Convergence and Complexity Guarantee for Inexact First-order Riemannian Optimization Algorithms

Yuchen Li [1] **Laura Balzano** [2] **Deanna Needell** [3] **Hanbaek Lyu** [1]

Abstract

We analyze inexact Riemannian gradient descent (RGD) where Riemannian gradients and retractions are inexactly (and cheaply) computed. Our focus is on understanding when inexact RGD converges and what is the complexity in the general nonconvex and constrained setting. We answer these questions in a general framework of tangential Block Majorization-Minimization (tBMM). We establish that tBMM converges to an ϵ-stationary point within $O(\epsilon^{-2})$ iterations. Under a mild assumption, the results still hold when the subproblem is solved inexactly in each iteration provided the total optimality gap is bounded. Our general analysis applies to a wide range of classical algorithms with Riemannian constraints including inexact RGD and proximal gradient method on Stiefel manifolds. We numerically validate that tBMM shows improved performance over existing methods when applied to various problems, including nonnegative tensor decomposition with Riemannian constraints, regularized nonnegative matrix factorization, and low-rank matrix recovery problems.

1. Introduction

A typical formulation of constrained Riemannian optimization takes the following form:

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\subseteq\mathcal{M}} \left(f(\boldsymbol{\theta}) := \varphi(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta})\right), \tag{1}$$

where $\mathcal{M}$ is a smooth Riemannian manifold embedded in a Euclidean space, $\boldsymbol{\Theta}$ is a closed subset of $\mathcal{M}$, and $f : \mathcal{M} \to \mathbb{R}$ is an objective function consisting of a smooth part φ and convex (and possibly nonsmooth) part ψ. Riemannian optimization problems of the form (1) have 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 & 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.

Motivation: Inexact RGD. In many Riemannian optimization problem instances, the dimension of the parameter space is much less than that of the ambient dimension. Such 'latent' low-dimensional structure in the parameter space can be utilized by using the Riemannian gradient of the objective that lives in the (low-dimensional) tangent space, instead of the full gradient in the (high-dimensional) ambient space. Thus many Riemannian optimization methods take the following form (Baker et al., 2008; Yang, 2007; Boumal et al., 2019; Edelman et al., 1998): Iteratively,

(i) Compute descent direction in the tangent space;

(ii) Take a step in that direction along a geodesic.

However, step **(ii)** is often challenging in practice so other approaches alleviate this burden by utilizing approximations or imposing additional assumptions (Absil et al., 2009; Boumal, 2023). A popular way to implement a similar idea with less computational burden for computing the geodesic is to make the parameter update first in the tangent spaces and then map the resulting point back onto the manifold using a retraction (which is like a projection from the tangent space onto the manifold).

Perhaps the simplest and most widely used Riemannian optimization algorithms for smooth objectives f of the above form is Riemannian gradient descent (RGD):

$$\textbf{(RGD)} \qquad \begin{cases} V_n & \leftarrow -\operatorname{grad} f(\boldsymbol{\theta}_{n-1}) \\ \boldsymbol{\theta}_n & \leftarrow \operatorname{Rtr}_{\boldsymbol{\theta}_{n-1}}(\alpha_n V_n). \end{cases}$$

Here $\operatorname{grad} f(\boldsymbol{\theta}_{n-1})$ denotes the Riemannian gradient of f at $\boldsymbol{\theta}_{n-1}$, $\operatorname{Rtr}_{\boldsymbol{\theta}_{n-1}}$ denotes the retraction map from the tangent space $T_{\boldsymbol{\theta}_{n-1}}\mathcal{M}$ at base point $\boldsymbol{\theta}_{n-1}$ onto the manifold $\mathcal{M}$

[1]Department of Mathematics, University of Wisconsin-Madison, WI 53706, USA [2]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA [3]Department of Mathematics, University of California, Los Angeles, CA 90025, USA. Correspondence to: Hanbaek Lyu <hlyu@math.wisc.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

(see Appendix A), and $\alpha_n > 0$ is a step size. In the literature, one typically assumes that computing Riemannian gradients and retractions are computationally feasible. However, there are several problem instances where either computing exact Riemannian gradient or the retraction is difficult (Wiersema & Killoran, 2023; Ablin & Peyré, 2022; Wang et al., 2021). For such situations, it is reasonable to consider the following 'inexact version' of RGD:

$$\textbf{(Inexact RGD)} \qquad \begin{cases} \hat{V}_n &= -\widehat{\operatorname{grad}} f(\boldsymbol{\theta}_{n-1}) \\ \boldsymbol{\theta}_n &= \widehat{\operatorname{Rtr}}_{\boldsymbol{\theta}_{n-1}}\left(\alpha_n \hat{V}_n\right), \end{cases} \tag{2}$$

where $\widehat{\operatorname{grad}}$ and $\widehat{\operatorname{Rtr}}$ are computationally feasible inexact Riemannian gradient and retraction operators, respectively. Our main question is the following. *When does this inexact RGD converge? What can we say about its complexity?* We answer these questions in a general framework of Riemannian tangential Block Majorization-Minimization (tBMM). Here we state a corollary of our general result for the context of inexact RGD (see proof in Appendix G).

Corollary 1.1 (Inexact RGD)**.** *Let $(\boldsymbol{\theta}_n)_n$ be the iterations generated by the inexact RGD (2) for solving (1) with $\psi = 0$. Assume the manifold is complete and the objective function is uniformly lower bounded with compact sub-level sets. Then each limit point of $(\boldsymbol{\theta}_n)_n$ is a stationary point and an ε-stationary point is obtained within $O(\varepsilon^{-2})$ iterations if the following holds:*

(i) *$\nabla\varphi$ is L-Lipschitz continuous for some $L > 0$.*

(ii) *For each n, choose $V_n \in T_{\boldsymbol{\theta}_{n-1}}\mathcal{M}$ such that*

$$\operatorname{Rtr}_{\boldsymbol{\theta}_{n-1}}(V_n) = \widehat{\operatorname{Rtr}}_{\boldsymbol{\theta}_{n-1}}\left(\hat{V}_n\right).$$

Denote $g_n(\eta) := \langle \operatorname{grad} f(\boldsymbol{\theta}_{n-1}), \eta\rangle + \frac{L}{2}\|\eta\|^2$ and define the optimality gap as

$$\Delta_n := g_n(\alpha_n V_n) - g_n(-\frac{1}{2L}\operatorname{grad} f(\boldsymbol{\theta}_{n-1})).$$

Then $\sum_{n=0}^{\infty} \Delta_n < \infty$.

General framework through tangential MM. We establish Corollary 1.1 for the broader class of first-order Riemannian optimization algorithms called *tangential Majorization-Minimization* (tMM). We first recall that the classical Majorization-Minimization algorithm generalizes gradient descent in Euclidean space (Mairal, 2013):

$$\begin{aligned} \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1} \leftarrow &= -\frac{1}{\lambda}\nabla f(\boldsymbol{\theta}_{n-1}) \\ = \operatorname*{arg\,min}_{\eta\in\mathbb{R}^p} &\left[g_n(\eta) := f(\boldsymbol{\theta}_{n-1}) + \langle \nabla f(\boldsymbol{\theta}_{n-1}),\, \eta\rangle + \frac{\lambda}{2}\|\eta\|^2\right]. \end{aligned}$$

Assuming $\lambda \geq L$, the Lipschitz parameter for ∇f, the quadratic function g_n above satisfies (majorization) $g_n(\eta) \geq f(\boldsymbol{\theta}_n + \eta)$ and (tightness) $\nabla g_n(0) = \nabla f(\boldsymbol{\theta}_{n-1})$ and is called the prox-linear surrogate of f at $\boldsymbol{\theta}_{n-1}$. Thus, we majorize objective f near $\boldsymbol{\theta}_{n-1}$ by the prox-linear surrogate g_n and minimize it to find a descent direction.

In the Riemannian setting, RGD can be thought of as a Riemannian version of MM. Notice that the prox-linear surrogate g_n above is in fact defined on the tangent space of $\mathbb{R}^p$ at $\boldsymbol{\theta}_{n-1}$, and what it majorizes is not the original objective f, but the 'pull-back' objective $f(\boldsymbol{\theta}_{n-1} + \eta)$ defined on the tangent space. Applying this observation to RGD, we seek to majorize the pull-back objective

$$\hat{\varphi}_n := \varphi_n \circ \operatorname{Rtr}_{\theta_{n-1}} : T_{\boldsymbol{\theta}_{n-1}}\mathcal{M} \to \mathbb{R}$$

obtained by precomposing the objective function $\varphi : \mathcal{M} \to \mathbb{R}$ with the retraction at $\boldsymbol{\theta}_{n-1}$. In this way, the Riemannian objective is now lifted to the Euclidean objective $\hat{\varphi}_n$ on the tangent space. Here we apply the usual MM strategy on the tangent space to find a descent direction, take a step on the tangent space, and retract back onto the manifold. We call this procedure 'tangential MM', which is concisely stated below:

$$\begin{aligned} \hat{g}_n &\leftarrow \left[\text{Majorizing surrogate of } \hat{\varphi}_n \text{ s.t. } \hat{g}_n(\mathbf{0}) = \hat{\varphi}(\mathbf{0})\right] \\ V_n &\leftarrow \operatorname*{arg\,min}_{\eta\in T_{\boldsymbol{\theta}_{n-1}}\mathcal{M}} \hat{g}_n(\eta) \\ \boldsymbol{\theta}_n &\leftarrow \operatorname{Rtr}_{\boldsymbol{\theta}_{n-1}}(\alpha_n V_n). \end{aligned} \tag{3}$$

To view the inexact RGD for smooth objectives (2) as a special case of tMM (3), let

$$\hat{g}_n(\eta) := \varphi(\boldsymbol{\theta}_{n-1}) + \langle \operatorname{grad}\varphi(\boldsymbol{\theta}_{n-1}), \eta\rangle + \frac{\lambda}{2}\|\eta\|^2 \tag{4}$$

where $\lambda \geq L$ and $L > 0$ is the Lipschitz continuity parameter of $\nabla\varphi$. Then one can verify the update of tMM using (4) is the same as (2) (see Section 4.2 for details).

In this work, we consider the more general setting when the manifold $\mathcal{M}$ is a product manifold given by $\mathcal{M} = \mathcal{M}^{(1)} \times \cdots \times \mathcal{M}^{(m)}$. Then problem (1) becomes a multi-block Riemannian constrained problem as follows:

$$\min_{\substack{\boldsymbol{\theta}=[\theta^{(1)},\ldots,\theta^{(m)}] \\ \theta^{(i)}\in\Theta^{(i)}\subseteq\mathcal{M}^{(i)} \text{ for } i=1,\ldots,m}} (f(\boldsymbol{\theta}) = \varphi(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta})). \tag{5}$$

Given that the problem (5) is typically nonconvex, expecting an algorithm to converge to a globally optimal solution from an arbitrary initialization might not always be reasonable. Instead, our goal is to ensure global convergence to stationary points from any initialization. In certain problem classes, stationary points can be practically and theoretically as good as global optimizers (Mairal et al., 2010; Sun et al., 2015). Additionally, determining the iteration complexity of such algorithms is crucial for both theoretical and practical

purposes. This involves bounding the worst-case number of iterations needed to achieve an ϵ-approximate stationary point (appropriately defined in Sec. 3.1).

In this work, we propose a block-extension of tMM in (3) that can also handle additional (geodesically convex) constraints within each manifold $\mathcal{M}^{(i)}$ as well as nonsmooth nonconvex objectives (see Algorithm 1).We carefully analyze tBMM in various settings and obtain first-order optimality guarantees and iteration complexity under inexact computations. From the general results, we can easily deduce Corollary 1.1.

1.1. Related Works

Under the Euclidean setting, i.e. when each $\mathcal{M}^{(i)}$ in (5) is a Euclidean space, the corresponding Euclidean block MM method has been well studied in the literature. For convex problems, the Euclidean block MM method is studied in (Xu & Yin, 2013) with prox-linear surrogates, and in (Razaviyayn et al., 2013; Hong et al., 2015) with general surrogates. For nonconvex problems, some variants of block MM are studied in some recent works, including BMM-DR in (Lyu & Li, 2023), and BCD-PR in (Kwon & Lyu, 2023).

Under the Riemannian setting, some recent work showed convergence and complexity of block MM methods for solving (5) when $\psi = 0$. Namely, in (Gutman & Ho-Nguyen, 2023), the authors established a sublinear convergence rate for an block-wise Riemannian gradient descent. However, the retraction considered there is restricted to the exponential map, which excludes many commonly used retractions in the literature. In (Peng & Vidal, 2023), the authors established convergence and complexity results of general block MM on compact manifolds. In (Li et al., 2023), convergence and complexity results are established for the Riemannian block MM methods on general Riemannian manifolds. Moreover, extra constrained sets on the manifolds and inexact computation of subproblems are allowed in the general framework of (Li et al., 2023). However, all these analyses are limited to the smooth problem ($\psi = 0$) and cannot be directly applied to the general problem (5).

For nonconvex nonsmooth problems, in (Chen et al., 2020), the authors studied a tangential type of MM method with prox-linear surrogates on Stiefel manifolds with complexity guarantees. However, the problem considered there is a single-block problem.

1.2. Our Contributions

In this work, we propose tBMM, which is a general framework of tangential type Riemannian block MM algorithms for solving nonconvex, nonsmooth, multi-block constrained Riemannian optimization problems (5), and allowing inexact computation of subproblems. We thoroughly analyze tBMM (3) and obtain asymptotic convergence to the set of stationary points and iteration complexity. The theoretical contributions of this work, compared to the aforementioned related work, lie especially in the following three aspects,

(1) (Iteration complexity) tBMM is applicable to nonsmooth, nonconvex, multi-block Riemannian optimization problems, and we derive the iteration complexity of $O(\varepsilon^{-2})$ along with asymptotic convergence to the set of stationary points. See Theorem 3.1.

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

(3) (Robustness) tBMM is robust in the face of inexact computations in each iteration. See (A1)**(ii)**.

tBMM entails various classical and practical algorithms. This includes inexact RGD (2), block Riemannian prox-linear updates (see Section 4.2), and nonsmooth proximal gradient method on Stiefel manifolds (see Section 4.3). We apply our results to the above classical algorithms and obtain the following including empirical findings:

(4) The inexact RGD (2) for smooth objectives converges to the set of stationary points and has iteration complexity of $O(\varepsilon^{-2})$. See Corollary 1.1.

(5) We give a convergence and complexity result of $O(\varepsilon^{-2})$ for nonsmooth proximal gradient method on Stiefel manifolds. See Section 4.3.

(6) We empirically verified tBMM is faster than the classical algorithm on various problems, including nonnegative tensor decomposition with Riemannian constraints, regularized nonnegative matrix factorization, and low-rank matrix recovery. See Section 5.

1.3. Preliminaries and Notations

The notations we use in this work are consistent with those of Riemannian optimization literature. In this section, we provide a brief introduction to the notations used in our paper, with further details provided in the Appendix A. We use $T_x\mathcal{M}$ or T_x to denote the tangent space at $x \in \mathcal{M}$ and Rtr_x to denote a retraction at 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* $\hat{g} := g \circ \mathrm{Rtr}_x : T_x \to \mathbb{R}$. Denote $\mathrm{r}_{\mathrm{inj}}(x)$ as the injectivity radius at x. For a subset $\mathbf{\Theta} \subseteq \mathcal{M}$ and $x \in \mathbf{\Theta}$, define the *lifted constraint set* T_x^* as

$$T_x^* := \left\{ u \in T_x \;\middle|\; \begin{array}{c} \mathrm{Rtr}_x(u) = x' \text{ for some} \\ x' \in \mathbf{\Theta} \text{ with } d(x, x') \le r_0/2 \end{array} \right\}, \quad (6)$$

where r_0 is the lower bound of the injectivity radius (see (A2)**(iii)**) and $d(x, x')$ is the geodesic distance between x and x'.

