Randomized Submanifold Subgradient Method for Optimization over Stiefel Manifolds

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

September 3, 2024

Abstract

Optimization over Stiefel manifolds has found wide applications in many scientific and engineering domains. Despite considerable research effort, high-dimensional optimization problems over Stiefel manifolds remain challenging, and the situation is exacerbated by nonsmooth objective functions. The purpose of this paper is to propose and study a novel coordinate-type algorithm for weakly convex (possibly nonsmooth) optimization problems over high-dimensional Stiefel manifolds, named randomized submanifold subgradient method (RSSM). 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 converges to the set of stationary points and attains ε-stationary points with respect to a natural stationarity measure in $\mathcal{O}(\varepsilon^{-4})$ iterations in both expectation and the almost-sure senses. To the best of our knowledge, these are the first convergence guarantees for coordinate-type algorithms to optimize nonconvex nonsmooth functions over Stiefel manifolds. An important technical tool in our convergence analysis is a new *Riemannian subgradient inequality* for weakly convex function on proximally smooth matrix manifolds, which could be of independent interest.

1 Introduction

In this paper, we consider the following constrained minimization problem:

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

where $n \geq p$ are positive integers, $\mathrm{St}(n,p) = \{X \in \mathbb{R}^{n\times p} : X^\top X = I_p\}$ is the Stiefel manifold, and $f : \mathbb{R}^{n\times p} \to \mathbb{R}$ is a function that is weakly convex on some convex neighborhood containing the Stiefel manifold $\mathrm{St}(n,p)$ but possibly nonsmooth. Recall that a function f is said to be τ-*weakly convex* on a convex set $\Omega \subseteq \mathbb{R}^{n\times p}$ if $f(\cdot)+\frac{\tau}{2}\|\cdot\|^2$ is convex on Ω. Problem (1) has attracted great attention from both optimization and machine learning communities due to its broad range of applications, including

*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) .

†Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong (wangjinxin68@gmail.com).

‡Department of Industrial and Manufacturing 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).

the orthogonal Procrustes problem [GD04, SI13], the joint diagonalization problem [TCA09], Kohn-Sham total energy minimization [LWW+15], robust subspace recovery [ZWR+18], and orthogonal dictionary learning [LCD+21]. For more details on Stiefel manifolds and the associated optimization problems, see [AMS09, HLWY20] and the references therein.

In modern applications, the dimension of the Stiefel manifold could be very high, i.e., n or p is a large number. For example, as pointed out in [LJW+19], deep neural networks attain optimal generalization error when the weight matrix has orthonormal columns. Constraining the columns of the weight matrix to be orthonormal in the training of deep neural networks naturally constitutes a high-dimensional instance of optimization over Stiefel manifolds. Unfortunately, despite considerable research effort, existing algorithms for optimization over Stiefel manifolds do not scale well and are only suitable for small- to medium-scale problems. When the problem is further complicated by a nonsmooth objective function, the situation becomes even more challenging. The main goal of this paper is to fill this gap in the literature and develop an efficient algorithm for nonsmooth optimization over high-dimensional Stiefel manifolds.

In the Euclidean setting, coordinate-type algorithms are a classical approach to tackling high-dimensional optimization problems and have shown promising performance in many applications [Nes12, Wri15, STXY16]. A natural idea for achieving our goal is therefore extending coordinate-type algorithms to Stiefel manifolds. This is not straightforward though. The development of coordinate-type algorithms relies critically on the separability of the feasible region, which enables that the feasibility can be maintained via simple operations. Such a separability condition is valid in the case of unconstrained or linearly constrained optimization problems. For Stiefel manifolds, the variables are coupled all together in a complicated, nonlinear manner, rendering the extension highly nontrivial.

Extending coordinate-type algorithms to manifold optimization is not entirely new and has been studied in a number of papers. When the manifold is a product manifold, coordinate-type algorithms can naturally be developed due to the built-in separability of the constraint; see, e.g., [HML21b, HML21a, TKRH21, PV23]. Without this product structure, the development of coordinate-type algorithms for manifold optimization is limited. In [GHN22], a coordinate-type algorithm, called the tangent subspace descent (TSD), for general manifolds has been developed, which deals with the coupling issue by decomposing the tangent space into lower-dimensional subspaces and updating the iterate along a chosen subspace at each iteration using the exponential map. Using a similar strategy as in [GHN22] to deal with the coupling issue, the recent paper [DR23] developed a coordinate-type algorithm for optimization over the manifold of positive definite matrices. Another strategy to tackle the coupling issue is to penalize the manifold constraint to the objective, essentially transforming the manifold optimization problem to an Euclidean optimization problem. Such an idea has been adopted in [GLY19] to develop coordinate-type algorithms for optimization over Stiefel manifolds. However, this leads to an infeasible algorithm, which might not be ideal in some applications. We should also point out that all these works however consider only smooth objective functions.

Our contributions can be summarized as follows.

- We devise a new coordinate-type algorithm, called the randomized submanifold subgradient method (RSSM), for solving nonsmooth weakly convex minimization over Stiefel manifolds. A key novel aspect of RSSM lies in the way it deals with coupling issue: instead of performing coordinate updates with respect to the intrinsic coordinates by decomposing the tangent space as in [GHN22], 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 restricted to a randomly chosen submanifold block.

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

- We theoretically analyze the convergence behaviour of RSSM. Specifically, we prove that our method converges to the set of stationary points and attains the iteration complexity $\mathcal{O}(\varepsilon^{-4})$ with respect to a natural stationarity measure in both expectation and the almost-sure senses.
- An important theoretical construct in our analysis is a positive definite operator on $\mathbb{R}^{n\times p}$, called the scaling operator. The scaling operator permeates the proofs via its associated metric. Remarkably, it allows us to derive a new Riemannian subgradient inequality for a weakly convex function on a proximally smooth manifold, which in turn plays a fundamental role in our convergence results and could be of independent interest.

We end the introduction by presenting a detailed comparison on coordinate-type algorithms over Stiefel manifolds, see Table 1.

Table 1: Coordinate-type methods in the Stiefel manifold setting.

References & Methods	**Coordinate**	**Objective Function**	**Columns**	**Feasible**	**Retraction**
[GHN22] (TSD)	intrinsic	C^1	multiple	✔	exponential map
[GLY19] (PCAL)	ambient	C^1 (C^2 in some theoretical results)	multiple	✗	N/A
Our work (**RSSM**)	ambient	weakly convex and nonsmooth	multiple	✔	polar decomposition-based retraction (9)

2 Notation and Preliminaries

For $\xi, \eta \in \mathbb{R}^{n\times p}$, we denote by $\langle \xi, \eta \rangle := \operatorname{tr}(\xi^\top \eta)$ the *Frobenius inner product* and by $\|\xi\| := \sqrt{\langle \xi, \xi \rangle}$ its induced norm. The *operator norm* is denoted by $\|\xi\|_{\mathsf{op}} := \sup_{v\in\mathbb{R}^p,\, \|v\|=1} \|\xi v\|$. The *nuclear norm* is denoted by $\|\xi\|_* := \operatorname{tr}((\xi^\top \xi)^{1/2})$. Given a self-adjoint positive definite linear operator $\mathcal{D} : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$, $\langle \xi, \eta \rangle_{\mathcal{D}}$ and $\|\xi\|_{\mathcal{D}}$ denote the *Mahalanobis inner product* and *norm* respectively, i.e., $\langle \xi, \eta \rangle_{\mathcal{D}} = \langle \mathcal{D}(\xi), \eta \rangle$ and $\|\xi\|_{\mathcal{D}} := \langle \mathcal{D}(\xi), \xi \rangle$. The symbols I and 0 represent the identity and zero matrices respectively. For a subset $C \subseteq [p] := \{1, \ldots, p\}$ and a matrix $X \in \mathbb{R}^{n\times p}$, $X_C \in \mathbb{R}^{n\times |C|}$ denotes the submatrix obtained by extracting all columns of X corresponding to indices in C. For $\ell \geq 2$, we let $\binom{[\ell]}{2}$ be the collection of 2-sets or unordered pairs in $[\ell]$, i.e., $\binom{[\ell]}{2} := \{\, \{i, j\} : i, j \in [\ell] \text{ and } i \neq j \,\}$. The *distance function* from a closed subset $\mathcal{S} \subseteq \mathbb{R}^{n\times p}$ is $\operatorname{dist}(X, \mathcal{S}) := \inf_{Y\in\mathcal{S}} \|X - Y\|$.

2.1 Stiefel Manifolds as Proximally Smooth Sets

Definition 2.1. *A closed set $\mathcal{S} \subseteq \mathbb{R}^{n\times p}$ is R-proximally smooth if the (nearest-point) projection mapping $\mathcal{P}_{\mathcal{S}}(X) := \operatorname{argmin}_{Y\in\mathcal{S}} \|X - Y\|$ is well-defined and single-valued whenever* $\operatorname{dist}(X, \mathcal{S}) < R$.

The proximally smooth set is an important class of nonconvex sets with wide applications. It has been proved in [BT22, Proposition 1] that the Stiefel manifold $\operatorname{St}(n, p)$ is 1-proximally smooth. Indeed, all convex sets, sublevel sets of weakly convex functions, and compact C^2 embedded submanifolds are proximally smooth (see [CSW95]). Besides, following [ANT16, Theorem 2.2], a submanifold

$\mathcal{M}$ is R-proximally smooth if and only if

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

for all $X, X' \in \mathcal{M}$ and $\zeta \in N_X\mathcal{M}$. Here, $N_X\mathcal{M}$ denotes the *normal space* of $\mathcal{M}$ at X.

2.2 Euclidean Clarke Subdifferential and Clarke Regularity

Definition 2.2 (Generalized Directional Derivative and Euclidean Clarke Subdifferential)**.** *[Cla90, §2.1] Let f be Lipschitz near $X \in \mathbb{R}^{n\times p}$, and let $V \in \mathbb{R}^{n\times p}$. The* generalized directional derivative *of f at X in the direction V is defined as*

$$f^\circ(X; V) := \limsup_{\substack{Y \to X \\ t \downarrow 0}} \frac{f(Y + tV) - f(Y)}{t}, \tag{3}$$

and the Euclidean Clarke subdifferential *or* generalized gradient *of f at X is defined as*

$$\partial f(X) := \{\, \zeta \in \mathbb{R}^{n\times p} : \langle \zeta, V \rangle \le f^\circ(X; V) \text{ for all } V \in \mathbb{R}^{n\times p} \,\}. \tag{4}$$

Definition 2.3 (Clarke Regularity)**.** *Let f be Lipschitz near $X \in \mathbb{R}^{n\times p}$. We say f is* Clarke regular *at X if, for every $V \in \mathbb{R}^{n\times p}$, the one-sided directional derivative*

$$f'(X; V) := \lim_{t \downarrow 0} \frac{f(X + tV) - f(X)}{t} \tag{5}$$

exists and is equal to $f^\circ(X; V)$. Moreover, f is Clarke regular *if f is Clarke regular at every X.*

2.3 Weakly Convex Optimization on Stiefel Manifolds

Given a τ-weakly convex function h, which must be Clarke regular [Via83, Proposition 4.5], its (Clarke) subdifferential at $X \in \mathbb{R}^{n\times p}$ can be reduced to $\partial h(X) := \partial g(X) - \tau X$, where $\partial g(X)$ is the convex subdifferential of g at $X \in \mathbb{R}^{n\times p}$; see [Via83, Proposition 4.6]. Moreover, if h is τ-weakly convex on $\mathbb{R}^{n\times p}$, we have

$$h(Y) \ge h(X) + \langle \widetilde{\nabla} h(X), Y - X \rangle - \frac{\tau}{2}\|Y - X\|^2 \tag{6}$$

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

$$\partial_{\mathcal{M}} h(X) := \mathcal{P}_{T_X\mathcal{M}}(\partial h(X)) := \{\, \mathcal{P}_{T_X\mathcal{M}}(\widetilde{\nabla} h(X)) : \widetilde{\nabla} h(X) \in \partial h(X) \,\}, \tag{7}$$

where $\mathcal{P}_{T_X\mathcal{M}}(\xi)$ denotes the orthogonal projection of ξ onto $T_X\mathcal{M}$; see [YZS14, Theorem 5.1]. 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, it is well known that $T_X\mathrm{St}(n,p) = \{\, \xi \in \mathbb{R}^{n\times p} : \xi^\top X + X^\top \xi = 0 \,\}$ and

$$\mathcal{P}_{T_X\mathrm{St}(n,p)}(\xi) := \xi - X\mathrm{sym}(X^\top \xi) = \xi - \tfrac{1}{2} X(X^\top \xi + \xi^\top X); \tag{8}$$

see [AMS09, Example 3.6.2].

