Riemannian Coordinate Descent Algorithms on Matrix Manifolds

Andi Han [1] **Pratik Jawanpuria** [2] **Bamdev Mishra** [2]

Abstract

Many machine learning applications are naturally formulated as optimization problems on Riemannian manifolds. The main idea behind Riemannian optimization is to maintain the feasibility of the variables while moving along a descent direction on the manifold. This results in updating all the variables at every iteration. In this work, we provide a general framework for developing computationally efficient coordinate descent (CD) algorithms on matrix manifolds that allows updating only a few variables at every iteration while adhering to the manifold constraint. In particular, we propose CD algorithms for various manifolds such as Stiefel, Grassmann, (generalized) hyperbolic, symplectic, and symmetric positive (semi)definite. While the cost per iteration of the proposed CD algorithms is low, we further develop a more efficient variant via a first-order approximation of the objective function. We analyze their convergence and complexity, and empirically illustrate their efficacy in several applications.

1. Introduction

In this work, we consider the optimization problem

$$\min_{X \in \mathbb{R}^{n \times p}} f(X) \qquad \text{s.t.} \quad X \in \mathcal{M}, \tag{1}$$

where $\mathcal{M}$ is a smooth, and often nonlinear constraint. Examples of $\mathcal{M}$ include orthogonality constraint (Edelman et al., 1998), positive (semi)definite constraint (Bhatia, 2009; Han et al., 2021), fixed-rank constraint (Vandereycken, 2013), hyperbolic constraint (Nickel & Kiela, 2018), doubly stochastic constraint (Douik & Hassibi, 2019), etc. Problem (1) has been explored in applications such as PCA (Zhang et al., 2016; Kasai et al., 2019), low-rank matrix/tensor completion (Jawanpuria & Mishra, 2018; Nimishakavi et al., 2018; Kressner et al., 2014), computer vision (Pennec et al., 2006), natural language processing (Jawanpuria et al., 2019a; 2020), optimal transport (Mishra et al., 2021; Han et al., 2022; Shi et al., 2021), and deep learning (Arjovsky et al., 2016; Wang et al., 2020). Problem (1) has also been studied in various settings such as stochastic optimization (Bonnabel, 2013; Zhang et al., 2016; Tripuraneni et al., 2018; Sato et al., 2019; Kasai et al., 2019; Han & Gao, 2021), differential privacy (Reimherr et al., 2021; Han et al., 2024a; Utpala et al., 2022), federated learning (Li & Ma, 2022; Huang et al., 2024), decentralized learning (Mishra et al., 2019), and saddle point and bilevel optimization (Han et al., 2023b;a; Zhang et al., 2023; Han et al., 2024b).

The smooth constraint set can be turned into a Riemannian manifold by endowing a properly chosen metric structure. The Riemannian optimization approach (Absil et al., 2008; Boumal, 2023) then provides a principled approach to solve (1) intrinsically on the manifold space. The main idea is to iteratively update the variable along a descent direction without leaving the manifold. The descent direction is often computed using the Riemannian gradient, which is then followed by a retraction update to ensure feasibility of the manifold constraint. As the dimensionality of the constraint set increases, ensuring feasibility via retraction becomes a key computational bottleneck, e.g., the complexity of ensuring orthogonality and positive definiteness scales as $O(n^3)$ with the input dimension n. This has led to many recent works (Gao et al., 2019; Xiao & Liu, 2021; Ablin et al., 2023) that develop infeasible methods for solving (1). However, such methods are largely limited to the orthogonality constraint and cannot be easily adapted to other manifolds.

In the Euclidean space, the coordinate descent (CD) method (Luo & Tseng, 1992; Nesterov, 2012; Wright, 2015) is a classic algorithm that successively solves a small-dimensional subproblem along a component of the vector variable while holding others fixed. Since each subproblem can be more easily solved than the original problem, this strategy leads to efficient variable update.

On manifolds, designing CD updates is inherently difficult (Gutman & Ho-Nguyen, 2023). A few works have proposed manifold specific CD updates, mainly for the orthogonal (Shalit & Chechik, 2014; Jiang et al., 2022; Massart & Abrol, 2022) and Stiefel (Gutman & Ho-Nguyen, 2023) manifolds.

[1]Riken AIP, Japan [2]Microsoft India. Correspondence to: Andi Han <andi.han@riken.jp>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Although Gutman & Ho-Nguyen (2023) discuss a general framework for developing CD methods on manifolds, concrete developments have been shown only for the Stiefel manifold. Recently, for a class of optimization objectives, Darmwal & Rajawat (2023) have proposed CD updates on the symmetric positive definite manifold with the affine-invariant metric.

In this work, we provide a general approach for developing CD algorithms on matrix manifolds. We summarize our contributions below.

- We introduce a framework for designing CD algorithms on manifolds. In particular, we find a basis spanning the tangent space such that a chosen retraction along the direction of such a basis admits an efficient computation. We discuss a simple expression for the coordinate derivative. Finally, we provide optimization ingredients for various matrix manifolds of interest.
- A nonlinear objective f in (1) requires computation of gradient for every CD update. Using a first-order approximation of f, we develop a more efficient CD algorithm which requires gradient computations one in every fixed number of CD updates. We analyze the convergence and complexity of the two algorithms with randomized and cyclic selection of coordinates.
- We show the benefits of the proposed CD algorithms on the orthogonal Procrustes, PCA, orthogonal deep network distillation, nearest matrix, and learning hyperbolic embeddings problems.

2. Preliminaries

Riemannian manifolds and optimization. For a Riemannian manifold $\mathcal{M}$, denote its tangent space at $X \in \mathcal{M}$ as $T_X\mathcal{M}$. A Riemannian metric is an inner product structure $g_X(\cdot,\cdot) = \langle \cdot,\cdot \rangle_X : T_X\mathcal{M} \times T_X\mathcal{M} \to \mathbb{R}$ that varies smoothly with the base point X. In this work, we particularly focus on matrix manifolds, i.e., where X can be represented in the ambient vector space $\mathbb{R}^{m\times n}$. The orthogonal projection $\mathrm{Proj}_X : \mathbb{R}^{m\times n} \to T_X\mathcal{M}$ projects arbitrary ambient vectors to the tangent space $T_X\mathcal{M}$ with respect to the Riemannian metric. For a differentiable function $f : \mathcal{M} \to \mathbb{R}$, the Riemannian gradient at X is defined as the tangent vector $\mathrm{grad} f(X) \in T_X\mathcal{M}$ such that $\langle U, \mathrm{grad} f(X)\rangle_X = \mathrm{D}f(X)[U], \forall U \in T_X\mathcal{M}$ where $\mathrm{D}f(X)[U] = \langle \nabla f(X), U\rangle$. A retraction $\mathrm{Retr}_X : T_X\mathcal{M} \to \mathcal{M}$ allows points to move along the manifold, which satisfies the conditions: $\mathrm{Retr}_X(0) = X$ and $\mathrm{DRetr}_X(0)[U] = U$.

Related works. We provide a detailed review of the existing coordinate descent (CD) algorithms on specific manifolds, along with other related works in Appendix B.

Notations. We use $\langle \cdot,\cdot \rangle$ without the subscript to represent the Euclidean inner product while we use $\langle \cdot,\cdot \rangle_X$ to denote the Riemannian inner product on $T_X\mathcal{M}$. The specific expression for $\langle \cdot,\cdot \rangle_X$ depends on both $\mathcal{M}$ and X. $\mathrm{Sym}(n)$ and $\mathrm{Skew}(n)$ denote the sets of $n \times n$ symmetric and skew-symmetric matrices, respectively. Let $\mathrm{sym}(A) := (A + A^\top)/2$, $\mathrm{skew}(A) := (A - A^\top)/2$, $\exp(\cdot)$ be the elementwise exponential, and $\mathrm{expm}(\cdot)$ be the matrix exponential. We also use e_i to represent the i-th basis vector with the dimension to be determined from the context. $[A]_{ij}$ denotes the i,j-th entry of a matrix A while A_{ij} represents a matrix with index i,j. We use I_n to denote the $n \times n$ identity matrix, 1_n to denote the size-n vector of all 1s, and define $[n] := \{1, 2, ..., n\}$.

3. Proposed CD Framework

As shown in (Shalit & Chechik, 2014; Gutman & Ho-Nguyen, 2023; Massart & Abrol, 2022; Jiang et al., 2022; Darmwal & Rajawat, 2023), for specific manifolds, the key in developing CD algorithms is the choice of the basis vectors B_ℓ ($\ell \in \mathcal{I}$ and $\mathcal{I}$ denotes the index set) spanning the tangent space that allow efficient retraction. In general, our chosen basis need not be orthonormal with respect to the Riemannian metric. Once the basis and retraction are chosen, the CD update is given by $\mathrm{Retr}_X(-\eta\theta B_\ell)$, where $\eta > 0$ is the stepsize and θ is the coordinate derivative, i.e.,

$$\theta := \frac{d}{d\theta} f(\mathrm{Retr}_X(\theta B_\ell))|_{\theta=0} = \langle \mathrm{grad} f(X), B_\ell \rangle_X. \quad (2)$$

It can be verified that $-\theta B_\ell$ is indeed a descent direction, i.e., $\langle \mathrm{grad} f(X), -\theta B_\ell \rangle_X = -\theta \frac{d}{d\theta} f(\mathrm{Retr}_X(\theta B_\ell))|_{\theta=0} = -\theta^2 \leq 0$. The CD algorithm then involves iteratively selecting coordinate index ℓ, computing θ, and updating in the coordinate descent direction $\mathrm{Retr}_X(-\eta\theta B_\ell)$.

The main challenges in developing CD algorithms on matrix manifolds are: 1) characterization of B_ℓ which facilitates efficient computation, 2) efficient computation of θ, and 3) easy generalization to different manifolds. We propose to leverage the following connection between the Riemannian and Euclidean gradients:

$$\theta = \langle \mathrm{grad} f(X), B_\ell \rangle_X = \langle \nabla f(X), B_\ell \rangle, \quad (3)$$

where $\nabla f(X)$ is the Euclidean gradient and the last equality follows from the definition of the Riemannian gradient. We exploit (3) to efficiently compute θ for several manifolds as it is independent of the Riemannian gradient and metric. In the subsequent sections, we develop concrete CD optimization ingredients for the manifolds of interest under the proposed approach. These are summarized in Table 1.

Table 1: Summary of CD ingredients over various manifolds. $H_{ij} = e_i e_j^\top - e_j e_i^\top$ and $E_{ij} = e_i e_j + e_j e_i^\top$ are the basis for skew-symmetric and symmetric matrices, respectively. $G_{ij}(\theta)$ and $R_{ij}(\theta)$ corresponds to the Givens and hyperbolic rotations, respectively. $\mathrm{P}_X \coloneqq I_n - XX^\top$, and $\nabla f(X)_k$ is the k-th column of $\nabla f(X)$. We use $\mathrm{Exp}_x^{\mathcal{S}}(v)$ to denote the exponential retraction over sphere. The complexity only considers the computation of coordinate derivative and coordinate update, while excluding the complexity of first-order oracle $\nabla f(X)$.

	Size	Basis B_ℓ	Coordinate derivative $\theta = \langle \mathrm{grad} f(X), B_\ell \rangle_X$	Coordinate descent update $\mathrm{Retr}_X(-\eta\theta B_\ell)$	Complexity (per update)
Orthogonal*	$n \times n$	$H_{ij}X$	$\langle \nabla f(X), H_{ij}X \rangle$	$G_{ij}\big(-\eta\theta\big)X$	$O(n)$
Stiefel†	$n \times p$	XH_{ij} , $v_X e_k^\top$	$\langle \nabla f(X), XH_{ij} \rangle$, $\mathrm{P}_X(\nabla f(X)_k)$	$XG_{ij}(-\eta\langle \nabla f(X), XH_{ij}\rangle)$, $\mathrm{Exp}_{X_k}^{\mathcal{S}}\big(-\eta \mathrm{P}_X(\nabla f(X)_k)\big)$	$O(n), O(np)$
SPD◇	$n \times n$	$LE_{ij}L^\top$	$\langle \nabla f(X), LE_{ij}L^\top \rangle$	$L\mathrm{expm}(-\eta\theta E_{ij})L^\top$	$O(n)$
Stiefel Grassmann	$n \times p$	$H_{ij}X$	$\langle \nabla f(X), H_{ij}X \rangle$	$G_{ij}\big(-\eta\theta\big)X$	$O(p)$
Hyperbolic	$n \times p$	$H_{ij}JX$	$\langle \nabla f(X), H_{ij}JX \rangle$	$\begin{cases} G_{ij}\big(-\eta\theta\big)X & \text{if } i \neq 1 \\ R_{ij}\big(-\eta\theta\big)X & \text{if } i = 1 \end{cases}$	$O(p)$
Symplectic	$2n \times 2p$	$E_{ij}\Omega_n X$	$\langle \nabla f(X), E_{ij}\Omega_n X \rangle$	Proposition 3.7	$O(p)$
Doubly stoch.	$m \times n$	$Ae_i e_j^\top B^\top$	$\langle \nabla f(YY^\top), Ae_i e_j^\top B^\top \rangle$	$\mathrm{cSK}(X \odot \exp(-\eta\theta B_\ell \oslash X))$	$O(1)$
Multinomial	$n \times p$	$e_i(e_j - e_{j+1})^\top$	$[\nabla f(X)]_{ij} - [\nabla f(X)]_{i(j+1)}$	$\mathcal{P}(X \odot \exp(-\eta\theta B_\ell \oslash X))$	$O(1)$
SPSD / SPD	$n \times n$	$e_i e_j^\top$	$\langle \nabla f(YY^\top), e_i e_j^\top \rangle$	$Y - \eta\theta e_i e_j^\top$	$O(1)$

*(Shalit & Chechik, 2014); †(Gutman & Ho-Nguyen, 2023); ◇(Darmwal & Rajawat, 2023).

3.1. CD on Stiefel manifold

The Stiefel manifold $\mathrm{St}(n,p)$ is the set of column orthonormal matrices of size $\mathbb{R}^{n\times p}$, i.e., $\mathrm{St}(n,p) \coloneqq \{X \in \mathbb{R}^{n\times p} : X^\top X = I_p\}$. When $p = n$, $\mathrm{St}(n,n) \equiv \mathcal{O}(n)$, the orthogonal manifold. The tangent space of Stiefel manifold is identified as $T_X\mathrm{St}(n,p) = \{U \in \mathbb{R}^{n\times p} : X^\top U + U^\top X = 0\}$. The Riemannian metric is defined as $\langle U, V\rangle_X \coloneqq \langle U, V\rangle$ for any $U, V \in T_X\mathcal{M}$. The Riemannian gradient is derived as $\mathrm{grad} f(X) = \nabla f(X) - X\mathrm{sym}(X^\top \nabla f(X))$.

Choice of basis. Taking inspiration from $\mathcal{O}(n)$, we adopt the ΩX parameterization of the tangent vectors (where $\Omega \in \mathrm{Skew}(n)$) and choose the basis as $B_\ell = H_{ij}X$ for $\ell \in \mathcal{I} = \{(i,j) : 1 \leq i < j \leq n\}$ and $H_{ij} \coloneqq e_i e_j^\top - e_j e_i^\top$. In contrast to $\mathcal{O}(n)$, the chosen basis is not orthonormal for $\mathrm{St}(n,p)$. This is expected as the manifold $\mathrm{St}(n,p)$ has a dimension $np - \frac{p(p-1)}{2}$ while we adopt an over-parameterization of the tangent space using $n(n-1)/2$ basis vectors.

Retraction. For the purpose of CD update, we first note that $\mathrm{Retr}_X(tU) = \mathrm{expm}(t\Omega)X$ is a valid retraction on $\mathrm{St}(n,p)$ because: 1) $\mathrm{Retr}_X(0) = X$ and 2) $\mathrm{DRetr}_X(0)[U] = \Omega X = U$ are satisfied (Siegel, 2020).

CD update. Based on the above choices of the basis vectors and retraction, the proposed CD update is $\mathrm{Retr}_X(-\eta\theta B_\ell) = G_{ij}(-\eta\theta)X$, where $\theta = \langle \nabla f(X), B_\ell\rangle = [\nabla f(X)X^\top - X\nabla f(X)^\top]_{ij}$. Here, $G_{ij}(\theta) = I_n + (\cos(\theta) - 1)(e_i e_i^\top + e_j e_j^\top) + \sin(\theta)(e_i e_j^\top - e_j e_i^\top)$ is known as the Givens rotation around axes i, j with angle $-\theta$. Overall, each CD update only requires $O(p)$ as we modify only two rows of X.

