Journal of Machine Learning Research 27 (2026) 1-77 Submitted 1/24; Published 2/26

Convergence and complexity of block majorization-minimization for constrained block-Riemannian optimization

Yuchen Li yli966@wisc.edu
Department of Mathematics
University of Wisconsin-Madison
480 Lincoln Dr., Madison, WI 53706, USA

Laura Balzano girasole@umich.edu
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109, USA

Deanna Needell deanna@math.ucla.edu
Department of Mathematics
University of California
Los Angeles, CA 90025, USA

Hanbaek Lyu hlyu@math.wisc.edu
Department of Mathematics
University of Wisconsin-Madison
480 Lincoln Dr., Madison, WI 53706, USA

Editor: Prateek Jain

Abstract

Block majorization-minimization (BMM) is a simple iterative algorithm for nonconvex optimization that sequentially minimizes a majorizing surrogate of the objective function in each block coordinate while the other block coordinates are held fixed. We consider a family of BMM algorithms for minimizing nonsmooth nonconvex objectives, where each parameter block is constrained within a subset of a Riemannian manifold. We establish that this algorithm converges asymptotically to the set of stationary points, and attains an ϵ-stationary point within $\widetilde{O}(\epsilon^{-2})$ iterations. In particular, the assumptions for our complexity results are completely Euclidean when the underlying manifold is a product of Euclidean or Stiefel manifolds, although our analysis makes explicit use of the Riemannian geometry. Our general analysis applies to a wide range of algorithms with Riemannian constraints: Riemannian MM, block projected gradient descent, Bures-JKO scheme for Wasserstein variational inference, optimistic likelihood estimation, geodesically constrained subspace tracking, robust PCA, and Riemannian CP-dictionary-learning. We experimentally validate that our algorithm converges faster than standard Euclidean algorithms applied to the Riemannian setting.

Keywords: Block Majorization-minimization, Riemannian Optimization, Constrained Optimization, Nonconvex Optimization, Nonsmooth Optimization, Iteration Complexity

Contents

1. Introduction

Optimization over Riemannian manifolds has a wide array of applications ranging from the computation of linear algebraic quantities and factorizations and problems with nonlinear differentiable constraints to the analysis of shape space and automated learning (Ring and Wirth, 2012; Jaquier et al., 2020). These applications arise either because of implicit constraints on the problem to be optimized or because the domain is naturally defined as a manifold. Although not nearly as abundant as in the Euclidean optimization setting, there are various methods employed for such optimization (Baker et al., 2008; Yang, 2007; Boumal et al., 2019; Edelman et al., 1998). Most of these approaches follow standard optimization techniques by iteratively computing a descent direction and taking a step in that direction along a geodesic. This computation is often challenging in practice so other approaches alleviate this burden by utilizing approximations or imposing additional assumptions (Absil et al., 2009; Boumal, 2023).

Majorization-minimization (MM) is a classical principle for minimizing objectives (Lange et al., 2000). The MM method, as its name suggests, consists of two main steps. The majorization step identifies an upper bounding surrogate for the objective function and the minimization step then minimizes that surrogate. The approach leads to iterations that decrease the objective function, with the benefit that each minimization of a well-chosen surrogate is computationally more efficient than the original problem. Such an approach has also been applied in the context of unconstrained manifold optimization with applications to robust sparse PCA on the Stiefel manifold (Breloy et al., 2021). Here unconstrained manifold optimization means the constraint set is only the manifold itself. When the underlying manifold is a Hadamard manifold (i.e., complete, simply connected manifold with nonpositive sectional curvature everywhere), then one can consider majorizing the objective function by adding the Riemannian distance squared function. The resulting MM algorithm corresponds to the Riemannian proximal point method (Li et al., 2009; Bento et al., 2017).

In this paper, we consider the minimization of a continuous function $F : \boldsymbol{\Theta} = \Theta^{(1)} \times \cdots \times \Theta^{(m)} \to \mathbb{R}$ which is the sum of a smooth function f (possibly non-convex) and a convex part (possibly non-smooth) $p(\boldsymbol{\theta})$. The concepts of smoothness and convexity are both with respect to the geometry of the manifold, which will be explained later in the paper. We assume the convex part p is separable, namely $p(\boldsymbol{\theta}) = \sum_{i=1}^{m} p^{(i)}(\theta^{(i)})$ where $\boldsymbol{\theta} = [\theta^{(1)}, \cdots, \theta^{(m)}]$. Each constraint set $\Theta^{(i)}$ is a closed subset of a Riemannian manifold $\mathcal{M}^{(i)}$. More precisely, we seek to solve

$$\min_{\substack{\boldsymbol{\theta}=[\theta^{(1)},\dots,\theta^{(m)}] \\ \theta^{(i)}\in\Theta^{(i)}\subseteq\mathcal{M}^{(i)} \text{ for } i=1,\dots,m}} \left(F(\boldsymbol{\theta}) := f(\boldsymbol{\theta}) + p(\boldsymbol{\theta})\right). \tag{1}$$

As the problem (1) is typically nonconvex, it is not always reasonable to expect that an algorithm would converge to a globally optimal solution starting from an arbitrary initialization. Instead, we aim to provide global convergence (from arbitrary initialization) to stationary points. In some problem classes, stationary points could be as good as global optimizers practically as well as theoretically (Mairal et al., 2010; Sun et al., 2015). Furthermore, obtaining the iteration complexity of such algorithms is of importance both for theoretical and practical purposes. In particular, one aims to bound the worst-case number of iterations to achieve an ε-approximate stationary point (defined appropriately, see Sec. 3.1).

In order to obtain a first-order optimal solution to (1), we consider various Riemannian generalizations of the *Block Majorization-Minimization* (BMM) algorithm in Euclidean space (Razaviyayn et al., 2013; Lyu and Li, 2025). The high-level idea of BMM is that, in order to minimize a multi-block objective, one can minimize a majorizing surrogate of the objective in each block in a cyclic order. The algorithm we study in the present work, which we call *Riemannian Block Majorization-Minimization* (RBMM), can be stated in a high-level as follows, where n denotes the iteration (see Algorithm 1 for the full statement):

$$\textbf{RBMM:}\qquad \begin{cases} \text{For } n=1,\cdots,N: \\ \quad \text{For } i=1,2,\dots,m: \\ \quad g_n^{(i)} \leftarrow \begin{bmatrix} \text{Majorizing surrogate of} \\ \theta \mapsto f_n^{(i)}(\theta) := f\left(\theta_n^{(1)},\cdots,\theta_n^{(i-1)},\theta,\theta_{n-1}^{(i+1)},\cdots,\theta_{n-1}^{(m)}\right); \end{bmatrix} \\ \quad \theta_n^{(i)} \in \arg\min_{\theta\in\Theta^{(i)}} \left(G_n^{(i)}(\theta) := g_n^{(i)}(\theta) + p(\theta, \boldsymbol{\theta}_{n:i})\right). \end{cases} \tag{2}$$

In this work, we carefully analyze RBMM (2) in various settings and obtain first-order optimality guarantees and iteration complexity. Moreover, the connection between our RBMM and other existing algorithms is studied, providing complexity results to some of them for the first time in the literature.

As a preview, we provide a special case of Corollary 11 of our main results concerning iteration complexity of RBMM on Stiefel manifolds (manifolds of orthonormal frames, see Ex. 1). Note the term L-smooth in the following corollary indicates the gradient of the function is L-Lipschitz continuous (see Sec. 3.2 for comparison with geodesic smoothness). The $\widetilde{O}(\cdot)$ notation is the variant of "big-O" notation that ignores the logarithmic factors.

Corollary 1 (Complexity of RBMM on Stiefel manifolds) *Suppose we are minimizing an L-smooth function f on the product of the Stiefel manifolds using RBMM (2) with L'-smooth surrogates with quadratic majorization gap*

$$g_n^{(i)}(\theta) - f_n^{(i)}(\theta) \geq c\|\theta - \theta_{n-1}^{(i)}\|^2$$

for all $n \geq 1$ and $i = 1, \dots, m$ for some $c > 0$. Then the iterates asymptotically converge to the set of stationary points and the algorithm has iteration complexity of $\widetilde{O}((1+c+c^{-1})\varepsilon^{-2})$.

For instance, Euclidean block-proximal updates or Euclidean prox-linear updates on the product of Stiefel manifolds have iteration complexity $\widetilde{O}(\varepsilon^{-2})$. To the best of our knowledge, these types of iteration complexity results for block Riemannian optimization methods are new to the literature. Note that the conditions we need to check for applying Corollary 1 (and its generalization Cor. 11) are completely Euclidean. However, our proof of these corollaries incorporates Riemannian geometry in a substantial manner.

1.1 Our contribution

In this work, we thoroughly analyze RBMM (2) and obtain asymptotic convergence to stationary points and iteration complexity. The novelty of this work, compared to the aforementioned related work, lies especially in the following aspects:

(1) (Non-smooth non-convex analysis) We obtain an iteration complexity of $\widetilde{O}(\varepsilon^{-2})$ of RBMM for solving nonsmooth nonconvex Riemannian optimization problems. See Theorems 7 and 10. Central to this analysis is a novel continuous first-order optimality measure, which we introduce in Section 3.1.

(2) (Constrained optimization) RBMM is applicable to constrained optimization problems on manifolds. Here, constrained optimization on manifolds means we allow the constraint set $\boldsymbol{\Theta}$ of the optimization problem to be a closed subset of the manifold, i.e., $\boldsymbol{\Theta} \subseteq \mathcal{M}$, which is not necessarily the entire manifold.

(3) (Surrogates flexibility) Our RBMM framework allows three different types of surrogates: g-smooth surrogates, Riemannian proximal surrogates, and Euclidean smooth surrogates. This provides flexibility for various Riemannian optimization problems. See examples in Sections 4 and 5.

(4) (Robustness) RBMM is robust in the face of *inexact* computation of minimizing block surrogates on the manifolds. This allows, for example, one to employ standard iterative g-convex minimization algorithms over manifolds in the absence of exact subproblem solutions. See (A0)**(ii)**.

RBMM, as a general Riemannian block optimization framework, entails many classical algorithms including Euclidean block MM, proximal updates on Hadamard manifolds, and MM methods on Stiefel manifolds. We apply our results to various stylized applications such as Bures-JKO scheme for Wasserstein variational inference, geodesically constrained subspace tracking, optimistic likelihood under Fisher-Rao distance, Riemannian CP-dictionary-learning, and robust PCA and obtain the following results:

(5) Asymptotic convergence and complexity of $\widetilde{O}(\varepsilon^{-2})$ of Bures-JKO scheme for Wasserstein variational inference is established for non-strongly convex potential functions. See Corollary 18.

(6) Asymptotic convergence and complexity of $\widetilde{O}(\varepsilon^{-2})$ for Euclidean block proximal methods when the manifolds are embedded in Euclidean spaces and the constraint sets on the manifolds are compact and g-convex. See Theorem 7.

(7) Asymptotic convergence and complexity of $\widetilde{O}(\varepsilon^{-2})$ for block MM with Euclidean L-smooth surrogate on the product of Stiefel and Euclidean manifolds with g-convex constraints. See Corollary 11.

(8) A Euclidean-regularized version of MM with linear surrogates on the Stiefel manifold is proposed, with guarantees on asymptotic convergence and complexity of $\widetilde{O}(\varepsilon^{-2})$, which can be applied to geodesically constrained subspace tracking problems (Blocker et al., 2023). See Corollary 17 and Corollary 20. It is worth noting that Grassmannian optimization is in general NP-hard.

1.2 Background and Related Work

At a high level, most Riemannian optimization algorithms iteratively solve a sequence of carefully constructed, yet simpler, sub-problems. These sub-problems are generally categorized into two types: "manifold-type," which involves another Riemannian optimization problem on the manifold itself, or "tangential-type," which involves a Euclidean optimization problem on the tangent spaces.

Riemannian gradient descent and trust region. Many popular Riemannian optimization methods, such as Riemannian (projected) gradient descent and Riemannian trust region (Baker et al., 2008; Yang, 2007; Boumal et al., 2019; Edelman et al., 1998; Absil et al., 2009; Boumal, 2023), are of the tangential type. Roughly speaking, these methods iterate a two-step process: (1) moving along a descent direction in the tangent space at the current iterate, and (2) projecting this update back onto the manifold using a retraction.

Zhang and Sra (2016) develops various Riemannian projected (sub)gradient methods for minimizing g-convex objectives on a compact subset $\mathcal{X}$ of a Hadamard manifold $\mathcal{M}$. Let D denote the diameter of the constraint set $\mathcal{X}$ and let $\kappa \leq 0$ be a uniform lower bound on the sectional curvature of $\mathcal{M}$. The iteration complexity of these Riemannian first-order methods depends explicitly on the geometry through the quantity

$$\zeta(\kappa, D) = \frac{\sqrt{|\kappa|}D}{\tanh(\sqrt{|\kappa|}D)}, \tag{3}$$

which appears in a nonlinear trigonometric distance bound (Zhang and Sra, 2016, Lem. 6). Clearly, this is an intrinsic geometric quantity of the manifold $\mathcal{M}$ in conjunction with the diameter D of the constraint set.

Boumal, Absil, and Cartis (Boumal et al., 2019) develop Riemannian gradient descent and trust region methods for minimizing nonconvex objectives φ on submanifolds $\mathcal{M}$ embedded in Euclidean space. A key assumption is the smoothness of the "pull-back objective" $\hat{\varphi}_x := \varphi \circ \mathrm{Rtr}_x : T_x\mathcal{M} \to \mathbb{R}$ defined on the tangent spaces, where Rtr_x denotes the retraction operator at x (see Sec. 2.1 for details). Specifically, there must exist a constant $L_g > 0$ such that for every $x \in \mathcal{M}$ and all sufficiently small $\eta \in T_x\mathcal{M}$,

$$|\hat{\varphi}_x(\eta) - \hat{\varphi}_x(0) - \langle \operatorname{grad} \varphi(x), \eta \rangle| \leq \frac{L_g}{2}\|\eta\|^2. \tag{4}$$

Noting that $\langle \operatorname{grad} \varphi(x), \eta \rangle = \langle \nabla \hat{\varphi}(0), \eta \rangle$ (see (6)), (4) is simply the standard Euclidean L_g-smoothness of the pull-back objective $\hat{\varphi}_x$.

Assuming $\mathcal{M}$ is a compact submanifold, Boumal et al. (2019, Pf. of Lem. 4) show that the uniform L_g-smoothness of the pull-back objectives holds with

$$L_g = \frac{L}{2}\alpha^2 + G\beta^2,$$

where L is the Euclidean smoothness parameter of the objective φ and G is the supremum of the Euclidean gradient $\nabla\varphi$ in the convex hull of $\mathcal{M}$. This property follows from two inequalities: For some constants $\alpha, \beta > 0$, and for each $\theta \in \mathcal{M}$ and $v \in T_\theta\mathcal{M}$,

$$\|\mathrm{Rtr}_\theta(v) - \theta\| \leq \alpha\|v\|, \qquad \|\mathrm{Rtr}_\theta(v) - (\theta + v)\| \leq \beta\|v\|^2. \tag{5}$$

These inequalities can be viewed as first- and second-order Euclidean approximations of the retraction operator. For the flat (Euclidean) case where $\mathrm{Rtr}_\theta(v) = \theta + v$, we may take $\alpha = 1$ and $\beta = 0$. The quantity β increases as the manifold becomes more curved, as one must travel a greater distance to retract from $\theta + v$ back to the manifold $\mathcal{M}$. An upper bound on β depending on the diameter of $\mathcal{M}$ is obtained in Boumal et al. (2019, Pf. of Lem. 4). In this sense, β acts as an extrinsic geometric quantity of the compact submanifold $\mathcal{M}$.

BMM, BCD, and Riemannian BCD. Block majorization-minimization (BMM) provides a flexible and unifying framework for optimization in Euclidean space by combining block optimization and MM. This framework encompasses popular algorithms such as block coordinate descent (BCD) and alternating minimization (Wright, 2015; Beck and Tetruashvili, 2013). Euclidean BMM for convex problems is studied in Hong et al. (2017) with general surrogates and in Xu and Yin (2013) with prox-linear surrogates. The iteration complexity of BMM for general nonconvex nonsmooth minimization was established as $O(\varepsilon^{-2})$ by Lyu and Li (2025), where an additional trust-region constraint with diminishing radius was shown to improve the constant factor. However, Riemannian generalizations of BMM have only been considered recently.

Tangential-type Riemannian BCD has appeared in recent literature. Gutman and Ho-Nguyen (2023) considered minimizing an objective via tangent subspace descent. In this approach, one moves within a chosen subspace of the tangent space (selecting a block in the tangent space) along the negative Riemannian gradient, then applies the exponential map to return to the manifold. For nonconvex g-smooth objectives (see Def. 3), they obtained an iteration complexity of $O(\varepsilon^{-2})$. The implied constant depends on the g-smoothness parameter of the objective, linking the convergence rate to the underlying geometry. Peng and Vidal (2023) also studied tangential Riemannian BCD for minimizing a smooth objective on the product of compact submanifolds in Euclidean space, utilizing compactness to ensure (5) holds globally. They established both asymptotic convergence to stationary points and iteration complexity for this method.

Tangential BMM in Li et al. (2024). It is well-known (see, e.g., Lyu and Li (2025)) that standard projected gradient descent in Euclidean space is a special instance of MM using a prox-linear surrogate (the first-order Taylor expansion of the objective plus a Euclidean proximal term). Similarly, standard Riemannian gradient descent is a special case of "tangential MM" with a prox-linear surrogate constructed on the tangent space (Li et al., 2024). Specifically, for an objective $\varphi : \mathcal{M} \to \mathbb{R}$, the global minimizer of the tangential prox-linear surrogate at $\theta \in \mathcal{M}$,

$$\hat{g}_\theta(\eta) := \varphi(\theta) + \langle \operatorname{grad} \varphi(\theta), \eta \rangle + \frac{\lambda}{2}\|\eta\|^2, \quad \eta \in T_\theta\mathcal{M},$$

is $-\frac{1}{\lambda} \operatorname{grad} \varphi(\theta)$. A sufficient condition for the descent of the objective value is that λ is large enough for $\hat{g}_n$ to majorize the pull-back objective $\hat{\varphi}_\theta := \varphi \circ \mathrm{Rtr}_\theta$ on the tangent space (e.g., $\lambda \geq L_g$, where L_g is the smoothness constant in (4)).

A natural generalization of this approach is to replace the tangential prox-linear surrogate $\hat{g}_\theta$ with a general tangential surrogate $\hat{g}_\theta : T_\theta\mathcal{M} \to \mathbb{R}$ such that $\hat{g}_n \geq \hat{\varphi}_\theta$. This yields a strategy where we find a descent direction on the tangent space via MM, take a step, and retract:

$$\begin{cases} \hat{g}_n & \leftarrow \Big[\text{Majorizing surrogate of } \hat{\varphi}_n \text{ s.t. } \hat{g}_n(\mathbf{0}) = \hat{\varphi}_n(\mathbf{0})\Big] \qquad (\triangleright \textit{ tangential surrogate}) \\ V_n & \leftarrow \arg\min_{\eta \in T_{\boldsymbol{\theta}_{n-1}}\mathcal{M}} \hat{g}_n(\eta), \qquad (\triangleright \textit{ descent direction}) \\ \boldsymbol{\theta}_n & \leftarrow \mathrm{Rtr}_{\boldsymbol{\theta}_{n-1}}(\alpha_n V_n) \qquad (\triangleright \textit{ Riemannian gradient update}) \end{cases}$$

where $\alpha_n > 0$ is a suitable stepsize.

In our prior work (Li et al., 2024), we proposed a cyclic block extension of this method called *tangential BMM* (tBMM) and obtained an iteration complexity of $\tilde{O}(\varepsilon^{-2})$ for minimizing nonconvex objectives on products of manifolds. A key contribution of that work is allowing both the minimization step on the tangent space and the retraction (or exponential map) to be computed *inexactly*. The accumulated sub-optimalities become part of the constant factor in the complexity bound. This offers practical computational benefits, even for the Riemannian BCD in Gutman and Ho-Nguyen (2023), as the exponential map need only be computed approximately.

As noted in Li et al. (2024), a minor modification of the argument in Boumal et al. (2019, Pf. of Lem. 4) shows that the uniform smoothness of pull-back objectives holds restricted to a compact subset $\mathcal{X}$ of a (not necessarily compact) submanifold $\mathcal{M}$. That is, (5) holds for all $\theta \in \mathcal{X}$ and sufficiently small $\eta \in T_\theta \mathcal{M}$ where $\mathrm{Rtr}_\theta(\eta) \in \mathcal{X}$, with β depending on the diameter of $\mathcal{X}$ rather than $\mathcal{M}$ (analogous to $\zeta(\kappa, D)$ in (3)). This property suffices for the convergence analysis of tBMM. However, a computational caveat of tBMM in this constrained setting is the need to compute the "lifted constraint set" $T_\theta^* = \{\eta \in T_\theta\mathcal{M} : \mathrm{Rtr}_\theta(\eta) \in \mathcal{X}\}$ at every iteration, which can be expensive.

Manifold-type BMM. While tangential MM conceptually generalizes Riemannian gradient descent, it excludes important classes of Riemannian algorithms, such as Riemannian proximal point methods on Hadamard manifolds (Li et al., 2009; Wang et al., 2016; Bento et al., 2017) and MM on the Stiefel manifold with linear surrogates (Breloy et al., 2021). Since proximal sub-problems on manifolds may not be easily solved exactly, establishing the iteration complexity of inexact Riemannian proximal point methods is crucial; to our knowledge, such results are currently unavailable. The method in Breloy et al. (2021) is an excellent example where a suitable choice of majorizing surrogate on the manifold results in easily solved sub-problems (minimizing a linear function on the Stiefel manifold reduces to a truncated SVD; see Sec. 4.5). However, no iteration complexity result exists for this method either.

In the present work, we develop the manifold counterpart of Riemannian BMM. Conceptually, this is simpler than the tangential variant as it avoids tangent spaces and retractions. However, this simplicity relies on constructing smart surrogates that allow for accessible minimizers on the manifold. Since exact minimizers may not always be available, we focus on providing convergence guarantees for a wide class of surrogates even when sub-problems are solved only approximately, quantifying how accumulated inaccuracies affect the convergence rate. The only other instance of manifold-type Riemannian BMM we are aware of is the Riemannian block exact minimization in Peng and Vidal (2023), which assumes exact solutions of block surrogate minimizers and requires restrictive assumptions such as objective smoothness and compact manifolds.

In Table 1, we provide a summary of the aforementioned related work, along with the details.

Methods	Manifold	Objective	Constraints	Blocks	Complexity	Inexact comp.
Euclidean BMM (Hong et al., 2017) Euclidean Block PGD (Beck and Tetruashvili, 2013)	Euclidean	convex	convex	many	$\widetilde{O}(\varepsilon^{-1})$	✗
Euclidean BMM-DR (Lyu and Li, 2025)	Euclidean	non-convex	convex	many	$\widetilde{O}(\varepsilon^{-2})$	✓
Riemannian prox. (Li et al., 2009)	Hadamard	g-convex	g-convex	1	-	✗
Riemannian prox. (Bento et al., 2017)	Hadamard	g-convex	g-convex	1	$\widetilde{O}(\varepsilon^{-1})$	✗
Riemannian Prox-linear (line search)(Chen et al., 2020)	Riemannian & Compact	non-convex & smooth†	N/A	1	$\widetilde{O}(\varepsilon^{-2})$	✗
Block Riemannian GD (Exp) (Gutman and Ho-Nguyen, 2023)	Riemannian	non-convex	N/A	many	$\widetilde{O}(\varepsilon^{-2})$	✗
BMM on manifolds (Peng and Vidal, 2023)	Riemannian & compact	non-convex	N/A	many	$\widetilde{O}(\varepsilon^{-2})$	✗
RBMM (**Ours**) with surr.:						
g-smooth (Thm. 10)	Riemannian	non-convex & non-smooth	g-convex	many	$\widetilde{O}(\varepsilon^{-2})$	✓
Riemannian proximal (Thm. 7)	Riemannian	non-convex & non-smooth	g-convex	many	$\widetilde{O}(\varepsilon^{-2})$	✓
Euclidean proximal (Thm. 7)	Riemannian $\subseteq$ Euclidean	non-convex & non-smooth	g-convex & compact	many	$\widetilde{O}(\varepsilon^{-2})$	✓
Smooth (Cor. 11)	Euclidean/ Stiefel	non-convex & non-smooth	convex/ g-convex	many	$\widetilde{O}(\varepsilon^{-2})$	✓

Table 1: Our main contributions and comparison to existing results. "g-smooth" means being geodesically smooth with respect to the geometry of the underlying manifold, and "smooth" means being smooth with respect to the Euclidean geometry. "g-convex" means geodesic convexity of subsets of manifolds. $\widetilde{O}(\cdot)$ notation means big-O up to logarithmic factors. The objective function marked by "smooth†" only needs to be smooth in the Euclidean sense; In all other cases, it is required to be g-smooth with respect to the underlying manifold. The last column shows whether the method allows the inexact solution to a subproblem, i.e., the robustness under inexact computation. Details of comparison to known results can be found in Section 4.

1.3 Organization

The paper is organized as follows. We introduce the preliminaries and the notation in Section 2.1 and 2.2. In Section 2.3, we give a precise statement of the RBMM algorithm. We detail the standing assumptions and state our main results in Section 3. In Section 4, we present the applications of RBMM on specific manifolds as special cases of our general framework. We present more applications of our results in Section 5. We prove the convergence results of RBMM, Algorithm 1, throughout Section 6. Figure 1 provides a structure diagram of the present paper.

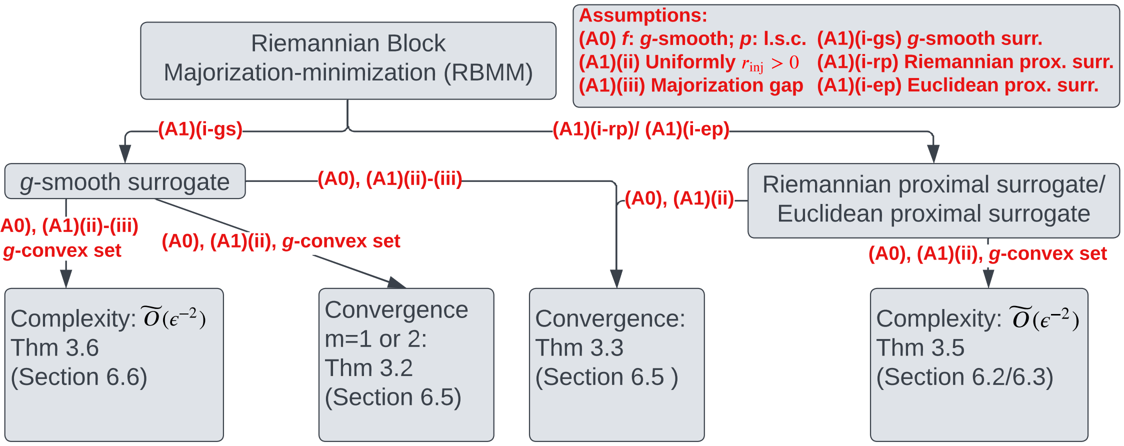

Riemannian Block Majorization-minimization (RBMM)
Assumptions:
(A0) f: g-smooth; p: l.s.c.
(A1)(i-gs) g-smooth surr.
(A1)(ii) Uniformly $r_{\text{inj}} > 0$
(A1)(i-rp) Riemannian prox. surr.
(A1)(iii) Majorization gap
(A1)(i-ep) Euclidean prox. surr.
(A1)(i-gs)
(A1)(i-rp)/ (A1)(i-ep)
g-smooth surrogate
(A0), (A1)(ii)-(iii)
Riemannian proximal surrogate/ Euclidean proximal surrogate
(A0), (A1)(ii)
(A0), (A1)(ii)-(iii) g-convex set
(A0), (A1)(ii), g-convex set
(A0), (A1)(ii), g-convex set
Complexity: $\widetilde{O}(\epsilon^{-2})$ Thm 3.6 (Section 6.6)
Convergence m=1 or 2: Thm 3.2 (Section 6.5)
Convergence: Thm 3.3 (Section 6.5)
Complexity: $\widetilde{O}(\epsilon^{-2})$ Thm 3.5 (Section 6.2/6.3)

Figure 1: Structure of the present paper.

2. Preliminaries and Algorithm

2.1 Preliminaries on Riemannian geometry

The notation throughout this paper is consistent with the common literature, see e.g., Absil et al. (2009) and Boumal (2023). For background knowledge on Riemannian geometry, we refer the readers to Sakai (1996), Lee (2003), Do Carmo and Flaherty Francis (1992), and Helgason (1979). In this section, we provide a brief introduction to the notation used in our paper, with further details provided in the Appendix A.

A *Riemannian manifold* $\mathcal{M}$ is a manifold endowed with a Riemannian metric $(\eta, \xi) \mapsto \langle \eta, \xi \rangle_x \in \mathbb{R}$, where η and ξ are tangent vectors in the tangent space $T_x\mathcal{M}$ (also denoted as T_x when it is clear from context) of $\mathcal{M}$ at x. This inner product in the tangent space is also denoted by $\langle \cdot, \cdot \rangle$ for convenience when the subscript is clear from the context. The induced norm on the tangent space is denoted by $\| \cdot \|_x$ or $\| \cdot \|$. The *Riemannian gradient* of a smooth function $f : \mathcal{M} \to \mathbb{R}$ at x is defined as the unique tangent vector, $\operatorname{grad} f(x) \in T_x\mathcal{M}$, such that $\langle \operatorname{grad} f(x), \xi_x \rangle = \mathrm{D} f(x) [\xi_x], \forall \xi_x \in T_x\mathcal{M}$, where $\mathrm{D} f(x) [\xi_x]$ is the differential of f at point x along the direction ξ_x. The geodesic distance between $x, y \in \mathcal{M}$ is denoted by $d_{\mathcal{M}}(x, y)$ or $d(x, y)$ when it is clear from the context.

A *retraction* on a manifold $\mathcal{M}$ is a locally defined smooth mapping Rtr from the tangent bundle $T\mathcal{M}$ to $\mathcal{M}$ with the following properties.

(i) For each $x \in \mathcal{M}$, let $\mathrm{r}_{\mathrm{Rtr}}(x) > 0$ be the 'retraction radius' such that the restriction $\mathrm{Rtr}_x : T_x\mathcal{M} \to \mathcal{M}$ of Rtr to $T_x\mathcal{M}$ is well-defined in a ball of radius $\mathrm{r}_{\mathrm{Rtr}}(x)$ around the origin $\mathbf{0} = \mathbf{0}_x$.

(ii) $\mathrm{Rtr}_x(\mathbf{0}) = x$; The differential of Rtr_x at $\mathbf{0}$, $D\,\mathrm{Rtr}_x(\mathbf{0})$, is the identity map on $T_x\mathcal{M}$.

For each $x \in \mathcal{M}$ and $\eta \in T_x\mathcal{M}$, the retraction curve $t \mapsto \mathrm{Rtr}_x(t\eta)$ agrees up to first order with geodesics passing through x with velocity η around x. Retractions provide a way to lift a function $g : \mathcal{M} \to \mathbb{R}$ onto the tangent space $T_x\mathcal{M}$ via its *pullback* $\hat{g} := g \circ \mathrm{Rtr}_x : T_x \to \mathbb{R}$. We use this construction to lift an upper-bounding surrogate defined on the manifold onto the tangent spaces in Algorithm 1. Note that for all $\eta \in T_x\mathcal{M}$,

$$\langle \nabla \hat{g}(\mathbf{0}),\, \eta \rangle = D\hat{g}(\mathbf{0})[\eta] = Dg(x)[D\,\mathrm{Rtr}_x(\mathbf{0})[\eta]] = Dg(x)[\eta] = \langle \mathrm{grad}\, g(x),\, \eta \rangle. \tag{6}$$

If for all $x \in \mathcal{M}$ and $\eta \in T_x\mathcal{M}$, the retraction curve $t \mapsto \mathrm{Rtr}_{t\eta}$ coincides with the geodesic curve $t \mapsto \gamma(t)$ with $\gamma(0) = x$ and $\gamma'(0) = \eta$ whenever $\|t\eta\| \le \mathrm{r}_{\mathrm{Exp}}(x)$ for some constant $\mathrm{r}_{\mathrm{Exp}}(x) > 0$, then the retraction Rtr is called the *exponential map* and denoted as Exp. Note that the exponential map is defined as the solution of a nonlinear ordinary differential equation. While every Riemannian manifold admits the exponential map, its computation is often challenging. Retractions provide computationally efficient alternatives to exponential maps. Some typical choices of retractions are $\mathrm{Rtr}_x(\eta) = x + \eta$ on Euclidean spaces and $\mathrm{Rtr}_x(\eta) = \frac{x+\eta}{\|x+\eta\|}$ on spheres. See Sec. 4.1. in Absil et al. (2009) for more examples of retractions. If the exponential map is defined on the entire tangent bundle (i.e., $\mathrm{r}_{\mathrm{Exp}}(x) = \infty$ for all $x \in \mathcal{M}$), then we say $\mathcal{M}$ is (geodesically) *complete*. Note by definition of exponential map we have $\|\eta\| = d(x, y)$ whenever $d(x, y) \le \mathrm{r}_{\mathrm{Exp}}(x)$ and $\mathrm{Exp}_x(\eta) = y$.

A Riemannian manifold is locally diffeomorphic to its tangent spaces, so it resembles the Euclidean space within a small metric ball around each point. Accordingly, there are several notions of radius functions $r : \mathcal{M} \to [0, \infty)$, including the injectivity and the convexity radii. For $x \in \mathcal{M}$, consider the open ball $B(x, r) = \{\eta \in T_x\mathcal{M} : \langle \eta, \eta \rangle < r\} \subseteq T_x\mathcal{M}$; the *injectivity radius* of $\mathcal{M}$ at x, denoted as $\mathrm{r}_{\mathrm{inj}}(x)$, is the supremum of values of r such that Exp_x defines a diffeomorphism from $B(x, r)$ to its image on $\mathcal{M}$. Thus, we can also define the inverse exponential map from $\mathcal{M}$ to $T_x\mathcal{M}$, denoted by $\mathrm{Exp}_x^{-1}(\cdot)$, within the injectivity radius. In particular, compact manifolds have uniformly positive injectivity radius (see Thm. III.2.3 in Chavel (2006)). Furthermore, it is worth noting that Hadamard manifolds (complete and simply connected manifolds with non-positive curvature), which include Euclidean space, hyperbolic space, and manifolds of positive definite matrices, also have uniformly positive injectivity radius (Afsari (2011), Sakai (1996, Theorem 4.1, p.221)). When the injectivity radius is uniformly positive, there exists a retraction with a uniformly positive retraction radius (e.g., the exponential map).

For a subset $\boldsymbol{\Theta} \subseteq \mathcal{M}$ and $x \in \boldsymbol{\Theta}$, define the *lifted constraint set* $T_x^{\boldsymbol{\Theta}}\mathcal{M}$ as

$$T_x^{\boldsymbol{\Theta}}\mathcal{M} := \{u \in T_x\mathcal{M} \,|\, \mathrm{Rtr}_x(u) = x' \text{ for some } x' \in \boldsymbol{\Theta} \text{ with } d(x, x') \le r_0/2\}, \tag{7}$$

where r_0 is the lower bound of the injectivity radius (see (A0) in Section 3). When we use exponential map Exp as the retraction in (7), one can think of the set $T_x^{\boldsymbol{\Theta}}\mathcal{M}$ as the 'lift' of the constraint set $\boldsymbol{\Theta}$ onto the tangent space $T_x\mathcal{M}$ in the sense that if $\boldsymbol{\Theta}$ is contained in the metric ball of radius $\mathrm{r}_{\mathrm{inj}}(x)$ centered at x, then $T_x^{\boldsymbol{\Theta}}\mathcal{M}$ equals the inverse image $\mathrm{Exp}_x^{-1}(\boldsymbol{\Theta})$ of $\boldsymbol{\Theta}$ under Exp_x. In particular, this formula holds for all subset $\boldsymbol{\Theta}$ of $\mathcal{M}$ if $\mathcal{M}$ is a complete Riemannian manifold since Exp_x is defined on the entire $T_x\mathcal{M}$. If $\mathcal{M}$ is a Euclidean space and $\boldsymbol{\Theta}$ is a convex subset of it, then $T_x^{\boldsymbol{\Theta}}\mathcal{M} = \boldsymbol{\Theta}$. Lastly, we note that when $\boldsymbol{\Theta}$ is strongly convex in $\mathcal{M}$, the set $T_x^{\boldsymbol{\Theta}}\mathcal{M}$ above is locally defined near x. That is, we can replace the injectivity radius r_0 in (7) by any constant $\delta \in (0, r_0)$.

