Riemannian Interior Point Methods for Constrained Optimization on Manifolds

Zhijian Lai[1] · Akiko Yoshise[1]

Abstract

We extend the classical primal-dual interior point method from the Euclidean setting to the Riemannian one. Our method, named the Riemannian interior point method, is for solving Riemannian constrained optimization problems. We establish its local superlinear and quadratic convergence under the standard assumptions. Moreover, we show its global convergence when it is combined with a classical line search. Our method is a generalization of the classical framework of primal-dual interior point methods for nonlinear nonconvex programming. Numerical experiments show the stability and efficiency of our method.

1 Introduction

In this paper, we consider the following Riemannian constrained optimization problem (RCOP):

$$\min_{x \in \mathcal{M}} f(x) \qquad \text{s.t.} \quad h(x) = 0, \text{ and } g(x) \leqslant 0, \tag{RCOP}$$

Communicated by Sándor Zoltán Németh.

An earlier version of this article has been circulated under the title "Superlinear and Quadratic Convergence of Riemannian Interior Point Methods".

✉ Akiko Yoshise
yoshise@sk.tsukuba.ac.jp

Zhijian Lai
galvin.lai@outlook.com

[1] University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan

where $\mathcal{M}$ is a connected, complete d-dimensional Riemannian manifold and $f : \mathcal{M} \rightarrow \mathbb{R}, h : \mathcal{M} \rightarrow \mathbb{R}^l$, and $g : \mathcal{M} \rightarrow \mathbb{R}^m$ are smooth functions. This problem appears in many applications, for instance, matrix approximation with nonnegative constraints on a fixed-rank manifold [31] and orthogonal nonnegative matrix factorization on the Stiefel manifold [19]; for more, see [23, 27].

The body of knowledge on (RCOP) without h, g, often called simply *Riemannian optimization*, has grown considerably in the last 20 years. The well-known methods in the Euclidean setting, such as steepest descent, Newton, and trust region, have been extended to the Riemannian setting [1, 5, 16]. By contrast, research on (RCOP) is still in its infancy. The earliest studies go back to ones on the optimality conditions. Yang et al. [36] extended the Karush–Kuhn–Tucker (KKT) conditions to (RCOP). Bergmann and Herzog [2] considered more constraint qualifications (CQs) on manifolds. Yamakawa and Sato [34] proposed sequential optimality conditions in the Riemannian case. Liu and Boumal [23] were the first to develop practical algorithms. They extended the augmented Lagrangian method and exact penalty method to (RCOP). Schiela and Ortiz [30] and Obara et al. [27] proposed the Riemannian sequential quadratic programming method. However, to our knowledge, interior point methods have yet to be considered for (RCOP).

The advent of interior point methods in the 1980s greatly advanced the field of optimization [33, 37]. By the early 1990s, the success of these methods in linear and quadratic programming ignited interest in using them on nonlinear nonconvex cases [11, 35]. From the 1990s to the first decade of the twenty-first century, a large number of interior point methods for nonlinear programming emerged. They proved to be as successful as the linear ones. A subclass known as *primal-dual* interior point methods is the most efficient and practical. As described in [24], the primal-dual approach to linear programming was introduced in [25]: it was first developed as an algorithm in [20] and eventually became standard for the nonlinear case as well [11, 35].

Contribution In this paper, we extend the primal-dual interior point algorithms from the Euclidean setting, i.e., $\mathcal{M} = \mathbb{R}^d$ in (RCOP), to the Riemannian setting. We call this extension the *Riemannian interior point method* (RIPM). Our contributions are summarized as follows:

1. To our knowledge, this is the first study to apply the primal-dual interior point method to the nonconvex constrained optimization problem on Riemannian manifolds. One significant contribution is that we establish many essential foundational concepts for the general interior point method in the Riemannian context, such as the KKT vector field and its covariant derivative. In addition, we build the first framework for the Riemannian version of the interior point method. These contributions will have uses in the future, especially in developing more advanced interior point methods.
2. We give a detailed theoretical analysis to ensure local and global convergence of RIPM. Considering that many practical problems involve minimizing a nonconvex function on Riemannian manifolds, the theoretical counterparts of our method are the early interior point methods for nonlinear nonconvex programming first proposed by El-Bakry et al. [11].

3. Our numerical experiments[1] demonstrate the great potential of RIPM. The method meets the challenges presented in these experiments with better stability and higher accuracy compared with the current Riemannian exact penalty, augmented Lagrangian, and sequential quadratic programming methods.

Organization The rest of this paper is organized as follows: In Sect. 2, we review the notation of Riemannian geometry and explain the Riemannian Newton method. In Sect. 3, we give a full interpretation of our RIPM and describe a prototype algorithm of RIPM. We also investigate the use of Krylov subspace methods to efficiently solve a condensed form of a perturbed Newton equation. This is particularly important for the numerical implementation of RIPM. Section 4 gives the necessary preliminaries and auxiliary results needed to prove convergence in our subsequent sections. In Sect. 5, we give the proof of local superlinear and quadratic convergence of the prototype algorithm of RIPM. Section 6 describes a globally convergent version of RIPM with a classical line search; then, Sect. 7 proves its global convergence. Section 8 is a collection of numerical experiments. Section 9 summarizes our research and presents future work.

2 Notation

2.1 Riemannian Geometry

Let us briefly review some concepts from Riemannian geometry, following the notation of [5]. $\mathcal{M}$ denotes a finite-dimensional smooth manifold. Let $p \in \mathcal{M}$ and $T_p\mathcal{M}$ be the tangent space at p with 0_p being its zero element. We use a canonical identification $T_p\mathcal{E} \cong \mathcal{E}$ for a vector space $\mathcal{E}$ and $p \in \mathcal{E}$. A vector field is a map $V: \mathcal{M} \to T\mathcal{M}$ with $V(p) \in T_p\mathcal{M}$, where $T\mathcal{M} := \bigcup_{p \in \mathcal{M}} T_p\mathcal{M}$ is the tangent bundle. $\mathfrak{X}(\mathcal{M})$ denotes the set of all smooth vector fields defined on $\mathcal{M}$. Furthermore, $\mathcal{M}$ is a Riemannian manifold if it is equipped with a Riemannian metric, that is, a choice of inner product $\langle \cdot, \cdot \rangle_p: T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$ for each tangent space at p on $\mathcal{M}$ such that for all $V, W \in \mathfrak{X}(\mathcal{M})$, the map

$$p \mapsto \langle V(p), W(p) \rangle_p \tag{1}$$

is a smooth function from $\mathcal{M}$ to $\mathbb{R}$. Riemannian metric induces the norm $\|\xi\|_p := \sqrt{\langle \xi, \xi \rangle_p}$ for $\xi \in T_p\mathcal{M}$. We often omit the subscript p if it is clear from the context. Throughout this paper, we assume that all the manifolds involved are *connected* and *complete*. Given a curve segment on $\mathcal{M}$, $c: [a, b] \to \mathcal{M}$, the length of c is defined as $L(c) := \int_a^b \|\dot{c}(t)\|_{c(t)} dt$, where $\dot{c}(t) \in T_{c(t)}\mathcal{M}$ is the velocity vector of c at t. Since $\mathcal{M}$ is connected, there exists a curve segment connecting any pair of points $p, q \in \mathcal{M}$. Indeed, $\mathcal{M}$ is a metric space under the Riemannian distance $d(p, q) := \inf_c L(c)$ where the infimum is taken over all curve segments joining p to q [22, Thm. 2.55]. For

---

[1] The code is freely available at https://doi.org/10.5281/zenodo.10612799.

two manifolds $\mathcal{M}_1$, $\mathcal{M}_2$ and a smooth map $F: \mathcal{M}_1 \to \mathcal{M}_2$, the differential of F at $p \in \mathcal{M}_1$ is a linear operator denoted as $\mathcal{D}F(p): T_p\mathcal{M}_1 \to T_{F(p)}\mathcal{M}_2$. Let $\mathfrak{F}(\mathcal{M})$ be the set of all smooth *scalar fields* (or say real-valued functions) $f: \mathcal{M} \to \mathbb{R}$. The Riemannian gradient of f at p, grad $f(p)$, is defined as the unique element of $T_p\mathcal{M}$ that satisfies $\langle \xi, \text{grad } f(p) \rangle_p = \mathcal{D}f(p)[\xi]$ for all $\xi \in T_p\mathcal{M}$, where $\mathcal{D}f(p): T_p\mathcal{M} \to T_{f(p)}\mathbb{R} \cong \mathbb{R}$. Note that for any $f \in \mathfrak{F}(\mathcal{M})$, the gradient vector field $x \mapsto \text{grad } f(x)$ is a smooth vector field on $\mathcal{M}$, i.e., grad $f \in \mathfrak{X}(\mathcal{M})$. A retraction $\mathrm{R}: T\mathcal{M} \to \mathcal{M}$ is a smooth map such that $\mathrm{R}_p(0_p) = p$ and $\mathcal{D}\mathrm{R}_p(0_p) = \text{id}_{T_p\mathcal{M}}$, i.e., the identity map on $T_p\mathcal{M}$, where R_p is the restriction of R to $T_p\mathcal{M}$ and $\mathcal{D}\mathrm{R}_p(0_p): T_{0_p}(T_p\mathcal{M}) \cong T_p\mathcal{M} \to T_p\mathcal{M}$ is the differential of R_p at 0_p. One theoretically perfect type of retraction is the exponential map, denoted as Exp. Since $\mathcal{M}$ is complete, the exponential map is well-defined on the whole tangent bundle. Let $\text{Exp}_p: T_p\mathcal{M} \to \mathcal{M}$ be the exponential map at p; then, $t \mapsto \text{Exp}_p(t\xi)$ is the unique geodesic that passes through p with velocity $\xi \in T_p\mathcal{M}$ when $t = 0$.

2.2 Riemannian Newton Method

The Newton method is a powerful tool for finding the zeros of nonlinear functions in the Euclidean setting. The generalized Newton method has been studied in the Riemannian setting; it aims to find a *singularity* for the vector field $F \in \mathfrak{X}(\mathcal{M})$, i.e., a point $p \in \mathcal{M}$ such that,

$$F(p) = 0_p \in T_p\mathcal{M}. \tag{2}$$

Let ∇ be the Levi–Civita connection on $\mathcal{M}$, i.e., the unique symmetric connection compatible with the Riemannian metric. The covariant derivative ∇F assigns each point $p \in \mathcal{M}$ a linear operator $\nabla F(p): T_p\mathcal{M} \to T_p\mathcal{M}$. In particular, the Riemannian Hessian of $f \in \mathfrak{F}(\mathcal{M})$ at p is a self-adjoint operator on $T_p\mathcal{M}$, defined as Hess $f(p) := \nabla(\text{grad } f)(p)$. The Riemannian Newton method for solving (2) is performed as Algorithm 1.

Algorithm 1: Riemannian Newton Method for (2)

Input: A vector field $F \in \mathfrak{X}(\mathcal{M})$, an initial point $p_0 \in \mathcal{M}$ and a retraction R on $\mathcal{M}$.
Output: Sequence $\{p_k\} \subset \mathcal{M}$ such that $\{p_k\} \to p^*$ and $F(p^*) = 0_{p^*}$.
Set $k \to 0$;
while *Stopping criterion not satisfied* **do**
    1. Obtain $\xi_k \in T_{p_k}\mathcal{M}$ by solving the Newton equation (a linear operator equation on tangent space $T_{p_k}\mathcal{M}$):

$$\nabla F(p_k)[\xi_k] = -F(p_k); \tag{3}$$

    2. Compute the next point as $p_{k+1} := \mathrm{R}_{p_k}(\xi_k)$;
    3. $k \to k + 1$;
end

---

3 Description of Riemannian Interior Point Methods (RIPM)

In this section, we will give a comprehensive interpretation of the Riemannian interior point method. Following common usage in the interior-point literature, big letters denote the associated diagonal matrix, e.g., $Z = \mathrm{diag}(z_1, \ldots, z_m)$ with $z \in \mathbb{R}^m$. e denotes the all-ones vector, and 0 stands for zero vector/matrix with proper dimensions.

3.1 KKT Vector Field and Its Covariant Derivative

The Lagrangian function of (RCOP) is

$$\mathcal{L}(x, y, z) := f(x) + \sum_{i=1}^{l} y_i h_i(x) + \sum_{i=1}^{m} z_i g_i(x),$$

where $y \in \mathbb{R}^l$ and $z \in \mathbb{R}^m$ are Lagrange multipliers corresponding to the equality and inequality constraints, respectively. With respect to the variable x, $\mathcal{L}(\cdot, y, z)$ is a real-valued function defined on $\mathcal{M}$, and its Riemannian gradient is:

$$\mathrm{grad}_x \, \mathcal{L}(x, y, z) = \mathrm{grad} \, f(x) + \sum_{i=1}^{l} y_i \, \mathrm{grad} \, h_i(x) + \sum_{i=1}^{m} z_i \, \mathrm{grad} \, g_i(x),$$

where $\{\mathrm{grad} \, h_i(x)\}_{i=1}^{l}$ and $\{\mathrm{grad} \, g_i(x)\}_{i=1}^{m}$ are the gradients of the component functions of h and g. The Riemannian KKT conditions (e.g., see [23, Definition 2.3] or [36]) for (RCOP) are given by

$$\mathrm{grad}_x \, \mathcal{L}(x, y, z) = 0_x; \ h(x) = 0, g(x) \leqslant 0, z \geqslant 0; \ Zg(x) = 0. \tag{4}$$

The above conditions can be written in terms of slack variables $\mathbb{R}^m \ni s := -g(x)$, as

$$F(w) := \begin{pmatrix} \mathrm{grad}_x \, \mathcal{L}(x, y, z) \\ h(x) \\ g(x) + s \\ ZSe \end{pmatrix} = 0_w = \begin{pmatrix} 0_x \\ 0 \\ 0 \\ 0 \end{pmatrix} \in T_w \mathcal{N}, \tag{5}$$

and $(z, s) \geqslant 0$, where $w := (x, y, z, s) \in \mathcal{N} := \mathcal{M} \times \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^m$. Here, we have generated a vector field F in (5) on the product manifold $\mathcal{N}$, i.e., for any $w \in \mathcal{N}$, $F(w) \in T_w \mathcal{N} \cong T_x \mathcal{M} \times \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^m$. We will call $F \colon \mathcal{N} \to T\mathcal{N}$ in (5) above the *KKT vector field* for (RCOP). Note that for $\xi = (\xi_x, \xi_y, \xi_z, \xi_s) \in T_w \mathcal{N}$,

$$\bar{\mathrm{R}}_w(\xi) := (\mathrm{R}_x(\xi_x), y + \xi_y, z + \xi_z, s + \xi_s) \tag{6}$$

is a well-defined retraction on the product manifold $\mathcal{N}$ as long as R is a retraction on $\mathcal{M}$.

Now, we aim to find a singularity of the KKT vector field F on $\mathcal{N}$ under some nonnegative constraints $(z, s) \geqslant 0$. If the Riemannian Newton method is to be applied to (5), we must formulate the covariant derivative of F at an arbitrary $w \in \mathcal{N}$. Before that, we need some new symbols. Fixing a point $x \in \mathcal{M}$, we can define two linear operators $\mathcal{H}_x \colon \mathbb{R}^l \to T_x\mathcal{M}$ and $\mathcal{G}_x \colon \mathbb{R}^m \to T_x\mathcal{M}$ by

$$
\mathcal{H}_x[v] := \sum_{i=1}^{l} v_i \ \mathrm{grad} \ h_i(x), \quad \mathcal{G}_x[u] := \sum_{i=1}^{m} u_i \ \mathrm{grad} \ g_i(x), \tag{7}
$$

respectively. Then, for $\mathcal{H}_x$, its adjoint operator $\mathcal{H}_x^* \colon T_x\mathcal{M} \to \mathbb{R}^l$ is given by:

$$
\mathcal{H}_x^*[\xi] = \left(\langle \mathrm{grad} \ h_1(x), \xi \rangle_x, \cdots, \langle \mathrm{grad} \ h_l(x), \xi \rangle_x \right)^T.
$$

Also, $\mathcal{G}_x^* \colon T_x\mathcal{M} \to \mathbb{R}^m$ with $\mathcal{G}_x^*[\xi] = \left(\langle \mathrm{grad} \ g_1(x), \xi \rangle_x, \cdots, \langle \mathrm{grad} \ g_m(x), \xi \rangle_x \right)^T$.