Remark 3.1. Gutman & Ho-Nguyen (2023) propose a column-wise CD update on the Stiefel manifold which costs $O(np)$ per iteration. On the other hand, our proposed CD update is row-wise and costs $O(p)$, which is cheaper. Furthermore, the CD update strategy of (Gutman & Ho-Nguyen, 2023) cannot be applied to the sphere manifold, i.e., when $p = 1$, it reduces to the full gradient update on the sphere. This, however, is not an issue for our update. Finally, the update of Gutman & Ho-Nguyen (2023) is not invariant to the right action of orthogonal group and hence does not yield a valid CD strategy for the Grassmann manifold. In contrast, as shown in the next section, our strategy can be readily generalized to the Grassmann manifold.

3.2. CD on Grassmann manifold

The Grassmann manifold $\mathrm{Gr}(n,p)$ represents the p-dimensional subspaces in $\mathbb{R}^n$, which can be represented by an $n \times p$ orthonormal matrix X, i.e., $X \in \mathrm{St}(n,p)$, where the columns span the subspace. The representation is not unique, with XQ representing the same subspace for arbitrary $Q \in \mathcal{O}(p)$. Thus, the Grassmann manifold can be identified as $\mathrm{Gr}(n,p) = \{[X] : X \in \mathbb{R}^{n\times p}, X^\top X = I_p\}$, where $[X] \coloneqq \{XQ : Q \in \mathcal{O}(p)\}$. The tangent space can be uniquely characterized by the horizontal space at $T_X\mathrm{St}(n,p)$, i.e., $T_{[X]}\mathrm{Gr}(n,p) = \{[U] : X^\top U = 0\}$. For a given $\xi \in T_{[X]}\mathrm{Gr}(n,p)$, its unique horizontal lift is $U = \mathrm{lift}_X(\xi)$, where $[X]$ is represented as X. The lift

operator satisfies $\mathrm{lift}_{XQ}(\xi) = \mathrm{lift}_X(\xi)Q$. On $\mathrm{Gr}(n,p)$, the Riemannian metric is pushed forward by the Euclidean metric on $\mathrm{St}(n,p)$ as $\langle \xi, \zeta \rangle_{[X]} = \langle \mathrm{lift}_X(\xi), \mathrm{lift}_X(\zeta) \rangle$ and the corresponding Riemannian gradient $\mathrm{grad} f([X])$ can be represented by $\mathrm{lift}_X(\mathrm{grad} f([X]) = (I_n - XX^\top)\nabla f(X)$. Retractions such as QR retraction for $\mathrm{St}(n,p)$ also work for $\mathrm{Gr}(n,p)$ as long as it preserves the equivalence class, i.e., $[\mathrm{Retr}_{XQ}(t\,\mathrm{lift}_X(\xi)Q)] = [\mathrm{Retr}_X(t\,\mathrm{lift}_X(\xi))]$ for any $Q \in \mathcal{O}(p)$. Below, we show that the proposed CD update for $\mathrm{St}(n,p)$ is also well-defined for $\mathrm{Gr}(n,p)$.

Proposition 3.2. *Consider a function $f : \mathrm{Gr}(n,p) \to \mathbb{R}$. Let the coordinate descent update at $[X]$ be given by $\mathrm{Retr}_X(-\eta\,\theta H_{ij} X) := G_{ij}(-\eta\theta)X$ for $1 \le i < j \le n$, where $\theta = \langle \nabla f(X), H_{ij}X \rangle$ and for some fixed stepsize $\eta > 0$. Then, $\mathrm{Retr}_{XQ}(-\eta\,\theta_{XQ} H_{ij} XQ) = \mathrm{Retr}_X(-\eta\theta H_{ij}X)Q$.*

3.3. CD on hyperbolic manifold

We now consider the generalized hyperbolic space (Bai & Li, 2014; Xiao et al., 2023) $\mathcal{H}(n,p) := \{X \in \mathbb{R}^{n\times p} : -X^\top JX = I_p\}$, where $J = \mathrm{diag}(-1, 1, ..., 1) \in \mathbb{R}^{n\times n}$ is the metric tensor. When $p = 1$, this reduces to the well-known hyperbolic space (the hyperboloid model). The tangent space at $X \in \mathcal{H}(n,p)$ is identified as

$$\begin{aligned} T_X\mathcal{H}(n,p) &= \{U \in \mathbb{R}^{n\times p} : U^\top JX + X^\top JU = 0\} \\ &= \{WJX : W \in \mathrm{Skew}(n)\}. \end{aligned}$$

The Riemannian metric on $T\mathcal{H}(n,p)$ is the generalized Lorentz inner product as $\langle U, V \rangle_{\mathcal{L}} := \mathrm{tr}(U^\top JV)$. The normal space is given by $N_X\mathcal{H}(d,r) = \{XS : S \in \mathrm{Sym}(p)\}$. The orthogonal projection to $T_X\mathcal{H}(n,p)$ and the Riemannian gradient are derived below.

Proposition 3.3. *The orthogonal projection of $A \in \mathbb{R}^{n\times p}$ to $T_X\mathcal{H}(n,p)$ is given by $\mathrm{Proj}_X(A) = A + X\mathrm{sym}(X^\top JA)$. The Riemannian gradient is $\mathrm{grad} f(X) = J\nabla f(X) + X\mathrm{sym}(X^\top \nabla f(X))$.*

Choice of basis. For generalized hyperbolic manifold, we consider the basis $B_\ell = H_{ij}JX = (e_ie_j^\top - e_je_i^\top)JX$, for $1 \le i < j \le n$.

Retraction. Taking inspiration from our Stiefel analysis in Section 3.1, we define the map $\mathrm{Retr}_X(tU) := \mathrm{expm}(tWJ)X$ for $U = WJX \in T_X\mathcal{H}(n,p)$. We next show such a map defines a valid retraction. As shown below, the retraction expression considerably simplifies along the chosen basis $H_{ij}JX$.

Proposition 3.4. *Given a tangent vector $U = WJX \in T_X\mathcal{H}(n,p)$ for some skew-symmetric matrix W, then $\mathrm{Retr}_X(tU) := \mathrm{expm}(tWJ)X$ is a retraction.*

In fact, $\mathrm{expm}(tWJ)$ is a Lorentz transform that satisfies $\mathrm{expm}(tWJ)^\top J\mathrm{expm}(tWJ) = J$, which preserves the Lorentz inner product as $(LX)^\top JLX = X^\top JX = -I_p$. Hence by following the direction $U = \theta H_{ij}JX$, we define a coordinate type of updates on (generalized) hyperbolic manifold as $\mathrm{expm}(\theta H_{ij}J)X$, which can be computed efficiently similar to the Givens rotation. Particularly, when $i, j \ne 1$, $H_{ij}J = H_{ij}$ and $\mathrm{expm}(\theta H_{ij}J) = G_{ij}(\theta)$ exactly recovers the Givens rotation. When $i = 1$, we have $H_{ij}J = E_{ij} := e_ie_j^\top + e_je_i^\top$. We show in the following lemma that this also leads to a rotation known as the hyperbolic rotation.

Lemma 3.5. *For $U = \theta H_{ij}JX$ with $1 \le i < j \le n$, $\mathrm{Retr}_X(tU) = \begin{cases} G_{ij}(\theta)X & \text{if } i \ne 1, \\ R_{ij}(\theta)X & \text{if } i = 1 \end{cases}$ where $R_{ij}(\theta) = I_n + (\cosh(\theta) - 1)(e_ie_i^\top + e_je_j^\top) + \sinh(\theta)(e_ie_j^\top + e_je_i^\top)$.*

When $d = 4$, $R_{ij}(\theta)$ is known as the Lorentz boost with rapidity $-\theta$ and can be thought of as rotation in the time domain. Hence, while the Givens rotation based CD updates have been explored for the orthogonal and Stiefel manifolds (Shalit & Chechik, 2014; Gutman & Ho-Nguyen, 2023), our approach generalizes the Givens rotation based CD updates to hyperbolic spaces.

CD update. Similar to the Stiefel case, the proposed CD update is $\mathrm{Retr}_X(-\eta\theta H_{ij}JX)$, where $\theta = \langle \nabla f(X), H_{ij}JX \rangle = [\nabla f(X)X^\top J - JX\nabla f(X)^\top]_{ij}$. In Appendix E.3, we additionally derive a canonical-type metric and a Cayley retraction.

3.4. CD on symplectic manifold

The symplectic manifold (Gao et al., 2021a;b; 2022) is defined as $\mathrm{Sp}(n,p) := \{X \in \mathbb{R}^{2n\times 2p} : X^\top \Omega_n X = \Omega_p\}$, where $\Omega_k := \begin{bmatrix} 0 & I_k \\ -I_k & 0 \end{bmatrix}$ is a $2k \times 2k$ skew-symmetric block matrix. The tangent space is given as

$$\begin{aligned} T_X\mathrm{Sp}(n,p) &= \{U \in \mathbb{R}^{2n\times 2p} : U^\top \Omega_n X + X^\top \Omega_n U = 0\} \\ &= \{S\Omega_n X : S \in \mathrm{Sym}(2n)\}. \end{aligned}$$

Here we consider the Euclidean metric (Gao et al., 2021a) as $\langle U, V \rangle_X = \mathrm{tr}(U^\top V)$ for any $X \in \mathrm{Sp}(n,p), U, V \in T_X\mathrm{Sp}(n,p)$. The Riemannian gradient (Gao et al., 2021a, Proposition 3) is given by $\mathrm{grad} f(X) = \nabla f(X) - \Omega_n XW_X$, where $W_X \in \mathrm{Skew}(2p)$ is the unique solution to the Lyapunov equation $X^\top XW + WX^\top X = 2\,\mathrm{skew}(X^\top \Omega_n^\top \nabla f(X))$.

Choice of basis. Similar to the Stiefel and hyperbolic manifolds, we consider the basis $B_\ell = E_{ij}\Omega_n X$ for the tangent space $T_X\mathrm{Sp}(n,p)$, where $E_{ij} = e_ie_j^\top + e_je_i^\top$, for $1 \le i \le j \le 2n$. Here, e_i is the i-th basis in $\mathbb{R}^{2n}$.

Retraction. We propose the following retraction for efficient CD updates.

Proposition 3.6. *For any $X \in \mathrm{Sp}(n,p)$ and $U = S\Omega_n X \in$*

$T_X\mathrm{Sp}(n,p)$, *the map* $\mathrm{Retr}_X(tU) = \mathrm{expm}(tS\Omega_n)X$ *is a retraction.*

The above retraction further simplifies when moving along the chosen basis direction.

Proposition 3.7. *Let* $U_{ij} = E_{ij}\Omega_n X \in T_X\mathrm{Sp}(n,p)$ *for* $1 \le i \le j \le 2n$. *Then, we have* $\mathrm{Retr}_X(\theta U_{ij}) = X + (\exp(-\theta) - 1)e_i e_i^\top X + (\exp(\theta) - 1)e_{n+i}e_{n+i}^\top X$, *if* $i = j-n, j > n$ *and* $\mathrm{Retr}_X(\theta U_{ij}) = X + \theta E_{ij}\Omega_n X$ *otherwise.*

Remark 3.8 (Block coordinate updates). We may also consider block coordinate updates. Let $E = \begin{bmatrix} A & C \\ C^\top & B \end{bmatrix}$ for $A, B \in \mathrm{Sym}(n)$ and $C \in \mathbb{R}^{n\times n}$, and we wish to update X in the direction of $U = E\Omega_n X \in T_X\mathrm{Sp}(n,p)$. First, we consider the upper-left and bottom-right blocks, i.e., where $E = \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix}$ or $E = \begin{bmatrix} 0 & 0 \\ 0 & B \end{bmatrix}$ for arbitrary $A, B \in \mathrm{Sym}(n)$. Here, $\mathrm{Retr}_X(\theta E\Omega_n X) = X + \theta E\Omega_n X$. Second, if $E = \begin{bmatrix} 0 & \mathrm{diag}(u) \\ \mathrm{diag}(v) & 0 \end{bmatrix}$, for $u, v \in \mathbb{R}^n$, then $E\Omega_n = \begin{bmatrix} -\mathrm{diag}(u) & 0 \\ 0 & \mathrm{diag}(v) \end{bmatrix}$, and therefore, $\mathrm{Retr}_X(\theta E\Omega_n X) = (I + \theta E\Omega_n + \frac{\theta^2}{2!}(E\Omega_n)^2 + \cdots)X = GX$, where $G = \begin{bmatrix} \mathrm{diag}(\exp(-\theta u)) & 0 \\ 0 & \mathrm{diag}(\exp(\theta v)) \end{bmatrix}$.

CD update. Finally, based on the above discussion our proposed CD update is $\mathrm{Retr}_X(-\eta\theta E_{ij}\Omega_n X)$, where, $\theta = \langle \nabla f(X), E_{ij}\Omega_n X\rangle = [\nabla f(X)X^\top\Omega_n^\top + \Omega_n X\nabla f(X)^\top]_{i,j}$.

3.5. CD on doubly stochastic and multinomial manifolds

Given two marginals $\mu \in \Delta_m, \nu \in \Delta_n$ where $\Delta_k := \{z \in \mathbb{R}^k : z \ge 0, z^\top 1_k = 1\}$ denotes the k-simplex, the doubly stochastic manifold (Douik & Hassibi, 2019; Shi et al., 2021; Mishra et al., 2021) is defined as $\Pi(\mu,\nu) := \{X \in \mathbb{R}^{m\times n} : X > 0, X1_n = \mu, X^\top 1_m = \nu\}$. The tangent space is $T_X\Pi(\mu,\nu) = \{U \in \mathbb{R}^{m\times n} : U1_n = 0, U^\top 1_m = 0\}$, which can be endowed with the Fisher metric as $\langle U, V\rangle_X = \mathrm{tr}(U^\top(V \oslash X))$, where $\odot$ and $\oslash$ represent the elementwise product and division operations, respectively. The orthogonal projection is $\mathrm{Proj}_X(A) = A - (\alpha 1_n^\top + 1_m\beta^\top) \odot X$, where $\alpha \in \mathbb{R}^m, \beta \in \mathbb{R}^n$ are solutions to the linear system: $\alpha \odot \mu + X\beta = A1_n, \beta \odot \nu + X^\top\alpha = A^\top 1_m$. The Riemannian gradient is given by $\mathrm{grad}f(X) = \mathrm{Proj}_X(X \odot \nabla f(X)) = X \odot \big(\nabla f(X) - (\alpha 1_n^\top + 1_m\beta^\top)\big)$.

Choice of basis. We consider the parameterization of the tangent space as $T_X\Pi(\mu,\nu) = \{ACB^\top : A \in \mathbb{R}^{m\times(m-1)}, B \in \mathbb{R}^{n\times(n-1)}, A^\top 1_m = 0, B^\top 1_n = 0, C \in \mathbb{R}^{(m-1)\times(n-1)}\}$. We notice that the tangent space has a dimension $(m-1)\times(n-1)$, and hence, we can let $A = [e_1 - e_2, ..., e_{m-1} - e_m] \in \mathbb{R}^{m\times(m-1)}, B = [e_1 - e_2, ..., e_{n-1} - e_n] \in \mathbb{R}^{n\times(n-1)}$, where we denote e_i as the i-th canonical basis for the corresponding vector space. Hence the tangent space is parameterized by $C \in \mathbb{R}^{(m-1)\times(n-1)}$. The basis we consider $B_\ell = Ae_ie_j^\top B^\top = (e_i - e_{i+1})(e_j - e_{j+1})^\top$ for $i \in [m-1], j \in [n-1]$.

Retraction. We consider the Sinkhorn retraction applied in the direction of the basis as $\mathrm{Retr}_X(-\eta\theta B_\ell) = \mathrm{SK}(X \odot \exp(-\eta\theta B_\ell \oslash X))$. Here, the Sinkhorn algorithm $\mathrm{SK}(U)$ iteratively normalize rows and columns of U according to the given marginals (Knight, 2008). We notice the input to the Sinkhorn algorithm only modifies a 2×2 sub-matrix of X. It, thus, suffices to apply the Sinkhorn algorithm to the 2×2 sub-matrix with the modified marginals, which largely simplifies the computation compared to running the Sinkhorn algorithm for the entire input. To this end, we define the coordinate Sinkhorn, denoted as $\mathrm{SK}^{ij}(U)$ or simply $\mathrm{cSK}(U)$ if the coordinates are clear from context, as performing the Sinkhorn algorithm for the 2×2 sub-matrix formed by indices $i, i+1$ and $j, j+1$ with marginals $([\mu]_i - \sum_{k\neq j,j+1}[U]_{ik}, [\mu]_{i+1} - \sum_{k\neq j,j+1}[U]_{(i+1)k})$ and $([\nu]_j - \sum_{l\neq i,i+1}[U]_{tj}, [\nu]_{j+1} - \sum_{l\neq i,i+1}[U]_{l(j+1)})$. The other entries of U remains unchanged. We show in the next proposition that applying coordinate Sinkhorn to the basis results in a valid retraction.