A set $C \subseteq \mathcal{M}$ is called (geodesically) *strongly convex* if for any x and y in C, there is a unique minimal geodesic γ in $\mathcal{M}$ joining x and y, and γ is contained in C. Within a g-convex set C, we can also define the *g-convex function* f as follows: For any $x = \gamma(0), y = \gamma(1) \in C$, and any $t \in [0, 1]$, it holds that

$$f(\gamma(t)) \le (1 - t)f(x) + tf(y). \tag{8}$$

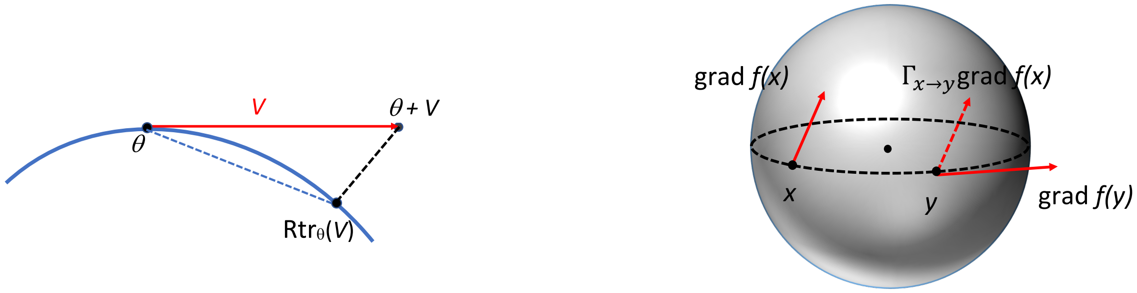


Figure 2: Illustration of (left) retraction and (right) parallel transport.

When $d(x, y) \le \mathrm{r_{inj}}(x)$, an equivalent definition of g-convex function is given by

$$f(y) \ge f(x) + \langle \partial_R f(x), \gamma'(0) \rangle, \tag{9}$$

where $\partial_R f(x)$ is a Riemannian subgradient of f at x, or the Riemannian gradient if f is differentiable (Zhang and Sra, 2016).

On Riemannian manifolds, parallel transport provides a way to transport a vector along a smooth curve. For each smooth curve $\gamma : [0, 1] \to \mathcal{M}$, denote the *parallel transport* along γ from $x = \gamma(0)$ to $y = \gamma(t)$ by $\Gamma^{\gamma}_{x\to y}$. If γ is clear from the context (e.g., the unique distance-minimizing geodesic from x to y), then we also write $\Gamma^{\gamma}_{x\to y} = \Gamma_{x\to y}$. Intuitively, a tangent vector $\eta \in T_x\mathcal{M}$ at x of γ is still a tangent vector $\Gamma(\gamma)\eta \in T_y\mathcal{M}$ of γ at y. Recall that, one important property of parallel transport on a Riemannian manifold is that it preserves the inner product, i.e.,

$$\left\langle \Gamma^{\gamma}_{\gamma(0)\to\gamma(t)}\,(\xi),\, \Gamma^{\gamma}_{\gamma(0)\to\gamma(t)}\,(\zeta) \right\rangle_{\gamma(t)} = \langle \xi,\, \zeta \rangle_{\gamma(0)} \quad \text{for all } t \in [0, 1],\ \xi, \zeta \in T_{\gamma(0)}.$$

See Figure 2 for an illustration.

2.2 Notation for block Riemannian optimization

In (1), we are interested in minimizing an objective function f within the product parameter space $\boldsymbol{\Theta} = \Theta^{(1)} \times \cdots \times \Theta^{(m)}$, where each constraint set $\Theta^{(i)}$ is a subset of a Riemannian manifold $\mathcal{M}^{(i)}$. It will be convenient to introduce the following notation: For $\boldsymbol{\theta} = [\theta^{(1)}, \dots, \theta^{(m)}]$,

$$\begin{aligned} \mathrm{grad}_i\, f(\boldsymbol{\theta}) &:= \text{Riemmanian gradient of } \theta \mapsto f(\theta^{(1)}, \dots, \theta^{(i-1)}, \theta, \theta^{(i+1)}, \dots, \theta^{(m)}), \\ \mathrm{grad}\, f(\boldsymbol{\theta}) &:= [\mathrm{grad}_1\, f(\boldsymbol{\theta}), \dots, \mathrm{grad}_m\, f(\boldsymbol{\theta})], \\ d(\mathbf{x}, \mathbf{y}) &:= \sqrt{\sum_{i=1}^{m} d(x^{(i)}, y^{(i)})^2} \quad \text{for } \mathbf{x} = (x^{(1)}, \dots, x^{(m)}),\ \mathbf{y} = (y^{(1)}, \dots, y^{(m)}) \in \prod_{i=1}^{m} \mathcal{M}^{(i)}. \end{aligned} \tag{10}$$

Note that if we endow the product $\prod_{i=1}^{m} \mathcal{M}^{(i)}$ of the manifolds a joint Riemannian structure, then we can interpret grad $f(\boldsymbol{\theta})$ above as the Riemannian gradient at $\boldsymbol{\theta}$ with respect to that joint Riemannian structure. However, we do not explicitly introduce or use such a product manifold structure in the manuscript.

Throughout this paper, we let $(\boldsymbol{\theta}_n)_{n\geq 1}$ denote an output of Algorithm 1 and write $\boldsymbol{\theta}_n = [\theta_n^{(1)}, \dots, \theta_n^{(m)}]$ for each $n \geq 1$. For each $n \geq 1$ and $i = 1, \dots, m$, denote

$$f_n^{(i)} : \theta \mapsto f(\theta_n^{(1)}, \dots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \dots, \theta_{n-1}^{(m)}),$$

which we will refer to as the ith marginal objective function at iteration n. Define

$$\begin{aligned} \boldsymbol{\theta}_{n;i} &:= \left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta_n^{(i)}, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right), \\ (\theta, \boldsymbol{\theta}_{n;i}) &:= \left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right). \end{aligned}$$

2.3 Statement of algorithm

Below in Algorithm 1, we give a precise statement of the RBMM algorithm we stated in high-level at (2). We first define majorizing surrogate functions defined on Riemannian manifolds.

Definition 2 (Majorizing surrogates on Riemannian manifolds) *Fix a function $h : \mathcal{M} \to \mathbb{R}$ and a point $\boldsymbol{\theta} \in \mathcal{M}$, where $\mathcal{M}$ is a Riemannian manifold. A function $g : \mathcal{M} \to \mathbb{R}$ is a majorizing surrogate of h at $\boldsymbol{\theta}$ if*

$$g(x) \geq h(x) \quad \text{for all } x \in \mathcal{M} \quad \text{and} \quad g(\boldsymbol{\theta}) = h(\boldsymbol{\theta}).$$

Note in Definition 2, the surrogate $g(x)$ may depend on the base point $\boldsymbol{\theta}$ in certain applications and is therefore denoted as $g(x|\boldsymbol{\theta})$ (Breloy et al., 2021). However, for simplicity, we omit $\boldsymbol{\theta}$ in the notation throughout this paper.

As mentioned before, the high-level idea is the following. In order to update the ith block of the parameter $\theta_n^{(i)}$ at iteration n, we use the RBMM we described in the introduction.

Algorithm 1 Riemannian Block Majorization-Minimization (RBMM)

```
Input: $\boldsymbol{\theta}_0 = (\theta_0^{(1)}, \cdots, \theta_0^{(m)}) \in \Theta^{(1)} \times \cdots \times \Theta^{(m)}$ (initial estimate); $N$ (number of iterations)
  for $n = 1, \dots, N$ do:
    Update estimate $\boldsymbol{\theta}_n = [\theta_n^{(1)}, \cdots, \theta_n^{(m)}]$ by
      For $i = 1, \cdots, m$ do:
      $F_n^{(i)}(\cdot) := F\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \cdot, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right) : \mathcal{M}^{(i)} \to \mathbb{R}$      (▷ marginal objective)
      $\begin{cases} g_n^{(i)} \leftarrow \left[\text{Majorizing surrogate of } f_n^{(i)} \text{ at } \theta_{n-1}^{(i)}\right] \\ \theta_n^{(i)} \in \arg\min_{\theta \in \Theta^{(i)}} \left(G_n^{(i)}(\theta) := g_n^{(i)}(\theta) + p_n^{(i)}(\theta)\right) \end{cases}$
      end for
  end for
output: $\boldsymbol{\theta}_N$
```

In Algorithm 1, the majorizing surrogate $g_n^{(i)}$ at each iteration n for each block i is chosen so that

(1) (Majorization) $g_n^{(i)}(x) - f_n^{(i)}(x) \geq 0$ for all $x \in \mathcal{M}^{(i)}$;

(2) (Sharpness) $g_n^{(i)}(\theta_{n-1}^{(i)}) = f_n^{(i)}(\theta_{n-1}^{(i)})$.

A direct implication from the above requirements on majorizing surrogates is the agreement of the Riemannian gradient between surrogates and marginal objective functions. Namely, for each n and i, we have $\operatorname{grad} g_n^{(i)}(\theta_{n-1}^{(i)}) = \operatorname{grad} f_n^{(i)}(\theta_{n-1}^{(i)})$. This is an essential observation for analyzing the complexity of RBMM.

3. Statement of Results

In this section, we state our main results concerning the convergence and complexity of our RBMM algorithm (Alg. 1) for the constrained block Riemannian optimization problem in (1). Figure 1 provides a structure diagram of the main results and assumptions.

3.1 Optimality and complexity measures

For iterative algorithms, first-order optimality conditions may hardly be satisfied exactly in a finite number of iterations, so it is more important to know how the worst-case number of iterations required to achieve an ε-approximate solution scales with the desired precision ε.

More precisely, for the multi-block problem (1), we introduce the following bi-variate function that we will use to define a generic first-order optimality measure for the block nonconvex and nonsmooth problem:

$$V(\boldsymbol{\theta}_*, \boldsymbol{u}) := \left(\sum_{i=1}^{m} \langle -\operatorname{grad}_i f(\boldsymbol{\theta}_*), u^{(i)} \rangle\right) + \frac{1}{\hat{r}}\left(p(\boldsymbol{\theta}_*) - \sum_{i=1}^{m} p^{(i)}\left(\operatorname{Exp}_{\theta_*^{(i)}}(\hat{r} u^{(i)})\right)\right), \tag{11}$$

where $\boldsymbol{\theta}_* = [\theta_*^{(1)}, \dots, \theta_*^{(m)}] \in \boldsymbol{\Theta}$ is a parameter and $\mathbf{u} = (u_1, \dots, u_m)$, $u_i \in T_{\theta_*^{(i)}}^{\Theta^{(i)}}$ is tuple of tangent vectors and $\hat{r} = \min\{r_0, 1\}$ denotes the minimum between 1 and the lower bound of injectivity radius r_0. Then we say $\boldsymbol{\theta}_*$ is an *ε-stationary point* of (1) if

$$\sup_{\mathbf{u}=(u^{(i)},\dots,u^{(m)}),\, \|u^{(i)}\| \le 1} V(\boldsymbol{\theta}_*, \mathbf{u}) \le \varepsilon, \tag{12}$$

and a *stationary point* of (1) if the above holds with $\varepsilon = 0$.

The Euclidean counterpart of the optimality measure (11) was first proposed in Lyu and Li (2025) for nonconvex nonsmooth optimization, and here we adapt it to the block Riemannian optimization. In the literature of nonsmooth optimization, another widely used stationarity measure is

$$\sup_{\boldsymbol{u}, \|u^{(i)}\| \le 1} \langle -\partial_R F(\boldsymbol{\theta}_*), \boldsymbol{u} \rangle \le 0. \tag{13}$$

In fact, (13) is equivalent to the stationarity measure by (11). To see this, first note by the g-convexity of p (9), we have that (13) implies (12) with $\varepsilon = 0$. Conversely, let $\phi(\boldsymbol{u}) = V(\boldsymbol{\theta}_*, \boldsymbol{u})$. Then $\phi(\boldsymbol{u})$ is a concave function with local maximum of $\phi(\mathbf{0}) = 0$. Hence by first-order optimality of $\mathbf{0}$ being the local minimizer of $-\phi$ and noting that $-\partial\phi = \partial_R F(\boldsymbol{\theta}_*)$ gives (13).

An important advantage of using the optimality measure V in (11) is that the bi-variate function V is *continuous* when the nonsmooth part p is continuous, while the corresponding measure in (13) is not.

In the unconstrained smooth block-Riemannian setting where $\Theta^{(i)} = \mathcal{M}^{(i)}$ for $i = 1, \dots, m$ and $p = 0$, the above equation (11) becomes

$$\sum_{i=1}^{m} \|\mathrm{grad}_i\, f(\boldsymbol{\theta}_*)\| \leq \varepsilon.$$

In the case of single-block $m = 1$, the above is the standard definition of ε-stationary points for unconstrained Riemannian optimization problems.

Next, for each $\varepsilon > 0$ we define the *(worst-case) iteration complexity* N_ε of an algorithm computing $(\boldsymbol{\theta}_n)_{n\geq 1}$ for solving (1) as

$$N_\varepsilon := \sup_{\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}} \inf \{ n \geq 1 \,|\, \boldsymbol{\theta}_n \text{ is an } \varepsilon\text{-approximate stationary point of } F \text{ over } \boldsymbol{\Theta} \}, \qquad (14)$$

where $(\boldsymbol{\theta}_n)_{n\geq 0}$ is a sequence of estimates produced by the algorithm with an initial estimate $\boldsymbol{\theta}_0$. Note that N_ε gives the *worst-case* bound on the number of iterations for an algorithm to achieve an ε-approximate solution due to the supremum over the initialization $\boldsymbol{\theta}_0$ in (14).

3.2 Statement of results

Here we state our main convergence results of solving the minimization problem (1) using Algorithm 1. First, we introduce a Riemannian counterpart of a smoothness property of a function $f : \mathcal{M} \to \mathbb{R}$. In the Euclidean setting, the function f is L-smooth if its gradient ∇f is L-Lipschitz continuous. In the Riemannian case, Riemannian gradients $\mathrm{grad}\, f(x)$ and $\mathrm{grad}\, f(y)$ at two base points $x, y \in \mathcal{M}$ live in different tangent spaces $T_x\mathcal{M}$ and $T_y\mathcal{M}$, so they have to be compared using a parallel transport (see Figure 2 for an illustration). We extend this notion of smoothness to the setting where the function f is defined on the product of Riemannian manifolds.

Definition 3 (Geodesic smoothness) *A function $f : \prod_{i=1}^m \mathcal{M}^{(i)} \to \mathbb{R}$ is geodesically smooth (g-smooth in short) with parameter $L > 0$ if f is block-wise continuously differentiable and for each $\mathbf{x} = (x^{(1)}, \dots, x^{(m)}), \mathbf{y} = (y^{(1)}, \dots, y^{(m)}) \in \prod_{i=1}^m \mathcal{M}^{(i)}$ where there exists a minimizing geodesic joining $x^{(i)}$ and $y^{(i)}$ for each $i = 1, \dots, m$,*

$$\left\| \mathrm{grad}_i\, f(\mathbf{x}) - \Gamma_{y^{(i)} \to x^{(i)}}(\mathrm{grad}_i\, f(\mathbf{y})) \right\| \leq \frac{L}{m} d(\mathbf{x}, \mathbf{y}),$$

where $\Gamma_{y^{(i)} \to x^{(i)}}$ is the parallel transport along a distance-minimizing geodesic joining $x^{(i)}$ and $y^{(i)}$ in $\mathcal{M}^{(i)}$, and $d(\mathbf{x}, \mathbf{y})$ is defined in (10).

An important consequence of the g-smoothness is the following quadratic bound on first-order approximation:

$$\left| f(y) - f(x) - \left\langle \mathrm{grad}\, f(x), \gamma'(0) \right\rangle_x \right| \leq \frac{L}{2} d^2(x, y),$$

where γ is any distance-minimizing geodesic in $\mathcal{M}$ from x to y and $d(x, y)$ is the Riemannian distance between x and y. See Lemma 38 for the proof. Throughout this paper, the term "g-smooth" denotes geodesic smooth (Def. 3), while "L-smooth" indicates the function L-smooth in Euclidean sense, meaning the Euclidean gradient of the function is L-Lipschitz continuous.

We start by stating some general assumptions. We allow inexact computation of the solution to the minimization sub-problems in Algorithm 1 (line 6). This is practical since minimizing the surrogates on the manifold (possibly with additional constraints) may not always be exactly solvable. To be precise, for each $n \geq 1$, we define the *optimality gap* Δ_n by

$$\Delta_n = \Delta_n(\boldsymbol{\theta}_0) := \max_{1\leq i\leq m}\left(G_n^{(i)}(\theta_n^{(i)}) - \inf_{\theta\in\Theta^{(i)}} G_n^{(i)}(\theta)\right) \tag{15}$$

For the convergence analysis to hold, we require that the optimal gaps decay fast enough so that they are summable, stated as Assumption (A0).

(A0) *For RBMM (Alg. 1), we make the following assumptions:*

(i) *(Objective) The objective $f : \boldsymbol{\Theta} = \prod_{i=1}^m \Theta^{(i)} \to \mathbb{R}$ is g-smooth with some parameter $L_f > 0$; p is block-separable , i.e., $p(\boldsymbol{\theta}) = \sum_i p^{(i)}(\theta^{(i)})$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ where each $p^{(i)}$ is g-convex. Futhermore, the objective value of F is uniformly lower bounded by some $F^* \in \mathbb{R}$, and the sublevel sets $F^{-1}((-\infty, a)) = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : F(\boldsymbol{\theta}) \leq a\}$ are compact for each $a \in \mathbb{R}$.*

(ii) *(Inexact computation) The optimality gaps Δ_n in (15) are summable, that is, $\sum_{n=1}^\infty \Delta_n < \infty$. Furthermore, let $\theta_n^{(i\star)}$ be an exact solution of the minimization step in Algorithm 1 and let $\theta_n^{(i)}$ be the inexact output. Then for all $i = 1, \dots, m$,*

$$\lim_{n\to\infty} d(\theta_n^{(i\star)}, \theta_n^{(i)}) = 0. \tag{16}$$

The separability of p in (A0)**(i)** is a common assumption in the literature of block optimization (Xu and Yin, 2013), as $p^{(i)}$ is often a regularizer applied to the i-th block, such as the l_1 regularizer.

If the surrogate function $g_n^{(i)}$ is geodesically strongly convex (see Definition 28), then (16) is a direct consequence of the summability of the optimality gaps. In particular, this is the case for Riemannian proximal surrogates on Hadamard manifolds as a special case of (A1)**(i-rp)** (see Prop. 30).

Next, we impose the following conditions on the underlying manifolds, majorizing surrogates, and constraint sets.

(A1) *For the manifolds and surrogates, at least one of the following holds:*

(i-gs) *(g-smooth surrogates) Each surrogate $g_n^{(i)} : \mathcal{M}^{(i)} \to \mathbb{R}$ is g-smooth with some parameter $L_g \geq 0$ for all $n \geq 1$ and $i = 1, \dots, m$.*

(i-rp) *(Riemannian Proximal surrogates) Each surrogate $g_n^{(i)}$ is a Riemannian proximal surrogate; that is, for each $n \geq 1$, $\lambda_{\min} \leq \lambda_n \leq \lambda_{\max}$ for some $\lambda_{\min}, \lambda_{\max} > 0$,*

$$g_n^{(i)}(\theta) = f_n^{(i)}(\theta) + \frac{\lambda_n}{2} d^2(\theta, \theta_{n-1}^{(i)}). \tag{17}$$

(i-ep) *(Euclidean Proximal surrogates) Each $\mathcal{M}^{(i)}$ is an embedded submanifold in an Euclidean space and the constraint set $\Theta^{(i)}$ is compact. Each surrogate $g_n^{(i)}$ is a Euclidean proximal surrogate, that is, for each $n \geq 1$, $\lambda_{\min} \leq \lambda_n \leq \lambda_{\max}$ for some $\lambda_{\min}, \lambda_{\max} > 0$,*

$$g_n^{(i)}(\theta) = f_n^{(i)}(\theta) + \frac{\lambda_n}{2} \|\theta - \theta_{n-1}^{(i)}\|^2. \tag{18}$$

Furthermore, we require the following for the constraint sets:

(ii) *For each $i = 1, \ldots, m$, there exists a uniform lower bound $r_0 > 0$ for the injectivity radius $\mathrm{r_{inj}}(x)$ over $x \in \Theta^{(i)}$, i.e., $\mathrm{r_{inj}}(x) \geq r_0$ for all $x \in \Theta^{(i)}$.*

When the manifolds $\mathcal{M}^{(i)}$ are non-Euclidean, constructing g-smooth surrogates may not always be easy. However, for Stiefel manifolds (see Ex. 1), Euclidean smoothness and g-smoothness are essentially equivalent (see Lem. 15) so standard smooth surrogates in the Euclidean space (e.g, proximal (18) and prox-linear (22)) verifies (A1)**(i-gs)**. This fact is crucially used in the proof of Corollaries 1 and 11.

Note that both the Riemannian and the Euclidean proximal surrogates in (17)-(18) are not necessarily g-smooth (see Section 6.2 for details). Our analysis of RBMM with these proximal surrogates uses explicit forms of the Riemannian/Euclidean gradient of the squared geodesic/Euclidean distance function (see Prop. 31) instead of g-smoothness of the surrogates.

One of the major benefits of RBMM is that the user has the freedom to choose problem-specific majorizing blockwise surrogates, as long as they satisfy the properties in (A1). An important design principle is that the problem of surrogate minimization in (2) should be easily executable up to a reasonable error Δ_n. The class of (geodesically) strongly convex functions is often a good candidate. In the Euclidean BMM, one can choose surrogates that are strongly convex that can be easily minimized by using standard Euclidean convex minimization algorithms (e.g., Projected GD). If the marginal objective $f_n^{(i)}$ is already (geodesically) strongly convex, then we can directly minimize it over the corresponding manifold. Otherwise, we may use large enough proximal regularization constant λ_n (e.g., $\lambda_n > L_f$) to warrant geodesic convexity of surrogates. For instance, if the manifold has non-positive sectional curvature (e.g., Hadamard manifolds, see Ex. 3), d^2 is g-strongly convex. However, for Riemannian BMM, the surrogates do not have to be g-convex to yield a practical algorithm (e.g., linear surrogates over Steifel manifold, see Sec. 4.5).

The surrogates in (A1)**(i-rp)** and (A1)**(i-ep)** are identical when the manifold $\mathcal{M}^{(i)}$ is the Euclidean space. For non-Euclidean manifolds, the difference between these two surrogates could be significant since the Riemannian proximal surrogates in (A1)**(i-rp)** use the geometry of the non-Euclidean manifold $\mathcal{M}^{(i)}$ while the Euclidean proximal surrogates in (A1)**(i-ep)** use the geometry of the ambient Euclidean space. For example, in the case of a sphere, (A1)**(i-rp)** uses the arc length of great circles while (A1)**(i-ep)** uses length of line segments. While the Riemannian proximal surrogates in (A1)**(i-rp)** can handle non-compact manifolds without compact constraints (e.g., see Sec. 5.3), the Euclidean proximal surrogates in (A1)**(i-ep)**, if applicable, are computationally simpler to handle (e.g., ease of differentiation). We also remark that (A1)**(i-ep)** holds for unconstrained optimization problems on compact manifolds embedded in Euclidean space, i.e., $\Theta^{(i)} = \mathcal{M}^{(i)}$ and $\mathcal{M}^{(i)}$ is a compact embedded submanifold of a Euclidean space (e.g., Stiefel manifolds).

We now state our first main result, concerning the asymptotic convergence of RBMM for the constrained single/two-block block Riemannian optimization problem (1). The proofs of the theorems are provided in subsequent sections, beginning with Section 6.

Theorem 4 (Asymptotic convergence to stationary points; one or two blocks) *Let f denote the objective function ($p = 0$) in (1) with $m = 1$ or 2. Let $(\boldsymbol{\theta}_n)_{n\ge 0}$ be an output of Algorithm 1 under (A0) and (A1)**(i-gs)**,**(ii)**. Further assume that the constraint set $\Theta^{(i)}$ of $\mathcal{M}^{(i)}$ is strongly g-convex for $i = 1, \ldots, m$. Then every limit point of $(\boldsymbol{\theta}_n)_{n\ge 0}$ is a stationary point of f over $\boldsymbol{\Theta}$.*

Next, we established similar asymptotic stationarity of RBMM for more than two blocks. In this case, a well-known counterexample by Powell (1973) shows that block coordinate descent methods for $m \ge 3$ with block optimization may not converge to stationary points. Indeed, block coordinate descent is a special case of block MM with the surrogate functions $g_n^{(i)}$ identical to $f_n^{(i)}$. To overcome this issue, the proximal regularized version of block coordinate descent was proposed (Grippo and Sciandrone, 2000; Xu and Yin, 2013; Attouch et al., 2010), which had the convergence guarantees under certain assumptions. With this insight, we need to require one more assumption on the surrogate function $g_n^{(i)}$. Namely, writing

$$g_n^{(i)}(\theta) = f_n^{(i)}(\theta) + h_n^{(i)}(\theta),$$

one can think of the nonnegative 'surrogate gap' function $h_n^{(i)}$ as a regularization term. A primary example of such regularization in the Euclidean case is the proximal point regularization, where we would take $h_n^{(i)}(\theta) = \lambda\|\theta - \theta_{n-1}^{(i)}\|^2$. This has the effect of penalizing a large change in the iterates $\theta_n^{(i)} - \theta_{n-1}^{(i)}$. In the general Riemannian case, the following assumption states that the surrogate gap function $h_n^{(i)}$ is penalizing a large geodesic distance $d(\theta, \theta_{n-1}^{(i)})$.

(A1) (iii) *(Distance-regularizing surrogates) There exists a strictly increasing function $\phi : [0, \infty) \to \mathbb{R}$ such that $\phi(0) = 0$ and*

$$h_n^{(i)}(\theta) := g_n^{(i)}(\theta) - f_n^{(i)}(\theta) \ge c\phi(d(\theta, \theta_{n-1}^{(i)}))$$

for all $n \ge 1$ and $i = 1, \ldots, m$.

Note that if we use Riemannian proximal surrogates as in (A1)**(i-rp)**, then (A1)**(iii)** is automatically satisfied with $c = \lambda/2$ and $\phi(x) = x^2$. This also holds with a possibly different parameter c when we use the Euclidean proximal surrogates within a compact constraint set on an embedded manifold as in (A1)**(i-ep)**. In fact, the geodesic distance and Euclidean distance are equivalent over compact sets, see Lemma 16 for details. Also, if the surrogates $g_n^{(i)}$ are ρ-strongly g-convex and if $\theta_{n+1}^{(i)}$ is the exact minimizer of $g_n^{(i)}$, then by the second-order growth property, one can verify

$$h_n^{(i)}(\theta) \ge \frac{\rho}{2} d^2(\theta, \theta_{n-1}^{(i)}).$$

Hence (A1)**(iii)** is verified with $c = \rho/2$ and $\phi(x) = x^2$ in this case as well. It is worth mentioning that assumption (A1)**(iii)** is standard in the Euclidean MM literature, see Hien et al. (2023).

Note a direct implication of (A1)**(iii)** is that if $g_n^{(i)}(\theta_n) - f_n^{(i)}(\theta_n) = o(1)$ then $d(\theta_n^{(i)}, \theta_{n-1}^{(i)}) = o(1)$. This fact will be crucial to our proof of asymptotic stationarity of RBMM for $m \geq 3$ blocks, which is our second main result stated below. Note that in the following result, we do not require geodesic convexity of the constraint sets.

Theorem 5 (Asymptotic convergence to stationary points; many blocks) *Let F denote the objective function in (1) with $m \geq 1$. Let $(\boldsymbol{\theta}_n)_{n\geq 0}$ be an output of Algorithm 1. Under (A0), (A1)**(ii)-(iii)**, and any of **(i-gs)**, **(i-rp)** and **(i-ep)**. Futher assume p is lower semi-continuous on $\boldsymbol{\Theta}$. Then every limit point of $(\boldsymbol{\theta}_n)_{n\geq 0}$ is a stationary point of F over $\boldsymbol{\Theta}$.*

Remark 6 (Convexity of the constraint set) The proof of Theorem 4 requires the constraint set $\Theta^{(i)} \subseteq \mathcal{M}^{(i)}$ to be strongly convex w.r.t. the geometry of the ambient manifold $\mathcal{M}^{(i)}$ (due to the Riemmanian line search used in Prop. 36). However, Theorem 5 does not require the constraint set $\Theta^{(i)} \subseteq \mathcal{M}^{(i)}$ to be strongly convex, since we can use Assumption (A1)**(iii)** to avoid using Prop. 36. Therefore, Theorem 5 applies when $\Theta^{(i)}$ is a manifold by itself (e.g., low-rank manifold or Stiefel manifold) and $\mathcal{M}^{(i)}$ is a Euclidean space in which $\Theta^{(i)}$ is embedded. See Section 4.2 for details.

Now we state our result concerning the rate of convergence of RBMM for the case of Riemannian/Euclidean proximal (and not necessarily g-smooth) surrogates.

Theorem 7 (Rate of convergence for Riemannian/Euclidean proximal surrogates) *Let F denote the objective function in (1) with $m \geq 1$. Let $(\boldsymbol{\theta}_n)_{n\geq 0}$ be an output of Algorithm 1 under (A0) and (A1)**(ii)**, and either **(i-rp)** or **(i-ep)**. Assume the geodesic convexity of the constraint sets. Then the following hold:*

(i) *(Rate of convergence) There exists constant $c > 0$ independent of $\boldsymbol{\theta}_0$ and the manifold such that*

$$\min_{1\leq k\leq n} \left[\sup_{\eta^{(i)} \in T_{\theta_k^{(i)}}^{\Theta^{(i)}}, \|\eta^{(i)}\| \leq 1} V(\boldsymbol{\theta}_k, \boldsymbol{\eta}) \right] \leq c \frac{L_f(1 + \lambda_{\min}^{-1})(1 + \sum_{k=1}^n \Delta_k) + \lambda_{\max}}{\sqrt{n}/\log n}.$$

See the explicit expression of the RHS constant in (60).

(ii) *(Worst-case iteration complexity) The worst-case iteration complexity N_ϵ for Algorithm 1 satisfies $N_\epsilon = O(\varepsilon^{-2} \left(\log \varepsilon^{-1}\right)^2)$*

Theorem 7 establishes that RBMM with either Riemannian or Euclidean proximal regularizer achieves an iteration complexity of $\widetilde{O}(\varepsilon^{-2})$. The case of Euclidean proximal regularizer in (A1)**(i-ep)** may be of wider practical interest than the Riemannian proximal regularizer. Note in (A1)**(i-rp)**, the Riemannian proximal regularizer is geodesic distance squared, which may introduce additional computational difficulties, especially when the close-form expression of end-point geodesic distance is unknown. When the underlying manifold is embedded in Euclidean space and the constraint set is compact (or the manifold itself is compact), as in (A1)**(i-ep)**, we can replace the Riemannian proximal regularizer by the Euclidean proximal regularizer, which is often much easier to deal with computationally.

In Theorem 7, the iteration complexity results depend on the optimality gap Δ_n arising from approximately solving the surrogate minimization subproblems. In practice, each surrogate minimization can be efficiently handled using standard Riemannian projected (sub)gradient methods (see Section 6.4 and (61)). Using the iteration complexity analysis due to Zhang and Sra (2016) to bound Δ_n, we obtain the following corollary on the total complexity guarantee of Riemannian block proximal point method.

Corollary 8 (Total complexity of Riemannian Block Proximal Point Method) *Keep the same assumptions as in Theorem 7 with $g_n^{(i)}$ being the Riemannian proximal surrogate as in (A1)**(i-rp)**. Assume further that each $\mathcal{M}^{(i)}$ is a Hadamard manifold with sectional curvature lower bounded by $\kappa \in (-\infty, 0]$. Also assume that each constraint set $\Theta^{(i)}$ is a bounded subset of $\mathcal{M}^{(i)}$ with diameter at most D.*

(i) *(Smooth objectives) Assume $p = 0$. Then $g_k^{(i)}$ is g-strongly convex with parameter $\lambda_k - L_f$ and g-smooth with parameter $L_f + 2\zeta(\kappa, D)$, where $\zeta(\kappa, D) = \sqrt{|\kappa|}D/\tanh(\sqrt{|\kappa|}D)$. Approximately solve each proximal sub-problem $\min_{\Theta^{(i)} \subseteq \mathcal{M}^{(i)}} g_k^{(i)}$ by using the Riemannian projected gradient descent (61) for $N_k^{(i)}$ sub-iterations with fixed step-size $\frac{1}{L_f + 2\zeta(\kappa, D)}$. Then denoting $\delta = \min\{\frac{1}{\zeta(\kappa,D)}, \frac{\lambda_k - L_f}{L_f + 2\zeta(\kappa,D)}\}$, we have*

$$\Delta_k^{(i)} \leq (1-\delta)^{N_k^{(i)} - 2} D^2 (L_f + 2\zeta(\kappa, D)).$$

In particular, if we take $N_k^{(i)} = C \log k$ with $C = 2/(-\log(1-\delta))$, then the algorithm reaches an ε-stationary point within $\widetilde{O}\big((1+\zeta(\kappa, D))^2\, \varepsilon^{-2}\big)$ Riemannian projected gradient descent steps.

(ii) *(Nonsmooth objectives) Assume each $F_k^{(i)} = f_k^{(i)} + p_k^{(i)}$ is geodesically $L^{(i)}$-Lipschitz continuous, i.e. $\|F_k^{(i)}(x) - F_k^{(i)}(y)\| \leq L^{(i)} d(x, y)$ for any $x, y \in \mathcal{M}^{(i)}$. Then $G_k^{(i)}$ is g-strongly convex with parameter $\lambda_k - L_f$ and geodesically Lipschitz continuous with parameter $L^{(i)} + \lambda_k D$. Approximately solve each proximal sub-problem $\min_{\Theta^{(i)} \subseteq \mathcal{M}^{(i)}} G_k^{(i)}$ by using the Riemannian projected subgradient with $N_k^{(i)}$ sub-iterations and auxiliary iterates (62). Then*

$$\Delta_k^{(i)} \leq \frac{2\zeta(\kappa, D)(L^{(i)} + \lambda_k D)^2}{(\lambda_k - L_f)(N_k^{(i)} + 1)}.$$

Specifically, if we take $N_k^{(i)} = k$, then the algorithm reaches an ε-stationary point within $\widetilde{O}\big((1+\zeta(\kappa, D))^4\, \varepsilon^{-4}\big)$ Riemannian projected subgradient descent steps.

Remark 9 *We compare the above total complexity result with the tangential BMM (tBMM) algorithm in our prior work (Li et al., 2024).*

Since Corollary 8 characterizes the total complexity by analyzing the cost of solving surrogate subproblems, we begin by examining the corresponding subproblems in tBMM. At each iteration, tBMM solves an Euclidean subproblem on a lifted constraint set using Euclidean projected gradient descent (PGD). For smooth objectives (on the tangent space), PGD enjoys a linear convergence rate, leading to a total complexity of $\widetilde{O}(\varepsilon^{-2})$ for tBMM. For nonsmooth objectives, PGD converges sublinearly, and the total complexity becomes $\widetilde{O}(\varepsilon^{-4})$. Thus, in terms of order, RBMM and tBMM exhibit the same $\widetilde{O}(\varepsilon^{-2})$ (g-smooth) and $\widetilde{O}(\varepsilon^{-4})$ (non-g-smooth) iteration-complexity guarantees.