Now, using the solutions of Exercises 5.4 and 5.13 on covariant derivatives of vector fields on product manifolds in monograph [5], the following results can be easily derived. Given KKT vector field F in (5) and $w \in \mathcal{N}$, the covariant derivative of F at w is the operator $\nabla F(w) \colon T_w\mathcal{N} \to T_w\mathcal{N}$ given by

$$
\nabla F(w)[\Delta w] = \begin{pmatrix} \mathrm{Hess}_x \ \mathcal{L}(w)[\Delta x] + \mathcal{H}_x[\Delta y] + \mathcal{G}_x[\Delta z] \\ \mathcal{H}_x^*[\Delta x] \\ \mathcal{G}_x^*[\Delta x] + \Delta s \\ Z\Delta s + S\Delta z \end{pmatrix}, \tag{8}
$$

where $\Delta w = (\Delta x, \Delta y, \Delta z, \Delta s) \in T_w\mathcal{N}$ and $\mathrm{Hess}_x \ \mathcal{L}(w)$ is the Riemannian Hessian of real-valued function $\mathcal{L}(\cdot, y, z)$. Indeed, $\mathrm{Hess}_x \ \mathcal{L}(w) \colon T_x\mathcal{M} \to T_x\mathcal{M}$ satisfies

$$
\mathrm{Hess}_x \ \mathcal{L}(w) = \mathrm{Hess} \ f(x) + \sum_{i=1}^{l} y_i \ \mathrm{Hess} \ h_i(x) + \sum_{i=1}^{m} z_i \ \mathrm{Hess} \ g_i(x), \tag{9}
$$

where $\{\mathrm{Hess} \ h_i(x)\}_{i=1}^{l}$ and $\{\mathrm{Hess} \ g_i(x)\}_{i=1}^{m}$ are Hessians of the component functions of h and g.

3.2 Implication of Standard Assumptions

Next, we will discuss an important result about the covariant derivative of the KKT vector field given in (8). Let $\mathbb{A}(x) := \{i \mid g_i(x) = 0\}$ denote the active set at $x \in \mathcal{M}$. A prior study on Riemannian optimality conditions [2, 36] showed that the following assumptions are meaningful. We call them the *standard Riemannian assumptions* for (RCOP). Note that the x^* and w^* in (**A2**)–(**A4**) all refer to those in (**A1**).

(**A1**) Existence. There exists $w^* = (x^*, y^*, z^*, s^*)$ satisfying the Riemannian KKT conditions (4). Here, we introduce the slack variables $s^* := -g(x^*)$.

(A2) Linear Independence Constraint Qualification (LICQ).
The set $\{\text{grad } h_i(x^*)\}_{i=1}^{l} \bigcup \{\text{grad } g_i(x^*)\}_{i \in \mathbb{A}(x^*)}$ is linearly independent in $T_{x^*}\mathcal{M}$.

(A3) Strict complementarity. $(z^*)_i > 0$ if $g_i(x^*) = 0$ for all $i = 1, \cdots, m$.

(A4) Second-order sufficiency. $\langle \text{Hess}_x \, \mathcal{L}(w^*)[\xi], \xi \rangle > 0$ for all nonzero $\xi \in T_{x^*}\mathcal{M}$ satisfying $\langle \xi, \text{grad } h_i(x^*) \rangle = 0$ for $i = 1, \ldots, l$, and $\langle \xi, \text{grad } g_i(x^*) \rangle = 0$ for $i \in \mathbb{A}(x^*)$.

Recall the Riemannian Newton method discussed in Sect. 2.2. It can be shown that if p^* is a solution of equation (2) and the covariant derivative $\nabla F(p^*)$ is *nonsingular*, then Algorithm 1 has the local superlinear convergence [12] and local quadratic convergence [13] under certain mild conditions on the map $p \mapsto \nabla F(p)$. Note that the requirement of nonsingularity for the covariant derivative at the solution point is of primary importance. Therefore, the next theorem motivates us to use the Riemannian Newton method to solve (5).

Theorem 3.1 *If the standard Riemannian assumptions (A1)–(A4) hold at some point w^*, then the operator $\nabla F(w^*)$ in (8) is nonsingular.*

Proof This proof omits all the asterisks of the variables. Define $\mathbb{E} := \{1, \ldots, l\}$ and $\mathbb{I} := \{1, \ldots, m\}$. Take some $w = (x, y, z, s) \in \mathcal{N}$ satisfying (A1)–(A4), then we have $s_i = -g_i(x)$ and $z_i s_i = 0$ for all $i \in \mathbb{I}$. For short, let $\mathbb{A} := \mathbb{A}(x) \subset \mathbb{I}$. Suppose that $\nabla F(w)[\Delta w] = 0$ for some $\Delta w = (\Delta x, \Delta y, \Delta z, \Delta s) \in T_w \mathcal{N} \cong T_x \mathcal{M} \times \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^m$. Δy_i denotes the components of the vector Δy, as do $\Delta z_i, \Delta s_i$. To prove its nonsingularity, we will show that $\Delta w = 0$. Expanding the equation $\nabla F(w)[\Delta w] = 0$ gives

$$
\begin{aligned}
0 &= \text{Hess}_x \, \mathcal{L}(w)[\Delta x] + \sum_{i \in \mathbb{E}} \Delta y_i \, \text{grad } h_i(x) + \sum_{i \in \mathbb{I}} \Delta z_i \, \text{grad } g_i(x), \\
0 &= \langle \text{grad } h_i(x), \Delta x \rangle, \text{ for all } i \in \mathbb{E}, \\
0 &= \langle \text{grad } g_i(x), \Delta x \rangle + \Delta s_i, \text{ for all } i \in \mathbb{I}, \\
0 &= z_i \Delta s_i + s_i \Delta z_i, \text{ for all } i \in \mathbb{I}.
\end{aligned}
\tag{10}
$$

Strict complementarity (A3) and the last equalities above imply that $\Delta s_i = 0$ for all $i \in \mathbb{A}$ and $\Delta z_i = 0$ for $i \in \mathbb{I}\backslash\mathbb{A}$. Substituting those values into the system (10) reduces it to

$$
\begin{aligned}
0 &= \text{Hess}_x \, \mathcal{L}(w)[\Delta x] + \sum_{i \in \mathbb{E}} \Delta y_i \, \text{grad } h_i(x) + \sum_{i \in \mathbb{A}} \Delta z_i \, \text{grad } g_i(x), \\
0 &= \langle \text{grad } h_i(x), \Delta x \rangle, \text{ for all } i \in \mathbb{E}, \\
0 &= \langle \text{grad } g_i(x), \Delta x \rangle, \text{ for all } i \in \mathbb{A},
\end{aligned}
\tag{11}
$$

and $\Delta s_i = -\langle \text{grad } g_i(x), \Delta x \rangle$ for all $i \in \mathbb{I}\backslash\mathbb{A}$. It follows from system (11) that

$$\begin{aligned}
0 &= \langle \text{Hess}_x\, \mathcal{L}(w)[\varDelta x] + \sum_{i \in \mathbb{E}} \varDelta y_i\, \text{grad}\, h_i(x) + \sum_{i \in \mathbb{A}} \varDelta z_i\, \text{grad}\, g_i(x),\, \varDelta x \rangle \\
&= \langle \text{Hess}_x\, \mathcal{L}(w)[\varDelta x],\, \varDelta x \rangle + \sum_{i \in \mathbb{E}} \varDelta y_i\, \langle \text{grad}\, h_i(x),\, \varDelta x \rangle \\
&\quad + \sum_{i \in \mathbb{A}} \varDelta z_i\, \langle \text{grad}\, g_i(x),\, \varDelta x \rangle \\
&= \langle \text{Hess}_x\, \mathcal{L}(w)[\varDelta x],\, \varDelta x \rangle.
\end{aligned}$$

Thus, from second-order sufficiency (**A4**), $\varDelta x$ must be zero elements. And then $\varDelta s_i = 0$ for all $i \in \mathbb{I}\backslash\mathbb{A}$. Next, substituting $\varDelta x = 0$ into the first equation in (11) yields $0 = \sum_{i \in \mathbb{E}} \varDelta y_i\, \text{grad}\, h_i(x) + \sum_{i \in \mathbb{A}} \varDelta z_i\, \text{grad}\, g_i(x)$. The LICQ (**A2**) implies that the coefficients $\varDelta y_i$ for $i \in \mathbb{E}$ and $\varDelta z_i$ for $i \in \mathbb{A}$ must be zero. This completes the proof. □

3.3 Prototype Algorithm of RIPM

Applying the Riemannian Newton method directly to the KKT vector field $F : \mathcal{N} \to T\mathcal{N}$ results in the following Newton equation (see (3) without iteration count k) at each iteration:

$$\nabla F(w)[\varDelta w] + F(w) = 0. \tag{12}$$

As with the usual interior point method in the Euclidean setting, once the iterates reach the boundary of the feasible region, they are forced to stick to it [33, P6]. For the iterates to maintain a sufficient distance from the boundary, we introduce a perturbed complementary equation with some barrier parameter $\mu > 0$ and define the *perturbed KKT vector field*:

$$F_\mu(w) := F(w) - \mu\hat{e}, \text{ and } \hat{e} := \hat{e}(w) := (0_x, 0, 0, e). \tag{13}$$

Notice that the perturbation term $\hat{e}$, indeed, is a special vector field on $\mathcal{N}$, not a constant, because 0_x is essentially dependent on w and/or x. In fact, the covariant derivative of the perturbed KKT vector field is the same as that of the original. From the linearity of the connection ∇, for any $w \in \mathcal{N}$ and any $\mu > 0$, we have

$$\nabla F_\mu(w) = \nabla F(w) - \mu\nabla\hat{e}(w) = \nabla F(w), \tag{14}$$

where the last equity comes from $\nabla\hat{e}(w)[\varDelta w] = (0_x, 0, 0, 0)$ for all $\varDelta w \in T_w\mathcal{N}$. Applying the Riemannian Newton method to perturbed KKT vector field $F_\mu(w)$ yields the *perturbed Newton equation*: $\nabla F_\mu(w)[\varDelta w] + F_\mu(w) = 0$. From (13) and (14), this equation is equivalent to

$$\nabla F(w)[\varDelta w] + F(w) = \mu\hat{e}, \tag{15}$$

which reduces to the ordinary Newton equation (12) as $\mu \to 0$. At this point, we can describe a prototype of the Riemannian interior point method (RIPM) in Algorithm 2.

Algorithm 2: Prototype Algorithm of RIPM for (RCOP)

Input: An initial point $w_0 = (x_0, y_0, z_0, s_0) \in \mathcal{N}$ with $(z_0, s_0) > 0$ and a retraction R on $\mathcal{M}$.
   $\hat{\gamma} \in (0, 1)$, $\mu_0 > 0$.
Output: Sequence $\{w_k\} \subset \mathcal{N}$ such that $\{w_k\} \to w^*$ and w^* satisfies the KKT conditions (4).
Set $k \to 0$;
while *Stopping criterion not satisfied* **do**

1. Obtain $\Delta w_k = (\Delta x_k, \Delta y_k, \Delta z_k, \Delta s_k) \in T_{w_k}\mathcal{N}$ by solving the perturbed Newton equation on $T_{w_k}\mathcal{N}$:

$$\nabla F(w_k)[\Delta w_k] = -F(w_k) + \mu_k \hat{e}; \qquad (16)$$

2. Choose $\gamma_k \in [\hat{\gamma}, 1]$ and compute the step size α_k defined by

$$\min\left\{1, \gamma_k \min_i \left\{-\frac{(s_k)_i}{(\Delta s_k)_i} \mid (\Delta s_k)_i < 0\right\}, \gamma_k \min_i \left\{-\frac{(z_k)_i}{(\Delta z_k)_i} \mid (\Delta z_k)_i < 0\right\}\right\}; \qquad (17)$$

3. Compute the next point as $w_{k+1} := \bar{\mathrm{R}}_{w_k}(\alpha_k \Delta w_k)$, see (6);
4. Choose $0 < \mu_{k+1} < \mu_k$;
5. $k \to k + 1$;

end

In the step 2 of Algorithm 2, we just wish to compute a step size $0 < \alpha_k \leqslant 1$ to ensure new point w_{k+1} with $(z_{k+1}, s_{k+1}) > 0$. There are many schemes to achieve this purpose. The scheme in (17) is simple but sufficient to guarantee the local convergence, as will be proved in Sect. 5.

3.4 Solving Perturbed Newton Equation Efficiently

The challenge of Algorithm 2 is how to solve the Newton equation (16) in an efficient manner. In this subsection, we will do this in two steps: the first step will be to turn the original full Newton equation, which is asymmetric and consists of four variables, into a condensed form, which is symmetric and consists of only two variables. In the second step, an iterative method, namely the Krylov subspace method, is used to solve the *operator equations* directly, avoiding the expensive computational effort of converting them into the usual *matrix equations*.

3.4.1 Condensed Form of Perturbed Newton Equation

Let us consider Algorithm 2 and omit the iteration count k. Given the current point $w \in \mathcal{N}$ with $(z, s) > 0$, for the KKT vector field $F(w)$ in (5), we denote its components by F_x, F_y, F_z, F_s in top-to-bottom order, namely $F_x := \mathrm{grad}_x \mathcal{L}(w)$, $F_y := h(x)$, $F_z := g(x) + s$, $F_s := ZSe$. By using these symbols and the formulation of $\nabla F(w)$ in (8), the *full* (perturbed) Newton equation (16) defined on $T_w\mathcal{N} \cong T_x\mathcal{M} \times \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^m$ is expanded as:

$$\begin{pmatrix} \text{Hess}_x\, \mathcal{L}(w)[\Delta x] + \mathcal{H}_x[\Delta y] + \mathcal{G}_x[\Delta z] \\ \mathcal{H}_x^*[\Delta x] \\ \mathcal{G}_x^*[\Delta x] + \Delta s \\ Z\Delta s + S\Delta z \end{pmatrix} = \begin{pmatrix} -F_x \\ -F_y \\ -F_z \\ -F_s + \mu e \end{pmatrix}. \tag{18}$$

Not only does this equation contain four variables, but there is no symmetry on the left side of the equation, so it would be unwise to solve it just like that. By $(z, s) > 0$ and the fourth line of (18), we deduce $\Delta s = Z^{-1} (\mu e - F_s - S\Delta z)$. Substituting this Δs into the third line of (18), we get $\mathcal{G}_x^*[\Delta x] - Z^{-1}S\Delta z = -Z^{-1}\mu e - g(x)$ and thus $\Delta z = S^{-1} \left[Z \left(\mathcal{G}_x^*[\Delta x] + F_z \right) + \mu e - F_s \right]$. Substituting this Δz further into the first line of (18) and combining it with the second line of (18) yield the following *condensed* Newton equation, which is defined on $T_x\mathcal{M} \times \mathbb{R}^l$:

$$\mathcal{T}(\Delta x, \Delta y) := \begin{pmatrix} \mathcal{A}_w[\Delta x] + \mathcal{H}_x[\Delta y] \\ \mathcal{H}_x^*[\Delta x] \end{pmatrix} = \begin{pmatrix} c \\ q \end{pmatrix}. \tag{19}$$

where

$$\begin{aligned} \mathcal{A}_w &:= \text{Hess}_x\, \mathcal{L}(w) + \mathcal{G}_x S^{-1} Z \mathcal{G}_x^*, \\ c &:= -F_x - \mathcal{G}_x S^{-1} \left(ZF_z + \mu e - F_s \right), q := -F_y. \end{aligned} \tag{20}$$

Here, c and q are constant vectors. If we defined $\Theta := \mathcal{G}_x S^{-1} Z \mathcal{G}_x^*$, then $\mathcal{A}_w = \text{Hess}_x\, \mathcal{L}(w) + \Theta$. Note that both Θ and $\text{Hess}_x\, \mathcal{L}(w)$ are operators from and to $T_x\mathcal{M}$.

From the discussion above, for any $w \in \mathcal{N}$ with $(z, s) > 0$, the operator $\nabla F(w)$ in (8) is nonsingular if and only if the newly defined operator $\mathcal{T}$ in (19) is nonsingular. Eventually, it is sufficient for us to solve the equation (19) containing only two variables Δx and Δy. In fact, when we consider the case of only inequality constraints in (RCOP), then Δy vanishes, and only a linear equation

$$\mathcal{A}_w[\Delta x] = c \tag{21}$$

on $T_x\mathcal{M}$ needs to be solved. More importantly, the operator $\mathcal{T}$ in the left side of (19) is symmetric, or say self-adjoint (although often indefinite). It is trivial to check that operators Θ and $\mathcal{A}_w$ are self-adjoint on $T_x\mathcal{M}$; and thus $\mathcal{T}$ is self-adjoint on the product vector space $T_x\mathcal{M} \times \mathbb{R}^l$ equipped with inner product $\langle (\xi_x, \xi_y), (\eta_x, \eta_y) \rangle := \langle \xi_x, \eta_x \rangle_x + \xi_y^T \eta_y$. We can also see that (19) is a saddle point problem defined on Hilbert spaces from its special structure.

3.4.2 Krylov Subspace Methods on Tangent Space

Next, how to solve (19) efficiently becomes critical. For simplicity, we consider the case of only inequality constraints in (RCOP), then we will solve operator equation (21) with a self-adjoint operator $\mathcal{A}_w : T_x\mathcal{M} \to T_x\mathcal{M}$. Let $d := \dim T_x\mathcal{M}$. Unfortunately, in most cases of practical applications, the Riemannian situation leaves us with no explicit matrix form available for $\mathcal{A}_w$. This means that we can only access $\mathcal{A}$ (subscript w omitted) by inputting a vector v to return $\mathcal{A}v$.

A general approach is first to find the matrix representation $\hat{\mathcal{A}}$ for $\mathcal{A}$ under some basis of $T_x\mathcal{M}$. In detail, the full process of this approach is as follows:

(Step 1) Obtain d random independent vectors on $T_x\mathcal{M}$.