With (7), we call a point $X \in \mathcal{M}$ satisfying $0 \in \partial_{\mathcal{M}} f(X)$ a *stationary point* of the weakly convex function f over $\mathcal{M}$.

To search for the next iterate along feasible curves at the current point X on an embedded manifold $\mathcal{M}$, the notion of *local retraction* was introduced in [AMS09, Definition 4.1.1], which is a differentiable mapping from $T_X\mathcal{M}$ onto $\mathcal{M}$ that locally approximates the exponential map on the manifold $\mathcal{M}$ up to first order. Given $\xi \in T_X\mathcal{M}$, we denote $\mathrm{Retr}_X(\xi)$ the local retraction at $X \in \mathcal{M}$ along the direction $\xi \in T_X\mathcal{M}$. In what follows, we would focus on polar decomposition-based retraction for the Stiefel manifold, i.e.,

$$\mathrm{Retr}_X(\xi) := (X + \xi)(I + \xi^\top \xi)^{-1/2} \tag{9}$$

for $X \in \mathrm{St}(n,p)$ and $\xi \in T_X\mathrm{St}(n,p)$. We note that the polar decomposition-based retraction $\mathrm{Retr}_X(\xi)$ gives the unique projection of $X+\xi$ onto $\mathrm{St}(n,p)$, i.e., $\mathrm{Retr}_X(\xi) := \mathcal{P}_{\mathrm{St}(n,p)}(X+\xi)$. Moreover, this projection $\mathcal{P}_{\mathrm{St}(n,p)}$ satisfies a Lipschitz-like property

$$\| \mathrm{Retr}_X(\xi) - Y \| = \| \mathcal{P}_{\mathrm{St}(n,p)}(X+\xi) - \mathcal{P}_{\mathrm{St}(n,p)}(Y) \| \le \| X + \xi - Y \| \tag{10}$$

for all $X, Y \in \mathrm{St}(n,p)$ and $\xi \in T_X\mathrm{St}(n,p)$; see [LCD+21, Lemma 1]. From [LSW19, Appendix E.1], $\xi \mapsto \mathcal{P}_{\mathrm{St}(n,p)}(X+\xi)$ also satisfies a second-order boundedness condition, i.e.,

$$\| \mathcal{P}_{\mathrm{St}(n,p)}(X+\xi) - X - \xi \| \le \|\xi\|^2 \tag{11}$$

for all $X \in \mathrm{St}(n,p)$ and $\xi \in T_X\mathrm{St}(n,p)$ with $\|\xi\| \le 1$.

When $\mathcal{M}$ is a proximally smooth manifold, we can define a local retraction in a similar fashion.

Lemma 2.4. *Let $\mathcal{M}$ be a proximally smooth manifold in $\mathbb{R}^{n\times p}$. Then, $\xi \mapsto \mathcal{P}_{\mathcal{M}}(X+\xi)$ is a local retraction from a neighborhood $\mathcal{U}$ of the origin of $T_X\mathcal{M}$ onto $\mathcal{M}$.*

Proof. Let $\mathcal{M}$ be R-proximally smooth. Suppose that $\xi \in \mathbb{R}^{n\times p}$ satisfies $\|\xi\| < R$. Then, we have $\mathrm{dist}(X+\xi, \mathcal{M}) \le \| X + \xi - X \| = \|\xi\| < R$, which implies $\mathcal{P}_{\mathcal{M}}$ is well-defined (i.e. single-valued). By taking $\mathcal{U} \subsetneq \{\xi \in T_X\mathcal{M} : \|\xi\| < R\}$, we can write $\mathcal{R}_X(\xi) := \mathcal{P}_{\mathcal{M}}(X+\xi)$ from $\mathcal{U} \subseteq T_X\mathcal{M}$ to $\mathcal{M}$. It is easy to see that $\mathcal{R}_X(0) = \mathcal{P}_{\mathcal{M}}(X) = X$ and $D\mathcal{R}_X(0)[\xi] := D\mathcal{P}_{\mathcal{M}}(X)[\xi] = \mathcal{P}_{T_X\mathcal{M}}(\xi) = \xi$ by [LM08, Lemma 4], so $\mathcal{R}_X$ is a local retraction from $\mathcal{U} \subseteq T_X\mathcal{M}$ to $\mathcal{M}$. □

3 Randomized Submanifold Subgradient Method

In this section, we present RSSM in Algorithm 1 for solving problem (1), which extends the idea of coordinate-type methods in the Euclidean setting to the Stiefel manifold setting. Our algorithm updates the iterates via a partial Riemannian subgradient oracle and a low-cost retraction onto a submanifold block, which will be defined appropriately in §3.1.

3.1 Submanifold Block

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

$$\mathcal{M}_{X_{[p]\setminus C}} := \{ Y \in \mathrm{St}(n, |C|) \; : \; X_{[p]\setminus C}^\top Y = 0 \}, \tag{12}$$

and the tangent space of the submanifold block at $X_C \in \mathcal{M}_{X_{[p]\setminus C}}$ is

$$T_{X_C}\mathcal{M}_{X_{[p]\setminus C}} := \{ \xi \in T_{X_C}\mathrm{St}(n,|C|) \; : \; X_{[p]\setminus C}^\top \xi = 0 \}. \tag{13}$$

By employing the submanifold block concept, we only allow the columns of X corresponding to indices in C to vary at one time, while keeping all other columns in $[p] \setminus C$ fixed, thereby preserving the feasibility. In other words, we extend the idea of a coordinate block in the Euclidean space to a submanifold block in the Stiefel manifold setting. It is easy to verify that $Y \in \mathcal{M}_{X_{[p]\setminus C}}$ if and only if $Y I_C^\top + X_{[p]\setminus C} I_{[p]\setminus C}^\top \in \mathrm{St}(n,p)$. This means searching along a submanifold block via, e.g., a retracted Riemannian subgradient step, preserves the feasibility. The tangent space (13) of the submanifold block (12) in Definition 3.1 contains the search direction.

Lemma 3.2 (Proximal Smoothness of Submanifold Block and Its Projection)**.** *Let $C \subseteq [p]$ and $X \in \mathrm{St}(n,p)$. For $\Xi \in \mathbb{R}^{n\times|C|}$ with $\mathrm{dist}(\Xi, \mathcal{M}_{X_{[p]\setminus C}}) < 1$, the projection is defined as*

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

This shows $\mathcal{M}_{X_{[p]\setminus C}}$ is 1-proximally smooth in $\mathbb{R}^{n\times|C|}$. If furthermore $X_{[p]\setminus C}^\top\Xi = 0$, then the projection $\mathcal{P}_{\mathcal{M}_{X_{[p]\setminus C}}}(\Xi) = \mathcal{P}_{\mathrm{St}(n,|C|)}(\Xi)$.

Proof. Let $\Xi \in \mathbb{R}^{n\times|C|}$ with $\mathrm{dist}(\Xi, \mathcal{M}_{X_{[p]\setminus C}}) < 1$. We can write $\Xi = X_{[p]\setminus C}X_{[p]\setminus C}^\top\Xi + (I - X_{[p]\setminus C}X_{[p]\setminus C}^\top)\Xi$ and

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

for all $U \in \mathcal{M}_{X_{[p]\setminus C}}$. 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 in $\mathbb{R}^{n\times|C|}$, we obtain

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

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

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

Therefore,

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

Thus, $\mathcal{M}_{X_{[p]\setminus C}}$ is 1-proximally smooth in $\mathbb{R}^{n\times|C|}$. Furthermore, if $X_{[p]\setminus C}^\top\Xi = 0$, then we have $\mathcal{P}_{\mathcal{M}_{X_{[p]\setminus C}}}(\Xi) = \mathcal{P}_{\mathrm{St}(n,|C|)}(\Xi)$. □

3.2 Partial Riemannian Subgradient Oracle

Let f be a τ-weakly convex and L-Lipschitz continuous function on some convex neighborhood of $\mathrm{St}(n,p)$ in $\mathbb{R}^{n\times p}$. Let $X \in \mathbb{R}^{n\times p}$ be fixed. With respect to a subset $C \subseteq [p]$, we can define a mapping $\Phi_C : \mathbb{R}^{n\times |C|} \to \mathbb{R}^{n\times p}$ by $\Phi_C(Y) := YI_C^\top + X_{[p]\setminus C} I_{[p]\setminus C}^\top$. Following §2.2 or [Cla90, §2.3], we can define the *partial Euclidean (Clarke) subdifferential* of f with respect to C as

$$\partial^C f(X) := \partial(f\circ\Phi_C)(X_C) \subseteq \mathbb{R}^{n\times|C|} \tag{14}$$

for $X \in \mathbb{R}^{n\times p}$. A *partial Euclidean (Clarke) subgradient* of f with respect to C can be written as $\widetilde{\nabla}^C f(X) := \widetilde{\nabla}(f\circ\Phi_C)(X_C) \in \partial^C f(X)$.

Proposition 3.3 (Relationship between Partial and Full Euclidean Subdifferentials)**.** *Let f be a Clarke regular function on $\mathbb{R}^{n\times p}$, $X \in \mathbb{R}^{n\times p}$. Then, for every $C \subseteq [p]$, it holds that $\partial^C f(X) = \partial f(X) I_C$, i.e., for every $\widetilde{\nabla}^C f(X) \in \partial^C f(X)$, there exists $\widetilde{\nabla} f(X) \in \partial f(X)$ such that $\widetilde{\nabla}^C f(X) = \widetilde{\nabla} f(X) I_C$.*

Proof. By definition of (14), $\partial^C f(X) = \partial(f\circ\Phi_C)(X_C)$. Since Φ_C is affine, the strict derivative of Φ_C at X_C is $D_s\Phi_C(X_C)[V] = VI_C^\top$ for $V \in \mathbb{R}^{n\times|C|}$; see [Cla90, §2.2]. By subdifferential chain rule [Cla90, Theorem 2.3.10] and since f is Clarke regular at $X = \Phi_C(X_C)$, we get $\partial^C f(X) := \partial(f\circ\Phi_C)(X_C) = D_s\Phi_C(X_C)^*\partial f(X) = \partial f(X) I_C$. □

Let $\mathfrak{C} := \{C_1,\dots,C_\ell\}$ be a partition of $[p]$ with $\ell \geq 2$, i.e., $C_i \cap C_j = \varnothing$ and $\bigcup_{i=1}^\ell C_i = [p]$. For the sake of convenience, in what follows we denote $p_i := |C_i|$, $X_i := X_{C_i}$, $X_{-i} := X_{[p]\setminus C_i}$ for $i \in [\ell]$, and $C_{ij} := C_i \cup C_j$, $p_{ij} := |C_{ij}|$, $X_{ij} := X_{C_{ij}}$, $X_{-ij} := X_{[p]\setminus C_{ij}}$ for $\{i,j\} \in \binom{[\ell]}{2}$.