Although the overall orders coincide, the two methods differ substantially in how these costs are incurred. Consider non-g-smooth objectives ($p \neq 0$) as an example. At iteration n for each block, tBMM requires $O(n)$ Euclidean projections while solving the lifted Euclidean subproblem, followed by a single retraction to map the solution back to the manifold. Consequently, tBMM performs in total $\widetilde{O}(\varepsilon^{-4})$ Euclidean projections and $\widetilde{O}(\varepsilon^{-2})$ retractions. In contrast, the particular implementation of RBMM in Corollary 8 applies the Riemannian projected (sub)gradient method of Zhang and Sra (2016) for solving each surrogate minimization. This leads to $O(n)$ exponential maps per iteration and a total of $\widetilde{O}(\varepsilon^{-4})$ exponential maps. Therefore, when Euclidean projections are significantly cheaper than exponential map, tBMM may be computationally favorable. However, tBMM also requires constructing the lifted constraint set at every iteration, which can be costly or infeasible in some cases.

An important advantage of RBMM is its flexibility in choosing subproblem solvers. If a problem-specific solver is available and converges faster than the general RGD method of Zhang and Sra (2016), then the $\widetilde{O}(\varepsilon^{-4})$ bound can be substantially improved in practice. RBMM is particularly efficient when the surrogate subproblems admit exact (closed-form) solutions. We provide several such examples, including the geodesically constrained subspace tracking problem (Section 5.2), MM on Stiefel manifolds (Section 4.5), and the Riemannian CP dictionary learning problem (Section 5.4). In these cases, no subroutine is needed for minimizing the surrogate, making RBMM especially attractive.

Lastly, the geometric dependence of RBMM and tBMM differs in an intrinsic–extrinsic manner consistent with the discussion in Section 1.2. For RBMM, the iteration complexity in Thm. 7 depends only on the g-smoothness constant L_f of the objective and not on any additional geometric regularity of the manifold. When the surrogate subproblems are solved via the RGD method of Zhang and Sra (2016), the cost of approximately minimizing each surrogate depends on the intrinsic quantity $\zeta(\kappa, D)$ in (3), where κ is an upper bound on the sectional curvature and D is the domain diameter.

In contrast, the iteration complexity of tBMM depends on the smoothness of the pull-back objectives on tangent spaces, which is controlled by the retraction approximation constant β in (5), which quantifies the second-order deviation of the retraction from the tangent space. This is an extrinsic *geometric quantity for the underlying manifold and directly affects the majorization parameter and the resulting complexity bound (Li et al., 2024). Since the subproblems of tBMM are Euclidean convex minimization problems, there is no further dependence on the geometry.*

Overall, neither method dominates the other in all settings; rather, each is preferable under different geometric and algorithmic conditions.

Next, the theorem below states similar rates of convergence results for g-smooth surrogates on general manifolds.

Theorem 10 (Rate of convergence for g-smooth surrogates) *Let f denote the objective function in (1) with $m \geq 1$. Let $(\boldsymbol{\theta}_n)_{n\geq 0}$ be an output of Algorithm 1 under (A0) and (A1)**(ii)**, and **(i-gs)**. Assume geodesic convexity of the constraint sets. Further assume that (A1)**(iii)** holds with $\phi(x) = c_\phi x^2$ for some constant $c_\phi > 0$. Then the following hold:*

(i) *(Worst-case rate of convergence) There exists constant $c > 0$ independent of $\boldsymbol{\theta}_0$ such that*

$$\min_{1\leq k\leq n}\left[\sup_{\eta^{(i)}\in T^{\Theta^{(i)}}_{\theta_k^{(i)}}, \|\eta^{(i)}\|\leq 1} V(\boldsymbol{\theta}_k, \boldsymbol{\eta})\right] \leq c\frac{L_g + L_f(1+c_\phi^{-1})(1+\sum_{k=1}^{n}\Delta_k)}{\sqrt{n}/\log n}.$$

See the explicit expression of the RHS constant in (69).

(ii) *(Worst-case iteration complexity) The worst-case iteration complexity N_ϵ for Algorithm 1 satisfies $N_\epsilon = O\left(\varepsilon^{-2}\left(\log\varepsilon^{-1}\right)^2\right)$.*

As we mentioned below (A1), the class of g-smooth surrogates offers flexibility and can be designed in a problem-specific manner. Theorem 10 for the g-smooth surrogates gives a unified treatment for such iteration complexity of RBMM with g-smooth surrogates, and in order to invoke this result, users only need to check the g-smoothness of the surrogates.

Below we give some additional remarks on Theorems 7 and 10. First, the iteration complexity results are not direct adaptations from the Euclidean setting. The complexity of Euclidean BMM for nonconvex objectives is established in Lyu and Li (2025) under the assumption that the constraint sets are convex in the Euclidean sense. However, even the embedded manifolds are in general non-convex when viewed as a constraint set in Euclidean spaces. Therefore, the results in Lyu and Li (2025) do not directly extend to our framework. Instead, our approach utilizes the geometry on the manifold to establish the complexity. It is worth mentioning that our results recover the Euclidean BMM complexity in Lyu and Li (2025) as a special case when the manifolds reduce to Euclidean spaces. Second, the influence of Riemannian geometry is reflected in the constants appearing in the complexity bound. In particular, the g-smoothness parameter L_f depends on the properties of the underlying manifold $\mathcal{M}$ as well as the objective function f. Moreover, with inexact computations allowed, the sum of the optimality gap $\sum_{n=1}^{\infty}\Delta_n(\boldsymbol{\theta}_0)$ also depends on $\mathcal{M}$, as stated in Corollary 8.

Lastly, we state a practical corollary of Theorem 10 for Riemannian optimization problems involving Stiefel manifolds (see Ex. 1). A special case of it was stated in Corollary 1 in the introduction. An important fact about Stiefel manifolds is that, if a function f is (Euclidean) L-smooth in the ambient Euclidean space, then it is g-smooth with respect to the geometry on the Stiefel manifold for some smoothness parameter L' (see Lem. 15).

Therefore, when the underlying manifolds are either Euclidean or Stiefel, we can apply Theorem 10 for Euclidean smooth objectives and surrogates and obtain iteration complexity of $\widetilde{O}(\varepsilon^{-2})$. This expands the applicability of our result to various optimization problems involving Stiefel manifolds.

Corollary 11 (Complexity of RBMM on Stiefel manifolds) *Suppose each underlying manifold $\mathcal{M}^{(i)}$ is either a Stiefel manifold or a Euclidean space. Assume the objective function f in (1) is Euclidean L-smooth for some $L > 0$. Suppose the surrogates $g_n^{(i)}$ are Euclidean L'-smooth for some constants $L' > 0$ and for some constant $c > 0$,*

$$h_n^{(i)}(\theta) := g_n^{(i)}(\theta) - f_n^{(i)}(\theta) \geq c\|\theta - \theta_{n-1}^{(i)}\|^2 \tag{19}$$

*for all $n \geq 1$ and $i = 1, \dots, m$. Assume the constraint sets $\Theta^{(1)}, \dots, \Theta^{(m)}$ are geodesic convex. Allow inexact computation in the sense of (A0)**(ii)** with $\mathrm{d}(\cdot, \cdot)$ in (16) replaced by the Euclidean distance. Then the iterates produced by Algorithm 1 asymptotically converge to the set of stationary points and the algorithm has iteration complexity of $\widetilde{O}(\varepsilon^{-2})$.*

It is important to note that the conditions required to apply Corollary 11 are completely Euclidean, except the g-convexity of constraint sets of Stiefel manifolds, which becomes vacuous when there are no additional constraints on the Stiefel manifolds. The results presented in Corollary 11 are applicable to various MM methods on Stiefel manifolds, as discussed later in Section 4.4, including the recent MM methods investigated in Breloy et al. (2021). Furthermore, in Section 5.2 we study the geodesically constrained subspace tracking problem as a stylized application.

4. RBMM on Specific Manifolds

In this section, we discuss some examples of our general framework of RBMM and its connection to other classical algorithms.

4.1 Examples of manifolds

In this section, we list several examples of manifolds that are typically used in various machine learning problems.

Example 1 (Stiefel Manifold) The Stiefel manifold $\mathcal{V}^{n\times k}$ is the set of all orthonormal k-frames in $\mathbb{R}^n$. That is, it is the set of ordered orthonormal k-tuples ($k \leq n$) of vectors in $\mathbb{R}^n$, i.e.,

$$\mathcal{V}^{n\times k} = \left\{ A \in \mathbb{R}^{n\times k} : A^T A = I_k \right\},$$

where I_k denotes the $k \times k$ identity matrix. For $X \in \mathbb{R}^{n\times k}$, denotes its SVD as $X = U\Sigma V^T$, where $U = [u_1, u_2, \dots, u_n]$ and $V = [v_1, v_2, \dots, v_k]$ are orthogonal matrices. Then the orthogonal projection of X onto $\mathcal{V}^{n\times k}$ is

$$\mathrm{Proj}_{\mathcal{V}^{n\times k}}(X) = UV^T. \tag{20}$$

▲

Example 2 (Fixed-rank Matrices Manifold) The set of matrices with fixed rank-r

$$\mathcal{R}_r = \left\{ X \in \mathbb{R}^{m \times n} : \operatorname{rank}(X) = r \right\} \tag{21}$$

is a smooth submanifold of $\mathbb{R}^{m\times n}$. Denote the SVD of X as $X = U\Sigma V^T$. The diagonal entries of Σ, which are the singular values of X, are written in nonincreasing order,

$$\sigma_1(X) \geq \sigma_2(X) \geq \cdots \geq \sigma_{\min\{n,m\}}(X) \geq 0.$$

Then the orthogonal projection of X onto $\mathcal{R}_r$ is given by

$$\operatorname{Proj}_{\mathcal{R}_r}(X) = \sum_{i=1}^{r} \sigma_i(X) u_i v_i^T := U\Sigma_r V^T.$$

Here $\Sigma_r \in \mathbb{R}^{m\times n}$ is a rectangular diagonal matrix by retaining the r leading diagonal elements of Σ and setting the remaining elements to 0. ▲

Example 3 (Hadamard manifolds) Hadamard manifolds are a class of manifolds that is widely studied in the literature, since it includes many commonly encountered manifolds. *Hadamard manifolds* are Riemannian manifolds with nonpositive sectional curvature that are complete and simply connected, see Burago et al. (2001) and Burago et al. (1992). Hadamard manifolds have infinite injectivity radii at every point.

Below we provide some examples of Hadamard manifolds (more details can be found in e.g., Bacak (2014)).

Example 4 (Euclidean spaces) The Euclidean space $\mathbb{R}^n$ with its usual metric is a Hadamard manifold with constant sectional curvature equal to 0. ▲

Example 5 (Hyperbolic spaces) We equip $\mathbb{R}^{n+1}$ with the $(-1, n)$-inner product

$$\langle x, y \rangle_{(-1,n)} := -x^0 y^0 + \sum_{i=1}^{n} x^i y^i$$

for $x := \left(x^0, x^1, \ldots, x^n\right)$ and $y := \left(y^0, y^1, \ldots, y^n\right)$. Define

$$\mathbb{H}^n := \left\{ x \in \mathbb{R}^{n+1} : \langle x, x \rangle_{(-1,n)} = -1, x_0 > 0 \right\}.$$

Then $\langle \cdot, \cdot \rangle$ induces a Riemannian metric g on the tangent spaces $T_p\mathbb{H}^n \subset T_p\mathbb{R}^{n+1}$ for $p \in \mathbb{H}^n$. The sectional curvature of $(\mathbb{H}^n, g)$ is -1 at every point. ▲

Example 6 (Manifolds of positive definite matrices) The space $\mathbb{S}^n_{++}$ of symmetric positive definite matrices $n \times n$ with real entries is a Hadamard manifold if it is equipped with the Riemannian metric

$$\langle \Omega_1, \Omega_2 \rangle_\Sigma \triangleq \frac{1}{2} \operatorname{Tr}\left(\Omega_1 \Sigma^{-1} \Omega_2 \Sigma^{-1}\right) \quad \forall \Omega_1, \Omega_2 \in T_\Sigma \mathbb{S}^n_{++},$$

where Σ is a $n \times n$ positive definite matrix. ▲

4.2 Euclidean block MM

When specialized on the standard Euclidean manifold, our RBMM becomes the standard Euclidean Block MM (e.g., see BSUM in Hong et al. (2015)), where convergence rate for convex and strongly convex objectives are known. Recently, Lyu and Li (Lyu and Li, 2025, Thm. 2.1) obtained convergence rates for Euclidean Block MM algorithms for nonconvex objectives with convex constraints. Our general results can recover part of their results:

Corollary 12 (Complexity of Euclidean Block MM) *Theorems 4, 5, 7, and 10 hold for Algorithm 1 when the underlying manifolds are Euclidean. In particular, the complexity result in Theorems 7 and 10 hold for the BSUM algorithm in Hong et al. (2015).*

We remark that Lyu and Li (2025, Thm. 2.1) also covers the case when convex surrogates with non-strongly-convex majorization gaps are used along with trust-regions and diminishing radii. Corollary 12 does not cover this case.

Below we give some examples of the Euclidean block MM. One primary example is Euclidean block proximal updates, namely, applying the surrogates in (17) when the underlying manifold is the Euclidean space. The Euclidean block proximal updates read

$$\theta_n^{(i)} \leftarrow \operatorname*{arg\,min}_{\theta\in\Theta^{(i)}} \left(g_n^{(i)}(\theta) = f_n^{(i)}(\theta) + \lambda_n \|\theta - \theta_{n-1}^{(i)}\|^2 \right).$$

Another example of Euclidean block MM is the following block prox-linear update proposed in Xu and Yin (2013): For minimizing a differentiable function f defined on the Euclidean space,

$$\theta_n^{(i)} \leftarrow \operatorname*{arg\,min}_{\theta\in\Theta^{(i)}} \left(g_n^{(i)}(\theta) := f_n^{(i)}(\theta_{n-1}^{(i)}) + \langle \nabla f_n^{(i)}(\theta_{n-1}^{(i)}),\, \theta - \theta_{n-1}^{(i)} \rangle + \frac{\lambda}{2}\|\theta - \theta_{n-1}^{(i)}\|^2 \right). \tag{22}$$

In Xu and Yin (2013), under a mild condition, it was shown that the above algorithm converges asymptotically to a Nash equilibrium (a weaker notion than stationary points) and also a local rate of convergence under the Kurdyka-Łojasiewicz condition is established. Notice that when f is block-wise L-smooth and if $\lambda \geq L$, then $g_n^{(i)}$ is a majorizing surrogate of $f_n^{(i)}$ at $\theta_{n-1}^{(i)}$. Thus (22) is a special instance of our RBMM algorithm in this case and hence Theorems 4 and 10 apply. That is, our general results imply that the block prox-linear algorithm (22) in the Euclidean space converges asymptotically to the stationary points (not only Nash equilibrium) and also has iteration complexity of $\widetilde{O}(\varepsilon^{-2})$.

In fact, the block prox-linear update (22) coincides with block projected gradient descent with a fixed step size. Indeed, denoting $\nabla := \nabla f_n^{(i)}(\theta_{n-1}^{(i)})$, (22) is equivalent to

$$\begin{aligned}\theta_n^{(i)} \leftarrow \operatorname*{arg\,min}_{\theta\in\Theta^{(i)}} \left(\langle \nabla, \theta\rangle + \frac{\lambda}{2}\|\theta\|^2 - \lambda\langle \theta, \theta_{n-1}^{(i)}\rangle \right) &= \operatorname*{arg\,min}_{\theta\in\Theta^{(i)}} \left\| \theta - \left(\theta_{n-1}^{(i)} - \frac{1}{\lambda}\nabla \right) \right\|^2 \\ &= \operatorname{Proj}_{\Theta^{(i)}} \left(\theta_{n-1}^{(i)} - \frac{1}{\lambda}\nabla \right).\end{aligned} \tag{23}$$

Notice that since $\lambda \geq L$, the above becomes the standard *block projected gradient descent* (Block PGD) update with step-size $\in (0, 1/L]$. For block PGD with convex objectives, convergence of function value with complexity $\tilde{O}(\varepsilon^{-1})$ is established in Beck and Tetruashvili (2013). Recently, in Lyu and Li (2025), the authors showed the block PGD for smooth nonconvex objectives converges to the set of stationary points with complexity $\tilde{O}(\varepsilon^{-2})$. Our general results apply to this classical algorithm as well and recover the same complexity as Lyu and Li (2025).

When the constraint set $\Theta^{(i)}$ is a Riemannian manifold embedded in the Euclidean space, in many cases it is non-convex as a subset of Euclidean space. Therefore, when applying the prox-linear updates as in (22) and (23), the complexity result in Theorem 7 and 10 do not hold, while the asymptotic convergence result in Theorem 5 still hold. Moreover, The projection in (23) can be solved exactly on some well-known manifolds, including low-rank manifolds and Stiefel manifolds, which can be found in Section 4.1. More examples and details can be found in Absil and Malick (2012).

Lastly, we provide two stylized applications of the Euclidean block MM in Section 5, namely the robust PCA (Section 5.5) and the Riemannian CP-dictionary learning (Section 5.4).

4.3 Block Riemannian proximal updates on Hadamard manifolds

We consider the Riemannian proximal surrogates in (A1)**(i-rp)** on Hadamard manifolds. As shown later in Section 6.2, this Riemannian proximal surrogate is geodesically strongly convex on Hadamard manifolds. Hence, the subproblems in the minimization step in Algorithm 1 (line 6), i.e., minimizing $g_n^{(i)}$ could be solved efficiently using classical Riemannian optimization methods (Udriste, 1994; Zhang and Sra, 2016; Liu et al., 2017).

Our general results in Theorems 5 and 7 (with Riemannian proximal surrogates) apply to Hadamard manifolds and we obtain the following corollary.

Corollary 13 (Block Riemannian proximal updates on Hadamard manifolds) *Theorems 5 and 7 hold for block Riemannian proximal updates (17) on Hadamard manifolds. That is, the RBMM with proximal surrogates converges asymptotically to the set of stationary points and has iteration complexity of $\widetilde{O}(\varepsilon^{-2})$.*

In Section 5.3, we give a stylized example of block Riemannian proximal updates solving the optimistic likelihood problem, where the manifold of PSD matrices is involved.

4.4 Block Riemannian/Euclidean proximal and prox-linear updates on Stiefel manifolds

In this section, we discuss both the Riemannian and Euclidean proximal/prox-linear updates on the Stiefel manifold, as well as their variants and applications.

It is known that the Stiefel manifold has non-negative sectional curvature (Ziller, 2007), so it is not a Hadamard manifold. Nevertheless, compact manifolds have a positive injectivity radius (Chavel, 2006, Thm. III.2.3). In particular, the Stiefel manifold has an injectivity radius of at least 0.89π (Rentmeesters, 2013, Eq. 5.13), which satisfies (A1)**(ii)**. Hence the results in Theorems 5 and 7 apply and we state the results in the following corollary.

Corollary 14 (Block proximal updates on Stiefel manifolds) *Theorems 5 and 7 hold for both block Riemannian proximal updates (17) and block Euclidean proximal updates (18)* *on Stiefel manifolds. That is, the RBMM with Riemannian/Euclidean proximal surrogates converges asymptotically to the set of stationary points and has iteration complexity of* $\widetilde{O}(\varepsilon^{-2})$.

We remark that in Corollary 14, both Riemannian proximal surrogates and Euclidean proximal surrogates are allowed. The Riemannian proximal surrogates involve the geodesic distance in the updates, which brings additional computational difficulty since the closed-form solution of the end-point geodesic is unknown for the Stiefel manifold. A survey of numerical methods on computing geodesics on the Stiefel manifold can be found in Edelman et al. (1998). Instead, the Euclidean proximal surrogates use the Euclidean distance function as a regularizer, which provides computational savings.

In fact, our general framework of RBMM can utilize standard (Euclidean) L-smooth surrogates for block optimization problems involving Stiefel manifolds, as stated in Corollary 11. Below we first give a key lemma from Chen et al. (2021) for deriving Corollary 11 from Theorems 5 and 10. This lemma states any L-smooth function in Euclidean space is a g-smooth function on the Stiefel manifold, and therefore an L-smooth objective function and surrogates satisfy the g-smoothness assumption in (A0) and (A1).

Lemma 15 (Smoothness on Stiefel manifolds; Lem. 2.4 in Chen et al. (2021)) *If f is L-smooth in Euclidean space $\mathbb{R}^{n\times d}$, then there exists a constant $L_g = L + L_N$ such that f is g-smooth with parameter L_g on the Stiefel manifold $\mathcal{V}^{n\times d}$, where $L_N = \max_{x\in\mathcal{V}^{n\times d}} \|\nabla f(x)\|$.*

In order for the iteration complexity results in Corollary 11 to hold, one also needs to verify assumption (A1)**(iii)**. Namely, the majorization gap should be lower bounded by an increasing function of geodesic distance. The following geometric lemma states that for compact embedded submanifolds of the Euclidean space, the Riemannian and the Euclidean distances are within constant multiples of each other. Therefore, in this case, the quadratic majorization gap in terms of the Euclidean distance in (19) implies the quadratic majorization gap in terms of the Riemannian distance in (A1)**(iii)**. This allows us to use a wide range of computationally efficient surrogates on Stiefel manifolds such as Euclidean proximal surrogates (18), Euclidean prox-linear surrogates (22), and *Euclidean regularized linear surrogates*, which will be discussed later in this section.

See Figure 3 for an illustration of Lemma 16 on Stiefel manifolds.

Lemma 16 (Equivalence of distance on compact sets; Lem. 4.1 in Michels (2019)) *Let $M \subset \mathbb{R}^n$ be a smooth submanifold, equipped with a Riemannian metric g. The geodesic distance between $x, y \in M$ induced by g is denoted as $d(x,y)$. Consider the Euclidean norm $\|\cdot\|$ on $\mathbb{R}^n$. Let $K \subseteq M$ be compact. Then there exists $c > 0$ such that for $x, y \in K$,*

$$cd(x,y) \le \|x-y\| \le d(x,y).$$

Below we give a proof of Corollary 11. Recall that Corollary 1 is a direct consequence of Corollary 11.

Proof of Corollary 11. To deduce Corollary 11 from Theorems 5 and 10, we verify the assumptions one by one. First, by the hypothesis on Euclidean smoothness of the objective f and the surrogates $g_n^{(i)}$ and Lemma 15 (and also using the definition of d on the product manifold in (10)), we have that f and $g_n^{(i)}$s are also g-smooth. This verifies (A0)**(i)**

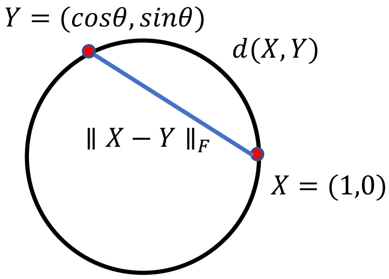


Figure 3: Illustration of distances on $\mathcal{V}^{n\times p}$ with $n = 2, p = 1$. The length of the blue segment is Euclidean distance $\|X - Y\|_F$. The length of the arc between X and Y is $d(X, Y)$.

and (A1)**(i-gs)**. By Lemma 16 and the hypothesis in Corollary 11, we verify (A0)**(ii)** and (A1)**(iii)**. Lastly, (A1)**(ii)** is satisfied since both Stiefel manifolds and Euclidean space have a uniformly positive injectivity radius. Then the assertion in Corollary 11 follows from Theorems 5 and 10. ■

4.5 Iteration complexity of MM on Stiefel manifolds (Breloy et al., 2021)

In order to solve minimization over Stiefel manifolds

$$\min_{U\in\mathcal{V}^{p\times k}} f(U), \tag{24}$$

Breloy et al. (2021) recently proposed an MM method incorporating linear surrogates and reducing the surrogate minimization problem to a projection onto the Stiefel manifold. More precisely, given an iterate $U_{n-1} \in \mathcal{V}^{p\times k}$ at iteration $n-1$, the surrogate g_n at iteration n takes the following form

$$g_n(U) = -\operatorname{Tr}\left(\mathbf{R}^H(U_{n-1})U\right) - \operatorname{Tr}\left(U^H\mathbf{R}(U_{n-1})\right) + \text{const}, \tag{25}$$

where ^{H} is the conjugate transpose operator and $\mathbf{R}(\cdot) : \mathbb{C}^{p\times k} \to \mathbb{C}^{p\times k}$ is a matrix function chosen so that $g_n(U_{n-1}) = f(U_{n-1})$ and $g_n \geq f$. Since g_n is linear and since $U \in \mathcal{V}^{p\times k}$, we have

$$U_{n+1} \overset{\text{def}}{=} \operatorname*{arg\,min}_{U\in\mathcal{V}^{p\times k}} g_n(U) = \operatorname*{arg\,min}_{U\in\mathcal{V}^{p\times k}} \|\mathbf{R}^H(U_{n-1}) - U\|_F.$$

When $\mathbf{R}^H(U_{n-1})$ is of full-rank, then the rightmost projection problem has a unique solution given by the projection operator in (20). See various applications of this MM method in Breloy et al. (2021).

In Breloy et al. (2021), the authors showed this MM method asymptotically converges to the set of stationary points by adapting a convergence result of Euclidean BMM algorithm in Razaviyayn et al. (2013). However, there is no known bound on the iteration complexity due to the non-convexity of the constraint set and the objective function. An iteration complexity of $\tilde{O}(\varepsilon^{-2})$ of Euclidean BMM for non-convex smooth objectives with convex constraints has been obtained recently in Lyu and Li (2025). However, this result is not applicable here since the Stiefel manifolds cannot be viewed as convex constraint sets within Euclidean spaces.

By applying Corollary 11, we can obtain an iteration complexity bound of $\tilde{O}(\varepsilon^{-2})$ for the above MM method on the Stiefel manifold with a slight modification via Euclidean proximal regularization. Namely, for a fixed proximal regularization parameter $\lambda \geq 0$, consider the following surrogate

$$\begin{aligned}\tilde{g}_n(U) &= g_n(U) + \lambda \|U - U_{n-1}\|_F^2 \\ &= -\operatorname{Tr}\left((\mathbf{R}(U_{n-1}) + \lambda U_{n-1})^H U\right) - \operatorname{Tr}\left(U^H(\mathbf{R}(U_{n-1}) + \lambda U_{n-1})\right) + \text{const.}\end{aligned} \tag{26}$$

Note that minimizing the above surrogate $\tilde{g}_n$ is as easy as minimizing g_n:

$$U_{n+1} \overset{\text{def}}{=} \underset{U \in \mathcal{V}^{p\times k}}{\arg\min}\, \tilde{g}_n(U) = \underset{U \in \mathcal{V}^{p\times k}}{\arg\min}\, \|\mathbf{R}^H(U_{n-1}) + \lambda U_{n-1} - U\|_F. \tag{27}$$

Note that the MM update in (27) satisfies the hypothesis of Corollary 11. Indeed, the surrogates g_n in (25) and $\tilde{g}_n$ in (26) are linear and hence are L-smooth in Euclidean space. The addition of Euclidean proximal regularization ensures that we have at least a quadratic majorization gap as in (19). Therefore, by Corollary 11, the iteration complexity of the MM update in (27) is of $\tilde{O}(\varepsilon^{-2})$. These results are formally stated in Corollary 17.

Corollary 17 (MM with regularized linear surrogates on Stiefel manifolds) *Consider the problem of minimizing a differentiable objective f on the Stiefel manifold $\mathcal{V}^{p\times k}$ as in* (24)*. Then the iterates generated by* (27) *with arbitrary initialization converges asymptotically to the set of stationary points of* (24)*. Moreover, the iteration complexity is $\widetilde{O}(\varepsilon^{-2})$.*

To the best of our knowledge, this is the first complexity result for MM on Stiefel manifolds in the literature. In particular, the iteration complexity in Corollary 17 applies for various problem instances discussed in Breloy et al. (2021, Sec. V) including power iteration for computing top eigenvector, generic non-homogenous quadratic form, and nonconvex subspace recovery. In Section 5.2, we use a similar idea to obtain the same iteration complexity bound for block MM algorithm for the geodesically constrained subspace tracking problem (Blocker et al., 2023).

5. Applications

In this section, we will discuss the following five applications of our general framework:

1. Variational inference under Wasserstein geometry (Lambert et al., 2022);
2. Geodesically constrained subspace tracking (Blocker et al., 2023);
3. Optimistic likelihood under Fisher-Rao distance (Nguyen et al., 2019);
4. Riemannian CP-dictionary-learning (Lyu et al., 2022; Dong et al., 2022);
5. Robust PCA (Candes et al., 2009; Rodriguez and Wohlberg, 2013).

The first problem above is an application of Riemannian proximal updates for a constrained nonsmooth nonconvex Riemannian optimization problem in the space of probability measures with 2-Wasserstein distance, and we are able to derive a new iteration complexity result (Cor. (18)) without assuming the strongly g-convexity of the objective function. The second problem is an application of block Euclidean proximal updates on Stiefel manifolds, which verifies our assumptions for Theorem 10 and we are able to derive a new iteration complexity result (Cor. 20), as mentioned in Corollary 11 and Section 4.4. The third problem above is an application of Riemannian proximal updates on Hadamard manifolds (see Section 4.3), which verifies our assumptions for Theorem 7 and we are able to derive a new iteration complexity result (Cor. 21). For the last two problems above, we only prove asymptotic convergence to stationary points by using Theorem 5. These two problems are formulated as minimizing a cost function involving the Euclidean distance function (e.g., the matrix Frobenius norm) over low-rank manifolds. The Euclidean distance function is not g-smooth over low-rank manifolds (see Appendix C.1 for details). Therefore, in order to satisfy the assumption of g-smoothness of the objective in (A0)**(i)**, we choose Euclidean geometry as the underlying manifold structure, and take the embedded submanifolds as constraints. In this case, the constraints are not g-convex with respect to the underlying Euclidean geometry, so we are not able to apply our iteration complexity results (Theorems 7 and 10). However, we can still deduce asymptotic convergence to stationary points by using Theorem 5 (see Section 4.2), which fortunately does not require g-convexity of the constraint sets (see Remark 6).

5.1 Variational inference under Wasserstein geometry (Lambert et al., 2022)

Given probability measures u, v on $\mathbb{R}^d$, 2-Wasserstein distance is given by

$$W_2(u,v) = \left[\inf_{\gamma \in C(u,v)} \int \|x-y\|^2 d\gamma(x,y) \right]^{1/2},$$

where $C(u,v)$ is the joint distribution on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginal distribution is u and v respectively. When $u, v \in \mathcal{P}_2(\mathbb{R}^d)$, which is the space of probability measure with finite second moment, we have $W_2(u,v)$ to be finite. It is shown in Otto (2001) that $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is a Riemannian manifold.

The *Variational Inference* (VI) aims to approximate a probability measure $\pi \propto \exp(-V)$ for a potential function $V : \mathbb{R}^d \to \mathbb{R}$ by minimizing the KL-divergence within a constraint set C:

$$\pi^*_C := \operatorname*{arg\,min}_{u \in C} \left(\mathrm{KL}(u \,\|\, \pi) = \underbrace{\int V du}_{=:\mathcal{V}(u)} + \underbrace{\int \log(u) du}_{=:\mathcal{H}(u)} + \mathrm{const} \right). \tag{28}$$

For instance, C may be taken as $\mathrm{BW}(\mathbb{R}^d) \subseteq \mathcal{P}(\mathbb{R}^d)$, which consists of Gaussian distributions and is known as Bures–Wasserstein manifold. The standard optimization procedure to solve (28) is the *Bures–JKO* scheme (Lambert et al., 2022; Jordan et al., 1998; Bures, 1969), which is an iterative process of minimizing the following Riemannian proximal surrogates:

$$u_{n+1} = \operatorname*{arg\,min}_{u \in \mathrm{BW}(\mathbb{R}^d)} \left(G_{n+1}(u) := \mathcal{V}(u) + \mathcal{H}(v) + \frac{\lambda_n}{2} W_2^2(u, u_n) \right). \tag{29}$$

A continuous variant of the above proximal point algorithm is known as the Wasserstein gradient flow, which has been extensively studied in the literature of VI. The existing analysis of Wasserstein gradient flow assumes that the potential function V is strongly convex, which warrants the integral $\mathcal{V}(\cdot)$ is strongly g-convex with respect to the Wassertein geometry (Lambert et al., 2022; Jiang et al., 2023). However, to our best knowledge, there is no iteration complexity result available in the literature for the original Bures-JKO scheme (29), especially when the potential function V is not necessarily strongly convex. In the following corollary, we provide an iteration complexity result without assuming the strong convexity.

Corollary 18 (Complexity of Bures-JKO scheme for Wasserstein VI) *Assume the potential function V is L-smooth over $\mathbb{R}^d$ (but not necessarily strongly convex) and the constraint set C is a g-convex subset of $\mathcal{P}(\mathbb{R}^d)$. Then the Bures-JKO scheme (29) converges asymptotically to the set of stationary points with iteration complexity of $O(\varepsilon^{-2})$.*

As we have noted above, the Bures-JKO scheme (29) can be viewed as a single-block RBMM with Riemannian proximal surrogates. We have established iteration complexity of the class of such algorithms in Theorem 7. Essentially what we need to verify is the g-smoothness of $\mathcal{V}(\cdot)$ and g-convexity of $\mathcal{H}(\cdot)$. The former is guaranteed if the potential function V is L-smooth (see Lem. 19), and the latter is a general fact about the entropy function. We provide further details of this derivation below.

Lemma 19 (Lambert et al. (2022)) *The following holds for $\mathcal{V}(u)$ and $\mathcal{H}(u)$,*

1. *$\mathcal{H}(u)$ is g-convex over $\mathcal{P}_2(\mathbb{R}^d)$.*
2. *$\mathcal{H}(u)$ is non-smooth over $\mathcal{P}_2(\mathbb{R}^d)$.*
3. *$\mathcal{V}(u)$ is g-smooth over $\mathcal{P}_2(\mathbb{R}^d)$ with parameter L if V is L-smooth over $\mathbb{R}^d$.*
4. *$\mathcal{V}(u)$ is strongly g-convex over $\mathcal{P}_2(\mathbb{R}^d)$ with parameter α if V is α-strongly convex over $\mathbb{R}^d$.*

Now we give a concise proof of Cor. 18.

Proof of Cor.18 Under the assumptions in Cor. 18, first note that Lemma 19 shows the objective in (28) satisfies the main assumptions in (A0). Namely, $\mathcal{V}(u)$ is g-smooth and $\mathcal{H}(u)$ is g-convex. Moreover, (A1)**(i-rp)** and **(iii)** are satisfied by the choice of surrogates in (29). The rest of the assumptions are trivially satisfied and we omit here. Hence, the complexity result in Thm. 7 holds. For asymptotic convergence, further notice that $\mathcal{H}(u)$ is lower semi-continuous over $\mathcal{P}(\mathbb{R}^d)$. So we can conclude. ■

We give some remark of the complexity results in Cor. 18. In fact, the surrogates we can apply for solving (28) is not limited to the Riemannian proximal surrogates as in (29). Both g-smooth surrogates and Euclidean proximal surrogates can be applied when the corresponding assumptions are satisfied, and we are able to deduce the same complexity for such algorithm from Thm. 7 and Thm. 10 similarly as above.

5.2 Geodesically constrained subspace tracking (Blocker et al., 2023)

Let $X_i \in \mathbb{R}^{d\times l}$ for $i = 1, \cdots, T$ be data generated from a low-rank model with noise,

$$X_i = U_i G_i + N_i,$$

where $U_i \in \mathbb{R}^{d\times k}$ has orthonormal columns representing a point on the Grassmannian $\mathcal{G}(k,d)$, the space of all rank-k subspaces in $\mathbb{R}^d$. More precisely, Grassmannian can be represented as a quotient space $\mathcal{G}(k,d) = \mathcal{V}^{d\times k}/O(k)$, where $\mathcal{V}^{d\times k}$ is the Stiefel manifold (see Example 1) and $O(k)$ is the orthogonal group of $k\times k$ matrices. $G_i \in \mathbb{R}^{k\times \ell}$ holds weight or loading vectors; and $N_i \in \mathbb{R}^{d\times \ell}$ is an independent additive noise matrix.

For the geodesic subspace tracking problem, we observe X_i and our objective is to estimate U_i for $i = 1, \cdots, T$. We model each U_i as an orthonormal basis whose span is sampled from a single continuous Grassmanian geodesic $U(t) : [0,1] \to \mathcal{V}^{d\times k}$, parameterized as follows: For $H, Y \in \mathcal{V}^{d\times k}$ consisting of orthogonal columns, i.e., $H^T Y = O$,

$$U_i = U(t_i) = H\cos(\Theta t_i) + Y \sin(\Theta t_i).$$

Here $\Theta \in \mathbb{R}^{k\times k}$ is a diagonal matrix where its jth diagonal entry, θ_j, is the j-th principal angle between the two endpoints of the geodesic. We assume either the time-points t_i are given, or the observed matrices X_i are equidistant along a geodesic curve.