For block Riemannian optimization, we introduce the following notations: For $\boldsymbol{\theta} = [\theta^{(1)}, \dots, \theta^{(m)}]$,

$$\operatorname{grad}_i f(\boldsymbol{\theta}) := \begin{bmatrix} \text{Riemmanian gradient of} \\ \theta \mapsto f(\theta^{(1)}, \dots, \theta^{(i-1)}, \theta, \theta^{(i+1)}, \dots, \theta^{(m)}) \end{bmatrix},$$

$$\operatorname{grad} f(\boldsymbol{\theta}) := [\operatorname{grad}_1 f(\boldsymbol{\theta}), \dots, \operatorname{grad}_m f(\boldsymbol{\theta})].$$

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.

2. Algorithm

In Algorithm 1, we give a precise statement of the tBMM algorithm we stated in high-level at (3). We first define majorizing surrogate functions on Riemannian manifolds.

Definition 2.1 (Tangential surrogates on Riemannian manifolds)**.** Fix a function $h : \mathcal{M} \to \mathbb{R}$ and $\boldsymbol{\theta} \in \mathcal{M}$. Denote the *pullback* $\hat{h} := h \circ \operatorname{Rtr}_{\boldsymbol{\theta}} : T_{\boldsymbol{\theta}}\mathcal{M} \to \mathbb{R}$. A function $\hat{g} : T_{\boldsymbol{\theta}}\mathcal{M} \to \mathbb{R}$ is a *tangential surrogate* of h at $\boldsymbol{\theta}$ if

$$\hat{g}(\eta) \geq \hat{h}(\eta) \quad \text{for all } \eta \in T_{\boldsymbol{\theta}}\mathcal{M} \quad \text{and} \quad \hat{g}(\mathbf{0}) = \hat{h}(\mathbf{0}).$$

If $\hat{g} : T_{\boldsymbol{\theta}}\mathcal{M} \to \mathbb{R}$ is a tangential surrogate of $h : \mathcal{M} \to \mathbb{R}$ and if $\hat{g}$, h are differentiable, then $\nabla \hat{g}(\mathbf{0}) = \hat{h}(\mathbf{0}) = \operatorname{grad} h(\boldsymbol{\theta})$. The high-level idea of tBMM is the following. In order to update the ith block of the parameter $\theta_n^{(i)}$ at iteration n, we use first minimize a tangential majorizer of the pullback of the ith block objective function, take a step in the resulting direction in the tangent space, and then use retraction to get back to the manifold. We remark that another formulation of Riemannian BMM (Li et al., 2023) uses majorizers on the manifold without using the tangent spaces. In fact, if we have a majorizer on the manifold, then its pullback is indeed a tangential majorizer. These two formulations of Riemannian BMM coincide on Euclidean spaces but not in general on non-Euclidean manifolds. See the details in Appendix B.

Below we give some remarks on the computational aspects of Algorithm 1. For Algorithm 1, the step size can be simply set as 1 since the lifted constraint set already restricts the length of tangent vectors (see definition in (6)). However, computing the lifted constraint set requires access to specific information about the manifold (see Section I for details). Alternatively, one can minimize the tangent surrogate over the 'tangent cone' (see Appendix A) and apply Riemannian line search in Algorithm 2 to help determine a step size, although it requires additional computational cost.

Algorithm 1 Tangential Block Majorization-Minimization (tBMM)

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); $\alpha \in [0, 1]$ (step size)
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**
    $f_n^{(i)}(\cdot) := f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \cdot, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right)$

$$\begin{cases} \hat{g}_n^{(i)} \leftarrow \left[\text{tangential surrogate of } \varphi_n^{(i)} \text{ at } \theta_{n-1}^{(i)}\right] \\ G_n^{(i)}(\eta) := \hat{g}_n^{(i)}(\eta) + \psi_n^{(i)}(\theta_{n-1}^{(i)} + \eta) \\ V_n^{(i)} \in \arg\min_{\eta \in T^*_{\theta_{n-1}^{(i)}}} G_n^{(i)}(\eta) \\ \alpha_n^{(i)} \leftarrow \alpha; \\ (\text{Optionally, } \alpha_n^{(i)} \leftarrow \text{line search using Alg. 2 }) \\ \theta_n^{(i)} = \operatorname{Rtr}_{\theta_{n-1}^{(i)}}\left(\alpha_n^{(i)} V_n^{(i)}\right) \end{cases} \tag{7}$$

  end for
end for
output: $\boldsymbol{\theta}_N$

3. Statement of Results

Here we state our main results concerning the convergence and complexity of our tBMM algorithm (Alg. 1) for the constrained block Riemannian optimization problem in (5).

3.1. Optimality and Complexity Measures

For iterative algorithms, a first-order optimality condition is unlikely to 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 (5), we say $\boldsymbol{\theta}_n = [\theta_n^{(1)}, \dots, \theta_n^{(m)}] \in \boldsymbol{\Theta}$ generated by Algorithm 1 is an *ε-stationary point* of $f = \varphi + \psi$ over $\boldsymbol{\Theta}$ if

$$\sum_{i=1}^{m} \begin{pmatrix} -\inf_{u \in T^*_{\theta_n^{(i)}}, \|u\| \leq 1} \left\langle \operatorname{grad}_i \varphi(\boldsymbol{\theta}_n) + \right. \\ \left. \operatorname{Proj}_{T_{\theta_n^{(i)}}} \partial \psi_n^{(i)}(\theta_n^{(i)} + V_n^{(i)}),\ \frac{u}{\hat{r}} \right\rangle \end{pmatrix} \leq \varepsilon, \tag{8}$$

where $\hat{r} = \min\{r_0, 1\}$, i.e. the minimum of 1 and the lower bound of injectivity radius r_0; $V_n^{(i)} \in \arg\min_{\eta \in T^*_{\theta_{n-1}^{(i)}}} \left(G_n^{(i)}(\eta) := \hat{g}_n^{(i)}(\eta) + \psi_n^{(i)}(\theta_{n-1}^{(i)} + \eta)\right)$. When the objective function is smooth ($\psi = 0$), (8) reduces

to the optimality measure for Riemannian constrained optimization problems used in (Li et al., 2023). Furthermore, in the Euclidean setting, (8) reduces to the optimality measure used in (Lyu, 2022; Lyu & Li, 2023) for constrained nonconvex smooth optimization and is also equivalent to Def.1 in (Nesterov, 2013) for smooth objectives.

In the unconstrained block-Riemannian setting where $\Theta^{(i)} = \mathcal{M}^{(i)}$ for $i = 1, \dots, m$, the above (8) becomes

$$\sum_{i=1}^{m} \|\mathrm{grad}_i\, \varphi(\boldsymbol{\theta}_n) + \mathrm{Proj}_{T_{\theta_n^{(i)}}} \partial \psi_n^{(i)}(\theta_n^{(i)} + V_n^{(i)})\| \le \varepsilon. \quad (9)$$

In the case of single-block $m = 1$, the above is consistent with the optimality measure discussed in (Chen et al., 2020; Yang et al., 2014). When the nonsmooth part $\psi = 0$, this becomes 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 $(\boldsymbol{\theta}_n)_{n\ge1}$ for solving (5) as

$$N_\varepsilon := \sup_{\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}} \inf \left\{ n \ge 1 \;\middle|\; \begin{array}{c} \boldsymbol{\theta}_n \text{ is an } \varepsilon\text{-approximate} \\ \text{stationary point of } f \text{ over } \boldsymbol{\Theta} \end{array} \right\}, \quad (10)$$

where $(\boldsymbol{\theta}_n)_{n\ge0}$ 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 (10).

3.2. Statement of Results

In this section, we consider solving the minimization problem (5) using Algorithm 1, utilizing tangent spaces and retraction for the majorization and minimization steps. One advantage of tBMM is that it utilizes tangent spaces and retractions so that one can bypass handling Riemannian geometry directly.

We start with stating some general assumptions. We allow inexact computation of the solution to the minimization sub-problems in Algorithm 1. This is practical since the minimization of the tangential surrogates may not always be exactly solvable. To be precise, for each $n \ge 1$, we define the *optimality gap* Δ_n by

$$\Delta_n := \max_{1\le i\le m} \left(G_n^{(i)}(V_n^{(i)}) - \inf_{\eta \in T^*_{\theta_{n-1}^{(i)}}} G_n^{(i)}(\eta) \right). \quad (11)$$

For the convergence analysis to hold, we require that the optimal gaps decay fast enough so that they are summable. See Assumption (A1).

(A1). *For Alg. 1, we make the following assumptions:*

(i) *(Objective) The smooth part of objective $\varphi : \boldsymbol{\Theta} = \prod_{i=1}^m \Theta^{(i)} \to \mathbb{R}$ is (block-wise) continuously differentiable and the possibly nonsmooth part ψ is convex and Lipschitz continuous with parameter L_ψ. The values of objective function f are uniformly lower bounded by some $f^* \in \mathbb{R}$. Furthermore, the sub-level sets $f^{-1}((-\infty, a)) = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : f(\boldsymbol{\theta}) \le a\}$ are compact for each $a \in \mathbb{R}$.*

(ii) *(Inexact computation) The optimality gaps Δ_n in (11) are summable, that is, $\sum_{n=1}^\infty \Delta_n < \infty$.*

(iii) *(Retraction) The retraction* Rtr *satisfies the following: There exists $M_2 > 0$ such that for all $\theta \in \Theta^{(i)}$ and $V \in T_\theta \mathcal{M}^{(i)}$,*

$$\|\mathrm{Rtr}_\theta(V) - (\theta + V)\| \le M_2 \|V\|^2.$$

Note (A1)**(iii)** is a common assumption in the literature of Riemannian optimization, which is used in e.g. (Absil et al., 2007; 2006; Boumal et al., 2019; Liu et al., 2019; Chen et al., 2020). This assumption is trivially satisfied in Euclidean spaces since the retraction becomes the identity there. Furthermore, (A1)**(iii)** holds when the constrained set $\Theta^{(i)}$ is compact, see Prop. 4.3.

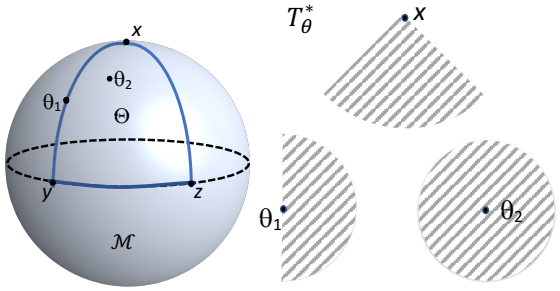

Figure 1. Illustration of the lifted constraint set T_θ^*.

For our analysis we impose the following assumption.

(A2). *For Algorithm 1, we assume the following: Each tangential surrogate $\hat{g}_n^{(i)} : \mathcal{M}^{(i)} \to \mathbb{R}$ of $\varphi_n^{(i)}$ at $\theta_{n-1}^{(i)}$ is differentiable and satisfies the following strong convexity and restricted smoothness properties:*

(i) *(Strong convexity) There exists a constant $\rho_n^{(i)} > 0$ such that for all $\eta, \xi \in T_{\boldsymbol{\theta}}$,*

$$\hat{g}_n^{(i)}(\xi + \eta) \ge \hat{g}_n^{(i)}(\xi) + \langle \nabla \hat{g}_n^{(i)}(\xi),\, \eta \rangle - \frac{\rho_n^{(i)}}{2} \|\eta\|^2.$$

(ii) *(Restricted smoothness) There exists a constant $L_n^{(i)} > 0$ such that for all $\eta \in T_{\boldsymbol{\theta}}$,*

$$|\hat{g}_n^{(i)}(\eta) - \hat{g}_n^{(i)}(\mathbf{0}) - \langle \nabla \hat{g}_n^{(i)}(\mathbf{0}),\, \eta \rangle| \le \frac{L_n^{(i)}}{2} \|\eta\|^2.$$

(iii) *Each $\Theta^{(i)} \subseteq \mathcal{M}^{(i)}$ is (geodesically) strongly convex and there exists a uniform lower bound $r_0 > 0$ for $\mathrm{r_{inj}}(x)$ over $x \in \Theta^{(i)}$. Moreover, the lifted constraint set $T^*_{\boldsymbol{\theta}^{(i)}}$ (see (6)) is convex for each $\theta^{(i)} \in \Theta^{(i)}$.*

The assumption of convexity of the lifted constraint set $T^*_{\boldsymbol{\theta}^{(i)}}$ in (A2)**(iii)** makes the sub-problem of finding $V_n^{(i)}$ in (7) a convex problem, which is easy to solve using classical methods (Boyd et al., 2004; Nesterov, 1998; 2018). For

Euclidean optimization problems, i.e. $\mathcal{M}^{(i)} = \mathbb{R}^{I_n}$ with $I_n \in \mathbb{Z}_+$, the tangent spaces are identical to $\mathbb{R}^{I_n}$. Hence $T^*_{\boldsymbol{\theta}^{(i)}} = \Theta^{(i)}$, whose convexity is already guaranteed by the first part of (A2)**(iii)**. More generally, for unconstrained optimization problems on Hadamard manifolds (see Section 4.1) which include Euclidean spaces, hyperbolic spaces, manifolds of positive definite matrices, when taking Exp as the retraction in (6), we have $T^*_{\boldsymbol{\theta}^{(i)}} = T_{\boldsymbol{\theta}^{(i)}}$ since the exponential map is a global diffeomorphism on Hadamard manifolds (see Theorem 4.1, p.221 in (Sakai, 1996)). Hence the convexity of $T^*_{\boldsymbol{\theta}^{(i)}}$ is guaranteed without further assumptions in this case. In fact, for the unconstrained problem on any Riemannian manifolds, when taking Exp as the retraction in (6), which is the same setting as in (Gutman & Ho-Nguyen, 2023), $T^*_{\boldsymbol{\theta}^{(i)}}$ becomes a ball of radius r_0 centered at $\boldsymbol{\theta}^{(i)}$ so the convexity is also guaranteed. In Figure 1, an illustration of $T^*_{\boldsymbol{\theta}}$ that satisfies (A2)**(iii)** is shown. Additional details about this concrete example can be found in Appendix D.

Our main result for Algorithm 1 is stated in Theorem 3.1.

Theorem 3.1 (Convergence of tBMM)**.** *Let f denote the objective function in (5) with $m \geq 1$. Let $(\boldsymbol{\theta}_n)_{n\geq 0}$ be an output of Algorithm 1. Suppose (A1) and (A2) hold. Let $\hat{r} = \min\{1, r_0\}$. Then the following hold.*

(i) *Suppose there exist constants $\rho, L > 0$ such that for all $n \geq 1$ and $i = 1, \ldots, m$, $\rho_n^{(i)} \geq \rho$, $L_n^{(i)} \leq L$, and $\alpha_n^{(i)} \in (0, \alpha]$ with $\alpha = \rho/(\rho + 2L_\psi M_2)$. Let $M = f(\boldsymbol{\theta}_0) - f^* + m\sum_{n=1}^{\infty}\Delta_n$. Then for $N \geq \frac{L^2}{(L+C')^2}\frac{2M}{\rho\alpha}$,*

$$\min_{1\leq k\leq N}\sum_{i=1}^{m} - \inf_{\eta\in T^*_{\theta^{(i)}_{k-1}},\|\eta\|\leq 1}\Big\langle \operatorname{grad}\varphi_k^{(i)}(\theta_{k-1}^{(i)}) + \operatorname{Proj}_{T_{\theta^{(i)}_{k-1}}}\partial\psi_k^{(i)}(\theta_{k-1}^{(i)} + V_k^{(i)}), \frac{\eta}{\min\{r_0,1\}}\Big\rangle \leq \frac{L+C}{\min\{r_0,1\}}\sqrt{\frac{2M}{\rho\alpha N}}. \tag{12}$$

(ii) *In addition to the hypothesis in* **(i)***, further assume that $\psi = 0$. Then every limit point of $(\boldsymbol{\theta}_n)_{n\geq 0}$ is a stationary point of (5).*

(iii) *In addition to the hypothesis in* **(i)***, further assume φ is joint smooth (Def. A.1). Let $M' = (4\alpha + 6)m\sum_{n=1}^{\infty}\Delta_n + 6(f(\boldsymbol{\theta}_0) - f^*)$. Then for each $N \geq \frac{L^2}{(L+C'+L')^2}\frac{M^2}{M'\rho\alpha}$,*

$$\min_{1\leq k\leq N}\sum_{i=1}^{m} - \inf_{\eta\in T^*_{\theta^{(i)}_{k-1}},\|\eta\|\leq 1}\Big\langle \operatorname{grad}_i\varphi(\boldsymbol{\theta}_{k-1}) + \operatorname{Proj}_{T_{\theta^{(i)}_{k-1}}}\partial\psi_k^{(i)}(\theta_{k-1}^{(i)} + V_k^{(i)}), \frac{\eta}{\min\{r_0,1\}}\Big\rangle \leq \frac{2(L+C+L')}{\min\{r_0,1\}}\sqrt{\frac{M'}{\rho\alpha N}}, \tag{13}$$

In particular, the worst case iteration complexity is $N_\varepsilon = O(\varepsilon^{-2})$.