Given a point $X \in \mathrm{St}(n,p)$ and $C_{ij} \subseteq [p]$ for an unordered block index pair $\{i,j\} \in \binom{[\ell]}{2}$, we can obtain a submanifold block $\mathcal{M}_{X_{-ij}} := \{Y \in \mathrm{St}(n,p_{ij}) : X_{-ij}^\top Y = 0\} \subseteq \mathrm{St}(n,p_{ij})$. Following §2.3, we can define the *partial Riemannian (Clarke) subdifferential* of f with respect to C_{ij} as $\partial_{ij} f(X) := \partial_{\mathcal{M}_{X_{-ij}}}(f\circ\Phi_{C_{ij}})(X_{ij})$ for $X \in \mathrm{St}(n,p)$, which means that a *partial Riemannian (Clarke) subgradient* of f with respect to C_{ij} can be written as

$$\widetilde{\nabla}_{ij} f(X) := \widetilde{\nabla}_{\mathcal{M}_{X_{-ij}}}(f\circ\Phi_{C_{ij}})(X_{ij}) = \mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}}\left(\widetilde{\nabla}^{C_{ij}} f(X)\right) \in \partial_{ij} f(X), \tag{15}$$

for some $\widetilde{\nabla}^{C_{ij}} f(X) \in \partial^{C_{ij}} f(X)$, where $\mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}}(\xi) := X_{ij}\,\mathrm{skew}(X_{ij}^\top\xi) + (I - XX^\top)\xi$ for all $\xi \in \mathbb{R}^{n\times p_{ij}}$.

We summarize our RSSM in Algorithm 1. We start from an initial point $X^0 \in \mathrm{St}(n,p)$ and a fixed partition $\mathfrak{C}$ of $[p]$. At each iteration k, our RSSM first randomly selects an unordered column block index pair $\{i,j\} \in \binom{[\ell]}{2}$ uniformly. We also choose a constant or diminishing stepsize $\gamma_k \in (0, \frac{1}{L})$. Next, we compute a partial Riemannian subgradient $\widetilde{\nabla}_{ij} f(X^k)$ (cf. (15)) of f with respect to C_{ij}. Then, we keep $X_{-ij}^{k+1} := X_{-ij}^k$ and perform a retracted partial Riemannian subgradient step

$$X_{ij}^{k+1} := \mathcal{P}_{\mathcal{M}_{X_{-ij}^k}}(X_{ij}^k - \gamma_k \widetilde{\nabla}_{ij} f(X^k)) = \mathcal{P}_{\mathrm{St}(n,p_{ij})}(X_{ij}^k - \gamma_k \widetilde{\nabla}_{ij} f(X^k)), \tag{16}$$

in which the last equality can be guaranteed by Lemma 3.2. We repeat the same process until convergence is achieved.

Proposition 3.3 guarantees that, given $X \in \mathrm{St}(n,p)$, we have $\partial^{C_{ij}} f(X) = \partial f(X) I_{ij}$ for each $\{i,j\} \in \binom{[\ell]}{2}$. With this relationship, we can describe the following *conditional independence* assumption on our partial subgradient oracle, which is required in the analysis of randomized subgradient algorithms.

Assumption 3.4 (Conditional Independence)**.** *Conditional on the current random iterate X^k, the full Euclidean subgradient $\widetilde{\nabla} f(X^k) \in \partial f(X^k)$ and the column block index pair $\{i,j\} \in \binom{[\ell]}{2}$ are independent random variables.*

Algorithm 1 Randomized Submanifold Subgradient Method on Stiefel Manifolds

Input: A partition $\mathfrak{C} := \{ C_1, \ldots, C_\ell \}$ of $[p]$ with $\ell \geq 2$, and $X^0 \in \mathrm{St}(n,p)$.
Output: A random sequence $\{X^k\} \subseteq \mathrm{St}(n,p)$

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

2: Pick $\{i,j\} \sim \mathrm{Uniform}\binom{[\ell]}{2}$ and a constant/diminishing stepsize $\gamma_k \in \left(0, \frac{1}{L}\right)$ (cf. Theorem 4.9 and Theorem 4.11).

3: Compute a partial Riemannian subgradient (cf. (15)) by

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

4: Perform the update (cf. (16))

$$X^{k+1}_{ij} := \mathcal{P}_{\mathcal{M}_{X^k_{-ij}}}(X^k_{ij} - \gamma_k \widetilde{\nabla}_{ij} f(X^k)) = \mathcal{P}_{\mathrm{St}(n,p_{ij})}(X^k_{ij} - \gamma_k \widetilde{\nabla}_{ij} f(X^k)), \quad X^{k+1}_{-ij} := X^k_{-ij}.$$

5: **end for**

3.3 Per-Iteration Complexity

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

Proposition 3.5 (Per-iteration Complexity)**.** *Let $\mathfrak{C} := \{C_1, \ldots, C_\ell\}$ be a partition of $[p]$, and $|C_i| \leq \lceil \frac{p}{\ell} \rceil$ for all $i \in [\ell]$. If, furthermore, the partial Euclidean subgradient $\widetilde{\nabla}^{C_{ij}} f(X)$ is given, then each iteration requires $\mathcal{O}\left(\frac{np^2}{\ell}\right)$ floating point operations.*

Proof. For a given $\xi_{ij} \in \mathbb{R}^{n \times p_{ij}}$, we can implement step (15) for finding $\mathcal{P}_{T_{X_{ij}} \mathcal{M}_{X_{-ij}}}(\xi_{ij})$ 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 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, computing the partial Riemannian subgradient $\widetilde{\nabla}_{ij} f(X)$ from the given partial Euclidean subgradient $\widetilde{\nabla}^{C_{ij}} f(X) \in \mathbb{R}^{n \times p_{ij}}$ via step (15) requires $\mathcal{O}(npp_{ij} + np_{ij} + p_{ij}^2) = \mathcal{O}(npp_{ij})$ floating point operations since $p_{ij} < p$. From [GVL13, §5.4.5], it is known that performing the polar decomposition $\mathcal{P}_{\mathrm{St}(n,p_{ij})}(\Xi)$ for a given $\Xi \in \mathbb{R}^{n \times p_{ij}}$ requires $\mathcal{O}(np_{ij}^2)$ floating point operations. This implies step (16) requires $\mathcal{O}(np_{ij} + np_{ij}^2) = \mathcal{O}(np_{ij}^2)$ floating point operations. In total, since $p_{ij} \leq \lceil \frac{2p}{\ell} \rceil = \mathcal{O}(\frac{p}{\ell})$, steps (15) and (16) in each iteration together require $\mathcal{O}(npp_{ij} + np_{ij}^2) = \mathcal{O}(npp_{ij}) = \mathcal{O}(\frac{np^2}{\ell})$ floating point operations. $\square$

4 Theoretical Analysis

In this section, we provide a theoretical study of our proposed RSSM. We first prove an inequality, which is a generalization of [LCD+21, Theorem 1] and plays an important role in the convergence analysis of subgradient-type methods for solving weakly convex optimization over compact proximally smooth manifolds, including Stiefel manifolds.

Theorem 4.1 (Riemannian Subgradient Inequality)**.** *Let $\mathcal{M}$ be a compact R-proximally smooth manifold in $\mathbb{R}^{n\times p}$. Suppose that $h : \mathbb{R}^{n\times p} \to \mathbb{R}$ is τ-weakly convex on $\mathbb{R}^{n\times p}$ for some $\tau \in \mathbb{R}$. Then, for any bounded open convex set $\mathcal{U}$ containing $\mathcal{M}$, there exists a constant $L > 0$ such that h is L-Lipschitz continuous on $\mathcal{U}$ and satisfies*

$$h(Y) \geq h(X) + \langle \widetilde{\nabla}_{\mathcal{M}} h(X), Y - X \rangle - \frac{\tau + \|\widetilde{\nabla} h(X)\|/R}{2} \|Y - X\|^2 \tag{17}$$

$$\geq h(X) + \langle \widetilde{\nabla}_{\mathcal{M}} h(X), Y - X \rangle - \frac{\tau + L/R}{2} \|Y - X\|^2 \tag{18}$$

for all $\widetilde{\nabla} h(X) \in \partial h(X)$, $\widetilde{\nabla}_{\mathcal{M}} h(X) := \mathcal{P}_{T_X\mathcal{M}} \widetilde{\nabla} h(X) \in \partial_{\mathcal{M}} h(X)$, and $X, Y \in \mathcal{M}$.

Proof. Note that h is τ-weakly convex on $\mathbb{R}^{n\times p}$. By [Via83, Proposition 4.4], h is L-Lipschitz on the bounded neighborhood $\mathcal{U}$ containing $\mathcal{M}$ for some $L > 0$. Furthermore, for $X, Y \in \mathcal{M} \subseteq \mathbb{R}^{n\times p}$, from (6) we have for all $\widetilde{\nabla} h(X) \in \partial h(X)$,

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

By uniform normal inequality (see §2.1 (2)), we obtain

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

Moreover, since $\partial_{\mathcal{M}} h(X) := \{ \mathcal{P}_{T_X\mathcal{M}} \widetilde{\nabla} h(X) \,:\, \widetilde{\nabla} h(X) \in \partial h(X) \}$, we have

$$\begin{aligned} h(Y) &\geq h(X) + \langle \widetilde{\nabla}_{\mathcal{M}} h(X), Y - X \rangle - \frac{\tau + \|\widetilde{\nabla} h(X)\|/R}{2} \| Y - X \|^2 \\ &\geq h(X) + \langle \widetilde{\nabla}_{\mathcal{M}} h(X), Y - X \rangle - \frac{\tau + L/R}{2} \| Y - X \|^2 \end{aligned}$$

for $X, Y \in \mathcal{M}$, $\widetilde{\nabla} h(X) \in \partial h(X)$, and $\widetilde{\nabla}_{\mathcal{M}} h(X) := \mathcal{P}_{T_X\mathcal{M}} \widetilde{\nabla} h(X) \in \partial_{\mathcal{M}} h(X)$. ☐

4.1 Adaptive Coordinate Representation

Given any $X \in \mathrm{St}(n, p)$, if we fix an $X_\perp \in \mathrm{St}(n, n-p)$ such that $[X \; X_\perp] \in \mathrm{St}(n, n) =: \mathrm{O}(n)$, for any $\xi \in \mathbb{R}^{n\times p}$, it can generate a coordinate representation $(A, B) \in \mathbb{R}^{p\times p} \times \mathbb{R}^{(n-p)\times p}$, namely, $\xi := XA + X_\perp B = X(X^\top \xi) + X_\perp(X_\perp^\top \xi)$. In other words, $(A, B) = (X^\top \xi, X_\perp^\top \xi)$.