The objective function is formulated as follows,

$$f(U) = f(H, Y, \Theta) = \min_{\{G_i\}_{i=1}^T} \|X_i - U(t_i)G_i\|_F^2 = -\sum_{i=1}^T \|X_i^T U(t_i)\|_F^2 + c, \tag{30}$$

where for the last equality we have substituted the optimal $G_i = U(t_i)^T X_i$ and c is a constant (see Golub and Perayra (2003)).

To approximately minimize (30), Blocker et al. proposed a two-block MM approach in Blocker et al. (2023), where one alternatively optimizes two block parameters $Q := (H, Y) \in \mathcal{V}^{d\times 2k}$ and Θ. Here we propose a proximal regularized version of the BMM method in Blocker et al. (2023) and establish its asymptotic convergence property in Corollary 20. For proximal regularization parameters $\lambda_Q, \lambda_\Theta \geq 0$, the proposed algorithm reads as

$$Q_{n+1} \leftarrow WV^T, \text{ where } W\Sigma V^T \text{ is the SVD of } \lambda_Q Q_n - \nabla_Q f(Q_n, \Theta_n) \tag{31}$$

$$(\theta_j)_{n+1} \leftarrow (\theta_j)_n - \frac{1}{w_j((\theta_j)_n) + \lambda_\Theta} \nabla f_{n+1,j}^{(2)}((\theta_j)_n) \text{ for } j = 1, \ldots, k, \tag{32}$$

where $(\theta_j)_n$ denotes the jth diagonal entry of the $k\times k$ diagonal matrix Θ_n at iteration n, $w_j = \sum_{i=1}^T w_{f_{i,j}}$ is the "weighting function" defined in Blocker et al. (2023), and

$$f_{n+1,j}^{(2)}(\theta_j) := -\sum_{i=1}^T (r_{i,j})_{n+1} \cos(2\theta_j t_i - (\phi_{i,j})_{n+1}) + (b_{i,j})_{n+1}. \tag{33}$$

The definition of the parameters $\phi_{i,j}$, $r_{i,j}$ and $b_{i,j}$ in (33) can be found in Blocker et al. (2023), which we also provide in Appendix C.2 for completeness. Note that when $\lambda_Q = 0$, (31) becomes the updates for Q in Blocker et al. (2023). Similarly, $\lambda_\Theta = 0$ gives the updates for Θ in Blocker et al. (2023). Thus our algorithm (31)-(32) generalizes the BMM algorithm proposed in Blocker et al. (2023).

We first argue why we can cast the above algorithm as RBMM. The discussion we provide here is a minor modification of the derivation in Blocker et al. (2023). First, we discuss the Q-update (31). Let $Z_i = [\cos(\Theta t_i)\ ;\ \sin(\Theta t_i)]$, which is the vertical concatenation of $\cos(\Theta t_i)$ and $\sin(\Theta t_i)$. The objective function can be rewritten as

$$f(Q,\Theta) = -\sum_{i=1}^{T} \|X_i^T Q Z_i\|_F^2 \tag{34}$$

and the gradient with respect to the first block Q is given by $\nabla_Q f = -2\sum_{i=1}^T X_i X_i^T Q Z_i Z_i^T$. The marginal objective function for updating Q can be rewritten as

$$f_{n+1}^{(1)}(Q) := -\left\langle \sum_{i=1}^{T} X_i X_i^T Q (Z_i)_n (Z_i)_n^T, Q \right\rangle,$$

which is a concave-up quadratic function in Q. Also note that $f_{n+1}^{(1)}$ is L-smooth for some constant $L > 0$ over the compact parameter space.

We consider the following proximal majorizer of $f_{n+1}^{(1)}$: For $\lambda \geq 0$,

$$\begin{aligned} g_{n+1}^{(1)}(Q) &:= -\left\langle \sum_{i=1}^{T} X_i X_i^T Q_n (Z_i)_n (Z_i)_n^T, Q \right\rangle + \frac{\lambda}{4}\|Q - Q_n\|_F^2 \\ &= \frac{1}{2}\langle \nabla_Q f(Q_n,\Theta_n), Q\rangle + \frac{\lambda}{4}\|Q - Q_n\|_F^2. \end{aligned}$$

We claim that

$$Q_{n+1} = \operatorname*{arg\,min}_{Q\in\mathcal{V}^{d\times 2k}} g_{n+1}^{(1)}(Q) = \begin{cases} \mathrm{Proj}_{\mathcal{V}^{d\times 2k}}\left(Q_n - \frac{1}{\lambda}\nabla_Q f(Q_n,\Theta_n)\right) & \text{if } \lambda > 0 \\ \mathrm{Proj}_{\mathcal{V}^{d\times 2k}}\left(-\nabla_Q f(Q_n,\Theta_n)\right) & \text{if } \lambda = 0. \end{cases} \tag{35}$$

Indeed, the above MM update for $\lambda > 0$ follows from (23). For $\lambda = 0$, we use the fact that $Q^T Q = I$ for all Q in the Stiefel manifold so that

$$\operatorname*{arg\,min}_{Q\in\mathcal{V}^{d\times 2k}} \langle \nabla_Q f(Q_n,\Theta_n), Q\rangle = \operatorname*{arg\,min}_{Q\in\mathcal{V}^{d\times 2k}} \|Q + \nabla_Q f(Q_n,\Theta_n)\|_F^2 = \mathrm{Proj}_{\mathcal{V}^{d\times 2k}}\left(-\nabla_Q f(Q_n,\Theta_n)\right).$$

Notice that projecting onto the Stiefel manifold can be easily done by SVD (see, e.g., Absil and Malick (2012)). Hence (35) coincides with (31).

Next, we discuss the Θ-update. For conciseness, we only put the expression of the loss function and surrogates here. For full details, we refer the readers to Blocker et al. (2023). The marginal loss function for Θ is separable for each diagonal element θ_j of Θ. Consider the following proximal majorizer of $f_{n+1,j}^{(2)}$ in (33): For $\lambda_\Theta \geq 0$,

$$g_{n+1,j}^{(2)}(\theta_j) = f_{n+1,j}^{(2)}\left((\theta_j)_n\right) + \nabla f_{n+1,j}^{(2)}\left((\theta_j)_n\right)\left(\theta_j - (\theta_j)_n\right) + \frac{w_j((\theta_j)_n) + \lambda_\Theta}{2}\left(\theta_j - (\theta_j)_n\right)^2$$

where $(\theta_j)_n$ is the value of θ_j at iteration n. Then by using (23), we see that (32) coincides with minimizing the majorizing surrogate $g^{(2)}_{n+1,j}$ above.

Now we discuss the convergence of this block MM algorithm. In fact, we can apply Corollary 11 and directly get the convergence and complexity results. The asymptotic convergence result in Corollary 11 for the updates (31)-(32) with $\lambda_Q, \lambda_\Theta \geq 0$. Note when $\lambda_Q = \lambda_\Theta = 0$, the updates (31)-(32) become the vanilla block MM method in Blocker et al. (2023). Moreover, when the proximal parameter $\lambda_Q, \lambda_\Theta$ are strictly positive, the complexity result in Corollary 11 holds. We state the convergence and complexity results of this proximal regularized MM method in the following corollary:

Corollary 20 (Complexity of regularized BMM for geodesic subspace tracking) *Given a sequence of data X_i for $i = 1, \cdots, T$. Let $\boldsymbol{\theta}_k = (Q_k, \Theta_k)$ be generated by (31)-(32) with arbitrary initialization $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} = \mathcal{V}^{d\times 2k} \times \mathbb{R}^{k\times k}$. Suppose the proximal parameters λ_Q and λ_Θ are non-negative. Then the limit points of $(\boldsymbol{\theta}_n)_{n\geq 0}$ are stationary points of problem (34). Moreover, if $\lambda_Q, \lambda_\Theta > 0$, then the iteration complexity is $\widetilde{O}(\varepsilon^{-2})$.*

Proof In fact, in order to apply Corollary 11, we only need to verify the Euclidean smoothness of the marginal loss function. Note

$$\nabla_Q f = -2\sum_{i=1}^{T} X_i X_i^T Q Z_i Z_i^T, \qquad \nabla f^{(2)}_{n+1,j}(\theta_j) = 2\sum_{i=1}^{T} (r_{i,j})_{n+1} t_i \sin\left(2\theta_j t_i - (\phi_{i,j})_{n+1}\right),$$

so f is block-wise Euclidean L-smooth. Hence, the convergence and complexity results in Corollary 11 follow. ∎

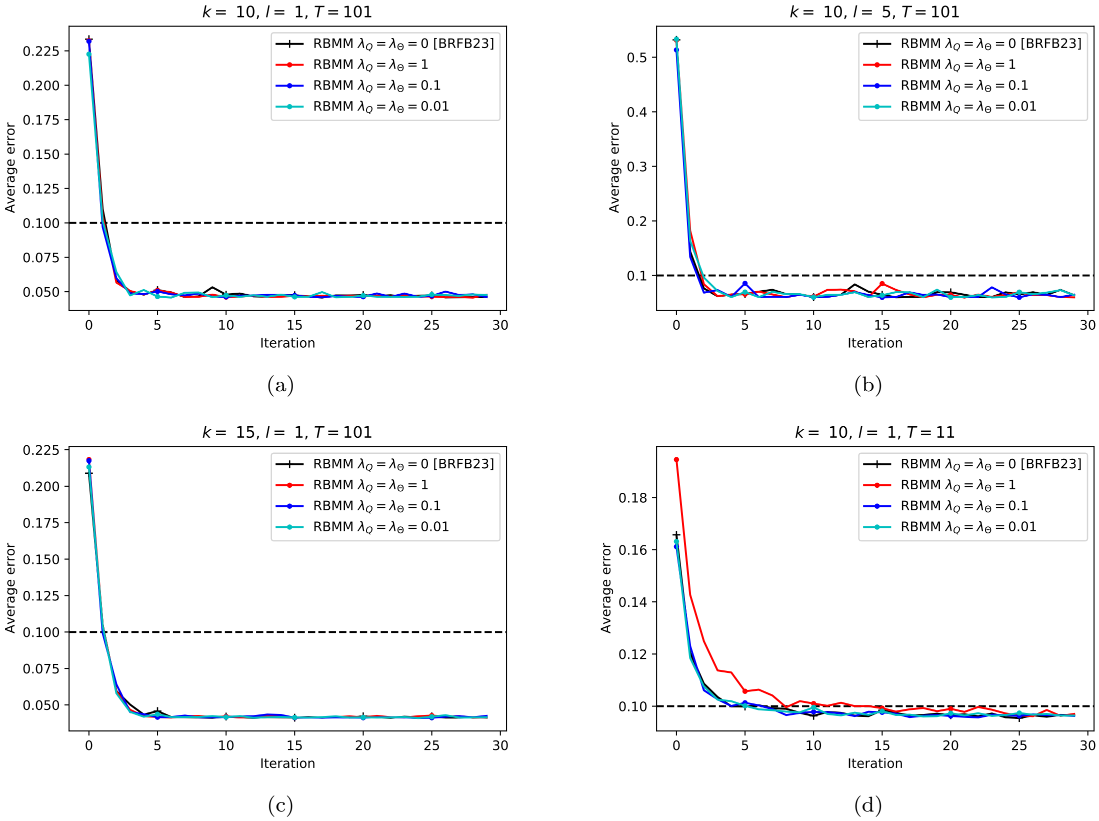
Now we compare our results with other existing works. We remark that the BMM algorithm in Blocker et al. (2023) is based on the MM algorithm on Stiefel manifold by Breloy et al. (2021). As mentioned in Section 4.4, the asymptotic convergence to the set of all stationary points in Breloy et al. (2021) was established by adopting a convergence result of Euclidean BMM algorithm in Razaviyayn et al. (2013). The complexity results are still unknown due to the non-convexity of the constraint set. Here, we give the first complexity result in the literature. Moreover, our method is computationally efficient with close-form updates as shown in (31)-(32).

Next, we discuss the results of numerical experiments. In the case of synthetic data, we have access to the true geodesic, so we could compare the estimated geodesic $\hat{U}(t)$ with the true geodesic $U(t)$. The following error metric is used, which is the square root of the average squared subspace error between corresponding points along the geodesic,

$$\text{Geodesic Error } = \sqrt{\int_0^1 \frac{1}{2k}\left\|\hat{U}(t)\hat{U}(t)^T - U(t)U(t)^T\right\|_{\mathrm{F}}^2 \,\mathrm{d}t}. \tag{36}$$

In practice, we approximate the integral by Riemann sum using the sample points. The geodesic error (36) takes minimum value of 0 when $\operatorname{span}(\hat{U}) = \operatorname{span}(U)$ and maximum value of 1 when $\operatorname{span}(\hat{U}) \perp \operatorname{span}(U)$.

In the numerical experiments, we set the dimension $d = 30$ and set the elements of the noise matrix N_i to be independent Gaussian noise with standard deviation $\sigma = 0.1$. The



(a)

(b)

(c)

(d)

Figure 4: Convergence of RBMM in geodesic error under different settings. Average geodesic error is computed over 50 independent trials. The dimension is $d = 30$ and the additive Gaussian noise has a standard deviation $\sigma = 0.1$. The value of other parameters is shown in the title for each panel. Adding Euclidean proximal regularizer does not accelerate the convergence, mainly due to the orthogonal property on Stiefel manifolds. Namely, any point Q on the Stiefel manifold satisfies $Q^T Q = \mathbf{I}$. Hence the proximal regularizer degenerates into a linear term providing with limited acceleration.

other parameters including the rank k, number of samples l, number of time points T, and proximal parameters λ_Q and λ_Θ are set differently for different experiments. We run each experiment 50 times independently to compute the average geodesic error. As shown in Figure 4, RBMM with different λ converges under different settings and is as good in terms of convergence speed as the block MM method in Blocker et al. (2023). Furthermore, Figure 4 shows that changing the proximal parameter λ_Q and λ_Θ only slightly affects the rate of convergence, this is due to the following two reasons: first, there are only two blocks for this problem, which gives a relatively simple setting for RBMM; second, note $Q \in \mathcal{V}^{d\times 2k}$ which gives $Q^TQ = I_{2k}$, therefore the quadratic term added would reduce to a linear term, which does not significantly accelerate convergence. For more numerical results of this geodesic subspace tracking problem, we refer the reader to Blocker et al. (2023).

5.3 Optimistic likelihood under Fisher-Rao distance (Nguyen et al., 2019)

Let C denote the finite set of classes, $C = \{1, 2, \ldots, |C|\}$. Each class $c \in C$ is associated with a Gaussian distribution $\mathbb{P}_c = \mathcal{N}(\mu_c, \Sigma_c)$, where $\mu_c \in \mathbb{R}^N$ is the mean vector for class c and the positive definite matrix $\Sigma_c \in \mathbb{R}^{N\times N}$ is the covariance matrix for class c. Consider a set of i.i.d data points $x^M \triangleq x_1, \ldots, x_M \in \mathbb{R}^N$ that are generated from one of $\mathbb{P}_c$. We want to determine the distribution $\mathbb{P}_{c^*}$, $c^* \in C$, under which the following log-likelihood function $\ell(x^M, \mathbb{P}_c)$ is maximized, i.e., we want to solve

$$c^* = \arg\max_{c\in C} \left\{ \ell\left(x^M, \mathbb{P}_c\right) \triangleq -\frac{1}{M}\sum_{m=1}^{M} \left(x_m - \mu_c\right)^T \Sigma_c^{-1} \left(x_m - \mu_c\right) - \log\det\Sigma_c \right\},$$

where μ_c and Σ_c denote the means and covariance matrices of $\mathbb{P}_c$. Since methods that sample points to get good estimators of μ_c and Σ_c are usually costly, we consider the following *optimistic likelihood problem* instead. Namely, instead of aiming to find the Gaussian distribution from the set, we look for a Gaussian distribution that maximizes the likelihood close to the empirical distribution under some distance measure. More precisely, we consider the following optimistic likelihood problem

$$\max_{\mathbb{P}\in\mathcal{P}_c} \ell\left(x^M, \mathbb{P}\right) \text{ with } \mathcal{P}_c = \left\{\mathbb{P} \in \mathcal{P} : \varphi\left(\hat{\mathbb{P}}_c, \mathbb{P}\right) \le \rho_c\right\},$$

where $\mathcal{P}$ is the set of all non-degenerate Gaussian distributions on $\mathbb{R}^N$, $\hat{\mathbb{P}}_c$ is the empirical distribution estimated from training data, φ is the Fisher-Rao distance, and $\rho_c \in \mathbb{R}_+$ are the radii of the ambiguity sets $\mathcal{P}_c$. For conciseness, we put the details of the Fisher-Rao distance in Appendix C.3. We refer the readers to Nguyen et al. (2019) for full details.

Now denoting the empirical mean and covariance from the data by $\hat{\mu}$ and $\hat{\Sigma}$, we can explicitly write down the optimization problem as

$$\min_{\mu,\Sigma} f(\mu, \Sigma) \triangleq \left\langle M^{-1}\sum_{m=1}^{M} \left(x_m - \mu\right)\left(x_m - \mu\right)^T, \Sigma^{-1} \right\rangle + \log\det\Sigma, \tag{37}$$

where $\mu \in \Theta^{(1)} = \left\{\mu \in \mathbb{R}^N : (\mu - \hat{\mu})^T(\mu - \hat{\mu}) \le \rho_1^2\right\}$ and $\Sigma \in \Theta^{(2)} = \left\{\Sigma \in \mathbb{S}_{++}^N : d(\Sigma, \hat{\Sigma}) \le \rho_2\right\}$. Here $d(\Sigma, \hat{\Sigma})$ is the Fisher-Rao distance of two Gaussian distributions with identical mean (see Appendix C.3).

Denote μ_n and Σ_n as the $\theta_n^{(1)}$ and $\theta_n^{(2)}$ generated by Algorithm 1 applied on $f(\mu, \Sigma)$. The marginal objective functions are denoted by

$$f_n^{(1)} := f(\mu, \Sigma_{n-1}) = \left\langle M^{-1} \sum_{m=1}^{M} (x_m - \mu)(x_m - \mu)^T, \Sigma_{n-1}^{-1} \right\rangle + \log\det \Sigma_{n-1}$$

$$f_n^{(2)} := f(\mu_n, \Sigma) = \langle S_n, \Sigma^{-1} \rangle + \log\det \Sigma \qquad \text{where} \qquad S_n = M^{-1} \sum_{m=1}^{M} (x_m - \mu_n)(x_m - \mu_n)^T .$$

As explained in Example 6, this manifold of positive definite matrices is a Hadamard manifold. Thus we could construct Riemannian proximal surrogate functions as in (17), more discussions can be found in Section 6.2, i.e.,

$$\mu_n = \operatorname*{arg\,min}_{\mu \in \mathbb{R}^N} g_n^{(1)}(\mu) := \left\langle M^{-1} \sum_{m=1}^{M} (x_m - \mu)(x_m - \mu)^T, \Sigma_{n-1}^{-1} \right\rangle + \log\det \Sigma_{n-1} + \frac{\lambda}{2} \|\mu - \mu_{n-1}\|^2$$

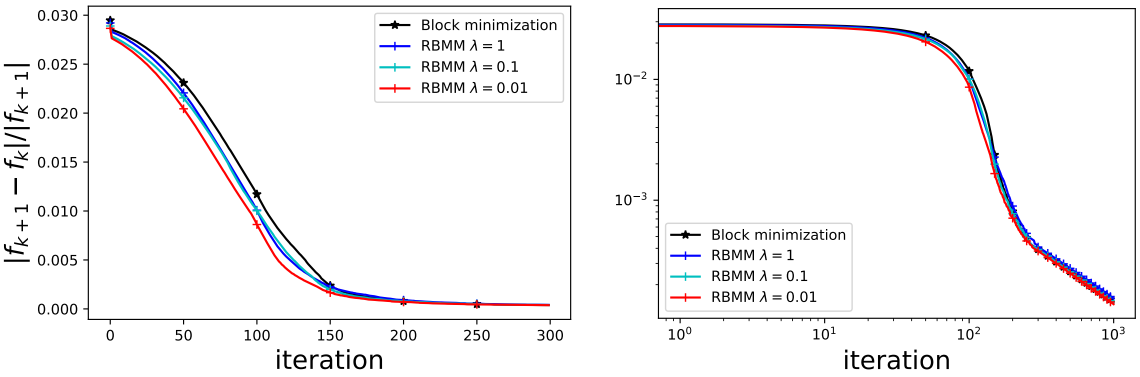
$$\Sigma_n = \operatorname*{arg\,min}_{\Sigma \in \mathbb{S}_{++}^N} g_n^{(2)}(\Sigma) := \langle S_n, \Sigma^{-1} \rangle + \log\det \Sigma + \frac{\lambda}{4} \left\| \log\left(\Sigma_{n-1}^{-\frac{1}{2}} \Sigma \Sigma_{n-1}^{-\frac{1}{2}} \right) \right\|_F^2 . \tag{38}$$

Note this block MM is a special instance of proximal methods on the Hadamard manifold discussed in Section 4.3. Hence Theorems 5 and 7 apply. We state it as a corollary as follows.

Corollary 21 (Complexity of Riemannian proximal updates for optimistic likelihood) *Given a set of data points $(x_i)_i$ for $i = 1, \cdots, M$. Let $\boldsymbol{\theta}_n = (\mu_n, \Sigma_n)$ be the iterates generated by (38) with arbitrary initialization $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subseteq \mathbb{R}^N \times \mathbb{S}_{++}^N$. Suppose the proximal regularization parameter λ is strictly positive. Then the limit points of $(\boldsymbol{\theta}_n)_n$ are stationary points of problem (37). Furthermore, it has iteration complexity of $\widetilde{O}(\varepsilon^{-2})$.*

Proof Note we only need to show (A0), (A1), and also the geodesic convexity of the constraint set are satisfied. The proof of (A0)**(i)** and geodesic convexity of the constraint sets are established based on some propositions adapted from Nguyen et al. (2019) which can be found in Appendix C.3. (A1)**(i-rp)** and (A0)**(ii)** holds by our choice of Riemannian proximal surrogates. (A1)**(ii)** holds since the underlying manifolds are Hadamard manifolds. Hence the corollary holds as a result of Theorems 5 and 7. ■

Now we show some numerical results. For numerical experiments, we use the same setup as in Nguyen et al. (2019) and study the empirical convergence behavior. We compare the performance of block minimization using $f_k^{(i)}(\theta)$ for i-th block and RBMM using $g_k^{(i)}(\theta) = f_k^{(i)}(\theta) + \lambda d^2(\theta, \theta_{k-1}^{(i)})$ for i-th block with different values of λ. We set the dimension of the data to be $N = 10$ and denote $f_k = f(\mu_k, \Sigma_k)$, the relative improvement is thus denoted as $|f_{k+1} - f_k|/f_{k+1}$, which is computed via 10 independent experiments. Numerical results are shown in Figure 5. While RBMM and block minimization both perform well, RBMM is slightly faster, where the fastest convergence of RBMM is achieved with $\lambda = 0.01$.



(a)

(b)

Figure 5: Comparison of block minimization and RBMM applied to optimistic likelihood problem under Fisher-Rao distance. RBMM is implemented with $\lambda = 0.01, 0.1, 1$ respectively. Panel (A) shows the averaged error for the first 300 iterations and Panel (B) shows the averaged error using a log-log plot for 1000 iterations.

5.4 Riemannian CP-dictionary-learning (CPDL)

In the CANDECOMP/PARAFAC (CP) decomposition problem (Kolda and Bader, 2009), given a data tensor $X \in \mathbb{R}^{I_1 \times \cdots \times I_m}$ and an integer $R > 0$, we would like to find the *loading matrices* $U^{(i)} \in \mathbb{R}^{I_i \times R}$ for $i = 1, \cdots, m$ such that

$$X \approx \sum_{k=1}^{R} \bigotimes_{i=1}^{m} U^{(i)}[:, k],$$

where $U^{(i)}[:, k]$ denotes the k^{th} column of the $I_i \times R$ loading matrix matrix $U^{(i)}$ and $\otimes$ denotes the outer product. We could formulate the above tensor decomposition problem as the following optimization problem:

$$\underset{U^{(1)} \in \Theta^{(1)}, \ldots, U^{(m)} \in \Theta^{(m)}}{\operatorname{argmin}} \left(f\left(U^{(1)}, \ldots, U^{(m)}\right) := \left\| X - \sum_{k=1}^{R} \bigotimes_{i=1}^{m} U^{(i)}[:, k] \right\|_F^2 \right), \tag{39}$$

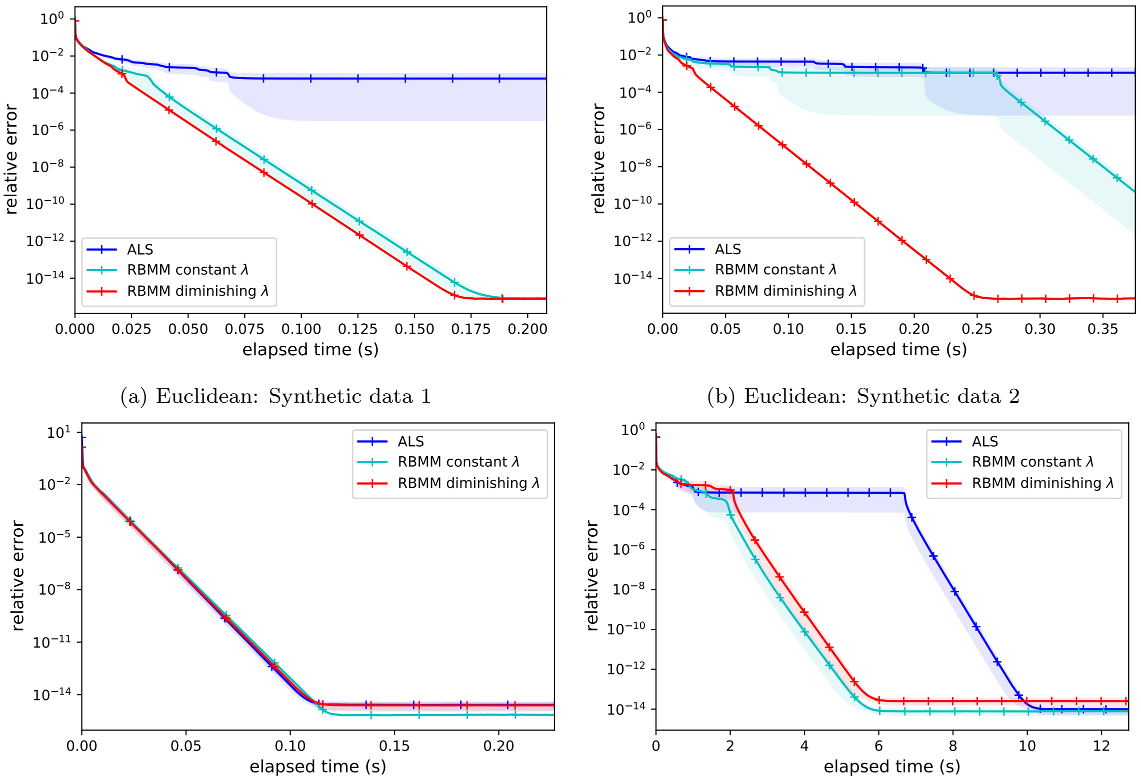
where $\Theta^{(i)} \subseteq \mathbb{R}^{I_i \times R}$ is an embedded manifold, which gives Riemannian constraints on the factor matrices. This setup of Riemmanian CP-dictionary learning is related to the recent work (Dong et al., 2022), where the authors used a CP-decomposition with Riemannian structure on the space of factor matrices as a pre-conditioning algorithm for tensor completion.

It is easy to see (39) is equivalent to

$$\underset{U^{(1)} \in \Theta^{(1)}, \ldots, U^{(m)} \in \Theta^{(m)}}{\operatorname{argmin}} \left\| X - \operatorname{Out}\left(U^{(1)}, \ldots, U^{(m-1)}\right) \times_m \left(U^{(m)}\right)^T \right\|_F^2, \tag{40}$$

which is the CP-dictionary-learning problem in Lyu et al. (2022). Here $\times_m$ denotes the mode-m product (see Kolda and Bader (2009)) and the outer product is given by

$$\operatorname{Out}\left(U^{(1)}, \ldots, U^{(m)}\right) := \left[\bigotimes_{k=1}^{m} U^{(k)}[:, 1], \bigotimes_{k=1}^{m} U^{(k)}[:, 2], \ldots, \bigotimes_{k=1}^{m} U^{(k)}[:, R] \right] \in \mathbb{R}^{I_1 \times \cdots \times I_m \times R}.$$



(a) Euclidean: Synthetic data 1

(b) Euclidean: Synthetic data 2

(c) Stiefel: Synthetic data 1

(d) Low-rank: Synthetic data 1

Figure 6: (A) and (B) are some typical cases of synthetic data in Euclidean case; (C) is the typical result when the first block is Stiefel manifold; (D) is a synthetic example where the first block is a point on low-rank manifold. The average relative reconstruction error with standard deviations are shown by the solid lines and shaded regions of respective colors.

The constrained CP-dictionary-learning problem (40) falls under the framework of block minimization. In fact, let

$$\begin{aligned}
&\mathbf{A} = \mathrm{Out}\left(U_{n-1}^{(1)}, \dots, U_{n-1}^{(i-1)}, U_{n-1}^{(i+1)}, \dots, U_{n-1}^{(m-1)}\right)^{(m)} \in \mathbb{R}^{(I_1 \times \cdots \times I_{i-1} \times I_{i+1} \times \cdots \times I_m) \times R} \\
&B = \mathrm{unfold}(\mathbf{A}, m) \in \mathbb{R}^{(I_1 \cdots I_{i-1} I_{i+1} \cdots I_m) \times R},
\end{aligned}$$

where unfold$(\cdot, \ i \ )$ denotes the mode- i tensor unfolding (see Kolda and Bader (2009)). Then the marginal objective function can be rewritten as

$$f_n^{(i)}(U^{(i)}) = \left\| \mathrm{unfold}(X, i) - B\left(U^{(i)}\right)^T \right\|_F^2. \tag{41}$$

Alternating block minimization of (39), i.e., directly minimizing (41) in each iteration, is known as *alternating least squares* (ALS), which is studied in Kolda and Bader (2009); Navasca et al. (2008); Hong et al. (2015); Razaviyayn et al. (2013) under Euclidean settings, i.e., each Θ is a convex subset of $\mathbb{R}^{I_i \times R}$. Recently, ALS with diminishing radius is also studied in Lyu and Li (2025).

Besides the standard Euclidean setting, we also study the more interesting setting when some blocks are Riemannian manifolds, e.g., Stiefel manifolds or rank-R manifolds. These Riemannian tensor decomposition problems have wide applications, including *low multilinear rank approximation.* We refer readers to Li and Zhang (2023) for further details and additional applications. In order to solve the corresponding Riemannian CP-dictionary learning problem (39), instead of directly minimizing $f_n^{(i)}$, we may apply RBMM, which cyclically minimizes a surrogate function $g_n^{(i)}$ in each iteration n given by

$$g_n^{(i)}(U^{(i)}) := f_n^{(i)}(U^{(i)})+\lambda_n\|U^{(i)}-U_{n-1}^{(i)}\|_F^2 = \left\|\text{unfold}(X,i) - B\left(U^{(i)}\right)^T\right\|_F^2+\lambda_n\|U^{(i)}-U_{n-1}^{(i)}\|_F^2. \tag{42}$$

As mentioned at the beginning of Section 5, the Euclidean distance function is not g-smooth with respect to the low-rank manifold. Hence when at least one $\Theta^{(i)}$ is a low-rank manifold, to ensure g-smoothness of the objective function, we need to take the underlying manifold to be the Euclidean space, i.e., $\mathcal{M}^{(i)} = \mathbb{R}^{I_i\times R}$, and allow the constraint set $\Theta^{(i)}$ to be an embedded submanifold of $\mathbb{R}^{I_i\times R}$. Note that $\Theta^{(i)}$ is not geodesically convex in $\mathbb{R}^{I_i\times R}$ with respect to the Euclidean geometry, so we cannot apply our asymptotic complexity results (Theorems 7 and 10). However, we can still apply Theorem 5 (see also Remark 6). When each $\Theta^{(i)}$ is either a Stiefel manifold or the Euclidean space, we can apply the results in Corollary 11 and therefore have the complexity of $\widetilde{O}(\varepsilon^{-2})$ along with the asymptotic convergence result.

Corollary 22 (Complexity of block Euclidean proximal updates for CPDL) *Let $(U_n^{(1)},\dots,U_n^{(m)})_{n\geq 0}$ be a sequence of factor matrices obtained by RBMM (Alg. 1) with Euclidean proximal surrogates $g_n^{(i)}$ as in (42). If $\inf_n \lambda_n > 0$, then the sequence of iterates converges to the set of stationary points of (39). Moreover, if all the $\Theta^{(i)}$ are either a Stiefel manifold or the Euclidean space, then the iteration complexity is $\widetilde{O}(\varepsilon^{-2})$.*

In the following numerical experiments, we present the advantage of using RBMM in both the Euclidean setting and the Riemannian setting. We generate synthetic data $X \in \mathbb{R}^{30\times 20\times 10}$, then apply ALS (Kolda and Bader, 2009), RBMM with constant $\lambda_n = 0.1$ and RBMM with diminishing $\lambda_n = 0.1 \times 0.5^n$ as suggested in Navasca et al. (2008) to find loading matrices $U^{(1)}, U^{(2)}, U^{(3)}$ with $R = 3$ columns. We run each algorithm 100 times from the independent random initialization and then compute the averaged relative error. The typical results of the Euclidean setting, i.e., $\mathcal{M}^{(i)} = \mathbb{R}^{I_i\times R}$ for $i = 1, 2, 3$ are shown in Figure 6 (A) and (B), where (A) is the case when two RBMM algorithms are significantly better than ALS. When RBMM reaches a relative error of order 10^{-14}, the relative error of ALS is still of order 10^{-3}; (B) shows RBMM with constant λ is better than ALS while RBMM with diminishing λ is much better than the other two. After 0.25s elapsed time, RBMM with diminishing λ reaches a relative error of order 10^{-14}, while that of the other two are still of order 10^{-2}. After 0.35s elapsed time, RBMM with constant λ reaches error of order 10^{-8} while that of ALS remains 10^{-2}.

Figure 6 (C) shows the result when $U^{(1)}$ is a point on the Stiefel manifold $\mathcal{V}^{I_1\times R}$ (see Example 1), i.e., we let $\Theta^{(1)} = \mathcal{V}^{I_1\times R}$, $\mathcal{M}^{(2)} = \mathbb{R}^{I_2\times R}$ and $\mathcal{M}^{(3)} = \mathbb{R}^{I_3\times R}$. In contrast to the Euclidean case, all three algorithms demonstrate good performance for random synthetic data with random initialization. One possible explanation of this phenomenon is that since $U_n^{(1)} \in \mathcal{V}^{I_1 \times R}$ for all n, we have $(U_n^{(1)})^T U_n^{(1)} = I_R$, the $R \times R$ identity matrix. As a result, the quadratic terms added in RBMM would reduce to a linear term that does not significantly accelerate convergence. A similar observation is made when dealing with Stiefel manifolds in the geodesic subspace tracking problem (see Section 5.2 for details).

Figure 6 (D) shows the convergence results for some synthetic data when $U^{(1)}$ is a point on fixed-rank matrices manifold $\mathcal{R}_r$ with $r < R$ (see Example 2) where $r = 5$ and $R = 10$, i.e., we let $\Theta^{(1)} = \mathcal{R}_r$, $\mathcal{M}^{(2)} = \mathbb{R}^{I_2 \times R}$ and $\mathcal{M}^{(3)} = \mathbb{R}^{I_3 \times R}$. There, RBMM with constant or diminishing λ performs much better than ALS. RBMM with constant and diminishing λ require only 5.16s and 5.51s of elapsed time respectively to reach a relative error of order 10^{-13} while that of ALS takes 9.66 seconds.