Note that (12) in Theorem 3.1 **(i)** gives a type of iteration complexity result but the measure of optimality (i.e., the left-hand side of (12)) is adapted to cyclic block optimization algorithms. In order to relate this to the more generic optimality measure in (9), we need to bound the difference between $\operatorname{grad}\varphi_k^{(i)}(\theta_{k-1}^{(i)})$ and $\operatorname{grad}_i\varphi(\boldsymbol{\theta}_{k-1})$. This can be achieved by the additional smoothness property of φ as in (24) and the resulting iteration complexity result is stated in (13) in Theorem 3.1 **(iii)**. We put the proof of Theorem 3.1 along with some key lemmas in Appendix I.

4. Applications

In this section, we discuss some examples of tBMM and its connection to some other classical algorithms.

4.1. Example Manifolds

In this section, we list several examples of manifolds that appear in various machine-learning problems.

Example 4.1 (Stiefel manifold)**.** The Stiefel manifold $\mathcal{V}^{n\times k}$ is the set of all orthonormal k-frames in $\mathbb{R}^n$. Namely, it is the set of all ordered orthonormal k-tuples of vectors in $\mathbb{R}^n$, i.e. $\mathcal{V}^{n\times k} = \{A \in \mathbb{R}^{n\times k} : A^TA = I_k\}$, where I_k is the $k \times k$ identity matrix. For $X \in \mathbb{R}^{n\times k}$, let $X = U\Sigma V^T$ be the SVD, then the projection of X onto $\mathcal{V}^{n\times k}$ is given by $\operatorname{Proj}_{\mathcal{V}^{n\times k}}(X) = UV^T$.

Example 4.2 (Manifold of fixed-rank matrices)**.** The set $\mathcal{R}_r = \{X \in \mathbb{R}^{n\times m} : \operatorname{rank}(X) = r\}$ is a smooth submanifold of $\mathbb{R}^{n\times m}$. For $X \in \mathbb{R}^{n\times m}$, let $X = U\Sigma V^T$ be the SVD, and let u_i and v_i be the i-th column of U and V, respectively. Write the singular values in Σ in nonincreasing order, $\sigma_1(X) \geq \sigma_2(X) \geq \cdots \geq \sigma_{\min\{n,m\}}(X) \geq 0$. Then the projection of X onto $\mathcal{R}_r$ is given by

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

Another class of manifolds that is widely studied in the literature is the Hadamard manifolds, which includes many commonly encountered manifolds. *Hadamard manifolds* are complete and simply connected Riemannian manifolds with nonpositive sectional curvature ((Burago et al., 2001) and (Burago et al., 1992)). The injectivity radius at every point is infinity on a Hadamard manifold. Examples of Hadamard manifolds can be found in Appendix E.

4.2. Block Riemannian Prox-linear Updates and Block Riemannian GD

A primary approach in tBMM is to use the Riemannian prox-linear surrogates. We begin our discussion from the

single-block case for (1) without constraints (i.e., $\boldsymbol{\Theta} = \mathcal{M}$) for simplicity. Given the iterate $\boldsymbol{\theta}_{n-1} \in \mathcal{M}$, the Riemannian prox-linear surrogate $\hat{g} : T_{\boldsymbol{\theta}_{n-1}}\mathcal{M} \to \mathbb{R}$ takes the form

$$\hat{g}_n(\eta) := f(\boldsymbol{\theta}_{n-1}) + \langle \operatorname{grad} f(\boldsymbol{\theta}_{n-1}),\, \eta\rangle + \frac{\lambda}{2}\|\eta\|^2, \quad (14)$$

where $\lambda > 0$ is the proximal regularization parameter. Note that $\hat{g}$ is defined only on the tangent space at $\boldsymbol{\theta}_{n-1}$ and not on the manifold. In order for $\hat{g}_n$ to be a tangential surrogate of f at $\boldsymbol{\theta}_{n-1}$, one can impose the following restricted smoothness property for the pullback objective $\hat{f} := f \circ \operatorname{Rtr}_{\boldsymbol{\theta}_{n-1}}$: For some constant $L > 0$,

$$\left|\hat{f}(\eta) - \hat{f}(\mathbf{0}) - \langle \operatorname{grad} f(\boldsymbol{\theta}_{n-1}),\, \eta\rangle\right| \le \frac{L}{2}\|\eta\|^2. \quad (15)$$

The above property was introduced in (Boumal et al., 2019) under the name of 'restricted Lipschitz-type gradients', and was critically used to analyze Riemannian gradient descent and Riemannian trust region methods. Hence under (15) and assuming $\lambda \ge L$, $\hat{g}_n$ in (14) is indeed a tangential majorizer of f at $\boldsymbol{\theta}_{n-1}$. The resulting tBMM with the Riemannian prox-linear surrogate above reads as

$$\boldsymbol{\theta}_n \leftarrow \operatorname{Rtr}_{\boldsymbol{\theta}_{n-1}}\left(-\frac{1}{\lambda}\operatorname{grad} f(\boldsymbol{\theta}_{n-1})\right),$$

which is the standard Riemannian gradient descent with a fixed step size $1/\lambda \le 1/L$. Our convergent result Theorem 3.1 holds for the above updates whenever $\lambda \ge L$.

Next, when we further impose an additional strongly convex constraint $\boldsymbol{\Theta} \subseteq \mathcal{M}$, the decent direction is not necessarily the negative Riemannian gradient $-\operatorname{grad} f(\boldsymbol{\theta}_{n-1})$ and one needs to solve a quadratic minimization over the tangent space. The portion of tBMM in this case reads as

$$\begin{cases} V_n \leftarrow \arg\min_{\eta \in T^*_{\boldsymbol{\theta}_{n-1}}} \left\|\eta - \left(-\frac{1}{\lambda}\operatorname{grad} f(\boldsymbol{\theta}_{n-1})\right)\right\|^2 \\ \qquad = \operatorname{Proj}_{T^*_{\boldsymbol{\theta}_{n-1}}}\left(-\frac{1}{\lambda}\operatorname{grad} f(\boldsymbol{\theta}_{n-1})\right), \\ \boldsymbol{\theta}_n \leftarrow \operatorname{Rtr}_{\boldsymbol{\theta}_{n-1}}(\alpha_n V_n). \end{cases} \quad (16)$$

where α_n is the step size that could be either a small enough constant or decided by Algorithm 2. Namely, the constrained descent direction V_n is the projection of the unconstrained descent direction $-\frac{1}{\lambda}\operatorname{grad} f(\boldsymbol{\theta}_{n-1})$ onto the lifted constraint set $T^*_{\boldsymbol{\theta}_{n-1}}$. The above is a Riemannian version of the projected gradient descent that our tBMM algorithm entails. As in the unconstrained case, our convergent result Theorem 3.1 holds for the above updates whenever $\lambda \ge L$.

We can apply a similar approach for the constrained block Riemannian optimization problem (5). Namely, for each iteration n and block i, use the following Riemannian prox-linear surrogate

$$\hat{g}_n^{(i)}(\eta) := f_n^{(i)}(\theta_{n-1}^{(i)}) + \langle \operatorname{grad} f_n^{(i)}(\theta_{n-1}^{(i)}),\, \eta\rangle + \frac{\lambda}{2}\|\eta\|^2. \quad (17)$$

Then the resulting portion of tBMM reads as the following block projected Riemannian gradient descent: For $i = 1, \ldots, m$,

$$\begin{cases} V_n^{(i)} \leftarrow \operatorname{Proj}_{T^*_{\theta_{n-1}^{(i)}}}\left(-\frac{1}{\lambda}\operatorname{grad} f(\theta_{n-1}^{(i)})\right), \\ \theta_n^{(i)} \leftarrow \operatorname{Rtr}_{\theta_{n-1}^{(i)}}(\alpha_n^{(i)} V_n^{(i)}), \end{cases}$$

where $\alpha_n^{(i)}$ is the step size that could be either a small enough constant or decided by Algorithm 2.

It is worth mentioning that utilizing the Riemannian gradient "grad" instead of the full Euclidean gradient "∇" can provide significant computational savings. One extensively investigated example is RPGD on fixed-rank manifolds (Example 4.2). There, computing $\operatorname{Proj}_{\mathcal{R}_r}\left(\theta_{n-1}^{(i)} - \frac{1}{\lambda}\nabla\right)$ involves computing the full SVD of a $n \times m$ matrix, while $\operatorname{Proj}_{\mathcal{R}_r}\left(\theta_{n-1}^{(i)} - \frac{1}{\lambda}\operatorname{grad}\right)$ only involves computing the SVD of a much smaller $2r \times 2r$ matrix (see e.g. (Wei et al., 2016)) when $r \ll \min\{m, n\}$.

4.3. Nonsmooth Proximal Gradient Method on the Stiefel Manifold (Chen et al., 2020)

In (Chen et al., 2020), the authors focus on the nonsmooth optimization problem on the Stiefel manifold as follows,

$$\min_{\boldsymbol{\theta} \in \mathcal{V}^{n\times p}} f(\boldsymbol{\theta}) := \phi(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}), \quad (18)$$

where ϕ is Euclidean smooth with parameter L_ϕ, ψ is convex and Lipschitz continuous with parameter L_ψ but is not necessarily Euclidean smooth. In the paper, the authors propose the following algorithm to solve (18),

$$\begin{cases} G(V) := \phi(\boldsymbol{\theta}_{n-1}) + \langle \operatorname{grad}\phi(\boldsymbol{\theta}_{n-1}), V\rangle \\ \quad +\lambda\|V\|_F^2 + \psi(\boldsymbol{\theta}_{n-1} + V) \\ V_n = \arg\min_{V \in T_{\boldsymbol{\theta}_{n-1}}} G(V) \\ \boldsymbol{\theta}_n = \operatorname{Rtr}_{\boldsymbol{\theta}_{n-1}}(\alpha V_n), \end{cases} \quad (19)$$

where $\alpha > 0$ is a step size. Recall the Stiefel manifold is compact, in order to compare and bound the difference of vectors before and after retraction, the authors in (Chen et al., 2020) use the following proposition substantially to establish the complexity result.

Proposition 4.3 (Compact manifold embedded in Euclidean space (Boumal et al., 2019; Liu et al., 2019))**.** *Let $\mathcal{M}$ be a compact Riemannian manifold embedded in an Euclidean space with norm $\|\cdot\|$. Then there exist constants $M_1, M_2 > 0$ such that for all $\theta \in \mathcal{M}^{(i)}$ and $V \in T_\theta\mathcal{M}$,*

$$\|\operatorname{Rtr}_\theta(V) - \theta\| \le M_1\|V\|,$$
$$\|\operatorname{Rtr}_\theta(V) - (\theta + V)\| \le M_2\|V\|^2.$$

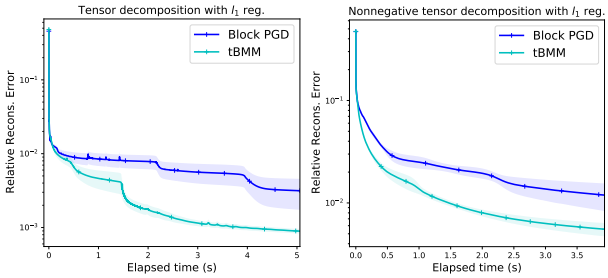

Figure 2. Comparison of tBMM and block projected gradient descent on (nonnegative) tensor decomposition problems with l_1 regularizations. Relative errors with standard deviation are shown by the solid lines and shaded regions of respective colors.

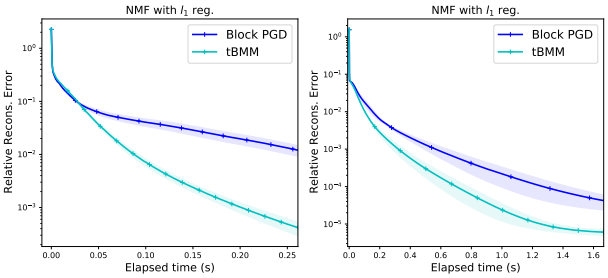

Figure 3. Comparison of tBMM and block projected gradient descent on nonnegative matrix factorization problems with l_1 regularizations. Relative errors with standard deviation are shown by the solid lines and shaded regions of respective colors.

In fact, when replacing the compact manifold with a compact constrained set on an arbitrary embedded submanifold, the bounds in Proposition 4.3 still hold. This is one situation that our assumption in (A1)**(iii)** holds. When ϕ in (18) is L_ϕ-smooth, the algorithm (19) with proper regularization parameter λ falls into our tBMM framework, and therefore we can apply the convergence and complexity results in Theorem 3.1. The complexity results are formally stated in the following corollary. The proof is in Appendix H.

Corollary 4.4 (Complexity of nonsmooth proximal gradient method on Stiefel manifolds)**.** *Let ϕ in (18) be L_ϕ-smooth. Then the complexity results in Theorem 3.1 hold for the algorithm (19) when $\lambda \geq L_\phi$. In particular, the algorithm (19) has iteration complexity of $O(\varepsilon^{-2})$.*

We remark that in (Chen et al., 2020), the authors established complexity results based on a different definition of ε-stationary point, namely, $\boldsymbol{\theta}_n$ is ε-stationary point if $\|V_n\| \leq \varepsilon/L_\phi$. Though this definition brings practical benefits, it is not a generic definition. The definition of ε-stationary point (8) we used is more generic, and is a natural generalization of the corresponding concept in the Euclidean setting. Moreover, our definition allows further constraints on the manifold, i.e. the constraint set $\boldsymbol{\Theta}$ is not necessarily the entire manifold $\mathcal{M}$.

5. Numerical Validation

In this section, we provide numerical validation of tBMM on various problems. Additional details of the experiments are provided in Appendix J.

Nonnegative Tensor Decomposition with Riemannian Constraints. Given a tensor $\mathbf{X} \in \mathbb{R}^{I_1 \times \cdots \times I_m}$ and a fixed integer $R \geq 1$, the tensor decomposition problem aims to find the loading matrices $U^{(i)} \in \mathbb{R}^{I_i \times R}$ for $i = 1, \ldots, m$ such that the sum of the outer products of their respective columns approximate $\mathbf{X}$: $\mathbf{X} \approx \sum_{k=1}^{R} \bigotimes_{i=1}^{m} U^{(i)}[:, k]$, where $U^{(i)}[:, k]$ denotes the k^{th} column of the $I_i \times R$ loading matrix $U^{(i)}$ and $\bigotimes$ denotes the outer product. One can also impose Riemannian constraints on the loading matrices, such that each $U^{(i)} \in \mathcal{M}^{(i)}$ resides on a Riemannian manifold. This Riemannian tensor decomposition problem can be formulated as the following optimization problem:

$$\underset{U^{(i)} \in \Theta^{(i)}}{\arg\min} \left(f(U^{(1)}, \ldots, U^{(m)}) := \left\| \mathbf{X} - \sum_{k=1}^{R} \bigotimes_{i=1}^{m} U^{(i)}[:, k] \right\|_F^2 + \sum_{i=1}^{m} \lambda_i \|U^{(i)}\|_1 \right), \tag{20}$$

where $\Theta^{(i)} \subseteq \mathcal{M}^{(i)}$ denotes a closed and geodesically convex constraint set and $\lambda_i \geq 0$ is the hyperparameter of the ℓ_1-regularizer.

