Riemannian Conjugate Gradient Methods with Inverse Retraction

Xiaojing Zhu[1*] · Hiroyuki Sato[2]

[1]*College of Mathematics and Physics, Shanghai University of Electric Power, Yangpu District, Shanghai 200090, China*
[2]*Department of Applied Mathematics and Physics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan*

Abstract We propose a new class of Riemannian conjugate gradient (CG) methods, in which inverse retraction is used instead of vector transport for search direction construction. In existing methods, differentiated retraction is often used for vector transport to move the previous search direction to the current tangent space. However, a different perspective is adopted here, motivated by the fact that inverse retraction directly measures the displacement from the current to the previous points in terms of tangent vectors at the current point. The proposed algorithm is implemented with the Fletcher–Reeves and the Dai–Yuan formulae, respectively, and global convergence is established using modifications of the Riemannian Wolfe conditions. Computational details of the practical inverse retractions over the Stiefel and fixed-rank manifolds are discussed. Numerical results obtained for the Brockett cost function minimization problem, the joint diagonalization problem, and the low-rank matrix completion problem demonstrate the potential effectiveness of Riemannian CG with inverse retraction.

Keywords: Riemannian optimization, conjugate gradient method, retraction, inverse retraction, Stiefel manifold, fixed-rank manifold

1 Introduction

The problem of minimizing a smooth function f over a Riemannian manifold $\mathcal{M}$, i.e.,

$$\min\ f(x)\ \text{ s.t. }\ x \in \mathcal{M} \tag{1}$$

(where by smooth we mean C^∞ or infinitely differentiable), has generated considerable interest in recent years because of its many important applications. The reader is referred to [2, 18, 19] and references therein for abundant applications of problem (1).

Riemannian optimization generalizes the concept of unconstrained optimization in Euclidean spaces. In addition to the basic Riemannian gradient descent and Newton's methods [2], various new Riemannian optimization methods have been developed in recent years, e.g., Riemannian trust region methods [1, 2, 19, 20], Riemannian conjugate gradient (CG) methods [2, 31, 33, 34], Riemannian quasi-Newton methods [19, 21, 22, 30, 31], and many other advanced Riemannian first-order methods targeting deterministic [14, 15, 16, 17, 23, 24, 43, 44] and stochastic [6, 25, 35, 39, 42] problems.

In this paper, we focus on Riemannian CG methods. In classical methodologies, a new search direction is constructed through addition of the negative gradient of the current point to a vector transport of the previous search direction. To the best of our knowledge, existing vector transports are only realized based on differentiated retraction or orthogonal projection. However, neither of these vector transport types is irreplaceable during construction of CG directions. Other methods of achieving the same effect exist. In this work, we propose a surrogate for vector transport, i.e., inverse retraction. An inverse retraction is simply the inverse map of a retraction. Therefore, it is not a new concept and follows from the concept of retraction. As mentioned in [4], inverse retractions are required in certain situations, e.g., for computation of the R-barycenter of a collection of points. From a geometric perspective,

*Corresponding author.
E-mail address: xjzhu2013@shiep.edu.cn

an inverse retraction directly measures the displacement between two points in terms of tangent vectors. From a numerical viewpoint, inverse retractions can be easily computed on a variety of manifolds. For example, when the manifold under consideration is a submanifold of a Euclidean space, an inverse orthographic retraction [3, 4] can be expressed as a simple projection onto tangent spaces. Inverse retraction is currently becoming a popular tool for Riemannian optimization, and is employed for algorithms such as the Riemannian stochastic variance reduced gradient method [35], the Riemannian stochastic averaging gradient descent method [39], and the Riemannian FISTA [23, 24].

The purpose of this paper is to enrich the theory of Riemannian CG methods by proposing inverse retraction as a competitive alternative to vector transport. We show that, in a theoretical framework, the proposed methods exhibit global convergence, similar to classical methods. Furthermore, in a practical framework (e.g., when implemented with state-of-the-art software such as Manopt [7]), we demonstrate that the proposed methods have the same efficiency as classical methods. This study makes three main contributions. First, new Riemannian CG directions are constructed by means of inverse retraction rather than vector transport. Second, for the proposed Riemannian CG algorithm with inverse retraction, modified Riemannian Wolfe conditions that involve inverse retraction instead of differentiated retraction are presented, and the global convergence of the new algorithm is confirmed. Third, we discuss the computational details of practical inverse retractions on the Stiefel and fixed-rank manifolds and show the effectiveness of Riemannian CG with inverse retraction in numerical experiments.