(Step 2) Obtain an orthonormal basis $\{u_i\}_{i=1}^d$ of $T_x\mathcal{M}$ by the modified Gram-Schmidt algorithm.

(Step 3) Compute $[\hat{\mathcal{A}}]_{ij} := \langle \mathcal{A}u_j, u_i \rangle_x$ for $1 \leqslant i \leqslant j \leqslant d$ due to symmetry, then we obtain the matrix representation $\hat{\mathcal{A}} \in \mathbb{R}^{d \times d}$.

(Step 4) Compute $[\hat{c}]_i := \langle c, u_i \rangle_x$ for $1 \leqslant i \leqslant d$, then we obtain the vector representation $\hat{c} \in \mathbb{R}^d$.

(Step 5) Using arbitrary linear solver to get solution $\Delta\hat{x} \in \mathbb{R}^d$ from matrix equation $\hat{\mathcal{A}}\Delta\hat{x} = \hat{c}$.

(Step 6) Recovery the tangent vector $\Delta x \in T_x\mathcal{M}$ by $\Delta x = \sum_{i=1}^d (\Delta\hat{x})_i u_i$.

In Algorithm 2, at each iteration, x is updated, and thus, the tangent space $T_x\mathcal{M}$ changes. Thus, the above six steps need to be done all over again. Obviously, this approach is so expensive that it is not feasible in practice.

An ideal approach is to use an iterative method, such as the Krylov subspace method (e.g., conjugate gradients method [5, Chapter 6.3]), on $T_x\mathcal{M}$ directly. Such a method does not explicitly require a matrix representation $\hat{\mathcal{A}}$ for $\mathcal{A}$. In general, it only needs to call an abstract linear operator $v \mapsto \mathcal{A}v$. Since $\mathcal{A}$ in (20) is self-adjoint but indefinite, for solving operator equation (21), we will use the Conjugate Residual (CR) method (see [29, ALGORITHM 6.20]) as stated in Algorithm 3.

Algorithm 3: Conjugate Residual (CR) Method on Tangent Spaces for (21)

Input: Symmetric invertible linear operator $\mathcal{A} : T_x\mathcal{M} \to T_x\mathcal{M}$, nonzero $c \in T_x\mathcal{M}$ and an initial point $v_0 \in T_x\mathcal{M}$.

Output: Sequence $\{v_n\} \subset T_x\mathcal{M}$ such that $\{v_n\} \to v^*$ and $\mathcal{A}v^* = c$.

Set $n \to 0$, $r_0 := c - Av_0$, $p_0 := r_0$ and compute $\mathcal{A}r_0$, $\mathcal{A}p_0$;

while *stopping criterion not satisfied* **do**

    1. Update number $\alpha_n := \langle r_n, \mathcal{A}r_n \rangle_x / \langle \mathcal{A}p_n, \mathcal{A}p_n \rangle_x$;   `// Step length`

    2. Set $v_{n+1} := v_n + \alpha_n p_n$;   `// Iterate point`

    3. Update $r_{n+1} := r_n - \alpha_n \mathcal{A}p_n$;   `// Residual`

    4. Compute $\mathcal{A}r_{n+1}$;   `// This is the only call to` $\mathcal{A}$ `in while loop`

    5. Update number $\beta_n := \langle r_{n+1}, \mathcal{A}r_{n+1} \rangle_x / \langle r_n, \mathcal{A}r_n \rangle_x$;

    6. Set $p_{n+1} := r_{n+1} + \beta_n p_n$;   `// Conjugate direction`

    7. Compute $\mathcal{A}p_{n+1} := \mathcal{A}r_{n+1} + \beta_n \mathcal{A}p_n$;   `// No need to call` $\mathcal{A}$ `here`

    8. $n \to n+1$;

end

A significant feature is that the iterate points v_k, conjugate directions p_k, and residual vectors $r_k := \mathcal{A}v_k - c$ are all contained in $T_x\mathcal{M}$. Usually, the initial point v_0 is the zero element of $T_x\mathcal{M}$; the iteration terminates when the relative residual $\|r_k\| / \|c\| \leqslant \epsilon$ for some threshold $\epsilon > 0$, or some maximum number of iterates is reached.

The discussion of the above two approaches can be naturally extended to the case containing equality constraints in (RCOP), where we consider $\mathcal{T}$ on $T_x\mathcal{M} \times \mathbb{R}^l$ instead of $\mathcal{A}$ on $T_x\mathcal{M}$.

4 Preliminaries and Auxiliary Results

This section introduces the useful results that are indispensable to our subsequent discussion.

Remark 4.1 For a retraction R on $\mathcal{M}$ and $x \in \mathcal{M}$, by $\mathcal{D}\mathrm{R}_x (0_x) = \mathrm{id}_{T_x\mathcal{M}}$ and the inverse function theorem, there exists a neighborhood V of 0_x in $T_x\mathcal{M}$ such that R_x is a diffeomorphism on V; thus, $\mathrm{R}_x^{-1}(y)$ is well defined for all $y \in \mathcal{M}$ sufficiently close to x. In this case, $\mathrm{R}_x(V) \subset \mathcal{M}$ is called a retractive neighborhood of x. Furthermore, the existence of a *totally* retractive neighborhood [38, Theorem 2] shows that for any $\bar{x} \in \mathcal{M}$ there is a neighborhood W of $\bar{x}$ such that $\mathrm{R}_x^{-1}(y)$ is well defined for all $x, y \in W$. In what follows, we will suppose that an appropriate neighborhood has been chosen by default for the well-definedness of $\mathrm{R}_x^{-1}(y)$.

4.1 Vector Transport and Parallel Transport

Define the Whitney sum $T\mathcal{M} \oplus T\mathcal{M} := \{(\eta, \xi) : \eta, \xi \in T_x\mathcal{M}, x \in \mathcal{M}\}$. A smooth map $\mathrm{T} : T\mathcal{M} \oplus T\mathcal{M} \to T\mathcal{M} : (\eta, \xi) \mapsto \mathrm{T}_\eta(\xi)$, is called a *vector transport* on $\mathcal{M}$ if there exists an associated retraction R on $\mathcal{M}$ such that T satisfies the following properties for all $x \in \mathcal{M}$:

1. Associated retraction. $\mathrm{T}_\eta(\xi) \in T_{\mathrm{R}_x(\eta)}\mathcal{M}$ for all $\eta, \xi \in T_x\mathcal{M}$.
2. Consistency. $\mathrm{T}_{0_x}(\xi) = \xi$ for all $\xi \in T_x\mathcal{M}$.
3. Linearity. $\mathrm{T}_\eta(a\xi + b\zeta) = a\mathrm{T}_\eta(\xi) + b\mathrm{T}_\eta(\zeta)$ for all $a, b \in \mathbb{R}$ and $\eta, \xi, \zeta \in T_x\mathcal{M}$.

Thus, fixing any $\eta \in T_x\mathcal{M}$, the map $\mathrm{T}_\eta : T_x\mathcal{M} \to T_{\mathrm{R}_x(\eta)}\mathcal{M} : \xi \mapsto \mathrm{T}_\eta(\xi)$, is a linear operator. Additionally, T is *isometric* if $\langle \mathrm{T}_\eta(\xi), \mathrm{T}_\eta(\zeta) \rangle = \langle \xi, \zeta \rangle$ holds, for all $x \in \mathcal{M}$ and all $\xi, \zeta, \eta \in T_x\mathcal{M}$. In other words, for any $\eta \in T_x\mathcal{M}$, the adjoint and the inverse of T_η coincide, i.e., $\mathrm{T}_\eta^* = \mathrm{T}_\eta^{-1}$. There are two important classes of vector transport as follows. Let R be a retraction on $\mathcal{M}$.

(1) The *differentiated retraction* defined by

$$\mathrm{T}_\eta(\xi) := \mathcal{D}\mathrm{R}_x(\eta)[\xi], \ \eta, \xi \in T_x\mathcal{M}, x \in \mathcal{M}, \tag{22}$$

is a valid vector transport [1, equation (8.6)].

(2) Given a smooth curve $\gamma : [0, 1] \to \mathcal{M}$ and $t_0, t_1 \in [0, 1]$, the *parallel transport* from the tangent space at $\gamma(t_0)$ to the tangent space at $\gamma(t_1)$ along γ, is a linear operator $\mathrm{P}_\gamma^{t_1 \to t_0} : T_{\gamma(t_0)}\mathcal{M} \to T_{\gamma(t_1)}\mathcal{M}$ defined by $\mathrm{P}_\gamma^{t_1 \to t_0}(\xi) = Z(t_1)$, where Z is the unique parallel vector field such that $Z(t_0) = \xi$. Then, for any $x \in \mathcal{M}, \eta \in T_x\mathcal{M}$, then

$$\mathrm{T}_\eta(\xi) := \mathrm{P}_\gamma^{1 \to 0}(\xi) \tag{23}$$

is a valid vector transport [1, equation (8.2)], where P_γ denotes the parallel transport along the curve $t \mapsto \gamma(t) := \mathrm{R}_x(t\eta)$. We often omit the superscript $1 \to 0$ if it is clear from the context. In particular, parallel transport is isometric.

4.2 Lipschitz Continuity with respect to Vector Transports

Multiple Riemannian versions of Lipschitz continuity have been defined, e.g., [5, Sect. 10.4]. Here, we consider Lipschitz continuity with respect to a vector transport. In what follows, let $\mathcal{M}$ be a Riemannian manifold endowed with a vector transport T and an associated retraction R. We first consider the Lipschitz continuous gradient of scale field f.

Definition 4.1 ([17, Definition 5.2.1]) A function $f : \mathcal{M} \to \mathbb{R}$ is *Lipschitz continuously differentiable with respect to* T *in* $\mathcal{U} \subset \mathcal{M}$ if it is differentiable and there exists a constant $\kappa > 0$ such that, for all $x, y \in \mathcal{U}$, $\left\| \text{grad } f(y) - T_\eta[\text{grad } f(x)] \right\| \leqslant \kappa \|\eta\|$, where $\eta = R_x^{-1} y$.

Going one degree higher, let us now discuss the Lipschitz continuity of Hessian operators. Throughout this paper, for a linear operator $\mathcal{A} : \mathcal{E} \to \mathcal{E}'$ between two finite-dimensional normed vector spaces $\mathcal{E}$ and $\mathcal{E}'$, the (operator) norm of $\mathcal{A}$ is defined by $\|\mathcal{A}\| := \sup \{ \|\mathcal{A}v\|_{\mathcal{E}'} : v \in \mathcal{E}, \|v\|_{\mathcal{E}} = 1, \text{ or, } \|v\|_{\mathcal{E}} \leqslant 1 \}$. The inverse of T_η is needed in the following definitions, so we can assume that vector transport T is isometric, e.g., parallel transport in (23). In fact, there are many ways to construct isometric vector transports; see [18, Sect. 2.3].

Definition 4.2 ([18, Assumption 3]) A function $f : \mathcal{M} \to \mathbb{R}$ is *twice Lipschitz continuously differentiable with respect to* T *in* $\mathcal{U} \subset \mathcal{M}$ if it is twice differentiable and there exists a constant $\kappa > 0$ such that, for all $x, y \in \mathcal{U}$, $\| \text{Hess } f(y) - T_\eta \text{ Hess } f(x) T_\eta^{-1} \| \leqslant \kappa d(x, y)$, where $\eta = R_x^{-1} y$.

Lemma 4.1 ([18, Lemma 4]) *If* $f : \mathcal{M} \to \mathbb{R}$ *is* C^3, *then for any* $\bar{x} \in \mathcal{M}$ *and any isometric vector transport* T, *there exists a neighborhood* $\mathcal{U}$ *of* $\bar{x}$ *such that* f *is twice Lipschitz continuously differentiable with respect to* T *in* $\mathcal{U}$.

If the operator, Hess $f(x)$, above is replaced by a general covariant derivative $\nabla F(x)$, we can get the next results in a similar way. Lemma 4.2 can be proven in the same way as Lemma 4.1.

Definition 4.3 Given a vector field F on $\mathcal{M}$. The map $x \mapsto \nabla F(x)$ is *Lipschitz continuous with respect to* T *in* $\mathcal{U} \subset \mathcal{M}$ if there exists a constant $\kappa > 0$ such that, for all $x, y \in \mathcal{U}$, it holds that $\|\nabla F(y) - T_\eta \nabla F(x) T_\eta^{-1}\| \leqslant \kappa d(x, y)$, where $\eta = R_x^{-1} y$.

Lemma 4.2 *If* F *is a* C^2 *vector field, then for any* $\bar{x} \in \mathcal{M}$ *and any isometric vector transport* T, *there exists a neighborhood* $\mathcal{U}$ *of* $\bar{x}$ *such that the map* $x \mapsto \nabla F(x)$ *is Lipschitz continuous with respect to* T *in* $\mathcal{U}$.

4.3 Auxiliary Lemmas

Notice that in the previous subsection on the definitions of Lipschitz continuity, we used $\|\eta\|$ with $\eta = R_x^{-1} y$ or $d(x, y)$ to denote the upper bound on the right-hand side. The next lemma shows that the two are not essentially different. When $\mathcal{M} = \mathbb{R}^n$, both reduce to $\|x - y\|$.

Lemma 4.3 ([18, Lemma 2]) *Let $\mathcal{M}$ be a Riemannian manifold with a retraction R and let $\bar{x} \in \mathcal{M}$. Then,*

(i) *there exist $a_0, a_1, \delta_{a_0, a_1} > 0$ such that for all x in a sufficiently small neighborhood of $\bar{x}$ and all $\xi, \eta \in T_x \mathcal{M}$ with $\|\xi\|, \|\eta\| \leqslant \delta_{a_0, a_1}$, one has $a_0 \|\xi - \eta\| \leqslant d(\mathrm{R}_x(\eta), \mathrm{R}_x(\xi)) \leqslant a_1 \|\xi - \eta\|$. In particular, $a_0 \|\xi\| \leqslant d(x, \mathrm{R}_x(\xi)) \leqslant a_1 \|\xi\|$ when $\eta = 0$;*

(ii) *there exist $a_0, a_1 > 0$ such that for all x in a sufficiently small neighborhood of $\bar{x}$, one has $a_0 \|\xi\| \leqslant d(x, \bar{x}) \leqslant a_1 \|\xi\|$ where $\xi = \mathrm{R}_{\bar{x}}^{-1}(x)$.*

The next lemma is the fundamental theorem of calculus in the Riemannian case.

Lemma 4.4 ([18, Lemma 8]) *Let F be a C^1 vector field and $\bar{x} \in \mathcal{M}$. Then there exist a neighborhood $\mathcal{U}$ of $\bar{x}$ and a constant $c_1 \geqslant 0$ such that for all $x, y \in \mathcal{U}$,*

$$\left\| \mathrm{P}_\gamma^{0 \to 1}[F(y)] - F(x) - \int_0^1 \mathrm{P}_\gamma^{0 \to t} \nabla F(\gamma(t)) \mathrm{P}_\gamma^{t \to 0}[\eta] \mathrm{d}t \right\| \leqslant c_1 \|\eta\|^2,$$

where $\eta = \mathrm{R}_x^{-1}(y)$ and P_γ is the parallel transport along the curve $\gamma(t) := \mathrm{R}_x(t\eta)$. Moreover, if $\mathrm{R} = \mathrm{Exp}$, then indeed $c_1 = 0$ above (see [13, equation (2.4)]).

The next lemma is a Riemannian extension of some important estimates, usually used to analyze Newton methods [8, Lemma 4.1.12].

Lemma 4.5 *Let F be a C^2 vector field and $\bar{x} \in \mathcal{M}$. Then, there exist a neighborhood $\mathcal{U}$ of $\bar{x}$ and a constant $c_2 > 0$ such that for all $x \in \mathcal{U}$,*

$$\left\| \mathrm{P}_\gamma^{0 \to 1}[F(x)] - F(\bar{x}) - \nabla F(\bar{x})[\eta] \right\| \leqslant c_2 d^2(\bar{x}, x),$$

where $\eta = \mathrm{R}_{\bar{x}}^{-1} x$ and P_γ is the parallel transport along the curve $\gamma(t) := \mathrm{R}_{\bar{x}}(t\eta)$.