Proposition 3.9. *The coordinate Sinkhorn applied to the basis* B_ℓ, *i.e.,* $\mathrm{cSK}(X \odot \exp(tB_\ell \oslash X))$ *is a valid retraction in the direction of* B_ℓ.

We can further simplify the computation of $\mathrm{cSK}(X \odot \exp(tB_\ell \oslash X))$, which is equivalent to performing Sinkhorn on a 2×2 matrix. Furthermore, in this case, we show in Lemma E.6 that the Sinkhorn admits a closed-form solution.

CD update. The CD update follows as $\mathrm{cSK}(X \odot \exp(-\eta\theta B_\ell \oslash X))$, where the coordinate derivative θ is computed as $\theta = \langle \nabla f(X), B_\ell\rangle = [\nabla f(X)]_{ij} - [\nabla f(X)]_{i(j+1)} - [\nabla f(X)]_{(i+1)j} + [\nabla f(X)]_{(i+1)(j+1)}$.

Remark 3.10 (CD on multinomial manifold). The developments in this section readily applies to the multinomial manifold (Douik & Hassibi, 2019), i.e., $\mathbb{M}^{n,p} := \{X \in \mathbb{R}^{n\times p} : X > 0, X1_p = v\}$ where we assume $v = 1_n$ without loss of generality. The multinomial constraint corresponds to n independent simplex constraints restricted to positive entries. The tangent space is $T_X\mathbb{M}^{n,p} = \{U \in \mathbb{R}^{n\times p} : U1_p = 0\} = \{VB^\top : V \in \mathbb{R}^{n\times(p-1)}, B \in \mathbb{R}^{p\times(p-1)}, B^\top 1_p = 0\}$. Thus, the basis is similarly given by $B_\ell = e_i(e_j - e_{j+1})^\top$. The Riemannian metric is the same Fisher metric. A retraction is given in $\mathrm{Retr}_X(tU) = \mathcal{P}(X \odot (\exp(tU \oslash X)))$, where $\mathcal{P}(V) = V \oslash (V1_p1_p^\top)$ denotes the row normalization. It should be noted that $\mathcal{P}(V)$ is a special case of the Sinkhorn algorithm without column normalization. Thus, in the basis direction, we can define the coordinate projection by modifying only two entries per row. The coordinate derivative can be computed as $\theta = \langle \nabla f(X), B_\ell\rangle = [\nabla f(X)]_{ij} - [\nabla f(X)]_{i(j+1)}$.

3.6. CD on positive (semi)definite manifold

The set of fixed-rank symmetric positive semi-definite manifold (SPSD) matrices (Vandereycken et al., 2009; 2013; Massart & Absil, 2020) is defined as $\mathbb{S}_+^{n,p} := \{X \in \mathbb{R}^{n\times n} : X \succeq 0, \mathrm{rank}(X) = p\}$. When $p = n$, we recover the set of symmetric positive definite (SPD) matrices as $\mathbb{S}_+^{n,n} \equiv \mathbb{S}_{++}^n$. For the purpose of developing efficient CD updates on SPSD, we follow the parameterization purposed in (Massart & Absil, 2020), i.e., $X \in \mathbb{S}_+^{n\times p}$ is factorized as $X = YY^\top$, $Y \in \mathbb{R}_*^{n\times p}$, which is unique up to the right-action of the orthogonal group $\mathcal{O}(p)$. The Riemannian gradient can be computed as $\mathrm{grad} f(Y) = \nabla f(YY^\top) = 2\,\mathrm{sym}(\nabla f(YY^\top))Y$ because the $T_Y\mathbb{R}_*^{n\times p} \simeq \mathbb{R}^{n\times p}$. The main advantage of this parameterization is its simple expression of retraction, i.e., $\mathrm{Retr}_Y(t\xi) = Y + t\xi$ (Massart & Absil, 2020).

Choice of basis, retraction, and CD update. Using $X = YY^\top$, the optimization problem is on $\mathbb{R}_*^{n\times p}$ with a simple retraction. For the objective $f : \mathbb{S}_+^{n,p} \to \mathbb{R}$, we initialize $Y \in \mathbb{R}_*^{n\times p}$ and update as $\mathrm{Retr}_Y(-\eta\nabla f(YY^\top)) = Y - \eta\nabla f(YY^\top)$ for some stepsize η. We choose the basis to be $e_ie_j^\top$ for $i \in [n], j \in [p]$, which is orthonormal for the tangent space $T_Y\mathbb{R}_*^{n\times p}$. The CD update is given by $\mathrm{Retr}_Y(-\eta\theta e_ie_j^\top)$, where $\theta = \langle \nabla f(YY^\top), e_ie_j^\top\rangle = [\nabla f(YY^\top)]_{ij}$. The simplicity of the geometry allows CD to be developed efficiently on $\mathbb{S}_+^{n,p}$, which coincides with the Euclidean CD update in the Euclidean space. When $p = n$, we obtain a CD update for the SPD manifold.

CD update with the BW metric. The Bures-Wasserstein (BW) metric for the SPD manifold ($p = n$) has been recently studied for various machine learning applications (Bhatia et al., 2019; Han et al., 2021). For the BW metric, the gradient descent update is $\mathrm{Exp}_X(-\eta\mathrm{grad} f(X)) = X - 2\eta(\nabla f(X)X + X\nabla f(X)) + 4\eta^2\nabla f(X)X\nabla f(X)$. Consider a basis $E_{ij}X + XE_{ij}$ where $E_{ij} = e_ie_j^\top + e_je_i^\top$. The coordinate derivative is computed as $\theta_{ij} = \langle E_{ij}X + XE_{ij}, \nabla f(X)\rangle$. Finally, the CD update is given by $X - 2\eta\theta_{ij}(E_{ij}X + XE_{ij}) + 4\eta^2\theta_{ij}^2E_{ij}XE_{ij}$. Each CD update modifies two rows and two columns of X.

Remark 3.11. For the SPD manifold ($p = n$), Darmwal & Rajawat (2023) propose CD updates based on the affine-invariant metric and Cholesky factorization. They specifically focus on a class of objective functions and show that the exponential map computations are efficient. In contrast, our choices of parameterization/metric directly leads to a faster retraction.

4. Algorithms and Analysis

RCD. We present the proposed Riemannian coordinate descent (RCD) algorithm in Algorithm 1. The complexity of RCD per iteration is the complexity of one first-order oracle and the update complexity in Table 1. Although for some problem settings, we may explore the structure of the objective to efficiently compute $\theta = \langle \nabla f(X), B_\ell\rangle$, for general problem instances, $\nabla f(X)$ becomes the main computational bottleneck.

Algorithm 1 Riemannian coordinate descent (RCD/RCDlin)

Initialize $X_0 \in \mathcal{M}$. $\mathcal{I}$ denotes the index set. Given hyper-parameters K, S, and η.
for $k = 0, ..., K-1$ **do**
  $X_k^0 = X_k$.
  for $s = 0, ..., S-1$ **do**
    Pick $\ell_k^s \in \mathcal{I}$ and construct basis $B_{\ell_k^s}$ at X_k^s.
    $\theta_k^s = \begin{cases} \langle \nabla f(X_k^s), B_{\ell_k^s}\rangle, & \text{if RCD,} \\ \langle \nabla f(X_k), B_{\ell_k^s}\rangle, & \text{if RCDlin.} \end{cases}$
    Update $X_k^{s+1} = \mathrm{Retr}_{X_k^s}(-\eta\theta_k^s B_{\ell_k^s})$.
  end for
  $X_{k+1} = X_k^S$.
end for

RCDLin. To reduce gradient computations in the RCD setup (especially for non-linear objectives), we also propose the Riemannian linearized coordinate descent (RCDlin) method in Algorithm 1. The main difference with RCD is that the variables are updated using an anchored gradient at X_k (which does not change for inner iterations). This scheme is equivalent to taking a linearization of the original cost function at X_k, and in the inner iterations, we solve: $\min_{X\in\mathcal{M}} \{g(X) := f(X_k) + \langle \nabla f(X_k), X - X_k\rangle\}$, where the inner product and subtraction are defined in the ambient Euclidean space. Subsequently, the Euclidean gradient at X_k^s is $\nabla g(X_k^s) = \nabla f(X_k)$, and thus, $\theta_k^s = \langle \nabla f(X_k), B_{\ell_k^s}\rangle$. For the randomized setting with $S = 1$, RCDlin is equivalent to RCD. Additionally, for linear problems where $\nabla f(X_k) = C$ (C is some constant matrix), RCDlin also reduces to RCD.

Convergence and complexity of Algorithm 1

We next discuss the convergence analysis of RCD and RCDlin. It follows the standard analysis for CD algorithms (Wright, 2015). Note that Gutman & Ho-Nguyen (2023) mostly consider the analysis of CD algorithms under exponential map and parallel transport operations. In contrast, we consider the more general retraction and vector transport operations. We also adapt our analysis for RCDlin. For brevity, our analysis is informally discussed here. The analysis is in a compact neighbourhood around a critical point, which is required for validating certain regularity assumptions, boundedness of basis and projection onto the basis, and smoothness of the objective (details in Appendix F).

On RCD. We start by showing the convergence of RCD under randomized selection of basis and certain regularity

assumptions. L is the smoothness constant of the objective.

Theorem 4.1 (Randomized RCD)**.** *Under mild assumptions, consider RCD with $S = 1$ and ℓ_k^s selected uniformly at random from $\mathcal{I}$. Then, choosing $\eta = \Theta(\frac{1}{L})$ leads to* $\min_{0\le k\le K-1} \mathbb{E}\|\mathrm{grad} f(X_k)\|_{X_k}^2 \le O(\frac{|\mathcal{I}|L}{K})$.

The convergence of RCD with cyclic selection of basis requires further assumptions that bound the difference of the constructed bases between tangent spaces. These are reasonable given the compactness of the domain.

Theorem 4.2 (Cyclic RCD)**.** *Under mild assumptions in addition to the ones required by randomized RCD, consider RCD with $S = |\mathcal{I}|$ and $\ell_k^s = s+1$ for $s = 0, ..., |\mathcal{I}|-1$. Then, for $\eta = \Theta(\frac{1}{L})$, we have* $\min_{0\le k\le K-1} \|\mathrm{grad} f(X_k)\|_{X_k} \le O(\frac{|\mathcal{I}|^2 L}{K})$.

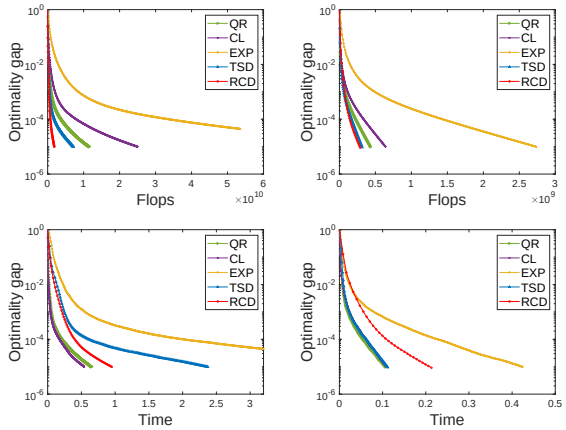
On RCDlin. The key idea here is to relate the coordinate derivative $\theta_k^s = \langle \nabla f(X_k), B_{\ell_k^s}\rangle$ to the correct descent derivative $\langle \nabla f(X_k^s), B_{\ell_k^s}\rangle$. In randomized settings, we can show the same convergence rate as RCD up to some additional constants regulated by the difference between θ_k^s and the descent direction. For the cyclic settings, however, we require $S = |\mathcal{I}|$ in order to cycle through all the basis.

Theorem 4.3 (Randomized and cyclic RCDlin)**.** *Under assumptions required in Theorems 4.1 & 4.2, suppose θ_k^s and $\langle \nabla f(X_k^s), B_{\ell_k^s}\rangle$ are positively related. Then, consider randomized RCDlin with ℓ_k^s selected uniformly at random from $\mathcal{I}$. Choosing $\eta = \Theta(\frac{1}{L})$ leads to* $\min_{0\le k\le K-1, 0\le s\le S-1} \mathbb{E}\|\mathrm{grad} f(X_k^s)\|_{X_k^s}^2 \le (\frac{|\mathcal{I}|L}{KS})$. *In addition, consider cyclic RCDlin with $S = |\mathcal{I}|$ and $\ell_k^s = s+1$ for $s = 0, ..., |\mathcal{I}|-1$. Also, if $\eta = \Theta(\frac{1}{L})$, then* $\min_{0\le k\le K-1} \|\mathrm{grad} f(X_k^s)\|_{X_k^s}^2 \le O(\frac{|\mathcal{I}|^2 L}{K})$.

Complexity analysis. Let the cost of computing the coordinate derivative θ and CD update be δ (last column of Table 1). Then, the total computational cost of RCD and RCDlin is $O(KS(\mathcal{F}+\delta))$ and $O(K(\mathcal{F}+S\delta))$, respectively, where $\mathcal{F}$ denotes the cost of computing $\nabla f(X)$. We note that the proposed algorithms can parallely update in disjoint basis directions. For example, in the Stiefel/Grassmann case, we can select $n/2$ non-overlapping index pairs, which results in $n/2$ independent Givens rotation, and can be parallelized.

5. Experiments

We now benchmark the performance of the proposed RCD and RCDlin algorithms in terms of computational efficiency (flop counts and/or runtime) and convergence quality (distance to optimality). One of the considered baselines is the Riemannian gradient descent method (RGD), a full gradient method. As RGD exploits the entire gradient direction, it has advantage over CD algorithms. However, RGD is significantly more costly than CD in every update. Our codes are implemented using the Manopt toolbox (Boumal et al., 2014) and run on a laptop with an i5-10500 3.1GHz CPU processor. The codes are available at `https://github.com/andyjm3`.



(a) $n = 200, p = 150$ (b) $n = 200, p = 50$

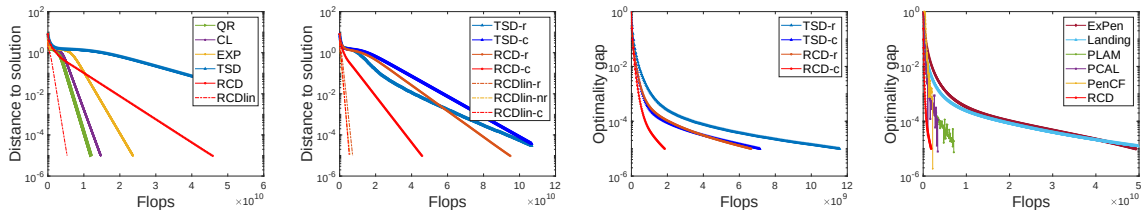
Figure 1: The Procrustes problem with varying p: (a) $p = 150$ and (b) $p = 50$. (Top row) Comparing various algorithms in terms of flop counts. (Bottom row) Comparing various algorithms in terms of runtime. We observe that our RCD algorithm obtains better flop counts than the baselines in flop counts and is competitive in terms of runtime.

5.1. Orthogonal Procrustes and PCA

Orthogonal Procrustes problem. We aim to solve $\min_{X\in \mathrm{St}(n,p)} \|XA - B\|^2 (\equiv -\langle XA, B\rangle)$ for given matrices $A \in \mathbb{R}^{p\times p}$, $B \in \mathbb{R}^{n\times p}$. There exists a closed-form solution provided by the (thin) SVD of $BA^\top$. For this, RCD and RCDlin have same updates as $\nabla f(X) = -BA^\top$ for $X \in \mathrm{St}(n,p)$. In experiments, we generate random matrices A, B and evaluate the performance against optimality gap computed as $|f(X_k) - f^*|/|f^*|$.

Baselines. The closest baseline to RCD is TSD (Gutman & Ho-Nguyen, 2023), which is a CD method under an alternative construction of bases. As discussed, while TSD updates the columns, the proposed RCD updates the rows. Since RCD is equivalent to TSD for $p = n$, we focus only on the $p < n$ setting. For both RCD and TSD, we use the cyclic selection of basis. We also compare against RGD methods with QR, Cayley (CL), and exponential (EXP) retractions. For all the methods, we tune the stepsize.

Results. In Figure 1, we show results with varying dimension p. While the proposed RCD obtains better flop counts than the baselines in flop counts, it is competitive in terms of runtime. We highlight that the runtime of RCD can be further improved with parallel implementation. In Figure 2c, we compare a variety of basis selection rules for both



(a) PCA (b) PCA (c) Procrustes (d) Procrustes

Figure 2: (a) & (b): Experiments on the PCA problem with $n = 200, p = 50$. In (a), we observe that our algorithm RCDlin achieves the fastest convergence due to low per-iteration cost. In (b), we compare various strategies for basis selection: cyclic selection (-c) and uniformly random selection (-r) of basis for TSD, RCD, and RCDlin, and selection without replacement (-nr) for RCDlin. We observe that cyclic and selection without replacement strategies are better than random selection. (c) & (d): Experiments on the Procrustes problem with $n = 200, p = 150$. In (c), we again observe that cyclic selection performs better than random selection. In (d), RCD performs competitively against the infeasible methods.