The remainder of this paper is organized as follows. Notation and preliminaries pertaining to Riemannian geometry and optimization are presented in Section 2. The new algorithm is proposed in Section 3 and its global convergence is analyzed in Section 4. Implementation details of several practical inverse retractions are discussed in Section 5 and numerical experiments are reported in Section 6. Conclusions are presented in Section 7.

2 Notation and preliminaries

2.1 Notation

Given a Riemannian manifold $\mathcal{M}$, a point $x \in \mathcal{M}$, and a function f defined over $\mathcal{M}$, $T_x\mathcal{M}$ denotes the tangent space to $\mathcal{M}$ at x, $T\mathcal{M} := \bigcup_{x\in\mathcal{M}} T_x\mathcal{M}$ denotes the tangent bundle of $\mathcal{M}$, $\langle\cdot,\cdot\rangle$ denotes the Riemannian metric on $\mathcal{M}$, $\langle\cdot,\cdot\rangle_x$ denotes the restriction of $\langle\cdot,\cdot\rangle$ to $T_x\mathcal{M}$, and ∇f denotes the Riemannian gradient of f. Given a matrix A, $\|A\|_2$ denotes the 2-norm of A, $\|A\|_{\mathrm{F}}$ denotes the Frobenius norm of A, and $\mathrm{tr}(A)$ denotes the trace of A if A is square. Given a subset S in a Euclidean space, $\mathrm{con}(S)$ denotes the convex hull of S.

2.2 Preliminaries

In Riemannian optimization, a general update scheme has the form

$$x_{k+1} = R_{x_k}(\alpha_k \xi_k), \tag{2}$$

where $\xi_k \in T_{x_k}\mathcal{M}$ is the search direction, $\alpha_k > 0$ is the step length, and R is a retraction, which is defined as follows.

Definition 1 [2] *A retraction R on a manifold $\mathcal{M}$ is a smooth map from the tangent bundle $T\mathcal{M}$ to $\mathcal{M}$ with the following properties, where R_x denotes the restriction of R to $T_x\mathcal{M}$.*

1. $R_x(0_x) = x$, where 0_x denotes the zero element of $T_x\mathcal{M}$.

2. With the canonical identification $T_{0_x}T_x\mathcal{M} \simeq T_x\mathcal{M}$, R_x satisfies $\mathrm{D}R_x(0_x) = \mathrm{id}_{T_x\mathcal{M}}$, where $\mathrm{D}R_x(0_x)$ denotes the differential of R_x at 0_x and $\mathrm{id}_{T_x\mathcal{M}}$ denotes the identity map on $T_x\mathcal{M}$.

Because vectors in different tangent spaces cannot be added, another important Riemannian optimization operation, i.e., vector transport, is defined as follows.

Definition 2 [2] *A vector transport $\mathcal{T}$ on a manifold $\mathcal{M}$ is a smooth map*

$$T\mathcal{M} \oplus T\mathcal{M} \to T\mathcal{M} : (\eta, \xi) \mapsto \mathcal{T}_\eta(\xi) \in T\mathcal{M}$$

satisfying the following properties for all $x \in \mathcal{M}$, where $\oplus$ denotes the Whitney sum

$$T\mathcal{M} \oplus T\mathcal{M} = \{(\xi_x, \eta_x) : \xi_x, \eta_x \in T_x\mathcal{M}, x \in \mathcal{M}\}.$$

1. There exists an associated retraction R such that $\mathcal{T}_{\eta_x}(\xi_x) \in T_{R_x(\eta_x)}\mathcal{M}$ for all $\eta_x, \xi_x \in T_x\mathcal{M}$.
2. $\mathcal{T}_{0_x}(\xi_x) = \xi_x$ for all $\xi_x \in T_x\mathcal{M}$.
3. $\mathcal{T}_{\eta_x}(a\xi_x + b\zeta_x) = a\mathcal{T}_{\eta_x}(\xi_x) + b\mathcal{T}_{\eta_x}(\zeta_x)$ for all $a, b \in \mathbb{R}$ and $\eta_x, \xi_x, \zeta_x \in T_x\mathcal{M}$.