Proof Let $LHS := \|\mathrm{P}_\gamma^{0 \to 1}[F(x)] - F(\bar{x}) - \nabla F(\bar{x})[\eta]\|$. It follows that

$$
\begin{aligned}
LHS \\
\leqslant & \left\| \mathrm{P}_\gamma^{0 \to 1}[F(x)] - F(\bar{x}) - \int_0^1 \mathrm{P}_\gamma^{0 \to t} \nabla F(\gamma(t)) \mathrm{P}_\gamma^{t \to 0}[\eta] \mathrm{d}t \right\| \\
& + \left\| \int_0^1 \mathrm{P}_\gamma^{0 \to t} \nabla F(\gamma(t)) \mathrm{P}_\gamma^{t \to 0}[\eta] \mathrm{d}t - \nabla F(\bar{x})[\eta] \right\| \\
\leqslant & c_1 \|\eta\|^2 + \left\| \int_0^1 \left(\mathrm{P}_\gamma^{0 \to t} \nabla F(\gamma(t)) \mathrm{P}_\gamma^{t \to 0} - \nabla F(\bar{x}) \right) [\eta] \mathrm{d}t \right\|. \text{ (by Lemma 4.4)}
\end{aligned}
$$

Let $\theta := \left\| \int_0^1 \left(P_\gamma^{0 \to t} \nabla F(\gamma(t)) P_\gamma^{t \to 0} - \nabla F(\bar{x}) \right) [\eta] dt \right\|$. Note that

$$
\begin{aligned}
\theta &\leqslant \int_0^1 \left\| P_\gamma^{0 \to t} \nabla F(\gamma(t)) P_\gamma^{t \to 0} - \nabla F(\bar{x}) \right\| \|\eta\| \, dt \\
&\leqslant \int_0^1 c_0 d(\bar{x}, R_{\bar{x}}(t\eta)) \|\eta\| \, dt \text{ (by Lemma 4.2)} \\
&\leqslant \int_0^1 c_0 a_1 t \|\eta\| \|\eta\| \, dt = \frac{1}{2} c_0 a_1 \|\eta\|^2. \text{ (by (i) of Lemma 4.3)}
\end{aligned}
$$

Combining the above results yields

$$
LHS \leqslant (c_1 + \frac{1}{2} c_0 a_1) \|\eta\|^2 \leqslant (c_1 + \frac{1}{2} c_0 a_1)/a_0^2 d^2(\bar{x}, x),
$$

where the last inequality comes from (ii) of Lemma 4.3. Letting $c_2 := (c_1 + \frac{1}{2} c_0 a_1)/a_0^2$ completes the proof. $\qquad\square$

We end this section with the following useful lemmas.

Lemma 4.6 ([12, Lemma 3.2]) *Given a vector field F on $\mathcal{M}$. If the map $p \mapsto \nabla F(p)$ is continuous at p^* and $\nabla F(p^*)$ is nonsingular, then there exist a neighborhood $\mathcal{U}$ of p^* and a constant $\Xi > 0$ such that, for all $p \in \mathcal{U}$, $\nabla F(p)$ is nonsingular and $\left\| \nabla F(p)^{-1} \right\| \leqslant \Xi$.*

Lemma 4.7 ([14, Lemma 14.5]) *Let F be a C^2 vector field on $\mathcal{M}$ and $p^* \in \mathcal{M}$. If $F(p^*) = 0$ and $\nabla F(p^*)$ is nonsingular, then there exist a neighborhood $\mathcal{U}$ of p^* and constants $c_3, c_4 > 0$ such that, for all $p \in \mathcal{U}$, $c_3 d(p, p^*) \leqslant \|F(p)\| \leqslant c_4 d(p, p^*)$.*

Lemma 4.8 ([7, Lemma 3.5]) *Let $u \in T_p\mathcal{M}$ such that $\operatorname{Exp}_p(u)$ exists and $v \in T_p\mathcal{M} \cong T_p(T_p\mathcal{M})$. Then, $\langle \mathcal{D} \operatorname{Exp}_p(u)[u], \mathcal{D} \operatorname{Exp}_p(u)[v] \rangle = \langle u, v \rangle$. In particular, $\left\| \mathcal{D} \operatorname{Exp}_p(\lambda u)[u] \right\| = \|u\|$ holds all $\lambda \geqslant 0$.*

5 Local Convergence

Here, for any two nonnegative sequences $\{u_k\}$ and $\{v_k\}$, we write $u_k = O(v_k)$ if there is a constant $M > 0$ such that $u_k \leqslant M v_k$ for all sufficiently large k, and we write $u_k = o(v_k)$ if $v_k > 0$ and the sequence of ratios $\{u_k/v_k\}$ approaches zero. In this section, we will establish local convergence of our prototype Algorithm 2 of RIPM.

5.1 Perturbed Damped Riemannian Newton Method

We will rely on an application of the so-called *perturbed damped* Riemannian Newton method for solving the singularity problem (2), which can be stated as Algorithm 4.

In contrast to the standard Riemannian Newton method described in Algorithm 1, the term "perturbed" means that we solve a Newton equation with a perturbed term

---

Algorithm 4: Perturbed Damped Riemannian Newton Method for (2)

Input: A vector field $F \in \mathfrak{X}(\mathcal{M})$, an initial point $p_0 \in \mathcal{M}$ and a retraction R on $\mathcal{M}$. Set $\mu_0 > 0$.
Output: Sequence $\{p_k\} \subset \mathcal{M}$ such that $\{p_k\} \to p^*$ and $F(p^*) = 0_{p^*}$.
Set $k \to 0$;
while *Stopping criterion not satisfied* **do**

    1. Obtain $\xi_k \in T_{p_k}\mathcal{M}$ by solving the perturbed Newton equation:

$$\nabla F(p_k)[\xi_k] = -F(p_k) + \mu_k \hat{e}; \tag{24}$$

    2. Choose a (damped) step size $0 < \alpha_k \leqslant 1$;
    3. Compute the next point as $p_{k+1} := \mathrm{R}_{p_k}(\alpha_k \xi_k)$;
    4. Choose $0 < \mu_{k+1} < \mu_k$;
    5. $k \to k + 1$;

end

$\mu_k \hat{e}$, while "damped" means using α_k instead of unit steps. It is well known that Algorithm 1 is locally superlinearly [12] and quadratically [13] convergent under the following Riemannian Newton assumptions:

(B1) There exists $p^* \in \mathcal{M}$ such that $F(p^*) = 0_{p^*}$;
(B2) The covariant derivative $\nabla F(p^*)$ is nonsingular;
(B3) The vector field F is C^2.

As Proposition 5.1 shows, Algorithm 4 also has the same convergence properties as Algorithm 1 if we control μ_k and α_k according to the two schemes that Proposition 5.1 gives. We can see that either scheme will have $\mu_k \to 0$ and $\alpha_k \to 1$, which makes Algorithm 4 eventually reduce to Algorithm 1 when k is sufficiently large.

Proposition 5.1 (Local convergence of Algorithm 4) *Consider the perturbed damped Riemannian Newton method described in Algorithm 4 for the singularity problem* (2). *Let* **(B1)**–**(B3)** *hold. Choose parameters μ_k, α_k as follows; then there exists a constant $\delta > 0$ such that for all $p_0 \in \mathcal{M}$ with $d(p_0, p^*) < \delta$, the sequence $\{p_k\}$ is well defined. Furthermore,*

(1) if we choose $\mu_k = o(\|F(p_k)\|)$ and $\alpha_k \to 1$, then $p_k \to p^*$ superlinearly;
(2) if we choose $\mu_k = O(\|F(p_k)\|^2)$ and $1 - \alpha_k = O(\|F(p_k)\|)$, then $p_k \to p^*$ quadratically.

Proof By **(B2)**–**(B3)**, Lemmas 4.2 and 4.6, we can let p_k be sufficiently close to p^* such that $\nabla F(p_k)$ is nonsingular, and $\left\| \nabla F(p_k)^{-1} \right\| \leqslant \varXi$ for some constant $\varXi$. Then, the next iterate point,

$$p_{k+1} := \mathrm{R}_{p_k}[\alpha_k \nabla F(p_k)^{-1}(-F(p_k) + \mu_k \hat{e})],$$

is well defined in Algorithm 4, and it follows from $p^* = \mathrm{R}_{p_k}(\eta)$ with $\eta := \mathrm{R}_{p_k}^{-1} p^*$ and (i) of Lemma 4.3 that

$$\begin{aligned}
d(p_{k+1}, p^*) &\leqslant a_1 \| \eta - \alpha_k \nabla F(p_k)^{-1}(-F(p_k) + \mu_k \hat{e}) \| \\
&= a_1 \| \eta + \alpha_k \nabla F(p_k)^{-1}(F(p_k) - \mu_k \hat{e}) \|.
\end{aligned} \tag{25}$$

Let $r_k := \eta + \alpha_k \nabla F(p_k)^{-1}(F(p_k) - \mu_k \hat{e})$. Algebraic manipulations show that

$$r_k = (1 - \alpha_k)\eta + \alpha_k \nabla F(p_k)^{-1}[\nabla F(p_k)\eta + F(p_k) - P_\gamma^{0 \to 1} F(p^*) - \mu_k \hat{e}],$$

where P_γ is the parallel transport along the curve $\gamma(t) = R_{p_k}(t\eta)$ and $F(p^*) = 0$. Thus, using $\|\eta\| \leqslant d(p_k, p^*)/a_0$ from (ii) of Lemmas 4.3 and 4.5, we have

$$\begin{aligned}
\|r_k\| &\leqslant (1 - \alpha_k)\|\eta\| + \alpha_k \|\nabla F(p_k)^{-1}\| \|P_\gamma^{0 \to 1} F(p^*) - F(p_k) - \nabla F(p_k)\eta\| \\
&\quad + \alpha_k \|\nabla F(p_k)^{-1}\| \|\hat{e}\|\mu_k \\
&\leqslant (1 - \alpha_k)d(p_k, p^*)/a_0 + \alpha_k \|\nabla F(p_k)^{-1}\| c_2 d^2(p_k, p^*) \\
&\quad + \alpha_k \|\nabla F(p_k)^{-1}\| \|\hat{e}\|\mu_k \\
&\leqslant (1 - \alpha_k)d(p_k, p^*)/a_0 + \Xi c_2 d^2(p_k, p^*) + \Xi \|\hat{e}\|\mu_k.
\end{aligned}$$

Combining the above with (25), we conclude that

$$d(p_{k+1}, p^*) \leqslant \kappa_1(1 - \alpha_k)d(p_k, p^*) + \kappa_2 d^2(p_k, p^*) + \kappa_3 \mu_k \tag{26}$$

for some positive constants $\kappa_1, \kappa_2, \kappa_3$. On the other hand, by Lemma 4.7, we have

$$\|F(p_k)\| = O(d(p_k, p^*)). \tag{27}$$

In what follows, we prove assertions (1) and (2).

(1) Suppose that $\alpha_k \to 1$ and $\mu_k = o(\|F(p_k)\|)$, which together with (27) imply $\mu_k = o(d(p_k, p^*))$. By (26), we have

$$\frac{d(p_{k+1}, p^*)}{d(p_k, p^*)} \leqslant \kappa_1(1 - \alpha_k) + \kappa_2 d(p_k, p^*) + \kappa_3 \frac{\mu_k}{d(p_k, p^*)}. \tag{28}$$

We can take δ sufficiently small and k sufficiently large, if necessary, to conclude that $d(p_{k+1}, p^*) < \frac{1}{2}d(p_k, p^*) < \delta$. Thus, $p_{k+1} \in B_\delta(p^*)$, the open ball of radius δ centered at p^* on $\mathcal{M}$. By induction, it is easy to show that the sequence $\{p_k\}$ is well defined and converges to p^*. Taking the limit of both sides of (28) proves superlinear convergence.

(2) Again, we start from (26) and rewrite it as:

$$d(p_{k+1}, p^*) = (1 - \alpha_k)O(d(p_k, p^*)) + O(d^2(p_k, p^*)) + O(\mu_k). \tag{29}$$

Suppose that $1 - \alpha_k = O(\|F(p_k)\|)$ and $\mu_k = O(\|F(p_k)\|^2)$. Using (27), the above reduces to $d(p_{k+1}, p^*) = O(d^2(p_k, p^*))$. This implies that there exists a constant ν such that $d(p_{k+1}, p^*) \leqslant \nu d^2(p_k, p^*)$, and hence, $d(p_{k+1}, p^*) \leqslant \nu d^2(p_k, p^*) \leqslant \nu \delta^2 < \delta$, if δ is sufficiently small. Again, by induction, $\{p_k\}$ converges to p^* quadratically. $\qquad \square$

5.2 Local Convergence of Algorithm 2

Next, lemma shows the relationship between the parameter γ_k and step size α_k in Algorithm 2.

Lemma 5.1 *Consider the Algorithm* 2 *for solving the problem* (RCOP). *Let* (**A1**) *and* (**A3**) *hold at some* $w^* = (x^*, y^*, z^*, s^*)$ *and* α_k *be as in* (17). *Define a constant,*

$$\Pi := 2 \max \left\{ \max_i \left\{ 1/(s^*)_i \mid (s^*)_i > 0 \right\}, \max_i \left\{ 1/(z^*)_i \mid (z^*)_i > 0 \right\} \right\}.$$

For $\gamma_k \in (0, 1)$, *if* $\Pi \|\Delta w_k\| \leqslant \gamma_k$, *then* $0 \leqslant 1 - \alpha_k \leqslant (1 - \gamma_k) + \Pi \|\Delta w_k\|$.

Proof Notice that the fourth line of (18) yields

$$S_k^{-1} \Delta s_k + Z_k^{-1} \Delta z_k = \mu_k (S_k Z_k)^{-1} e - e,$$

which is exactly the same as in the usual interior point method in the Euclidean setting. Thus, the proof entails directly applying [35, Lemma 3 and 4] for the Euclidean case to the Riemannian case. □

Now, let us establish the local convergence of our Algorithm 2 in a way that replicates Proposition 5.1 except for taking account of parameter γ_k.

Theorem 5.1 (Local convergence of prototype Algorithm 2) *Consider the Algorithm* 2 *for solving the problem* (RCOP). *Let* (**A1**)–(**A4**) *hold at some* w^*. *Choose parameters* μ_k, γ_k *as follows; then there exists a constant* $\delta > 0$ *such that, for all* $w_0 \in \mathcal{N}$ *with* $d(w_0, w^*) < \delta$, *the sequence* $\{w_k\}$ *is well defined. Furthermore,*

(1) *if we choose* $\mu_k = o(\|F(w_k)\|)$ *and* $\gamma_k \to 1$, *then* $w_k \to w^*$ *superlinearly;*
(2) *if we choose* $\mu_k = O(\|F(w_k)\|^2)$ *and* $1 - \gamma_k = O(\|F(w_k)\|)$, *then* $w_k \to w^*$ *quadratically.*

Proof We only prove (2) because (1) can be proven in the same way. Let w_k be such that $d(w_k, w^*) < \delta$ for sufficiently small δ. From Proposition 3.1, (**A1**)–(**A4**) shows that the KKT vector field F satisfies (**B1**)–(**B3**); thus, the discussion in the proof of Proposition 5.1 applies to KKT vector field F as well, simply by replacing the symbol p with w. Since we choose $\mu_k = O(\|F(w_k)\|^2)$, and $\|F(w_k)\| = O(d(w_k, w^*))$ by (27), we obtain $\mu_k = O(d^2(w_k, w^*))$. Thus,

$$
\begin{aligned}
\|\Delta w_k\| &= \left\| \nabla F(w_k)^{-1} (-F(w_k) + \mu_k \hat{e}) \right\| \text{ (by (16))} \\
&\leqslant \Xi (\|F(w_k)\| + \mu_k \|\hat{e}\|) \text{ (by Lemma 4.6)} \\
&\leqslant O(\|F(w_k)\|) + O(\mu_k) \\
&= O(d(w_k, w^*)) + O(d^2(w_k, w^*)) = O(d(w_k, w^*)).
\end{aligned}
$$

Since δ is sufficiently small, from the above inequalities, the condition of Lemma 5.1 is satisfied, i.e., $\|\Delta w_k\| \leqslant \hat{\gamma}/\Pi \leqslant \gamma_k/\Pi$ for a constant $\hat{\gamma} \in (0, 1)$. Hence, by

$1 - \gamma_k = O(\|F(w_k)\|)$, one has $1 - \alpha_k \leqslant (1 - \gamma_k) + \Pi \|\Delta w_k\| = (1 - \gamma_k) + O(d(w_k, w^*)) = O(d(w_k, w^*))$. Finally, from (29), we have $d(w_{k+1}, w^*) = (1 - \alpha_k)O(d(w_k, w^*)) + O(d^2(w_k, w^*)) + O(\mu_k) = O(d^2(w_k, w^*))$. This completes the proof. □

Algorithm 2 guaranteed the local convergence, but we are more interested in its globally convergent version (described in the next section). We still provide a simple example of Algorithm 2 online.[2] where from (2) of Theorem 5.1, in practice we set parameters $\mu_{k+1} = \min\{\mu_k/1.5, 0.5 \|F(w_k)\|^2\}$; and $\gamma_{k+1} = \max\{\hat{\gamma}, 1 - \|F(w_k)\|\}$ with $\hat{\gamma} = 0.5$.

6 Global Algorithm

The globally convergent version of our RIPM uses the classical line search described in [11]. The following considerations and definitions are needed in order to describe it compactly. For simplicity, we often omit the subscript of iteration count k.