In our experiments shown in Figure 2, we let the first block $\mathcal{M}^{(1)}$ be the fixed-rank manifold $\mathcal{R}_r$ and the other blocks be Euclidean spaces. We consider two different cases: with or without nonnegativity constraints for each loading matrix. In both cases, we use the prox-linear surrogates (see Sec. 4.3) for tBMM. Due to the ℓ_1 regularizer in (20), the subproblem of minimizing the tangential surrogate $\hat{G}$ does not have a closed-form solution. Furthermore, when nonnegativity is imposed, another projection onto the positive orthant is required when solving the subproblems. Hence, one needs to implement an iterative algorithm for solving the subproblems. This brings inexactness to the solution to the subproblems. In Figure 2, the left figure is the reconstruction error plot of solving problem (20) without nonnegativity constraints, and the right figure is that with nonnegativity constraints. In both cases, tBMM outperforms block projected gradient descent (block PGD).

Regularized Nonnegative Matrix Factorization with Riemannian Constraints. Given a data matrix $X \in \mathbb{R}^{p \times N}$, the constrained matrix factorization problem is formulated as the following,

$$\underset{W \in \mathcal{M}^{(1)} \subseteq \mathbb{R}^{p \times R},\, H \in \mathcal{M}^{(2)} \subseteq \mathbb{R}^{R \times N}}{\arg\min} \|X - WH\|_F^2 + \lambda \|H\|_1.$$

We consider a similar setting as the tensor decomposition problem in the last section. Namely, the first block $\mathcal{M}^{(1)}$ is a fixed-rank manifold $\mathcal{R}_r$ with nonnegativity constraint

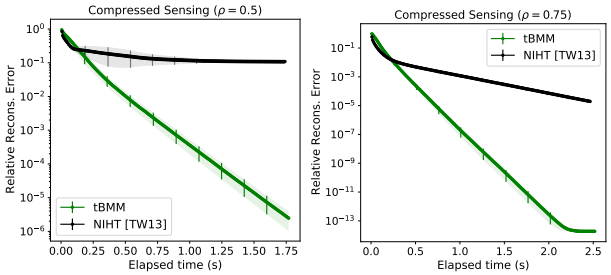

Figure 4. Comparison of tBMM and NIHT on the problem of low-rank matrix recovery. Relative errors with standard deviation are shown by the solid lines and shaded regions of respective colors.

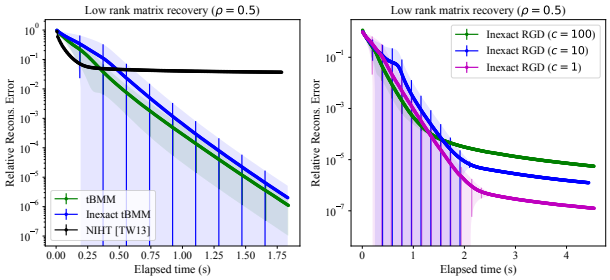

Figure 5. Illustration of the convergence of inexact RGD on the low-rank matrix recovery problem. Relative errors with standard deviation are shown by the solid lines and shaded regions.

and the second block is Euclidean space. Similarly, the nonnegativity constraints and the ℓ_1 regularizer make the subproblems have no closed-form solution. Therefore, the iterative algorithms for solving it bring inexactness to the solution to the subproblem. In Figure 3, we show the error plot with slightly different settings. In the left figure, there is no nonnegativity constraint to the second block H. In the right figure, we impose a nonnegativity constraint on both blocks. In both cases, tBMM using prox-linear surrogates outperforms block PGD.

Low-rank Matrix Recovery. Consider the *low-rank matrix recovery* problem (Donoho, 2006; Candes & Plan, 2009) formulated as the following in (Tanner & Wei, 2013; Wei et al., 2016),

$$\min_{X \in \mathcal{R}_r \subseteq \mathbb{R}^{m \times n}} \left(f(X) := \frac{1}{2} \|\mathcal{A}(X) - b\|^2 \right) \tag{21}$$

where $\mathcal{R}_r$ is the manifold of fixed-rank matrices, see Example 4.2. The linear map $\mathcal{A} : \mathbb{R}^{m \times n} \to \mathbb{R}^p$ maps an $m \times n$ matrix to a p dimensional vector, given by

$$\mathcal{A}(X)_i := \langle A_i, X \rangle, \quad \text{for } i = 1, \cdots, p.$$

The p-dimensional vector b represents the observations. Given the observations from the sensing operator $\mathcal{A}$, low-rank matrix recovery aims to find the fixed-rank matrix X. We define the umdersampling ratio as $\rho = p/mn$.

We compare the performance of the proposed tBMM (Algorithm 1) applied to the low-rank matrix recovery problem against *normalized iterative hard thresholding* (NIHT) (Tanner & Wei, 2013), see details in Appendix J.3. To solve (21), we use the prox-linear surrogates (17) for tBMM (Algorithm 1) as discussed in Section 4.2. The resulting updates for this single block problem are shown in (16).

In Figure 4, we compare the performance of tBMM and NIHT for $\rho = 0.5$ and 0.75 respectively. Under both settings, tBMM outperforms NIHT in terms of relative reconstruction error. Moreover, tBMM is observed to be more robust for different dimensions of the matrices. The better performance of tBMM over NIHT can be attributed to two main factors: Firstly, as discussed in Section 4.2, NIHT needs to solve a full SVD of a matrix of dimension $m \times n$ in each iteration, while tBMM only needs to solve that of a $2r \times 2r$ matrix, which provides computational savings. Secondly, the performance of NIHT is affected by the *oversampling ratio* defined as $\mu = r(m + n - r)/p$. When $\mu > 1/2$, the recovery of all low-rank matrices by the algorithm is impossible (Tanner & Wei, 2013; Eldar et al., 2011), which leads to the relatively poor performance of NIHT in the left panel of Figure 4.

Inexact RGD. In the introduction, we state that the convergence and complexity of inexact RGD can be deduced as a corollary of our Theorem 3.1. Here, we provide numerical validations based on the low-rank matrix recovery problem. In Figure 5, we verify the performance of inexact tBMM and inexact RGD. We pose inexactness by adding a noise term $\Delta_n = c/(n+1)^2$ to the Riemannian gradient, where n is the iteration number. The Δ_ns are summable, which satisfies (A1)**(ii)**. The inexact tBMM and inexact RGD show compelling convergence speed.

6. Conclusion and Limitations

In this paper, we proposed a general framework named tBMM which entails many classical first-order Riemannian optimization algorithms, including the inexact RGD and the proximal gradient method on Stiefel manifolds. tBMM is applicable to solving multi-block Riemannian optimization problems, and can handle additional (geodesically convex) constraints within each manifold, as well as nonsmooth nonconvex objectives. We established the convergence and complexity results of tBMM. Namely, tBMM converges to a ε-stationary point within $O(\varepsilon^{-2})$ iterations. The convergence and complexity results still hold when the subproblem in each iteration is computed inexactly, as long as the optimality gap is summable. We validated our theoretical results on tBMM through various numerical experiments. We also demonstrated that tBMM can show improved performance on various Riemannian optimization problems compared to existing methods.

Impact Statement

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

Acknowledgements

We thank all the reviewers for the thoughtful comments and suggestions. 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. LB is partially supported by NSF CAREER award CCF-1845076 and ARO YIP award W911NF1910027.

References

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

Absil, P., Baker, C., and Gallivan, K. Convergence analysis of riemannian trust-region methods, 2006.

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

Absil, P.-A., Baker, C. G., and Gallivan, K. A. Trust-region methods on riemannian manifolds. *Foundations of Computational Mathematics*, 7:303–330, 2007.

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009. ISBN 9781400830244.

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

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

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

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

Boumal, N., Absil, P.-A., and Cartis, C. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Burago, D., Burago, Y., and Ivanov, S. *A Course in Metric Geometry*. Crm Proceedings & Lecture Notes. American Mathematical Society, 2001. ISBN 9780821821299.

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

Candes, E. J. and Plan, Y. Accurate low-rank matrix recovery from a small number of linear measurements. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1223–1230. IEEE, 2009.

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

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

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

Chen, S., Garcia, A., Hong, M., and Shahrampour, S. Decentralized riemannian gradient descent on the stiefel manifold. In *International Conference on Machine Learning*, pp. 1594–1605. PMLR, 2021.

Do Carmo, M. P. and Flaherty Francis, J. *Riemannian geometry*, volume 6. Springer, 1992.

Donoho, D. L. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

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

Eldar, Y. C., Needell, D., and Plan, Y. Unicity conditions for low-rank matrix recovery. *arXiv preprint arXiv:1103.5479*, 2011.

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

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

Hong, M., Razaviyayn, M., Luo, Z.-Q., and Pang, J.-S. 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.

Jaquier, N., Rozo, L., Calinon, S., and Bürger, M. Bayesian optimization meets riemannian manifolds in robot learning. In *Conference on Robot Learning*, pp. 233–246. PMLR, 2020.

Kwon, D. and Lyu, H. Complexity of block coordinate descent with proximal regularization and applications to wasserstein cp-dictionary learning. *arXiv preprint arXiv:2306.02420*, 2023.

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

Li, Y., Balzano, L., Needell, D., and Lyu, H. Convergence and complexity of block majorization-minimization for constrained block-riemannian optimization. *arXiv preprint arXiv:2312.10330*, 2023.

Liu, H., So, A. M.-C., and Wu, W. Quadratic optimization with orthogonality constraint: explicit łojasiewicz exponent and linear convergence of retraction-based line-search and stochastic variance-reduced gradient methods. *Mathematical Programming*, 178:215–262, 2019.

Lyu, H. Convergence and complexity of stochastic block majorization-minimization. *arXiv preprint arXiv:2201.01652*, 2022.

Lyu, H. and Li, Y. Block majorization-minimization with diminishing radius for constrained nonconvex optimization. *arXiv preprint arXiv:2012.03503*, 2023.

Mairal, J. Optimization with first-order surrogate functions. In *International Conference on Machine Learning*, pp. 783–791, 2013.

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

Nesterov, Y. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 3(4):5, 1998.

Nesterov, Y. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

Nesterov, Y. *Lectures on Convex Optimization*. Springer Optimization and Its Applications. Springer International Publishing, 2018. ISBN 9783319915784.

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

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

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

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

Song, G., Ng, M. K., and Jiang, T.-X. Tangent space based alternating projections for nonnegative low rank matrix approximation. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Song, G.-J. and Ng, M. K. Nonnegative low rank matrix approximation for nonnegative matrices. *Applied Mathematics Letters*, 105:106300, 2020.

Sun, J., Qu, Q., and Wright, J. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

Tanner, J. and Wei, K. Normalized iterative hard thresholding for matrix completion. *SIAM Journal on Scientific Computing*, 35(5):S104–S125, 2013.

Wang, X., Tu, Z., Hong, Y., Wu, Y., and Shi, G. No-regret online learning over riemannian manifolds. *Advances in Neural Information Processing Systems*, 34:28323–28335, 2021.

Wei, K., Cai, J.-F., Chan, T. F., and Leung, S. Guarantees of riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37 (3):1198–1222, 2016.

Wiersema, R. and Killoran, N. Optimizing quantum circuits with riemannian gradient flow. *Physical Review A*, 107 (6):062421, 2023.

Xu, Y. and Yin, W. 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.

Yang, W. H., Zhang, L.-H., and Song, R. Optimality conditions for the nonlinear programming problems on riemannian manifolds. *Pacific Journal of Optimization*, 10(2): 415–434, 2014.

Yang, Y. 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.

A. Preliminaries on Riemannian Optimization

In this paper, we use the same notations as the common Riemannian optimization literature, see e.g. (Absil et al., 2009) and (Boumal, 2023). We refer the readers to (Sakai, 1996), (Lee, 2003), (Do Carmo & Flaherty Francis, 1992), and (Helgason, 1979) for the background knowledge on Riemannian geometry. Below we give some preliminaries on Riemannian optimization.

We call a manifold to be a *Riemannian manifold* if it is equipped with a Riemannian metric $(\xi, \eta) \mapsto \langle \xi, \eta \rangle_x \in \mathbb{R}$, where the tangent vectors ξ and η are on the tangent space $T_x\mathcal{M}$. We denote $T_x\mathcal{M}$ as T_x when the corresponding manifold is clear from the context. Moreover, we drop the subscript x of the inner product $\langle \cdot, \cdot \rangle_x$ when it is clear from the context. This inner product induces a norm on the tangent space, which is denoted by $\| \cdot \|_x$ or $\| \cdot \|$. The geodesic distance generalizes the concept of distance by measuring the shortest path between two points on a curved surface, accounting for the geometry of the manifold. We denote the geodesic distance between $x, y \in \mathcal{M}$ by $d_{\mathcal{M}}(x, y)$ or simply $d(x, y)$. An important concept in Riemannian optimization is the *Riemannian gradient* of a smooth function $f : \mathcal{M} \to \mathbb{R}$ denoted as $\operatorname{grad} f(x)$ at $x \in \mathcal{M}$. It is defined as the tangent vector satisfying $\langle \operatorname{grad} f(x), \xi_x \rangle = \mathrm{D}f(x)[\xi_x], \forall \xi_x \in T_x\mathcal{M}$. Here $\mathrm{D}f(x)[\xi_x]$ is the directional derivative along the direction ξ_x. As aforementioned in the introduction, the widely used Riemannian gradient descent algorithms use retractions to map the Riemannian gradient to the manifold. A *retraction* denoted as Rtr is a smooth mapping from the tangent bundle $T\mathcal{M}$ to $\mathcal{M}$ that satisfying the following key properties.

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

(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 $x \in \mathcal{M}$ and $\eta \in T_x\mathcal{M}$, the retraction curve $t \mapsto \mathrm{Rtr}_x(t\eta)$ agrees with the geodesic passing x with initial velocity η to the first order. Furthermore, if this retraction curve coincides with the geodesic for all $x \in \mathcal{M}$ and $\eta \in T\mathcal{M}$, then this specific retraction is called an *exponential map*. In fact, the definition of exponential map involves solving a nonlinear ordinary differential equation, which brings computational burden. Hence, using computationally efficient retractions instead of exponential map is more desirable. The widely used retractions including $\mathrm{Rtr}_x(\eta) = x + \eta$ in Euclidean spaces and $\mathrm{Rtr}_x(\eta) = \frac{x+\eta}{\|x+\eta\|}$ on spheres. For more examples and details, we refer the readers to Sec. 4.1 in (Absil et al., 2009). A manifold $\mathcal{M}$ is called (geodesically) *complete* if the domain of exponential map is the entire tangent bundle. The exponential map by its definition preserves distance, i.e. $\|\eta\| = d(x, y)$ for $\mathrm{Exp}_x(\eta) = y$.