The exponential map and parallel transport [10] are special examples of retraction and vector transport, respectively. In early works, such as the seminal reference [11], these tools were used in algorithmic design for Riemannian optimization methods. The exponential map $\exp_x : T_x\mathcal{M} \to \mathcal{M}$ at x is defined by $\exp_x(\xi_x) = \gamma_e(1)$, where $\gamma_e(t)$ is the geodesic such that $\gamma_e(0) = x$ and $\dot{\gamma}_e(0) = \xi_x$. The parallel transport $P_\gamma^{\gamma(t)\leftarrow\gamma(a)}$ along a curve γ transports $\xi_{\gamma(a)} \in T_{\gamma(a)}\mathcal{M}$ to $P_\gamma^{\gamma(t)\leftarrow\gamma(a)}\xi_{\gamma(a)} \in T_{\gamma(t)}\mathcal{M}$ with the property $\nabla_{\dot{\gamma}(t)}\left(P_\gamma^{\gamma(t)\leftarrow\gamma(a)}\xi_{\gamma(a)}\right) = 0$, where ∇ is the Levi-Civita connection. The exponential map, geodesics, and parallel transports are perfect in theory, e.g., $\nabla_{\dot{\gamma}_e(t)}\dot{\gamma}_e(t) = 0$, $\gamma_e(t) = \exp_x(t\xi_x)$, $\dot{\gamma}_e(t) = \mathrm{D}\exp_x(t\xi_x)[\xi_x] = P_{\gamma_e}^{\gamma_e(t)\leftarrow x}\xi$, and $\left\langle P_\gamma^{\gamma(t)\leftarrow\gamma(a)}\xi_{\gamma(a)}, P_\gamma^{\gamma(t)\leftarrow\gamma(a)}\eta_{\gamma(a)}\right\rangle_{\gamma(t)} \equiv \text{Const}$ for t. However, although these geometric properties are appealing, they are computationally intractable on many concrete matrix manifolds.

We next introduce two important topological concepts in Riemannian geometry: the normal neighborhood and totally normal neighborhood [10]. Because $\mathrm{D}\exp_x(0_x) = \mathrm{id}_{T_x\mathcal{M}}$, by the inverse function theorem, there exists a neighborhood V of 0_x in $T_x\mathcal{M}$ such that $\exp_x$ is a diffeomorphism on V. We call $U = \exp_x(V)$ a normal neighborhood of x. The definition of a totally normal neighborhood is given by the following theorem.

Theorem 1 [10] *For any $p \in \mathcal{M}$, there exists a so-called totally normal neighborhood W of p and a number $\delta > 0$ such that, for every $x \in W$, $\exp_x$ is a diffeomorphism on the open ball $B_\delta(0_x) \subset T_x\mathcal{M}$ centered at 0_x with radius δ, $W \subset \exp_x(B_\delta(0_x))$, and $\exp : (x, \xi_x) \mapsto (x, \exp_x(\xi_x))$ is a diffeomorphism on $\{(x, \xi_x) : x \in W, \xi_x \in B_\delta(0_x)\} \subset T\mathcal{M}$.*

Corollary 1 [10] *For any two points $x, y \in W$, where W is a totally normal neighborhood, there exists a unique minimizing geodesic joining x with y.*

As a retraction is a first-order approximation to the exponential, we can also define a retractive neighborhood and totally retractive neighborhood [20], analogous to a normal neighborhood and totally normal neighborhood. Thus, a retractive neighborhood of x with respect to R is $\tilde{U} = R_x(\tilde{V})$, where $\tilde{V}$ is a neighborhood of 0_x in $T_x\mathcal{M}$ such that R_x is a diffeomorphism on $\tilde{V}$. The local diffeomorphisms of the exponential map and retraction naturally induce two families of coordinate systems around $x \in \mathcal{M}$, called exponential coordinates and retractive coordinates, respectively, by identifying $T_x\mathcal{M}$ with $\mathbb{R}^n$, where n is the dimension of $\mathcal{M}$. The definition of a totally retractive neighborhood is given in the following theorem.