Under the viewpoint of this coordinate representation, we can rewrite the projections

$$\begin{aligned} (I - X_{-ij} X_{-ij}^\top)\xi &= X(I_{ij}\, X_{ij}^\top \xi) + X_\perp(X_\perp^\top \xi), \\ \mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}}(\xi_{ij}) I_{ij}^\top &= X(I_{ij} \mathrm{skew}(X_{ij}^\top \xi_{ij}) I_{ij}^\top) + X_\perp(X_\perp^\top \xi_{ij} I_{ij}^\top), \\ \mathcal{P}_{T_X \mathrm{St}(n,p)}(\xi) &= X(\mathrm{skew}(X^\top \xi)) + X_\perp(X_\perp^\top \xi), \end{aligned}$$

where X_{ij}, X_{-ij}, and I_{ij} are defined with respect to a given partition $\mathfrak{C}$ of $[p]$ (cf. §3 for details).

In the convergence analysis of our algorithm, it is crucial to study the simple average of the partial Riemannian subgradient $\mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}}(\xi_{ij}) I_{ij}^\top$. Note that, for a given partition $\mathfrak{C}$, it holds that $X_{-ij}^\top \left(\mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}}(\xi_{ij}) \right) = 0$ for each pair $\{i, j\}$. We therefore define a linear operator $\mathcal{A}_X : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$, called *scaling operator*, by

$$\mathcal{A}_X(\xi) := \frac{1}{\ell - 1} \sum_{(i,j): i<j} (I - X_{-ij} X_{-ij}^\top) \xi_{ij} I_{ij}^\top. \tag{19}$$

Under the adaptive coordinate representation associated to X, we can rewrite $\mathcal{A}_X$ as

$$\begin{aligned}\mathcal{A}_X(\xi) &= X\left(\tfrac{1}{\ell-1}\sum_{(i,j):i<j} I_{ij}X_{ij}^\top \xi_{ij} I_{ij}^\top\right) + X_\perp\left(X_\perp^\top\left(\tfrac{1}{\ell-1}\sum_{(i,j):i<j}\xi_{ij}I_{ij}^\top\right)\right)\\ &= X\left(\tfrac{1}{\ell-1}Q \boxdot (X^\top\xi)\right) + X_\perp(X_\perp^\top \xi). \end{aligned} \tag{20}$$

Here, Q denotes the *signless graph Laplacian matrix* of the complete graph on $[\ell]$, and we define $\boxdot$ the *block-splitting 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 $A \boxdot B := (\,a_{ij}B_{ij}\,)$. That is to say,

$$Q := \begin{bmatrix} \ell-1 & 1 & \cdots & 1 \\ 1 & \ell-1 & & 1 \\ \vdots & \ddots & \ddots & \vdots \\ 1 & 1 & \cdots & \ell-1 \end{bmatrix},$$

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

We see from Lemma A.1 that $\mathcal{A}_X$ and $\mathcal{P}_{T_X\mathrm{St}(n,p)}$ are commutative, and $\mathcal{A}_X \circ \mathcal{P}_{T_X\mathrm{St}(n,p)}(\xi) = \frac{1}{\ell-1}\sum_{(i,j):i<j}\mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}}(\xi_{ij})I_{ij}^\top$. We also learn from Lemma A.2 that $\mathcal{A}_X$ is self-adjoint and has eigenvalues 1 and $\frac{1}{\ell-1}$ with multiplicities $(n-p)p+\sum_{i=1}^\ell p_i^2$ and $p^2-\sum_{i=1}^\ell p_i^2$ respectively. Therefore, $\mathcal{A}_X$ is positive definite. With this scaling operator $\mathcal{A}_X$, we can define the *Mahalanobis inner products* and *norms* on $\mathbb{R}^{n\times p}$, namely $\langle\,\xi,\eta\,\rangle_{\mathcal{A}_X^{-1}} := \langle\,\mathcal{A}_X^{-1}(\xi),\eta\,\rangle$ and $\|\xi\|_{\mathcal{A}_X^{-1}} := \langle\,\mathcal{A}_X^{-1}(\xi),\xi\,\rangle$.

The following lemma highlights our main purpose of introducing the inner product $\langle\,\cdot\,,\,\cdot\,\rangle_{\mathcal{A}_X^{-1}}$ and the norm $\|\cdot\|_{\mathcal{A}_X^{-1}}$:

Lemma 4.2 (Simple Average of Partial Subgradient)**.** *Let $X \in \mathrm{St}(n,p)$ and $\mathfrak{C}$ be a partition of $[p]$ with $\ell := |\mathfrak{C}| \geq 2$. Suppose $\{i,j\} \sim \mathrm{Uniform}\binom{[\ell]}{2}$ (conditional on X). Then, for any $\eta \in \mathbb{R}^{n\times p}$ that is conditionally deterministic given X and $\widetilde{\nabla} f(X)$, we have*

$$\mathbb{E}\left[\langle\,\widetilde{\nabla}_{ij}f(X)I_{ij}^\top,\eta\,\rangle_{\mathcal{A}_X^{-1}}\,\middle|\,X,\widetilde{\nabla}f(X)\right] = \tfrac{2}{\ell}\langle\,\widetilde{\nabla}_{\mathrm{St}(n,p)}f(X),\eta\,\rangle, \tag{21}$$

$$\mathbb{E}\left[\|\widetilde{\nabla}_{ij}f(X)I_{ij}^\top\|^2_{\mathcal{A}_X^{-1}}\,\middle|\,X,\widetilde{\nabla}f(X)\right] = \tfrac{2}{\ell}\|\widetilde{\nabla}_{\mathrm{St}(n,p)}f(X)\|^2, \tag{22}$$

where $\widetilde{\nabla} f(X) \in \partial f(X)$ is a full Euclidean subgradient of f at X.

Proof. Conditional on X and $\widetilde{\nabla} f(X) \in \partial f(X)$, we have $\widetilde{\nabla}_{ij}f(X) := \mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}}(\widetilde{\nabla}f(X)I_{ij})$ for $\{i,j\} \in \binom{[\ell]}{2}$. Thus, by Lemma A.1, for any conditionally deterministic $\eta \in \mathbb{R}^{n\times p}$,

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

i.e, (21) is proved. By (35) in Lemma A.1 and $\mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}} \circ \mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}} = \mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}}$, we have

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

Therefore, (22) follows from (21) by putting $\eta := \widetilde{\nabla} f(X)$. □

Now we state below a lemma that relates the two Mahalanobis norms $\|\cdot\|_{\mathcal{A}_X^{-1}}$ and $\|\cdot\|_{\mathcal{A}_{X^+}^{-1}}$, where $X_{ij}^+ := \mathcal{P}_{\mathcal{M}_{X_{-ij}}}(X_{ij} - \gamma \widetilde{\nabla}_{ij} f(X))$ and $X_{-ij}^+ := X_{-ij}$ with a given $\{i,j\} \in \binom{[\ell]}{2}$.

Lemma 4.3 (Almost Isometry Property)**.** *Let $X \in \mathrm{St}(n,p)$ and $\mathfrak{C}$ be a partition of $[p]$ with $\ell := |\mathfrak{C}| \geq 2$. Suppose $X^+ \in \mathrm{St}(n,p)$ such that $X_{ij}^+ := \mathcal{P}_{\mathcal{M}_{X_{-ij}}}(X_{ij} - \gamma \widetilde{\nabla}_{ij} f(X))$, $\gamma \in \left(0, \frac{1}{L}\right)$ and $X_{-ij}^+ := X_{-ij}$ with $\{i,j\} \sim \mathrm{Uniform}\binom{[\ell]}{2}$. Then, for any $Y \in \mathrm{St}(n,p)$ that is conditionally deterministic given X, we have*

$$\begin{aligned}&\mathbb{E}\left[\|Y - X^+\|^2_{\mathcal{A}_{X^+}^{-1}} \,\middle|\, X\right] \\ &\leq \mathbb{E}\left[\|Y - X^+\|^2_{\mathcal{A}_X^{-1}} \,\middle|\, X\right] \\ &\quad + \tfrac{4\gamma}{\ell}(\ell-2)L\|Y - X\|^2 + \tfrac{2\gamma^2 L^2}{\ell}(\ell-2)\left[\,\|Y - X\|^2 + \gamma L(\|Y - X\|^2 + 1) + 2\|Y - X\|\,\right]. \qquad (23)\end{aligned}$$

Before going to the proof, we state a useful lemma that characterizes the operator norm of the difference between two orthogonal projections. We refer the readers to [GVL13, Theorem 2.5.1]) for the proof of Lemma 4.4.

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 $[X\ X_\perp], [Y\ Y_\perp] \in \mathrm{O}(n)$. 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. Let $X \in \mathrm{St}(n,p)$ and $X^+ \in \mathrm{St}(n,p)$ such that $X_{ij}^+ := \mathcal{P}_{\mathcal{M}_{X_{-ij}}}(X_{ij} - \gamma \widetilde{\nabla}_{ij} f(X))$ and $X_{-ij}^+ := X_{-ij}$ with a given $\{i,j\} \in \binom{[\ell]}{2}$. For the sake of clarity, we denote by $\Psi_X(\xi) := X\left[(J - I) \boxdot (X^\top \xi)\right]$ for any $\xi \in \mathbb{R}^{n \times p}$, which is self-adjoint. By (38) in Lemma A.3, we have

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

For all $\xi \in \mathbb{R}^{n \times p}$, we can write

$$(\Psi_{X^+} - \Psi_X)(\xi) = (X_i^+ X_i^{+\top} - X_i X_i^\top)\xi_{-i} I_{-i}^\top + (X_j^+ X_j^{+\top} - X_j X_j^\top)\xi_{-j} I_{-j}^\top \qquad (24)$$

and

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

Since $\gamma<\frac{1}{L}$, by the second-order boundedness condition (11), we have

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

This means, especially when $\xi\in\mathbb{R}^{n\times p}$ is conditionally deterministic given X, we have

$$\langle (\Psi_{X^+}-\Psi_X)(\xi),\xi\rangle\le 2\gamma\|\widetilde{\nabla}_{ij}f(X)\|\|X_{ij}^\top\xi\|\|\xi\|+\gamma^2\|\widetilde{\nabla}_{ij}f(X)\|^2\,\|\xi\|^2+\gamma^3\|\widetilde{\nabla}_{ij}f(X)\|^3\,\|\xi\|^2.$$

Since $Y-X$ is conditionally deterministic given X, we have

$$\begin{aligned}&\langle (\Psi_{X^+}-\Psi_X)(Y-X),Y-X\rangle\\ &\le 2\gamma\|\widetilde{\nabla}_{ij}f(X)\|\|X_{ij}^\top(Y-X)\|\|Y-X\|+\gamma^2\|\widetilde{\nabla}_{ij}f(X)\|^2\,\|Y-X\|^2+\gamma^3\|\widetilde{\nabla}_{ij}f(X)\|^3\,\|Y-X\|^2.\end{aligned}\tag{25}$$