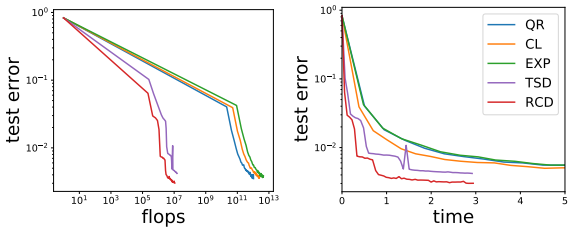
TSD and RCD: cyclic selection ('c') and uniformly random selection ('r'). We observe that cyclic selection is more favourable than the random selection rule for both the methods. We compare against full gradient infeasible methods in Figure 2d, including PLAM (Gao et al., 2019), PCAL (Gao et al., 2019), PenCF (Xiao et al., 2022), ExPen (Xiao & Liu, 2021; Xiao et al., 2023), and Landing (Ablin & Peyré, 2022; Ablin et al., 2023). RCD is performs competitively against infeasible methods for orthogonality constraints.

PCA problem. The PCA problem solves a quadratic maximization problem as $\max_{X^\top X = I_p} \operatorname{tr}(X^\top A X)$ for some positive definite matrix A, i.e., $A \in \mathbb{S}^n_{++}$. This problem is in fact an optimization problem over the Grassmann manifold because the objective is invariant to basis change of X. Hence, we use the Riemannian distance to the optimal solution on the Grassmann manifold to measure the performance. As discussed in Section 3.2, our proposed RCD has well-defined updates on the Grassmann manifold. In contrast, TSD is not invariant to the basis change. For experiments, we generate A with a condition number 10^3 and with exponential decay of eigenvalues. For TSD, RCD, and RCDlin, we implement the cyclic selection of basis.

Results. In Figure 2a, we observe that RCDlin achieves the best performance due to its low per-iteration cost. We note that TSD converges slowly due to non-invariance of the CD updates. In Figure 2b, we compare the cyclic and uniformly random selection of the basis of RCD, RCDlin, and TSD. For RCDlin, we also implement the selection without replacement ('nr') strategy. We observe that cyclic and 'nr' strategies are better than random selection.

5.2. Orthogonal deep networks for distillation

We next evaluate RCD on a deep learning based distillation problem (Hinton et al., 2015). Let Θ denote the parameters of the student network (S) while Θ_{T} be the optimal parameters of the teacher network (T). Then, the aim



(a) Distill (flops) (b) Distill (time)

Figure 3: Experiments on the distillation problem. We observe that the proposed RCD algorithm performs better than the baselines both in terms of flop counts and runtime.

is to learn S that approximates T, i.e., minimize $\mathcal{L}(\Theta) = \|\Psi_\Theta(X) - \Psi_{\Theta_{\mathrm{T}}}(X)\|^2$, where $\Psi_\Theta(X) \in \mathbb{R}^{N \times d_{\mathrm{out}}}$ represent the output of the network for some input $X \in \mathbb{R}^{N \times d_{\mathrm{in}}}$. The network architecture is detailed in Appendix A.1. Here, we constrain all the weights to be orthonormal, thus posing the problem as optimization over the joint space of Stiefel and Euclidean manifolds. For experiments, we consider a six-layer network and set $d_{\mathrm{in}} = 500$, $d_{\mathrm{out}} = 200$. We use stochastic versions of RGD and RCD where the input samples are randomly generated. In Figure 3, RCD outperforms the baselines in terms of flop counts and runtime. This is because RCD has the most cost-efficient update per iteration, while maintaining a competitive convergence rate.

5.3. Nearest matrix problem

We consider the problem: $\min_{X \in \mathrm{Sp}(n,p)} \|X - A\|^2$ on the symplectic manifold (Gao et al., 2021b). We follow the setting in (Gao et al., 2021b) by generating A as a random matrix with $p = n = 200$. The algorithms are evaluated on optimality gap $|f(X_k) - f^*|/|f^*|$, where f^* is obtained by running the conjugate gradient algorithm with the Cayley retraction. We implement RCD and RCDlin with both CD and block CD updates (discussed in Remark 3.8). As there

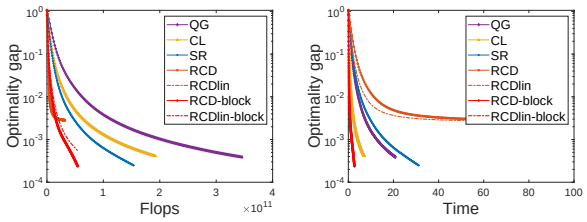


Figure 4: Experiments on the nearest matrix problem. We notice the utility of the block-update variants of our RCD and RCDlin algorithms in obtaining faster convergence.

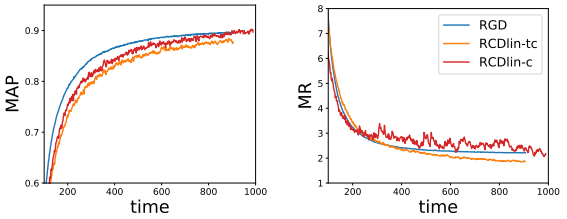


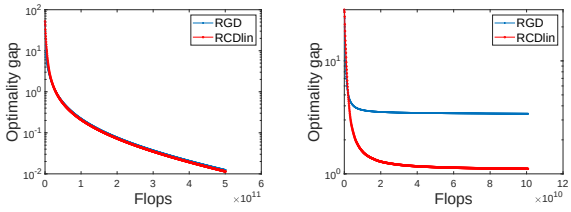
Figure 5: Experiments on learning Lorentz (hyperbolic) embeddings. The performance of our RCDlin algorithms (with cyclic and time-cyclic basis selection) is competitive to RGD.

is no CD baseline on the symplectic manifold, we compare against the full gradient RGD algorithms with three retractions: Cayley ('CL'), quasi-geodesic ('QG'), and SR ('SR'). In Figure 4, RCD with block update shows clear advantage in both flop counts and runtime.

5.4. Learning Lorentz embeddings

We consider the task of learning embeddings for word hierarchies, which is formulated on the hyperbolic manifold using the hyperboloid model (Nickel & Kiela, 2017; 2018; Jawanpuria et al., 2019b). The goal is to map word pairs with hypernymy relations closer while separate those without. We follow the formulation in (Nickel & Kiela, 2018), and the details are in Appendix A.2.

For experiment settings, we train 5-dimensional embeddings ($n = 5$) for WordNet mammals subtree (Miller, 1998). We adopt the RCDlin algorithm with two selection rules for the basis $H_{ij}JX$: cyclic ('RCDlin-c') and time cyclic ('RCDlin-tc'). The cyclic selection loops through all $n(n-1)/2$ pairs per iteration. The time cyclic selection only loops through all the space-time coordinate pairs, namely $(1,2),(1,3),...,(1,n)$, which reduces the computational cost to scale linearly with dimension n. For RCDlin-c and RCDlin-tc, we use a linearly-decaying stepsize, i.e., $\eta/(1+0.1\times\text{epoch})$. For RGD we use a fixed stepsize η which generally leads to better convergence. We tune and set $\eta = 1.0$ for RCDlin and 0.5 for RGD. We use the metrics for evaluating the convergence: mean average precision (MAP) and mean rank (MR) (Nickel & Kiela, 2017; 2018). In Figure 5, we see that RCDlin converges at a similar rate compared to RGD in terms of runtime.



(a) Optimization over SPD manifold with dense A

(b) Optimization over SPSD manifold with sparse A

Figure 6: Experiments on the weighted least squares problem in two settings: (a) $n = p = 500$ and $A = 1_n 1_n^\top$ is a dense matrix and (b) $n = 500, p = 100$ and A is a random symmetric matrix with 70% entries as 1 and others are 0. While RGD and the proposed RCDlin have similar convergence rate in (a), RCDlin has clear advantage in (b).

5.5. Weighted least squares (SPSD manifold)

The weighted least squares problem is $\min_{X\in\mathbb{S}_+^{n\times p}} \|A\odot X - B\|$, where $A\in\{0,1\}^{n\times n}$ masks the known entries in B. It is an instance of the matrix completion problem (Han et al., 2021). For the experiment, we follow (Han et al., 2021) by generating $B = A\odot X^*$ where X^* is an SPD/SPSD matrix with exponentially decaying eigenvalues. We consider two settings: (left) $n = p = 500$, $A = 1_n 1_n^\top$ is a dense matrix and (right) $n = 500, p = 100$ and A is a random symmetric matrix with 70% entries as 1 and others are 0. We compare RCDlin with RGD (for $X = YY^\top$ factorization). We set $S = np/5$ and select the coordinates randomly without replacement. In Figure 6, we observe that RCDlin performs competitively with RGD on the SPD manifold with dense A while performing significantly better on the SPSD manifold with sparse A.

6. Conclusion

In this work, we discuss how to develop computationally efficient CD updates for a number of matrix manifolds. The main bottleneck in developing CD methods is on finding the right basis parameterization of the tangent space. We show precise constructions for various manifolds and propose two CD algorithms: RCD and RCDlin. RCDlin specifically reduces the gradient computations of RCD further. Our experiments show the benefit of our proposed CD updates on a number of problem instances.

Impact Statement

This paper presents work whose goal is to advance optimization methods with applications in Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ablin, P. and Peyré, G. Fast and accurate optimization on the orthogonal manifold without retraction. In *International Conference on Artificial Intelligence and Statistics*, pp. 5636–5657. PMLR, 2022.

Ablin, P., Vary, S., Gao, B., and Absil, P.-A. Infeasible deterministic, stochastic, and variance-reduction algorithms for optimization under orthogonality constraints. *arXiv:2303.16510*, 2023.

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

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

Bai, Z. and Li, R.-C. Minimization principles and computation for the generalized linear response eigenvalue problem. *BIT Numerical Mathematics*, 54:31–54, 2014.

Bhatia, R. *Positive definite matrices*. Princeton University Press, 2009.

Bhatia, R., Jain, T., and Lim, Y. On the bures–wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

Bonnabel, S. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58 (9):2217–2229, 2013.

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

Boumal, N., Mishra, B., Absil, P.-A., and Sepulchre, R. Manopt, a matlab toolbox for optimization on manifolds. *The Journal of Machine Learning Research*, 15(1):1455–1459, 2014.

Cardoso, J. R. and Leite, F. S. Exponentials of skew-symmetric matrices and logarithms of orthogonal matrices. *Journal of Computational and Applied Mathematics*, 233(11):2867–2875, 2010.

Darmwal, Y. and Rajawat, K. Low-complexity subspace-descent over symmetric positive definite manifold. Technical report, arXiv preprint arXiv:2305.02041, 2023.

Douik, A. and Hassibi, B. Manifold optimization over the set of doubly stochastic matrices: A second-order geometry. *IEEE Transactions on Signal Processing*, 67 (22):5761–5774, 2019.

Edelman, A., Arias, T. A., and Smith, S. T. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Gao, B., Liu, X., and Yuan, Y.-x. Parallelizable algorithms for optimization problems with orthogonality constraints. *SIAM Journal on Scientific Computing*, 41(3):A1949–A1983, 2019.

Gao, B., Son, N. T., Absil, P.-A., and Stykel, T. Geometry of the symplectic Stiefel manifold endowed with the Euclidean metric. In *Geometric Science of Information*, pp. 789–796. Springer, 2021a.

Gao, B., Son, N. T., Absil, P.-A., and Stykel, T. Riemannian optimization on the symplectic Stiefel manifold. *SIAM Journal on Optimization*, 31(2):1546–1575, 2021b.

Gao, B., Son, N. T., and Stykel, T. Optimization on the symplectic Stiefel manifold: SR decomposition-based retraction and applications. Technical report, arXiv preprint arXiv:2211.09481, 2022.

Gutman, D. H. and Ho-Nguyen, N. Coordinate descent without coordinates: Tangent subspace descent on Riemannian manifolds. *Mathematics of Operations Research*, 48(1):127–159, 2023.

Han, A. and Gao, J. Improved variance reduction methods for Riemannian non-convex optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (11):7610–7623, 2021.

Han, A., Mishra, B., Jawanpuria, P. K., and Gao, J. On Riemannian optimization over positive definite matrices with the Bures-Wasserstein geometry. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8940–8953, 2021.

Han, A., Mishra, B., Jawanpuria, P., and Gao, J. Riemannian block SPD coupling manifold and its application to optimal transport. *Machine Learning*, pp. 1–28, 2022.

Han, A., Mishra, B., Jawanpuria, P., and Gao, J. Nonconvex-nonconcave min-max optimization on riemannian manifolds. *Transactions on Machine Learning Research*, 2023a. ISSN 2835-8856.

Han, A., Mishra, B., Jawanpuria, P., Kumar, P., and Gao, J. Riemannian hamiltonian methods for min-max optimization on manifolds. *SIAM Journal on Optimization*, 33(3): 1797–1827, 2023b.

Han, A., Mishra, B., Jawanpuria, P., and Gao, J. Differentially private Riemannian optimization. *Machine Learning*, 113(3):1133–1161, 2024a.

Han, A., Mishra, B., Jawanpuria, P., and Takeda, A. A framework for bilevel optimization on Riemannian manifolds. Technical report, arXiv preprint arXiv:2402.03883, 2024b.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. Technical report, arXiv preprint arXiv:1503.02531, 2015.

Huang, M., Ma, S., and Lai, L. A Riemannian block coordinate descent method for computing the projection robust Wasserstein distance. In *International Conference on Machine Learning*, pp. 4446–4455. PMLR, 2021.

Huang, W., Absil, P.-A., and Gallivan, K. A. A Riemannian symmetric rank-one trust-region method. *Mathematical Programming*, 150(2):179–216, 2015.

Huang, Z., Huang, W., Jawanpuria, P., and Mishra, B. Federated learning on Riemannian manifolds with differential privacy. Technical report, arXiv preprint arXiv:2404.10029, 2024.

Jawanpuria, P. and Mishra, B. A unified framework for structured low-rank matrix learning. In *International Conference on Machine Learning (ICML)*, 2018.

Jawanpuria, P., Balgovind, A., Kunchukuttan, A., and Mishra, B. Learning multilingual word embeddings in a latent metric space: a geometric approach. *Transactions of the Association for Computational Linguistics*, 7(3): 107–120, 2019a.

Jawanpuria, P., Meghwanshi, M., and Mishra, B. Low-rank approximations of hyperbolic embeddings. In *IEEE Conference on Decision and Control*, 2019b.

Jawanpuria, P., Meghwanshi, M., and Mishra, B. Geometry-aware domain adaptation for unsupervised alignment of word embeddings. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

Jiang, Y., Zhang, H., Qiu, Y., Xiao, Y., Long, B., and Yang, W.-Y. Givens coordinate descent methods for rotation matrix learning in trainable embedding indexes. In *International Conference on Learning Representations*, 2022.

Kasai, H., Jawanpuria, P., and Mishra, B. Riemannian adaptive stochastic gradient algorithms on matrix manifolds. In *International Conference on Machine Learning (ICML)*, 2019.

Knight, P. A. The Sinkhorn–Knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

Kressner, D., Steinlechner, M., and Vandereycken, B. Low-rank tensor completion by Riemannian optimization. *BIT Numerical Mathematics*, 54:447–468, 2014.

Lezcano Casado, M. Trivializations for gradient-based optimization on manifolds. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Li, J. and Ma, S. Federated learning on Riemannian manifolds. Technical report, arXiv preprint arXiv:2206.05668, 2022.

Luo, Z.-Q. and Tseng, P. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, 1992.

Massart, E. and Abrol, V. Coordinate descent on the orthogonal group for recurrent neural network training. In *AAAI Conference on Artificial Intelligence*, volume 36, pp. 7744–7751, 2022.

Massart, E. and Absil, P.-A. Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices. *SIAM Journal on Matrix Analysis and Applications*, 41(1):171–198, 2020.

Miller, G. A. *WordNet: An electronic lexical database*. MIT press, 1998.

Mishra, B., Kasai, H., Jawanpuria, P., and Saroop, A. A Riemannian gossip approach to subspace learning on Grassmann manifold. *Machine Learning*, 108(10):1783–1803, 2019.

Mishra, B., Satyadev, N., Kasai, H., and Jawanpuria, P. Manifold optimization for non-linear optimal transport problems. Technical report, arXiv preprint arXiv:2103.00902, 2021.

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

Nickel, M. and Kiela, D. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Nickel, M. and Kiela, D. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, pp. 3779–3788. PMLR, 2018.