Theorem 2 [20] *Let R be a retraction on $\mathcal{M}$. For any $p \in \mathcal{M}$, there exists a so-called totally retractive neighborhood $\tilde{W}$ of p and a number $\tilde{\delta} > 0$ such that, for every $x \in \tilde{W}$, R_x is a diffeomorphism on the open ball $B_{\tilde{\delta}}(0_x) \subset T_x\mathcal{M}$ centered at 0_x with radius $\tilde{\delta}$, $\tilde{W} \subset R_x(B_{\tilde{\delta}}(0_x))$, and $R : (x, \xi_x) \mapsto (x, R_x(\xi_x))$ is a diffeomorphism on $\{(x, \xi_x) : x \in \tilde{W}, \xi_x \in B_{\tilde{\delta}}(0_x)\} \subset T\mathcal{M}$.*

We now move to our main topic, i.e., Riemannian CG methods. Existing CG directions in Riemannian optimization have the form

$$\xi_{k+1} = -\nabla f(x_{k+1}) + \beta_k \mathcal{T}_{\alpha_k\xi_k}(\xi_k). \tag{3}$$

Moreover, a commonly used vector transport for general manifolds is differentiated retraction

$$\mathcal{T}_{\eta_x}^{\mathrm{dr}}(\xi_x) = \mathrm{D}R_x(\eta_x)[\xi_x]. \tag{4}$$

After the search direction ξ_k is determined, an appropriate step length $\alpha_k > 0$ for global convergence should be computed. In classical Riemannian CG methods, α_k must satisfy the Wolfe conditions

$$f(R_{x_k}(\alpha_k\xi_k)) \le f(x_k) + c_1\alpha_k\langle\nabla f(x_k), \xi_k\rangle_{x_k}, \tag{5}$$

$$\langle\nabla f(R_{x_k}(\alpha_k\xi_k)), \mathrm{D}R_{x_k}(\alpha_k\xi_k)[\xi_k]\rangle_{R_{x_k}(\alpha_k\xi_k)} \ge c_2\langle\nabla f(x_k), \xi_k\rangle_{x_k}, \tag{6}$$

where c_1, c_2 are two constants with $0 < c_1 < c_2 < 1$, or the strong Wolfe conditions, for which (6) is replaced with

$$\left|\langle\nabla f(R_{x_k}(\alpha_k\xi_k)), \mathrm{D}R_{x_k}(\alpha_k\xi_k)[\xi_k]\rangle_{R_{x_k}(\alpha_k\xi_k)}\right| \le -c_2\langle\nabla f(x_k), \xi_k\rangle_{x_k}. \tag{7}$$

Note that these conditions make sense under the premise that ξ_k is a descent direction, i.e., $\langle\nabla f(x_k), \xi_k\rangle_{x_k} < 0$.

3 RCG with inverse retraction

Given two retractions R^{fw} and R^{bw}, we propose a new Riemannian CG scheme with inverse retraction

$$x_{k+1} = R^{\mathrm{fw}}_{x_k}(\alpha_k \xi_k), \tag{8}$$

$$\xi_{k+1} = -\nabla f(x_{k+1}) - \beta_k s_k \alpha_k^{-1} {R^{\mathrm{bw}}_{x_{k+1}}}^{-1}(x_k), \tag{9}$$

where R^{fw} is called the forward retraction, R^{bw} is called the backward retractions, ${R^{\mathrm{bw}}}^{-1}$ is the inverse map of R^{bw} and is called the inverse retraction associated with R^{bw}, s_k is a scaling number, β_k is the conventional CG parameter, and α_k^{-1} performs a role similar to "normalization."

According to Section 2, the inverse retraction ${R^{\mathrm{bw}}}^{-1}$ is well-defined if x_k is sufficiently close to x_{k+1}. Note also that R^{fw} and R^{bw} are allowed to be different. For a clearer understanding of inverse retraction, see Figure 1. The