Now, by applying (24) on $\xi:=X-X^+$ to get the first inequality and Lemma 4.4 to get the second inequality, we have

$$\begin{aligned}\|(\Psi_{X^+}-\Psi_X)(X-X^+)\| &\le \|X_{ij}^+X_{ij}^{+\top}-X_{ij}X_{ij}^\top\|_{\mathsf{op}}\|X-X^+\|\le\|(I-X_{ij}X_{ij}^\top)X_{ij}^+\|_{\mathsf{op}}\|X_{ij}-X_{ij}^+\|\\ &= \|(I-X_{ij}X_{ij}^\top)(X_{ij}^+-X_{ij})\|_{\mathsf{op}}\|X_{ij}-X_{ij}^+\|\\ &\le \|X_{ij}-X_{ij}^+\|^2\le\gamma^2\|\widetilde{\nabla}_{ij}f(X)\|^2.\end{aligned}\tag{26}$$

Combining (25) and (26) yields

$$\begin{aligned}&\|Y-X^+\|^2_{\mathcal{A}^{-1}_{X^+}}-\|Y-X^+\|^2_{\mathcal{A}^{-1}_X}\\ &\le(\ell-2)\big(\langle (\Psi_{X^+}-\Psi_X)(Y-X),Y-X\rangle\\ &\qquad+2\,\|(\Psi_{X^+}-\Psi_X)(X-X^+)\|\,\|Y-X\|+\|(\Psi_{X^+}-\Psi_X)(X-X^+)\|\,\|X-X^+\|\big)\\ &\le(\ell-2)\Big(2\gamma\|\widetilde{\nabla}_{ij}f(X)\|\|X_{ij}^\top(Y-X)\|\|Y-X\|+\gamma^2\|\widetilde{\nabla}_{ij}f(X)\|^2\,\|Y-X\|^2\\ &\qquad+\gamma^3\|\widetilde{\nabla}_{ij}f(X)\|^3\,\|Y-X\|^2+2\gamma^2\|\widetilde{\nabla}_{ij}f(X)\|^2\|Y-X\|+\gamma^3\|\widetilde{\nabla}_{ij}f(X)\|^3\Big).\end{aligned}$$

Conditional on X and $\widetilde{\nabla} f(X)$, from (21) and (22) in Lemma 4.2 we have

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

i.e., (23) follows by taking expectation over $\widetilde{\nabla} f(X) \in \partial f(X)$. □

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

Given a partition $\mathfrak{C}$ of $[p]$ and a fixed $X \in \mathrm{St}(n,p)$, the Bregman divergence associated to the convex function $\frac{1}{2}\|\cdot\|^2_{\mathcal{A}_X^{-1}}$ is given by $D_{\|\cdot\|_{\mathcal{A}_X^{-1}}}(Y,Z) = \frac{1}{2}\|Y-Z\|^2_{\mathcal{A}_X^{-1}}$, which is symmetric. We can consider the Bregman-Moreau envelope and the Bregman-proximal map of a weakly convex function f over $\mathrm{St}(n,p)$ associated to $\frac{1}{2}\|\cdot\|^2_{\mathcal{A}_X^{-1}}$, and then evaluate them at $X \in \mathrm{St}(n,p)$ to construct our *adaptive Moreau envelope*, *adaptive proximal mapping* and the *surrogate stationarity measure* of f respectively as follows:

$$
f_\lambda^{\mathfrak{C}}(X) := \min_{Y\in\mathrm{St}(n,p)} \left\{ f(Y) + \tfrac{1}{2\lambda}\|Y-X\|^2_{\mathcal{A}_X^{-1}} \right\}, \tag{27}
$$

$$
P_{\lambda f}^{\mathfrak{C}}(X) := \operatorname*{argmin}_{Y\in\mathrm{St}(n,p)} \left\{ f(Y) + \tfrac{1}{2\lambda}\|Y-X\|^2_{\mathcal{A}_X^{-1}} \right\}, \tag{28}
$$

$$
\Theta_\lambda^{\mathfrak{C}}(X) := \tfrac{1}{\lambda}\|P_{\lambda f}^{\mathfrak{C}}(X) - X\|. \tag{29}
$$

Following the similar idea in [WHC+23, Lemma 4.2], we show that, like the standard proximal map, our adaptive proximal map in (28) is also single-valued and Lipschitz continuous over consecutive iterates $X, X^+ \in \mathrm{St}(n,p)$.

Lemma 4.5. *Let $f : \mathbb{R}^{n\times p} \to \mathbb{R}$ be a τ-weakly convex function which is automatically L-Lipschitz continuous on some convex neighborhood of $\mathrm{St}(n,p)$ for some $L > 0$. For $\lambda < \frac{1}{\tau+(2\ell-1)L}$, the adaptive proximal map $P_{\lambda f}^{\mathfrak{C}}$ is single-valued and Lipschitz continuous over consecutive iterates $X, X^+ \in \mathrm{St}(n,p)$ as in Lemma 4.3 with a constant $\frac{\ell-1}{1-\lambda(\tau+(2\ell-1)L)}$.*

Proof. For any $X \in \mathrm{St}(n,p)$, we have $h_X := f + \frac{1}{2\lambda}\|\cdot - X\|^2_{\mathcal{A}_X^{-1}}$ is $\left(\frac{1}{\lambda}-\tau\right)$-strongly convex where $\frac{1}{\lambda} > \tau$ is to be determined, and $\partial h_X(Z) = \partial f(Z) + \frac{1}{\lambda}\mathcal{A}_X^{-1}(Z-X)$ for any $Z \in \mathbb{R}^{n\times p}$.

If $Z \in P_{\lambda f}^{\mathfrak{C}}(X) = \operatorname{argmin}_{Z'\in\mathrm{St}(n,p)} h_X(Z')$, then $h_X(Z) := f(Z) + \frac{1}{2\lambda}\|Z-X\|^2_{\mathcal{A}_X^{-1}} \leq f(X)$, which implies $\frac{1}{2\lambda}\|Z-X\|^2 \leq \frac{1}{2\lambda}\|Z-X\|^2_{\mathcal{A}_X^{-1}} \leq f(X)-f(Z) \leq L\,\|Z-X\|$, i.e., $\|Z-X\| \leq 2\lambda L$. Furthermore, we obtain

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

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

To prove that $P_{\lambda f}^{\mathfrak{C}}$ is single-valued, letting $Z, Y \in P_{\lambda f}^{\mathfrak{C}}(X)$ and applying (17) of Theorem 4.1 to h_X at Z and Y yield

$$h_X(Y) \geq h_X(Z) + \frac{\frac{1}{\lambda}-\tau-(2\ell-1)L}{2}\|Y - Z\|^2 \text{ and } h_X(Z) \geq h_X(Y) + \frac{\frac{1}{\lambda}-\tau-(2\ell-1)L}{2}\|Y - Z\|^2,$$

which implies $\|Y - Z\|^2 \leq 0$, i.e, $Z = Y$ by setting $\lambda < \frac{1}{\tau+(2\ell-1)L}$.

Let $X, X^+ \in \mathrm{St}(n, p)$ be consecutive iterates as in Lemma 4.3, and we want to show the Lipschitz property of $P_{\lambda f}^{\mathfrak{C}}$. Applying (17) of Theorem 4.1 to h_X and h_{X^+} at $P_{\lambda f}^{\mathfrak{C}}(X)$ and $P_{\lambda f}^{\mathfrak{C}}(X^+)$ respectively, we have

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

Summing these two inequalities, we get

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

where the second equality follows from the facts that $\mathcal{A}_X^{-1}(X) = X$ and $\mathcal{A}_{X^+}^{-1}(X^+) = X^+$, and the second inequality follows from (39) in Lemma A.4 and the Cauchy-Schwarz inequality. Thus, we have $P_{\lambda f}^{\mathfrak{C}}$ is Lipschitz over consecutive iterates X and X^+ with a constant $\frac{\ell-1}{1-\lambda(\tau+(2\ell-1)L)}$. □

Proposition 4.6 (Surrogate Stationarity Measure)**.** *Let $f : \mathbb{R}^{n\times p} \to \mathbb{R}$ be a τ-weakly convex function and be L-Lipschitz continuous on some neighborhood of $\mathrm{St}(n, p)$ for some $L > 0$. Suppose $\lambda < \frac{1}{\tau+(2\ell-1)L}$ so that $P_{\lambda f}^{\mathfrak{C}}$ is well-defined and single-valued. Then, the following assertions hold:*

(a) $f_\lambda^{\mathfrak{C}}(X) \leq f(X) - \frac{\frac{1}{\lambda}-\tau-(2\ell-1)L}{2}\|P_{\lambda f}^{\mathfrak{C}}(X) - X\|^2$.

(b) $\mathrm{dist}(0, \partial_{\mathrm{St}(n,p)} f(P_{\lambda f}^{\mathfrak{C}}(X))) \leq \frac{\ell-1}{\lambda}\|P_{\lambda f}^{\mathfrak{C}}(X) - X\| = (\ell - 1)\Theta_\lambda^{\mathfrak{C}}(X)$, *and*
$(1 - \lambda(\tau + \ell L))\,\Theta_\lambda^{\mathfrak{C}}(X) = \frac{1-\lambda(\tau+\ell L)}{\lambda}\|P_{\lambda f}^{\mathfrak{C}}(X) - X\| \leq \mathrm{dist}(0, \partial_{\mathrm{St}(n,p)} f(X))$.

(c) $\Theta_\lambda^{\mathfrak{C}}(X) = 0 \;\Leftrightarrow\; X = P_{\lambda f}^{\mathfrak{C}}(X) \;\Leftrightarrow\; 0 \in \partial_{\mathrm{St}(n,p)} f(X)$.

Proof. (a) Let $X \in \mathrm{St}(n, p)$. Due to $\left(\frac{1}{\lambda} - \tau\right)$-strong convexity and bounded subgradient norm of $h_X := f + \frac{1}{2\lambda}\|\cdot - X\|^2_{\mathcal{A}_X^{-1}}$ at $P_{\lambda f}^{\mathfrak{C}}(X)$, applying (17) of Theorem 4.1 on h_X at $P_{\lambda f}^{\mathfrak{C}}(X)$ yields

$$\begin{aligned} f(X) &\geq f(P_{\lambda f}^{\mathfrak{C}}(X)) + \frac{1}{2\lambda}\|P_{\lambda f}^{\mathfrak{C}}(X) - X\|^2_{\mathcal{A}_X^{-1}} + \frac{\frac{1}{\lambda}-\tau-(2\ell-1)L}{2}\|P_{\lambda f}^{\mathfrak{C}}(X) - X\|^2 \\ &\geq f_\lambda^{\mathfrak{C}}(X) + \frac{\frac{1}{\lambda}-\tau-(2\ell-1)L}{2}\|P_{\lambda f}^{\mathfrak{C}}(X) - X\|^2. \end{aligned}$$

(b) Let $Z := P^{\mathfrak{C}}_{\lambda f}(X)$. By optimality, $\mathcal{P}_{T_Z \mathrm{St}(n,p)}\big(\frac{1}{\lambda}\mathcal{A}_X^{-1}(X - Z)\big) \in \partial_{\mathrm{St}(n,p)} f(Z)$. This implies $\mathrm{dist}(0, \partial_{\mathrm{St}(n,p)} f(Z)) \le \frac{1}{\lambda}\|\mathcal{A}_X^{-1}\|_{\mathsf{op}}\|X - Z\| \le \frac{\ell-1}{\lambda}\|X - Z\|$. Now, let $V \in \partial_{\mathrm{St}(n,p)} f(X)$. By (18) of Theorem 4.1,

$$f(Z) \ge f(X) + \langle V, Z - X \rangle - \tfrac{\tau+L}{2}\|Z - X\|^2,$$
$$\langle V, X - Z \rangle \ge \tfrac{\frac{2}{\lambda} - \tau - (2\ell-1)L}{2}\|Z - X\|^2 - \tfrac{\tau+L}{2}\|Z - X\|^2 \ge \big(\tfrac{1}{\lambda} - \tau - \ell L\big)\|Z - X\|^2$$

by (a). This implies $\|V\| \ge \big(\frac{1}{\lambda} - \tau - \ell L\big)\|Z - X\|$, where $\frac{1}{\lambda} - \tau - \ell L > (2\ell - 1)L - \ell L = (\ell - 1)L \ge L > 0$. Hence, the result follows.