Nimishakavi, M., Jawanpuria, P., and Mishra, B. A dual framework for low-rank tensor completion. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Peng, L. and Vidal, R. Block coordinate descent on smooth manifolds. Technical report, arXiv preprint arXiv:2305.14744, 2023.

Pennec, X., Fillard, P., and Ayache, N. A Riemannian Framework for Tensor Computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

Reimherr, M., Bharath, K., and Soto, C. Differential privacy over riemannian manifolds. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12292–12303, 2021.

Sato, H., Kasai, H., and Mishra, B. Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport. *SIAM Journal on Optimization*, 29(2): 1444–1472, 2019.

Shalit, U. and Chechik, G. Coordinate-descent for learning orthogonal matrices through Givens rotations. In *International Conference on Machine Learning*, pp. 548–556. PMLR, 2014.

Shi, D., Gao, J., Hong, X., Boris Choy, S., and Wang, Z. Coupling matrix manifolds assisted optimization for optimal transport problems. *Machine Learning*, 110:533–558, 2021.

Siegel, J. W. Accelerated optimization with orthogonality constraints. *Journal of Computational Mathematics*, 39 (2):207–206, 2020.

Sinkhorn, R. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

Tripuraneni, N., Flammarion, N., Bach, F., and Jordan, M. I. Averaging stochastic gradient descent on Riemannian manifolds. In *Conference on Learning Theory*, pp. 650–687. PMLR, 2018.

Utpala, S., Han, A., Jawanpuria, P., and Mishra, B. Improved differentially private Riemannian optimization: Fast sampling and variance reduction. *Transactions on Machine Learning Research*, 2022.

Vandereycken, B. Low-rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization*, 23(2): 1214–1236, 2013.

Vandereycken, B., Absil, P.-A., and Vandewalle, S. Embedded geometry of the set of symmetric positive semidefinite matrices of fixed rank. In *IEEE/SP Workshop on Statistical Signal Processing*, pp. 389–392. IEEE, 2009.

Vandereycken, B., Absil, P.-A., and Vandewalle, S. A Riemannian geometry with complete geodesics for the set of positive semidefinite matrices of fixed rank. *IMA Journal of Numerical Analysis*, 33(2):481–514, 2013.

Wang, J., Chen, Y., Chakraborty, R., and Yu, S. X. Orthogonal convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition*, pp. 11505–11515, 2020.

Wen, Z. and Yin, W. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013.

Wilson, B. and Leimeister, M. Gradient descent in hyperbolic space. Technical report, arXiv preprint arXiv:1805.08207, 2018.

Wright, S. J. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

Xiao, N. and Liu, X. Solving optimization problems over the Stiefel manifold by smooth exact penalty function. Technical report, arXiv preprint arXiv:2110.08986, 2021.

Xiao, N., Liu, X., and Yuan, Y.-x. A class of smooth exact penalty function methods for optimization problems with orthogonality constraints. *Optimization Methods and Software*, 37(4):1205–1241, 2022.

Xiao, N., Liu, X., and Toh, K.-C. Dissolving constraints for Riemannian optimization. *Mathematics of Operations Research*, 2023.

Yuan, G. A block coordinate descent method for nonsmooth composite optimization under orthogonality constraints. Technical report, arXiv preprint arXiv:2304.03641, 2023.

Zhang, H., J Reddi, S., and Sra, S. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Zhang, P., Zhang, J., and Sra, S. Sion's minimax theorem in geodesic metric spaces and a Riemannian extragradient algorithm. *SIAM Journal on Optimization*, 33(4):2885–2908, 2023.

Zhu, X. A Riemannian conjugate gradient method for optimization on the Stiefel manifold. *Computational Optimization and Applications*, 67:73–110, 2017.

A. Additional experiment details

The experiments are run on a laptop with an i5-10500 3.1GHz CPU processor.

A.1. Orthogonal deep networks for distillation

Here we provide the detailed network architecture for the distillation task. In particular, we define a L-layer feed-forward neural network with Tanh activation function, i.e., $X_{\ell+1} = \tanh(W_\ell X_\ell + b_\ell)$ for $\ell \in [L]$, where $W_1 \in \mathbb{R}^{d_{\text{in}} \times d}, W_L \in \mathbb{R}^{d \times d_{\text{out}}}$, and $W_\ell \in \mathbb{R}^{d \times d}$ for $\ell \neq 1, L$. In the experiment, we set $L = 6$.

A.2. Learning Lorentz embeddings

Here we provide the problem formulation for the task of learning Lorentz embeddings. Let $\mathcal{D} = \{(u, v)\}$ be the related word pairs and construct $\text{Neg}(u) = \{v : (u, v) \notin \mathcal{D}\}$ as the negative samples of word u. The objective is to learn embeddings x_u for all word u by solving

$$\min_{\{x_u \in \mathcal{H}(n,1)\}_u} \sum_{(u,v)\in\mathcal{D}} \log \frac{\exp\left(-\text{dist}(x_u, x_v)\right)}{\sum_{v' \in \text{Neg}(u)} \exp\left(-\text{dist}(u, v')\right)},$$

where $\text{dist}(x_u, x_v) = \text{arccosh}(-\langle x_u, x_v \rangle_{\mathcal{L}})$ is the Lorentz Riemannian distance.

B. A review on coordinate descent for orthogonal and SPD manifold

We start by reviewing the developments of coordinate descent on the orthogonal manifold (Shalit & Chechik, 2014; Massart & Abrol, 2022; Jiang et al., 2022), Stiefel manifold (Gutman & Ho-Nguyen, 2023), and symmetric positive definite (SPD) manifold (Darmwal & Rajawat, 2023), which motivate the proposed general framework for other manifolds.

Some other works (Huang et al., 2021; Peng & Vidal, 2023) study (block) coordinate descent on a product of manifolds, where each update concerns a component manifold. This is different to our considered setting, where the update is defined for coordinate on the tangent space for a single manifold.

B.1. Orthogonal manifold

Orthogonal manifold $\mathcal{O}(n)$ is the smooth space formed by the orthogonality constraints, i.e., $\mathcal{O}(n) \coloneqq \{X \in \mathbb{R}^{n\times n} : XX^\top = X^\top X = I_n\}$. The tangent space can be identified as $T_X\mathcal{O}(n) \coloneqq \{\Omega X : \Omega \in \text{Skew}(n)\}$. The Riemannian metric coincides with the Euclidean metric, i.e., $\langle U, V\rangle_X = \frac{1}{2}\langle U, V\rangle$. The $\frac{1}{2}$ is added to ensure consistency with the canonical metric for the Stiefel manifold, as we shall see later. This leads to the Riemannian gradient $\text{grad} f(X) = \nabla f(X) - X\nabla f(X)^\top X$ and the exponential retraction is given by $\text{Retr}_X(\theta\Omega X) = \text{expm}(\theta\Omega)X$, for some $\theta \in \mathbb{R}$.

Remark B.1. We remark that in all the existing works (Shalit & Chechik, 2014; Massart & Abrol, 2022; Jiang et al., 2022), the tangent space is parameterized as $T_X\mathcal{O}(n) \coloneqq \{X\Omega' : \Omega' \in \text{Skew}(n)\}$ and thus the exponential retraction amounts to $\text{Retr}_X(\theta X\Omega') = X\text{expm}(\theta\Omega')$. Such a formulation is equivalent to the above by letting $\Omega = X\Omega' X^\top \in \text{Skew}(n)$. Our reformulation allows natural generalization of the coordinate descent framework to column orthonormal matrices (the Stiefel manifold), where the orthogonal matrix is a special case.

The manifold has a dimension of $n(n-1)/2$ and its tangent space can be provided with an orthonormal basis $H_{ij}X$ where $H_{ij} \coloneqq e_i e_j^\top - e_j e_i^\top$ is the basis for the skew-symmetric matrices. In each basis direction $H_{ij}X$, the exponential retraction reduces to the Givens rotation, which allows efficient updates, i.e., $\text{Retr}_X(\theta H_{ij}X) = G_{ij}(\theta)X$ where $G_{ij}(\theta) = I_n + (\cos(\theta) - 1)(e_i e_i^\top + e_j e_j^\top) + \sin(\theta)(e_i e_j^\top - e_j e_i^\top)$ is known as the Givens rotation matrix around axes i, j with angle $-\theta$.

In order to minimize a function $f : \mathcal{O}(n) \to \mathbb{R}$, one needs to update the variables in the negative gradient direction. Here along the basis direction, coordinate descent aims to minimize the function $f(G_{ij}(\theta)X)$ with respect to θ. One strategy is to solve this one-variable optimization problem directly as in (Shalit & Chechik, 2014). When the objective is more involved, we can approximately solve this problem by following a descent direction (Massart & Abrol, 2022; Jiang et al., 2022), which is given by $-\frac{d}{d\theta} f(G_{ij}(\theta)X)|_{\theta=0} = -\langle \text{grad} f(X), H_{ij}X\rangle_X$. This leads to the coordinate descent update in the direction of $-\langle \text{grad} f(X), H_{ij}X\rangle_X H_{ij}X$, which modifies two rows of X every iteration, resulting in an $O(n)$ complexity per update. One pass over all the coordinates requires $\frac{n(n-1)}{2}$ iterations, leading to $O(n^3)$ complexity in total, which is comparable to the commonly considered retractions, including the exponential, Cayley and QR retractions.

B.2. Stiefel manifold

The Stiefel manifold $\mathrm{St}(n,p)$ is the set of column orthonormal matrices of size $\mathbb{R}^{n\times p}$, i.e., $\mathrm{St}(n,p) := \{X \in \mathbb{R}^{n\times p} : X^\top X = I_p\}$. When $p = n$, $\mathrm{St}(n,n) \equiv \mathcal{O}(n)$. The tangent space of Stiefel manifold is identified as $T_X\mathrm{St}(n,p) = \{U \in \mathbb{R}^{n\times p} : X^\top U + U^\top X = 0\} = \{X\Omega + X_\perp K : \Omega \in \mathrm{Skew}(p), K \in \mathbb{R}^{(n-p)\times p}\}$. The Euclidean metric turns out to be a valid Riemannian metric (Edelman et al., 1998), $\langle U, V\rangle_X = \langle U, V\rangle$ for any $U, V \in T_X\mathcal{M}$. The Riemannian gradient is derived as $\mathrm{grad} f(X) = \nabla f(X) - X\mathrm{sym}(X^\top \nabla f(X))$.

In (Gutman & Ho-Nguyen, 2023), a coordinate (subspace) descent algorithm has been developed for general manifolds, via selecting proper subspaces for the tangent space. Although showing theoretical guarantees, the paper only provides a concrete developments for Stiefel manifold (thus including the orthogonal manifold). The basis considered for the tangent space of Stiefel manifold is $\{XH_{ij}\}_{1\le i<j\le n} \cup \{ve_k^\top : X^\top v = 0\}_{k\in[p]}$, where $H_{ij} = e_ie_j^\top - e_je_i^\top$. They show along the direction of basis, the exponential retraction can be computed efficiently. That is, for basis XH_{ij}, we compute $\mathrm{Retr}_X(tXH_{ij}) = XG_{ij}(t)$, which also leads to the Givens rotation. The projection of Riemannian gradient onto the basis is given by $\langle \mathrm{grad} f(X), XH_{ij}\rangle XH_{ij}$. For basis $ve_k^\top$, $\mathrm{Retr}_X(ve_k^\top) = X + \big(\cos(\|v\|)[X]_{:,j} + \sin(\|v\|)\frac{v}{\|v\|} - [X]_{:,j}\big)e_k^\top$ and the projection of Riemannian gradient is $(I_n - XX^\top)[\nabla f(X)]_{:,k}e_k^\top$. This essentially updates j-th column by Riemannian gradient descent over sphere. We notice that each coordinate update costs $O(n)$ while the projection onto second basis costs $O(np)$.

Further, we highlight a recent paper (Yuan, 2023), which considers block coordinate descent updates for Stiefel manifold by modifying k rows. The strategy is to decompose the the rows of the variable into two sets and solve for a subproblem that updates k rows instead. Nevertheless, each subproblem can be difficult to solve in general.

B.3. SPD manifold

A recent work (Darmwal & Rajawat, 2023) develops coordinate (subspace) descent algorithm for symmetric positive definite (SPD) manifold, i.e., $\mathbb{S}^n_{++} := \{X \in \mathbb{R}^{n\times n} : X \succ 0\}$, with tangent space given by $T_X\mathbb{S}^n_{++} = \{X \in \mathbb{R}^{n\times n} : X \in \mathrm{Sym}(n)\}$. Optimization with SPD constraint is difficult as the cost of maintaining positive definiteness requires at least $O(n^3)$. Under the affine-invariant metric (Bhatia, 2009), i.e., $\langle U, V\rangle_X = \mathrm{tr}(UX^{-1}VX^{-1})$, the paper verifies that the tangent space can be provided with an orthonormal basis $LE_{ij}L^\top$ where $LL^\top = X$ is the Cholesky decomposition and $E_{ij} = (e_ie_j^\top + e_je_i^\top)/\sqrt{2}$ for $i \neq j$ and $e_ie_j^\top$ for $i = j$. The exponential retraction along the basis direction can be simplified as $\mathrm{Retr}_X(tLE_{ij}L^\top) = L\mathrm{expm}(tE_{ij})L^\top$, where the Cholesky factor of $\mathrm{expm}(tE_{ij})$ has a simple form. Thus it suffices to update the Cholesky factor. The coordinate descent then updates as $\mathrm{Retr}_X(-\eta\theta LE_{ij}L^\top)$ where θ is computed as the coordinate derivative $\theta = \langle \mathrm{grad} f(X), LE_{ij}L^\top\rangle_X = \langle \nabla f(X), LE_{ij}L^\top\rangle$ because the Riemannian gradient has the form $\mathrm{grad} f(X) = X\nabla f(X)X$.

C. Other related works

This section summarizes the existing works for optimization under the respective manifold constraints. These methods exploit the full gradient information and in general have advantage over CD methods.

Optimization on orthogonal, Stiefel, and Grassmann manifolds. Optimization with orthogonality constraint has been widely studied. Apart from the Riemannian optimization approach, many recent works turn to infeasible methods, either through converting the orthogonality constrained problem into an unconstrained counterpart in the Euclidean space (Xiao et al., 2022; Lezcano Casado, 2019; Xiao & Liu, 2021; Xiao et al., 2023), or following a direction that leads to the same critical point on the manifold (Gao et al., 2019; Ablin & Peyré, 2022; Ablin et al., 2023). We provide a through review of these methods in Appendix section D. Although the per-iteration cost is smaller without using a retraction to ensure feasibility, the methods are sensitive to the choice of stepsize and other parameters, mostly a regularization parameter. In the experiment sections, we observe the infeasible methods either require careful tuning of multiple parameters or some stepsize sequence to show good performance.

Optimization on hyperbolic, symplectic and SPSD manifold. Optimization over hyperbolic space $\mathcal{H}(n,p)$ has mostly focused on the case where $p = 1$ (Nickel & Kiela, 2018; Wilson & Leimeister, 2018). Given the exponential map is already efficient in this case, few works have explored more efficient alternatives to Riemannian optimization. Similarly for the symplectic manifold, existing works (Gao et al., 2021b;a; 2022) have focused on the retraction-based Riemannian

optimization approach. Optimization over SPD matrices has a long history and can be solved with semidefinite programming if the objective is convex. For general cost functions and with additional rank constraint, many works (Vandereycken et al., 2009; 2013; Massart & Absil, 2020) have defined Riemannian geometries for the constraint set to leverage the tools from Riemannian optimization.

D. Review of infeasible methods for optimization under orthogonality constraints

This section reviews various infeasible methods for solving optimization problems under orthogonality constraints, i.e.,

$$\min_X f(X), \quad \text{s.t.} X^\top X = I_p. \tag{4}$$

Recall the first-order optimality conditions of the problem are

$$\begin{cases} \nabla f(X) = X\Lambda, \\ X^\top X = I_p \end{cases}$$

where $\Lambda = \Lambda^\top$ is the Lagrangian multipliers. Because Λ is symmetric, we can compute $\Lambda = X^\top \nabla f(X) = \nabla f(X)^\top X = \mathrm{sym}(\nabla f(X)^\top X)$ (by left-multiplying the first equation by $X^\top$). Subsequently, the first condition becomes $\nabla f(X) - X\nabla f(X)^\top X = \mathrm{skew}(\nabla f(X) X^\top)X = 0$, which in fact corresponds to the first-order condition of Riemannian optimization (under the canonical metric). The augmented Lagrangian of (4) is given by

$$\mathcal{L}_\mu(X, \Lambda) = f(X) - \frac{1}{2}\langle \Lambda, X^\top X - I_p\rangle + \frac{\mu}{4}\|X^\top X - I_p\|^2$$

and the Augmented Lagrangian Method (ALM) outlined in (Gao et al., 2019) considers alternately updating X and Λ. However the numerical performance of ALM is sensitive to the choice of μ, which has been empirically verified in (Gao et al., 2019).