5.5 Robust PCA

Principal component analysis (PCA) is a popular technique for reducing the dimensionality of a data set, while preserving the maximum amount of information, which is done by seeking the best low-rank approximation of the high dimensional data set (Jolliffe, 1986). Mathematically, consider the data matrix $M \in \mathbb{R}^{m \times n}$, PCA seeks the best r-rank approximation of M where $r < \min\{m, n\}$, by solving $\min_L \{\|L - M\|, \operatorname{rank}(L) \leq r\}$, where $\|X\|$ denotes the spectral norm. However, it is well known that classical PCA would break down when entries of M are grossly corrupted. The PCA problem with grossly corrupted entries is called *robust PCA* (RPCA) problem (Candes et al., 2009). More specifically, it considers the matrix M of the form $M = L_0 + S_0$ where L_0 is the ideal low-rank matrix and S_0 is a sparse matrix. This problem can be mathematically formulated as

$$\min_{L,S} \operatorname{rank}(L) + \lambda \|S\|_0, \quad \text{subject to} \quad M = L + S, \tag{43}$$

where $\lambda > 0$ is a trade-off parameter and $\|S\|_0$ is the number of non-zero entries of S.

Since the objective function in (43) is the sum of two nonconvex functions, it is natural to consider a convex relaxation of (43) by relaxing the rank function by the nuclear norm and the ℓ_0-norm by the ℓ_1-norm. This gives the following *Principal Component Pursuit* (PCP) (Candes et al., 2009; Ma et al., 2015):

$$\min_{L,S} \|L\|_* + \lambda \|S\|_1, \quad \text{subject to} \quad M = L + S \tag{44}$$

where $\|X\|_*$ is the nuclear norm of X, defined to be the sum of the singular values. Under certain conditions on the unknown parameters L_0 and S_0, it is possible to recover L_0 and S_0 by solving the relaxed problem (44). A more general and realistic setting is that the observations are noisy, i.e., $M = L + S + Z$ where Z represents the noise term. This is the *stable PCP* (SPCP) problem proposed in Zhou et al. (2010), where the authors showed the following optimization problem efficiently recovers the true L and S with properly chosen μ,

$$\min_{L,S} \|L\|_* + \lambda \|S\|_1 + \frac{1}{2\mu} \|M - L - S\|_F^2 \tag{45}$$

More studies on solving (45) can also be found in Yin et al. (2019); Aravkin et al. (2014).

While (45) and (44) are well-studied formulations of RPCA, they do not always guarantee a solution with a fixed rank since they promote the rank of L to be small indirectly through the nuclear norm penalization. Rodriguez and Wohlberg (2013) proposed the following alternative formulation of RPCA by using a hard low-rank constraint on L instead of the soft nuclear-norm penalization:

$$\min_{L\in\mathcal{R}_r,\, S\in\mathbb{R}^{m\times n}} F(L,S) := \lambda\|S\|_1 + \frac{1}{2\mu}\|M-L-S\|_F^2 \tag{46}$$

where $\mathcal{R}_r$ is the rank-r matrix manifold (see Example 2). In Rodriguez and Wohlberg (2013), an alternating minimization algorithm for solving (46) was proposed. Here, we revisit it through our general framework of RBMM. The alternating minimization algorithm reads as follows. For any $k = 1, \cdots, T$, denoting the solution at iteration k as L_k and S_k, the next iterates are computed by

$$L_{k+1} = \arg\min_{L\in\mathcal{R}_r} F(L, S_k), \qquad S_{k+1} = \arg\min_{S\in\mathbb{R}^{m\times n}} F(L_{k+1}, S). \tag{47}$$

Note that (47) is a special instance of our RBMM with zero majorization gap. Hence we can generalize it by using proximal surrogates as below:

$$\begin{aligned} L_{k+1} &= \arg\min_{L\in\mathcal{R}_r} G_{k+1}^{(1)}(L) : = F(L, S_k) + \frac{\lambda_{k+1}}{2}\|L-L_k\|_F^2 \\ S_{k+1} &= \arg\min_{S\in\mathbb{R}^{m\times n}} G_{k+1}^{(2)}(S) : = F(L_{k+1}, S) + \frac{\lambda_{k+1}}{2}\|S-S_k\|_F^2. \end{aligned} \tag{48}$$

The iterative updates in (48) fall under the framework of RBMM. Namely, since $\mathcal{M}^{(1)} = \mathcal{M}^{(2)} = \Theta^{(2)} = \mathbb{R}^{m\times n}$ and $\Theta^{(1)} = \mathcal{R}_r$, the assumptions (A0) and (A1) are satisfied trivially. Therefore, we deduce the following corollary of Theorem 5:

Corollary 23 (Convergence of Block MM for PCP) *Fix a matrix $M \in \mathbb{R}^{m\times n}$. Let $\boldsymbol{\theta}_k := (L_k, S_k)$ be generated by (48) with arbitrary initialization $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} := \mathcal{R}_r \times \mathbb{R}^{m\times n}$. Suppose the proximal regularization parameters λ_k are strictly positive for all $k \geq 0$. For the PCP problem (46), the updates in (48) converge to the set of stationary points.*

We remark that in Rodriguez and Wohlberg (2013), the authors numerically verified the efficiency of the alternating minimization for solving (46), but no theoretical results on asymptotic convergence or iteration complexity are established. Corollary 23 provides the asymptotic convergence of the algorithm (48) by using the general RBMM framework.

6. Convergence Analysis

In this section, we prove the convergence and complexity of RBMM (Algorithm 1) stated in Theorems 4, 5, 7, and 10. Throughout this section, we let $(\boldsymbol{\theta}_n)_{n\geq 1}$ denote the sequence of iterates in $\boldsymbol{\Theta} \subseteq \prod_{i=1}^m \mathcal{M}^{(i)}$ generated by Algorithm 1. We will also use the notation introduced in Section 2.2. Note in this section, we use the exponential map as the retraction in the definition of lifted constraint set (7), i.e.,

$$T_x^{\boldsymbol{\Theta}}\mathcal{M} := \{u \in T_x\mathcal{M} \,|\, \mathrm{Exp}_x(u) = x' \text{ for some } x' \in \boldsymbol{\Theta} \text{ with } d(x,x') \leq r_0/2\}.$$

It is worth mentioning that we will be using the exponential map only for the convergence analysis. When implementing RBMM, access to the exponential map is not required, as shown in Algorithm 1.

6.1 Preliminary analysis

In this section, we present some preliminary results that will be used in the analysis of RBMM with both g-smooth and proximal surrogates.

We first establish basic monotonicity properties of the iterates $(\boldsymbol{\theta}_n)_{n\geq 1}$.

Proposition 24 (Monotonicity of objective and Stability of iterates) *Suppose (A0), except for the g-smoothness of f. Then the following hold:*

(i) $F(\boldsymbol{\theta}_{n-1}) - F(\boldsymbol{\theta}_n) \geq \sum_{i=1}^m \left(G_n^{(i)}(\theta_n^{(i)}) - F_n^{(i)}(\theta_n^{(i)}) - \Delta_n \right) \geq -m\Delta_n$*;*

(ii) $\sum_{n=1}^N \sum_{i=1}^m \left(G_n^{(i)}(\theta_n^{(i)}) - F_n^{(i)}(\theta_n^{(i)}) \right) < F(\boldsymbol{\theta}_0) - F^* + m\sum_{n=1}^N \Delta_n < \infty$.

(iii) *Further assume (A1)**(iii)** holds. Then the following also holds:*

$$\sum_{n=1}^N \sum_{i=1}^m \phi\left(d\left(\theta_{n-1}^{(i)}, \theta_n^{(i)}\right)\right) \leq \sum_{n=1}^N \sum_{i=1}^m \left(G_n^{(i)}(\theta_n^{(i)}) - F_n^{(i)}(\theta_n^{(i)}) \right) < F(\boldsymbol{\theta}_0) - F^* + m\sum_{n=1}^N \Delta_n < \infty.$$

In particular, $d(\theta_{n-1}^{(i)}, \theta_n^{(i)}) = o(1)$ *for all* $i = 1, \dots, m$*.*

Proof Fix $i \in \{1, \dots, m\}$. Since $\theta_n^{(i\star)}$ is a minimizer of $G_n^{(i)}$ over $\Theta^{(i)}$ (see (A0)**(ii)**), we get $G_n^{(i)}(\theta_n^{(i\star)}) \leq G_n^{(i)}(\theta_{n-1}^{(i)}) = F_n^{(i)}(\theta_{n-1}^{(i)})$, for $n \geq 1$. Hence we deduce

$$\begin{aligned} F_n^{(i)}(\theta_{n-1}^{(i)}) - F_n^{(i)}(\theta_n^{(i)}) &= G_n^{(i)}(\theta_{n-1}^{(i)}) - G_n^{(i)}(\theta_n^{(i\star)}) + G_n^{(i)}(\theta_n^{(i\star)}) - G_n^{(i)}(\theta_n^{(i)}) + G_n^{(i)}(\theta_n^{(i)}) - F_n^{(i)}(\theta_n^{(i)}) \\ &\geq -\Delta_n + G_n^{(i)}(\theta_n^{(i)}) - F_n^{(i)}(\theta_n^{(i)}) \geq -\Delta_n. \end{aligned}$$

It follows that

$$\begin{aligned} &F(\boldsymbol{\theta}_{n-1}) - F(\boldsymbol{\theta}_n) \\ &\quad = \sum_{i=1}^m F(\theta_n^{(1)}, \dots, \theta_n^{(i-1)}, \theta_{n-1}^{(i)}, \theta_{n-1}^{(i+1)}, \dots, \theta_{n-1}^{(m)}) - F(\theta_n^{(1)}, \dots, \theta_n^{(i-1)}, \theta_n^{(i)}, \theta_{n-1}^{(i+1)}, \dots, \theta_{n-1}^{(m)}) \\ &\quad = \sum_{i=1}^m F_n^{(i)}(\theta_{n-1}^{(i)}) - F_n^{(i)}(\theta_n^{(i)}) \\ &\quad \geq \sum_{i=1}^m G_n^{(i)}(\theta_n^{(i)}) - F_n^{(i)}(\theta_n^{(i)}) - \Delta_n \\ &\quad \geq -m\Delta_n. \end{aligned}$$

This shows **(i)**.

Next, to show **(ii)**, adding up the above inequality for $n = 1, \cdots, N$,

$$\sum_{n=1}^{N}\sum_{i=1}^{m}\left(G_n^{(i)}(\theta_n^{(i)}) - F_n^{(i)}(\theta_n^{(i)})\right) \le \left(\sum_{n=1}^{N} F(\boldsymbol{\theta}_{n-1}) - F(\boldsymbol{\theta}_n)\right) + m\sum_{n=1}^{\infty}\Delta_n$$
$$\le F(\boldsymbol{\theta}_0) - F^* + m\sum_{n=1}^{N}\Delta_n < \infty.$$

To show **(iii)**, note that by (A1)**(iii)**,

$$G_n^{(i)}(\theta_n^{(i)}) - F_n^{(i)}(\theta_n^{(i)}) = g_n^{(i)}(\theta_n^{(i)}) - f_n^{(i)}(\theta_n^{(i)}) \ge \phi\left(d\left(\theta_{n-1}^{(i)}, \theta_n^{(i)}\right)\right).$$

Together with **(ii)** we get **(iii)**. ■

The following is an immediate consequence of Proposition 24 and sub-level compactness.

Proposition 25 (Boundedness of iterates) *Under (A0) except for the g-smoothness of f, the set $\{\boldsymbol{\theta}_n : n \ge 1\}$ is bounded.*

Proof Let $m\sum_{n=1}^{\infty}\Delta_n = T < \infty$, by Prop. 24 we immediately have $F(\boldsymbol{\theta}_n) \le F(\boldsymbol{\theta}_0) + m\sum_{k=1}^{n}\Delta_k < F(\boldsymbol{\theta}_0) + T$. Consider $K = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : F(\boldsymbol{\theta}) \le F(\boldsymbol{\theta}_0) + T\}$, by (A0)**(i)**, K is compact. Hence the set $\{\boldsymbol{\theta}_n : n \ge 1\}$ is bounded. ■

The following two propositions to be introduced will only be used in deriving the rate of convergence results, and they assume the geodesic convexity of the constraint sets.

If $\mathcal{M}$ is a Riemannian manifold and if $x, y \in \mathcal{M}$ such that $d(x,y) < \mathrm{r_{inj}}(x)$, then the tangent vector $\eta_x(y) := \mathrm{Exp}_x(y) \in T_x\mathcal{M}$ is uniquely defined. Now we introduce the following notation:

$$\eta_{n+1}^{(i)}(n) := \eta_{\theta_{n+1}^{(i)}}(\theta_n^{(i)}), \quad \eta_n^{(i)}(n+1) := \eta_{\theta_n^{(i)}}(\theta_{n+1}^{(i)}). \tag{49}$$

Note that the above tangent vectors are well-defined for all n sufficiently large under (A0) and (A1)**(i)-(iii)**. Indeed, by Proposition 24 we have $d(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}) = o(1)$. Also, the injectivity radius is uniformly lower bounded by some constant $r_0 > 0$ (see (A1)**(ii)**). Hence

$$d(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}) < r_0 \quad \text{for all } n \ge N_0 \text{ for some } N_0 \in \mathbb{N},$$

so the tangent vectors in (49) are well-defined for $n \ge N_0$.

First, we show that the first-order variation of the objective function along the trajectory of iterates $\boldsymbol{\theta}_n$ is summable.

Proposition 26 (Finite first-order variation) *Suppose (A0), (A1)**(i)-(ii)**, and (A1)**(iii)** with $\phi(x) = x^2$ hold. Further assume that the constraint set $\Theta^{(i)}$ is strongly convex in $\mathcal{M}^{(i)}$ for $i = 1, \ldots, m$. Then*

$$\sum_{n=N_0}^{\infty}\sum_{i=1}^{m}\left\langle -\operatorname{grad} f_{n+1}^{(i)}(\theta_n^{(i)}), \eta_n^{(i)}(n+1)\right\rangle + p_{n+1}^{(i)}(\theta_n^{(i)}) - p_{n+1}^{(i)}(\theta_{n+1}^{(i)})$$

$$\leq \sum_{n=N_0}^{\infty} \frac{L_f}{2} d^2(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}) + F(\boldsymbol{\theta}_0) - F^*.$$

Proof By g-smoothness of f,

$$\begin{aligned}\left\langle -\operatorname{grad} f_{n+1}^{(i)}(\theta_n^{(i)}), \eta_n^{(i)}(n+1)\right\rangle &\leq \frac{L_f}{2} d^2(\theta_n^{(i)}, \theta_{n+1}^{(i)}) + f_{n+1}^{(i)}(\theta_n^{(i)}) - f_{n+1}^{(i)}(\theta_{n+1}^{(i)}) \\ &\leq \frac{L_f}{2} d^2(\theta_n^{(i)}, \theta_{n+1}^{(i)}) + F_{n+1}^{(i)}(\theta_n^{(i)}) - F_{n+1}^{(i)}(\theta_{n+1}^{(i)}) \\ &\qquad - p_{n+1}^{(i)}(\theta_n^{(i)}) + p_{n+1}^{(i)}(\theta_{n+1}^{(i)}).\end{aligned}$$

Summing up for all $i = 1, \dots, m$ and then for n gives the result. ■

Next, we relate the first-order approximation of the objective difference $F(\boldsymbol{\theta}_{n+1}) - F(\boldsymbol{\theta}_n)$ with that of the worst-case difference $F(\boldsymbol{\theta}) - F(\boldsymbol{\theta}_n)$. This will be used crucially in the proof of Theorem 7 and 10.

Proposition 27 (Bound on first-order optimality gap) *Suppose (A0) and (A1)**(i)-(iii)** hold. Further assume that the constraint set $\Theta^{(i)}$ is strongly convex in $\mathcal{M}^{(i)}$ for $i = 1, \dots, m$. Fix a sequence $(b_n)_{n\geq 1}$ of positive reals with $b_n \leq \hat{r} := \min\{r_0, 1\}$ for all n. Then the following hold for all $n \geq N_0$:*

(i) *(g-smooth surrogates) Suppose (A1)**(i-gs)** holds. Then there exists constant $c_1, c_2 > 0$ independent of $\boldsymbol{\theta}_0$ such that*

$$\begin{aligned}b_{n+1} \sup_{\|\eta^{(i)}\|\leq 1, \eta^{(i)} \in T_{\theta_n^{(i)}}^{\Theta^{(i)}}} V(\boldsymbol{\theta}_n, \eta) &\leq \sum_{i=1}^{m} \left(\left\langle \operatorname{grad} f_{n+1}^{(i)}(\theta_n^{(i)}),\, \eta_n^{(i)}(n+1) \right\rangle + p^{(i)}(\theta_{n+1}^{(i)}) - p^{(i)}(\theta_n^{(i)}) \right) \\ &\quad + m\frac{L_g}{2} b_{n+1}^2 + \frac{L_f}{2} d^2(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}) + m\Delta_n + mL_f b_{n+1} d(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}).\end{aligned}$$

(ii) *(Riemannian proximal surrogates) Suppose (A1)**(i-rp)** holds. Then*

$$\begin{aligned}b_{n+1} \sup_{\|\eta^{(i)}\|\leq 1, \eta^{(i)} \in T_{\theta_n^{(i)}}^{\Theta^{(i)}}} V(\boldsymbol{\theta}_n, \eta) &\leq \sum_{i=1}^{m} \left(\left\langle \operatorname{grad} f_{n+1}^{(i)}(\theta_n^{(i)}),\, \eta_n^{(i)}(n+1) \right\rangle + p^{(i)}(\theta_{n+1}^{(i)}) - p^{(i)}(\theta_n^{(i)}) \right) \\ &\quad + m\frac{L_f + \lambda_n}{2} b_{n+1}^2 + \frac{L_f}{2} d^2(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}) + m\Delta_n \\ &\quad + mL_f b_{n+1} d(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}).\end{aligned}$$

(iii) *(Euclidean proximal surrogates) Suppose (A1)**(i-ep)** holds. Then there exists constant $c_{\boldsymbol{\Theta}} > 0$ independent of $\boldsymbol{\theta}_0$ such that*

$$b_{n+1} \sup_{\|\eta^{(i)}\|\leq 1, \eta^{(i)} \in T_{\theta_n^{(i)}}^{\Theta^{(i)}}} V(\boldsymbol{\theta}_n, \eta) \leq \sum_{i=1}^{m} \left(\left\langle \operatorname{grad} f_{n+1}^{(i)}(\theta_n^{(i)}),\, \eta_n^{(i)}(n+1) \right\rangle + p^{(i)}(\theta_{n+1}^{(i)}) - p^{(i)}(\theta_n^{(i)}) \right)$$

$$+ m\frac{L_f + \lambda_n}{2} b_{n+1}^2 + \frac{L_f}{2c_{\boldsymbol{\Theta}}}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|^2 + m\Delta_n \\ + mL_f b_{n+1} d(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}).$$

Proof We first show **(ii)**. Let $\hat{r} = \min\{r_0, 1\}$, where r_0 is the lower bound of injectivity radius (see (A1) **(ii)**). Fix arbitrary $\boldsymbol{\theta} = [\theta^{(1)}, \cdots, \theta^{(m)}] \in \boldsymbol{\Theta}$ such that

$$d(\theta^{(i)}, \theta_n^{(i)}) \le b_{n+1} \le \hat{r} \le \mathrm{r_{inj}}(\theta^{(i)})$$

for all i. For each i, let $\gamma^{(i)} : [0,1] \to \mathcal{M}^{(i)}$ be the unique distance-minimizing geodesic from $\theta_n^{(i)}$ to $\theta^{(i)}$. Denote $\eta^{(i)} := (\gamma^{(i)})'(0)$. Then $\eta^{(i)} \in T_{\theta_n^{(i)}}^{\Theta^{(i)}}$. Denote $h_n^{(i)} := g_n^{(i)} - f_n^{(i)}$.

Fix $i = 1, \ldots, m$. We will drop the superscripts (i) indicating block i in the proof until we mention otherwise. As in (A0)**(ii)**, let $\theta_{n+1}^{\star}$ be an exact minimizer of G_{n+1} over Θ. Then

$$\begin{aligned} &\langle \mathrm{grad}\, f_{n+1}(\theta_n), \eta_n(n+1)\rangle - \frac{L_f}{2} d^2(\theta_n, \theta_{n+1}) + p(\theta_{n+1}) - p(\theta_n) \\ &\overset{(a)}{\le} f_{n+1}(\theta_{n+1}) - f_{n+1}(\theta_n) + p(\theta_{n+1}) - p(\theta_n) \\ &\overset{(b)}{\le} g_{n+1}(\theta_{n+1}) - g_{n+1}(\theta_n) + p(\theta_{n+1}) - p(\theta_n) \\ &\overset{(c)}{\le} g_{n+1}(\theta_{n+1}^{\star}) + \Delta_n - g_{n+1}(\theta_n) + p(\theta_{n+1}^{\star}) - p(\theta_n) \\ &\overset{(d)}{\le} f_{n+1}(\gamma(1)) - f_{n+1}(\theta_n) + \Delta_n + h_{n+1}(\gamma(1)) + p(\gamma(1)) - p(\theta_n) \\ &\overset{(e)}{\le} \langle \mathrm{grad}\, f_{n+1}(\theta_n), \eta\rangle + \frac{L_f}{2} d^2(\theta_n, \theta) + \Delta_n + h_{n+1}(\gamma(1)) + p(\gamma(1)) - p(\theta_n). \end{aligned} \tag{50}$$

Here, (a) follows from geodesic L_f-smoothness of f (see (A0)**(i)**), (b) follows from definition of majorizing surrogates (Def. 2), (c) follows from the definition of $\theta_n^{\star}$ and the optimality gap Δ_n (see (A0)**(ii)**), (d) follows from the fact that $g_{n+1}(\theta_n) = f_{n+1}(\theta_n)$ and $G_{n+1}(\theta_{n+1}^{\star}) \le G_{n+1}(\gamma(t))$ for all $t \in [0,1]$, (e) uses geodesic L_f-smoothness of f. Also,

$$h_{n+1}(\gamma(1)) = \frac{\lambda_n}{2} d^2(\theta, \theta_n) \le \frac{\lambda_n}{2} b_{n+1}^2. \tag{51}$$

This gives

$$\begin{aligned} &\langle \mathrm{grad}\, f_{n+1}(\theta_n),\, \eta_n(n+1)\rangle + p(\theta_{n+1}) - p(\theta_n) \\ &\qquad \le \langle \mathrm{grad}\, f_{n+1}(\theta_n),\, \eta\rangle + c_1 b_{n+1}^2 + c_2 d^2(\theta_n, \theta_{n+1}) + \Delta_n + p(\gamma(1)) - p(\theta_n), \end{aligned} \tag{52}$$

where $c_1 := \frac{L_f + \lambda_n}{2}$ and $c_2 = \frac{L_f}{2}$. By g-smoothness of f, we have

$$\|\Gamma_{\theta_{n+1} \to \theta_n} \mathrm{grad}\, f_{n+1}(\theta_n) - \mathrm{grad}\, f(\boldsymbol{\theta}_n)\| \le L d(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}).$$

Note (52) holds for all $\theta \in \Theta$ with $d(\theta, \theta_n) \le b_{n+1}$. Hence summing over $i = 1, \cdots, m$ gives

$$\sup_{\|\eta^{(i)}\| \le b_{n+1}, \eta^{(i)} \in T^{\Theta^{(i)}}_{\theta_n^{(i)}}} \left(\langle -\operatorname{grad} f(\boldsymbol{\theta}_n),\, \eta \rangle + \sum_{i=1}^m p^{(i)}(\theta_n^{(i)}) - p^{(i)}(\operatorname{Exp}_{\theta_n^{(i)}}(\eta^{(i)})) \right) \tag{53}$$
$$\le \sum_{i=1}^m \left(\left\langle \operatorname{grad} f_{n+1}^{(i)}(\theta_n^{(i)}),\, \eta_n^{(i)}(n+1) \right\rangle + p^{(i)}(\theta_{n+1}^{(i)}) - p^{(i)}(\theta_n^{(i)}) \right)$$
$$+ mc_1 b_{n+1}^2 + c_2 d^2(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}) + m\Delta_n + mL_f b_{n+1} d(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}).$$

Now fix $\eta \in T^{\Theta^{(i)}}_{\theta_n^{(i)}}$ with $\|\eta\| \le \hat{r}$ and $b \in (0,1)$. Denote $\operatorname{Exp}_{\theta_n^{(i)}}(\eta) = \theta'$ and $\operatorname{Exp}_{\theta_n^{(i)}}(b\eta) = \theta$. Then $d(\theta_n^{(i)}, \theta) = b\|\eta\| = b\, d(\theta_n^{(i)}, \theta')$. Since $p^{(i)}$ is g-convex (see (8))

$$p^{(i)}(\theta) \le b\, p^{(i)}(\theta') + (1-b) p^{(i)}(\theta_n^{(i)}).$$

Rearranging, we get

$$p^{(i)}(\theta_n^{(i)}) - p^{(i)}(\theta) \ge b \left(p^{(i)}(\theta_n^{(i)}) - p^{(i)}(\theta') \right).$$

Hence with $b = b_{n+1}/\hat{r}$,

$$b_{n+1} \sup_{\|\eta^{(i)}\| \le 1, \eta^{(i)} \in T^{\Theta^{(i)}}_{\theta_n^{(i)}}} \left(\langle -\operatorname{grad} f(\boldsymbol{\theta}_n),\, \eta \rangle + \frac{1}{\hat{r}} \left(\sum_{i=1}^m p^{(i)}(\theta_n^{(i)}) - p^{(i)}(\operatorname{Exp}_{\theta_n^{(i)}}(\hat{r}\eta^{(i)})) \right) \right)$$
$$\le \sup_{\|\eta^{(i)}\| \le 1, \eta^{(i)} \in T^{\Theta^{(i)}}_{\theta_n^{(i)}}} \left(\langle -\operatorname{grad} f(\boldsymbol{\theta}_n),\, b_{n+1}\eta \rangle + \sum_{i=1}^m p^{(i)}(\theta_n^{(i)}) - p^{(i)}(\operatorname{Exp}_{\theta_n^{(i)}}(b_{n+1}\eta^{(i)})) \right)$$
$$\le \sup_{\|\eta^{(i)}\| \le b_{n+1}, \eta^{(i)} \in T^{\Theta^{(i)}}_{\theta_n^{(i)}}} \left(\langle -\operatorname{grad} f(\boldsymbol{\theta}_n),\, \eta \rangle + \sum_{i=1}^m p^{(i)}(\theta_n^{(i)}) - p^{(i)}(\operatorname{Exp}_{\theta_n^{(i)}}(\eta^{(i)})) \right).$$

Combining with (53) and using block separability of p, we get

$$b_{n+1} \sup_{\|\eta^{(i)}\| \le 1, \eta^{(i)} \in T^{\Theta^{(i)}}_{\theta_n^{(i)}}} V(\boldsymbol{\theta}_n, \eta) \le \sum_{i=1}^m \left(\left\langle \operatorname{grad} f_{n+1}^{(i)}(\theta_n^{(i)}),\, \eta_n^{(i)}(n+1) \right\rangle + p^{(i)}(\theta_{n+1}^{(i)}) - p^{(i)}(\theta_n^{(i)}) \right)$$
$$+ mc_1 b_{n+1}^2 + c_2 d^2(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}) + m\Delta_n + mL_f b_{n+1} d(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}).$$

We note that **(i)** can be shown by a similar argument. The main difference is that we do not have an explicit expression for the majorization gap h_{n+1} as in (51) but instead, we have g-smoothness of the surrogates. Hence we can proceed by using the g-smoothness of $g_{n+1}^{(i)}$ in line (50) instead of decomposing $g_{n+1}^{(i)} = f_{n+1}^{(i)} + h_{n+1}^{(i)}$ and using the g-smoothness of $f_{n+1}^{(i)}$. The rest of the analysis is identical.

Lastly, **(iii)** can be shown similarly as **(ii)**. Note we only need to replace $d^2(\theta, \theta_n)$ by $\|\theta - \theta_n\|^2$ in (51). Then further bound the geodesic distance by Euclidean distance in (50) with Lemma 16, so there exists $c_\Theta > 0$ that $d^2(\theta_n, \theta_{n+1}) \le \frac{1}{c_\Theta}\|\theta_n - \theta_{n+1}\|^2$. The rest of the analysis is identical. ∎

6.2 RBMM with Riemannian proximal surrogates

In this section, we analyze our RBMM algorithm with Riemannian proximal surrogates. Under (A1)**(i-rp)**, Algorithm 1 is the Riemannian analog of the Euclidean block majorization-minimization with Euclidean proximal regularizer.

Before we present the proof of our general results in Theorem 5 and 7, we will first delve into the special case of block Riemannian proximal methods on Hadamard manifolds, as briefly mentioned in Section 4.3. Examining this special case will be helpful for understanding RBMM with Riemannian proximal surrogates, particularly in two key aspects. First, the Riemannian proximal surrogate $g_n^{(i)}$ is geodesically strongly convex on Hadamard manifolds, which makes the subproblem for the minimization step in Algorithm 1 (line 6) easy to solve. Second, $d^2(\theta, \theta_{n-1}^{(i)})$ is not g-smooth even on Hadamard manifolds, highlighting the necessity of analyzing the case of g-smooth surrogates and Riemannian/Euclidean proximal surrogates separately.

We first recall some useful facts on Hadamard manifolds. First, recall that on Hadamard manifolds, the exponential map is a global diffeomorphism, and therefore its injectivity radius is $+\infty$ (Sakai, 1996, Theorem 4.1, p.221). Second, $d^2(\cdot, p) : \mathcal{M} \to \mathbb{R}$ is geodesically 2-strongly convex for fixed $p \in \mathcal{M}$ where $\mathcal{M}$ is a Hadamard manifold (Bento et al., 2015). We give the formal definition of geodesically strongly convex below.

Definition 28 (Geodesic strong convexity) *Suppose $\mathcal{M}$ is a Hadamard manifold. A function $f : \mathcal{M} \to \mathbb{R}$ is geodesically α-strongly convex if it satisfies*

$$f(y) - f(x) - \left\langle \operatorname{grad} f(x), \operatorname{Exp}_x^{-1}(y) \right\rangle_x \ge \frac{\alpha}{2} d^2(x, y)$$

for all x, $y \in \mathcal{M}$, where $d(x, y)$ is the geodesic distance between x and y on $\mathcal{M}$.

We can also show by definition that the Riemannian proximal surrogate $g_n^{(i)}(\theta)$ is geodesically strongly convex with appropriate Riemannian proximal regularization parameter. This is formally stated in the following proposition.

Proposition 29 (Geodesic strong convexity of Riemannian proximal surrogates) *Let $\mathcal{M}$ be a Hadamard manifold and $f : \mathcal{M} \to \mathbb{R}$ is geodesically L_f-smooth (see Definition 3). Fix $p \in \mathcal{M}$, let $g : \mathcal{M} \to \mathbb{R}$*

$$g(x) = f(x) + \frac{c}{2} d^2(x, p)$$

where $c > L_f$. Then g is geodesically strongly convex with parameter $c - L_f$.

Proof First by g-smoothness of f using Lemma 38 and the geodesical 2-strong convexity of $d^2(\cdot, p)$,

$$f(y) - f(x) - \left\langle \operatorname{grad} f(x), \operatorname{Exp}_x^{-1}(y) \right\rangle_x \ge -\frac{L_f}{2} d^2(x, y)$$

$$d^2(y,p) - d^2(x,p) - \left\langle \operatorname{grad} d^2(x,p), \operatorname{Exp}_x^{-1}(y) \right\rangle_x \geq d^2(x,y)$$

for any x, $y \in \mathcal{M}$. Finally multiplying the second inequality by $\frac{c}{2}$ and adding up with the first inequality finish the proof. ■

Hence, a major benefit of using Riemannian proximal surrogates on Hadamard manifolds is that the sub-problems in RBMM are geodesically strongly convex so can be efficiently solved by using standard convex Riemannian optimization methods (see, e.g., (Udriste, 1994; Zhang and Sra, 2016; Liu et al., 2017)).

Another benefit from the geodesic strong convexity of surrogates is related to the inexact computation, which is stated in the following proposition. Namely, the square of the geodesic distance between the inexact and the exact minimizer of the proximal surrogate (17) is upper bounded by the optimality gap (15), when the surrogate is geodesically strongly convex. Therefore, (16) in (A0)**(ii)** is automatically satisfied on Hadamard manifolds.

Proposition 30 (Optimality gap for iterates) *For each $n \geq 1$ and $i \in \{1, \dots, m\}$, let $\theta_n^{(i\star)}$ be the exact minimizer of the geodesically $(\lambda_n - L_f)$-strongly convex function $\theta \mapsto g_n^{(i)}(\theta)$ in (17) over the strongly convex set $\Theta^{(i)}$. Then*

$$\frac{\lambda_n - L_f}{2} d^2\left(\theta_n^{(i\star)}, \theta_n^{(i)}\right) \leq \Delta_n.$$

Proof Note the optimality condition of $\theta_n^{(i\star)}$ over $\Theta^{(i)}$ is

$$\langle g_n^{(i)}(\theta_n^{(i\star)}), \operatorname{Exp}^{-1}_{\theta_n^{(i\star)}}(\theta) \rangle \geq 0, \quad \forall \theta \in \Theta^{(i)}.$$

Then we have

$$\begin{aligned}\frac{\lambda_n - L_f}{2} d^2\left(\theta_n^{(i\star)}, \theta_n^{(i)}\right) &\leq g_n^{(i)}(\theta_n^{(i)}) - g_n^{(i)}(\theta_n^{(i\star)}) - \langle g_n^{(i)}(\theta_n^{(i\star)}), \operatorname{Exp}^{-1}_{\theta_n^{(i\star)}}(\theta_n^{(i)}) \rangle \\ &\leq g_n^{(i)}(\theta_n^{(i)}) - g_n^{(i)}(\theta_n^{(i\star)}) \leq \Delta_n.\end{aligned}$$

■

Next, we give an illustration of $d^2(\cdot, p)$ is not g-smooth in general. The following proposition gives the expression of the Riemannian gradient of geodesic distance.

Proposition 31 (Riemannian gradient of geodesic distance) *Let $\mathcal{M}$ be a complete Riemannian manifold, $p \in \mathcal{M}$ with $r_{\text{inj}}(p) = r$. Consider the function $h : \mathcal{M} \to \mathbb{R}$ where $h(x) = d^2_{\mathcal{M}}(x,p)$, where $p \in \mathcal{M}$ is fixed. If $d(x,p) < r$, then $\operatorname{grad}(h) = -2\operatorname{Exp}_x^{-1}(p)$ as a vector in $T_x\mathcal{M}$.*

Proof See e.g., Sakai (1996, Proposition 4.8, p. 108). ■

With Prop. 31, we can now explain why the surrogate given in (A1)**(i-rp)** is not g-smooth in general. Recall Def. 3, in order $\frac{1}{2}d^2(x,p)$ in Prop. 31 to be g-smooth, we need the following to hold,

$$\frac{1}{2}\left\|\operatorname{grad} d^2(y,p) - \Gamma_{x\to y}(\operatorname{grad} d^2(x,p))\right\| = \left\|-\operatorname{Exp}_y^{-1}(p) + \Gamma_{x\to y}(\operatorname{Exp}_x^{-1}(p))\right\| \leq L d(x,y). \tag{54}$$

In Euclidean space, parallel transport $\Gamma_{x\to y}$ is identical, and (54) would degenerate to equality with $L = 1$. Hence $\frac{1}{2}d^2(x,p)$ is g-smooth in Euclidean space, see Fig. 7(a) for an illustrative example. However, on a general Riemannian manifold, even on the Hadamard manifold, this is no longer true in general, see Fig. 7(b) for a counterexample of hyperbolic space (see Example 5). Some more examples of S^1 are shown in Fig. 8.