(c) This result directly follows from part (b). □

4.3 Convergence of RSSM to Stationary Points

In this subsection, we will show that every accumulation point of the random iterates of Algorithm 1 is a stationary point of (1) almost surely. To this end, we first derive a basic recursion result.

Lemma 4.7 (Basic Recursion Result). *Let $f : \mathbb{R}^{n\times p} \to \mathbb{R}$ be a τ-weakly convex function which is L-Lipschitz continuous on some neighborhood of $\mathrm{St}(n,p)$ for some $L > 0$, and $\mathfrak{C}$ be a partition of $[p]$ with $\ell := |\mathfrak{C}| \ge 2$. Let $\{X^k\} \subseteq \mathrm{St}(n,p)$ be the random sequence of iterates generated by Algorithm 1. Then, for all $Y \in \mathrm{St}(n,p)$ that is conditionally deterministic given X^k, we have*

$$\begin{aligned}
&\mathbb{E}\left[\, \|Y - X^{k+1}\|^2_{\mathcal{A}^{-1}_{X^{k+1}}} \,\middle|\, X^k \right] \\
&\le \|Y - X^k\|^2_{\mathcal{A}^{-1}_{X^k}} + \tfrac{4\gamma_k}{\ell}\left[f(Y) - f(X^k) + \tfrac{\tau+(2\ell-3)L}{2}\|Y - X^k\|^2\right] \\
&\quad + \tfrac{2\gamma_k^2 L^2}{\ell}\,\big(1 + (\ell-2)\,[\,(1+\gamma_k L)\|Y - X^k\|^2 + 4\,\|Y - X^k\| + \gamma_k L(3 + \gamma_k L)]\,\big). \qquad (30)
\end{aligned}$$

Proof. Let $Y \in \mathrm{St}(n,p)$ and $\{i,j\} \sim \mathrm{Uniform}\binom{[\ell]}{2}$ be selected. By the second-order boundedness condition (11) due to $\gamma_k L < 1$, we have $\|X^{k+1}_{ij} - X^k_{ij} + \gamma_k \widetilde{\nabla}_{ij} f(X^k)\| \le \gamma_k^2 \|\widetilde{\nabla}_{ij} f(X^k)\|^2$. Then, we obtain

$$\begin{aligned}
&\|Y - X^{k+1}\|^2_{\mathcal{A}^{-1}_{X^k}} = \|Y - X^{k+1}\|^2 + \|Y - X^{k+1}\|^2_{\mathcal{A}^{-1}_{X^k} - \mathcal{I}} \\
&\le \|Y - X^k + \gamma_k \widetilde{\nabla}_{ij} f(X^k) I_{ij}^\top\|^2 \\
&\qquad + \left(\|Y - X^k + \gamma_k \widetilde{\nabla}_{ij} f(X^k) I_{ij}^\top\|_{\mathcal{A}^{-1}_{X^k} - \mathcal{I}} + \|\mathcal{A}^{-1}_{X^k} - \mathcal{I}\|^{1/2}_{\mathsf{op}}\, \|X^{k+1} - X^k + \gamma_k \widetilde{\nabla}_{ij} f(X^k) I_{ij}^\top\|\right)^2 \\
&\le \|Y - X^k + \gamma_k \widetilde{\nabla}_{ij} f(X^k) I_{ij}^\top\|^2_{\mathcal{A}^{-1}_{X^k}} \\
&\qquad + 2(\ell-2)\,\gamma_k^2\,\|\widetilde{\nabla}_{ij} f(X^k)\|^2\,\|Y - X^k + \gamma_k \widetilde{\nabla}_{ij} f(X^k) I_{ij}^\top\| + (\ell-2)\,\gamma_k^4 \|\widetilde{\nabla}_{ij} f(X^k)\|^4 \\
&\le \|Y - X^k\|^2_{\mathcal{A}^{-1}_{X^k}} + 2\gamma_k\,\langle \widetilde{\nabla}_{ij} f(X^k) I_{ij}^\top, Y - X^k \rangle_{\mathcal{A}^{-1}_{X^k}} + \gamma_k^2\,\|\widetilde{\nabla}_{ij} f(X^k) I_{ij}^\top\|^2_{\mathcal{A}^{-1}_{X^k}} \\
&\qquad + (\ell-2)\left(2\gamma_k^2\,\|\widetilde{\nabla}_{ij} f(X^k)\|^2\,\|Y - X^k\| + 2\gamma_k^3 \|\widetilde{\nabla}_{ij} f(X^k)\|^3 + \gamma_k^4 \|\widetilde{\nabla}_{ij} f(X^k)\|^4\right).
\end{aligned}$$

Here, $\mathcal{I} : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ in the first equality means the identity operator, the first inequality holds because of the Lipschitz-like property (10) and the fact that $Y - X^{k+1} = (Y - X^k + \gamma_k \widetilde{\nabla}_{ij} f(X^k) I_{ij}^\top) - (X^{k+1} - X^k + \gamma_k \widetilde{\nabla}_{ij} f(X^k) I_{ij}^\top)$, and the second inequality is due to $\|\mathcal{A}^{-1}_{X^k} - \mathcal{I}\|_{\mathsf{op}} = \ell - 2$.

Conditional on X^k and $\widetilde{\nabla} f(X^k)$, we have

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

where the first inequality follows from the fact $\|\widetilde{\nabla}_{ij} f(X^k)\| \leq \|\widetilde{\nabla}_{\mathrm{St}(n,p)} f(X^k)\|$ and (21) and (22) in Lemma 4.2, the second inequality holds because $\|\widetilde{\nabla}_{\mathrm{St}(n,p)} f(X^k)\| \leq L$, and the last inequality comes from the Riemannian subgradient inequality (18) in Theorem 4.1.

Therefore, after taking expectation over $\widetilde{\nabla} f(X^k) \in \partial f(X^k)$, it follows from (23) that

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

which is (30). □

The following proposition presents a sufficient decrease result for the surrogate stationarity measure $\Theta_\lambda^{\mathfrak{C}}$, which plays an important role in proving the convergence results of Algorithm 1.

Proposition 4.8 (Sufficient Decrease)**.** *Let $f : \mathbb{R}^{n \times p} \to \mathbb{R}$ be a τ-weakly convex function which is L-Lipschitz continuous on some neighborhood of $\mathrm{St}(n,p)$ for some $L > 0$, and $\mathfrak{C}$ be a partition of $[p]$ with $\ell := |\mathfrak{C}| \geq 2$. Suppose that $\{X^k\} \subseteq \mathrm{St}(n,p)$ is the random sequence generated by the Algorithm 1. Then, for $\lambda \in \left(0, \frac{1}{\tau + (2\ell - 1)L}\right)$ and $k \geq 0$, we have, conditional on X^k,*

$$
\gamma_k \Theta_\lambda^{\mathfrak{C}}(X^k)^2 := \frac{\gamma_k}{\lambda^2} \|P_{\lambda f}^{\mathfrak{C}}(X^k) - X^k\|^2 \leq \frac{f_\lambda^{\mathfrak{C}}(X^k) - \mathbb{E}[\, f_\lambda^{\mathfrak{C}}(X^{k+1}) \,|\, X^k \,] + \frac{\gamma_k^2 L^2}{\lambda \ell} \left(1 + \frac{68}{9}(\ell - 2) \right)}{\frac{\lambda}{\ell} \left[\frac{1}{\lambda} - (\tau + (2\ell - 3)L) \right]}. \tag{31}
$$

Proof. Putting $Y := P_{\lambda f}^{\mathfrak{C}}(X^k) \in \mathrm{St}(n,p)$ into the basic recursion result (30) in Lemma 4.7, we have

$$\begin{aligned}
&\mathbb{E}[\, f_\lambda^{\mathfrak{C}}(X^{k+1}) \,|\, X^k \,] \le \mathbb{E}\left[f(P_{\lambda f}^{\mathfrak{C}}(X^k)) + \tfrac{1}{2\lambda}\|P_{\lambda f}^{\mathfrak{C}}(X^k) - X^{k+1}\|^2_{\mathcal{A}^{-1}_{X^{k+1}}} \,\middle|\, X^k \right] \\
&\le f_\lambda^{\mathfrak{C}}(X^k) + \tfrac{2\gamma_k}{\lambda\ell}\left[f(P_{\lambda f}^{\mathfrak{C}}(X^k)) - f(X^k) + \tfrac{\tau+(2\ell-3)L}{2}\|P_{\lambda f}^{\mathfrak{C}}(X^k) - X^k\|^2 \right] \\
&\qquad + \tfrac{\gamma_k^2 L^2}{\lambda\ell}\left(1 + (\ell-2)\left[(1+\gamma_k L)\|P_{\lambda f}^{\mathfrak{C}}(X^k) - X^k\|^2 + 4\,\|P_{\lambda f}^{\mathfrak{C}}(X^k) - X^k\| + \gamma_k L(3+\gamma_k L)\right]\right) \\
&\le f_\lambda^{\mathfrak{C}}(X^k) + \tfrac{\lambda}{\ell}\left(\tau + (2\ell-3)L - \tfrac{1}{\lambda}\right)\cdot \gamma_k \Theta_\lambda^{\mathfrak{C}}(X^k)^2 \\
&\qquad + \tfrac{\gamma_k^2 L^2}{\lambda\ell}\left(1 + (\ell-2)\left[(1+\gamma_k L)\, 4\lambda^2 L^2 + 8\lambda L + \gamma_k L(3+\gamma_k L)\right]\right) \\
&< f_\lambda^{\mathfrak{C}}(X^k) + \tfrac{\lambda}{\ell}\left(\tau + (2\ell-3)L - \tfrac{1}{\lambda}\right)\cdot \gamma_k \Theta_\lambda^{\mathfrak{C}}(X^k)^2 \\
&\qquad + \tfrac{\gamma_k^2 L^2}{\lambda\ell}\left(1 + (\ell-2)\left[\tfrac{4(1+\gamma_k L)}{(2\ell-1)^2} + \tfrac{8}{2\ell-1} + \gamma_k L(3+\gamma_k L)\right]\right) \\
&< f_\lambda^{\mathfrak{C}}(X^k) + \tfrac{\lambda}{\ell}\left(\tau + (2\ell-3)L - \tfrac{1}{\lambda}\right)\cdot \gamma_k \Theta_\lambda^{\mathfrak{C}}(X^k)^2 + \tfrac{\gamma_k^2 L^2}{\lambda\ell}\left(1 + (\ell-2)\cdot\tfrac{4(4\ell^2+1)}{(2\ell-1)^2}\right) \\
&\le f_\lambda^{\mathfrak{C}}(X^k) + \tfrac{\lambda}{\ell}\left(\tau + (2\ell-3)L - \tfrac{1}{\lambda}\right)\cdot \gamma_k \Theta_\lambda^{\mathfrak{C}}(X^k)^2 + \tfrac{\gamma_k^2 L^2}{\lambda\ell}\left(1 + \tfrac{68}{9}(\ell-2)\right)
\end{aligned}$$

for $\lambda \in \left(0, \frac{1}{\tau+(2\ell-1)L}\right)$. Here, the third inequality is due to the facts that $f(P_{\lambda f}^{\mathfrak{C}}(X^k)) + \frac{1}{2\lambda}\|P_{\lambda f}^{\mathfrak{C}}(X^k) - X^k\|^2_{\mathcal{A}^{-1}_{X^k}} \le f(X^k)$ and $\|P_{\lambda f}^{\mathfrak{C}}(X^k) - X^k\| \le 2\lambda L$. The fourth inequality follows from $\lambda < \frac{1}{\tau+(2\ell-1)L} < \frac{1}{(2\ell-1)L}$ and $\gamma_k < \frac{1}{L}$.