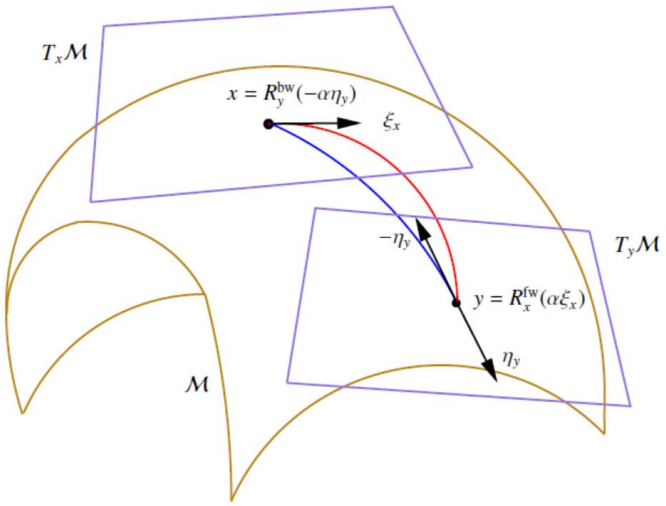

Figure 1: Illustration for $\eta_y = -\alpha^{-1}{R^{\mathrm{bw}}_y}^{-1}(x)$. The red curve is the forward retraction curve $\gamma^{\mathrm{fw}}(t) = R^{\mathrm{fw}}_x(t\xi_x), t \in [0, \alpha]$ which starts at x with velocity ξ_x and ends at y, and the blue curve is the backward retraction curve $\gamma^{\mathrm{bw}}(t) = R^{\mathrm{bw}}_y(-t\eta_y), t \in [0, \alpha]$ which starts at y with velocity $-\eta_y$ and ends at x.

number s_k is set as

$$s_k = \min\left\{ \frac{\|\alpha_k \xi_k\|_{x_k}}{\left\| {R^{\mathrm{bw}}_{x_{k+1}}}^{-1}(x_k) \right\|_{x_{k+1}}},\ 1 \right\} \tag{10}$$

and is introduced to ensure

$$\left\| s_k \alpha_k^{-1} {R^{\mathrm{bw}}_{x_{k+1}}}^{-1}(x_k) \right\|_{x_{k+1}} \leq \|\xi_k\|_{x_k}, \tag{11}$$

which is crucial for global convergence analysis of Riemannian CG methods. This scaling technique is also used in [33, 34]. There are various choices for β_k, e.g., the Fletcher–Reeves formula [12]

$$\beta_k^{\mathrm{FR}} = \frac{\|\nabla f(x_{k+1})\|^2_{x_{k+1}}}{\|\nabla f(x_k)\|^2_{x_k}}, \tag{12}$$

and the Dai–Yuan formula [9]

$$\beta_k^{\mathrm{DY}} = \frac{\langle \nabla f(x_{k+1}), \xi_{k+1} \rangle_{x_{k+1}}}{\langle \nabla f(x_k), \xi_k \rangle_{x_k}}, \tag{13}$$

which is implicit and, in our case, equivalent to the following explicit form:

$$\beta_k^{\mathrm{DY}} = -\frac{\|\nabla f(x_{k+1})\|^2_{x_{k+1}}}{s_k \left\langle \nabla f(x_{k+1}), \alpha_k^{-1} {R^{\mathrm{bw}}_{x_{k+1}}}^{-1}(x_k) \right\rangle_{x_{k+1}} + \langle \nabla f(x_k), \xi_k \rangle_{x_k}}. \tag{14}$$

The equivalence between (13) and (14) can be easily verified using (9).

Because the proposed search direction (9) is established using an inverse retraction instead of differentiated retraction (as in (3)), to ensure global convergence, we modify the Wolfe conditions ((5) and (6)) as follows:

$$f(R^{\mathrm{fw}}_{x_k}(\alpha_k\xi_k)) \le f(x_k) + c_1\alpha_k\langle\nabla f(x_k),\xi_k\rangle_{x_k}, \tag{15}$$

$$c_2\langle\nabla f(x_k),\xi_k\rangle_{x_k} \le -\left\langle\nabla f(R^{\mathrm{fw}}_{x_k}(\alpha_k\xi_k)),\alpha_k^{-1}R^{\mathrm{bw}}_{R^{\mathrm{fw}}_{x_k}(\alpha_k\xi_k)}{}^{-1}(x_k)\right\rangle_{R^{\mathrm{fw}}_{x_k}(\alpha_k\xi_k)} \le -c_3\langle\nabla f(x_k),\xi_k\rangle_{x_k}, \tag{16}$$