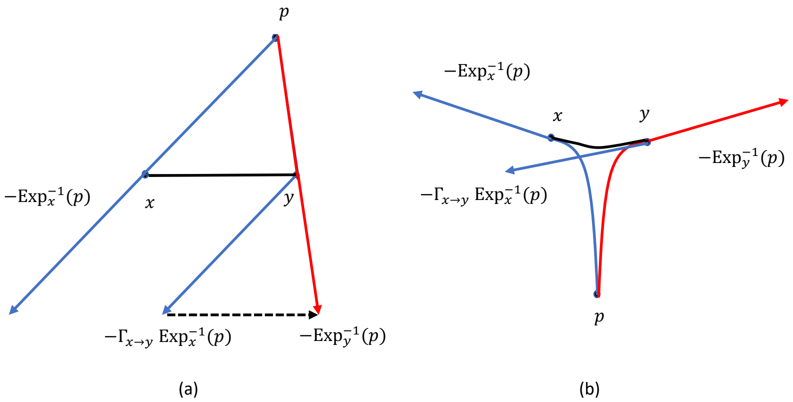


Figure 7: Examples on g-smoothness of $d^2(x,p)$. Panel (a) is an example in Euclidean space, the length of the dashed black arrow represents the LHS of (54) and the length of the black segment is $d(x,y)$; Panel (b) is a counterexample in hyperbolic space. In this case, for fixed points x, y on the manifold, the $d(x,y)$ on the RHS of (54) remains constant. In contrast, the left-hand side of (54), which is the sum of the lengths of the blue and red geodesic segments, can grow arbitrarily large as the point p moves farther away from x and y.

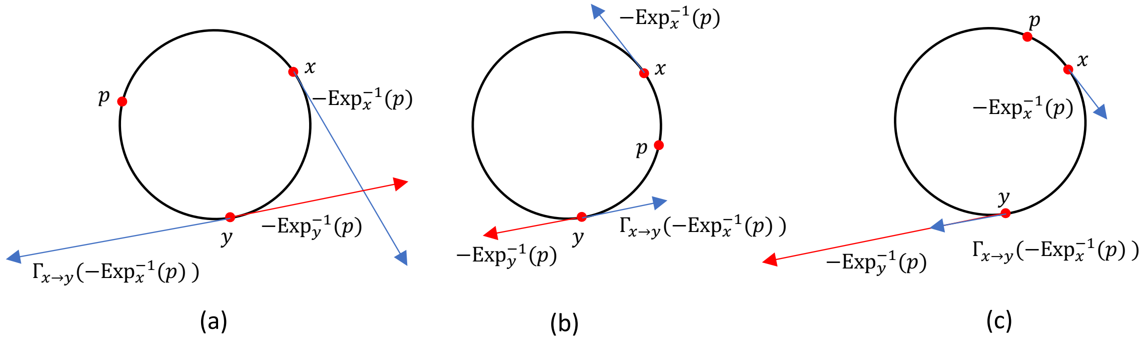


Figure 8: Examples on g-smoothness of $d^2(x,p)$ on S^1. Panel (a) is the case when (54) does not hold; Panel (b) (c) is the case when (54) becomes an equality with $L = 2$.

Now we come back to establish the proof of Theorems 5 and 7. Let $\boldsymbol{\theta}_\infty = [\theta_\infty^{(1)}, \cdots, \theta_\infty^{(m)}]$ be a limit point of $(\boldsymbol{\theta}_n)_n$, for simplicity of the notation, replace $(\boldsymbol{\theta}_n)_n$ by the convergent subsequence. We introduce the following notation of tangent vectors and sets, which will be used throughout the remaining proofs. For $n \geq 1$, let

$$\Theta_n^{(i)} = \{\theta \in \Theta^{(i)} : d(\theta, \theta_n^{(i)}) \leq r_0/2\}$$

where r_0 is the uniform lower bound of injectivity radius (see (A1)**(ii)**). Similarly, denote

$$\Theta_\infty^{(i)} = \{\theta \in \Theta^{(i)} : d(\theta, \theta_\infty^{(i)}) \leq r_0/2\}. \tag{55}$$

For any $\theta^{(i)} \in \Theta_{\infty}^{(i)}$, let

$$\eta_n^{(i)} := \eta_{\theta_n^{(i)}}(\theta^{(i)}), \quad \eta_n^{(i\star)} := \eta_{\theta_n^{(i\star)}}(\theta^{(i)}), \text{ and } \eta_{\infty}^{(i)} := \eta_{\theta_{\infty}^{(i)}}(\theta^{(i)}). \tag{56}$$

Note that the above tangent vectors are well-defined for all n sufficiently large under (A0) and (A1)**(i)-(ii)**. In fact, since $\boldsymbol{\theta}_n$ converges to $\boldsymbol{\theta}_{\infty}$,

$$d(\theta_n^{(i)}, \theta_{\infty}^{(i)}) < r_0/4 \quad \text{for all } n \geq N_1 \text{ for some } N_1 \in \mathbb{N}.$$

Also, by (A0),

$$d(\theta_n^{(i)}, \theta_n^{(i\star)}) < r_0/4 \quad \text{for all } n \geq N_2 \text{ for some } N_2 \in \mathbb{N}.$$

Therefore by triangle inequality, the tangent vectors in (56) are well defined for $n \geq \max\{N_1, N_2\}$. The following proposition shows the asymptotic first-order optimality of inexact solutions, which is the key proposition to prove Theorem 5.

Proposition 32 (Asymptotic first order optimality of inexact solutions) *Assume (A1)**(i-rp), (ii)**, (A0). Let $\boldsymbol{\theta}_{\infty} = [\theta_{\infty}^{(1)}, \cdots, \theta_{\infty}^{(m)}]$ be a limit point of $(\boldsymbol{\theta}_n)_n$, for simplicity of the notation, replace $(\boldsymbol{\theta}_n)_n$ by the convergent subsequence. Consider any $\theta^{(i)} \in \boldsymbol{\Theta}_{\infty}^{(i)}$ (see (55)) such that $\eta_{\infty}^{(i)}$ defined in (56) satisfies $\|\eta_{\infty}^{(i)}\| \leq 1$. Let $\eta_n^{(i\star)}$ and $\eta_n^{(i)}$ be tangent vectors defined in (56). For each $i = 1, \cdots, m$, the following holds,*

$$\left| \left\langle \operatorname{grad} g_n^{(i)}(\theta_n^{(i\star)}), \eta_n^{(i\star)} \right\rangle - \left\langle \operatorname{grad} g_n^{(i)}(\theta_n^{(i)}), \eta_n^{(i)} \right\rangle \right| = o(1).$$

Proof First by triangle inequality and (A0),

$$d(\theta_n^{(i\star)}, \theta_{\infty}^{(i)}) \leq d(\theta_n^{(i)}, \theta_{\infty}^{(i)}) + d(\theta_n^{(i\star)}, \theta_n^{(i)}) = o(1),$$

so $\theta_n^{(i\star)} \to \theta_{\infty}^{(i)}$ as $n \to \infty$.

Next, note that by Prop. 24,

$$d(\theta_n^{(i)}, \theta_{n-1}^{(i)}) = o(1). \tag{57}$$

This means when $n >> 1$, $d(\theta_n^{(i)}, \theta_{n-1}^{(i)}) \leq r^{(i)} \leq r_{\text{inj}}(\theta_n^{(i)})$.

Therefore, by (57) and (A0),

$$\| \operatorname{Exp}_{\theta_n^{(i\star)}}^{-1}(\theta_{n-1}^{(i)}) \| = d(\theta_n^{(i\star)}, \theta_{n-1}^{(i)}) \leq d(\theta_n^{(i\star)}, \theta_n^{(i)}) + d(\theta_n^{(i)}, \theta_{n-1}^{(i)}) = o(1).$$

Hence

$$\operatorname{Exp}_{\theta_n^{(i\star)}}^{-1}(\theta_{n-1}^{(i)}) \to \mathbf{0}. \tag{58}$$

Similarly, we have $\operatorname{Exp}_{\theta_n^{(i)}}^{-1}(\theta_{n-1}^{(i)}) \to \mathbf{0}$.

Let $\Gamma_{\theta_n^{(i\star)} \to \theta_{\infty}^{(i)}}$ be the parallel transport along the minimal geodesic joining $\theta_n^{(i\star)}$ and $\theta_{\infty}^{(i)}$. Since $\theta_n^{(i\star)} \to \theta_{\infty}^{(i)}$, by smoothness of the exponential map, we have $\Gamma_{\theta_n^{(i\star)} \to \theta_{\infty}^{(i)}} \eta_n^{(i\star)} \to \eta_{\infty}^{(i)}$ where $\operatorname{Exp}_{\theta_{\infty}^{(i)}}(\eta_{\infty}^{(i)}) = \theta^{(i)}$ (see Azagra et al. (2005, Remark 6.11) for a concise conclusion).

Also note $\|\eta_\infty^{(i)}\| \le 1$. Therefore by continuity of the Riemann metric together with (57), (58), and the geodesical smoothness of f as well as $\lambda_n = O(1)$, we have

$$\begin{aligned}\left\langle \operatorname{grad} g_n^{(i)}(\theta_n^{(i\star)}), \eta_n^{(i\star)} \right\rangle &= \left\langle \operatorname{grad} f_n^{(i)}(\theta_n^{(i\star)}), \eta_n^{(i\star)} \right\rangle - \left\langle \lambda_n \operatorname{Exp}_{\theta_n^{(i\star)}}^{-1}(\theta_{n-1}^{(i)}), \eta_n^{(i\star)} \right\rangle \\ &\to \left\langle \operatorname{grad}_i f(\boldsymbol{\theta}_\infty), \eta_\infty^{(i)} \right\rangle.\end{aligned}$$

Similarly, we have $\left\langle \operatorname{grad} g_n^{(i)}(\theta_n^{(i)}), \eta_n^{(i)} \right\rangle \to \left\langle \operatorname{grad}_i f(\boldsymbol{\theta}_\infty), \eta_\infty^{(i)} \right\rangle$. The assertion follows. ■

We are now ready to prove Theorem 5.

Proof of Theorem 5 under (A1)(i-rp). Assume (A0), and (A1) **(i-rp), (ii)**. Futher assume p is lower semi-continuous on $\boldsymbol{\Theta}$. Fix a convergent subsequence $(\boldsymbol{\theta}_{n_k})_{k\ge 1}$ of $(\boldsymbol{\theta}_n)_{n\ge 1}$. We wish to show that $\boldsymbol{\theta}_\infty = \lim_{k\to\infty} \boldsymbol{\theta}_{n_k}$ is a stationary point of f over $\boldsymbol{\Theta}$. Fix any $\theta^{(i)} \in \Theta_\infty^{(i)}$ (see (55)) with $d(\theta^{(i)}, \theta_\infty^{(i)}) \le 1$, by first-order optimality of $\theta_{n_k}^{(i\star)}$,

$$\left\langle \operatorname{grad} g_{n_k}^{(i)}(\theta_{n_k}^{(i\star)}), \eta_{n_k}^{(i\star)} \right\rangle + p_{n_k}^{(i)}(\operatorname{Exp}(\eta_{n_k}^{(i\star)})) - p_{n_k}^{(i)}(\theta_{n_k}^{(i\star)}) \ge 0.$$

Note that by Prop. 24,

$$d(\theta_n^{(i)}, \theta_{n-1}^{(i)}) = o(1),$$

which means when $n \gg 1$, $d(\theta_n^{(i)}, \theta_{n-1}^{(i)}) \le r_{\text{inj}}^{(i)} \le r_{\text{inj}}(\theta_n^{(i)})$, so the inverse exponential map is well defined.

Then by Prop. 31 and Prop. 32,

$$\begin{aligned}&\liminf_{k\to\infty} \left\langle \operatorname{grad} g_{n_k}^{(i)}(\theta_{n_k}^{(i)}), \eta_{n_k}^{(i)} \right\rangle + p_{n_k}^{(i)}(\theta^{(i)}) - p_{n_k}^{(i)}(\theta_{n_k}^{(i\star)}) \\ &= \liminf_{k\to\infty} \left\langle \operatorname{grad} f_{n_k}^{(i)}(\theta_{n_k}^{(i)}) - \lambda_{n_k} \operatorname{Exp}_{\theta_{n_k}^{(i)}}^{-1}(\theta_{n_k-1}^{(i)}), \eta_{n_k}^{(i)} \right\rangle + p_{n_k}^{(i)}(\theta^{(i)}) - p_{n_k}^{(i)}(\theta_{n_k}^{(i\star)}) \ge 0.\end{aligned}$$

Note by Prop. 24 and $\lambda_n = O(1)$, we get $\lambda_{n_k} d(\boldsymbol{\theta}_{n_k}, \boldsymbol{\theta}_{n_k-1}) = o(1)$ and therefore $\|\lambda_{n_k} \operatorname{Exp}_{\theta_{n_k}^{(i)}}^{-1}(\theta_{n_k-1}^{(i)})\| = \lambda_{n_k} d(\theta_{n_k}^{(i)}, \theta_{n_k-1}^{(i)}) = o(1)$. Also recall $\|\eta_\infty^{(i)}\| \le 1$, so $\|\eta_{n_k}^{(i)}\|$ is uniformly bounded. Thus

$$\liminf_{k\to\infty} \left\langle \operatorname{grad} f_{n_k}^{(i)}(\theta_{n_k}^{(i)}), \eta_{n_k}^{(i)} \right\rangle + p_{n_k}^{(i)}(\theta^{(i)}) - p_{n_k}^{(i)}(\theta_{n_k}^{(i)}) \ge 0.$$

Denote $\boldsymbol{\theta}_\infty = \left[\theta_\infty^{(1)}, \dots, \theta_\infty^{(m)}\right]$. Let $\Gamma_{\theta_{n_k}^{(i)} \to \theta_\infty^{(i)}}$ be the parallel transport along the minimal geodesic joining $\theta_{n_k}^{(i)}$ and $\theta_\infty^{(i)}$. By smoothness of exponential map, we have $\Gamma_{\theta_{n_k}^{(i)} \to \theta_\infty^{(i)}} \eta_{n_k}^{(i)} \to \eta_\infty^{(i)} \in T_{\theta_\infty^{(i)}}$ satisfying $\operatorname{Exp}_{\theta_\infty^{(i)}}(\eta_\infty^{(i)}) = \theta^{(i)}$. Then the above together with continuity of grad f and lower semi-continuity of p give

$$\begin{aligned}&\left\langle \operatorname{grad}_i f\left(\theta_\infty^{(1)}, \dots, \theta_\infty^{(i-1)}, \theta_\infty^{(i)}, \theta_\infty^{(i+1)}, \dots, \theta_\infty^{(m)}\right), \eta_\infty^{(i)} \right\rangle + p_\infty^{(i)}(\theta^{(i)}) - p_\infty^{(i)}(\theta_\infty^{(i)}) \\ &\ge \liminf_{k\to\infty} \left\langle \operatorname{grad} f_{n_k}^{(i)}(\theta_{n_k}^{(i)}), \eta_{n_k}^{(i)} \right\rangle + p_{n_k}^{(i)}(\theta^{(i)}) - p_{n_k}^{(i)}(\theta_{n_k}^{(i)}) \ge 0.\end{aligned}$$

This holds for all $i = 1, \dots, m$. Therefore we verified $V(\boldsymbol{\theta}_\infty), \boldsymbol{\eta}) \le 0$ for all $\eta^{(i)} \in T^{\Theta^{(i)}}_{\theta^{(i)}_\infty}$ with $\|\eta^{(i)}\| \le \hat{r}$, which means that $\boldsymbol{\theta}_\infty$ is a stationary point of F over $\boldsymbol{\Theta}$, as desired. ■

Proof of Theorem 7 under (A1)(i-rp). Let b_n be any square-summable positive sequence. First, by Cauchy-Schwartz inequality, we have

$$\sum_{n=1}^{N} b_{n+1} d(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}) \le \left(\sum_{n=1}^{N} b_n^2 \right)^{1/2} \left(\sum_{n=1}^{N} d^2(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}) \right)^{1/2}. \tag{59}$$

Then by Prop. 27 with $h_n^{(i)}(\theta) = \frac{\lambda_n}{2} d^2(\theta, \theta^{(i)}_{n-1})$, using Prop. 24 and Prop. 26,

$$\begin{aligned}
&\sum_{n=1}^{N} b_{n+1} \sup_{\eta^{(i)} \in T^{\Theta^{(i)}}_{\theta^{(i)}_n}, \|\eta^{(i)}\| \le 1} V(\boldsymbol{\theta}_n, \boldsymbol{\eta}) \\
&\qquad\qquad \le C \left(F(\boldsymbol{\theta}_0) - \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} F(\boldsymbol{\theta}) + \sum_{n=1}^{N} b_n^2 + \sum_{n=1}^{N} d^2(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}) + \sum_{n=1}^{N} \Delta_n(\boldsymbol{\theta}_0) \right)
\end{aligned}$$

for some constant $C > 0$ independent of $\boldsymbol{\theta}_0$ and the right hand side is finite. Recall that by Proposition 24 and the hypothesis,

$$\sum_{n=1}^{N} d^2(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}) \le 2\lambda_{\min}^{-1} \left(F(\boldsymbol{\theta}_0) - F^* + m \sum_{n=1}^{N} \Delta_n \right).$$

Now take $b_n = \frac{\hat{r}}{\sqrt{n}\log(1+n)}$. Then $\sum_n b_n^2 < 5\hat{r}^2$, $\sum_{k=1}^{n} b_k \sim \hat{r} n^{1/2} / \log n$. Using Lemma 40, we deduce

$$\min_{1 \le k \le n} \left[\sup_{\eta^{(i)} \in T^{\Theta^{(i)}}_{\theta^{(i)}_k}, \|\eta^{(i)}\| \le 1} V(\boldsymbol{\theta}_k, \boldsymbol{\eta}) \right] \le \frac{M}{\sqrt{n}/\log n},$$

where

$$\begin{aligned}
M = \hat{r}^{-1} \left(L_f \lambda_{\min}^{-1} + 1 \right) \left(F(\boldsymbol{\theta}_0) - F^* + m \sum_{n=1}^{N} \Delta_n \right) + \frac{m(L_f + \lambda_{\max})}{2} 5\hat{r} \\
+ m L_f \lambda_{\min}^{-1/2} \sqrt{5} \left(F(\boldsymbol{\theta}_0) - F^* + m \sum_{n=1}^{N} \Delta_n \right)^{1/2}.
\end{aligned} \tag{60}$$

This shows **(i)**. Note that M above depends on the manifold geometry implicitly: L_f depends on the properties of the underlying manifold $\mathcal{M}$ as well as the objective function f. Moreover, when inexact computations are allowed, the sum of the optimality gap $\sum_{n=1}^{\infty} \Delta_n(\boldsymbol{\theta}_0)$ also depends on $\mathcal{M}$.

Assume $\sup_{\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}} \sum_{n=1}^{\infty} \Delta_n(\boldsymbol{\theta}_0) < \infty$. Then we conclude **(ii)** by using the fact that $n \ge 2\varepsilon^{-2} \left(\log \varepsilon^{-2} \right)^2$ implies $\log n / \sqrt{n} \le \varepsilon$ for all sufficiently small $\varepsilon > 0$. This completes the proof.

■

6.3 RBMM with Euclidean proximal surrogates

In this section, we prove Theorems 5 and 7 under (A1)**(i-ep)**. The proof is very similar to that of Theorems 5 and 7 under (A1)**(i-rp)**, as shown above. However, it is not a direct corollary from the aforementioned theorems, due to the difference of assumptions. In the following proof, we omit the repeated details for conciseness and only show the key parts.

Proof of Theorems 5 and 7 under (A1)(i-ep). We first show the parallel results from Prop. 24 and 27. Using the same analysis as in Prop. 24 with the Euclidean proximal surrogate in (A1)**(i-ep)**, we have

$$\sum_{n=1}^{N}\sum_{i=1}^{m}\lambda_n\|\theta_{n-1}^{(i)}-\theta_n^{(i)}\|^2 \le \sum_{n=1}^{N}\sum_{i=1}^{m}\left(g_n^{(i)}(\theta_n^{(i)})-f_n^{(i)}(\theta_n^{(i)})\right) < f(\boldsymbol{\theta}_0)-f^*+m\sum_{n=1}^{N}\Delta_n.$$

In particular, $\|\theta_{n-1}^{(i)}-\theta_n^{(i)}\|=o(1)$ for all $i=1,\dots,m$. By Prop. 27**(iii)**, we can show the complexity results using the analysis in the proof of Theorem 7 under (A1)**(i-rp)** and we omit it here.

To show the asymptotic convergence, first note that

$$\operatorname{grad} g_n^{(i)}(\theta_n^{(i)}) = \operatorname{grad} f_n^{(i)}(\theta_n^{(i)}) - \lambda_n \operatorname{Proj}_{T_{\theta_{n-1}^{(i)}}}(\theta_n^{(i)}-\theta_{n-1}^{(i)}).$$

Also note $\|\operatorname{Proj}_{T_{\theta_{n-1}^{(i)}}}(\theta_n^{(i)}-\theta_{n-1}^{(i)})\|\le\|\theta_n^{(i)}-\theta_{n-1}^{(i)}\|=o(1)$. Hence, by an identical argument as in the proof of Theorem 5 under (A1)**(i-rp)**, we conclude $(\boldsymbol{\theta}_n)_n$ asymptotically converges to the set of stationary points. ■

6.4 Total complexity of RBMM with Riemannian proximal surrogates

In this section, we provide the proof of Corollary 8.

To solve each surrogate subproblem, we apply the Riemannian projected (sub)gradient descent (RPGD) method. We omit the superscript (i) for conciseness when introducing RPGD. On a Hadamard manifold $\mathcal{M}$, assume we have access to a projection $\operatorname{Proj}_\Theta$ that maps a point $\theta\in\mathcal{M}$ to $\operatorname{Proj}_\Theta(\theta)\in\Theta\subseteq\mathcal{M}$ such that

$$d\big(\theta,\operatorname{Proj}_\Theta(\theta)\big) < d\big(\theta,\theta'\big) \quad ,\forall\theta'\in\Theta\backslash\{\operatorname{Proj}_\Theta(\theta)\}.$$

RPGD takes the following form for minimizing an objective function $g(\theta)$ on $\mathcal{M}$,

$$\textbf{(RPGD)} \qquad \theta_{k+1}=\operatorname{Proj}_\Theta\left(\operatorname{Exp}_{\theta_k}(-\eta_k g_k)\right), \tag{61}$$

where k is the iteration index, η_k is the step-size, g_k is the Riemannian subgradient at θ_k. Namely, $g_k\in T_{\theta_k}$ such that

$$g(\theta)\ge g(\theta_k)+\langle g_k,\operatorname{Exp}_{\theta_k}(\theta)\rangle.$$

Note that for a g-convex function, a Riemannian subgradient always exists (Zhang and Sra, 2016). Following (Zhang and Sra, 2016, Thm. 11), we also define an auxiliary iterates $\bar{\theta}_k$ by $\bar{\theta}_1 := \theta_1$ and

$$\bar{\theta}_k = \mathrm{Exp}_{\bar{\theta}_{k-1}}\left(\frac{2}{k}\,\mathrm{Exp}^{-1}_{\bar{\theta}_{k-1}}(\theta_k)\right). \tag{62}$$

Proof of Corollary 8. We first prove the corollary with smooth objectives, i.e., when $p = 0$. First, the g-strong convexity of the Riemannian proximal surrogate $g_n^{(i)}$ follows directly from Proposition 29, while its g-smoothness is established in Alimisis et al. (2020); Zhang and Sra (2016). Next, by Theorem 15 in Zhang and Sra (2016), and under the assumptions stated in the corollary, the optimality gap of the surrogate subproblem satisfies

$$\Delta_k^{(i)} \le \frac{(1-\delta)^{N_k^{(i)}-2} D^2 (L_f + 2\zeta(\kappa, D))}{2}. \tag{63}$$

Now choose $N_k^{(i)} = C \log k$ with $C = 2/(-\log(1-\delta))$. Substituting this choice into (63) gives

$$\begin{aligned}
\sum_{k=1}^{\infty} \Delta_k &\le m \sum_{k=1}^{\infty} \frac{(1-\delta)^{C\log k - 2}(L_f + 2\zeta(\kappa, D))D^2}{2} \\
&= \frac{m(L_f + 2\zeta(\kappa, D))D^2}{2(1-\delta)^2} \sum_{k=1}^{\infty} k^{C\log(1-\delta)} \\
&= \frac{m(L_f + 2\zeta(\kappa, D))D^2}{2(1-\delta)^2} \sum_{k=1}^{\infty} k^{-2} \\
&\le \frac{m(L_f + 2\zeta(\kappa, D))D^2}{2(1-\delta)^2} \frac{\pi^2}{6}.
\end{aligned}$$

Therefore, plugging this into the result of Theorem 7**(i)**, we obtain

$$\begin{aligned}
\min_{1\le k\le n}\left[\sup_{\eta^{(i)}\in T_{\theta_k^{(i)}}\Theta^{(i)}, \|\eta^{(i)}\|\le 1} V(\boldsymbol{\theta}_k, \boldsymbol{\eta})\right] &\le c\frac{L_f(1+\lambda_{\min}^{-1})(1+\sum_{k=1}^{\infty}\Delta_k) + \lambda_{\max}}{\sqrt{n}/\log n} \\
&\le c\frac{L_f(1+\lambda_{\min}^{-1})(1+K) + \lambda_{\max}}{\sqrt{n}/\log n},
\end{aligned}$$

where $K = \frac{m(L_f + 2\zeta(\kappa,D))D^2}{2(1-\delta)^2}\frac{\pi^2}{6}$. In order that

$$c\frac{L_f(1+\lambda_{\min}^{-1})(1+K) + \lambda_{\max}}{\sqrt{n}/\log n} \le \varepsilon,$$

it suffices to take $N_\varepsilon = O\Big((1+\zeta(\kappa, D))^2\,\varepsilon^{-2}\left(\log\frac{1}{\varepsilon}\right)^2\Big) = \widetilde{O}\big((1+\zeta(\kappa, D))^2\,\varepsilon^{-2}\big)$.

The total complexity needed is therefore

$$\sum_{k=1}^{N_\varepsilon}\sum_{i=1}^{m} N_k^{(i)} = \sum_{k=1}^{N_\varepsilon}\sum_{i=1}^{m} C\log k = O(N_\varepsilon \log N_\varepsilon) = \widetilde{O}\big((1+\zeta(\kappa,D))^2\,\varepsilon^{-2}\big).$$

Next, we prove the corollary with nonsmooth objectives. The g-strong convexity and geodesical Lipschitz continuity of $g_n^{(i)}$ follow similarly from the assumptions. Next, by Theorem 11 in Zhang and Sra (2016), the optimality gap satisfies

$$\Delta_k^{(i)} \le \frac{2\zeta(\kappa,D)(L^{(i)}+\lambda_k D)^2}{(\lambda_k - L_f)(N_k^{(i)}+1)}. \tag{64}$$

Now choose $N_k^{(i)} = k$. Substituting into (64) gives

$$\sum_{k=1}^{n}\Delta_k \le \sum_{k=1}^{n}\frac{2\zeta(\kappa,D)(L^{(i)}+\lambda_{\max}D)^2}{(\lambda_{\min}-L_f)(k+1)} \le \frac{2\zeta(\kappa,D)(L^{(i)}+\lambda_{\max}D)^2}{(\lambda_{\min}-L_f)}\log n$$

Plugging this into the result of Theorem 7**(i)** gives

$$\begin{aligned}
\min_{1\le k\le n}\left[\sup_{\eta^{(i)}\in T^{\Theta^{(i)}}_{\theta_k^{(i)}},\|\eta^{(i)}\|\le 1} V(\boldsymbol{\theta}_k,\boldsymbol{\eta})\right] &\le c\frac{L_f(1+\lambda_{\min}^{-1})(1+\sum_{k=1}^n\Delta_k)+\lambda_{\max}}{\sqrt{n}/\log n}\\
&= c\frac{L_f(1+\lambda_{\min}^{-1})\sum_{k=1}^n\Delta_k}{\sqrt{n}/\log n} + c\frac{L_f(1+\lambda_{\min}^{-1})+\lambda_{\max}}{\sqrt{n}/\log n}\\
&\le c\frac{C(\kappa,D)L_f(1+\lambda_{\min}^{-1})\log n}{\sqrt{n}/\log n} + c\frac{L_f(1+\lambda_{\min}^{-1})+\lambda_{\max}}{\sqrt{n}/\log n}
\end{aligned}$$

where $C(\kappa,D) = \frac{2\zeta(\kappa,D)(L^{(i)}+\lambda_{\max}D)^2}{(\lambda_{\min}-L_f)}$. It suffices to guarantee

$$A_{\mathrm{ns}}(\kappa,D)\,\frac{(\log n)^2}{\sqrt{n}} \le \varepsilon,$$

where

$$\begin{aligned}
A_{\mathrm{ns}}(\kappa,D) &:= c\Big[L_f(1+\lambda_{\min}^{-1})C(\kappa,D) + L_f(1+\lambda_{\min}^{-1})+\lambda_{\max}\Big]\\
&:= \alpha_0 + \alpha_1\,\zeta(\kappa,D).
\end{aligned}$$

Therefore, it suffices to choose

$$N_\varepsilon := \left\lceil \frac{16\,A_{\mathrm{ns}}(\kappa,D)^2}{\varepsilon^2}\left(\log\frac{4A_{\mathrm{ns}}(\kappa,D)}{\varepsilon}\right)^4\right\rceil = \widetilde{O}\big((1+\zeta(\kappa,D))^2\,\varepsilon^{-2}\big)$$

The total complexity needed is

$$\sum_{k=1}^{N_\varepsilon}\sum_{i=1}^{m} N_k^{(i)} = \sum_{k=1}^{N_\varepsilon}\sum_{i=1}^{m} N_k^{(i)} = O(N_\varepsilon^2) = \widetilde{O}\big((1+\zeta(\kappa,D))^4\,\varepsilon^{-4}\big).$$

■

6.5 RBMM with g-smooth surrogates

In this section, we prove Theorems 4, 5, and 10 for smooth surrogates. Throughout this section, we assume (A1)**(i-gs)**, that is, the surrogates are g-smooth.

One of the most important tools in our analysis for the general smooth surrogate case is stated in Proposition 33. Recall that in Proposition 24, we have shown that the surrogate gaps $h_n^{(i)}(\theta_n^{(i)})$ are summable. In the next proposition, we show that this is enough to deduce that the norm of the Riemannian gradient of the surrogate gaps is also summable. This will be used later to deduce asymptotic optimality with respect to the objective function from that with respect to the surrogates.

Proposition 33 (Bound on surrogate optimality gap) *Assume (A0), (A1)**(i-gs), (ii)**. Then for any sequence of nonnegative real numbers $(w_n)_{n\geq 1}$,*

$$\sum_{n=1}^{N}\sum_{i=1}^{m} w_n \|\mathrm{grad}\, g_n^{(i)}(\theta_n^{(i)}) - \mathrm{grad}\, f_n^{(i)}(\theta_n^{(i)})\|^2 \leq 2(L_f+L_g)\sum_{n=1}^{N}\sum_{i=1}^{m} w_n (g_n^{(i)}(\theta_n^{(i)}) - f_n^{(i)}(\theta_n^{(i)})).$$

Proof Denote $\alpha_n^{(i)} = \mathrm{grad}\, g_n^{(i)}(\theta_n^{(i)}) - \mathrm{grad}\, f_n^{(i)}(\theta_n^{(i)})$. Fix a tangent vector $\eta \in T_{\theta_n^{(i)}}\mathcal{M}^{(i)}$ and $\varepsilon > 0$. By Lemma 38, we can write

$$\left| g_n^{(i)}(\mathrm{Exp}_{\theta_n^{(i)}}(\epsilon\eta)) - g_n^{(i)}(\theta_n^{(i)}) - \langle \nabla g_n^{(i)}(\theta_n^{(i)}), \varepsilon\eta\rangle \right| \leq \frac{L_g\varepsilon^2}{2}\|\eta\|^2,$$
$$\left| f_n^{(i)}(\mathrm{Exp}_{\theta_n^{(i)}}(\epsilon\eta)) - f_n^{(i)}(\theta_n^{(i)}) - \langle \nabla f_n^{(i)}(\theta_n^{(i)}), \varepsilon\eta\rangle \right| \leq \frac{L_f\varepsilon^2}{2}\|\eta\|^2,$$

for constants $L_f, L_g \geq 0$ in (A1). Hence

$$\begin{aligned} -\frac{L_f\varepsilon^2}{2}\|\eta\|^2 + f_n^{(i)}(\theta_n^{(i)}) + \langle \nabla f_n^{(i)}(\theta_n^{(i)}), \varepsilon\eta\rangle &\leq f_n^{(i)}(\theta_n^{(i)} + \varepsilon\eta) \\ &\leq g_n^{(i)}(\theta_n^{(i)} + \varepsilon\eta) \\ &\leq g_n^{(i)}(\theta_n^{(i)}) + \langle \nabla g_n^{(i)}(\theta_n^{(i)}, \varepsilon\eta\rangle + \frac{L_g\varepsilon^2}{2}\|\eta\|^2. \end{aligned}$$

Thus, denoting $c = (L_f + L_g)/2$, we obtain the following inequality

$$\langle \alpha_n, \varepsilon\eta\rangle \geq f_n^{(i)}(\theta_n^{(i)}) - g_n^{(i)}(\theta_n^{(i)}) - c\varepsilon^2\|\eta\|^2.$$

Choosing $\eta = -\alpha_n$, we get

$$-\varepsilon\|\alpha_n\|^2 \geq f_n^{(i)}(\theta_n^{(i)}) - g_n^{(i)}(\theta_n^{(i)}) - c\varepsilon^2\|\alpha_n\|^2.$$

Rearranging, we get

$$(\varepsilon - c\varepsilon^2)\|\alpha_n\|^2 \leq g_n^{(i)}(\theta_n^{(i)}) - f_n^{(i)}(\theta_n^{(i)}).$$

Then the assertion follows by multiplying both sides of the above inequality by w_n and summing up in $i = 1, \ldots, m$ and then for $n \geq 1$. ■

6.6 Asymptotic stationarity

In this section, we prove Theorems 4 and 5. We start with Theorem 4 where the number of blocks is $m \leq 2$. Below we focus on the proof when $m = 2$. In fact, the proof of the single block case ($m = 1$) follows from a similar analysis and is much simpler.

For conciseness, denote $w(k,i) = (\theta_k^{(1)}, \ldots, \theta_k^{(i-1)}, \theta_k^{(i)}, \theta_{k-1}^{(i+1)}, \ldots, \theta_{k-1}^{(m)})$, which is the variable value in the k-th cycle after updating i-th block. Denote $w(k,i)^{(j)}$ as the j-th component of $w(k,i)$. The following observations would be helpful: $w(k,i)^{(i+1)} = \theta_{k-1}^{(i+1)}$, $w(k,i+1)^{(i+1)} = \theta_k^{(i+1)}$. The following proposition gives an observation related to the limit behavior of surrogate values.

Proposition 34 *Assume (A1)**(i-gs), (ii)** and (A0). Suppose that for some $i \in \{1, \ldots, m\}$ the sequence $\{w(k,i)\}$ admits a limit point $\bar{w}$and the subsequence converging to $\bar{w}$ is denoted by $w(k(n),i)$, $n \geq 1$. Then we have*

$$\lim_{n\to\infty} g_{k(n)^+}^{(i^+)}\left(w(k(n),i)^{(i^+)}\right) = \lim_{n\to\infty} g_{k(n)^+}^{(i^+)}\left(w(k(n),i^+)^{(i^+)}\right) = f(\bar{w}),$$

where $i^+ = i (\operatorname{mod} m) + 1$, $k(n)^+ = k(n)$ or $k(n)+1$ for $i \in \{1, \cdots, m-1\}$ or $i = m$, respectively.

Proof For conciseness, let $i \in \{1, \cdots, m-1\}$ so that $i^+ = i+1$. Note that the proof for $i = m$ would be exactly the same after modifying the notation.