PLAM. Gao et al. (2019) propose an alternative update of the Lagrangian multipliers Λ by setting it to $\mathrm{sym}(\nabla f(X)^\top X)$, which is optimal at first-order stationarity, and the proximal linearized augmented Lagrangian (PLAM) algorithm takes the update

$$X_{k+1} = X_k - \eta\Big(\nabla f(X_k) - X_k \mathrm{sym}(\nabla f(X_k)^\top X_k) + \lambda X_k (X_k^\top X_k - I_p)\Big), \tag{5}$$

which corresponds to the $X_{k+1} = \arg\min_X \left\langle \nabla_X \mathcal{L}_\mu(X_k, \Lambda_k), X - X_k \right\rangle + \frac{1}{2\eta}\|X - X_k\|^2$ with $\Lambda_k = \mathrm{sym}(\nabla f(X_k)^\top X_k)$. However, the boundedness of the iterates cannot be expected without setting λ to be sufficiently large and η to be sufficiently small.

PCAL. Gao et al. (2019) further constrain the proximal linearized update in (5) to a redundant column-sphere constraint. This is used to reduce the sensitivity of convergence to λ, μ. The update can be implemented in a column-wise parallel fashion as

$$(X_{k+1})_i = \frac{(X_k)_i - \eta \nabla_{X_i} \mathcal{L}_\mu(X_k, \Lambda_k)}{\|(X_k)_i - \eta \nabla_{X_i} \mathcal{L}_\mu(X_k, \Lambda_k)\|}.$$

PenCF Xiao et al. (2022) consider the merit function (used to analyze function value decrease in (Gao et al., 2019)) as an exact penalty function to be minimized. Specifically, the merit function is given by

$$h(X) = f(X) - \frac{1}{2}\langle \mathrm{sym}(X^\top \nabla f(X)), X^\top X - I_p\rangle + \frac{\beta}{4}\|X^\top X - I_p\|.$$

Without proper constraints, directly minimizing $h(X)$ can be problematic because $h(X)$ can be unbounded. Hence, (Xiao et al., 2022) incorporates constraints to the minimization of $h(X)$ (called PenC), including a ball, convex hull of Stiefel/oblique manifold, which includes the orthogonality constraint set. Then the paper shows the first-order and second-order critical points of PenC match those of the original problem (provided β is sufficiently large). Nevertheless, the gradient $\nabla h(X)$ involves second-order derivatives $\nabla^2 f(X)$. Hence, the paper similarly considers approximating $\nabla h(X)$ with $\nabla f(X) - X\mathrm{sym}(\nabla f(X)^\top X) + \beta X(X^\top X - I_p)$, which is the same as in (Gao et al., 2019) and then project the update to the constraint set.

Landing field Ablin et al. (2023) consider the update

$$X_{k+1} = X_k - \eta\Big(\mathrm{skew}(\nabla f(X_k)X_k^\top)X_k + \lambda X_k(X_k^\top X_k - I_p)\Big),$$

where the direction takes a combination of the Riemannian gradient and a landing direction orthogonal to the Riemannian gradient.

ExPen Xiao & Liu (2021); Xiao et al. (2023) convert the constrained optimization problem to an unconstrained problem by minimizing the following objective:

$$h(X) = f\Big(X(\frac{3}{2}I_P - \frac{1}{2}X^\top X)\Big) + \frac{\beta}{4}\|X^\top X - I_p\|^2.$$

It can be shown that when β is sufficiently large, first-order and second-order critical points recover those of the original problem. When compared to the framework of PCAL/PenC, the problem is now unconstrained and the gradient can be computed as

$$\nabla h(X) = G(X)(\frac{3}{2}I_p - \frac{1}{2}X^\top X) - X\mathrm{sym}(X^\top G(X)) + \beta X(X^\top X - I_p)$$

where $G(X) = \nabla f(Y)|_{Y=X(\frac{3}{2}I_p - \frac{1}{2}X^\top X)}$. This allows unconstrained solvers to be directly applied.

E. Additional developments and proofs for Section 3

E.1. Stiefel manifold

In the main text, we focus on the Euclidean metric on Stiefel manifold, i.e., $\langle U, V\rangle_X^E = \langle U, V\rangle$, which leads to the Riemannian gradient $\mathrm{grad} f(X) = \nabla f(X) - X\mathrm{sym}(X^\top \nabla f(X))$. We here review another popular metric on Stiefel manifold, namely the canonical metric (Edelman et al., 1998), defined as $\langle U, V\rangle_X^C = \langle U, (I_n - \frac{1}{2}XX^\top)V\rangle$. The corresponding Riemannian gradient is derived as $\mathrm{grad}_C f(X) = \nabla f(X) - X\nabla f(X)^\top X$.

There exists a variety of retractions for the Stiefel manifold, including the QR retraction $\mathrm{Retr}_X^{\mathrm{qr}}(tU) = \mathrm{qf}(X + tU)$ where $\mathrm{qf}(A)$ extracts the q-factor from the thin QR decomposition. The Cayley retraction (Wen & Yin, 2013; Zhu, 2017) is $\mathrm{Retr}_X^{\mathrm{cl}}(tU) = (I - \frac{t}{2}W)^{-1}(I + \frac{t}{2}W)X$, where $U = WX$ for some $W \in \mathrm{Skew}(n)$. The exponential retraction is given by $\mathrm{Retr}_X^{\mathrm{exp}}(tU) = \begin{bmatrix} X & U \end{bmatrix} \mathrm{expm}\Big(\begin{bmatrix} X^\top U & -U^\top U \\ I & X^\top U \end{bmatrix}\Big)\begin{bmatrix} \mathrm{expm}(-X^\top U) \\ 0 \end{bmatrix}$.

Here we verify the same coordinate derivative is recovered under the canonical metric.

Lemma E.1. *Let $B_\ell = H_{ij}X$ for $1 \le i < j \le n$, then we have $\theta = \langle \mathrm{grad}_E f(X), B_\ell\rangle_X^E = \langle \mathrm{grad}_C f(X), B_\ell\rangle_X^C = \langle \nabla f(X), B_\ell\rangle = [\nabla f(X)X^\top - X\nabla f(X)^\top]_{ij}$.*

Proof of Lemma E.1. Recall the $\mathrm{grad}_E f(X) = \nabla f(X) - X\mathrm{sym}(X^\top \nabla f(X))$. Then

$$\theta = \langle \mathrm{grad}_E f(X), B_\ell\rangle_X^E = \langle \nabla f(X) - X\mathrm{sym}(X^\top \nabla f(X)), H_{ij}X\rangle = \langle \nabla f(X), H_{ij}X\rangle = 2[\mathrm{skew}(\nabla f(X)X^\top)]_{ij},$$

where we use the fact that skew-symmetric matrix is orthogonal to symmetric matrix with respect to the Euclidean inner product. Similarly for the canonical metric,

$$\begin{aligned}\theta = \langle \mathrm{grad}_C f(X), B_\ell\rangle_X^C &= \langle \nabla f(X) - X\nabla f(X)X^\top, (I_n - \frac{1}{2}XX^\top)H_{ij}X\rangle \\ &= \langle \nabla f(X) - X\mathrm{sym}(X^\top \nabla f(X)), H_{ij}X\rangle \\ &= \langle \nabla f(X), H_{ij}X\rangle.\end{aligned}$$

This verifies θ is the same under the two metrics. □

E.2. Grassmann manifold

Proof of Proposition 3.2. It suffices to verify that $\nabla f(XQ) = \nabla f(X)Q$ as the Euclidean inner product $\langle \nabla f(X)Q, H_{ij}XQ\rangle = \langle \nabla f(X), H_{ij}X\rangle$. We start by noticing $f(X) = f(XQ)$ for any $Q \in \mathcal{O}(p)$. Then taking derivative on both sides gives $\nabla f(X) = \nabla f(XQ)Q^\top$ and thus $\nabla f(XQ) = \nabla f(X)Q$. □

E.3. Hyperbolic manifold

E.3.1. PROOFS

Proof of Proposition 3.3. By the decomposition, we can express $A = X\Omega + X_{J_\perp}K + XS$ for some $\Omega \in \mathrm{Skew}(p), S \in \mathrm{Sym}(p), K \in \mathbb{R}^{(n-p)\times p}$. Left-multiplying $X^\top J$ gives $X^\top JA = -\Omega - S$. Summing both sides with the transposes yields $S = -\mathrm{sym}(X^\top JA)$. Hence the projection to $T_X\mathcal{H}(d,r)$ is given by $\mathrm{Proj}_X(A) = A + X\mathrm{sym}(X^\top JA)$.

From the definition of Riemannian gradient, we have $\mathrm{D}f(X)[U] = \langle \nabla f(X), U\rangle_F = \langle J\nabla f(X), U\rangle_{\mathcal{L}} = \langle J\nabla f(X), \mathrm{Proj}_X(U)\rangle_{\mathcal{L}} = \langle \mathrm{Proj}_X(J\nabla f(X)), U\rangle_{\mathcal{L}}$, where we use the self-adjoint property of orthogonal projection with respect to the metric. Thus the Riemannian gradient is $\mathrm{grad}f(X) = \mathrm{Proj}_X(J\nabla f(X)) = J\nabla f(X) + X\mathrm{sym}(X^\top \nabla f(X))$. □

Proof of Proposition 3.4. First, we see $c(0) = X$ and we compute $c'(t) = \mathrm{expm}(tWJ)WJX$ and hence $c'(0) = WJX = U$. The only part left is to show $c(t) \in \mathcal{H}(n,p)$. This can be verified by showing $\mathrm{expm}(tWJ)$ is a Lorentz transform. To see this, let $L(t) \coloneqq \mathrm{expm}(tWJ)^\top J\mathrm{expm}(tWJ)$. Then we have

$$\begin{aligned} L'(t) &= (WJ\mathrm{expm}(tWJ))^\top J\mathrm{expm}(tWJ) + \mathrm{expm}(tWJ)^\top JWJ\mathrm{expm}(tWJ) \\ &= \mathrm{expm}(tWJ)^\top (JW^\top J + JWJ)\mathrm{expm}(tWJ) = 0, \quad \forall t. \end{aligned}$$

Thus $L(t) = L(0) = J, \forall t$. This completes the proof as $c(t) = \mathrm{expm}(tWJ)X \in \mathcal{H}(n,p)$ because $c(t)^\top Jc(t) = X^\top L(t)X = X^\top JX = -I_p$. □

Proof of Lemma 3.5. For the case where $i,j \neq 1$, we can see $H_{ij}J = H_{ij}$ and thus $\mathrm{expm}(\theta H_{ij})$ leads to the Givens rotation. However in the case where $i = 1$, $H_{ij}J = E_{ij} \coloneqq e_ie_j^\top + e_je_i^\top$. Thus, it remains to show $\mathrm{expm}(\theta E_{ij})$ can be simplified. To this end, we see $E_{ij}^{2t} = \frac{1}{2}(E_{ii} + E_{jj})$, $E_{ij}^{2t-1} = E_{ij}$, for $t \in \mathbb{N}$. Then we can show $\mathrm{expm}(\theta E_{ij}) = I + \theta E_{ij} + \frac{1}{2!}\theta^2 E_{ij}^2 + \frac{1}{3!}\theta^3 E_{ij}^3 + \cdots = \frac{1}{2}\sum_{k\neq i,j} E_{kk} + (1 + \frac{\theta^2}{2!} + \frac{\theta^4}{4!} + \cdots)\frac{1}{2}(E_{ii} + E_{jj}) + (\theta + \frac{\theta^3}{3!} + \frac{\theta^5}{5!} + \cdots)E_{ij} = \frac{1}{2}\sum_{k\neq i,j} E_{kk} + \frac{\cosh(\theta)}{2}(E_{ii} + E_{jj}) + \sinh(\theta)E_{ij}$. □

E.3.2. A CANONICAL-TYPE METRIC

In addition to the Euclidean metric, we define the canonical metric on hyperbolic manifold as for $U, V \in T_X\mathcal{H}(n,p)$,

$$\begin{aligned} \langle U, V\rangle_X = -\mathrm{tr}\big(U^\top(J + \frac{1}{2}JXX^\top J)V\big) &= -\mathrm{tr}\big((X\Omega_u + X_{J_\perp}K_u)^\top(J + \frac{1}{2}JXX^\top J)(X\Omega_v + X_{J_\perp}K_v)\big) \\ &= \frac{1}{2}\mathrm{tr}(\Omega_u^\top\Omega_v) + \mathrm{tr}(K_u^\top K_v). \end{aligned}$$

The normal space under the canonical metric is the same as that for the Euclidean metric $N_X\mathcal{H}(d,r) = \{XS : S \in \mathrm{Sym}(p)\}$ and thus the orthogonal projection can be derived also to be the same. The Riemannian gradient can be derived as follows.

Proposition E.2. *The orthogonal projection of $A \in \mathbb{R}^{n\times p}$ to $T_X\mathcal{H}(n,p)$ is given by* $\mathrm{Proj}_X(A) = A + X\mathrm{sym}(X^\top JA)$. *The Riemannian gradient with respect to the canonical metric is* $\mathrm{grad}f(X) = -J\nabla f(X) - X\nabla f(X)^\top X = 2\mathrm{skew}(J\nabla f(X)X^\top)JX$.

Proof of Proposition E.2. First we notice that $(J + \frac{1}{2}JXX^\top J)^{-1} = J - XX^\top$. This implies $\langle U, V\rangle = \mathrm{tr}(U^\top V) = \mathrm{tr}(U^\top(J - XX^\top)(J + \frac{1}{2}JXX^\top J)V) = -\langle (J - XX^\top)U, V\rangle_X$. Then by definition of Riemannian gradient, we have $\langle \nabla f(X), U\rangle = -\langle (J - XX^\top)\nabla f(X), U\rangle_X = \langle \mathrm{Proj}_X\big((XX^\top - J)\nabla f(X)\big), U\rangle_X$. Hence, $\mathrm{grad}f(X) = \mathrm{Proj}_X\big((XX^\top - J)\nabla f(X)\big) = (XX^\top - J)\nabla f(X) = -J\nabla f(X) - X\nabla f(X)^\top X$. □

E.3.3. CAYLEY RETRACTION

Motivated by the Cayley transform for (generalized) Stiefel manifold, we define the Cayley transform for generalized hyperbolic manifold as follows. Then in Proposition E.4, we show the Cayley transform naturally leads to a valid retraction on the generalized hyperbolic manifold.

Definition E.3. For $X \in \mathcal{H}(d,r)$ and $U = WJX \in T_X\mathcal{H}(d,r)$, the Cayley transform is defined as $\mathrm{Cay}_X(U) = (I - \frac{1}{2}WJ)^{-1}(I + \frac{1}{2}WJ)X$, which is well-defined if $I - \frac{1}{2}WJ$ is nonsingular.

Proposition E.4 (Cayley retraction)**.** *For $X \in \mathcal{H}(d,r)$ and $U \in T_X\mathcal{H}(d,r)$, The map $\mathrm{Retr}_X^C(tU) := \mathrm{Cay}_X(tU) = (I - \frac{t}{2}W_U J)^{-1}(I + \frac{t}{2}W_U J)X$ is a retraction, where $W_U = XU^\top P_X - P_X^\top U X^\top$, with $P_X = I_n + \frac{1}{2}JXX^\top$.*

Proof of Proposition E.4. First, because $U \in T_X\mathcal{H}(d,r)$, we can express $U = WJX$ for some $W \in \mathrm{Skew}(n)$. Since $\mathcal{H}(d,r)$ is an embedded submanifold of $\mathbb{R}^{d\times r}$, let the curve $c(t) = \mathrm{Retr}_X^C(tU)$ and we have $c(0) = X$ and

$$\begin{aligned} c'(t) &= \frac{1}{2}(I - \frac{t}{2}WJ)^{-1}WJ(I - \frac{t}{2}WJ)^{-1}(I + \frac{t}{2}WJ)X + \frac{1}{2}(I - \frac{t}{2}WJ)^{-1}WJX \\ &= \frac{1}{2}(I - \frac{t}{2}WJ)^{-1}WJ(X + c(t)) \end{aligned}$$

with $c'(0) = WJX = U$. It remains to show $c(t) \in \mathcal{H}(d,r)$. First we notice that

$$c(t) = (I - \frac{t}{2}WJ)^{-1}(I + \frac{t}{2}WJ)X = (J - \frac{t}{2}JWJ)^{-1}(J + \frac{t}{2}JWJ)X.$$

Then

$$\begin{aligned} c(t)^\top Jc(t) &= X^\top(J + \frac{t}{2}JW^\top J)(J - \frac{t}{2}JW^\top J)^{-1}J(J - \frac{t}{2}JWJ)^{-1}(J + \frac{t}{2}JWJ)X \\ &= X^\top(J - \frac{t}{2}JWJ)(J + \frac{t}{2}JWJ)^{-1}J(J - \frac{t}{2}JWJ)^{-1}(J + \frac{t}{2}JWJ)X \\ &= X^\top(I - \frac{t}{2}JW)(I + \frac{t}{2}JW)^{-1}(I - \frac{t}{2}JW)^{-1}(I + \frac{t}{2}JW)JX \\ &= X^\top JX = -I_r, \end{aligned}$$

where the second equality uses the fact that $W \in \mathrm{Skew}(d)$ and the last equality is due to $(I+A)(I-A) = (I-A)(I+A)$ for any A.