Therefore, for $\lambda \in \left(0, \frac{1}{\tau+(2\ell-1)L}\right)$, we have

$$\gamma_k \Theta_\lambda^{\mathfrak{C}}(X^k)^2 \le \frac{f_\lambda^{\mathfrak{C}}(X^k) - \mathbb{E}[\, f_\lambda^{\mathfrak{C}}(X^{k+1}) \,|\, X^k \,] + \frac{\gamma_k^2 L^2}{\lambda\ell}\left(1 + \frac{68}{9}(\ell-2)\right)}{\frac{\lambda}{\ell}\left[\frac{1}{\lambda} - (\tau + (2\ell-3)L)\right]},$$

which is (31). □

By choosing suitable constant or diminishing stepsizes, we can obtain the following iteration complexity result for our RSSM (Algorithm 1):

Theorem 4.9. *Let $f : \mathbb{R}^{n\times p} \to \mathbb{R}$ be a τ-weakly convex function which is L-Lipschitz continuous on some neighborhood of $\mathrm{St}(n,p)$ for some $L > 0$, and $\mathfrak{C}$ be a partition of $[p]$ with $\ell := |\mathfrak{C}| \ge 2$. Suppose $\{X^k\} \subseteq \mathrm{St}(n,p)$ is the random sequence generated by the Algorithm 1, and $\lambda \in \left(0, \frac{1}{\tau+(2\ell-1)L}\right)$.*

(a) If we choose the constant stepsize $\gamma_k := \frac{\Delta}{\sqrt{T+1}} \in \left(0, \frac{1}{L}\right)$ for some $\Delta > 0$, then

$$\min_{0\le k\le T} \mathbb{E}[\Theta_\lambda^{\mathfrak{C}}(X^k)^2] \le \frac{f_\lambda^{\mathfrak{C}}(X^0) - \min\limits_{X\in\mathrm{St}(n,p)} f_\lambda^{\mathfrak{C}}(X) + \frac{\Delta^2 L^2}{\lambda\ell}\left(1 + \frac{68}{9}(\ell-2)\right)}{\frac{\lambda}{\ell}\left(\frac{1}{\lambda} - (\tau + (2\ell-3)L)\right)\cdot \Delta\sqrt{T+1}}.$$

(b) If we choose the diminishing stepsize $\gamma_k := \frac{\Delta}{\sqrt{k+1}} \in \left(0, \frac{1}{L}\right)$ for some $\Delta > 0$, then

$$\min_{0\le k\le T} \mathbb{E}[\Theta_\lambda^{\mathfrak{C}}(X^k)^2] \le \frac{f_\lambda^{\mathfrak{C}}(X^0) - \min\limits_{X\in\mathrm{St}(n,p)} f_\lambda^{\mathfrak{C}}(X) + \frac{\Delta^2 L^2}{\lambda\ell}\left(1 + \frac{68}{9}(\ell-2)\right)(1 + \log T)}{\frac{\lambda}{\ell}\left(\frac{1}{\lambda} - (\tau + (2\ell-3)L)\right)\cdot \Delta\sqrt{T+1}}.$$

Here, $\mathbb{E}[\Theta_\lambda^{\mathfrak{C}}(X^k)^2]$ is taken over all random paths of random iterate X^k in the algorithm.

Proof. Taking expectation on X^k in the result of Proposition 4.8, we have

$$\gamma_k \, \mathbb{E}[\Theta_\lambda^{\mathfrak{c}}(X^k)^2] \leq \frac{\mathbb{E}\big[f_\lambda^{\mathfrak{c}}(X^k)\big] - \mathbb{E}\big[f_\lambda^{\mathfrak{c}}(X^{k+1})\big] + \frac{\gamma_k^2 L^2}{\lambda \ell}\left(1 + \frac{68}{9}(\ell-2)\right)}{\frac{\lambda}{\ell}\left(\frac{1}{\lambda} - (\tau + (2\ell-3)L)\right)}. \tag{32}$$

Summing (32) over $k = 0, \ldots, T$ gives

$$\min_{0 \leq k \leq T} \mathbb{E}[\,\Theta_\lambda^{\mathfrak{c}}(X^k)^2\,] \leq \frac{f_\lambda^{\mathfrak{c}}(X^0) - \mathbb{E}[\,f_\lambda^{\mathfrak{c}}(X^{T+1})] + \frac{L^2}{\lambda\ell}\left(1 + \frac{68}{9}(\ell-2)\right)\sum_{k=0}^T \gamma_k^2}{\frac{\lambda}{\ell}\left(\frac{1}{\lambda} - (\tau + (2\ell-3)L)\right)\sum_{k=0}^T \gamma_k}. \tag{33}$$

The result in (a) follows from the fact that $\sum_{k=0}^T \gamma_k = \Delta\sqrt{T+1}$ and $\sum_{k=0}^T \gamma_k^2 = \Delta^2$ by putting $\gamma_k := \frac{\Delta}{\sqrt{T+1}} \in \left(0, \frac{1}{L}\right)$ in the inequality (33). On the other hand, if we choose $\gamma_k := \frac{\Delta}{\sqrt{k+1}} \in \left(0, \frac{1}{L}\right)$ in the inequality (33), then we have $\sum_{k=0}^T \gamma_k^2 = \sum_{k=0}^T \frac{\Delta^2}{k+1} \leq \Delta^2\left(1 + \int_1^T \frac{du}{u}\right) = \Delta^2(1 + \log T)$ and $\gamma_k \geq \gamma_T = \frac{\Delta}{\sqrt{T+1}}$ for $k \leq T$, which implies $\sum_{k=0}^T \gamma_k \geq \Delta\sqrt{T+1}$. Hence, the result in (b) follows. ☐

If we take $\lambda := \frac{1}{2(\tau + (2\ell-3)L)}$, and choose the constant stepsizes $\gamma_k := \frac{\Delta}{\sqrt{T+1}}$, from Theorem 4.9,

$$\begin{aligned}\min_{0 \leq k \leq T} \mathbb{E}[\Theta_\lambda^{\mathfrak{c}}(X^k)] &\leq \min_{0 \leq k \leq T} \sqrt{\mathbb{E}[\Theta_\lambda^{\mathfrak{c}}(X^k)^2]} \\ &\leq \frac{\sqrt{\frac{2\ell}{\Delta}\left(f_\lambda(X^0) - \min_{X \in \mathrm{St}(n,p)} f_\lambda(X)\right) + 32\ell L^2\Delta(\tau + (2\ell-3)L)}}{\sqrt[4]{T+1}}.\end{aligned}$$

This means we can prove that our RSSM can compute an ε-nearly stationary point in terms of the expectation of the surrogate stationarity measure of iterates within $\mathcal{O}(\binom{\ell}{2}^2 \varepsilon^{-4})$ iterations.

Lemma 4.10 (Convergence Lemma)**.** *(see [ZCZL22, Lemma 3.1]) Consider the sequences $\{Y^k\}$ in $\mathbb{R}^{n\times p}$ and $\{\mu_k\}$ in $\mathbb{R}_+$. Let $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ be L_Φ-Lipschitz continuous over consecutive iterates in $\{Y^k\}$. Suppose further there exists $p > 0$ such that*

- *$M > 0$ such that $\|Y^k - Y^{k+1}\| \leq M\mu_k$ for all k*
- *$\sum_k \mu_k = \infty$*
- *there exists $\bar{\Phi} \in \mathbb{R}^m$ such that $\sum_k \mu_k \|\Phi(Y^k) - \bar{\Phi}\|^p < \infty$.*

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

By employing the supermartingale convergence theorem (see [ZCZL22, Theorem 2.1]) and a convergence lemma (see Lemma 4.10) modified from [ZCZL22, Lemma 3.1], we obtain the following almost-sure convergence result for $\Theta_\lambda^{\mathfrak{c}}(X^k)$ of the iterates:

Theorem 4.11 (Almost-Sure Convergence)**.** *Let $f : \mathbb{R}^{n\times p} \to \mathbb{R}$ be a τ-weakly convex function which is L-Lipschitz continuous on some neighborhood of $\mathrm{St}(n,p)$ for some $L > 0$. Suppose $\{X^k\} \subseteq \mathrm{St}(n,p)$ is the random sequence generated by the Algorithm 1 with the stepsizes $\{\gamma_k\}$ satisfying $\sum_k \gamma_k = \infty$ and $\sum_k \gamma_k^2 \leq \bar{\gamma} < \infty$. Then, for $\lambda \in \left(0, \frac{1}{\tau + (2\ell-1)L}\right)$, $\lim_{k\to\infty} \Theta_\lambda^{\mathfrak{c}}(X^k) = 0$ almost surely. Hence, every accumulation point of $\{X^k\}$ is a stationary point of (1) almost surely.*

Proof. By Proposition 4.8, we have

$$\begin{aligned}&\mathbb{E}\left[\, f_\lambda^{\mathfrak{C}}(X^{k+1}) - f^* \,\middle|\, X^k \,\right]\\ &\le \left(f_\lambda^{\mathfrak{C}}(X^k) - f^*\right) - \tfrac{\lambda}{\ell}\left(\tfrac{1}{\lambda} - (\tau + (2\ell-3)L)\right)\cdot \gamma_k \Theta_\lambda^{\mathfrak{C}}(X^k)^2 + \tfrac{\gamma_k^2 L^2}{\lambda \ell}\left(1 + \tfrac{68}{9}(\ell-2)\right).\end{aligned}$$