Given the current point $w = (x, y, z, s)$ and $\Delta w = (\Delta x, \Delta y, \Delta z, \Delta s)$, the next iterate is obtained along a curve on product manifold $\mathcal{N}$, i.e., $\alpha \mapsto w(\alpha) := \bar{R}_w(\alpha \Delta w)$ with some step size $\alpha > 0$, see (6) for $\bar{R}_w$. By introducing $w(\alpha) = (x(\alpha), y(\alpha), z(\alpha), s(\alpha))$, we have $x(\alpha) = R_x(\alpha \Delta x)$, $y(\alpha) = y + \alpha \Delta y$, $z(\alpha) = z + \alpha \Delta z$, and $s(\alpha) = s + \alpha \Delta s$. For a given starting point $w_0 \in \mathcal{N}$ with $(z_0, s_0) > 0$, let us set two constants

$$\tau_1 := \min(Z_0 S_0 e)/(z_0^T s_0/m), \ \ \tau_2 := z_0^T s_0 / \|F(w_0)\|.$$

As well, define two functions $f^I(\alpha) := \min(Z(\alpha)S(\alpha)e) - \gamma \tau_1 z(\alpha)^T s(\alpha)/m$, and $f^{II}(\alpha) := z(\alpha)^T s(\alpha) - \gamma \tau_2 \|F(w(\alpha))\|$, where $\gamma \in (0, 1)$ is a constant. For $i = I, II$, define

$$\alpha^i := \max_{\alpha \in (0,1]} \left\{ \alpha : f^i(t) \geqslant 0, \ \text{for all} \ t \in (0, \alpha] \right\}, \tag{30}$$

i.e., α^i are either one or the smallest positive root for the functions $f^i(\alpha)$ in $(0, 1]$.

Moreover, we define the merit function $\varphi \colon \mathcal{N} \to \mathbb{R}$ by $\varphi(w) := \|F(w)\|^2$; accordingly, we have $\text{grad} \, \varphi(w) = 2\nabla F(w)^*[F(w)]$, where symbol $*$ means its adjoint operator. Let $\| \cdot \|_1, \| \cdot \|_2$ be l_1, l_2 vector norms. Note that $\|F(w)\|_w^2 = \left\|\text{grad}_x \, \mathcal{L}(w)\right\|_x^2 + \|h(x)\|_2^2 + \|g(x) + s\|_2^2 + \|ZSe\|_2^2$ by (5). Moreover, for any non-negative $z, s \in \mathbb{R}^m$, one has $\|ZSe\|_2 \leqslant z^T s = \|ZSe\|_1 \leqslant \sqrt{m} \|ZSe\|_2$. Hence,

$$\|ZSe\|_2 / \sqrt{m} \leqslant z^T s / \sqrt{m} \leqslant \|ZSe\|_2 \leqslant \|F(w)\|. \tag{31}$$

Now, using the above definitions, the globally convergent RIPM can be stated as Algorithm 5.

Compared to the prototype Algorithm 2, the global Algorithm 5 involves a more elaborate choice of step size. Regarding the centrality condition (3a), it does not differ

---

[2] See .../LocalRIPM/PrototypeRIPM.m in https://doi.org/10.5281/zenodo.10612799.

Algorithm 5: Global Convergent Algorithm of RIPM for (RCOP)

Input: An initial point $w_0 = (x_0, y_0, z_0, s_0) \in \mathcal{N}$ with $(z_0, s_0) > 0$ and a retraction R on $\mathcal{M}$.
        $\theta \in (0, 1)$, $\beta \in (0, 0.5]$, $\gamma_{-1} \in (0.5, 1)$.
Output: Sequence $\{w_k\} \subset \mathcal{N}$ such that $\{w_k\} \to w^*$ and w^* satisfies the KKT conditions (4).
Set $k \to 0$;
while *Stopping criterion not satisfied* **do**
    1. Set $\sigma_k \in (0, 1)$, $\rho_k \in [z_k^T s_k / m, \|F(w_k)\| / \sqrt{m}]$;
    2. Obtain $\Delta w_k = (\Delta x_k, \Delta y_k, \Delta z_k, \Delta s_k) \in T_{w_k} \mathcal{N}$ by solving the following linear equation:

$$\nabla F(w_k)[\Delta w_k] = -F(w_k) + \sigma_k \rho_k \hat{e}; \qquad (32)$$

    3. Step size selection:
    (3a) Centrality condition: Set $\gamma_k \in (0.5, \gamma_{k-1})$ and $\bar{\alpha}_k = \min\{\alpha_k^I, \alpha_k^{II}\}$ from (30);
    (3b) Sufficient decrease condition: Let $\alpha_k := \bar{\alpha}_k$;
    while α_k *does not satisfy the condition* $\varphi(\bar{R}_{w_k}(\alpha_k \Delta w_k)) - \varphi(w_k) \leqslant \alpha_k \beta \langle \text{grad } \varphi_k, \Delta w_k \rangle$ **do**
        $\alpha_k := \theta \alpha_k$;
    end
    4. Compute the next point as $w_{k+1} := \bar{R}_{w_k}(\alpha_k \Delta w_k)$;
    5. $k \to k + 1$;
end

---

in any way from the Euclidean setting; thus, references [3, 9, 11] show that $\bar{\alpha}_k$ is well-defined, thereby ensuring that z_k, s_k are positive. In the following, we focus on the sufficient decrease (Armijo) condition (3b):

$$\varphi(\bar{R}_{w_k}(\alpha_k \Delta w_k)) - \varphi(w_k) \leqslant \alpha_k \beta \langle \text{grad } \varphi_k, \Delta w_k \rangle, \qquad (33)$$

where grad $\varphi_k \equiv$ grad $\varphi(w_k)$ for short. With a slight abuse of notation φ, at the current point w and direction Δw, we define a real-to-real function $\alpha \mapsto \varphi(\alpha) := \varphi(\bar{R}_w(\alpha \Delta w))$; then, it follows from the definition of a retraction and the chain rule that

$$\varphi'(0) = \mathcal{D}\varphi(\bar{R}_w(0)) \left[\mathcal{D}\bar{R}_w(0)[\Delta w] \right] = \mathcal{D}\varphi(w)[\Delta w] = \langle \text{grad } \varphi(w), \Delta w \rangle.$$

Hence, $\varphi_k'(0) = \langle \text{grad } \varphi_k, \Delta w_k \rangle$ at the k-th iterate w_k, and then, the Armijo condition (33) reads $\varphi_k(\alpha_k) - \varphi_k(0) \leqslant \alpha_k \beta \varphi_k'(0)$ as usual. If $\varphi_k'(0) < 0$, the backtracking loop in (3b) of Algorithm 5 will eventually stop [26, Lemma 3.1]. The next lemma shows the condition under which the Newton direction Δw_k generated by (32) ensures the descent of the merit function.

Lemma 6.1 *If the direction Δw_k is the solution of equation (32), then*

$$\langle \text{grad } \varphi(w_k), \Delta w_k \rangle = 2(-\|F(w_k)\|^2 + \sigma_k \rho_k z_k^T s_k).$$

In this case, Δw_k is a descent direction for $\varphi(w)$ at w_k if and only if $\rho_k < \|F(w_k)\|^2 / \sigma_k z_k^T s_k$.

Proof The iteration count k is omitted. Let Δw be given by (32). Then, we have $\langle \text{grad } \varphi(w), \Delta w \rangle = \langle 2\nabla F(w)^*[F(w)], \Delta w \rangle = 2\langle F(w), \nabla F(w)[\Delta w] \rangle =$

$2\langle F(w), -F(w) + \sigma\rho\hat{e}\rangle = 2(-\langle F(w), F(w)\rangle + \sigma\rho\langle F(w), \hat{e}\rangle)$. Then, by definition of $\hat{e}$ in (13), $\langle F(w), \hat{e}\rangle = \langle ZSe, e\rangle = z^T s$ completes the proof. □

The next proposition shows that Algorithm 5 can generate the monotonically non-increasing sequence $\{\varphi_k\}$. Note that $\varphi_{k+1} \equiv \varphi_k(\alpha_k)$ and $\varphi_k \equiv \varphi_k(0) \equiv \varphi(w_k)$.

Proposition 6.1 *If $\|F(w_k)\| \neq 0$, then the direction Δw_k generated by Algorithm 5 is a descent direction for $\varphi(w)$ at w_k. Moreover, if the Armijo condition (33) is satisfied, then*

$$\varphi_k(\alpha_k) \leqslant [1 - 2\alpha_k\beta(1 - \sigma_k)]\varphi_k(0).$$

Thus, the sequence $\{\varphi_k\}$ is monotonically nonincreasing.

Proof The iteration count k is omitted. Suppose that $\rho \leqslant \|F(w)\|/\sqrt{m}$ and Δw is given by (32), we have

$$\begin{aligned} \varphi'(0) &= \langle \text{grad } \varphi(w), \Delta w\rangle = 2(-\varphi(w) + \sigma\rho z^T s) \text{ (by Lemma 6.1)} \\ &\leqslant 2(-\varphi(w) + \sigma\|F(w)\|z^T s/\sqrt{m}) \\ &\leqslant 2(-\varphi(w) + \sigma\|F(w)\|^2) \text{ (by (31))} \\ &= -2(1 - \sigma)\varphi(w) < 0. \end{aligned} \tag{34}$$

Thus, in Algorithm 5, Δw is a descent direction for the merit function φ at w. Alternatively, by Lemma 6.1, it is sufficient to show that $\|F(w)\|/\sqrt{m} < \|F(w)\|^2/\sigma z^T s$. By $\sigma z^T s < z^T s \leqslant \sqrt{m}\|F(w)\|$; then, $1/\sqrt{m} < \|F(w)\|/\sigma z^T s$. Multiplying both sides by $\|F(w)\|$ yields the result.

Moreover, if condition (33) is satisfied, then by (34), we have $\varphi(\alpha) \leqslant \varphi(0) + \alpha\beta\langle \text{grad } \varphi(w), \Delta w\rangle \leqslant \varphi(0) + \alpha\beta(-2(1 - \sigma)\varphi(0)) = [1 - 2\alpha\beta(1 - \sigma)]\varphi(0)$. Note that in Algorithm 5, we set $\beta \in (0, 1/2]$, $\sigma \in (0, 1)$, and $\alpha \in (0, 1]$, which imply that the sequence $\{\varphi_k\}$ is monotonically nonincreasing. □

Finally, we need to make the following assumptions. For $\epsilon \geqslant 0$, let $\Omega(\epsilon) := \{w \in \mathcal{N} \mid \epsilon \leqslant \varphi(w) \leqslant \varphi_0, \min(ZSe)/(z^T s/m) \geqslant \tau_1/2, z^T s/\|F(w)\| \geqslant \tau_2/2\}$.

- **(C1)** In the set $\Omega(0)$, f, h, and g are smooth functions; $\{\text{grad } h_i(x)\}_{i=1}^l$ is linearly independent for all x; and the map $w \mapsto \nabla F(w)$ is Lipschitz continuous (with respect to parallel transport);
- **(C2)** The sequences $\{x_k\}$ and $\{z_k\}$ are bounded [3, 10];
- **(C3)** In any compact subset of $\Omega(0)$, $\nabla F(w)$ is nonsingular.

Given the above assumptions, we can now prove the following statement.

Theorem 6.1 *(Global Convergence of RIPM) Let $\{w_k\}$ be generated by Algorithm 5 with $R = \text{Exp}$ and $\{\sigma_k\} \subset (0, 1)$ be bounded away from zero and one. Let φ be Lipschitz continuous on $\Omega(0)$. If (C1)–(C3) hold, then $\{\|F(w_k)\|\}$ converges to zero. Moreover, if w^* is a limit point of sequence $\{w_k\}$, then w^* satisfies Riemannian KKT conditions (4).*

The proof of the above theorem will be given in the next section. Note that although the exponential map is used in the proof, the numerical experiments indicate that global convergence may hold for a general retraction R.

7 Proof of Global Convergence

In this section, our goal is to prove the global convergence Theorem 6.1. We will follow
the proof procedure in [11], which discussed Algorithm 5 when $\mathcal{M} = \mathbb{R}^n$ in (RCOP),
and we call the algorithm in [11] the Euclidean Interior Point Method (EIPM). In what
follows, we will omit similar content because of space limitations and focus on the
difficulties encountered when adapting the proof of EIPM to RIPM. In particular, we
will make these difficulties as tractable as in EIPM by proving a series of propositions
in Sect. 7.1.

7.1 Continuity of Some Special Scalar Fields

To show the boundedness of the sequences generated by Algorithm 5, we need the
continuity of some special scalar fields on manifold $\mathcal{M}$. The claims of this subsection
are trivial if $\mathcal{M} = \mathbb{R}^n$, but they need to be treated carefully for general $\mathcal{M}$.

If we assign a linear operator $\mathcal{A}_x : T_x\mathcal{M} \to T_x\mathcal{M}$ to each $x \in \mathcal{M}$, then the map
$x \mapsto \|\mathcal{A}_x\| := \sup\{\|\mathcal{A}_x v\|_x \mid v \in T_x\mathcal{M}, \|v\|_x = 1, \text{ or, } \|v\|_x \leqslant 1\}$ is a scalar field
on $\mathcal{M}$; but notice that the operator norm $\|\cdot\|$ depends on x. Let $\mathrm{Sym}\,(d)$ denote the
set of symmetric matrices of order d, and $\|\cdot\|_F$, $\|\cdot\|_2$ denote the Frobenius norm and
the spectral norm, respectively, applied to a given matrix.

Lemma 7.1 *Let $\mathcal{M}$ be an n-dimensional Riemannian manifold. Let $x \in \mathcal{M}$ and $\mathcal{A}_x$
be a linear operator on $T_x\mathcal{M}$. Choose an orthonormal basis of $T_x\mathcal{M}$ with respect to
$\langle \cdot, \cdot \rangle_x$, and let $\hat{\mathcal{A}}_x \in \mathbb{R}^{n \times n}$ denote the matrix representation of $\mathcal{A}_x$ under the basis.
Then, $\|\hat{\mathcal{A}}_x\|_2$, $\|\hat{\mathcal{A}}_x\|_F$ are* invariant *under a change of orthonormal basis; moreover,
$\|\mathcal{A}_x\| = \|\hat{\mathcal{A}}_x\|_2 \leqslant \|\hat{\mathcal{A}}_x\|_F$.*

Proof Suppose that there are two orthonormal bases $\{E_i\}_{i=1}^n$, $\{E_i'\}_{i=1}^n$ on $T_x\mathcal{M}$. With
respect to them, let $P \in \mathbb{R}^{n \times n}$ denote the change-of-basis matrix, i.e., $[P]_{kj} :=
\langle E_j', E_k \rangle_x$, for $1 \leqslant k, j \leqslant n$; then, P is orthogonal. Let $\hat{\mathcal{A}}_x, \hat{\mathcal{A}}_x' \in \mathbb{R}^{n \times n}$ denote the
matrix representations of $\mathcal{A}_x$ under the two bases, respectively.

We have $\hat{\mathcal{A}}_x' = P^{-1}\hat{\mathcal{A}}_x P$. Then, $\|\hat{\mathcal{A}}_x'\| = \|P^{-1}\hat{\mathcal{A}}_x P\| = \|\hat{\mathcal{A}}_x\|$ holds for the
Frobenius norm or the spectral norm. Therefore, the values $\|\hat{\mathcal{A}}_x\|_2$ and $\|\hat{\mathcal{A}}_x\|_F$ are
invariant under a change of orthonormal basis. Now, consider an orthonormal basis
$\{E_i\}_{i=1}^n$ on $T_x\mathcal{M}$. For any $y \in T_x\mathcal{M}$, its vector representation $\hat{y} \in \mathbb{R}^n$ is defined by
$y = \sum_{i=1}^n \hat{y}_i E_i$. Accordingly, we have $\widehat{\mathcal{A}_x y} = \hat{\mathcal{A}}_x \hat{y}$, i.e., $\mathcal{A}_x y = \sum_{i=1}^n (\hat{\mathcal{A}}_x \hat{y})_i E_i$,
and from the orthonormal property of the basis, we have

$$\|\mathcal{A}_x y\|_x^2 = \left\|\sum_{i=1}^n (\hat{\mathcal{A}}_x \hat{y})_i E_i\right\|_x^2 = \sum_{i=1}^n (\hat{\mathcal{A}}_x \hat{y})_i^2 = \|\hat{\mathcal{A}}_x \hat{y}\|_2^2.$$

Thus, $\|\mathcal{A}_x y\|_x = \|\hat{\mathcal{A}}_x \hat{y}\|_2$ for any $y \in T_x\mathcal{M}$. Finally, we have

$$\|\mathcal{A}_x\| = \sup_{y \in T_x\mathcal{M}, \|y\|_x=1} \|\mathcal{A}_x y\|_x = \sup_{\hat{y} \in \mathbb{R}^n, \|\hat{y}\|_2=1} \|\hat{\mathcal{A}}_x \hat{y}\|_2 = \|\hat{\mathcal{A}}_x\|_2.$$

It is clear that $\|X\|_2 \leqslant \|X\|_F$ for any matrix X. $\qquad\qquad\qquad\qquad$ □

Yet, we have not clarified the continuity about $x \mapsto \|A_x\|$. The following proposition proves the continuity of an important case of A_x that appears in (RCOP).