By Prop. 24 along with the tightness and upper-boundedness of surrogate functions, we have the following inequalities,

$$\begin{aligned} f\left(w(k(n),i)\right) &= f_{k(n)}^{(i+1)}\left(\theta_{k(n)-1}^{(i+1)}\right) = g_{k(n)}^{(i+1)}\left(\theta_{k(n)-1}^{(i+1)}\right) = g_{k(n)}^{(i+1)}\left(w(k(n),i)^{(i+1)}\right) \geq g_{k(n)}^{(i+1)}\left(\theta_{k(n)}^{(i\star)}\right) \\ &\geq g_{k(n)}^{(i+1)}\left(w(k(n),i+1)^{(i+1)}\right) - \Delta_{k(n)} \geq f_{k(n)}^{(i+1)}\left(w(k(n),i+1)^{(i+1)}\right) - \Delta_{k(n)} \\ &= f\left(w(k(n),i+1)\right) - \Delta_{k(n)} \geq f\left(w(k(n+1),i)\right) - m \sum_{i=k(n)}^{k(n+1)} \Delta_i. \end{aligned}$$

Recall

$$\lim_{n\to\infty} f\left(w(k(n),i)\right) = \lim_{n\to\infty} f\left(w(k(n+1),i)\right) = f(\bar{w}) \quad \text{and} \quad \lim_{n\to\infty} \sum_{i=n}^{\infty} \Delta_i = 0.$$

This completes the proof. ∎

Proposition 35 (Asymptotic first order optimality of inexact solutions) *Suppose (A1)**(i-gs), (ii)**, and (A0) hold. Consider any $\theta^{(i)} \in \boldsymbol{\Theta}_\infty^{(i)}$ (see (55)) such that $\eta_\infty^{(i)}$ defined in (56) satisfies $\|\eta_\infty^{(i)}\| \leq 1$. Let $\eta_n^{(i\star)} \in T_{\theta_n^{(i\star)}}$ and $\eta_n^{(i)} \in T_{\theta_n^{(i)}}$ be the tangent vectors defined in (56)). For each $i = 1, \cdots, m$, the following holds:*

$$\left|\left\langle \operatorname{grad} g_n^{(i)}(\theta_n^{(i\star)}), \eta_n^{(i\star)} \right\rangle - \left\langle \operatorname{grad} g_n^{(i)}(\theta_n^{(i)}), \eta_n^{(i)} \right\rangle\right| = o(1).$$

Proof Let $\boldsymbol{\theta}_\infty = [\theta_\infty^{(1)}, \cdots, \theta_\infty^{(m)}]$ be a limit point of $(\boldsymbol{\theta}_n)_n$. For simplicity of the notation, replace $(\boldsymbol{\theta}_n)_n$ by the convergent subsequence.

First by triangle inequality and (A0),

$$d(\theta_n^{(i\star)}, \theta_\infty^{(i)}) \le d(\theta_n^{(i)}, \theta_\infty^{(i)}) + d(\theta_n^{(i\star)}, \theta_n^{(i)}) = o(1), \tag{65}$$

so $\theta_n^{(i\star)} \to \theta_\infty^{(i)}$ as $n \to \infty$.

Let $T = m \sum_{k=1}^{\infty} \Delta_k < \infty$, by Prop. 24 we have

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} f(\boldsymbol{\theta}) \le f(\boldsymbol{\theta}_0) + T.$$

Now let $K = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : f(\boldsymbol{\theta}) \le f(\boldsymbol{\theta}_0) + T\}$. By (A0), K is compact. Denote $K = K^{(1)} \times \cdots \times K^{(m)}$ where $K^{(i)} \subseteq \mathcal{M}^{(i)}$, then $K^{(i)}$ is compact for each $i = 1, \cdots, m$. Since $g_n^{(i)}$ is geodesically smooth, $\| \operatorname{grad} g_n^{(i)}(\theta^{(i)}) \|$ is uniformly bounded for $\theta^{(i)} \in K^{(i)}$ by some constant, say $L_k > 0$.

Let $\Gamma_{\theta_n^{(i\star)} \to \theta_\infty^{(i)}}$ be the parallel transport along a minimal geodesic joining $\theta_n^{(i\star)}$ and $\theta_\infty^{(i)}$. By smoothness of the exponential map, we have we have $\Gamma_{\theta_n^{(i\star)} \to \theta_\infty^{(i)}} \eta_n^{(i\star)} \to \eta_\infty^{(i)} \in T_{\theta_\infty^{(i)}}$ where $\operatorname{Exp}_{\theta_\infty^{(i)}}(\eta_\infty^{(i)}) = \theta^{(i)}$.

Then

$$\begin{aligned}
&\left| \left\langle \operatorname{grad} g_n^{(i)}(\theta_n^{(i\star)}), \eta_n^{(i\star)} \right\rangle - \left\langle \operatorname{grad} g_n^{(i)}(\theta_\infty^{(i)}), \eta_\infty^{(i)} \right\rangle \right| \\
&\quad = \left| \left\langle \Gamma_{\theta_n^{(i\star)} \to \theta_\infty^{(i)}} \operatorname{grad} g_n^{(i)}(\theta_n^{(i\star)}), \Gamma_{\theta_n^{(i\star)} \to \theta_\infty^{(i)}} \eta_n^{(i\star)} \right\rangle - \left\langle \operatorname{grad} g_n^{(i)}(\theta_\infty^{(i)}), \eta_\infty^{(i)} \right\rangle \right| \\
&\quad \le \left| \left\langle \Gamma_{\theta_n^{(i\star)} \to \theta_\infty^{(i)}} \operatorname{grad} g_n^{(i)}(\theta_n^{(i\star)}), \Gamma_{\theta_n^{(i\star)} \to \theta_\infty^{(i)}} \eta_n^{(i\star)} \right\rangle - \left\langle \operatorname{grad} g_n^{(i)}(\theta_\infty^{(i)}), \Gamma_{\theta_n^{(i\star)} \to \theta_\infty^{(i)}} \eta_n^{(i\star)} \right\rangle \right| \\
&\quad + \left| \left\langle \operatorname{grad} g_n^{(i)}(\theta_\infty^{(i)}), \Gamma_{\theta_n^{(i\star)} \to \theta_\infty^{(i)}} \eta_n^{(i\star)} \right\rangle - \left\langle \operatorname{grad} g_n^{(i)}(\theta_\infty^{(i)}), \eta_\infty^{(i)} \right\rangle \right| \\
&\quad \le d(\theta_n^{(i\star)}, \theta_\infty^{(i)}) + \| \operatorname{grad} g_n^{(i)}(\theta_\infty^{(i)}) \| \cdot \left\| \Gamma_{\theta_n^{(i\star)} \to \theta_\infty^{(i)}} \eta_n^{(i\star)} - \eta_\infty^{(i)} \right\| \\
&\quad = o(1),
\end{aligned}$$

where the last inequality is by g-smoothness of $g_n^{(i)}$, the last equality is by (65) and the boundedness of $\| \operatorname{grad} g_n^{(i)}(\theta_\infty^{(i)}) \|$.

Similarly, we also have $\left| \left\langle \operatorname{grad} g_n^{(i)}(\theta_n^{(i)}), \eta_n^{(i)} \right\rangle - \left\langle \operatorname{grad} g_n^{(i)}(\theta_\infty^{(i)}), \eta_\infty^{(i)} \right\rangle \right| = o(1)$. Therefore we have

$$\begin{aligned}
&\left| \left\langle \operatorname{grad} g_n^{(i)}(\theta_n^{(i\star)}), \eta_n^{(i\star)} \right\rangle - \left\langle \operatorname{grad} g_n^{(i)}(\theta_n^{(i)}), \eta_n^{(i)} \right\rangle \right| \\
\le& \left| \left\langle \operatorname{grad} g_n^{(i)}(\theta_n^{(i\star)}), \eta_n^{(i\star)} \right\rangle - \left\langle \operatorname{grad} g_n^{(i)}(\theta_\infty^{(i)}), \eta_\infty^{(i)} \right\rangle \right| \\
&+ \left| \left\langle \operatorname{grad} g_n^{(i)}(\theta_n^{(i)}), \eta_n^{(i)} \right\rangle - \left\langle \operatorname{grad} g_n^{(i)}(\theta_\infty^{(i)}), \eta_\infty^{(i)} \right\rangle \right| \\
=& o(1).
\end{aligned}$$

■

Now we show the following proposition, which is the Riemannian version of Prop. 3 in Grippo and Sciandrone (2000).

Proposition 36 *Assume (A1)**(i-gs), (ii)**, and (A0). Further assume that the constraint set $\Theta^{(i)}$ is strongly convex in $\mathcal{M}^{(i)}$ for $i = 1, \dots, m$. Suppose that for some fixed $i \in \{1, \dots, m\}$ the sequence $(w(k,i))_{k\geq 1}$ admits a limit point $\bar{w}$ and the subsequence converging to $\bar{w}$ is denoted by $w(k(n), i)$, $n \geq 1$. Denote $i^+ = i (\operatorname{mod} m) + 1$, $k(n)^+ = k(n)$ or $k(n)+1$ for $i \in \{1, \cdots, m-1\}$ or $i = m$ respectively and $w(k(n), i)^{(i)}$ as the i-th coordinate of $w(k(n), i)$. For $n \gg 1$, consider any $\theta^{(i)} \in \Theta^{(i)}$ such that $d(\theta^{(i)}, \bar{w}^{(i)}) \leq r_0/2$ and $\theta^{(i^+)} \in \Theta^{(i^+)}$ such that $d(\theta^{(i^+)}, \bar{w}^{(i^+)}) \leq r_0/2$, denote $\eta_n^{(i)} \in T_{\theta_n^{(i)}}$ such that $\operatorname{Exp}_{\theta_n^{(i)}}(\eta_n^{(i)}) = \theta^{(i)}$ and $\eta_n^{(i^+)} \in T_{w(n,i)^{(i^+)}}$ such that $\operatorname{Exp}_{w(n,i)^{(i^+)}}(\eta_n^{(i^+)}) = \theta^{(i^+)}$. Then we have*

$$\begin{aligned}\lim_{n\to\infty} \left\langle \operatorname{grad} g_{k(n)}^{(i)} \left(w(k(n), i)^{(i)} \right), \eta_{k(n)}^{(i)} \right\rangle \geq 0 \\ \textit{and} \quad \liminf_{n\to\infty} \left\langle \operatorname{grad} g_{k(n)^+}^{(i^+)} \left(w(k(n), i)^{(i^+)} \right), \eta_{k(n)}^{(i^+)} \right\rangle \geq 0.\end{aligned} \tag{66}$$

Proof Firstly, by the first-order optimality condition of $\theta_{k(n)}^{(i\star)}$ with respect to $g_{k(n)}^{(i)}$, for any $\theta^{(i)} \in \Theta^{(i)}$, we have

$$\left\langle \operatorname{grad} g_{k(n)}^{(i)} \left(\theta_{k(n)}^{(i\star)} \right), \eta_{k(n)}^{(i\star)} \right\rangle \geq 0, \quad \text{for all } n,$$

where $\eta_n^{(i\star)} \in T_{\theta_n^{(i\star)}}$ such that $\operatorname{Exp}_{\theta_n^{(i\star)}}(\eta_n^{(i\star)}) = \theta^{(i)}$. Then by Prop. 35, we get the first identity of (66).

In order to prove the second identity in (66), suppose towards a contradiction that there exists $\theta^{i^+} \in \Theta^{(i^+)}$ such that

$$\liminf_{n\to\infty} \left\langle \operatorname{grad} g_{k(n)^+}^{(i^+)} \left(w(k(n), i)^{(i^+)} \right), \eta_{k(n)}^{(i^+)} \right\rangle < 0.$$

Then there exists an infinite index set $K_1 \subset K = \{k(n) : n \geq 1\}$ such that for any $k \in K_1$,

$$\left\langle \operatorname{grad} g_{k(n)^+}^{(i^+)} \left(w(k(n), i)^{(i^+)} \right), \eta_{k(n)}^{(i^+)} \right\rangle < 0.$$

For conciseness, from now on we only consider the case $i \in \{1, \cdots, m-1\}$, then $i^+ = i + 1$. Note that when $i = m$, the following proof would be exactly the same after some modification of notation.

Let the search direction

$$d_k^{(i+1)} = \begin{cases} \eta_k^{(i+1)} & \text{if } \|\eta_k^{(i+1)}\| \leq 1 \\ \eta_k^{(i+1)} / \|\eta_k^{(i+1)}\| & \text{if } \|\eta_k^{(i+1)}\| > 1. \end{cases}$$

Now for all $k \in K_1$ suppose we compute the step size $\alpha_k^{(i+1)}$ by the line search algorithm 2, we have

$$\begin{aligned} g_k^{(i+1)} \left(\operatorname{Exp}_{w(k,i)^{(i+1)}}(\alpha_k^{(i+1)} d_k^{(i+1)}) \right) &= g_k^{(i+1)} \left(\operatorname{Exp}_{\theta_{k-1}^{(i+1)}}(\alpha_k^{(i+1)} d_k^{(i+1)}) \right) \\ &\leq g_k^{(i+1)} \left(\theta_{k-1}^{(i+1)} \right) = g_k^{(i+1)} \left(w(k,i)^{(i+1)} \right). \end{aligned}$$

Note here we have $\operatorname{Exp}_{\theta_{k-1}^{(i+1)}}(\alpha_k^{(i+1)} d_k^{(i+1)}) \in \boldsymbol{\Theta}^{(i+1)}$ by geodesic convexity of $\boldsymbol{\Theta}^{(i+1)}$.

Recall by definition, the optimality of $\theta_k^{(i+1)}$ gives

$$g_k^{(i+1)}\left(w(k,i+1)^{(i+1)}\right)-\Delta_n\leq g_k^{(i+1)}\left(\theta_k^{(i+1\star)}\right)=\min_{\theta\in\boldsymbol{\Theta}^{(i+1)}}g_k^{(i+1)}(\theta).$$

We can then write

$$g_k^{(i+1)}\left(w(k,i+1)^{(i+1)}\right)-\Delta_n\leq g_k^{(i+1)}\left(\operatorname{Exp}_{w(k,i)^{(i+1)}}(\alpha_k^{(i+1)}d_k^{(i+1)})\right)\leq g_k^{(i+1)}\left(w(k,i)^{(i+1)}\right).$$

Then by noting that $k\in K_1\subset K=\{k(n):n\geq 1\}$ and use Proposition 34 we get

$$\lim_{k\to\infty}g_k^{(i+1)}\left(w(k,i)^{(i+1)}\right)-g_k^{(i+1)}\left(\operatorname{Exp}_{w(k,i)^{(i+1)}}(\alpha_k^{(i+1)}d_k^{(i+1)})\right)=0.$$

Finally by Proposition 42 we have

$$\lim_{k\to\infty}\operatorname{grad}g_k^{(i+1)}\left(w(k,i)^{(i+1)}\right)=\mathbf{0},$$

which gives a contradiction. ■

The preceding proposition indicates that every limit point of the sequence generated by Algorithm 1 is a critical point with respect to the first and last component, i.e., $\theta^{(1)}$ and $\theta^{(m)}$, which is formally stated below.

Corollary 37 *Under the same assumptions as in Prop. 36. Let $\boldsymbol{\theta}_n$ be a sequence generated by Algorithm 1 which admits a limit point $\boldsymbol{\theta}_\infty$, denote the subsequence converging to $\boldsymbol{\theta}_\infty$ by $\boldsymbol{\theta}_{n_k}$, then for any $\theta^{(m)}\in\Theta_\infty^{(m)}$ and $\theta^{(1)}\in\Theta_\infty^{(1)}$*

$$\lim_{k\to\infty}\langle\operatorname{grad}g_{n_k}^{(m)}\left(\theta_{n_k}^{(m)}\right),\eta_{n_k}^{(m)}\rangle\geq 0\quad\textit{and}\quad\liminf_{k\to\infty}\langle\operatorname{grad}g_{n_k+1}^{(1)}\left(\theta_{n_k}^{(1)}\right),\eta_{n_k}^{(1)}\rangle\geq 0.$$

We are now ready to give a proof of Theorem 4 for the case (A1)**(i-gs)** of g-smooth surrogates on general manifolds.

Proof of Theorem 4 for g-smooth surrogates. Assume (A1)**(i-gs)**, **(ii)**, and (A0). Further assume that the constraint set $\Theta^{(i)}$ is strongly convex in $\mathcal{M}^{(i)}$ for $i=1,\dots,m$. Fix a convergent subsequence $(\boldsymbol{\theta}_{k_n})_{k\geq 1}$ of $(\boldsymbol{\theta}_n)_{n\geq 1}$. We wish to show that $\boldsymbol{\theta}_\infty=\lim_{k\to\infty}\boldsymbol{\theta}_{k_n}$ is a stationary point of f over $\boldsymbol{\Theta}$ when $m=2$. The proof of single block case ($m=1$) follows from the same analysis and is much simpler, which can be done without Corollary 37. Below we omit the $m=1$ case and focus on proving it for $m=2$.

First, we apply Proposition 33 with $w_n\equiv 1$ and Proposition 24 **(ii)** to deduce that

$$\sum_{i=1}^{m}\|\operatorname{grad}g_n^{(i)}(\theta_n^{(i)})-\operatorname{grad}f_n^{(i)}(\theta_n^{(i)})\|^2=o(1).\tag{67}$$

Hence, by Corollary 37, for any $\theta^{(1)}\in\boldsymbol{\Theta}_\infty^{(1)}$ and $\theta^{(m)}\in\boldsymbol{\Theta}_\infty^{(m)}$ we get

$$\liminf_{k\to\infty}\langle\operatorname{grad}f_{n_k+1}^{(1)}(\theta_{n_k}^{(1)}),\eta_{n_k}^{(1)}\rangle\geq 0\qquad\textit{and}\qquad\liminf_{k\to\infty}\langle\operatorname{grad}f_{n_k}^{(m)}(\theta_{n_k}^{(m)}),\eta_{n_k}^{(m)}\rangle\geq 0,$$

where $\mathrm{Exp}_{\theta^{(m)}_{n_k}}(\eta^{(m)}_{n_k}) = \theta^{(m)}$ and $\mathrm{Exp}_{\theta^{(1)}_{n_k}}(\eta^{(1)}_{n_k}) = \theta^{(1)}$.

Note that $\mathrm{grad}\, f^{(1)}_{n_k+1}(\boldsymbol{\theta}^{(1)}_{n_k}) = \mathrm{grad}_1\, f(\boldsymbol{\theta}_{n_k})$ and $\mathrm{grad}\, f^{(m)}_{n_k}(\boldsymbol{\theta}^{(m)}_{n_k}) = \mathrm{grad}_m\, f(\boldsymbol{\theta}_{n_k})$. Thus for the case when $m = 2$, by continuity of Riemannian metric and the continuity of $\mathrm{grad}\, f$, we verified for each $i = 1, 2$ and any $\theta \in \Theta^{(i)}_\infty$ we have $\langle \mathrm{grad}_i\, f(\boldsymbol{\theta}_\infty), \eta \rangle \ge 0$, where $\mathrm{Exp}_{\theta^{(i)}_\infty}(\eta) = \theta$. This shows $\boldsymbol{\theta}_\infty$ is a stationary point of f over $\boldsymbol{\Theta}$, as desired. ■

Next, we prove Theorem 5 for smooth surrogates.

Proof of Theorem 5 for g-smooth surrogates. Suppose (A1)**(i-gs), (ii), (iii)** and (A0) hold. Futher assume p is lower semi-continuous on $\boldsymbol{\Theta}$. Fix a convergent subsequence $(\boldsymbol{\theta}_{n_k})_{k\ge1}$ of $(\boldsymbol{\theta}_n)_{n\ge1}$. We wish to show that $\boldsymbol{\theta}_\infty = \lim_{k\to\infty} \boldsymbol{\theta}_{n_k}$ is a stationary point of f over $\boldsymbol{\Theta}$. First, we apply Proposition 33 with $w_n \equiv 1$ and Proposition 24 **(ii)** to deduce that

$$\sum_{n=1}^{\infty}\sum_{i=1}^{m} \phi\left(d\left(\theta^{(i)}_{n-1}, \theta^{(i)}_n\right)\right) < \infty, \qquad \sum_{n=1}^{\infty}\sum_{i=1}^{m} \|\mathrm{grad}\, g^{(i)}_n(\theta^{(i)}_n) - \mathrm{grad}\, f^{(i)}_n(\theta^{(i)}_n)\|^2 < \infty.$$

In particular, this yields

$$\sum_{i=1}^{m} \phi\left(d\left(\theta^{(i)}_{n-1}, \theta^{(i)}_n\right)\right) = o(1), \qquad \sum_{i=1}^{m} \|\mathrm{grad}\, g^{(i)}_n(\theta^{(i)}_n) - \mathrm{grad}\, f^{(i)}_n(\theta^{(i)}_n)\|^2 = o(1). \tag{68}$$

Fix $\theta^{(i)} \in \Theta^{(i)}_\infty$ (see (55)) such that $\eta^{(i)}_\infty$ defined in (56) satisfies $\|\eta^{(i)}_\infty\| \le 1$. Let $\eta^{(i\star)}_n \in T_{\theta^{(i\star)}_n}$ and $\eta^{(i)}_n \in T_{\theta^{(i)}_n}$ be the tangent vectors defined in (56)). Since $\theta^{(i\star)}_n$ minimizes $G^{(i)}_n$ over $\Theta^{(i)}$, we have

$$\langle \mathrm{grad}\, g^{(i)}_n(\theta^{(i\star)}_n), \eta^{(i\star)}_n \rangle + p^{(i)}_{n_k}(\theta^{(i)}) - p^{(i)}_{n_k}(\theta^{(i\star)}_{n_k}) \ge 0.$$

Note by Prop. 35 and since p is block-separable due to (A0), we have

$$\liminf_{k\to\infty} \langle \mathrm{grad}\, g^{(i)}_{n_k}(\theta^{(i)}_{n_k}), \eta^{(i)}_{n_k} \rangle + p^{(i)}(\theta^{(i)}) - p^{(i)}(\theta^{(i)}_{n_k}) \ge 0.$$

Then recall the second part of (68), $\|\mathrm{grad}\, g^{(i)}_n(\theta^{(i)}_n) - \mathrm{grad}\, f^{(i)}_n(\theta^{(i)}_n)\| = o(1)$. Note since $\theta^{(i)}_{n_k} \to \theta^{(i)}_\infty$ and $\|\eta^{(i)}_\infty\| \le 1$, we have $\|\eta^{(i)}_{n_k}\|$ is uniformly bounded by some constant $C_0 > 1$. Hence, for each $i = 1, \dots, m$, we get

$$\left\| \left\langle \mathrm{grad}\, g^{(i)}_n(\theta^{(i)}_n), \eta^{(i)}_n \right\rangle - \left\langle \mathrm{grad}\, f^{(i)}_n(\theta^{(i)}_n), \eta^{(i)}_n \right\rangle \right\| \le C_0 \|\, \mathrm{grad}\, g^{(i)}_n(\theta^{(i)}_n) - \mathrm{grad}\, f^{(i)}_n(\theta^{(i)}_n)\| = o(1),$$

so $\liminf\limits_{k\to\infty} \left\langle \mathrm{grad}\, f^{(i)}_{n_k}(\theta^{(i)}_{n_k}), \eta^{(i)}_{n_k} \right\rangle + p^{(i)}(\theta^{(i)}) - p^{(i)}(\theta^{(i)}_{n_k}) \ge 0$.

Let $\Gamma_{\theta^{(i)}_{n_k} \to \theta^{(i)}_\infty}$ be the parallel transport along a minimal geodesic joining $\theta^{(i)}_{n_k}$ and $\theta^{(i)}_\infty$. By smoothness of exponential map, we have $\Gamma_{\theta^{(i)}_{n_k} \to \theta^{(i)}_\infty} \eta^{(i)}_{n_k} \to \eta^{(i)}_\infty \in T_{\theta^{(i)}_\infty}$. By the first part of (68) and the fact ϕ is strictly increasing, we have $d\left(\theta^{(i)}_{n-1}, \theta^{(i)}_n\right) = o(1)$ and hence $\theta^{(j)}_{n_k-1} \to \theta^{(j)}_\infty$ as $k \to \infty$ for all $j = 1, \dots, m$. Hence by continuity of $\mathrm{grad}\, f$ and lower semi-continuity of p, we get

$$\left\langle \mathrm{grad}_i\, f(\theta^{(1)}_\infty, \dots, \theta^{(i-1)}_\infty, \theta^{(i)}_\infty, \theta^{(i+1)}_\infty, \dots, \theta^{(m)}_\infty), \eta^{(i)}_\infty \right\rangle + p^{(i)}(\theta^{(i)}) - p^{(i)}(\theta^{(i)}_\infty)$$

$$
\begin{aligned}
&\geq \liminf_{k\to\infty} \left\langle \Gamma_{\theta^{(i)}_{n_k}\to\theta^{(i)}_{\infty}} \operatorname{grad} f^{(i)}_{n_k}(\theta^{(i)}_{n_k}), \Gamma_{\theta^{(i)}_{n_k}\to\theta^{(i)}_{\infty}} \eta^{(i)}_{n_k} \right\rangle + p^{(i)}(\theta^{(i)}) - p^{(i)}(\theta^{(i)}_{n_k}) \\
&= \liminf_{k\to\infty} \left\langle \operatorname{grad} f^{(i)}_{n_k}(\theta^{(i)}_{n_k}), \eta^{(i)}_{n_k} \right\rangle + p^{(i)}(\theta^{(i)}) - p^{(i)}(\theta^{(i)}_{n_k}) \geq 0.
\end{aligned}
$$

Since this holds for any $\theta^{(i)} \in \Theta^{(i)}_{\infty}$ with $d(\theta^{(i)}, \theta^{(i)}_{\infty}) \leq 1$ and also holds for any $i = 1, \cdots, m$, we conclude

$$
V(\boldsymbol{\theta}_{\infty}, \boldsymbol{\eta}) \geq 0, \quad \forall \boldsymbol{\eta} \quad \text{such that} \quad \eta^{(i)} \in T^{\Theta^{(i)}}_{\theta^{(i)}_{\infty}} \quad \text{with} \quad \|\eta^{(i)}\| \leq 1,
$$

which means $\boldsymbol{\theta}_{\infty}$ is a stationary point of F in $\boldsymbol{\Theta}$, as desired. ■

6.7 Rate of convergence with g-smooth surrogates

In this section, we prove Theorem 10.

Proof of Theorem 10. We first show **(i)**. By Prop. 27, Prop. 24 and using (59), we have

$$
\begin{aligned}
\sum_{n=0}^{N} b_{n+1} \left[\sup_{\eta^{(i)} \in T^{\Theta^{(i)}}_{\theta^{(i)}_k}, \|\eta\| \leq 1} V(\boldsymbol{\theta}_k, \boldsymbol{\eta}) \right] \leq & C \left(\sum_{n=0}^{N} b^2_{n+1} + \sqrt{\sum_{n=0}^{N} b^2_{n+1} \sum_{k=1}^{N} d^2(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k+1})} + \sum_{n=1}^{N} \Delta_n(\boldsymbol{\theta}_0) \right. \\
& \left. + (F(\boldsymbol{\theta}_0) - F^*) + \sum_{k=1}^{N} d^2(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k+1}) \right).
\end{aligned}
$$

for some constant $C > 0$ independent of $\boldsymbol{\theta}_0$ and the right hand side is finite.

Take $b_n = \frac{\hat{r}}{\sqrt{n}\log(1+n)}$. Then $\sum_n b_n^2 < 5\hat{r}^2$, $\sum_{k=1}^{n} b_k \sim \hat{r} n^{1/2}/\log n$. Hence by using Lemma 40, it follows that there exists some $M > 0$ such that for all $n \geq 1$,

$$
\min_{1 \leq k \leq n} \left[\sup_{\eta^{(i)} \in T^{\Theta^{(i)}}_{\theta^{(i)}_k}, \|\eta\| \leq 1} V(\boldsymbol{\theta}_k, \boldsymbol{\eta}) \right] \leq \frac{M}{n^{1/2}/\log n},
$$

where

$$
\begin{aligned}
M = \hat{r}^{-1} \left(L_f C_{\phi}^{-1} + 1 \right) \left(F(\boldsymbol{\theta}_0) - F^* + m \sum_{n=1}^{N} \Delta_n \right) + \frac{m L_g}{2} 5\hat{r} \qquad (69) \\
+ m L_f C_{\phi}^{-1/2} \sqrt{5} \left(F(\boldsymbol{\theta}_0) - F^* + m \sum_{n=1}^{N} \Delta_n \right)^{1/2}.
\end{aligned}
$$

This shows **(i)**. Then we could conclude **(ii)** by using the fact that $n \geq 2\varepsilon^{-2} \left(\log \varepsilon^{-2}\right)^2$ implies $\log n/\sqrt{n} \leq \varepsilon$ for all sufficiently small $\varepsilon > 0$. This completes the proof. ■

7. Concluding Remark

In this paper, we develop and analyze RBMM, a general framework for solving nonsmooth nonconvex block Riemannian optimization problems. We establish the iteration complexity of RBMM with different types of surrogates, which incorporate a wide range of algorithms, including block Riemannian prox-linear updates, Bures-JKO scheme for Wasserstein variational inference, geodesically constrained subspace tracking, optimistic likelihood estimation, robust PCA, and Riemannian CP-dictionary-learning. While we allow only cyclic updates in RBMM, it is natural to consider other update rules, e.g., randomized updating rule with i.i.d. sampling, which has been extensively studied in Euclidean MM literature Hong et al. (2015); Razaviyayn et al. (2014). We believe that extending our results to the randomized updates is straightforward. Most key lemmas will remain nearly identical, with the primary difference being taking conditional expectations of the result. In order to maintain the focus of the paper, we do not pursue that direction.

In the Euclidean MM literature, certain types of acceleration have been explored in recent papers (Hien et al., 2023). Acceleration techniques for Riemannian optimization methods have also been investigated (Ahn and Sra, 2020; Zhang and Sra, 2018; Kong and Tao, 2024). Studying the acceleration of RBMM is an interesting topic, which we leave as a direction for future work.

Acknowledgments

YL is partially supported by the Institute for Foundations of Data Science RA fund through NSF Award DMS-2023239 and by the National Science Foundation through grants DMS-2206296. HL is partially supported by the National Science Foundation through grants DMS-2206296 and DMS-2010035. DN is partially supported by NSF DMS-2011140 and NSF DMS-2108479. LB is partially supported by NSF CAREER award CCF-1845076 and NSF award CCF-2331590.

References

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

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009. ISBN 9781400830244. URL https://books.google.com/books?id=NSQGQeLN3NcC.

P-A Absil, Robert Mahony, and Jochen Trumpf. An extrinsic look at the riemannian hessian. In *International conference on geometric science of information*, pages 361–368. Springer, 2013.

Bijan Afsari. Riemannian l_p center of mass: Existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139(2):655–673, 2011.

Kwangjun Ahn and Suvrit Sra. From nesterov's estimate sequence to riemannian acceleration. In *Conference on Learning Theory*, pages 84–118. PMLR, 2020.

Foivos Alimisis, Antonio Orvieto, Gary Bécigneul, and Aurelien Lucchi. A continuous-time perspective for modeling acceleration in riemannian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1297–1307. PMLR, 2020.

Aleksandr Aravkin, Stephen Becker, Volkan Cevher, and Peder Olsen. A variational approach to stable principal component pursuit. *arXiv preprint arXiv:1406.1089*, 2014.

C. Atkinson and A.F. Mitchell. Rao's distance measure. *Sankhyā: The Indian Journal of Statistics, Series A*, 43(3):345–365, 1981.

Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of operations research*, 35(2):438–457, 2010.

Daniel Azagra, Juan Ferrera, and Fernando López-Mesas. Nonsmooth analysis and hamilton–jacobi equations on riemannian manifolds. *Journal of Functional Analysis*, 220(2):304–361, 2005.

Miroslav Bacak. *Convex analysis and optimization in Hadamard spaces*. De Gruyter, 2014. ISBN 9783110361629. doi: doi:10.1515/9783110361629.

Christopher G Baker, P-A Absil, and Kyle A Gallivan. An implicit trust-region method on riemannian manifolds. *IMA journal of numerical analysis*, 28(4):665–689, 2008.

Amir Beck and Luba Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM journal on Optimization*, 23(4):2037–2060, 2013.

G.C. Bento, O.P. Ferreira, and P.R. Oliveira. Proximal point method for a special class of nonconvex functions on hadamard manifolds. *Optimization*, 64(2):289–319, 2015. doi: 10.1080/02331934.2012.745531.

Glaydston C Bento, Orizon P Ferreira, and Jefferson G Melo. Iteration-complexity of gradient, subgradient and proximal point methods on riemannian manifolds. *Journal of Optimization Theory and Applications*, 173(2):548–562, 2017.

Cameron J Blocker, Haroon Raja, Jeffrey A Fessler, and Laura Balzano. Dynamic subspace estimation with grassmannian geodesics. *arXiv preprint arXiv:2303.14851*, 2023.

N. Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

Nicolas Boumal, Pierre-Antoine Absil, and Coralia Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019.

Arnaud Breloy, Sandeep Kumar, Ying Sun, and Daniel P Palomar. Majorization-minimization on the stiefel manifold with application to robust sparse pca. *IEEE Transactions on Signal Processing*, 69:1507–1520, 2021.

Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A Course in Metric Geometry*. Crm Proceedings & Lecture Notes. American Mathematical Society, 2001. ISBN 9780821821299. URL https://books.google.com/books?id=dRmIAwAAQBAJ.

Yu Burago, M Gromov, and G Perel'man. A.d. alexandrov spaces with curvature bounded below. *Russian Mathematical Surveys*, 47(2):1, apr 1992.

Donald Bures. An extension of kakutani's theorem on infinite product measures to the tensor product of semifinite w*-algebras. *Transactions of the American Mathematical Society*, 135:199–212, 1969.

Emmanuel J. Candes, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *arXiv preprint arXiv:0912.3599*, 2009. URL https://arxiv.org/abs/0912.3599.

I. Chavel. *Riemannian Geometry: A Modern Introduction*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2006.

Jeff Cheeger and Detlef Gromoll. On the structure of complete manifolds of nonnegative curvature. *Annals of Mathematics*, 96(3):413–443, 1972.

Shixiang Chen, Shiqian Ma, Anthony Man-Cho So, and Tong Zhang. Proximal gradient method for nonsmooth optimization over the stiefel manifold. *SIAM Journal on Optimization*, 30(1):210–239, 2020.

Shixiang Chen, Alfredo Garcia, Mingyi Hong, and Shahin Shahrampour. Decentralized riemannian gradient descent on the stiefel manifold. In *International Conference on Machine Learning*, pages 1594–1605. PMLR, 2021.

Manfredo Perdigao Do Carmo and J Flaherty Francis. *Riemannian geometry*, volume 6. Springer, 1992.

Shuyu Dong, Bin Gao, Yu Guan, and François Glineur. New riemannian preconditioned algorithms for tensor completion via polyadic decomposition. *SIAM Journal on Matrix Analysis and Applications*, 43(2):840–866, 2022.

Alan Edelman, T.A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 1998.

Gene Golub and Victor Perayra. Separable nonlinear least squares: the variable projection method and its applications. *Inverse Problems*, 19(2):R1, feb 2003.

Luigi Grippo and Marco Sciandrone. On the convergence of the block nonlinear gauss–seidel method under convex constraints. *Operations research letters*, 26(3):127–136, 2000.

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

Sigurdur Helgason. *Differential geometry, Lie groups, and symmetric spaces*. Academic press, 1979.

Le Thi Khanh Hien, Duy Nhat Phan, and Nicolas Gillis. An inertial block majorization minimization framework for nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 24(18):1–41, 2023.

Mingyi Hong, Meisam Razaviyayn, Zhi-Quan Luo, and Jong-Shi Pang. A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing. *IEEE Signal Processing Magazine*, 33(1):57–77, 2015.

Mingyi Hong, Xiangfeng Wang, Meisam Razaviyayn, and Zhi-Quan Luo. Iteration complexity analysis of block coordinate descent methods. *Mathematical Programming*, 163: 85–114, 2017.

Noémie Jaquier, Leonel Rozo, Sylvain Calinon, and Mathias Bürger. Bayesian optimization meets riemannian manifolds in robot learning. In *Conference on Robot Learning*, pages 233–246. PMLR, 2020.

Yiheng Jiang, Sinho Chewi, and Aram-Alexandre Pooladian. Algorithms for mean-field variational inference via polyhedral optimization in the wasserstein space. *arXiv preprint arXiv:2312.02849*, 2023.

Ian T Jolliffe. Principal components in regression analysis. In *Principal component analysis*, pages 129–155. Springer, 1986.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

Lingkai Kong and Molei Tao. Quantitative convergences of lie group momentum optimizers. In *Advances in Neural Information Processing Systems*, 2024.

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 35:14434–14447, 2022.

Kenneth Lange, David R Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions. *Journal of computational and graphical statistics*, 9(1):1–20, 2000.

J.M. Lee. *Introduction to Smooth Manifolds.* Graduate Texts in Mathematics. Springer, 2003.