A function $g : \mathcal{M} \to \mathbb{R}$ can be lifted to the tangent spaces via retractions. Namely, consider the composition of g and a retraction Rtr, define the *pullback* as $\hat{g} := g \circ \mathrm{Rtr}_x : T_x \to \mathbb{R}$. In tBMM, we seek a majorizing surrogate of the lifted marginal loss function to minimize in each iteration. One important observation used in the analysis is the following,

$$\langle \nabla \hat{g}(\mathbf{0}),\, \eta \rangle = D\hat{g}(\mathbf{0})[\eta] = Dg(x)[D\,\mathrm{Rtr}_x(\mathbf{0})[\eta]] = Dg(x)[\eta] = \langle \operatorname{grad} g(x),\, \eta \rangle. \tag{22}$$

At each point $x \in \mathcal{M}$, the Riemannian manifold $\mathcal{M}$ resembles an Euclidean space within a small metric ball with radius r since it is locally diffeomorphic to the tangent space. The *injectivity radius*, denoted as $\mathrm{r}_{\mathrm{inj}}(x)$, is defined as the supremum of values of r such that this diffemorphism holds. Namely, Exp_x is a diffemorphism of $B(x, r)$ and its image on $\mathcal{M}$ when $r \leq \mathrm{r}_{\mathrm{inj}}(x)$ where $B(x, r) = \{\eta \in T_x\mathcal{M} : \langle \eta, \eta \rangle < r\} \subseteq T_x\mathcal{M}$. Therefore, the inverse exponential map Exp_x^{-1} is well defined within the small ball of radius $\mathrm{r}_{\mathrm{inj}}(x)$. Many widely studied manifolds have uniformly positive injectivity radius, including compact manifolds (see Thm. III.2.3 in (Chavel, 2006)) and Hadamard manifolds (see Appendix E, (Afsari, 2011) and Theorem 4.1, p.221 in (Sakai, 1996)). We call a subset $C \subseteq M$ to be (geodesically) *strongly convex* if there is a unique minimal geodesic γ connecting any two points $x, y \in C$ and that $\gamma \subseteq C$.

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

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

$$\mathcal{N}_{\mathbf{\Theta}}(x) := \left\{ u \in T_x\mathcal{M} \,\middle|\, \langle u, \eta \rangle \leq 0 \text{ for all } \eta \in T_x\mathcal{M} \text{ s.t. } \mathrm{Exp}_x\left(t\frac{\eta}{\|\eta\|}\right) \in \mathbf{\Theta} \text{ for some } t \in (0, \mathrm{r}_{\mathrm{inj}}(x)) \right\}.$$

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

The *lifted constraint set* $T_x^*\mathcal{M}$ is defined as

$$T_x^*\mathcal{M} := \{u \in T_x\mathcal{M} \,|\, \mathrm{Rtr}_x(u) = x' \text{ for some } x' \in \boldsymbol{\Theta} \text{ with } d(x, x') \le r_0/2\}, \tag{23}$$

where r_0 is the uniform lower bound of injectivity as in (A2)**(iii)**. See an example of the lifted constraint set in Appendix D and Figure 1. In (23), when exponential map is used as the retraction, the set $T_x^*\mathcal{M}$ can be viewed as the 'lift' of the restricted constraint set $\boldsymbol{\Theta}$ within a small ball to the tangent space $T_x\mathcal{M}$, which equals the image $\mathrm{Exp}_x^{-1}(\boldsymbol{\Theta} \cap B(x, r_0/2))$ by the inverse exponential map. For the special case when $\mathcal{M}$ is an Euclidean space, then the retraction becomes identity. Therefore if $\boldsymbol{\Theta}$ is a convex subset, we have $T_x^*\mathcal{M} = \boldsymbol{\Theta}$ is also convex. We remark that if $\boldsymbol{\Theta}$ is (geodesically) strongly convex, then the lifted constraint set $T_x^*\mathcal{M}$ is locally well defined, i.e. we can replace r_0 in (23) by any value $r' \in (0, r_0)$.

At different points on a Riemannian manifold, the corresponding tangent spaces are different. The *parallel transport* allows one to transport a tangent vector to another tangent space. For a smooth curve $\gamma : [0, 1] \to \mathcal{M}$, the parallel transport along γ from the point $x = \gamma(0)$ to the point $y = \gamma(1)$ is denoted as $\Gamma^{\gamma}_{x\to y}$. We drop the superscript γ when it is clear from the context. Intuitively, for a tangent vector $\eta \in T_x\mathcal{M}$, the vector $\Gamma^{\gamma}_{x\to y}\eta$ is a tangent vector in $T_y\mathcal{M}$. The key property of parallel transport is that it is a linear isomorphism that preserves inner product. Namely,

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

A.1. Notations for Block Riemannian Optimization

In (5), we aim to minimize an objective function f within the product constraint set $\boldsymbol{\Theta} = \Theta^{(1)} \times \cdots \times \Theta^{(m)}$, where each $\Theta^{(i)}$ is a subset on a Riemannian manifold $\mathcal{M}^{(i)}$. For convenience, we introduce the following notations: For $\boldsymbol{\theta} = [\theta^{(1)}, \ldots, \theta^{(m)}]$, let

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

We remark that one can endow the product manifold $\prod_{i=1}^m \mathcal{M}^{(i)}$ a joint Riemannian structure, and therefore the above $\mathrm{grad}\, f(\boldsymbol{\theta})$ can be viewed as the full Riemannian gradient at $\boldsymbol{\theta}$ w.r.t the joint Riemannian structure. However, in the present paper we do not explicitly use the product manifold structure.

Throughout the paper, we denote $(\boldsymbol{\theta}_n)_{n\ge1}$ as an output of Algorithm 1 and write $\boldsymbol{\theta}_n = [\theta_n^{(1)}, \ldots, \theta_n^{(m)}]$ for $n \ge 1$. For each $n \ge 1$ and $i = 1, \ldots, m$, denote the marginal objective function as

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

which is at iteration n and block i.

Below we define the *joint smoothness* used in Thm.3.1**(iii)**.

Definition A.1 (Joint smoothness)**.** Let $\mathcal{M} = \mathcal{M}^{(1)} \times \cdots \times \mathcal{M}^{(m)}$ be a product manifold where all the $\mathcal{M}^{(i)}$s are smooth Riemannian manifold. Let $\varphi : \mathcal{M} \to \mathbb{R}$ be a smooth function in each block. We say φ is *jointly smooth* if there exists a constant $L' > 0$ such that for each distinct $i, j \in \{1, \ldots, m\}$ and $(\theta^{(1)}, \ldots, \theta^{(m)}) \in \boldsymbol{\Theta}$, whenever $V \in T_{\theta^{(j)}}$ such that $\mathrm{Rtr}_{\theta^{(j)}}(V) \in \Theta^{(j)}$,

$$\left\| \begin{array}{l} \mathrm{grad}_i\, \varphi(\theta^{(1)}, \ldots, \theta^{(j)}, \ldots, \theta^{(m)}) \\ \qquad - \mathrm{grad}_i\, \varphi(\theta^{(1)}, \ldots, \mathrm{Rtr}_{\theta^{(j)}}(V), \ldots, \theta^{(m)}) \end{array} \right\| \le L'\|V\|. \tag{24}$$

B. Relation Between Types of Majorizing Surrogates

In (Li et al., 2023), the authors studied the convergence of *RBMM*, which is a Riemannian block majorization-minimization algorithm using the majorizing surrogates on the manifold. A function $g : \mathcal{M} \to \mathbb{R}$ is a *majorizing surrogate* of h at $\boldsymbol{\theta}$ if

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

We remark that if a function $g : \mathcal{M} \to \mathbb{R}$ is a majorizing surrogate of $h : \mathcal{M} \to \mathbb{R}$ at some point $\boldsymbol{\theta} \in \mathcal{M}$, then the pullback surrogate $\hat{g} := g \circ \mathrm{Rtr}_\theta : T_{\boldsymbol{\theta}}\mathcal{M} \to \mathbb{R}$ is a tangential surrogate of h at $\boldsymbol{\theta}$. Indeed,

$$\hat{g}(\eta) = g(\mathrm{Rtr}_{\boldsymbol{\theta}}(\eta)) \geq h(\mathrm{Rtr}_{\boldsymbol{\theta}}(\eta)) = \hat{h}(\eta) \ , \forall \eta \in T_{\boldsymbol{\theta}}\mathcal{M},$$
$$\hat{g}(\mathbf{0}) = g(\mathrm{Rtr}_{\boldsymbol{\theta}}(\mathbf{0})) = g(\boldsymbol{\theta}) = h(\boldsymbol{\theta}) = h(\mathrm{Rtr}_{\boldsymbol{\theta}}(\mathbf{0})) = \hat{h}(\mathbf{0}).$$

Note that minimizing the majorizer g directly on the manifold $\mathcal{M}$ and minimizing its pullback $\hat{g}$ on the tangent space are in general two different optimization problems. However, if the manifold $\mathcal{M}$ is Euclidean, then we can take the trivial retraction of translating the base point to the origin, in which case the two viewpoints coincide. Hence in the Euclidean case, tBMM agrees with RBMM in (Li et al., 2023). Also in the Euclidean setting, tBMM agrees with the standard Euclidean BMM (Hong et al., 2015). Hence tBMM generalizes the Euclidean BMM.

C. Line Search Algorithm

Below is the optional line search for choosing the step size α in Algorithm 1,

Algorithm 2 Constrained Riemannian Line Search

Input: $\alpha \in [0, 1]$ (initial step size); $\theta_n^{(i)} \in \mathcal{M}^{(i)}$ (previous block iterate); $V_n^{(i)} \in \mathcal{T}_{\Theta^{(i)}}(\theta_{n-1}^{(i)})$ (admissible search direction in the tangent cone); $\gamma \in (0, 1)$ (shrinkage parameter)
set $\alpha = 1$
while $f_n^{(i)}(\mathrm{Rtr}_{\theta_{n-1}^{(i)}}(\alpha V_n^{(i)})) - f_n^{(i)}(\theta_{n-1}^{(i)}) > \frac{\rho\alpha}{4}\|V_n^{(i)}\|^2$ or $\mathrm{Rtr}_{\theta_{n-1}^{(i)}}(\alpha V_n^{(i)}) \notin \Theta^{(i)}$ **do**
    $\alpha = \gamma\alpha$
end while
output: α

D. An Example of the Lifted Constraint Set

In this section, we provide a concrete example of a constrained problem on the manifold. An illustration of this example is shown in Figure 1. Consider the 2-sphere $S^2 = \{x \in \mathbb{R}^3 : \|x\| = 1\}$, which can also be viewed as a simple Stiefel manifold. Consider three points $x, y, z \in S^2$ that x is at the north pole; y and z are on the equator such that $d(x, y) = \frac{\pi}{2}$. Then the region on the sphere bounded by the geodesic triangle $\triangle xyz$ is a geodesic convex constraint set. Let us call this constraint set Θ. Now for a point $\theta \in \Theta \subset S^2$, the corresponding lifted constraint set T_θ^* is one of the following three cases: 1) If θ is a vertex of the geodesic triangle (e.g. x in Figure 1), then the lifted constraint set is a circular sector with central angle $\frac{\pi}{2}$; 2) If θ is at the boundary but not the vertex (e.g. θ_1 in Figure 1), then T_θ^* is a half-disk; 3) If θ is in the interior of the triangle (e.g. θ_2 in Figure 1), then T_θ^* is a disk. In all the three cases, the lifted constraint set is convex on the tangent space. In fact, one could generalize this example of the geodesic triangle to the geodesic polyhedron on a submanifold of $\mathbb{R}^n$. One can then show the corresponding lifted constraint sets satisfy our assumptions using the same argument.

E. Examples of Hadamard Manifolds

The complete and simply connected Riemannian manifolds with nonpositive sectional curvature at every point is called *Hadamard manifolds* ((Burago et al., 2001) and (Burago et al., 1992)). The injectivity radius is infinity at each point on a Hadamard manifold. In fact, many widely studied spaces are Hadamard manifold and we give some examples below. We refer the readers to (Bacak, 2014) for more details.

Example E.1 (Euclidean spaces)**.** The Euclidean space $\mathbb{R}^n$ with its usual metric is a Hadamard manifold. The sectional curvature at each point is constant and equal to 0. ▲

Example E.2 (Hyperbolic spaces)**.** Equip $\mathbb{R}^{n+1}$ with the $(-1, n)$-inner product given by

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

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

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

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

Example E.3 (Manifolds of positive definite matrices)**.** The space $\mathbb{S}^n_{++}$ of all symmetric positive definite matrices of size $n \times n$ with real entries is a Hadamard manifold when equipped with the Riemannian metric given by

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

▲

F. Preliminary Lemmas

Lemma F.1. *Let $\psi : \mathbb{R}^n \to \mathbb{R}$ be a convex and Lipschitz continuous function with respect to the norm $\|\cdot\|$ with parameter L_ψ. Then for any x in the domain of ψ and any subgradient $\partial\psi$, we have $\|\partial\psi(x)\| \leq L_\psi$.*

Proof. Fix any x in the domain of ψ and any $\partial\psi$. Let

$$u^* \in \arg\max_{u, \|u\| \leq 1} \langle u, \partial\psi(x) \rangle.$$

Let $y = x + u^*$. Then $u^* = \frac{\partial\psi(x)}{\|\partial\psi(x)\|}$ and $\langle u^*, \partial\psi(x) \rangle = \|\partial\psi(x)\|$, so

$$\|\partial\psi(x)\| = \langle u^*, \partial\psi(x) \rangle \leq \psi(x + u^*) - \psi(x) \leq L_\psi \|u^*\| = L_\psi.$$

where the first inequality is by convexity of ψ, the second inequality is by Lipschitz continuity of ψ and the last equality is by definition of y. □

Proposition F.2 (Euclidean smoothness implies geodesic smoothness on the Stiefel manifold)**.** *If f is L-smooth in Euclidean space $\mathbb{R}^{n\times p}$, then there exists a constant $\tilde{L}_g = M_0^2 L + M_2 L_N$ such that f satisfying restricted smoothness (15) with parameter $\tilde{L}_g$ on the Stiefel manifold $\mathcal{V}^{n\times p}$, where $L_N = \max_{x \in \mathcal{V}^{n\times p}} \|\nabla f(x)\|$, M_0 is a constant related to the retraction, and M_2 is the same constant as in Prop. 4.3.*

Proof. See (Chen et al., 2021), Lemma 2.4 and Appendix C.1. □

G. Proof for Section 1

Proof of Corollary 1.1. Under the assumptions of Cor. 1.1, the inexact retraction step in (2) can be viewed as an exact retraction on V_n. Therefore, the inexactness of the second step in (2) is transferred to the first step. Namely, let

$$\hat{g}_n(\eta) = \varphi(\boldsymbol{\theta}_{n-1}) + \langle \operatorname{grad} \varphi(\boldsymbol{\theta}_{n-1}), \eta \rangle + \frac{L}{2}\|\eta\|^2.$$

Then inexact RGD can be viewed as tMM with tangential surrogate $\hat{g}_n(\eta)$. The exact solution of minimizing $\hat{g}_n$ is $-\frac{1}{2L} \operatorname{grad} \varphi(\boldsymbol{\theta}_{n-1})$ and the inexact solution used for inexact RGD is V_n. Recall $\Delta_n = \hat{g}_n(\alpha_n V_n) - \hat{g}_n(-\frac{1}{2L} \operatorname{grad} \varphi(\boldsymbol{\theta}_{n-1}))$ is summable. Hence (A1)**(ii)** is satisfied. Note (A1)**(i)** is satisfied by the assumptions in Cor. 1.1 and (A1)**(iii)** is not needed in the analysis when $\psi = 0$ (see the proof of Theorem 3.1 in Section I). By definition of $\hat{g}$, (A2)**(i),(ii)** are satisfied. (A2)**(iii)** again is satisfied by assumptions in Cor. 1.1. Hence Thm. 3.1 holds. The convergence and complexity results follow.