where c_1, c_2, and c_3 are three constants with $0 < c_1 < c_2 < 1$, $c_3 > c_2$. Note that the second inequality in (16), which is needed for Lemma 7 (Section 4), is established for purely theoretical purposes. When $c_3 \gg c_2$, (16) is very close to the condition

$$-\left\langle\nabla f(R^{\mathrm{fw}}_{x_k}(\alpha_k\xi_k)),\alpha_k^{-1}R^{\mathrm{bw}}_{R^{\mathrm{fw}}_{x_k}(\alpha_k\xi_k)}{}^{-1}(x_k)\right\rangle_{R^{\mathrm{fw}}_{x_k}(\alpha_k\xi_k)} \ge c_2\langle\nabla f(x_k),\xi_k\rangle_{x_k},$$

which is an analogy of the traditional Riemannian Wolfe condition (6) used in [33]. For the modified strong Wolfe conditions, (16) is replaced by

$$\left|\left\langle\nabla f(R^{\mathrm{fw}}_{x_k}(\alpha_k\xi_k)),\alpha_k^{-1}R^{\mathrm{bw}}_{R^{\mathrm{fw}}_{x_k}(\alpha_k\xi_k)}{}^{-1}(x_k)\right\rangle_{R^{\mathrm{fw}}_{x_k}(\alpha_k\xi_k)}\right| \le -c_2\langle\nabla f(x_k),\xi_k\rangle_{x_k}. \tag{17}$$

Note that, if the first inequality of (16) holds and $\langle\nabla f(x_k),\xi_k\rangle_{x_k} < 0$, the denominator of (14) is nonzero because, in that case,

$$s_k\left\langle\nabla f(x_{k+1}),\alpha_k^{-1}R^{\mathrm{bw}}_{x_{k+1}}{}^{-1}(x_k)\right\rangle_{x_{k+1}} \le -c_2\langle\nabla f(x_k),\xi_k\rangle_{x_k} < -\langle\nabla f(x_k),\xi_k\rangle_{x_k},$$

where $s_k \le 1$ is used in the first inequality.

A complete description of the proposed algorithm is presented in Algorithm 1. Note that we do not specify the formula for β_k. Furthermore, the step length conditions depend on the formula for β_k. For example, if we choose the Dai–Yuan formula (14), the weak Wolfe conditions (15) and (16) are sufficient to guarantee global convergence. However, if we choose the Fletcher–Reeves formula (12), the strong Wolfe conditions (15) and (17) are required.

Algorithm 1: RCG with inverse retraction

Input: parameters c_1, c_2, and c_3, with $0 < c_1 < c_2 < 1$, $c_3 > c_2$, and initial point $x_0 \in \mathcal{M}$.

Compute $\xi_0 = -\nabla f(x_0)$.

for $k = 0, 1, 2, \ldots$ **do**

  Compute $\alpha_k > 0$ satisfying (15) and (16) or satisfying (15) and (17), and obtain $x_{k+1} = R^{\mathrm{fw}}_{x_k}(\alpha_k\xi_k)$, $\eta_k = -\alpha_k^{-1}R^{\mathrm{bw}}_{x_{k+1}}{}^{-1}(x_k)$, and $g_{k+1} = \nabla f(x_{k+1})$.

  Compute s_k using (10) and β_k using a selected formula, e.g., (12) or (14).

  Compute $\xi_{k+1} = -g_{k+1} + \beta_k s_k \eta_k$.

As discussed above, the main novelty of the search direction (9) is that inverse retraction is used instead of vector transport. Note that vector transport is linear but inverse retraction is nonlinear. Our motivation for employing the inverse retraction $R^{\mathrm{bw}}_{x_{k+1}}{}^{-1}(x_k)$ is twofold. First, an inverse retraction directly measures the displacement from x_{k+1} to x_k in the sense of tangent vectors in $T_{x_{k+1}}\mathcal{M}$, with respect to the retraction R^{bw}. Second, an inverse retraction is computationally comparable to vector transport on many matrix manifolds. Given $\xi_x \in T_x\mathcal{M}$, if $\exp_x(t\xi_x)$ is in a normal neighborhood of x,