Proposition 7.1 *Consider f in (RCOP). Let $\hat{\text{Hess}}\, f(x) \in \text{Sym}\,(d)$ denote the matrix representation of $\text{Hess}\, f(x)$ under an arbitrary orthonormal basis of $T_x\mathcal{M}$. Then, $x \mapsto \|\hat{\text{Hess}}\, f(x)\|$ is a continuous scalar field on $\mathcal{M}$ for $\|\cdot\|_F$ or $\|\cdot\|_2$. Moreover, $x \mapsto \|\text{Hess}\, f(x)\|$ is a continuous scalar field on $\mathcal{M}$.*

Proof Lemma 7.1 shows that $x \mapsto \|\hat{\text{Hess}}\, f(x)\|$ is well defined, so it suffices to prove its continuity. From Corollary 13.8 in [21], for each $\bar{x} \in \mathcal{M}$ there is a smooth, orthonormal local frame $\{E_i\}_{i=1}^d$ on a neighborhood $\mathcal{U}$ of $\bar{x}$; namely, $\{E_1(x), \ldots, E_d(x)\}$ forms an orthonormal basis on $T_x\mathcal{M}$ for all $x \in \mathcal{U}$. Choose such a local frame $\{E_i\}_{i=1}^d$ around $\bar{x}$; then, the matrix representation of $\text{Hess}\, f(x)$ is given by $[\hat{\text{Hess}}\, f(x)]_{kj} := \langle \text{Hess}\, f(x)[E_j(x)], E_k(x) \rangle_x = \langle (\nabla_{E_j}\, \text{grad}\, f)(x), E_k(x) \rangle_x$ for $1 \leqslant k, j \leqslant d$. From the smoothness of the Riemannian metric (1), it follows that $x \mapsto \hat{\text{Hess}}\, f(x)$ is a continuous function from $\mathcal{U} \subset \mathcal{M}$ to $\text{Sym}\,(d)$. Since matrix norms are continuous, $\|\hat{\text{Hess}}\, f(x)\|$ is continuous on $\mathcal{U} \ni \bar{x}$. This argument holds for any $\bar{x} \in \mathcal{M}$. From Lemma 7.1, $\|\hat{\text{Hess}}\, f(x)\|_2 = \|\text{Hess}\, f(x)\|$ for any $x \in \mathcal{M}$, which completes the proof. $\qquad\qquad\qquad\qquad$ □

The above result can be applied verbatim to the Hessian of constraint functions $\{h_i\}_{i=1}^l, \{g_i\}_{i=1}^m$ in (RCOP). The next proposition can be proved similarly, as in Lemma 7.1 and Proposition 7.1.

Proposition 7.2 *Consider h, g in (RCOP) and the linear operators $\mathcal{H}_x$, $\mathcal{G}_x$ defined in (7). Then, $x \mapsto \|\mathcal{H}_x\|$ and $x \mapsto \|\mathcal{G}_x\|$ are continuous scalar fields on $\mathcal{M}$.*

Proposition 7.3 *Given $w = (x, y, z, s) \in \mathcal{N}$, consider the operator $\nabla F(w)$ in (8). Let $\{E_i\}_{i=1}^d$ be an orthonormal basis of $T_x\mathcal{M}$ and $\{e_i\}_{i=1}^l, \{\dot{e}_i\}_{i=1}^m$ be the standard bases of $\mathbb{R}^l, \mathbb{R}^m$, respectively. If we choose an orthonormal basis of $T_w\mathcal{N}$ as follows:*

$$\{(E_i, 0, 0, 0)\}_{i=1}^d \cup \{(0_x, e_i, 0, 0)\}_{i=1}^l \cup \{(0_x, 0, \dot{e}_i, 0)\}_{i=1}^m \cup \{(0_x, 0, 0, \dot{e}_i)\}_{i=1}^m, \quad (35)$$

then, the matrix representation of $\nabla F(w)$ is given by

$$\hat{\nabla} F(w) = \begin{pmatrix} Q & B & C & 0 \\ B^T & 0 & 0 & 0 \\ C^T & 0 & 0 & I \\ 0 & 0 & S & Z \end{pmatrix},$$

i.e., a matrix of order $(d + l + 2m)$ and, where
    $Q := Q(w) \in \text{Sym}\,(d)$ *is given by* $[Q]_{kj} := \langle \text{Hess}_x\, \mathcal{L}(w)[E_j], E_k \rangle_x$ *for* $1 \leqslant k, j \leqslant d$;
    $B := B(x) = [\hat{\text{grad}}\, h_1(x), \cdots, \hat{\text{grad}}\, h_l(x)] \in \mathbb{R}^{d \times l}$;
    $C := C(x) = [\hat{\text{grad}}\, g_1(x), \cdots, \hat{\text{grad}}\, g_m(x)] \in \mathbb{R}^{d \times m}$ *and the "hat" symbol above means the corresponding vector representation under the basis* $\{E_i\}_{i=1}^d$.

In this case, there is a continuous scalar field $T : \mathcal{N} \to \mathbb{R}$ such that for any w, $\|Q(w)\|_{\mathrm{F}} \leqslant T(w)$. Moreover, $x \mapsto \|B(x)\|_{\mathrm{F}}$ and $x \mapsto \|C(x)\|_{\mathrm{F}}$ are continuous scalar fields on $\mathcal{M}$.

Proof The matrix $\hat{\nabla} F(w)$ under the basis (35) is obtained through a trivial process, so we will omit its description. From relation (9), we have $Q(w) := \hat{\mathrm{Hess}}_x \mathcal{L}(w) = \hat{\mathrm{Hess}} f(x) + \sum_{i=1}^{l} y_i \hat{\mathrm{Hess}} h_i(x) + \sum_{i=1}^{m} z_i \hat{\mathrm{Hess}} g_i(x)$, under the same basis. Thus,

$$\|Q(w)\|_{\mathrm{F}} \leqslant \|\hat{\mathrm{Hess}} f(x)\|_{\mathrm{F}} + \sum_{i=1}^{l} |y_i| \|\hat{\mathrm{Hess}} h_i(x)\|_{\mathrm{F}} + \sum_{i=1}^{m} |z_i| \|\hat{\mathrm{Hess}} g_i(x)\|_{\mathrm{F}} =: T(w).$$

From Proposition 7.1, $\|\hat{\mathrm{Hess}} f(x)\|_{\mathrm{F}}$, $\|\hat{\mathrm{Hess}} h_i(x)\|_{\mathrm{F}}$, and $\|\hat{\mathrm{Hess}} g_i(x)\|_{\mathrm{F}}$ are all continuous with respect to x, thus, T is continuous. As for $x \mapsto \|B(x)\|_{\mathrm{F}}$, since the basis $\{E_i\}_{i=1}^{d}$ is orthonormal,

$$\|B(x)\|_{\mathrm{F}}^2 = \sum_{i=1}^{l} \|\hat{\mathrm{grad}} \, h_i(x)\|_2^2 = \sum_{i=1}^{l} \|\mathrm{grad} \, h_i(x)\|_x^2 ,$$

which implies continuity by (1). The claim for $x \mapsto \|C(x)\|_{\mathrm{F}}$ can be proven similarly. $\square$

7.2 Global Convergence Theorem

Now, we are ready to prove the global convergence Theorem 6.1 by following the procedure in [11]. In what follows, we will omit similar content in [11] and focus on the difficulties encountered when adapting the proof of EIPM to RIPM.

Proposition 7.4 *(Boundedness of the sequences) Let $\{w_k\}$ be a sequence generated by Algorithm 5 and suppose that **(C1)**– **(C3)** hold. If $\epsilon > 0$ and $w_k \in \Omega(\epsilon)$ for all k, then*

(a) $\{z_k^T s_k\}$, $\{(z_k)_i (s_k)_i\}$, $i = 1, \ldots, m$, *are all bounded above and below away from zero;*
(b) $\{z_k\}$ *and $\{s_k\}$ are bounded above and component-wise bounded away from zero;*
(c) $\{w_k\}$ *is bounded;*
(d) $\{\|\nabla F(w_k)^{-1}\|\}$ *is bounded;*
(e) $\{\Delta w_k\}$ *is bounded.*

Proof The proofs in [11, Lemma 6.1] and/or [3, Theorem 2 (a)] can be applied verbatim to **(a)**, **(b)** and **(e)**.

(c) By (b), it suffices to prove that $\{y_k\}$ is bounded. The iteration count k is omitted in what follows. By using the notation $\mathcal{H}_x$ and $\mathcal{G}_x$ as defined in (7), we have $\mathcal{H}_x y = \mathrm{grad}_x \mathcal{L}(w) - \mathrm{grad} \, f(x) - \mathcal{G}_x z =: b$. By **(C1)**, $\mathcal{H}_x$ is an injection; then, there exists a unique solution y to $\mathcal{H}_x y = b$. Indeed, we have

$$y = \left[\left(\mathcal{H}_x^* \mathcal{H}_x \right)^{-1} \mathcal{H}_x^* \right] \left(\mathrm{grad}_x \mathcal{L}(w) - \mathrm{grad} \, f(x) - \mathcal{G}_x z \right). \tag{36}$$

Define $\mathcal{C}_x : T_x\mathcal{M} \to \mathbb{R}^l$ as $\mathcal{C}_x := \left(\mathcal{H}_x^*\mathcal{H}_x\right)^{-1}\mathcal{H}_x^*$. Under an arbitrary orthonormal basis of $T_x\mathcal{M}$ and standard basis of $\mathbb{R}^l$, if $\hat{\mathcal{H}}_x$ is the matrix corresponding to $\mathcal{H}_x$, then $\hat{\mathcal{C}}_x = (\hat{\mathcal{H}}_x^T\hat{\mathcal{H}}_x)^{-1}\hat{\mathcal{H}}_x^T$. It is easy to show that $\|\mathcal{C}_x\| = \|\hat{\mathcal{C}}_x\|_2$ for any x. We can see that for each $\bar{x} \in \mathcal{M}$ there is a neighborhood $\mathcal{U}$ of $\bar{x}$ such that $x \mapsto \hat{\mathcal{H}}_x$ is continuous over $\mathcal{U}$. Then, by function composition, $x \mapsto \hat{\mathcal{C}}_x$ is also continuous over $\mathcal{U}$. This implies that $x \mapsto \|\mathcal{C}_x\| = \|\hat{\mathcal{C}}_x\|_2$ is continuous at each $\bar{x}$, and hence, on $\mathcal{M}$. Finally, by Proposition 7.2, $\|\mathcal{C}_x\|$, $\|\text{grad } f(x)\|$, and $\|\mathcal{G}_x\|$ are all continuous on $\mathcal{M}$. Because $\{x_k\}$ is bounded, by (36) we have $\|y_k\| \leqslant \|\mathcal{C}_{x_k}\| \left(\|\text{grad}_x \mathcal{L}(w_k)\| + \|\text{grad } f(x_k)\| + \|\mathcal{G}_{x_k}\| \|z_k\|\right) \leqslant c_1\left(\sqrt{\varphi_0} + c_2 + c_3\|z_k\|\right)$, for some positive constants c_1, c_2, c_3. Then, $\{y_k\}$ is bounded because $\{z_k\}$ is bounded.

(d) For each w_k, choose an arbitrary orthonormal basis of $T_{w_k}\mathcal{N}$. If the matrix representation $\hat{\nabla}F(w_k)$ corresponds to $\nabla F(w_k)$, then $[\hat{\nabla}F(w_k)]^{-1}$ corresponds to $\nabla F(w_k)^{-1}$. By Lemma 7.1, we have $\|\nabla F(w_k)^{-1}\| \leqslant \|[\hat{\nabla}F(w_k)]^{-1}\|_F$; thus, it is sufficient to show that $\{\|[\hat{\nabla}F(w_k)]^{-1}\|_F\}$ is bounded. For convenience, we will choose the basis of $T_{w_k}\mathcal{N}$ given in (35). Then, we have

$$\hat{\nabla}F(w_k) = \begin{pmatrix} Q_k & B_k & C_k & 0 \\ B_k^T & 0 & 0 & 0 \\ C_k^T & 0 & 0 & I \\ 0 & 0 & S_k & Z_k \end{pmatrix}.$$

By Proposition 7.3, there is a continuous scalar field $T : \mathcal{N} \to \mathbb{R}$ such that $\|Q(w)\|_F \leqslant T(w)$ for all $w \in \mathcal{N}$; and $\|B(x)\|_F$, $\|C(x)\|_F$ are continuous on $\mathcal{M}$. It follows from the boundedness of $\{x_k\}$ and $\{w_k\}$ that for all k, $\|Q_k\|_F \equiv \|Q(w_k)\|_F \leqslant T(w_k) \leqslant c_4$, $\|B_k\|_F \equiv \|B(x_k)\|_F \leqslant c_5$, and $\|C_k\|_F \equiv \|C(x_k)\|_F \leqslant c_6$, for some positive constants c_4, c_5, and c_6.

On the other hand, whichever basis is used in the form of (35), the structure of $\hat{\nabla}F(w_k)$ and the properties of its block submatrices remain unchanged, e.g., symmetry of Q_k; full rank of B_k; identity matrix I in the third row; all zero matrices; diagonal matrices S_k, Z_k; etc. This ensures that we can obtain the desired result by performing an appropriate decomposition of $\hat{\nabla}F(w_k)$. Up to this point, we have created all the conditions needed in the proof of the Euclidean version, namely EIPM. We can make the claim that $\{\|[\hat{\nabla}F(w_k)]^{-1}\|_F\}$ is bounded by applying the proofs in [11, Lemma 6.2] and/or [3, Theorem 2 (c)] directly. $\qquad\square$

Lemma 7.2 ($\{\bar{\alpha}_k\}$ bounded away from zero) *Let $\{w_k\}$ be generated by Algorithm 5 with $\text{R} = \text{Exp}$ and let **(C1)–(C3)** hold. If $\epsilon > 0$ and $w_k \in \Omega(\epsilon)$ for all k, $\{\sigma_k\}$ is bounded away from zero; then, $\{\bar{\alpha}_k\}$ is bounded away from zero.*

Proof Since $\bar{\alpha}_k = \min\{\alpha_k^I, \alpha_k^{II}\}$, it is sufficient to show that $\{\alpha_k^I\}$ and $\{\alpha_k^{II}\}$ are bounded away from zero. For α_k^I, see [11, Lemma 6.3] and/or [9, Theorem 3.1]. The proofs in those references apply verbatim to the Riemannian case. On the other hand, for α_k^{II}, we need to adapt the proofs in [9, 11], since Lipschitz continuity on manifolds is more complicated, see Subsection 4.2.

Let us suppress the subscript k. Recall that $w(\alpha) = \text{Exp}_w(\alpha\Delta w)$. Fix $\alpha\Delta w$ and let P_γ be the parallel transport along the geodesic $\gamma(t) = \text{Exp}_w(t\alpha\Delta w)$. By Lemma 4.4

where $c_1 = 0$, we obtain

$$
\begin{aligned}
&\mathrm{P}_\gamma^{0\to 1}[F(w(\alpha))]\\
&= F(w) + \alpha \nabla F(w)[\Delta w] - \alpha \nabla F(w)[\Delta w] + \int_0^1 \mathrm{P}_\gamma^{0\to t} \nabla F(\gamma(t)) \mathrm{P}_\gamma^{t\to 0}[\alpha \Delta w]\mathrm{d}t\\
&= F(w) + \alpha \left(\sigma\rho\hat{e} - F(w)\right) + \alpha \int_0^1 \left(\mathrm{P}_\gamma^{0\to t} \nabla F(\gamma(t)) \mathrm{P}_\gamma^{t\to 0} - \nabla F(w)\right)[\Delta w]\mathrm{d}t\\
&= (1-\alpha)F(w) + \alpha\sigma\rho\hat{e} + \alpha \int_0^1 \left(\mathrm{P}_\gamma^{0\to t} \nabla F(\gamma(t)) \mathrm{P}_\gamma^{t\to 0} - \nabla F(w)\right)[\Delta w]\mathrm{d}t.
\end{aligned}
$$

Taking the norm on both sides above gives

$$
\begin{aligned}
&\|F(w(\alpha))\|\\
&= \left\|\mathrm{P}_\gamma^{0\to 1}[F(w(\alpha))]\right\| \text{ (since parallel transport (23) is isometric)}\\
&\leqslant (1-\alpha)\|F(w)\| + \alpha\sigma\rho\|\hat{e}\| + \alpha \int_0^1 \left\|\mathrm{P}_\gamma^{0\to t} \nabla F(\gamma(t)) \mathrm{P}_\gamma^{t\to 0} - \nabla F(w)\right\| \|\Delta w\|\,\mathrm{d}t\\
&\leqslant (1-\alpha)\|F(w)\| + \alpha\sigma\rho\sqrt{m} + \alpha \int_0^1 \kappa_2 \|t\alpha\Delta w\| \|\Delta w\|\,\mathrm{d}t\\
&= (1-\alpha)\|F(w)\| + \alpha\sigma\rho\sqrt{m} + \alpha^2 \|\Delta w\|^2 \kappa/2.
\end{aligned}
$$

The rest of the proof is the same as [11, Lemma 6.3] and/or [9, Theorem 3.1], so we omit it. □

Proof of Theorem 6.1 By Proposition 6.1, we know that $\{\|F(w_k)\|\}$ is monotonically nonincreasing, hence convergent. Assume that $\{\|F(w_k)\|\}$ does not converge to zero. Then, there exists $\epsilon > 0$ such that $\{w_k\} \subset \Omega(\epsilon)$ for infinitely many k. We will show that the following two cases both lead to contradictions, and thus, the hypothesis $\|F(w_k)\| \nrightarrow 0$ is not valid.