□

H. Proof for Section 4

Proof of Corollary 4.4. In order the complexity results in Thoerem 3.1 hold, we need to show assumptions (A1) and (A2) hold. (A1)**(i)** holds by the assumptions on objective function of problem (18). (A1)**(ii)** holds trivially. (A1)**(iii)** holds by Proposition 4.3. To see (A2) holds, let

$$\hat{g}(V) := \phi(\boldsymbol{\theta}_{n-1}) + \langle \operatorname{grad} \phi(\boldsymbol{\theta}_{n-1}), V \rangle + \lambda \|V\|_F^2.$$

Then since $\hat{g}$ is a quadratic function, (A2)**(i)**, **(ii)** hold. (A2)**(iii)** holds since Stiefel manifold is complete and (18) is an unconstrained problem. Namely, the constrained set is the manifolds itself, i.e. $\Theta = \mathcal{V}^{n\times p}$. Lastly, $\hat{g}(V)$ is indeed a tangential majorizer by the same analysis in Section 4.2 and Prop. F.2. Therefore, all the assumptions of Theorem 3.1 are satisfied. □

I. Convergence Analysis for tBMM

In this section, we prove Theorem 3.1. We will first prove the statement for the single-block case ($m = 1$) for notational simplicity. The main ingredient in our analysis is the following descent lemma for constrained tBMM.

Lemma I.1 (An inexact descent lemma for constrained tBMM)**.** *Suppose $\boldsymbol{\Theta}$ is a strongly convex subset of a Riemannian manifold $\mathcal{M}$. Fix a function $f : \mathcal{M} \to \mathbb{R}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Denote the pullback objective $\hat{f} := f \circ \mathrm{Rtr}_{\boldsymbol{\theta}} : T_{\boldsymbol{\theta}}\mathcal{M} \to \mathbb{R}$. Suppose we have a function $\hat{g} : T_{\boldsymbol{\theta}}\mathcal{M} \to \mathbb{R}$ such that the following hold:*

(i) *(Tangential marjorization) $\hat{g}$ is a majorizing surrogate of the pullback smooth part of the objective $\hat{\varphi}$:*

$$\hat{g}(\eta) \geq \hat{\varphi}(\eta) \text{ for all } \eta \in T_{\boldsymbol{\theta}}\mathcal{M} \text{ and } \hat{g}(\mathbf{0}) = \hat{\varphi}(\mathbf{0}).$$

(ii) *(Strong convexity of surrogate) The surrogate $\hat{g} : T_{\boldsymbol{\theta}} \to \mathbb{R}$ is ρ-strongly convex for some $\rho > 0$: for all $\eta \in T_{\boldsymbol{\theta}}\mathcal{M}$,*

$$\hat{g}(\eta) \geq \hat{g}(\mathbf{0}) + \langle \nabla \hat{g}(\mathbf{0}),\, \eta\rangle - \frac{\rho}{2}\|\eta\|^2.$$

Let $T^*_{\boldsymbol{\theta}}$ denote the lifted constrained set $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}$ (see (6)). Define $V^\star \in \arg\min_{\eta \in T^*_{\boldsymbol{\theta}}} (\hat{g}(\eta) + \psi(\theta + \eta))$. Fix $V \in T^*_{\boldsymbol{\theta}}$ denote $\Delta := \hat{g}(V) + \psi(\theta + V) - \hat{g}(V^\star) - \psi(\theta + V^\star)$. Then for each $\alpha \in [0, 1]$, $\mathrm{Rtr}_{\boldsymbol{\theta}}(\alpha V) \in \boldsymbol{\Theta}$ and

$$\begin{aligned} &f(\mathrm{Rtr}_{\boldsymbol{\theta}}(\alpha V)) - f(\boldsymbol{\theta}) \\ \leq &-\frac{\rho\alpha}{2}\|V^\star\|^2 - \frac{1}{2}\rho\alpha\left(1 - \frac{2L_\psi M_2 + \rho}{\rho}\alpha\right)\|V\|^2 + \alpha\Delta. \end{aligned} \tag{25}$$

Proof of Lemma I.1. That $\mathrm{Rtr}_{\boldsymbol{\theta}}(\alpha V)$ is well-defined and belongs to $\boldsymbol{\Theta}$ is due to the definitions of the retraction and the lifted constrained set. Hence we only need to show (25).

Let $G := \hat{g} + \psi : T_\theta \to \mathbb{R}$. Then since $\hat{g}$ is strongly convex and ψ is convex, we have G is strongly convex. Recall $T^*_{\boldsymbol{\theta}}$ is a convex subset of tangent space $T_{\boldsymbol{\theta}}$ by (A2)**(iii)**. The optimization problem that defines $V^\star$ is to minimize a strongly convex function in a convex subset of Euclidean space, so it admits a unique solution $V^\star$. Then by the first order optimality condition and since $\mathbf{0} \in T_{\boldsymbol{\theta}}$,

$$\langle \nabla \hat{g}(V^\star) + \partial\psi(\boldsymbol{\theta} + V^\star),\, \mathbf{0} - V^\star\rangle \geq 0.$$

Since $G : T_\theta \to \mathbb{R}$ is ρ-strongly convex,

$$\begin{aligned} &G(\mathbf{0}) - G(V^\star) \\ \geq &\langle \nabla \hat{g}(V^\star) + \partial\psi(\boldsymbol{\theta} + V^\star), \mathbf{0} - V^\star\rangle + \frac{\rho}{2}\|V^\star\|^2 \\ \geq &\frac{\rho}{2}\|V^\star\|^2. \end{aligned}$$

Hence

$$\begin{aligned} &G(\mathbf{0}) - G(V) \\ = &(G(\mathbf{0}) - G(V^\star)) + (G(V^\star) - G(V)) \geq \frac{\rho}{2}\|V^\star\|^2 - \Delta. \end{aligned}$$

Again using the ρ-strong convexity of $G : T_{\boldsymbol{\theta}} \to \mathbb{R}$ and the above inequality, for $\alpha \in [0, 1]$,

$$\begin{aligned} G(\alpha V) - G(\mathbf{0}) &= G(\alpha V + (1-\alpha)\mathbf{0}) - G(\mathbf{0}) \\ &\leq \alpha G(V) + (1-\alpha)G(\mathbf{0}) + \frac{\rho\alpha(1-\alpha)}{2}\|V\|^2 - G(\mathbf{0}) \\ &= \alpha(G(V) - G(\mathbf{0})) + \frac{\rho\alpha(\alpha-1)}{2}\|V\|^2 \\ &\leq -\frac{\rho\alpha}{2}\|V^\star\|^2 - \frac{\rho\alpha(1-\alpha)}{2}\|V\|^2 + \alpha\Delta. \end{aligned} \tag{26}$$

In order to conclude, note the following:

$$\begin{aligned} f(\mathrm{Rtr}_{\boldsymbol{\theta}}(\alpha V)) - f(\boldsymbol{\theta}) &= \hat{\varphi}(\alpha V) + \psi(\mathrm{Rtr}_{\boldsymbol{\theta}}(\alpha V)) - \hat{\varphi}(\mathbf{0}) - \psi(\boldsymbol{\theta}) \\ &\overset{(a)}{\leq} \hat{\varphi}(\alpha V) + \psi(\alpha V) + L_\psi \|\mathrm{Rtr}_{\boldsymbol{\theta}}(\alpha V) - \alpha V\| - \hat{\varphi}(\mathbf{0}) - \psi(\boldsymbol{\theta}) \\ &\overset{(b)}{\leq} G(\alpha V) - G(\mathbf{0}) + L_\psi M_2 \|\alpha V\|^2 \\ &\overset{(c)}{\leq} -\frac{\rho\alpha}{2}\|V^\star\|^2 - \frac{1}{2}\rho\alpha\left(1 - \frac{2L_\psi M_2 + \rho}{\rho}\alpha\right)\|V\|^2 + \alpha\Delta. \end{aligned}$$

Namely, (a) follows from the Lipschitz continuity of ψ, (b) uses Proposition 4.3, (c) uses (26). This shows (25), as desired. □

A direct consequence of Lemma I.1 is the boundedness of iterates, which is stated in the following proposition.

Proposition I.2 (Boundedness of iterates)**.** *Under (A1) and (A2), the set $\{\boldsymbol{\theta}_n : n \geq 1\}$ is bounded.*

Proof of Proposition I.2. Let $m\sum_{n=1}^{\infty}\Delta_n = T < \infty$. By Lemma I.1 we immediately have, for all $n \geq 1$, $f(\boldsymbol{\theta}_n) \leq f(\boldsymbol{\theta}_{n-1}) + m\Delta_{n-1}$ and therefore $f(\boldsymbol{\theta}_n) \leq f(\boldsymbol{\theta}_0) + m\sum_{i=1}^{n-1}\Delta_i \leq f(\boldsymbol{\theta}_0) + T$. Consider $K = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : f(\boldsymbol{\theta}) \leq f(\boldsymbol{\theta}_0) + T\}$, by (A1)**(i)**, K is compact. Hence the set $\{\boldsymbol{\theta}_n : n \geq 1\}$ is bounded. □

The following lemma states the local property of retractions, which is essential in the asymptotic convergence analysis.

Lemma I.3. *For $\boldsymbol{\theta} \in \mathcal{M}$, $V \in T_{\boldsymbol{\theta}}$ and $\alpha > 0$, when $\alpha\|V\|$ is small, we have*

$$d\left(\mathrm{Rtr}_{\boldsymbol{\theta}}(\alpha V), \boldsymbol{\theta}\right) = O(\alpha\|V\|).$$

Proof of Lemma I.3. By triangle inequality we have

$$\begin{aligned} d\left(\mathrm{Rtr}_{\boldsymbol{\theta}}(\alpha V), \boldsymbol{\theta}\right) &\leq d\left(\mathrm{Rtr}_{\boldsymbol{\theta}}(\alpha V), \mathrm{Exp}_{\boldsymbol{\theta}}(\alpha V)\right) + d\left(\mathrm{Exp}_{\boldsymbol{\theta}}(\alpha V), \boldsymbol{\theta}\right) \\ &= o(\alpha\|V\|) + \|\alpha V\| \\ &= O(\alpha\|V\|), \end{aligned}$$

where the first equality is by definition of the exponential map and properties of retraction, see (Absil & Malick, 2012). □

Now we are ready to prove Theorem 3.1 and we first prove it for $m = 1$. Then in the following proof, we will omit all superscripts (i) in Algorithm 1 for indicating the blocks since we are in the single-block case.

Proof of Theorem 3.1 for $m = 1$. By the hypothesis, the step-size α_n for $n \geq 1$ is set as $\alpha_n = \alpha \leq 1$. Therefore, for each $n \geq 1$, $\mathrm{Rtr}_{\boldsymbol{\theta}_{n-1}}(\alpha V_n)$ is well-defined and belongs to $\boldsymbol{\Theta}$.

Let $V_n^\star = \arg\min_{\eta \in T^*_{\boldsymbol{\theta}_{n-1}}} \left(G_n(\eta) := \hat{g}_n(\eta) + \psi(\boldsymbol{\theta}_{n-1} + \eta)\right)$. By Lemma I.1 and the hypothesis that $\alpha \leq \frac{\rho}{\rho + 2L_\psi M_2}$, for each $N \geq 0$,

$$\begin{aligned} 0 \leq \frac{\alpha\rho}{2}\sum_{n=0}^{N}\|V_n^\star\|^2 &\leq \sum_{n=0}^{N} f(\boldsymbol{\theta}_n) - f(\boldsymbol{\theta}_{n+1}) + \Delta_n \\ &\leq f(\boldsymbol{\theta}_0) - f^* + \sum_{n=1}^{\infty}\Delta_n < \infty. \end{aligned}$$

It follows that $\|V_n^\star\| = o(1)$ and

$$\min_{0\le k\le N} \|V_k^\star\|^2 \le \frac{2(f(\boldsymbol{\theta}_0) - f^* + \sum_{n=1}^\infty \Delta_n)}{\rho\alpha(N+1)}. \tag{27}$$

Recall that $V_n^\star$ is the minimizer of the tangential surrogate $\hat{g}_n$ on the lifted constraint set $T^*_{\boldsymbol{\theta}_{n-1}} \subseteq T_{\boldsymbol{\theta}_{n-1}}\mathcal{M}$, which is convex by (A2)**(iii)**. Hence by the first-order optimality condition

$$\langle \nabla \hat{g}(V_n^\star) + \mathrm{Proj}_{T_{\theta_{n-1}}} \partial\psi(\theta_{n-1} + V_n^\star)\ ,\ \eta - V_n^\star\rangle \ge 0 \quad \text{for all } \eta \in T^*_{\theta_{n-1}}.$$

Therefore for all $\eta \in T^*_{\theta_{n-1}}$ with $\|\eta\| \le 1$ we have

$$\begin{aligned}\langle \nabla \hat{g}_n(V_n^\star) + \mathrm{Proj}_{T_{\theta_{n-1}}} \partial\psi(\theta_{n-1} + V_n^\star),\ \eta\rangle &\ge \langle \nabla \hat{g}_n(V_n^\star) + \mathrm{Proj}_{T_{\theta_{n-1}}} \partial\psi(\theta_{n-1} + V_n^\star),\ V_n^\star\rangle \\ &\ge -\|\nabla \hat{g}_n(V_n^\star) + \mathrm{Proj}_{T_{\theta_{n-1}}} \partial\psi(\theta_{n-1} + V_n^\star)\| \cdot \|V_n^\star\|\end{aligned} \tag{28}$$

Next, we bound the term $\|\nabla \hat{g}_n(V_n^\star) + \mathrm{Proj}_{T_{\theta_{n-1}}} \partial\psi(\theta_{n-1} + V_n^\star)\|$ by a linear term with respect to $\|V_n^\star\|$. Note by Lemma I.2, for all $n \ge 1$ there exists constant C such that $\|\operatorname{grad}\varphi(\boldsymbol{\theta}_{n-1})\| \le C$. Since $\hat{g}_n$ majorizes $\hat{\varphi}$ on $T_{\boldsymbol{\theta}}\mathcal{M}$ and and is exact at $\mathbf{0}$, and also noting (22), we get $\nabla \hat{g}_n(\mathbf{0}) = \nabla \hat{\varphi}(\mathbf{0}) = \operatorname{grad}\varphi(\boldsymbol{\theta}_{n-1})$. Moreover, denoting

$$\zeta_n := \nabla \hat{g}_n(\mathbf{0}) - \nabla \hat{g}_n(V_n^\star),$$

by the restricted L_n-smoothness of $\hat{g}_n : T_{\boldsymbol{\theta}_{n-1}} \to \mathbb{R}$,

$$\|\zeta_n\| \le L_n\|V_n^\star\| \le L\|V_n^\star\|.$$

The above and triangle inequality gives

$$\|\nabla \hat{g}_n(V_n^\star)\| \le \|\operatorname{grad}\varphi(\boldsymbol{\theta}_{n-1})\| + L\|V_n^\star\| \le C + L\|V_n^\star\|. \tag{29}$$

Then, note

$$\|\mathrm{Proj}_{T_{\theta_{n-1}}} \partial\psi(\theta_{n-1} + V_n^*)\| \le \|\partial\psi(\theta_{n-1} + V_n^*)\| \le L_\psi \tag{30}$$

where the last inequality follows from Lemma F.1. Hence, (28) together with (29) and (30) gives

$$\langle \nabla \hat{g}_n(V_n^\star) + \mathrm{Proj}_{T_{\theta_{n-1}}} \partial\psi(\theta_{n-1} + V_n^\star),\ \eta\rangle \ge -(C' + L\|V_n^\star\|) \cdot \|V_n^\star\|$$

where $C' = C + L_\psi$. It follows

$$\begin{aligned}&\langle \operatorname{grad}\varphi(\boldsymbol{\theta}_{n-1}) + \mathrm{Proj}_{T_{\theta_{n-1}}} \partial\psi(\theta_{n-1} + V_n^\star), \eta\rangle \\ &= \langle \operatorname{grad}\varphi(\boldsymbol{\theta}_{n-1}) - \nabla \hat{g}_n(V_n^\star), \eta\rangle + \langle \nabla \hat{g}_n(V_n^\star) + \mathrm{Proj}_{T_{\theta_{n-1}}} \partial\psi(\theta_{n-1} + V_n^\star), \eta\rangle \\ &\ge -\|\zeta_n\| \cdot \|\eta\| - C'\|V_n^\star\| - L\|V_n^\star\|^2 \\ &\ge -(L + C')\|V_n^\star\| - L\|V_n^\star\|^2\end{aligned}$$

for all $\eta \in T^*_{\boldsymbol{\theta}_{n-1}}$ such that $\|\eta\| \le 1$. Therefore, (27) and the above give

$$\begin{aligned}\min_{1\le k\le N}\Bigg[-\inf_{\eta\in T^*_{\boldsymbol{\theta}_{k-1}},\,\|\eta\|\le 1} \langle \operatorname{grad}\varphi(\boldsymbol{\theta}_{k-1}) + \mathrm{Proj}_{T_{\theta_{k-1}}} \partial\psi(\theta_{k-1} + V_k^\star), \frac{\eta}{\min\{1, r_0\}}\rangle\Bigg] \\ \le \frac{L + C'}{\min\{1, r_0\}}\sqrt{\frac{2(f(\boldsymbol{\theta}_0) - f^* + \sum_{n=1}^\infty \Delta_n)}{\rho\alpha(N+1)}} + \frac{L}{\min\{1, r_0\}}\frac{2(f(\boldsymbol{\theta}_0) - f^* + \sum_{n=1}^\infty \Delta_n)}{\rho\alpha(N+1)}.\end{aligned}$$

When $N \geq \frac{L^2}{(L+C')^2}\frac{2(f(\boldsymbol{\theta}_0)-f^*+\sum_{n=1}^{\infty}\Delta_n)}{\rho\alpha}$, the above gives

$$\min_{1\leq k\leq N}\left[-\inf_{\eta\in T^*_{\boldsymbol{\theta}_{k-1}},\,\|\eta\|\leq 1}\langle \operatorname{grad}\varphi(\boldsymbol{\theta}_{k-1})+\operatorname{Proj}_{T_{\theta_{k-1}}}\partial\psi(\theta_{k-1}+V_k^\star),\frac{\eta}{\min\{1,r_0\}}\rangle\right] \leq \frac{2(L+C')}{\min\{1,r_0\}}\sqrt{\frac{2(f(\boldsymbol{\theta}_0)-f^*+\sum_{n=1}^{\infty}\Delta_n)}{\rho\alpha(N+1)}}.$$

This shows **(i)**. Note that **(iii)** is equivalent to **(i)** for the single-block ($m=1$) case.

