}{2} Q' \boxdot (Y^\top \xi)\right) - Y(\tfrac{\ell}{2}\, Y^\top \xi) - X\left(\binom{\ell}{2} Q' \boxdot (X^\top \xi)\right) + X(\tfrac{\ell}{2}\, X^\top \xi) \\ &= \tfrac{\ell}{2}\left(Y\left[((\ell-1)Q' - J)\boxdot(Y^\top \xi)\right] - X\left[((\ell-1)Q' - J)\boxdot(X^\top \xi)\right]\right) \\ &= \tfrac{\ell(\ell-2)}{2}\left(Y\left[(J-I)\boxdot(Y^\top \xi)\right] - X\left[(J-I)\boxdot(X^\top \xi)\right]\right),\end{aligned}$$

which completes the proof. □

Lemma A.4. *Consider the setting of Lemma 4.3. For any $\xi, \zeta \in \mathrm{St}(n,p)$, we have*

$$\left| \left(\|\xi\|^2_{\mathcal{A}^{-1}_{X^+}} - \|\xi\|^2_{\mathcal{A}^{-1}_{X}} \right) - \left(\|\zeta\|^2_{\mathcal{A}^{-1}_{X^+}} - \|\zeta\|^2_{\mathcal{A}^{-1}_{X}} \right) \right| \leq \tfrac{\ell(\ell-2)}{\sqrt{2}} \left\| X^+ - X \right\| \cdot \left\| \xi - \zeta \right\|. \tag{33}$$

Proof. Recall that $\Psi_X(\xi) = X\left[(J-I) \boxdot (X^\top \xi)\right]$. By (32) and (20), we have

$$\begin{aligned}
&\left(\|\xi\|^2_{\mathcal{A}^{-1}_{X^+}} - \|\xi\|^2_{\mathcal{A}^{-1}_{X}} \right) - \left(\|\zeta\|^2_{\mathcal{A}^{-1}_{X^+}} - \|\zeta\|^2_{\mathcal{A}^{-1}_{X}} \right) \\
&= \tfrac{\ell(\ell-2)}{2} \left(\langle (\Psi_{X^+} - \Psi_X)(\xi), \xi \rangle - \langle (\Psi_{X^+} - \Psi_X)(\zeta), \zeta \rangle \right) \\
&= \tfrac{\ell(\ell-2)}{2} \Big(\langle (X_i^+ X_i^{+\top} - X_i X_i^\top)\xi_{-i} I_{-i}^\top + (X_j^+ X_j^{+\top} - X_j X_j^\top)\xi_{-j} I_{-j}^\top, \xi \rangle \\
&\qquad\qquad - \langle (X_i^+ X_i^{+\top} - X_i X_i^\top)\zeta_{-i} I_{-i}^\top + (X_j^+ X_j^{+\top} - X_j X_j^\top)\zeta_{-j} I_{-j}^\top, \zeta \rangle \Big) \\
&= \tfrac{\ell(\ell-2)}{2} \Big(\langle X_i^+ X_i^{+\top} - X_i X_i^\top, \xi_{-i}\xi_{-i}^\top - \zeta_{-i}\zeta_{-i}^\top \rangle + \langle X_j^+ X_j^{+\top} - X_j X_j^\top, \xi_{-j}\xi_{-j}^\top - \zeta_{-j}\zeta_{-j}^\top \rangle \Big) \\
&\leq \tfrac{\ell(\ell-2)}{2} \Big(\|X_i^+ X_i^{+\top} - X_i X_i^\top\|_{\mathsf{op}} \cdot \|\xi_{-i}\xi_{-i}^\top - \zeta_{-i}\zeta_{-i}^\top\| + \|X_j^+ X_j^{+\top} - X_j X_j^\top\|_{\mathsf{op}} \cdot \|\xi_{-j}\xi_{-j}^\top - \zeta_{-j}\zeta_{-j}^\top\| \Big) \\
&= \tfrac{\ell(\ell-2)}{2} \Big(\|(I - X_i X_i^\top) X_i^+\|_{\mathsf{op}} \cdot \|(I - \zeta_{-i}\zeta_{-i}^\top)\xi_{-i}\| + \|(I - X_j X_j^\top) X_j^+\|_{\mathsf{op}} \cdot \|(I - \zeta_{-j}\zeta_{-j}^\top)\xi_{-j}\| \Big) \\
&\leq \tfrac{\ell(\ell-2)}{2} \Big(\|X_i^+ - X_i\|_{\mathsf{op}} \cdot \|\xi_{-i} - \zeta_{-i}\| + \|X_j^+ - X_j\|_{\mathsf{op}} \cdot \|\xi_{-j} - \zeta_{-j}\| \Big) \\
&\leq \tfrac{\ell(\ell-2)}{\sqrt{2}} \|X_{ij}^+ - X_{ij}\| \cdot \|\xi - \zeta\|,
\end{aligned}$$

where the fourth equality follows from Lemma 4.4. This completes the proof. ☐

Lemma A.5 (Convergence Lemma; see [ZCZL22, Lemma 4.1])**.** *Let $\{Y^k\} \subseteq \mathbb{R}^{n\times p}$ and $\{\mu_k\} \subseteq \mathbb{R}_+$ be given sequences. Let $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ be a map and $L_\Phi > 0$ be a constant such that $\|\Phi(Y^k) - \Phi(Y^{k+1})\| \leq L_\Phi \|Y^k - Y^{k+1}\|$ for all $k \geq 0$. Suppose there exist constants $M, p > 0$ and a vector $\bar{\Phi} \in \mathbb{R}^m$ such that*

- $\|Y^k - Y^{k+1}\| \leq M\mu_k$ *for all $k \geq 0$,*
- $\sum_{k\geq 0} \mu_k = \infty$,
- $\sum_{k\geq 0} \mu_k \|\Phi(Y^k) - \bar{\Phi}\|^p < \infty$.

Then, we have $\lim_{k\to\infty} \|\Phi(Y^k) - \bar{\Phi}\|^p = 0$.