Case 1. For infinitely many k, if step (3b) in Algorithm 5 is executed with $\alpha_k \equiv \bar{\alpha}_k$, it follows from Proposition 6.1 that $\varphi(w_{k+1})/\varphi(w_k) \leqslant \lambda_k := [1 - 2\bar{\alpha}_k\beta(1-\sigma_k)]$. Since $\{\bar{\alpha}_k\}$ is bounded away from zero by Lemma 7.2 and $\{\sigma_k\}$ is bounded away from one, then $\{\lambda_k\}$ is bounded away from one, and hence, $\varphi(w_k) \to 0$; this is a contradiction.

Case 2. On the other hand, for infinitely many k, if $\alpha_k < \bar{\alpha}_k$, we have that $\alpha_k \leqslant \theta\bar{\alpha}_k$. Then, condition (33) fails to hold for an $\tilde{\alpha}_k$ with $\alpha_k < \tilde{\alpha}_k \leqslant \alpha_k/\theta = \theta^{t-1}\bar{\alpha}_k$. Notice that α_k/θ is the value corresponding to the last failure. Recall that the derivative of the real-valued function $\alpha \mapsto \varphi(\alpha) := \varphi\left(\bar{\mathrm{Exp}}_{w_k}(\alpha\Delta w_k)\right)$ is

$$
\varphi'(\alpha) = \mathcal{D}\varphi(\bar{\mathrm{Exp}}_{w_k}(\alpha\Delta w_k))[\mathcal{D}\bar{\mathrm{Exp}}_{w_k}(\alpha\Delta w_k)[\Delta w_k]]. \tag{37}
$$

Applying the mean value theorem to $\varphi(\alpha)$ on the interval $[0, \tilde{\alpha}_k]$ yields a number $\xi \in (0, 1)$ such that $\tilde{\alpha}_k \varphi'(\xi \tilde{\alpha}_k) = \varphi(\tilde{\alpha}_k) - \varphi(0)$. For short, let $u := \xi \tilde{\alpha}_k \Delta w_k$. Hence,

$$
\begin{aligned}
&\tilde{\alpha}_k \beta \langle \mathrm{grad}\ \varphi_k, \Delta w_k \rangle \\
&< \varphi(\tilde{\alpha}_k) - \varphi(0) \ \text{(as condition (33) fails for } \tilde{\alpha}_k) \\
&= \tilde{\alpha}_k \varphi'(\xi \tilde{\alpha}_k) \\
&= \tilde{\alpha}_k \mathcal{D}\varphi(\bar{\mathrm{Exp}}_{w_k}(u)) \left[\mathcal{D}\bar{\mathrm{Exp}}_{w_k}(u) \left[\Delta w_k \right] \right] \ \text{(by equation (37))} \\
&= \tilde{\alpha}_k \langle \mathrm{grad}\ \varphi(\bar{\mathrm{Exp}}_{w_k}(u)), \mathcal{D}\bar{\mathrm{Exp}}_{w_k}(u) \left[\Delta w_k \right] \rangle.
\end{aligned}
\tag{38}
$$

On the other hand, note that

$$
\begin{aligned}
&\langle \mathrm{grad}\ \varphi_k, \Delta w_k \rangle \\
&= \langle \mathrm{grad}\ \varphi_k, u \rangle / \xi \tilde{\alpha}_k \\
&= \langle \mathcal{D}\bar{\mathrm{Exp}}_{w_k}(u) \left[\mathrm{grad}\ \varphi_k \right], \mathcal{D}\bar{\mathrm{Exp}}_{w_k}(u) \left[u \right] \rangle / \xi \tilde{\alpha}_k \ \text{(by Lemma 4.8)} \\
&= \langle \mathcal{D}\bar{\mathrm{Exp}}_{w_k}(u) \left[\mathrm{grad}\ \varphi_k \right], \mathcal{D}\bar{\mathrm{Exp}}_{w_k}(u) \left[\Delta w_k \right] \rangle.
\end{aligned}
\tag{39}
$$

Subtracting $\tilde{\alpha}_k \langle \mathrm{grad}\ \varphi_k, \Delta w_k \rangle$ from both sides of (38) and using equalities (39) gives

$$
\begin{aligned}
&\tilde{\alpha}_k (\beta - 1) \langle \mathrm{grad}\ \varphi_k, \Delta w_k \rangle \\
&< \tilde{\alpha}_k \left[\langle \mathrm{grad}\ \varphi(\bar{\mathrm{Exp}}_{w_k}(u)), \mathcal{D}\bar{\mathrm{Exp}}_{w_k}(u) \left[\Delta w_k \right] \rangle - \langle \mathrm{grad}\ \varphi_k, \Delta w_k \rangle \right] \\
&= \tilde{\alpha}_k \langle \mathrm{grad}\ \varphi(\bar{\mathrm{Exp}}_{w_k}(u)) - \mathcal{D}\bar{\mathrm{Exp}}_{w_k}(u) \left[\mathrm{grad}\ \varphi_k \right], \mathcal{D}\bar{\mathrm{Exp}}_{w_k}(u) \left[\Delta w_k \right] \rangle \\
&\leqslant \tilde{\alpha}_k \left\| \mathrm{grad}\ \varphi(\bar{\mathrm{Exp}}_{w_k}(u)) - \mathcal{D}\bar{\mathrm{Exp}}_{w_k}(u) \left[\mathrm{grad}\ \varphi_k \right] \right\| \left\| \mathcal{D}\bar{\mathrm{Exp}}_{w_k}(u) \left[\Delta w_k \right] \right\| \\
&= \tilde{\alpha}_k \left\| \mathrm{grad}\ \varphi(y) - \mathcal{D}\bar{\mathrm{Exp}}_{w_k}(u) \left[\mathrm{grad}\ \varphi(w_k) \right] \right\| \left\| \mathcal{D}\bar{\mathrm{Exp}}_{w_k}(u) \left[\Delta w_k \right] \right\| \\
&\quad \text{(by letting } y := \bar{\mathrm{Exp}}_{w_k}(u)) \\
&\leqslant \tilde{\alpha}_k \kappa \left\| u \right\| \left\| \Delta w_k \right\| \ \text{(by (22), Definition 4.1 and Lemma 4.8)} \\
&= \kappa \xi \tilde{\alpha}_k^2 \left\| \Delta w_k \right\|^2.
\end{aligned}
$$

Finally, we obtain $(\beta - 1) \langle \mathrm{grad}\ \varphi_k, \Delta w_k \rangle / (\kappa \xi \left\| \Delta w_k \right\|^2) < \tilde{\alpha}_k$.

Because $\langle \mathrm{grad}\ \varphi_k, \Delta w_k \rangle < 0$ and α_k satisfies (33), we have

$$
\begin{aligned}
\varphi_k(0) - \varphi_k(\alpha_k) &\geqslant -\alpha_k \beta \langle \mathrm{grad}\ \varphi_k, \Delta w_k \rangle \\
&\geqslant -\theta \beta \tilde{\alpha}_k \langle \mathrm{grad}\ \varphi_k, \Delta w_k \rangle \\
&\geqslant -\theta \beta \langle \mathrm{grad}\ \varphi_k, \Delta w_k \rangle (\beta - 1) \langle \mathrm{grad}\ \varphi_k, \Delta w_k \rangle / (\kappa \xi \left\| \Delta w_k \right\|^2) \\
&\geqslant [\theta \beta (1 - \beta) / \kappa \xi] (\langle \mathrm{grad}\ \varphi_k, \Delta w_k \rangle / \left\| \Delta w_k \right\|)^2 \\
&= \omega (\langle \mathrm{grad}\ \varphi_k, \Delta w_k \rangle / \left\| \Delta w_k \right\|),
\end{aligned}
$$

where $\omega(\cdot)$ is an F-function (see [28, Definition 14.2.1 & 14.2.2 in P479]). Since $\{\varphi_k\}$ is bounded below and $\varphi_k \geqslant \varphi_{k+1}$, it follows that $\lim_{k \to \infty}(\varphi_k - \varphi_{k+1}) = 0$. By the definition of F-functions, we obtain $\langle \mathrm{grad}\ \varphi_k, \Delta w_k \rangle / \left\| \Delta w_k \right\| \to 0$. Since

$\{\|\Delta w_k\|\}$ is bounded (Proposition 7.4), we have $\langle \text{grad } \varphi_k, \Delta w_k \rangle \to 0$. Choosing ρ_k with $z_k^T s_k / m \leqslant \rho_k \leqslant \|F(w_k)\| / \sqrt{m}$ in Algorithm 5 implies that

$$\langle \text{grad } \varphi_k, \Delta w_k \rangle / (-2) = \varphi_k - \sigma_k \rho_k z_k^T s_k \geqslant \varphi_k - \sigma_k \|F(w_k)\| z_k^T s_k / \sqrt{m}$$
$$\geqslant \varphi_k - \sigma_k \|F(w_k)\|^2 \geqslant (1 - \sigma_k)\varphi_k.$$

This shows that $\varphi(w_k) \to 0$, because $\{\sigma_k\}$ is bounded away from one; this is a contradiction. $\qquad\square$

8 Numerical Experiments

The numerical experiments compared the performance of the globally convergent RIPM (Algorithm 5) with those of other Riemannian methods in solving two problems. They were conducted in Matlab R2022a on a computer with an Intel Core i7-10700 (2.90GHz) and 16GB RAM. Algorithm 5 utilized Manopt 7.0 [6], a Riemannian optimization toolbox for MATLAB. The problems involve three manifolds:

- fixed-rank manifold, $\mathcal{M}_r = \{X \in \mathbb{R}^{m \times n} : \text{rank}(X) = r\}$;
- Stiefel manifold, $\text{St}(n, k) = \{X \in \mathbb{R}^{n \times k} : X^T X = I_k\}$;
- oblique manifold, $\text{Ob}(n, k) = \{X \in \mathbb{R}^{n \times k} : (X^T X)_{ii} = 1, \forall i = 1, \ldots, k\}$.

We only consider their embedded geometry and we apply the default retractions in Manopt, e.g., the retraction based on QR decomposition for the Stiefel manifold.

Problem I. Recently, [31] proposed the nonnegative low-rank matrix (NLRM) approximation. Formally, NLRM aims to solve

$$\min_{X \in \mathcal{M}_r} \|A - X\|_F^2 \quad \text{s.t. } X \geqslant 0. \tag{NLRM}$$

Input. We tested three cases of integer triples (m, n, r): $(20, 16, 2)$, $(30, 24, 3)$, and $(40, 32, 4)$. For each (m, n, r), we generated nonnegative $L \in \mathbb{R}^{m \times r}$, $R \in \mathbb{R}^{r \times n}$ whose entries follow a uniform distribution in [0,1]. Then, we added the Gaussian noise with zero mean and different standard deviation ($\sigma = 0, 0.001, 0.01$) to $A = LR$. When there is no noise (i.e., $\sigma = 0$), the input data matrix A itself is exactly a solution.

Problem II. Given $C \in \mathbb{R}^{n \times k}$, [19] computed its projection onto the nonnegative part of the Stiefel manifold. If the distance is measured in terms of $\|C - X\|_F^2$, we can express it equivalently as

$$\min_{X \in \text{St}(n,k)} -2 \operatorname{trace}(X^T C) \quad \text{s.t. } X \geqslant 0. \tag{Model\_St}$$

In [19], it is shown that (Model_St) can be equivalently reformulated into

$$\min_{X \in \text{Ob}(n,k)} -2 \operatorname{trace}(X^T C) \quad \text{s.t. } X \geqslant 0, \ \|XV\|_F = 1, \tag{Model\_Ob}$$

where the positive integer p and $V \in \mathbb{R}^{k \times p}$ can be arbitrary as long as $\|V\|_F = 1$ and VV^T is entry-wise positive. We examined both models. *Input.*

We considered four cases of integer pairs (n, k): $(40, 8)$, $(50, 10)$, $(60, 16)$, and $(70, 14)$. For a general C, it is always difficult to seek nonnegative projections globally. Fortunately, Proposition 1 in [19] showed a way to construct C such that Model_St has a unique, known solution X^*. In our experiments, we generated a feasible point B of Model_St; then, we obtained C by using codes: `X1=(B>0).*(1+rand(n,k));Xstar=X1./sqrt(sum(X1.*X1));L` `=rand(k,k);L=L+k*eye(k);C=Xstar*L'`. The initial point was computed by projecting C onto the Stiefel manifold. In addition, for Model_Ob we set `p=1;V=ones(k,p);V=V/norm(V,"fro")`.

8.1 Implementation Details

The experimental implementation of Algorithm 5 (i.e.,RIPM) initialized z_0 and s_0 from a uniform distribution in [0,1] and set $y_0 = 0$ if equality constraints exist. We used $\rho_k = z_k^T s_k / m$, $\sigma_k = \min\{0.5, \|F(w_k)\|^{1/2}\}$ and Algorithm 3 to solve the condensed form of Newton equation (19). Algorithm 3 stopped when the relative residual went below 10^{-9}, or it reached 1000 iterations. We used a backtracking line search simultaneously for the central conditions and sufficient decreasing conditions. We set $\gamma_{-1} = 0.9$, $\gamma_{k+1} = (\gamma_k + 0.5)/2$; and $\beta = 10^{-4}$, $\theta = 0.5$. We compared RIPM with the following Riemannian methods:

– RALM: Riemannian augmented Lagrangian method [23].
– REPM$_{lqh}$: Riemannian exact penalty method with smoothing functions of linear-quadratic and pseudo-Huber [23].
– REPM$_{lse}$: Riemannian exact penalty method with smoothing functions of log-sum-exp [23].
– RSQP: Riemannian sequential quadratic programming [27].

Let $[t]_+ := \max(0, t)$ and $[t]_- := \min(0, t)$ for any $t \in \mathbb{R}$. The experimental settings followed those of [27], where they used the KKT residual defined as

$$\sqrt{\left\|\operatorname{grad}_x \mathcal{L}(w)\right\|^2 + \sum_{i=1}^{m}\{[z_i]_-^2 + [g_i(x)]_+^2 + |z_i g_i(x)|^2\} + \sum_{i=1}^{l} |h_i(x)|^2}$$

to measure the deviation of an iterate from the KKT conditions. For the parameters of RALM, REPMs, and RSQP, we utilized the experimental setting and Matlab codes provided by [27].

  We conducted 20 random trials of each problem and model. All the algorithms ran with the same initial point. The experiment is considered successfully terminated if it finds a solution with a KKT residual lower than ϵ_{kkt} before triggering any of the stopping conditions. For the *first-order* algorithms (including RALM and the REPMs), the stopping conditions are: elapsed time exceeding t_{max} seconds, number of outer iterations exceeding 1,000, or failure of the algorithm to update any parameters. For the *second-order* algorithms (including RSQP and RIPM), the stopping conditions are elapsed time exceeding t_{max} seconds or a number of outer iterations exceeding 10,000. Here, considering that some problems might not have converged easily, the maximum

number of iterations was chosen to be 1,000 (10,000), which was a sufficiently large value. The selection of $t_{\max}$ is related to the actual time it took to run all the codes on the computer. Setting $t_{\max}$ too large resulted in excessive time spent on poorly performing algorithms. On the other hand, ϵ_{kkt} was chosen to better demonstrate that second-order algorithms could achieve more accurate solutions. Therefore, we chose the appropriate values for $t_{\max}$ and ϵ_{kkt} according to the problem that was to be solved.

8.2 Results and Analysis

The tables in this subsection report the success rate (**Success**), the average time (**Time**), and the average iteration number (**Iter.**) among the successful trials. In order to capture the combination of stability and speed, the algorithms with 0.9 or higher are first highlighted in bold in the "Success" column, and then, of those algorithms, the fastest result is highlighted in bold in the "Time" column. Here, "successful convergence in numerical experiments" means that the algorithm can generate a relatively accurate solution in a relatively reasonable amount of time. It is not exactly the same as the "theoretical convergence in any global convergence theorem". Numerical experiments reflect the actual performance of the algorithm in the application. For example, the first-order algorithms (RALM, REPMs) also have theoretical global convergence, but under our high criterion, it is difficult for them to generate a high-precision solution within a certain period of time. The second-order algorithm (RIPM, RSQP), on the other hand, is excellent in terms of accuracy, although it takes quite a bit of time.

Problem I. Here, we set $t_{\max} = 180$, $\epsilon_{\mathrm{kkt}} = 10^{-8}$. The numerical results in Table 1 show that RIPM had the best performance, except for cases (30, 40, 3) and (40, 32, 4) without noise. The time spent by RALM and the REPMs grew slowly with the problem size m and n, but their success rates dropped sharply as the noise level (standard deviation σ) intensified, eventually leading to non-convergence. In contrast, RSQP and RIPM were more stable and robust, while RIPM was much faster than RSQP. The cost of RSQP increased drastically with the problem size because a *quadratic programming subproblem* (defined over tangent space of current point) had to be solved in each iteration. Unlike RIPM, which uses Krylov subspace methods introduced in Sect. 3.4.2 to avoid expensive computations, RSQP had to transform the subproblem into a matrix representation (similar to Steps 1–6 in Sect. 3.4.2). As can be seen from Table 1, RIPM took about the same amount of time as RALM and the REPMs did.