Finally we can verify that $W_U = XU^\top P_X - P_X^\top U X^\top$ where $P_X = I_n + \frac{1}{2}JXX^\top$ satisfies $W_U JX = U$. That is,

$$W_U JX = \frac{1}{2}XU^\top JX + (I_n + \frac{1}{2}XX^\top J)U = U + \frac{1}{2}X(U^\top JX + X^\top JU) = U$$

where we use $U \in T_X\mathcal{H}(n,p)$. □

E.4. Symplectic manifold

E.4.1. Review of canonical metric and retractions

The canonical metric of symplectic manifold is developed in (Gao et al., 2021b), as $\langle U, V\rangle_X = \frac{1}{\rho}\langle S_u, S_v\rangle + \langle K_u, K_v\rangle$ for some chosen $\rho > 0$, where $U = X\Omega_p S_u + \Omega_n X_\perp K_u$ and $V = X\Omega_p S_v + \Omega_n X_\perp K_v$. The Riemannian gradient associated with the metric is $\mathrm{grad}_\rho f(X) = \rho X\Omega_p \mathrm{sym}(\Omega_p^\top X^\top \nabla f(X)) + \Omega_n X_\perp X_\perp^\top \Omega_n^\top \nabla f(X)$. The quasi-geodesic retraction is derived by replacing the covariance derivative with the Euclidean derivative, given by $\mathrm{Retr}_X^{\mathrm{qgeo}}(tU) = [X, U]\mathrm{expm}\Big(t\begin{bmatrix} -\Omega_p W & \Omega_p U^\top \Omega_n U \\ I_{2p} & -\Omega_p W\end{bmatrix}\Big)\begin{bmatrix}\mathrm{expm}(t\Omega_n W) \\ 0\end{bmatrix}$ where $W = X^\top \Omega_n U$. The symplectic Cayley retraction is derived to be $\mathrm{Retr}_X^{\mathrm{cay}}(tU) = \Big(I_{2n} - \frac{t}{2}S_{X,U}\Omega_n\Big)^{-1}\Big(I_{2n} + \frac{t}{2}S_{X,U}\Omega_n\Big)X$ where $S_{X,U} = G_X U(X\Omega_p)^\top + X\Omega_{2p}(G_X U)^\top$, $G_X = I_{2n} - \frac{1}{2}X\omega_p X^\top \Omega_n^\top$. In (Gao et al., 2022), a SR decomposition based retraction is proposed. That is, let $P_{2p} := [e_1, e_3, ..., e_{2p-1}, e_2, ..., e_{2p}]$ where e_j is the j-th basis vector of $\mathbb{R}^{2p}$. Then denote a congruence matrix set as $T_{sk}(P_{2p}) := \{P_{2p}^\top \hat{R} P_{2p} : \hat{R} \in \mathbb{R}^{2p\times 2p}$ is upper triangular$\}$. Then the $\mathrm{Retr}_X^{\mathrm{sr}}(tU) = \mathrm{sf}(X + tU)$ where $A = SR$ is the SR decomposition of $A \in \mathbb{R}^{2n\times 2p}$, with $S \in \mathrm{Sp}(2n, 2p)$ and $R \in T_{2p}(P_{2p})$ and $\mathrm{sf}(A)$ extracts the S factor.

E.4.2. Proofs

The proof of Proposition 3.6 follows immediately from the following Lemma.

Lemma E.5. *For any $S \in \mathrm{Sym}(2n)$, we have $\mathrm{expm}(t\Omega_n S), \mathrm{expm}(tS\Omega_n) \in \mathrm{Sp}(n,n)$.*

Proof of Lemma E.5. The proof follows from (Gao et al., 2021b, Proposition 4.6). □

Proof of Proposition 3.6. Similar to the previous sections, let $c(t) := \mathrm{Retr}_X(tU)$ and we have $c(0) = X$ and $c'(0) = S\Omega_n X = U$. Finally from Lemma E.5, we have $c(t)^\top \Omega_n c(t) = X^\top \mathrm{expm}(tS\Omega_n)^\top \Omega_n \mathrm{expm}(tS\Omega_n) X = X^\top \Omega_n X = \Omega_p$, which verifies $c(t) \in \mathrm{Sp}(n,p)$ □

Proof of Proposition 3.7. We can partition the basis $E_{ij} = \begin{bmatrix} E_{ij,1} & E_{ij,2} \\ E_{ij,2}^\top & E_{ij,3} \end{bmatrix} \in \{0,1\}^{2n\times 2n}$, where $E_{ij,1}, E_{ij,3} \in \mathrm{Sym}(n)$ and $E_{ij,2} \in \{0,1\}^{n\times n}$. Hence $E_{ij}\Omega_n = \begin{bmatrix} -E_{ij,2} & E_{ij,1} \\ -E_{ij,3} & E_{ij,2}^\top \end{bmatrix}$ and the aim is to express $\mathrm{expm}(\theta E_{ij}\Omega_n)$ in compact form.

For $1 \le i \le j \le n$, we have $E_{ij,2}, E_{ij,3} = 0$ and thus we obtain $\exp(\theta E_{ij}\Omega_n) = I + \theta E_{ij}\Omega_n + \frac{\theta^2}{2}(E_{ij}\Omega_n)^2 + \cdots = I + \theta E_{ij}\Omega_n$. Similarly for $n+1 \le i < j \le 2n$, we have $E_{ij,1}, E_{ij,2} = 0$ and $\mathrm{expm}(\theta E_{ij}\Omega_n) = I + \theta E_{ij}\Omega_n$. This verifies $\forall 1 \le i \le j \le n$ or $n+1 \le i < j \le 2n$

$$\mathrm{Retr}_X(\theta E_{ij}\Omega_n X) = X + \theta E_{ij}\Omega_n X.$$

For $1 \le i \le n < j \le 2n$, we have $E_{ij,1} = E_{ij,3} = 0$ and $E_{ij,2} = e_i e_{j-n}^\top$. Then $E_{ij}\Omega_n = \begin{bmatrix} -E_{ij,2} & 0 \\ 0 & E_{ij,2}^\top \end{bmatrix}$. We first notice that for $k = 2,3,4,...$

$$(E_{ij,2})^k = (e_i e_{j-n}^\top)^k = \begin{cases} e_i e_i^\top, & \text{if } i = j-n \\ 0, & \text{otherwise} \end{cases}, \qquad (E_{ij,2}^\top)^k = (e_{j-n} e_i^\top)^k = \begin{cases} e_i e_i^\top, & \text{if } i = j-n \\ 0, & \text{otherwise} \end{cases}$$

This suggests for $k = 2,3,4,...$,

$$(E_{ij}\Omega_n)^k = \begin{cases} (-1)^k e_i e_i^\top + e_{n+i} e_{n+i}^\top, & \text{if } i = j-n \\ 0, & \text{otherwise} \end{cases}$$

$$\text{and} \quad \mathrm{expm}(\theta E_{ij}\Omega_n) = \begin{cases} I + (e^{-\theta}-1) e_i e_i^\top + (e^{\theta}-1) e_{n+i} e_{n+i}^\top & \text{if } i = j-n \\ I + \theta E_{ij}\Omega_n, & \text{otherwise} \end{cases}$$

This verifies for $1 \le i \le n < j \le 2n$,

$$\mathrm{Retr}_X(\theta E_{ij}\Omega_n X) = \begin{cases} X + (e^{-\theta}-1) e_i e_i^\top X + (e^{\theta}-1) e_{n+i} e_{n+i}^\top X, & \text{if } i = j-n \\ X + \theta E_{ij}\Omega_n X, & \text{otherwise} \end{cases}$$

The proof is now complete. □

E.5. Doubly stochastic manifold

For the doubly stochastic manifold, the retraction applies the Sinkhorn algorithm (Knight, 2008) for matrix balancing, i.e., $\mathrm{Retr}_X(tU) = \mathrm{SK}(X \odot \exp(tU \oslash X))$, where U is a tangent vector belonging to the tangent space $T_X\Pi(\mu,\nu)$ and the Sinkhorn algorithm $\mathrm{SK}(U)$ iteratively normalize rows and columns of U according to the given marginals (Shi et al., 2021; Cardoso & Leite, 2010).

E.5.1. Proofs

Proof of Proposition 3.9. In fact, we can show for any $H_{ijkl} := (e_i - e_j)(e_k - e_l)^\top$ for $i \ne j$, $k \ne l$. The coordinate Sinkhorn is a valid retraction along the direction H_{ijkl}. Let $c(t) := \mathrm{cSK}(X \odot \exp(tH_{ijkl} \oslash X))$ and we can immediately see $c(0) = X$. Also, Let $\widetilde{X} := X \odot \exp(tH_{ijkl} \oslash X)$. Then $\widetilde{X}$ differs with X in only the entries at $(i,k),(j,k),(i,l),(k,l)$, which forms the 2×2 sub-matrix that we wish to balance. Also, by definition, the marginals are given by $\tilde{\mu} := ([X]_{ik} + [X]_{il}, [X]_{jk} + [X]_{jl})$ and $\tilde{\nu} := ([X]_{ik} + [X]_{jk}, [X]_{il} + [X_{jl}])$. It readily holds that $\tilde{\mu}^\top 1_2 = \tilde{\nu}^\top 1_2$. For notational purposes, for any matrix $A \in \mathbb{R}^{m\times n}$, let $A_{ijkl} \in \mathbb{R}^{m\times n}$ be the matrix that zeros out the entries except for the 2×2 sub-matrix. Also, we denote $A^\flat_{ijkl} := \begin{bmatrix} [A]_{ik} & [A]_{il} \\ [A]_{jk} & [A]_{jl} \end{bmatrix}$ that extracts the corresponding 2×2 sub-matrix. Then $\mathrm{cSK}(\widetilde{X})$ reduces to performing Sinkhorn on $\widetilde{X}^\flat_{ijkl}$ with marginals $\tilde{\mu}, \tilde{\nu}$ with other entries of $\widetilde{X}$ unchanged. This is well-defined as the Sinkhorn algorithm

converges to the unique doubly stochastic matrix of the form $\mathrm{diag}(u)\widetilde{X}^{\flat}_{ijkl}\mathrm{diag}(v)$ for some positive vectors u, v (Sinkhorn, 1967). This verifies that $\mathrm{cSK}(\widetilde{X})$ results in a doubly stochastic matrix, which remains on the manifold. Lastly, it remains to show that $c'(0) = H_{ijkl}$. For this, we first have

$$c'(0) = \lim_{t\to 0} \frac{c(t) - c(0)}{t} = \frac{\mathrm{cSK}(X \odot \exp(tH_{ijkl} \oslash X)) - X}{t} = \frac{\mathrm{cSK}(X + tH_{ijkl}) - X}{t},$$

where we use the first-order approximation of the exponential operations. Notice that $\mathrm{cSK}(X + tH_{ijkl})$ only modifies the 2×2 sub-matrix of X by $\mathrm{SK}(X^{\flat}_{ijkl}+tH^{\flat}_{ijkl})$. From (Douik & Hassibi, 2019; Shi et al., 2021), we have $\mathrm{SK}(X^{\flat}_{ijkl}+tH^{\flat}_{ijkl}) \approx X^{\flat}_{ijkl} + tH^{\flat}_{ijkl}$. This suggests $\lim_{t\to 0}(\mathrm{SK}(X^{\flat}_{ijkl} + tH^{\flat}_{ijkl}) - X^{\flat}_{ijkl})/t = H^{\flat}_{ijkl}$, which verifies $c'(0) = H_{ijkl}$. □

Lemma E.6. *Given a positive 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, the Sinkhorn algorithm on A with marginals $p = [p_1, p_2], q = [q_1, q_2] \in \Delta_2$ converges to $\begin{bmatrix} c_{11}a & c_{12}b \\ c_{21}c & c_{22}d \end{bmatrix}$ where $c_{12} = p_1/(\kappa a + b), c_{22} = p_2/(\kappa c + d)$, $c_{11} = \kappa c_{12}, c_{21} = \kappa c_{22}$ where κ is the positive root of the equation $q_2 ac\kappa^2 + \big((bc + ad)q_2 - bcp_1 - adp_2\big)\kappa - bdq_1 = 0$.*

Proof of Lemma E.6. Sinkhorn algorithm converges to the unique doubly stochastic matrix of the form $\mathrm{diag}(u)A\mathrm{diag}(v)$ for $u = [u_1, u_2], v = [v_1, v_2]$. From the constraints, $\mathrm{diag}(u)A\mathrm{diag}(v)1_2 = p$ and $\mathrm{diag}(v)A^{\top}\mathrm{diag}(u)1_2 = q$ we need to solve the quadratic problem

$$\begin{cases} (u_1v_1)a + (u_1v_2)b = p_1 \\ (u_2v_1)c + (u_2v_2)d = p_2 \\ (u_1v_1)a + (u_2v_1)c = q_1 \\ (u_1v_2)b + (u_2v_2)d = q_2. \end{cases} \tag{6}$$

Let $c_{11} = u_1v_1, c_{12} = u_1v_2, c_{21} = u_2v_1, c_{22} = u_2v_2$ which transforms (6) into a set of linear equations for the variables $c_{11}, c_{12}, c_{21}, c_{22}$. The equation system, however, is under-determined and has many solutions. The unique solution that is sought should satisfy $c_{11}/c_{12} = c_{21}/c_{22} = \kappa$. To this end, from the first two equations, we obtain

$$c_{12} = p_1/(\kappa a + b), \quad c_{22} = p_2/(\kappa c + d).$$

Substituting the expressions to the last equation yields $\frac{bp_1}{\kappa a+b} + \frac{dp_2}{\kappa c+d} = q_2$, which we solve for κ as the positive root of $q_2 ac\kappa^2 + \big((bc + ad)q_2 - bcp_1 - adp_2\big)\kappa - bdq_1 = 0$. □

F. Formal developments and proofs for Section 4

F.1. Developments

Assumption F.1. Consider a neighbourhood $\mathcal{X} \subseteq \mathcal{M}$ that contains a critical point.

F.1.1 The basis and its projection are bounded. Let the projection onto the basis $B_{\ell,X}$ be $\mathcal{P}_{B_{\ell,X}}(U) := \langle U, B_{\ell,X}\rangle_X B_{\ell,X}$. There exists constant $c_b, c_p > 0$ such that $\forall X \in \mathcal{X}, U \in T_X\mathcal{M}, \ell \in \mathcal{I}$, $\|B_{\ell,X}\|_X^2 \le c_b$ and $\sum_{\ell\in\mathcal{I}} \|\mathcal{P}_{B_{\ell,X}}U\|_X^2 \ge c_p\|U\|_X^2$.

F.1.2 The objective f is retraction L-smooth in $\mathcal{X}$, i.e., $f(\mathrm{Retr}_X(U)) - f(U) - \langle \mathrm{grad} f(X), U\rangle_X \le \frac{L}{2}\|U\|_X^2, \forall X \in \mathcal{X}$ and $U \in T_X\mathcal{M}$ such that $\mathrm{Retr}_X(U) \in \mathcal{X}$.

Remark F.2. Assumption F.1.1 requires the basis has a bounded norm and the projection of any tangent vector onto the basis does not vanish. Such an assumption is manifold-specific and we can verify that $\|B_{\ell,X}\|_X^2$ has an upper bound (e.g., for Stiefel and Grassmann, $\|B_{\ell,X}\|_X^2 \le \|H_{i,j}\|_F^2 = 2$). Then we note that the second requirement trivially holds for orthonormal basis due to the decomposition of U and Jensen's inequality. For non-orthonormal basis, this assumption also holds as long as projection of a tangent vector does not vanish. Assumption F.1.2 can also be satisfied by the compactness of the domain, e.g., we can take L in Assumption F.1.2 to be $L = \max_{X\in\mathcal{X}, U\in T_X\mathcal{M}:\|U\|_X=1} \frac{d^2 f(\mathrm{Retr}_X(tU))}{dt^2}$. These are all reasonable assumptions within a compact neighbourhood $\mathcal{X}$.