Next, we show **(ii)**. Recall $\psi=0$, so $f=\varphi$ and is continuously differentiable. For asymptotic stationarity, suppose a subsequence of the iterates $(\boldsymbol{\theta}_{n_k})_{k\geq 1}$ converges to some limit point $\boldsymbol{\theta}_\infty\in\boldsymbol{\Theta}$. Note by first-order optimality of $V_n^\star$ and the ρ-strongly convexity of $\hat{g}_n$,

$$\Delta_n=\hat{g}_n(V_n)-\hat{g}_n(V_n^\star)\geq\langle\nabla\hat{g}_n(V_n^\star),V_n-V_n^\star\rangle+\frac{\rho}{2}\|V^\star\|^2\geq\frac{\rho}{2}\|V_n-V_n^\star\|^2.$$

Since $\sum_{n=1}^{\infty}\Delta_n<\infty$ by (A1), we have $\|V_n-V_n^\star\|=o(1)$. Hence by triangle inequality $\|V_n\|\leq\|V_n-V_n^\star\|+\|V_n^\star\|=o(1)$. Therefore, we have $\|\boldsymbol{\theta}_n-\boldsymbol{\theta}_{n-1}\|=o(1)$ by Lemma I.3. Hence $\boldsymbol{\theta}_{n_k-1}\to\boldsymbol{\theta}_\infty$ as $k\to\infty$. For each $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, let $B_{\boldsymbol{\theta}}$ denote the metric ball of radius half of the injectivity radius $\mathrm{r_{inj}}(\boldsymbol{\theta})$ centered at $\boldsymbol{\theta}$. Fix $\boldsymbol{\theta}'\in\boldsymbol{\Theta}\cap B_{\boldsymbol{\theta}_\infty}$ be arbitrary. Note that $\boldsymbol{\theta}'\in\boldsymbol{\Theta}\cap B_{\boldsymbol{\theta}_\infty}\cap B_{\boldsymbol{\theta}_{n_k}-1}$ for all sufficiently large k since $\boldsymbol{\theta}_n\in\boldsymbol{\Theta}$ for all $n\geq 1$ and $\boldsymbol{\theta}_{n_k-1}\to\boldsymbol{\theta}_\infty$. Denote by Γ_k the parallel transport $\Gamma_{\boldsymbol{\theta}_{n_k-1}\to\boldsymbol{\theta}_\infty}$. Then

$$\langle-\Gamma_k\left(\operatorname{grad}f(\boldsymbol{\theta}_{n_k-1})\right),\,\Gamma_k\big(\eta_{\boldsymbol{\theta}_{n_k-1}}(\boldsymbol{\theta}')\big)\rangle=\langle-\operatorname{grad}f(\boldsymbol{\theta}_{n_k-1}),\,\eta_{\boldsymbol{\theta}_{n_k-1}}(\boldsymbol{\theta}')\rangle\leq 0.$$

Now since f is continuously differentiable, we have $\operatorname{grad}f(\boldsymbol{\theta}_{n_k-1})\to\operatorname{grad}f(\boldsymbol{\theta}_\infty)$ as $k\to\infty$. Also, by the continuity of Riemannian metric, the left-hand side above converges to $\langle-\operatorname{grad}f(\boldsymbol{\theta}_\infty),\,\eta_{\boldsymbol{\theta}_\infty}(\boldsymbol{\theta}')\rangle$, so we obtain

$$\langle-\operatorname{grad}f(\boldsymbol{\theta}_\infty),\,\eta_{\boldsymbol{\theta}_\infty}(\boldsymbol{\theta}')\rangle\leq 0.$$

Since $\boldsymbol{\theta}'\in\boldsymbol{\Theta}\cap B_{\boldsymbol{\theta}_\infty}$ is arbitrary, the above show that $\boldsymbol{\theta}_\infty$ is a stationary point of f over $\boldsymbol{\Theta}$. □

Now we prove Theorem 3.1 for the multi-block case $m\geq 2$. The proof is almost identical to the single-block case.

Proof of Theorem 3.1 for $m\geq 2$. By the hypothesis, the step-size $\alpha_n^{(i)}$ for $n\geq 1$ and $i\in\{1,\dots,m\}$ satisfies $\alpha_n^{(i)}=\alpha\leq 1$. Therefore, for each $n\geq 1$, $\mathrm{Rtr}_{\theta_{n-1}^{(i)}}(\alpha V_n^{(i)})$ is well-defined and belongs to $\Theta^{(i)}$.

Let $V_n^{(i\star)}=\arg\min_{\eta\in T^*_{\theta_{n-1}^{(i)}}}\Big(G_n^{(i)}(\eta):=\hat{g}_n^{(i)}(\eta)+\psi_n^{(i)}(\theta_{n-1}^{(i)}+\eta)\Big)$. By Lemma I.1 and the hypothesis that $\alpha\leq\frac{\rho}{\rho+2L_\psi M_2}$, for each $N\geq 0$,

$$\begin{aligned}0<\frac{\alpha\rho}{2}\sum_{n=0}^{N}\sum_{i=1}^{m}\|V_n^{(i\star)}\|^2&\leq\sum_{n=0}^{N}\sum_{i=1}^{m}f_n^{(i)}(\theta_{n-1}^{(i)})-f_n^{(i)}(\theta_n^{(i)})+\Delta_n\\&\leq\sum_{n=0}^{N}f(\boldsymbol{\theta}_n)-f(\boldsymbol{\theta}_{n+1})+m\Delta_n\\&\leq f(\boldsymbol{\theta}_0)-f^*+m\sum_{n=1}^{\infty}\Delta_n=:M<\infty.\end{aligned}$$

It follows that $\sum_{i=1}^{m}\|V_n^{(i\star)}\|=o(1)$ and

$$\min_{0\leq k\leq N}\sum_{i=1}^{m}\|V_k^{(i\star)}\|^2\leq\frac{2M}{\rho\alpha N}.\tag{31}$$

Recall that $V_n^{(i\star)}$ is the minimizer of the surrogate $G_n^{(i)}$ on the lifted constrained set $T^*_{\theta_{n-1}^{(i)}} \subseteq T_{\theta_{n-1}^{(i)}}\mathcal{M}^{(i)}$, which is convex by (A2)**(iii)**. Hence by the first-order optimality condition

$$\langle \nabla \hat{g}_n^{(i)}(V_n^\star) + \mathrm{Proj}_{T_{\theta_{n-1}^{(i)}}} \partial\psi_n^{(i)}(\theta_{n-1}^{(i)} + V_n^{(i\star)}) \ , \ \eta - V_n^{(i\star)} \rangle \geq 0 \quad \text{for all } \eta \in T^*_{\theta_{n-1}^{(i)}}.$$

Therefore for all $\eta \in T^*_{\theta_{n-1}^{(i)}}$ with $\|\eta\| \leq 1$ we have

$$\begin{aligned}\langle \nabla \hat{g}_n^{(i)}(V_n^{(i\star)}) + \mathrm{Proj}_{T_{\theta_{n-1}^{(i)}}} \partial\psi_n^{(i)}(\theta_{n-1}^{(i)} + V_n^{(i\star)}), \eta \rangle &\geq \langle \nabla \hat{g}_n^{(i)}(V_n^{(i\star)}) + \mathrm{Proj}_{T_{\theta_{n-1}^{(i)}}} \partial\psi_n^{(i)}(\theta_{n-1}^{(i)} + V_n^{(i\star)}), V_n^{(i\star)} \rangle \\ &\geq -\|\nabla \hat{g}_n^{(i)}(V_n^{(i\star)}) + \mathrm{Proj}_{T_{\theta_{n-1}^{(i)}}} \partial\psi_n^{(i)}(\theta_{n-1}^{(i)} + V_n^{(i\star)})\| \cdot \|V_n^{(i\star)}\| \\ &\geq -(C' + L\|V_n^{(i\star)}\|) \cdot \|V_n^{(i\star)}\| \end{aligned} \tag{32}$$

for some constant $C' > 0$, which can be determined following the same line of $m = 1$ case.

Since $\hat{g}_n^{(i)}$ majorizes $\hat{\varphi}_n^{(i)}$ on $T_{\theta_{n-1}}\mathcal{M}^{(i)}$ and is exact at $\theta_{n-1}^{(i)}$, and also noting and also noting (22), we get $\nabla \hat{g}_n^{(i)}(\mathbf{0}) = \nabla \hat{\varphi}_n^{(i)}(\mathbf{0}) = \mathrm{grad}\, \varphi_n^{(i)}(\theta_{n-1}^{(i)})$. Moreover, denoting

$$\zeta_n^{(i)} := \nabla \hat{g}_n^{(i)}(\mathbf{0}) - \nabla \hat{g}_n^{(i)}(V_n^{(i\star)}), \tag{33}$$

by the restricted L_n-smoothness of $\hat{g}_n : T_{\boldsymbol{\theta}_{n-1}} \to \mathbb{R}$,

$$\|\zeta_n^{(i)}\| \leq L_n^{(i)}\|V_n^{(i\star)}\| \leq L\|V_n^{(i\star)}\|. \tag{34}$$

Therefore it follows

$$\begin{aligned}&\langle \mathrm{grad}\, \varphi_n^{(i)}(\theta_{n-1}^{(i)}) + \mathrm{Proj}_{T_{\theta_{n-1}^{(i)}}} \partial\psi_n^{(i)}(\theta_{n-1}^{(i)} + V_n^{(i)}), \eta \rangle \\ &= \langle \mathrm{grad}\, \varphi_n^{(i)}(\theta_{n-1}^{(i)}) - \nabla \hat{g}_n(V_n^{(i\star)}), \eta \rangle + \langle \nabla \hat{g}_n(V_n^{(i\star)}) + \mathrm{Proj}_{T_{\theta_{n-1}^{(i)}}} \partial\psi_n^{(i)}(\theta_{n-1}^{(i)} + V_n^{(i)}), \eta \rangle \\ &\geq -\|\zeta_n^{(i)}\| \cdot \|\eta\| - (C' + L\|V_n^{(i\star)}\|)\|V_n^{(i\star)}\| \\ &\geq -(L + C')\|V_n^{(i\star)}\| - L\|V_n^{(i\star)}\|^2 \end{aligned}$$

for all $\eta \in T^*_{\theta_{n-1}^{(i)}}$ and $\|\eta\| \leq 1$. Thus it follows that

$$\begin{aligned}\sum_{i=1}^m - \inf_{\eta \in T^*_{\theta_{n-1}^{(i)}}, \|\eta\| \leq 1} &\langle \mathrm{grad}\, \varphi_n^{(i)}(\theta_{n-1}^{(i)}) + \mathrm{Proj}_{T_{\theta_{n-1}^{(i)}}} \partial\psi_n^{(i)}(\theta_{n-1}^{(i)} + V_n^{(i)}), \frac{\eta}{\min\{r_0, 1\}} \rangle \\ &\leq \frac{L + C'}{\min\{r_0, 1\}} \sum_{i=1}^m \|V_n^{(i\star)}\| + \frac{L}{\min\{r_0, 1\}} \sum_{i=1}^m \|V_n^{(i\star)}\|^2. \end{aligned}$$

Therefore, the above with (31) implies

$$\begin{aligned}\min_{1 \leq n \leq N} \sum_{i=1}^m - \inf_{\eta \in T^*_{\theta_{n-1}^{(i)}}, \|\eta\| \leq 1} &\langle \mathrm{grad}\, \varphi_n^{(i)}(\theta_{n-1}^{(i)}) + \mathrm{Proj}_{T_{\theta_{n-1}^{(i)}}} \partial\psi_n^{(i)}(\theta_{n-1}^{(i)} + V_n^{(i)}), \frac{\eta}{\min\{r_0, 1\}} \rangle \\ &\leq \frac{L + C'}{\min\{r_0, 1\}} \sqrt{\frac{2M}{\rho\alpha N}} + \frac{L}{\min\{r_0, 1\}} \frac{2M}{\rho\alpha N}. \end{aligned}$$

When $N \geq \frac{L^2}{(L+C')^2} \frac{2M}{\rho\alpha}$, the above gives

$$\min_{1 \leq n \leq N} \sum_{i=1}^m - \inf_{\eta \in T^*_{\theta_{n-1}^{(i)}}, \|\eta\| \leq 1} \langle \mathrm{grad}\, \varphi_n^{(i)}(\theta_{n-1}^{(i)}) + \mathrm{Proj}_{T_{\theta_{n-1}^{(i)}}} \partial\psi_n^{(i)}(\theta_{n-1}^{(i)} + V_n^{(i)}), \frac{\eta}{\min\{r_0, 1\}} \rangle \leq \frac{2(L + C')}{\min\{r_0, 1\}} \sqrt{\frac{2M}{\rho\alpha N}}.$$

This shows **(i)**.