If $\sum_k \gamma_k^2 < \infty$, by Robbins-Siegmund Theorem or Supermartingale Convergence Theorem (cf. [ZCZL22, Theorem 2.1]) we can conclude $f_\lambda^{\mathfrak{C}}(X^k) - f^*$ almost surely converges to a finite random variable and $\sum_k \gamma_k \Theta_\lambda^{\mathfrak{C}}(X^k)^2 < \infty$ almost surely.
Let $\Omega_1 := \left\{\, \omega \in \Omega : \sum_k \gamma_k \Theta_\lambda^{\mathfrak{C}}(X^k(\omega))^2 < \infty \,\right\}$, i.e., $\mathbb{P}(\Omega_1) = 1$. By Lemma 4.5, $P_{\lambda f}^{\mathfrak{C}}$ is Lipschitz over consecutive iterates. Hence, $\Theta_\lambda^{\mathfrak{C}}$ is also Lipschitz continuous over consecutive iterates. Furthermore, $\sum_k \gamma_k = \infty$ and $\|X^k - X^{k+1}\| \le L\gamma_k$ for all k. By Lemma 4.10, we have $\lim_{k\to\infty} \Theta_\lambda^{\mathfrak{C}}(X^k(\omega)) = 0$ for all $\omega \in \Omega_1$, i.e., $\lim_{k\to\infty} \Theta_\lambda^{\mathfrak{C}}(X^k) = 0$ almost surely. This means every accumulation point of $\{X^k\}$ is a stationary point of (1) almost surely. □

We also derive the following almost-sure asymptotic convergence rate result:

Theorem 4.12 (Almost-Sure Asymptotic Convergence Rate)**.** *Let $f : \mathbb{R}^{n\times p} \to \mathbb{R}$ be a τ-weakly convex function which is L-Lipschitz continuous on some neighborhood of* $\mathrm{St}(n,p)$ *for some $L > 0$. Suppose $\{X^k\} \subseteq \mathrm{St}(n,p)$ is the random sequence generated by the Algorithm 1 with the stepsizes $\gamma_k := \frac{\Delta}{\sqrt{k+2}\log(k+2)}$ for some $\Delta > 0$. Then, for $\lambda \in \left(0, \frac{1}{\tau+(2\ell-1)L}\right)$, we have $\liminf_{k\to\infty} \sqrt[4]{k+2}\,\Theta_\lambda^{\mathfrak{C}}(X^k) = 0$ almost surely.*

Proof. Observe that $\sum_{k=0}^K \gamma_k^2 \le \gamma_0^2 + \int_0^K \frac{\Delta^2 du}{(u+2)\log(u+2)^2} \le \gamma_0^2 + \frac{\Delta^2}{\log 2} < \frac{2\Delta^2}{(\log 2)^2}$, and $\sum_{k=0}^K \gamma_k \ge \sum_{k=0}^K \frac{\Delta}{\sqrt{K+2}\log(K+2)} = \frac{\Delta(K+1)}{\sqrt{K+2}\log(K+2)} \to +\infty$. Let $\widetilde{\Omega}_\delta := \{\, \omega \in \Omega : \liminf_{k\to\infty} \sqrt{k+2}\,\Theta_\lambda^{\mathfrak{C}}(X^k(\omega))^2 \ge \delta \,\}$ and assume $\mathbb{P}(\widetilde{\Omega}_\delta) > 0$ for some $\delta > 0$. This means there exists a large $\bar{k}$ such that $\mathbb{P}(\bigcap_{k\ge\bar{k}} \widetilde{\Omega}_{\delta,k}) > 0$, where

$$\widetilde{\Omega}_{\delta,k} := \{\, \omega \in \Omega : \sqrt{k+2}\,\Theta_\lambda^{\mathfrak{C}}(X^k(\omega))^2 \ge \delta \,\}.$$

Hence, we have

$$\mathbb{P}\left(\left\{\, \omega \in \Omega \,\middle|\, \sum_{k\ge\bar{k}} \frac{\Theta_\lambda^{\mathfrak{C}}(X^k(\omega))^2}{\sqrt{k+2}\,\log(k+2)} \ge \sum_{k\ge\bar{k}} \frac{\delta}{(k+2)\log(k+2)} \,\right\}\right) \ge \mathbb{P}\left(\bigcap_{k\ge\bar{k}} \widetilde{\Omega}_{\delta,k}\right) > 0.$$

Note that $\sum_{k\ge\bar{k}} \frac{1}{(k+2)\log(k+2)} \ge \int_{\bar{k}+3}^\infty \frac{du}{u\log u} = \infty$. This means the above inequality is contradicting to $\mathbb{P}(\Omega_1) = 1$ where $\Omega_1 := \left\{\, \omega \in \Omega : \sum_k \gamma_k \Theta_\lambda^{\mathfrak{C}}(X^k(\omega))^2 < \infty \,\right\}$, which has been appeared inside the proof of Theorem 4.11. □

5 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 especially suitable for high-dimensional applications. Furthermore, we showed that RSSM converges to the set of stationary points at a sublinearly rate. To the best of our knowledge, RSSM is the first feasible, coordinate-type algorithm for nonsmooth weakly convex optimization over Stiefel manifolds.

References

[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.

[BT22] MV Balashov and AA Tremba. Error bound conditions and convergence of optimization methods on smooth and proximally smooth manifolds. *Optimization*, 71(3):711–735, 2022.

[Cla90] Frank H Clarke. *Optimization and Nonsmooth Analysis*. SIAM, 1990.

[CSW95] Francis H Clarke, Ronald J Stern, and Peter R Wolenski. Proximal smoothness and the lower-C2 property. *Journal of Convex Analysis*, 2(1-2):117–144, 1995.

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

[GD04] John C Gower and Garmt B Dijksterhuis. *Procrustes Problems*, volume 30. OUP Oxford, 2004.

[GHN22] David H Gutman and Nam Ho-Nguyen. Coordinate descent without coordinates: Tangent subspace descent on Riemannian manifolds. *Mathematics of Operations Research*, 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.

[GVL13] Gene H Golub and Charles F Van Loan. *Matrix Computations*. JHU press, 2013.

[HLWY20] Jiang Hu, Xin Liu, Zai-Wen Wen, and Ya-Xiang Yuan. A brief introduction to manifold optimization. *Journal of the Operations Research Society of China*, 8:199–248, 2020.

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

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

[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.

[LM08] Adrian S Lewis and Jérôme Malick. Alternating projections on manifolds. *Mathematics of Operations Research*, 33(1):216–234, 2008.

[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.

[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.

[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.

[STXY16] Hao-Jun Michael Shi, Shenyinying Tu, Yangyang Xu, and Wotao Yin. A primer on coordinate descent algorithms. *arXiv preprint arXiv:1610.00040*, 2016.

[TCA09] Fabian J Theis, Thomas P Cason, and P A Absil. Soft dimension reduction for ICA by joint diagonalization on the Stiefel manifold. In *Independent Component Analysis and Signal Separation: 8th International Conference, ICA 2009, Paraty, Brazil, March 15-18, 2009. Proceedings 8*, pages 354–361. Springer, 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.

[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 Some Technical Lemmas

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

$$\mathcal{P}_{T_X\mathrm{St}(n,p)} \circ \mathcal{A}_X(\xi) = \mathcal{A}_X \circ \mathcal{P}_{T_X\mathrm{St}(n,p)}(\xi) = \tfrac{1}{\ell-1} \sum_{(i,j):i<j} \mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}}(\xi_{ij}) I_{ij}^\top, \tag{34}$$

$$\mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}}\left(\mathcal{A}_X(\xi) I_{ij}\right) I_{ij}^\top = \mathcal{A}_X\left(\mathcal{P}_{T_{X_{ij}}\mathcal{M}_{X_{-ij}}}(\xi_{ij}) I_{ij}^\top\right). \tag{35}$$

Proof. For every $\xi \in \mathbb{R}^{n\times p}$, direct computation gives

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

Besides, since $\mathcal{A}_X(\xi) = X\left(\frac{1}{\ell-1} Q \boxdot (X^\top\xi)\right) + (I - XX^\top)\xi$, we have

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

i.e., $\mathcal{P}_{T_X\mathrm{St}(n,p)} \circ \mathcal{A}_X(\xi) = \mathcal{A}_X \circ \mathcal{P}_{T_X\mathrm{St}(n,p)}(\xi)$ and result (34) follows. Furthermore,

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

i.e., result (35) is shown. □

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

Proof. Let $X \in \mathrm{St}(n,p)$. Fix any $X_\perp$ such that $[X\ X_\perp] \in \mathrm{O}(n)$. Use adaptive coordinate representation associated to X, for any $\xi, \eta \in \mathbb{R}^{n\times p}$,

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

i.e., $\mathcal{A}_X$ is self-adjoint. In addition, in terms of adaptive coordinate representation, the following subspaces

$$\left\{ X \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & A_{\ell\ell} \end{bmatrix} + X_\perp B \,:\, A_{ii} \in \mathbb{R}^{p_i \times p_i}, B \in \mathbb{R}^{(n-p)\times p} \right\}, \tag{36}$$

$$\left\{ X \begin{bmatrix} 0 & A_{12} & \cdots & A_{1\ell} \\ A_{21} & 0 & & A_{2\ell} \\ \vdots & & \ddots & \vdots \\ A_{\ell 1} & A_{\ell 2} & \cdots & 0 \end{bmatrix} \,:\, A_{ij} \in \mathbb{R}^{p_i \times p_j} \right\} \tag{37}$$

are the eigenspaces of $\mathcal{A}_X$ with respect to eigenvalues 1 and $\frac{1}{\ell-1}$ respectively. ∎

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

$$\mathcal{A}_Y^{-1}(\xi) - \mathcal{A}_X^{-1}(\xi) = (\ell-2)\left\{ Y\left[(J-I)\boxdot(Y^\top\xi)\right] - X\left[(J-I)\boxdot(X^\top\xi)\right]\right\}, \tag{38}$$

where $I \in \mathbb{R}^{\ell\times\ell}$ is the identity matrix, $J \in \mathbb{R}^{\ell\times\ell}$ is the matrix with all entries equal to 1.

Proof. Let Q^{-1} denote the *entrywise reciprocal* of the signless graph Laplacian matrix Q. For any $\xi \in \mathbb{R}^{n\times p}$, by definition of $\mathcal{A}_X^{-1}$, we have

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

i.e., (38) is shown. ∎

Lemma A.4. *Let $X \in \mathrm{St}(n,p)$ and $\mathfrak{C}$ be a partition of $[p]$ with $\ell := |\mathfrak{C}| \geq 2$. Suppose $X^+ \in \mathrm{St}(n,p)$ such that $X_{ij}^+ := \mathcal{P}_{\mathcal{M}_{X_{-ij}}}(X_{ij} - \gamma\widetilde{\nabla}_{ij} f(X))$ and $X_{-ij}^+ := X_{-ij}$ with a given $\{i,j\} \in \binom{[\ell]}{2}$. Then, for any $\xi, \zeta \in \mathrm{St}(n,p)$, we have*

$$\left|\left(\|\xi\|^2_{\mathcal{A}^{-1}_{X^+}} - \|\xi\|^2_{\mathcal{A}^{-1}_X}\right) - \left(\|\zeta\|^2_{\mathcal{A}^{-1}_{X^+}} - \|\zeta\|^2_{\mathcal{A}^{-1}_X}\right)\right| \leq \sqrt{2}\,(\ell-2)\,\|X^+ - X\|\,\|\,\xi - \zeta\,\|. \tag{39}$$

Proof. By (38) in Lemma A.3, and from (24) and $\Psi_X(\xi) := X\left[(J-I)\boxdot(X^\top\xi)\right]$, we have

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

where the fourth equality is guaranteed by Lemma 4.4. Thus, (39) follows. ∎