Theorem 4.1 (Formal). *Under Assumption F.1, consider RCD algorithm with $S = 1$ and ℓ_k^s selected uniformly at random from $\mathcal{I}$. Then suppose $\eta = \frac{1}{Lc_b}$, it satisfies $\min_{0\le k\le K-1} \mathbb{E}\|\mathrm{grad} f(X_k)\|_{X_k}^2 \le \frac{2Lc_b^2 c_p^{-1}|\mathcal{I}|\Delta_0}{K}$, where $\Delta_0 \coloneqq f(X_0) - f^*$.*

To analyze the cyclic variant, we further require the assumption that bounds the difference between Riemannian distance and distance induced by general retraction. In addition, we require the gradient Lipschitzness.

Assumption F.3. Under the same settings as in Assumption F.1,

F.3.1 For all $X, Y = \mathrm{Retr}_X(U) \in \mathcal{X}$, there exists constants $\vartheta_0, \vartheta_1 > 0$ such that $\vartheta_0\|U\|_X^2 \le \mathrm{dist}^2(X, Y) \le \vartheta_1\|U\|_X^2$.

F.3.2 The objective has retraction L_g-Lipschitz gradient, i.e., $\|\mathrm{grad} f(X) - \mathcal{T}_Y^X \mathrm{grad} f(Y)\|_X^2 \le L_g\|U\|_X^2$, $\forall X, Y = \mathrm{Retr}_X(U) \in \mathcal{X}$ and $\mathcal{T}_X^Y$ is the an isometric vector transport that satisfies $\langle \mathcal{T}_X^Y U, \mathcal{T}_X^Y V\rangle_Y = \langle U, V\rangle_X$, $\forall X, Y \in \mathcal{X}, U, V \in T_X\mathcal{M}$.

F.3.3 For any fixed coordinate index $\ell \in \mathcal{I}$, there exists a constant $\upsilon > 0$ such that for all $X, Y \in \mathcal{X}$, $V \in T_Y\mathcal{M}$, $\|\mathcal{P}_{B_{\ell,X}} \mathcal{T}_Y^X V\|_X^2 \ge \upsilon\|\mathcal{P}_{B_{\ell,Y}} V\|_Y^2$.

Remark F.4. Assumption F.3.1 bounds the difference Riemannian distance (relating to inverse exponential map) and the inverse retraction. Because retraction is a first-order approximation to the exponential map, this assumption naturally holds when the domain is sufficiently small (see (Huang et al., 2015; Sato et al., 2019)). Assumption F.3.2 is further required because when general retraction is used, gradient Lipschitzness is not equivalent to function smoothness. Assumption F.3.3 further claims that the difference between the same coordinate basis on different tangent spaces is bounded. We note that the RHS is identical to $\|\mathcal{P}_{\mathcal{T}_Y^X B_{\ell,Y}} \mathcal{T}_Y^X V\|$ due to the isometric vector transport. Then it reduces to whether $\mathcal{T}_Y^X B_{\ell,Y}$ and $B_{\ell,X}$ are related, which is expected because due to the compactness of the domain, X is bounded from Y. This allows to establish the convergence for cyclic selection of basis.

Theorem 4.2 (Formal). *Under Assumption F.1 and consider RCD algorithm with $S = |\mathcal{I}|$ and $\ell_k^s = s+1$ for $s = 0, ..., |\mathcal{I}|-1$. Then selecting $\eta = \frac{1}{Lc_b}$ gives $\min_{0\le k\le K-1} \|\mathrm{grad} f(X_k)\|_{X_k} \le \frac{C\Delta_0}{K}$, where $C = 4Lc_b^2 c_p^{-1}\upsilon^{-1}(1+|\mathcal{I}|^2 c_b^{-1} L^{-2} L_g \vartheta_1 \vartheta_0^{-1})$.*

Theorem 4.3 (Formal). *Under Assumption F.1, F.3 and further let $\omega_0, \omega_1 > 0$, such that for any fixed epoch k, $\omega_0\langle \nabla f(X_k^s), B_{\ell_k^s}\rangle^2 \le \theta_k^s\langle \nabla f(X_k^s), B_{\ell_k^s}\rangle \le \omega_1\langle \nabla f(X_k^s), B_{\ell_k^s}\rangle^2$, $\forall s \le S_{\max} - 1$.*

(Randomized). Consider RCDlin algorithm with $1 < S \le S_{\max}$ and ℓ_k^s selected uniformly at random from $\mathcal{I}$. Then choosing $\eta = \frac{\omega_0}{Lc_b\omega_1^2}$, we obtain $\min_{0\le k\le K-1, 0\le s\le S-1} \mathbb{E}\|\mathrm{grad} f(X_k^s)\|_{X_k^s}^2 \le \frac{2Lc_b^2 c_p^{-1} w_0^{-2} w_1^2 |\mathcal{I}|\Delta_0}{KS}$.

(Cyclic). Suppose $S_{\max} \ge |\mathcal{I}|$ and consider RCDlin algorithm with $S = |\mathcal{I}|$ and $\ell_k^s = s + 1$ for $s = 0, ..., |\mathcal{I}| - 1$. Then choosing $\eta = \frac{\omega_0}{Lc_b\omega_1^2}$, we have $\min_{0\le k\le K-1} \|\mathrm{grad} f(X_k^s)\|_{X_k^s}^2 \le \frac{\widetilde{C}\Delta_0}{K}$, where $\widetilde{C} = 4Lc_b^2\omega_1^2\omega_0^{-2}c_p^{-1}\nu^{-1}(1 + |\mathcal{I}|^2 c_b^{-1} L^{-2} L_g \vartheta_1 \vartheta_0^{-1})$.

Remark F.5. We finally remark that the proof ideas of cyclic and randomized RCD follow from classic developments of coordinate descent (Wright, 2015) in the Euclidean space by showing sufficient descent in the objective function. On general manifolds, in order to generalize the proof ideas, we further require the assumptions outlined in Assumption F.1, F.3. In particular, for cyclic selection rule, we require Assumption F.3.3 to relate bases from different tangent spaces. Similar assumptions have been considered in (Gutman & Ho-Nguyen, 2023) for showing convergence of deterministic subspace descent algorithms on manifolds (see (C, r)-norm condition).

F.2. Proofs

Proof of Theorem 4.1. Because $S = 1$ and by retraction L-smoothness, we have

$$
\begin{aligned}
f(X_{k+1}) - f(X_k) &\le -\eta\langle \mathrm{grad} f(X_k), \theta_k B_{\ell_k}\rangle_{X_k} + \frac{\eta^2\theta_k^2 L}{2}\|B_{\ell_k}\|_{X_k}^2 \\
&\le -\eta\theta_k^2 \frac{\|B_{\ell_k}\|_{\ell_k}^2}{c_b} + \frac{\eta^2\theta_k^2 L}{2}\|B_{\ell_k}\|_{X_k}^2 \\
&= -\frac{\|\mathcal{P}_{B_{\ell_k}} \mathrm{grad} f(X_k)\|_{X_k}^2}{2Lc_b^2},
\end{aligned}
$$

where we use the assumption $\|B_{\ell_k}\|_{X_k}^2 \leq c_b$, and choose $\eta = \frac{1}{Lc_b}$. Taking expectation with respect to ℓ_k, we have

$$\begin{aligned}\mathbb{E}_{\ell_k}[f(X_{k+1})] - f(X_k) &\leq -\frac{1}{2Lc_b^2}\mathbb{E}_{\ell_k}[\|\mathcal{P}_{B_{\ell_k}}\mathrm{grad}f(X_k)\|_{X_k}^2] \\ &= -\frac{1}{2|\mathcal{I}|Lc_b^2}\sum_{\ell\in\mathcal{I}}\|\mathcal{P}_{B_{\ell,X_k}}\mathrm{grad}f(X_k)\|_{X_k}^2 \\ &\leq -\frac{c_p}{2|\mathcal{I}|Lc_b^2}\|\mathrm{grad}f(X_k)\|_{X_k}^2.\end{aligned}$$

Telescoping this inequality and taking full expectation yields

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\mathrm{grad}f(X_k)\|_{X_k}^2 \leq \frac{2Lc_b^2c_p^{-1}|\mathcal{I}|\Delta_0}{K},$$

where we let $\Delta_0 \coloneqq f(X_0) - f^*$. □

Lemma F.6. *Under Assumption F.1.1, we have* $\|\mathcal{P}_{B_{\ell,X}}U\|_X \leq c_b\|U\|_X$, $\forall X \in \mathcal{X}, \ell \in \mathcal{I}$.

Proof of Lemma F.6. $\|\mathcal{P}_{B_{\ell,X}}U\|_X^2 \leq c_b\langle U, B_\ell(X)\rangle_X^2 = c_b\big\langle\langle U, B_\ell(X)\rangle_X B_\ell(X), U\big\rangle_X \leq c_b\|\mathcal{P}_{B_{\ell,X}}U\|_X\|U\|_X$. Cancelling $\|\mathcal{P}_{B_{\ell,X}}U\|_X$ on both sides completes the proof. □

Proof of Theorem 4.2. We first focus on a single epoch k and for notation simplicity, we let $\mathcal{T}_{X_k^s}^{X_k} = \mathcal{T}_{s\to 0}$ and $\mathcal{T}_{X_k}^{X_k^s} = \mathcal{T}_{0\to s}$. Similarly from retraction L-smoothness,

$$f(X_k^{s+1}) - f(X_k^s) \leq -\frac{1}{2Lc_b^2}\|\mathcal{P}_{B_{\ell_k^s}}\mathrm{grad}f(X_k^s)\|_{X_k^s}^2$$

with stepsize $\eta = \frac{1}{Lc_b}$. Summing over $s = 0, ..., S-1$, we can bound

$$f(X_{k+1}) - f(X_k) \leq -\frac{1}{2Lc_b^2}\sum_{s=0}^{S-1}\|\mathcal{P}_{B_{\ell_k^s}}\mathrm{grad}f(X_k^s)\|_{X_k^s}^2.$$

Then it remains to bound the RHS. From L_g-Lipschitz,

$$\begin{aligned}\|\mathrm{grad}f(X_k^s) - \mathcal{T}_{0\to s}\mathrm{grad}f(X_k)\|_{X_k^s}^2 \leq L_g\|\mathrm{Retr}_{X_k^s}^{-1}(X_k)\|_{X_k^s}^2 &\leq L_g\vartheta_0^{-1}\mathrm{dist}^2(X_k, X_k^s) \\ &\leq L_g\vartheta_0^{-1}\big(\sum_{j=1}^{s}\mathrm{dist}(X_k^{j-1}, X_k^j)\big)^2 \\ &\leq \eta^2L_g\vartheta_1\vartheta_0^{-1}s\sum_{j=0}^{s-1}\|\mathcal{P}_{B_{\ell_k^j}}(\mathrm{grad}f(X_k^j))\|_{X_k^j}^2.\end{aligned}$$

where we use Assumption F.3 and triangle inequality of Riemannian distance.

Now we can show

$$\begin{aligned}\|\mathcal{P}_{B_{\ell_k^s}}\mathcal{T}_{0\to s}\mathrm{grad}f(X_k)\|_{X_k^s}^2 &\leq 2\|\mathcal{P}_{B_{\ell_k^s}}\mathcal{T}_{0\to s}\mathrm{grad}f(X_k) - \mathcal{P}_{B_{\ell_k^s}}\mathrm{grad}f(X_k^s)\|_{X_k^s}^2 + 2\|\mathcal{P}_{B_{\ell_k^s}}\mathrm{grad}f(X_k^s)\|_{X_k^s}^2 \\ &\leq 2c_b\|\mathcal{T}_{0\to s}\mathrm{grad}f(X_k) - \mathrm{grad}f(X_k^s)\|_{X_k^s}^2 + 2\|\mathcal{P}_{B_{\ell_k^s}}\mathrm{grad}f(X_k^s)\|_{X_k^s}^2 \\ &\leq 2c_b\eta^2L_g\vartheta_1\vartheta_0^{-1}S\sum_{j=0}^{s-1}\|\mathcal{P}_{B_{\ell_k^j}}(\mathrm{grad}f(X_k^j))\|_{X_k^j}^2 + 2\|\mathcal{P}_{B_{\ell_k^s}}\mathrm{grad}f(X_k^s)\|_{X_k^s}^2\end{aligned}$$

where the second inequality is due to Lemma F.6. Summing this inequality from $s = 0, ..., S-1$ gives

$$\begin{aligned}
&\sum_{s=0}^{S-1} \|\mathcal{P}_{B_{\ell_k^s}} \mathcal{T}_{0\to s} \mathrm{grad} f(X_k)\|_{X_k^s}^2 \\
&\leq 2\sum_{s=0}^{S-1} \|\mathcal{P}_{B_{\ell_k^s}} \mathrm{grad} f(X_k^s)\|_{X_k^s}^2 + 2c_b\eta^2 L_g \vartheta_1 \vartheta_0^{-1} |\mathcal{I}| \sum_{s=0}^{S-1}\sum_{j=0}^{s-1} \|\mathcal{P}_{B_{\ell_k^j}} (\mathrm{grad} f(X_k^j))\|_{X_k^j} \\
&\leq 2(1 + |\mathcal{I}|^2 c_b \eta^2 L_g \vartheta_1 \vartheta_0^{-1}) \sum_{s=0}^{S-1} \|\mathcal{P}_{B_{\ell_k^s}} \mathrm{grad} f(X_k^s)\|_{X_k^s}^2
\end{aligned}$$

where we notice $S = |\mathcal{I}|$. Also due to the cyclic selection of ℓ_k^s, we can see the LHS is

$$\sum_{s=0}^{S-1} \|\mathcal{P}_{B_{\ell_k^s}} \mathcal{T}_{0\to s} \mathrm{grad} f(X_k)\|_{X_k^s}^2 \geq \upsilon \sum_{\ell\in\mathcal{I}} \|\mathcal{P}_{B_{\ell,X_k}} \mathrm{grad} f(X_k)\|_{X_k}^2 \geq c_p \upsilon \|\mathrm{grad} f(X_k)\|_{X_k}^2,$$

where we use Assumption F.1.1 and F.3.3. Combining with previous results, we finally obtain

$$f(X_{k+1}) - f(X_k) \leq -\frac{1}{2Lc_b^2} \sum_{s=0}^{S-1} \|\mathcal{P}_{B_{\ell_k^s}} \mathrm{grad} f(X_k^s)\|_{X_k^s}^2 \leq -\frac{1}{C} \|\mathrm{grad} f(X_k)\|_{X_k}^s,$$

where $C \coloneqq 4Lc_b^2 c_p^{-1} \upsilon^{-1} (1 + |\mathcal{I}|^2 c_b^{-1} L^{-2} L_g \vartheta_1 \vartheta_0^{-1})$. Telescoping this inequality completes the proof. □

Proof of Theorem 4.3. For the randomized setting, by retraction L-smoothness,

$$\begin{aligned}
f(X_k^{s+1}) - f(X_k^s) &\leq -\eta \langle \mathrm{grad} f(X_k^s), \theta_k^s B_{\ell_k^s} \rangle_{X_k^s} + \frac{\eta^2 {\theta_k^s}^2 L}{2} \|B_{\ell_k^s}\|_{X_k^s}^2 \\
&= -\eta \theta_k^s \langle \nabla f(X_k^s), B_{\ell_k^s} \rangle + \frac{\eta^2 {\theta_k^s}^2 L}{2} \|B_{\ell_k^s}\|_{X_k^s}^2 \\
&\leq -\eta \omega_0 \langle \nabla f(X_k), B_{\ell_k^s} \rangle^2 + \frac{\eta^2 \omega_1^2 L}{2} \|\langle \nabla f(X_k^s), B_{\ell_k^s} \rangle B_{\ell_k^s}\|_{X_k^s}^2 \\
&\leq \Big(-\frac{\eta\omega_0}{c_b} + \frac{\eta^2 \omega_1^2 L}{2} \Big) \|\langle \nabla f(X_k^s), B_{\ell_k^s} \rangle B_{\ell_k^s}\|_{X_k^s}^2 \\
&= -\frac{\omega_0^2}{2Lc_b^2 \omega_1^2} \|\mathcal{P}_{B_{\ell_k^s}} \mathrm{grad} f(X_k^s)\|_{X_k^s}^2
\end{aligned}$$

where we choose $\eta = \frac{\omega_0}{Lc_b\omega_1^2}$. The second inequality follows from the assumption $\omega_0 \langle \nabla f(X_k^s), B_{\ell_k^s} \rangle^2 \leq \theta_k^s \langle \nabla f(X_k^s), B_{\ell_k^s} \rangle \leq \omega_1 \langle \nabla f(X_k^s), B_{\ell_k^s} \rangle^2$ and the third inequality is due to Assumption F.1.1. Following the similar proof strategy, we obtain the desired result. For the cyclic setting, the bound also readily follows by using the above result. □