Next, we show **(ii)**. Recall $\psi = 0$, so $f = \varphi$ and is continuously differentiable. For asymptotic stationarity, suppose a subsequence of the iterates $(\boldsymbol{\theta}_{n_k})_{k\geq 1}$ converges to some limit point $\boldsymbol{\theta}_\infty \in \boldsymbol{\Theta}$. Note by first-order optimality of $V_n^{(i\star)}$ and the ρ-strongly convexity of $\hat{g}_n$,

$$\Delta_n = \hat{g}_n^{(i)}(V_n) - \hat{g}_n^{(i)}(V_n^{(i\star)}) \geq \langle \nabla \hat{g}_n^{(i)}(V_n^{(i\star)}), V_n^{(i)} - V_n^{(i\star)} \rangle + \frac{\rho}{2}\|V^{(i\star)}\|^2 \geq \frac{\rho}{2}\|V_n^{(i)} - V_n^{(i\star)}\|^2. \tag{35}$$

Since $m\sum_{n=1}^{\infty}\Delta_n \leq \infty$ by (A1), we have $\sum_{i=1}^m \|V_n^{(i)} - V_n^{(i\star)}\| = o(1)$. Hence by triangle inequality $\sum_{i=1}^m \|V_n^{(i)}\| \leq \sum_{i=1}^m \|V_n^{(i)} - V_n^{(i\star)}\| + \sum_{i=1}^m \|V_n^{(i\star)}\| = o(1)$. Therefore, we have $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = o(1)$ by Lemma I.3. Hence $\boldsymbol{\theta}_{n_k-1} \to \boldsymbol{\theta}_\infty$ as $k \to \infty$. Therefore since f is continuously differentiable, for each $i = 1, \ldots, m$,

$$\begin{aligned}
\lim_{k\to\infty} \operatorname{grad} f_{n_k}^{(i)}(\theta_{n_k-1}^{(i)}) &= \lim_{k\to\infty} \operatorname{grad}_i f(\theta_{n_k}^{(1)}, \cdots, \theta_{n_k}^{(i-1)}, \theta_{n_k-1}^{(i)}, \theta_{n_k-1}^{(i+1)}, \cdots, \theta_{n_k-1}^{(m)}) \\
&= \operatorname{grad}_i f(\theta_\infty^{(1)}, \cdots, \theta_\infty^{(i-1)}, \theta_\infty^{(i)}, \theta_\infty^{(i+1)}, \cdots, \theta_\infty^{(m)}) \\
&= \operatorname{grad}_i f(\boldsymbol{\theta}_\infty).
\end{aligned}$$

Fix $i \in \{1, \ldots, m\}$. For each $\theta \in \Theta^{(i)}$, let B_θ denote the metric ball of radius half of the injectivity radius $\mathrm{r_{inj}}^{(i)}(\theta)$ centered at θ in $\mathcal{M}^{(i)}$. Fix $\theta' \in \Theta^{(i)} \cap B_{\boldsymbol{\theta}_\infty^{(i)}}$ be arbitrary. Note that $\boldsymbol{\theta}_{n_k-1} \in \Theta^{(i)} \cap B_{\theta_\infty^{(i)}} \cap B_{\theta_{n_k-1}^{(i)}}$ for all sufficiently large k since $\theta_n^{(i)} \in \Theta^{(i)}$ for all $n \geq 1$ and $\theta_{n_k-1}^{(i)} \to \theta_\infty^{(i)}$. Denote by Γ_k the parallel transport $\Gamma_{\theta_{n_k-1}^{(i)} \to \theta_\infty^{(i)}}$ on $\mathcal{M}^{(i)}$. Then

$$\left\langle -\Gamma_k\left(\operatorname{grad} f_{n_k}^{(i)}(\theta_{n_k-1}^{(i)})\right), \Gamma_k\big(\eta_{\theta_{n_k-1}^{(i)}}^{(i)}(\theta')\big)\right\rangle = \left\langle -\operatorname{grad} f_{n_k}^{(i)}(\theta_{n_k-1}^{(i)}), \eta_{\theta_{n_k-1}^{(i)}}^{(i)}(\theta')\right\rangle \leq 0.$$

Also note that by the continuity of Riemannian metric, the left-hand side above converges to $\langle -\operatorname{grad}_i f(\boldsymbol{\theta}_\infty), \eta_{\theta_\infty^{(i)}}(\theta')\rangle$, so we obtain

$$\langle -\operatorname{grad}_i f(\boldsymbol{\theta}_\infty), \eta_{\theta_\infty^{(i)}}(\theta')\rangle \leq 0.$$

Since $\theta' \in \Theta^{(i)} \cap B_{\theta_\infty^{(i)}}$ and $i \in \{1, \ldots, m\}$ are arbitrary, the above shows that $\boldsymbol{\theta}_\infty$ is a stationary point of f over $\boldsymbol{\Theta}$.

Lastly, we show **(iii)**. For each $n \geq 1$ and $i = 1, \ldots, m$, denote

$$\xi_n^{(i)} := \operatorname{grad} f_n^{(i)}(\theta_{n-1}^{(i)}) - \operatorname{grad}_i f(\boldsymbol{\theta}_{n-1}).$$

Now by writing

$$\begin{aligned}
\nabla \hat{g}_n^{(i)}(V_n^{(i\star)}) - \operatorname{grad}_i \varphi(\boldsymbol{\theta}_{n-1}) &= \left(\nabla \hat{g}_n^{(i)}(V_n^{(i\star)}) - \nabla \hat{g}_n^{(i)}(\mathbf{0})\right) + \left(\operatorname{grad} \varphi_n^{(i)}(\theta_{n-1}^{(i)}) - \operatorname{grad}_i \varphi(\boldsymbol{\theta}_{n-1})\right) \\
&= \zeta_n^{(i)} + \xi_n^{(i)},
\end{aligned}$$

from (32), for all $\eta \in T^*_{\theta_{n-1}^{(i)}}$ with $\|\eta\| \leq 1$,

$$\begin{aligned}
&\langle \operatorname{grad}_i \varphi(\boldsymbol{\theta}_{n-1}) + \operatorname{Proj}_{T_{\theta_{n-1}^{(i)}}} \partial\psi_n^{(i)}(\theta_{n-1}^{(i)} + V_n^{(i)}), \eta\rangle \\
&= -\langle \xi_n^{(i)}, \eta\rangle - \langle \zeta_n^{(i)}, \eta\rangle + \langle \nabla \hat{g}_n(V_n^{(i\star)}) + \operatorname{Proj}_{T_{\theta_{n-1}^{(i)}}} \partial\psi_n^{(i)}(\theta_{n-1}^{(i)} + V_n^{(i)}), \eta\rangle \\
&\geq -(\|\zeta_n^{(i)}\| + \|\xi_n^{(i)}\|) \cdot \|\eta\| - C'\|V_n^{(i\star)}\| - L\|V_n^{(i\star)}\|^2 \\
&\geq -(\|\zeta_n^{(i)}\| + \|\xi_n^{(i)}\|) - C'\|V_n^{(i\star)}\| - L\|V_n^{(i\star)}\|^2.
\end{aligned} \tag{36}$$

According to the hypothesis, by using a triangle inequality,

$$\|\xi_n^{(i)}\| \leq L'(\|V_n^{(1)}\| + \cdots + \|V_n^{(i-1)}\|) \leq L'\sum_{i=1}^m \|V_n^{(i)}\|.$$

Combining with (34),

$$\|\zeta_n^{(i)}\| + \|\xi_n^{(i)}\| \leq L\|V_n^{(i\star)}\| + L'\sum_{i=1}^{m}\|V_n^{(i)}\|,$$

where $\zeta_n^{(i)}$ is defined in (33). From (35), we have

$$\frac{\rho}{2}\sum_{n=1}^{N}\sum_{i=1}^{m}\|V_n^{(i)} - V_n^{(i\star)}\|^2 < m\sum_{n=1}^{\infty}\Delta_n < \infty.$$

Therefore,

$$\sum_{n=1}^{N}\sum_{i=1}^{m}\|V_n^{(i)}\|^2 \leq 2\sum_{n=1}^{N}\sum_{i=1}^{m}\|V_n^{(i)} - V_n^{(i\star)}\|^2 + 2\sum_{n=1}^{N}\sum_{i=1}^{m}\|V_n^{(i\star)}\|^2 < \frac{4\alpha m\sum_{n=1}^{\infty}\Delta_n + 4M}{\alpha\rho}.$$

And also,

$$\sum_{n=1}^{N}\sum_{i=1}^{m}\|V_n^{(i)}\|^2 + \sum_{i=1}^{m}\|V_n^{(i\star)}\|^2 < \frac{4\alpha m\sum_{n=1}^{\infty}\Delta_n + 6M}{\alpha\rho}.$$

Hence

$$\min_{0\leq n\leq N}\sum_{i=1}^{m}\|V_n^{(i)}\|^2 + \sum_{i=1}^{m}\|V_n^{(i\star)}\|^2 < \frac{4\alpha m\sum_{n=1}^{\infty}\Delta_n + 6M}{\alpha\rho N}. \tag{37}$$

From (37), there exists a subsequence $(k_n)_{n\geq 1}$ such that $1 \leq k_n \leq n$ and

$$\max\left\{\sum_{i=1}^{m}\|V_{k_n}^{(i\star)}\|^2, \sum_{i=1}^{m}\|V_{k_n}^{(i)}\|^2\right\} \leq \sum_{i=1}^{m}\|V_{k_n}^{(i\star)}\|^2 + \|V_{k_n}^{(i)}\|^2 \leq \frac{M'}{\alpha\rho n},$$

where $M' := 4\alpha m\sum_{n=1}^{\infty}\Delta_n + 6M$.

Hence we get

$$\begin{aligned}(\|\zeta_n^{(i)}\| + \|\xi_n^{(i)}\|) + C'\|V_n^{(i\star)}\| + L\|V_n^{(i\star)}\|^2 &\leq (L+C')\|V_n^{(i\star)}\| + L'\sum_{i=1}^{m}\|V_n^{(i)}\| + L\sum_{i=1}^{m}\|V_n^{(i\star)}\|^2\\ &\leq (L+C')\sqrt{\frac{M'}{\rho\alpha n}} + L'\sqrt{\frac{M'}{\rho\alpha n}} + L\frac{M}{\rho\alpha n}\\ &= (L+C'+L')\sqrt{\frac{M'}{\rho\alpha n}} + L\frac{M}{\rho\alpha n}.\end{aligned}$$

This combining (36) gives

$$\begin{aligned}\min_{1\leq n\leq N}\sum_{i=1}^{m} -\inf_{\eta\in T^*_{\theta_{n-1}^{(i)}},\|\eta\|\leq 1}\langle \mathrm{grad}_i\,\varphi(\boldsymbol{\theta}_{n-1}) + \mathrm{Proj}_{T_{\theta_{n-1}^{(i)}}}\partial\psi_n^{(i)}(\theta_{n-1}^{(i)} + V_n^{(i)}), \frac{\eta}{\min\{r_0,1\}}\rangle&\\ \leq \frac{L+C'+L'}{\min\{r_0,1\}}\sqrt{\frac{M'}{\rho\alpha N}} + \frac{L}{\min\{r_0,1\}}\frac{M}{\rho\alpha N}.&\end{aligned}$$

When $N \geq \frac{L^2}{(L+C'+L')^2}\frac{M^2}{M'\rho\alpha}$, the above gives,

$$\min_{1\leq n\leq N}\sum_{i=1}^{m} -\inf_{\eta\in T^*_{\theta_{n-1}^{(i)}},\|\eta\|\leq 1}\langle \mathrm{grad}_i\,\varphi(\boldsymbol{\theta}_{n-1}) + \mathrm{Proj}_{T_{\theta_{n-1}^{(i)}}}\partial\psi_n^{(i)}(\theta_{n-1}^{(i)} + V_n^{(i)}), \frac{\eta}{\min\{r_0,1\}}\rangle \leq \frac{2(L+C'+L')}{\min\{r_0,1\}}\sqrt{\frac{M'}{\rho\alpha N}},$$

the desired complexity result. □

J. Details of Numerical Experiments

For all the numerical experiments in Section 5, we test the algorithms with random initial points. Each experiment is repeated for 10 times with i.i.d. Gaussian/uniform random initial points, depending on the setting of the problem. The *relative reconstruction error* defined as $\mathrm{Error}(X) = \|X - X^*\|/\|X^*\|$ is computed as a function of elapsed time (on a Macbook Pro 2020 with 1.4 GHz Quad-Core Intel Core i5). Averaged relative reconstruction error with standard deviation are shown by the solid lines and shaded regions in all the plots.

J.1. Nonnegative Tensor Decomposition with Riemannian Constraints

For the (nonnegative) tensor decomposition problem with Riemannian constraints we studied in Section 5, we pose Riemannian constraints to the first block. Specifically, in the experiments, we let $\mathcal{M}^{(1)} = \mathcal{R}_r \subseteq \mathbb{R}^{50\times 10}$ with $r = 2$. The other two blocks are Euclidean spaces, i.e. $\mathcal{M}^{(2)} = \mathbb{R}^{40\times 10}$ and $\mathcal{M}^{(3)} = \mathbb{R}^{30\times 10}$. When the loading matrices are nonnegative, the nonnegativity can be considered as an additional constraint set within each block. Namely, the constraint sets $\Theta^{(2)}$ and $\Theta^{(3)}$ are nonnegative orthants of the corresponding Euclidean spaces. As for the first low-rank block $M^{(1)}$, it is shown in (Song & Ng, 2020) that the intersection of the fixed-rank manifold and nonnegative orthant remains a smooth manifold. Therefore, one can also simply let $\mathcal{M}^{(1)}$ to be the nonnegative fixed-rank manifold. The ℓ_1-regularizer in (20) brings nonsmoothness to the problem. In the numerical experiments, we set the regularization parameter $\lambda_i = 10^{-2}$ for $i = 1, 2, 3$.

J.2. Regularized Nonnegative Matrix Factorization with Riemannian Constraints

Similar to the tensor decomposition problem described in the previous section, we let $\mathcal{M}^{(1)} = \mathcal{R}_r \subseteq \mathbb{R}^{50\times 10}$ with $r = 5$ and $\mathcal{M}^{(2)} = \mathbb{R}^{10\times 40}$. The nonnegativity can be viewed as constraint sets, as stated in the previous section. This low-rank matrix factorization (without ℓ_1 regularization) is recently studied in (Song et al., 2022), where the authors proposed a tangent space based alternating projection method. In our setting, we further include the ℓ_1-regularization term that brings nonsmoothness. The regularization parameter is set to be $\lambda = 10^{-2}$.

J.3. Low-rank Matrix Recovery

We state the NIHT for solving (21) in Alg. 3.

Algorithm 3 Normalized Iterative Hard Thresholding (NIHT) (Tanner & Wei, 2013)

Input: Initial point X_0, left singular vector space U_0
for $i = 1, \cdots, n$ **do**
$\nabla f(X_i) = \mathcal{A}^*(\mathcal{A}(X_i) - b)$
$\alpha_i = \frac{\|\mathrm{Proj}_{U_i}(\nabla f(X_i))\|^2}{\|\mathcal{A}\,\mathrm{Proj}_{U_i}(\nabla f(X_i))\|^2}$
$X_{i+1} = \mathrm{Proj}_{\mathcal{R}_r}(X_i - \alpha_i \nabla f(X_i))$
end for
output: X_n

Here $\mathcal{A}^* : \mathbb{R}^p \to \mathbb{R}^{m\times n}$ is the Hermitian adjoint operator of $\mathcal{A}$, $\mathrm{Proj}_{U_i} = U_i U_i^*$ is the projection onto the left singular vector subspace of X_i. Note NIHT is a projected gradient descent algorithm with adaptive step size.

In the numerical validation in Figure 4, we set the dimensions to be $m = 50$, $n = 12$, $p = 300$ or 450 respectively. The true solution X^* is a low-rank matrix in $\mathbb{R}^{m\times n}$ with rank $r = 3$. The sensing matrices A_i for $i = 1, \cdots, p$ are i.i.d. Gaussian random matrices. The observations b is generated by $b = \mathcal{A}(X^*)$.

J.4. Inexact RGD

We show the performance of inexact RGD based on the low-rank matrix recovery problem in Figure 5. The dimensions are set to be $m = 50$, $n = 12$, and $p = 300$. The inexactness is simulated by adding noise to the Riemannian gradient with $\Delta_n = c/(n+1)^2$, where n is the iteration number. The inexact tBMM is implemented with $c = 1$. In the left plot of Figure 5, inexact tBMM shows similar performance as exact tBMM where both of them outperform NIHT. In the right plot of Figure 5, it is shown that inexact RGD with different noise parameter c converges at similar speed.