Problem II. Here, we set $t_{\max} = 600$, $\epsilon_{\mathrm{kkt}} = 10^{-6}$ for both (Model_St) and (Model_Ob). Since the true solution X^* is known, we added a column showing the average error $\|\tilde{X} - X^*\|_{\mathrm{F}}$, where $\tilde{X}$ denotes the final iterate. The numerical results are listed in Tables 2 and 3. The error columns show that if the KKT residual is sufficiently small, then $\tilde{X}$ does approximate the true solution. In particular, RSQP and RIPM yield a more accurate solution. (The error is less than 10^{-7}.) From Table 2, we can see that RALM is stable and fast for (Model_St). However, from Table 3, the success rate of RALM for (Model_Ob) decreases as the problem size becomes larger. The REPMs do not work at all on either model. RSQP also does not perform well on either model. In contrast, RIPM successfully solved all instances of both models and was only slightly

Table 1 Performance of various Riemannian methods on problem I ([NLRM](#))

(m, n, r)	(20,16,2)			(30,24,3)			(40,32,4)		
No noise	Success	Time (s)	Iter	Success	Time (s)	Iter	Success	Time (s)	Iter
RALM	0.4	1.115	31	0.65	1.813	31	0.75	2.800	31
REPM$_\text{lqh}$	**1**	5.165×10^{-1}	31	**1**	**1.009**	31	**1**	**1.747**	31
REPM$_\text{lse}$	**1**	2.242	31	**1**	4.041	31	**0.95**	6.952	31
RSQP	**0.9**	6.429	7	**0.9**	3.944×10	8	**0.9**	1.254×10^2	8
RIPM	**1**	$\mathbf{4.920\times10^{-1}}$	19	**1**	2.247	27	**1**	5.277	32

(m, n, r)	(20,16,2)			(30,24,3)			(40,32,4)		
$\sigma = 0.001$	Success	Time (s)	Iter	Success	Time (s)	Iter	Success	Time (s)	Iter
RALM	0.2	1.001	31	0.15	2.050	31	0.05	2.758	31
REPM$_\text{lqh}$	0.1	4.983×10^{-1}	32	0.25	1.035	31	0.15	1.787	31
REPM$_\text{lse}$	0.15	2.444	31	0.1	4.867	31	0.05	8.371	31
RSQP	**0.95**	6.619	7	**0.95**	3.848×10	8	**0.9**	1.299×10^2	8
RIPM	**1**	$\mathbf{5.376\times10^{-1}}$	20	**1**	**2.342**	27	**1**	**4.631**	29

(m, n, r)	(20,16,2)			(30,24,3)			(40,32,4)		
$\sigma = 0.01$	Success	Time (s)	Iter	Success	Time (s)	Iter	Success	Time (s)	Iter
RALM	0	–	–	0	–	–	0	–	–
REPM$_\text{lqh}$	0	–	–	0	–	–	0	–	–
REPM$_\text{lse}$	0	–	–	0	–	–	0	–	–
RSQP	**1**	7.295	8	**0.95**	4.114×10	8	**0.95**	1.430×10^2	9
RIPM	**1**	$\mathbf{5.980\times10^{-1}}$	21	**0.95**	**1.883**	25	**0.95**	**4.602**	29

Algorithms with values greater than or equal to 0.9 in the "Success" column are bolded. Among these, the one with the fastest result in the "Time" column is also highlighted in bold

slower than RALM in some cases. Overall, our RIPM was relatively fast and most stable.

9 Conclusions

In this paper, we proposed a Riemannian version of the classical interior point method and established its local and global convergence. To our knowledge, this is the first study to apply the primal-dual interior point method to the nonconvex constrained optimization problem on a Riemannian manifold. Numerical experiments showed the stability and efficiency of our method.

Recently, Hirai et al. [15] extended the self-concordance-based interior point methods to Riemannian manifolds. They aimed to minimize a *geodesically convex* (i.e., convex on manifolds) objective $f : D \to \mathbb{R}$ defined on a geodesically convex subset $D \subset \mathcal{M}$. In contrast, in (RCOP), we do not require any convexity. In practice, many convex functions (in the Euclidean sense) are not geodesically convex on some interested manifolds. For example, for any geodesically convex function defined on a connected, compact Riemannian manifold (e.g., Stiefel manifold), it must be constant [5, Corollary 11.10], which is not of interest in the field of optimization. Thus, (RCOP) has a wider applicability.

In closing, let us make a comparison with the Euclidean interior point method (EIPM) to illustrate the theoretical advantages of our RIPM and discuss two future directions of research on more advanced RIPM methods.
Comparison: Riemannian IPM (RIPM) v.s. Euclidean IPM (EIPM).

1. RIPM generalizes EIPM from Euclidean space to general Riemannian manifolds. EIPM is a special case of RIPM when $\mathcal{M} = \mathbb{R}^n$ or $\mathbb{R}^{m \times n}$ in (RCOP).
2. RIPM inherits all the advantages of Riemannian optimization. For example, we can exploit the geometric structure of $\mathcal{M}$, which is usually regarded as a set of constraints from the Euclidean viewpoint.
3. Note that in both RIPM and EIPM, we have to solve the condensed Newton equation (19) at each iteration. However, if the equality constraints can be considered to be a manifold, RIPM can solve (19) with a *smaller* order on $T_x \mathcal{M} \times \mathbb{R}^l$. For example, the problem II (Model_St) can be rewritten as:

$$\min_{X \in \mathbb{R}^{n \times k}} -2 \operatorname{trace}(X^T C) \quad \text{s.t.} \ X^T X = I_k, \ X \geqslant 0.$$

Here, Stiefel manifold is replaced by the equality constraints, i.e., we define $h : \mathbb{R}^{n \times k} \to \operatorname{Sym}(k) : X \mapsto h(X) := X^T X - I_k$; and $\mathcal{M} = \mathbb{R}^{n \times k}$, $l = \dim \operatorname{Sym}(k) = k(k+1)/2$ in (RCOP). Then, when we apply EIPM, it requires us to solve (19) of order $nk + k(k+1)/2$. On the other hand, if we apply RIPM to (Model_St), then (19) reduces to (21) since there are only inequality constraints on $\mathcal{M} = \operatorname{St}(n, k)$. In this case, we solve the equation of order $nk - k(k+1)/2$, i.e., the dimension of $\operatorname{St}(n, k)$. Compared to EIPM, using RIPM reduces our dimensionality by $k(k+1)$.

Table 2 Performance of various Riemannian methods on problem II (Model_St)

(n,k)	(40,8)				(50,10)			
	Success	Time (s)	Iter	Error	Success	Time (s)	Iter	Error
RALM	**1**	2.347	45	5.41×10^{-7}	**1**	4.344	54	5.21×10^{-7}
REPM$_{\text{lqh}}$	0	–	–	–	0	–	–	–
REPM$_{\text{lse}}$	0	–	–	–	0	–	–	–
RSQP	**0.9**	1.352×10	7	2.05×10^{-9}	0.7	3.097×10	6	2.47×10^{-9}
RIPM	**1**	**2.225**	31	3.72×10^{-8}	**1**	**3.785**	32	3.38×10^{-8}

(n,k)	(60,12)				(70,14)			
	Success	Time (s)	Iter	Error	Success	Time (s)	Iter	Error
RALM	**1**	**4.097**	34	4.93×10^{-7}	**1**	**6.234**	37	5.34×10^{-7}
REPM$_{\text{lqh}}$	0	–	–	–	0	–	–	–
REPM$_{\text{lse}}$	0	–	–	–	0	–	–	–
RSQP	0.65	7.802×10	7	6.48×10^{-9}	0.85	1.661×10^{2}	7	2.64×10^{-9}
RIPM	**1**	5.555	32	2.81×10^{-8}	**1**	7.574	33	2.45×10^{-8}

Algorithms with values greater than or equal to 0.9 in the "Success" column are bolded. Among these, the one with the fastest result in the "Time" column is also highlighted in bold

Table 3 Performance of various Riemannian methods on problem II (Model_Ob)

(n, k)	(40,8)				(50,10)			
	Success	Time (s)	Iter	Error	Success	Time (s)	Iter	Error
RALM	**1**	**2.510**	51	5.04×10^{-7}	**0.95**	**4.727**	64	4.94×10^{-7}
REPM$_{\text{lqh}}$	0	–	–	–	0	–	–	–
REPM$_{\text{lse}}$	0	–	–	–	0	–	–	–
RSQP	0.65	8.618	5	2.30×10^{-10}	0.7	2.782×10	6	1.12×10^{-10}
RIPM	**1**	3.791	22	5.62×10^{-9}	**1**	5.880	23	7.93×10^{-9}

(n, k)	(60,12)				(70,14)			
	Success	Time (s)	Iter	Error	Success	Time (s)	Iter	Error
RALM	0.6	5.725	49	3.82×10^{-7}	0.6	8.223	52	3.85×10^{-7}
REPM$_{\text{lqh}}$	0	–	–	–	0	–	–	–
REPM$_{\text{lse}}$	0	–	–	–	0	–	–	–
RSQP	0.7	4.446×10	5	1.17×10^{-9}	0.5	9.138×10	5	1.82×10^{-9}
RIPM	**1**	**7.134**	23	9.69×10^{-9}	**1**	**9.268**	24	1.06×10^{-8}

Algorithms with values greater than or equal to 0.9 in the "Success" column are bolded. Among these, the one with the fastest result in the "Time" column is also highlighted in bold

4. RIPM can solve some problems that EIPM cannot. For example, the problem I (NLRM) can be rewritten as:

$$\min_{X \in \mathbb{R}^{m \times n}} \|A - X\|_{\mathrm{F}}^2 \quad \text{s.t. } \mathrm{rank}\,(X) = r, \ X \geqslant 0.$$

Since the rank function, $X \mapsto \mathrm{rank}\,(X)$, is not even continuous, we cannot apply EIPM.

Future Work I: Preconditioner for linear operator equation.
With regard to the complementary condition, $S_k^{-1} Z_k$ values display a huge difference in magnitude as $k \to \infty$. The operator $\Theta := G_x S^{-1} Z G_x^*$ causes the system (19) to be ill-conditioned, risking failure of the iterative method without preconditioning. Matrix-decomposition-based preconditioner methods cannot be applied to a problem that does not have a matrix form. A possible alternative is to find a nonsingular $\mathcal{P}$ such that the condition number of $\mathcal{P}^{-1}\mathcal{T}$ is smaller.

Future Work II: Treatment of more state-of-the-art interior point methods.
While we have considered interior point methods on a manifold for the first time, our Euclidean theoretic counterpart is an early nonlinear interior point method algorithm [11]; however, the counterpart now appears to be obsolete compared with more recent interior point methods. For example, our method does not drive the iteration toward minimizers but only toward stationary points; globalization is done by monitoring only the KKT residuals; moreover, the boundedness assumption **(C2)** of $\{z_k\}$ is too strong to hold in some simple cases (see Wächter–Biegler effect [32]). It remains an important issue to adapt more modern interior point methods to manifolds, although we may encounter various difficulties in Riemannian geometry.

Data Availability The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

References

1. Absil, P.-A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton (2009)
2. Bergmann, R., Herzog, R.: Intrinsic formulation of KKT conditions and constraint qualifications on smooth manifolds. SIAM J. Optim. **29**, 2423–2444 (2019)
3. Bonettini, S., Galligani, E., Ruggiero, V.: An inexact Newton method combined with Hestenes multipliers' scheme for the solution of Karush–Kuhn–Tucker systems. Appl. Math. Comput. **168**, 651–676 (2005)
4. Bortoloti, MAd.A., Fernandes, T.A., Ferreira, O.P., Yuan, J.: Damped Newton's method on Riemannian manifolds. J. Glob. Optim. **77**, 643–660 (2020)
5. Boumal, N.: An Introduction to Optimization on Smooth Manifolds. Cambridge University Press, Cambridge (2023)
6. Boumal, N., Mishra, B., Absil, P.-A., Sepulchre, R.: Manopt, a Matlab toolbox for optimization on manifolds. J. Mach. Learn. Res. **15**, 1455–1459 (2014)
7. Do Carmo, M.P.: Riemannian Geometry. Birkhäuser, Boston (1992)

8. Dennis, J.E., Schnabel, R.B.: Numerical Methods for Unconstrained Optimization and Nonlinear Equations. SIAM (1996)
9. Durazzi, C.: On the Newton interior-point method for nonlinear programming problems. J. Optim. Theory Appl. **104**, 73–90 (2000)
10. Durazzi, C., Ruggiero, V.: Global convergence of the Newton interior-point method for nonlinear programming. J. Optim. Theory Appl. **120**, 199–208 (2004)
11. El-Bakry, A., Tapia, R.A., Tsuchiya, T., Zhang, Y.: On the formulation and theory of the Newton interior-point method for nonlinear programming. J. Optim. Theory Appl. **89**, 507–541 (1996)
12. Fernandes, T.A., Ferreira, O.P., Yuan, J.: On the superlinear convergence of Newton's method on Riemannian manifolds. J. Optim. Theory Appl. **173**, 828–843 (2017)
13. Ferreira, O.P., Svaiter, B.F.: Kantorovich's theorem on Newton's method in Riemannian manifolds. J. Complex. **18**, 304–329 (2002)
14. Gallivan, K.A., Qi, C., Absil, P.-A.: A Riemannian Dennis–Moré condition. In: Berry, M.W., Gallivan, K.A., Gallopoulos, E., Grama, A., Philippe, B., Saad, Y., Saied, F. (eds.) High-Performance Scientific Computing: Algorithms and Applications, pp. 281–293. Springer, London (2012)
15. Hirai, H., Nieuwboer, H., Walter, M.: Interior-point methods on manifolds: theory and applications, arXiv:2303.04771, (2023) to appear in FOCS (2023)
16. Hu, J., Liu, X., Wen, Z.W., Yuan, Y.X.: A brief introduction to manifold optimization. J. Oper. Res. Soc. China **8**, 199–248 (2020)
17. Huang, W.: Optimization Algorithms on Riemannian Manifolds with Applications, Ph.D. thesis, Florida State University, Tallahassee, FL (2013)
18. Huang, W., Absil, P.-A., Gallivan, K.A.: A Riemannian symmetric rank-one trust-region method. Math. Program. **150**, 179–216 (2015)
19. Jiang, B., Meng, X., Wen, Z.W., Chen, X.: An exact penalty approach for optimization with nonnegative orthogonality constraints. Math. Program. 1–43 (2022)
20. Kojima, M., Mizuno, S., Yoshise, A.: A primal-dual interior point algorithm for linear programming, in Progress in Mathematical Programming, N. Megiddo ed., Springer, pp. 29–47 (1989)
21. Lee, J.M.: Introduction to Smooth Manifolds, 2nd ed., Grad. Texts in Math. 218, Springer (2013)
22. Lee, J.M.: Introduction to Riemannian Manifolds, 2nd ed., Grad. Texts in Math. 176, Springer (2018)
23. Liu, C., Boumal, N.: Simple algorithms for optimization on Riemannian manifolds with constraints. Appl. Math. Optim. **82**, 949–981 (2020)
24. Lustig, I., Marsten, R.E., Shanno, D.F.: Computational experience with a primal-dual interior point method for linear programming. Linear Algebra Appl. **152**, 191–222 (1991)
25. Megiddo, N.: Pathways to the optimal set in linear programming, in Progress in Mathematical Programming, N. Megiddo ed., Springer, pp. 131–158 (1989)
26. Nocedal, J., Wright, S.J., II (eds.): Numerical Optimization. Springer, New York (2006)
27. Obara, M., Okuno, T., Takeda, A.: Sequential quadratic optimization for nonlinear optimization problems on Riemannian manifolds. SIAM J. Optim. **32**, 822–853 (2022)
28. Ortega, J.M., Rheinboldt, W.C.: Iterative Solution of Nonlinear Equations in Several Variables. SIAM (2000)
29. Saad, Y., 2nd ed., Iterative Methods for Sparse Linear Systems, Other Titles in Applied Mathematics SIAM (2003)
30. Schiela, A., Ortiz, J.: An SQP method for equality constrained optimization on Hilbert manifolds. SIAM J. Optim. **31**, 2255–2284 (2021)
31. Song, G.J., Ng, M.K.: Nonnegative low rank matrix approximation for nonnegative matrices. Appl. Math. Lett. **105**, 106300 (2020)
32. Wächter, A., Biegler, L.T.: Failure of global convergence for a class of interior point methods for nonlinear programming. Math. Program. **88**, 565–574 (2000)
33. Wright, S.J.: Primal-dual Interior-point Methods. SIAM (1997)
34. Yamakawa, Y., Sato, H.: Sequential optimality conditions for nonlinear optimization on Riemannian manifolds and a globally convergent augmented Lagrangian method. Comput. Optim. Appl. **81**, 397–421 (2022)
35. Yamashita, H., Yabe, H.: Superlinear and quadratic convergence of some primal-dual interior point methods for constrained optimization. Math. Program. **75**, 377–397 (1996)
36. Yang, W.H., Zhang, L.H., Song, R.: Optimality conditions for the nonlinear programming problems on Riemannian manifolds. Pacific J. Optim. **10**, 415–434 (2014)
37. Ye, Y.: Interior Point Algorithms: Theory and Analysis. Wiley, London (1997)

38. Zhu, X., Sato, H.: Riemannian conjugate gradient methods with inverse retraction. Comput. Optim. Appl. **77**, 779–810 (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.