Chong Li, Genaro López, and Victoria Martín-Márquez. Monotone vector fields and the proximal point algorithm on hadamard manifolds. *Journal of the London Mathematical Society*, 79(3):663–683, 2009.

Jian-Ze Li and Shu-Zhong Zhang. Polar decomposition-based algorithms on the product of stiefel manifolds with applications in tensor approximation. *Journal of the Operations Research Society of China*, pages 1–47, 2023.

Yuchen Li, Laura Balzano, Deanna Needell, and Hanbaek Lyu. Convergence and complexity guarantee for inexact first-order riemannian optimization algorithms. In *Proceedings of the 41st International Conference on Machine Learning*, pages 27376–27398, 2024.

Yuanyuan Liu, Fanhua Shang, James Cheng, Hong Cheng, and Licheng Jiao. Accelerated first-order methods for geodesically convex optimization on riemannian manifolds. *Advances in Neural Information Processing Systems*, 30, 2017.

Hanbaek Lyu and Yuchen Li. Block majorization-minimization with diminishing radius for constrained nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 35(2): 842–871, 2025.

Hanbaek Lyu, Christopher Strohmeier, and Deanna Needell. Online nonnegative cp-dictionary learning for markovian data. *The Journal of Machine Learning Research*, 23(1):6630–6679, 2022.

Yi Ma, Shankar Sastry, and Rene Vidal. *Generalized Principal Component Analysis*. Interdisciplinary Applied Mathematics. Springer New York, 2015. ISBN 9780387879253.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(Jan):19–60, 2010.

Bart Michels. Riemannian distances are locally equivalent. 2019.

Carmeliza Navasca, Lieven De Lathauwer, and Stefan Kindermann. Swamp reducing technique for tensor decomposition. In *2008 16th European Signal Processing Conference*, pages 1–5. IEEE, 2008.

Viet Nguyen, Soroosh Shafieezadeh-Abadeh, Man-Chung Yue, Daniel Kuhn, and Wolfram Wiesemann. Calculating optimistic likelihoods using (geodesically) convex optimization. *NeurIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, (1249):13943–13954, 2019.

Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. 2001.

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

Michael JD Powell. On search directions for minimization algorithms. *Mathematical programming*, 4(1):193–201, 1973.

Meisam Razaviyayn, Mingyi Hong, and Zhi-Quan Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

Meisam Razaviyayn, Mingyi Hong, Zhi-Quan Luo, and Jong-Shi Pang. Parallel successive convex approximation for nonsmooth nonconvex optimization. *Advances in neural information processing systems*, 27, 2014.

Quentin Rentmeesters. *Algorithms for data fitting on some common homogeneous spaces.* PhD thesis, Université catholique de Louvain, 2013.

Wolfgang Ring and Benedikt Wirth. Optimization methods on riemannian manifolds and their application to shape space. *SIAM Journal on Optimization*, 22(2):596–627, 2012.

Paul Rodriguez and Brendt Wohlberg. Fast principal component pursuit via alternating minimization. In *2013 IEEE International Conference on Image Processing*, pages 69–73. IEEE, 2013.

Takashi Sakai. *Riemannian geometry*, volume 149. American Mathematical Soc., 1996.

Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

Constantin Udriste. *Convex Functions and Optimization Methods on Riemannian Manifolds.* Mathematics and Its Applications. Springer Netherlands, 1994. ISBN 9780792330028. URL https://books.google.com/books?id=Xtq1RF9lkg4C.

Jinhua Wang, Chong Li, Genaro Lopez, and Jen-Chih Yao. Proximal point algorithms on hadamard manifolds: linear convergence and finite termination. *SIAM Journal on Optimization*, 26(4):2696–2729, 2016.

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

Yangyang Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

Yaguang Yang. Globally convergent optimization algorithms on riemannian manifolds: Uniform framework for unconstrained and constrained optimization. *Journal of Optimization Theory and Applications*, 132(2):245–265, 2007.

Lei Yin, Ankit Parekh, and Ivan Selesnick. Stable principal component pursuit via convex analysis. *IEEE Transactions on Signal Processing*, 67(10):2595–2607, 2019.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pages 1617–1638. PMLR, 2016.

Hongyi Zhang and Suvrit Sra. Towards riemannian accelerated gradient methods. *arXiv preprint arXiv:1806.02812*, 2018.

Zihan Zhou, Xiaodong Li, John Wright, Emmanuel Candes, and Yi Ma. Stable principal component pursuit. In *2010 IEEE international symposium on information theory*, pages 1518–1522. IEEE, 2010.

Wolfgang Ziller. Examples of riemannian manifolds with non-negative sectional curvature. *arXiv preprint math/0701389*, 2007.

Appendix A. Background on Riemannian Optimization

Let $\gamma : [a, b] \to \mathcal{M}$ be a piece-wise differentiable curve, then it assigns to each $t \in (a, b)$ a vector $\gamma'(t)$ in the vector space $T_{\gamma(t)}\mathcal{M}$, the size of which can be measured by the norm $\|\cdot\|_{\gamma(t)}$. The length of γ is given by $L(\gamma) = \int_a^b \|\gamma'(t)\|_{\gamma(t)}\, dt$. The distance for any $x, y \in \mathcal{M}$ is then given by

$$d_{\mathcal{M}}(x, y) = \inf\{L(\gamma) : \gamma \text{ a piecewise continuously differentiable curve from } x \text{ to } y\}$$

We drop the subscript $\mathcal{M}$ when it is clear from context. If $L(\gamma) = d_{\mathcal{M}}(x, y)$, then γ is called a *distance-minimizing geodesic* joining x and y.

For each $\theta, \theta' \in \mathcal{M}$, define $\eta = \eta_\theta(\theta')$ to be the (Such tangent vector need not be unique unless θ' is within the injectivity radius at θ) tangent vector in T_θ such that $\mathrm{Exp}_\theta(\eta) = \theta'$.

For a subset $\boldsymbol{\Theta} \subseteq \mathcal{M}$ and $x \in \boldsymbol{\Theta}$, define the *tangent cone* $\mathcal{T}_{\boldsymbol{\Theta}}(x)$ and the *normal cone* $\mathcal{N}_{\boldsymbol{\Theta}}(x)$ at x as

$$\mathcal{T}_{\boldsymbol{\Theta}}(x) := \left\{ u \in T_x\mathcal{M} \,\middle|\, \mathrm{Exp}_x\left(t\frac{u}{\|u\|}\right) \in \boldsymbol{\Theta} \text{ for some } t \in (0, \mathrm{r_{inj}}(x)) \right\} \cup \{\mathbf{0}\}, \tag{70}$$

$$\mathcal{N}_{\boldsymbol{\Theta}}(x) := \left\{ u \in T_x\mathcal{M} \,\middle|\, \begin{array}{c} \langle u, \eta\rangle \le 0 \text{ for all } \eta \in T_x\mathcal{M} \text{ s.t.} \\ \mathrm{Exp}_x\left(t\frac{\eta}{\|\eta\|}\right) \in \boldsymbol{\Theta} \text{ for some } t \in (0, \mathrm{r_{inj}}(x)) \end{array} \right\}.$$

Note that $\mathcal{T}_{\boldsymbol{\Theta}}(x) = T_x\mathcal{M}$ and $\mathcal{N}_{\boldsymbol{\Theta}}(x) = \{\mathbf{0}\}$ if x is in the interior of $\boldsymbol{\Theta}$. When $\boldsymbol{\Theta}$ is strongly convex, then the tangent cone $\mathcal{T}_{\boldsymbol{\Theta}}(x)$ is a convex cone in the tangent space $T_x\mathcal{M}$ (see Cheeger and Gromoll (1972, Prop.1.8) and Afsari (2011)).

Appendix B. Auxiliary Lemmas

Recall that for each $\theta, \theta' \in \mathcal{M}$, we define $\eta_\theta(\theta')$ to be the set of all tangent vectors $\eta \in T_\theta\mathcal{M}$ such that $\mathrm{Exp}_\theta(\eta) = \theta'$. Define the (Riemannian) subdifferential of $f : \mathcal{M} \to \mathbb{R}$ by

$$\partial f(\theta) := \left\{ v \in T_\theta\mathcal{M} \mid f(\theta') - f(\theta) \ge \langle v, \eta_\theta(\theta')\rangle + o\left(d(\theta, \theta')\right) \text{ as } \theta' \to \theta \right\}.$$

Recall that for a subset $\boldsymbol{\Theta} \subseteq \mathcal{M}$, its (Riemannian) *normal cone* at $\theta \in \Theta$ is defined as (equivalent to (70))

$$\mathcal{N}_{\boldsymbol{\Theta}}(\theta) = \left\{ u \in T_\theta \mid \langle u, \eta\rangle \le 0 \quad \forall\, \theta' \in \Theta\ \forall \eta \in \eta_\theta(\theta') \right\}.$$

Lemma 38 (Bound on the linear approximation for g-smooth functions) *Suppose the function $f : \mathcal{M} \to \mathbb{R}$ is geodesically L-smooth (see Definition 3) and $\mathcal{M}$ is a Riemannian manifold. Suppose $x, y \in \mathcal{M}$ and there exists a distance-minimizing geodesic $\gamma : [0, 1] \to \mathcal{M}$ from x to y. Then*

$$\left| f(y) - f(x) - \left\langle \mathrm{grad}\, f(x), \gamma'(0)\right\rangle_x \right| \le \frac{L}{2} d^2(x, y),$$

where $d(x, y)$ is the Riemannian distance between x and y. Moreover, if $\mathrm{Exp}_x^{-1}(y)$ is well defined, the above can be rewritten as

$$\left| f(y) - f(x) - \left\langle \mathrm{grad}\, f(x), \mathrm{Exp}_x^{-1}(y)\right\rangle_x \right| \le \frac{L}{2} d^2(x, y).$$

Proof Denote the minimal geodesic from x to y as $\gamma : [0,1] \to \mathcal{M}$. That is, $\gamma(0) = x$, $\gamma(1) = y$, and $\int_0^1 \|\gamma'(s)\|\, ds = d(x,y)$. Since the geodesic has a constant speed, we have $\|\gamma'(s)\| \equiv d(x,y)$. Then by the fundamental theorem of calculus,

$$f(y) - f(x) = f(\gamma(1)) - f(\gamma(0)) = \int_0^1 (f \circ \gamma)'(s)\, ds = \int_0^1 \left\langle \operatorname{grad} f\left(\gamma(s)\right), \gamma'(s) \right\rangle_{\gamma(s)} ds.$$

By Cauchy-Schwarz inequality and geodesic L-smoothness of f,

$$\begin{aligned}
&\left| \int_0^1 \left\langle \operatorname{grad} f\left(\gamma(s)\right), \gamma'(s) \right\rangle_{\gamma(s)} - \int_0^1 \left\langle \operatorname{grad} f\left(\gamma(0)\right), \gamma'(0) \right\rangle_{\gamma(0)} ds \right| \\
&\quad = \left| \int_0^1 \left\langle \operatorname{grad} f\left(\gamma(s)\right), \gamma'(s) \right\rangle_{\gamma(s)} - \int_0^1 \left\langle \Gamma_{\gamma(0)\to\gamma(s)} \operatorname{grad} f\left(\gamma(0)\right), \gamma'(s) \right\rangle_{\gamma(s)} ds \right| \\
&\quad = \left| \int_0^1 \left\langle \operatorname{grad} f\left(\gamma(s)\right) - \Gamma_{\gamma(0)\to\gamma(s)} \operatorname{grad} f\left(\gamma(0)\right), \gamma'(s) \right\rangle_{\gamma(s)} ds \right| \\
&\quad \le \int_0^1 \left\| \operatorname{grad} f\left(\gamma(s)\right) - \Gamma_{\gamma(0)\to\gamma(s)} \operatorname{grad} f\left(\gamma(0)\right) \right\| \, \|\gamma'(s)\|\, ds \\
&\quad \le \int_0^1 L d(\gamma(s), \gamma(0)) \|\gamma'(s)\|\, ds \\
&\quad \overset{(*)}{=} L d^2(\gamma(1), \gamma(0)) \int_0^1 s\, ds \\
&\quad = \frac{L}{2} d^2(y, x).
\end{aligned}$$

Now the assertion follows by noting that

$$\int_0^1 \left\langle \operatorname{grad} f\left(\gamma(0)\right), \gamma'(0) \right\rangle_{\gamma(0)} ds = \left\langle \operatorname{grad} f\left(x\right), \gamma'(0) \right\rangle_x = \left\langle \operatorname{grad} f\left(x\right), \frac{\gamma'(0)}{\|\gamma'(0)\|} \right\rangle_x d(x,y).$$

■

Lemma 39 (Additivity of g-smooth functions) *Suppose the function $f, g : \mathcal{M} \to \mathbb{R}$ are geodesically smooth function with positive constants L_f and L_g, respectively (see definition 3). Then $f + g$ is geodesically smooth with constant $L_f + L_g$.*

Proof First note that by definition of $\operatorname{grad} f$ linearity of $D(\cdot)$, the operator of directional derivative, for any $\eta \in T_x\mathcal{M}$,

$$\begin{aligned}
\langle \operatorname{grad} f(x) + \operatorname{grad} g(x), \eta \rangle_x &= \langle \operatorname{grad} f(x), \eta \rangle_x + \langle \operatorname{grad} g(x), \eta \rangle_x \\
&= D(f)(x)[\eta] + D(g)(x)[\eta] \\
&= D(f+g)(x)[\eta].
\end{aligned}$$

Therefore by definition, $\operatorname{grad} f(x) + \operatorname{grad} g(x) = \operatorname{grad}(f+g)(x)$.

Also note that the parallel transport $\Gamma_{x\to y} : T_x\mathcal{M} \to T_y\mathcal{M}$ is a linear isomorphism, therefore we have

$$\begin{aligned}
&\| \operatorname{grad}(f+g)(x) - \Gamma_{x\to y} \operatorname{grad}(f+g)(y)\| \\
\le &\| \operatorname{grad}(f)(x) - \Gamma_{x\to y} \operatorname{grad}(f)(y)\| + \| \operatorname{grad}(g)(x) - \Gamma_{x\to y} \operatorname{grad}(g)(y)\| \\
\le &(L_g + L_f) d(x,y).
\end{aligned}$$

■

Lemma 40 *Let $(a_n)_{n\geq 0}$ and $(b_n)_{n\geq 0}$ be sequences of nonnegative real numbers such that $\sum_{n=0}^{\infty} a_n b_n < \infty$. Then*

$$\min_{1\le k\le n} b_k \le \frac{\sum_{k=0}^{\infty} a_k b_k}{\sum_{k=1}^{n} a_k} = O\left(\left(\sum_{k=1}^{n} a_k\right)^{-1}\right).$$

Proof The assertion follows by noting that

$$\left(\sum_{k=1}^{n} a_k\right) \min_{1\le k\le n} b_k \le \sum_{k=1}^{n} a_k b_k \le \sum_{k=1}^{\infty} a_k b_k < \infty.$$

■

The following propositions are about the line search method on Riemannian manifolds and are parallel to the Euclidean versions in Grippo and Sciandrone (2000).

Consider a sequence $\{x_k\} \in \mathbf{\Theta}$ with partition $x_k = (x_k^{(1)}, \cdots, x_k^{(m)})$ and the searching directions $d_k^{(i)} \in T_{\Theta^{(i)}}$ satisfying the following assumptions:

Definition 41 (Gradient related searching directions) *Let $\{d_k^{(i)}\} \in T^*_{\Theta^{(i)}}$ be the sequence of searching directions such that they are gradient-related, i.e.,*

1. *there exists a number $M > 0$ such that $\left\| d_k^{(i)} \right\| \leqslant M$ for all k;*

2. $\liminf_{k\to\infty} \langle \operatorname{grad}_i g_k^{(i)}(x_k), d_k^{(i)} \rangle < 0$.

Then we call $\{d_k^{(i)}\}$ gradient related.

An Armijo-type line search method can be described as follows.

Algorithm 2 Armijo-type line search algorithm for surrogates

```
Input: $\sigma \in (0,1), \beta \in (0,1)$; $d_k^{(i)}$ (search direction)
Compute  $\alpha_k = \min_{j\geq 0} \left\{ \beta^j : g_k^{(i)}\left(\operatorname{Exp}_{x_k^{(i)}}(\beta^j d_k^{(i)})\right) \leqslant g_k^{(i)}(x_k) + \sigma\beta^j \langle \operatorname{grad} g_k^{(i)}(x_k), d_k^{(i)} \rangle \right\}$
output: $\alpha_k$
```

Next, we show some well-known results on the Riemannian line search algorithm. It is worthwhile to point out that, in what follows, the sequence $\{x_k\}$ is a given sequence that may not depend on the line search algorithm, in the sense that x_{k+1} may not be generated by line search along d_k. Nevertheless, this has no substantial effect on the convergence proof, which can be deduced easily from known results (see e.g., Absil et al. (2009)).

Proposition 42 *Let $\{x_k\}$ be a sequence in $\Theta^{(i)}$ and let $\{d_k\} \in T^*_{\Theta^{(i)}}$ be the sequence of searching directions satisfies Definition 41. Let α_k be computed by Algorithm 2, then*

(i) There exists a finite integer $j \geq 0$ such that $\alpha_k = (\beta_i)^j$ satisfies the acceptability condition (2);

(ii) Suppose $\{x_k\}$ converges to $\bar{x}$ and

$$\lim_{k\to\infty} g_{k+1}^{(i)}\left(\mathrm{Exp}_{x_k}(\alpha_k d_k)\right) - g_{k+1}^{(i)}(x_k) = 0.$$

Then

$$\lim_{k\to\infty} \operatorname{grad} g_{k+1}^{(i)}(x_k) = \mathbf{0}.$$

Proof (i) is obvious by $\langle \operatorname{grad}_i g_k^{(i)}(x_k), d_k^{(i)} \rangle < 0$ and smoothness of $g_k^{(i)}$. To prove (ii), suppose for a contradiction that $\lim_{k\to\infty} \operatorname{grad} g_{k+1}^{(i)}(x_k) \neq \mathbf{0}$. By the choice of search directions d_k, there exists $\delta > 0$ such that

$$\left\langle \operatorname{grad} g_{k+1}^{(i)}(x_k), d_k \right\rangle_{x_k} < -\delta < 0 \quad \text{for all sufficiently large } k.$$

By the choice of α_k, we have

$$g_{k+1}^{(i)}(x_k) - g_{k+1}^{(i)}\left(\mathrm{Exp}_{x_k}(\alpha_k d_k)\right) \geq -\sigma\alpha_k \left\langle \operatorname{grad} g_{k+1}^{(i)}(x_k), d_k \right\rangle_{x_k} > \sigma\alpha_k\delta > 0$$

for all sufficiently large k. Since $g_{k+1}^{(i)}(x_k) - g_{k+1}^{(i)}\left(\mathrm{Exp}_{x_k}(\alpha_k d_k)\right)$ goes to zero, we must have $\alpha_k \to 0$. Recall that α_k 's are determined from the Armijo rule, so $\alpha_k = \beta^{m_k}$ for some integer $\mathtt{m}_k \geq 0$ all $k \geq 1$. Since $\alpha_k = o(1)$, m_k must diverge, so $m_k \geq 1$ for all $k \geq \bar{k}$ for some integer $\bar{k} \geq 1$. Then $\alpha_k/\beta = \beta^{m_k - 1} \leq 1$, and the step-size $\frac{\alpha_k}{\beta}$ did not satisfy the Armijo condition. Hence

$$g_{k+1}^{(i)}(x_k) - g_{k+1}^{(i)}\left(\mathrm{Exp}_{x_k}(\frac{\alpha_k}{\beta} d_k)\right) < -\sigma\frac{\alpha_k}{\beta}\left\langle \operatorname{grad} g_{k+1}^{(i)}(x_k), d_k \right\rangle_{x_k}, \quad k \geq \bar{k}.$$

Denoting

$$\hat{g}_x = g_{k+1}^{(i)} \circ \mathrm{Exp}_x \quad \text{and} \quad \tilde{\alpha}_k = \frac{\alpha_k}{\beta}.$$

The inequality above reads

$$\frac{\hat{g}_{x_k}(\mathbf{0}) - \hat{g}_{x_k}(\tilde{\alpha}_k d_k)}{\tilde{\alpha}_k} < -\sigma\left\langle \operatorname{grad} g_{k+1}^{(i)}(x_k), d_k \right\rangle_{x_k} < \sigma\delta \quad \forall\ k \geq \bar{k}.$$

The mean value theorem ensures that there exists $t \in [0, \tilde{\alpha}_k]$ such that

$$-\operatorname{D}\hat{g}_{x_k}(td_k)[d_k] < -\sigma \left\langle \operatorname{grad} g^{(i)}_{k+1}(x_k), d_k \right\rangle_{x_k}, \quad k \geq \bar{k}.$$

Now since $\tilde{\alpha}_k \to 0$ and recall that $\operatorname{D}\hat{g}_{x_k}(0)[d_k] = \left\langle \operatorname{grad} g^{(i)}_{k+1}(x_k), d_k \right\rangle_{x_k}$, we obtain

$$-\liminf_{k\to\infty} \left\langle \operatorname{grad} g^{(i)}_{k+1}(x_k), d_k \right\rangle_{x_k} \leq -\sigma \liminf_{k\to\infty} \left\langle \operatorname{grad} g^{(i)}_{k+1}(x_k), d_k \right\rangle_{x_k}.$$

Since $\sigma < 1$, it follows that $\liminf_{k\to\infty} \left\langle \operatorname{grad} g^{(i)}_{k+1}(x_k), d_k \right\rangle_{x_k} \geq 0$, which is a contradiction. ■

Proposition 43 (Properties of inverse exponential map on Hadamard manifold.) *Let $\mathcal{M}$ be a Hadamard manifold , $(x_n)_{n\geq 1} \subset \mathcal{M}$ and $x_0 \in \mathcal{M}$,*

(i) For any $y \in \mathcal{M}$, we have

$$\operatorname{Exp}^{-1}_{x_n}(y) \longrightarrow \operatorname{Exp}^{-1}_{x_0}(y) \quad \textit{and} \quad \operatorname{Exp}^{-1}_{y}(x_n) \longrightarrow \operatorname{Exp}^{-1}_{y}(x_0).$$

(ii) If $v_n \in T_{x_n}\mathcal{M}$ and $v_n \to v_0$, then $v_0 \in T_{x_0}\mathcal{M}$

(iii) Given $u_n, v_n \in T_{x_n}\mathcal{M}$ and $u_0, v_0 \in T_{x_0}\mathcal{M}$, if $u_n \to u_0$ and $v_n \to v_0$, then

$$\langle u_n, v_n \rangle \longrightarrow \langle u_0, v_0 \rangle.$$

Proof See e.g., Li et al. (2009, Lemma 2.4). ■

Appendix C. Details of Section 5

C.1 Riemannian Hessian on fixed-rank manifold

Here we provide details about why the Euclidean distance function is not g -smooth over low-rank manifolds, as discussed in Section 5. Let $\mathcal{R}_r \subseteq \mathbb{R}^{m\times n}$ be the manifold of rank-r matrices as in (21). Let $X \in \mathcal{R}_r$, and without loss of generality, let $X = U\Sigma V^T$ where $U \in \mathcal{V}^{m\times r}$, $V \in \mathcal{V}^{n\times r}$ and $\Sigma = \operatorname{diag}(\sigma_1, \cdots, \sigma_r)$ with $\sigma_1 \geq \cdots \geq \sigma_r > 0$. Following from Absil et al. (2013, Section 4.3), the Riemannian Hessian of f at X for $Z \in T_X\mathcal{R}_r$ is given by

$$\begin{aligned} \operatorname{Hess} f(X)[Z] &= \mathcal{P}_X \nabla^2 f(X) Z + \mathcal{P}_X \operatorname{D}_Z \mathcal{P}_X \nabla f(X) \\ &= (\nabla^2 f(X)\mathrm{P}_V + \mathrm{P}_U \nabla^2 f(X) - \mathrm{P}_U \nabla^2 f(X) \mathrm{P}_V) Z \\ &\qquad + \nabla f(X) Z^T (X^+)^T + (X^+)^T Z^T \nabla f(X), \end{aligned} \tag{71}$$

where ∇f and $\nabla^2 f$ are the Euclidean gradient and Euclidean Hessian of f, $\mathrm{P}_U = UU^T$, $\mathrm{P}_V = VV^T$, $X^+ = V\Sigma^{-1}U^T$. $\mathcal{P}_X$ is the projection operator onto the tangent space at X.

D_Z is the directional derivative following the tangent vector Z. A detailed discussion of these operators can be found in Absil et al. (2013).

Note for a simple Euclidean distance squared function $f(X) = \|X - X_0\|_F^2$, we have $\nabla f = 2(X - X_0)$ and $\nabla^2 f = 2\mathbf{I}$. In order to show that f is g-smooth on the fixed-rank manifold, one needs to verify

$$\sup_{X\in\mathcal{R}_r} \max_{Z\in T_X} (\langle \operatorname{Hess} f(X)[Z], Z\rangle) \leq C$$

using (71), where $C \in (0,\infty)$ is a constant (see Nguyen et al. (2019, Lemma C.6)). Below we give a counterexample to show such a constant C does not exist.

Take $Z = UV^T$ and $X_0 = -UV^T$. Note the inner product of the first term in (71) with Z is bounded. Hence we only need to show the inner product of the last two terms in (71) with Z can be unbounded. In fact, the last two terms are the same by the cyclic property of trace,

$$\begin{aligned}\langle (X^+)^T Z^T \nabla f(X), Z\rangle &= \operatorname{Tr}((X^+)^T Z^T \nabla f(X) Z^T) \\ &= \operatorname{Tr}(\nabla f(X) Z^T (X^+)^T Z^T) = \langle \nabla f(X) Z^T (X^+)^T, Z\rangle.\end{aligned}$$

Therefore we only compute the second term,

$$\begin{aligned}\langle \nabla f(X) Z^T (X^+)^T, Z\rangle &= 2\langle U\Sigma V^T V U^T U \Sigma^{-1} V^T + UV^T V U^T U \Sigma^{-1} V^T, Z\rangle \\ &= 2\operatorname{Tr}((UV^T)^T UV^T) + 2\operatorname{Tr}((U\Sigma^{-1}V^T)^T UV^T) \\ &= 2\|UV^T\|^2 + 2\operatorname{Tr}(V\Sigma^{-1}V^T) \\ &= 2\|UV^T\|^2 + 2\operatorname{Tr}(\Sigma^{-1}).\end{aligned} \tag{72}$$

Now we take a sequence of $X^{(k)} \in \mathcal{R}_r$ such that the smallest singular value of $X^{(k)}$ goes to zero, i.e., $\sigma_r^{(k)} \to 0$ as $k \to \infty$. Therefore we have $(\sigma_r^{(k)})^{-1} \to \infty$ as $k \to \infty$. Hence (72) is unbounded. We conclude that $f(X) = \|X - X_0\|_F^2$ is not g-smooth on the fixed-rank manifold.

C.2 Details of Section 5.2

We give the details of the MM update for Θ (Blocker et al., 2023). For fixed H and Y, we first simplify (34),

$$\begin{aligned}f(Q,\Theta) &= -\sum_{i=1}^{T} \left\| X_i^T \left(H\cos(\Theta t_i) + Y\sin(\Theta t_i)\right)\right\|_{\mathrm{F}}^2 \\ &= -\sum_{i=1}^{T}\sum_{j=1}^{k} r_{i,j}\cos(2\theta_j t_i - \phi_{i,j}) + b_{i,j},\end{aligned} \tag{73}$$

where θ_j is the j-th diagonal element of Θ. Defining $\arctan 2(y,x)$ as the angle of the point (x,y) in the 2D plane counter-clockwise from the positive x-axis, the associated constants $r_{i,j}, \phi_{i,j}, b_{i,j}$ in the above equation are defined as

$$\phi_{i,j} = \arctan 2\left(\beta_{i,j}, \frac{\alpha_{i,j} - \gamma_{i,j}}{2}\right), \qquad \alpha_{i,j} = \left[H^T X_i X_i^T H\right]_{j,j},$$

$$r_{i,j} = \sqrt{\left(\frac{\alpha_{i,j} - \gamma_{i,j}}{2}\right)^2 + \beta_{i,j}^2}, \qquad \beta_{i,j} = \text{real}\left\{\left[Y^T X_i X_i^T H\right]_{j,j}\right\},$$
$$b_{i,j} = \frac{\alpha_{i,j} + \gamma_{i,j}}{2}, \qquad \gamma_{i,j} = \left[Y^T X_i X_i^T Y\right]_{j,j}.$$

Note that (73) is separable for each diagonal element θ_j of Θ, so we could find the minimizer separately by a univariate minimization. Let

$$f_{n+1,j}^{(2)}(\theta_j) := -\sum_{i=1}^{T} (r_{i,j})_{n+1} \cos\left(2\theta_j t_i - (\phi_{i,j})_{n+1}\right) + (b_{i,j})_{n+1},$$

which is the marginal objective function at iteration $n+1$ for the j-th diagonal component of Θ. The subscript $n+1$ outside the parenthesis denotes values of the parameters at iteration $n+1$. The gradient, which actually becomes scalar, is given by

$$\nabla f_{n+1,j}^{(2)}(\theta_j) = \dot{f}_{n+1,j}^{(2)}(\theta_j) = 2\sum_{i=1}^{T} (r_{i,j})_{n+1} t_i \sin\left(2\theta_j t_i - (\phi_{i,j})_{n+1}\right),$$

which is Lipschitz continuous with parameter $L_{n+1,j} = 4\sum_{i=1}^{T}(r_{i,j})_{n+1} t_i^2$. We consider the following prox-linear majorizer of $f_{n+1,j}^{(2)}$: For $\lambda > L_{n+1,j}$,

$$g_{n+1,j}^{(2)}(\theta_j) = f_{n+1,j}^{(2)}\left((\theta_j)_n\right) + \nabla f_{n+1,j}^{(2)}\left((\theta_j)_n\right)\left(\theta_j - (\theta_j)_n\right) + \frac{\lambda}{2}\left(\theta_j - (\theta_j)_n\right)^2,$$

where $(\theta_j)_n$ is the value of θ_j at iteration n.

Then by using (22),

$$\begin{aligned}(\theta_j)_{n+1} &= \arg\min g_{n+1,j}^{(2)}(\theta_j)\\ &= \text{Proj}_{\mathbb{R}}\left((\theta_j)_n - \frac{1}{\lambda}\nabla f_{n+1,j}^{(2)}\left((\theta_j)_n\right)\right)\\ &= (\theta_j)_n - \frac{1}{\lambda}\nabla f_{n+1,j}^{(2)}\left((\theta_j)_n\right).\end{aligned}$$

C.3 Details of Section 5.3

The following propositions from Atkinson and Mitchell (1981) give the closed form of Fisher-Rao (FR) distance under certain circumstances,

Proposition 44 (FR distance for Gaussian distributions with identical mean) *If $\mathcal{N}(\hat{\mu}, \Sigma_0)$ and $\mathcal{N}(\hat{\mu}, \Sigma_1)$ are Gaussian distributions with identical mean $\hat{\mu} \in \mathbb{R}^n$ and covariance matrices $\Sigma_0, \Sigma_1 \in \mathbb{S}_{++}^n$, the set of $n \times n$ positive definite matrices, we have*

$$d(\Sigma_0, \Sigma_1) = \frac{1}{\sqrt{2}}\left\|\log\left(\Sigma_1^{-\frac{1}{2}}\Sigma_0\Sigma_1^{-\frac{1}{2}}\right)\right\|_F,$$

where $\log(\cdot)$ represents the matrix logarithm, and $\|\cdot\|_F$ stands for the Frobenius norm.

Remark 45 *It is worthwhile to point out the FR metric on the tangent space $T_\Sigma \mathbb{S}^n_{++}$ at $\Sigma \in \mathbb{S}^n_{++}$ can be re-expressed as (see Atkinson and Mitchell (1981, p. 382))*

$$\langle \Omega_1, \Omega_2 \rangle_\Sigma \triangleq \frac{1}{2} \operatorname{Tr}\left(\Omega_1 \Sigma^{-1} \Omega_2 \Sigma^{-1}\right) \quad \forall \Omega_1, \Omega_2 \in T_\Sigma \mathbb{S}^n_{++}.$$

Proposition 46 (FR distance for Gaussian distributions with identical covariance) *If $\mathcal{N}\left(\mu_0, \hat{\Sigma}\right)$ and $\mathcal{N}\left(\mu_1, \hat{\Sigma}\right)$ are Gaussian distributions with identical covariance matrix $\hat{\Sigma} \in \mathbb{S}^n_{++}$ and mean vectors $\mu_0, \mu_1 \in \mathbb{R}^n$, we have*

$$\bar{d}\left(\mu_0, \mu_1\right) = \sqrt{\left(\mu_0 - \mu_1\right)^T \hat{\Sigma}^{-1} \left(\mu_0 - \mu_1\right)}.$$

Next, we show the optimization problem in (37) satisfies our assumptions for RBMM. We first cite the following proposition from Nguyen et al. (2019), which states the geodesical convexity of the constraint sets.

Lemma 47 (Convexity of constraint sets) *$\Theta^{(1)}$ and $\Theta^{(2)}$ are strongly convex.*

The following lemma shows the geodesical smoothness of $f_n^{(1)}$ and $f_n^{(2)}$, aiming to show that $f(\mu, \Sigma)$ satisfies (A0)**(i)**.

Lemma 48 (Geodesic smoothness of marginal objective function) *For problem (37), we have*

1. *$f_n^{(1)}$ is (geodesically) $L_n^{(1)}$-smooth with $L_n^{(1)} = 1/\lambda_{\min}(\Sigma_n)$.*
2. *$f_n^{(2)}$ is geodesically $L_n^{(2)}$-smooth with $L_n^{(2)} = 2\lambda_{\max}(S_n)\lambda_{\min}(\widehat{\Sigma})^{-1} \exp(\sqrt{2}\rho_2)$.*

We also need a uniform upper bound for $L_n^{(1)}$ and $L_n^{(2)}$. The following proposition gives lower and upper bounds for the eigenvalues of $\Sigma \in \Theta^{(2)}$.

Proposition 49 (Property of Fisher-Rao ball (Nguyen et al., 2019)) *The FR ball $\Theta^{(2)}$ has the following property. For any $\Sigma \in \Theta^{(2)}$, we have $\lambda_{\min}(\hat{\Sigma})e^{-\sqrt{2}\rho_2} \cdot I_n \preceq \Sigma \preceq \lambda_{\max}(\hat{\Sigma})e^{\sqrt{2}\rho_2} \cdot I_n$.*

We further set a fixed ρ_1 given in the following proposition in order to upper bound $\lambda_{\max}(S_n)$.

Proposition 50 *Let the radius ρ_1 in $\Theta^{(1)}$ to be*

$$\rho_1 = \max_{m=1,\cdots,M} \|x_m - \hat{\mu}\|_2.$$

Then $\lambda_{\max}(S_n) \leq 4\rho_1^2$.

Proof Recall that $\Theta^{(1)} = \left\{\mu \in \mathbb{R}^n : (\mu - \hat{\mu})^T(\mu - \hat{\mu}) \leq \rho_1^2\right\}$ and $\hat{\mu} = \frac{1}{M}\sum_{m=1}^M x_m$.

Now with $\rho_1 = \max_m \|x_m - \hat{\mu}\|_2$ we have

$$\operatorname{tr}\left(\left(x_m - \mu_n\right)\left(x_m - \mu_n\right)^T\right) = \operatorname{tr}\left(\left(x_m - \mu_n\right)^T\left(x_m - \mu_n\right)\right) \leq 4\rho_1^2, \qquad \text{for} \quad 1 \leq m \leq M.$$

therefore $\lambda_{\max}(S_n) \leq \operatorname{tr}\left(\max_m \left(x_m - \mu_n\right)^T\left(x_m - \mu_n\right)\right) \leq 4\rho_1^2$. ∎

Lemma 48 together with Prop. 49 and Prop. 50 gives

$$L_1^{(n)} \leq 1/\lambda_{\min}(\hat{\Sigma})e^{-\sqrt{2}\rho_2} \quad \text{and} \quad L_2^{(n)} \leq 8\rho_1^2\lambda_{\min}(\widehat{\Sigma})^{-1}\exp(\sqrt{2}\rho_2) \quad \text{for all } n.$$

which shows $f(\mu, \Sigma)$ satisfies (A0)**(i)**.