$$\exp^{-1}_{\exp_x(t\xi_x)}(x) = -t\mathrm{D}\exp_x(t\xi_x)[\xi_x]. \tag{18}$$

Therefore, when $R^{\mathrm{fw}} = R^{\mathrm{bw}} = \exp$, (9) is the same as (3) in cases where the vector transport $\mathcal{T}$ in (3) is the differentiated retraction, and $R^{\mathrm{bw}-1}$ is simply the logarithm $\log \equiv \exp^{-1}$. Note that the right-hand side of (18) can be rewritten in terms of parallel transport as $-t\mathrm{D}\exp_x(t\xi_x)[\xi_x] = P^{\gamma_e(t)\leftarrow x}_{\gamma_e}(-t\xi_x)$, where $\gamma_e(t) = \exp_x(t\xi_x)$. Note also that $s_k = 1$ if $R^{\mathrm{fw}} = R^{\mathrm{bw}} = \exp$, because parallel transport conserves the norm.

In what follows, we estimate the deviation between the differentiated retraction $\mathrm{D}R^{\mathrm{fw}}_x(\alpha\xi_x)[\xi_x]$ and the "normalized" negative inverse retraction $-\alpha^{-1}R^{\mathrm{bw}-1}_y(x)$, where $y = R^{\mathrm{fw}}_x(\alpha\xi_x)$. Before presenting the final evaluation (Proposition 1), we first prepare three technical lemmas.

Lemma 1 [13] *Let $\mathcal{N}$ be a compact coordinate neighborhood of some point, where a hat denotes a coordinate expression. There exist $m_2 > m_1 > 0$ such that, for all $x, y \in \mathcal{N}$,*

$$m_1\|\hat{x} - \hat{y}\|_2 \leq \mathrm{dist}(x, y) \leq m_2\|\hat{x} - \hat{y}\|_2.$$

Lemma 2 *Let ξ be a smooth vector field on $\mathcal{M}$ and let $\mathcal{N}$ be a compact subset of a totally normal neighborhood of some point on $\mathcal{M}$. There exists a constant $C_1 > 0$ such that, for all $x, y \in \mathcal{N}$,*

$$\left\|P_{\gamma_e}^{x\leftarrow y}\xi_y - \xi_x\right\|_x \leq C_1\mathrm{dist}(x, y),$$

where γ_e is the unique minimizing geodesic joining x with y.

Proof As parallel transport is invariant under affine reparametrization of the corresponding curve, we can assume γ_e is normalized with the initial condition $\gamma_e(0) = x$. Then, we have

$$P_{\gamma_e}^{x\leftarrow y}\xi_y - \xi_x = \int_0^{\mathrm{dist}(x,y)} \frac{\mathrm{d}}{\mathrm{d}t} P_{\gamma_e}^{x\leftarrow\gamma_e(t)}\xi_{\gamma_e(t)}dt = \int_0^{\mathrm{dist}(x,y)} P_{\gamma_e}^{x\leftarrow\gamma_e(t)}(\nabla_{\dot{\gamma}_e}\xi)_{\gamma_e(t)}dt. \tag{19}$$

As $P_{\gamma_e}^{x\leftarrow\gamma_e(t)}\left(\nabla_{\dot{\gamma}_e}\xi\right)_{\gamma_e(t)}$ is smooth and $\mathcal{N}$ is compact, there exists a constant $C_1 > 0$ such that

$$\sup\left\{\left\|P_{\gamma_e}^{x\leftarrow\gamma_e(t)}(\nabla_{\dot{\gamma}_e}\xi)_{\gamma_e(t)}\right\|_{\gamma_e(t)} :\ t \in [0, \mathrm{dist}(x, y)],\ x, y \in \mathcal{N}\right\} \leq C_1. \tag{20}$$

Combining (19) and (20), we obtain $\left\|P_{\gamma_e}^{x\leftarrow y}\xi_y - \xi_x\right\|_x \leq C_1\mathrm{dist}(x, y)$ for all $x, y \in \mathcal{N}$. □

Lemma 3 *Let $\mathcal{N}$ be a compact subset of a totally normal neighborhood of $p \in \mathcal{M}$, and let $\tilde{\mathcal{N}}$ be a compact subset of a totally retractive neighborhood of $p \in \mathcal{M}$ with respect to some retraction R on $\mathcal{M}$. There exists a constant $C_2 > 0$ such that, for all $x, y \in \mathcal{N} \cap \tilde{\mathcal{N}}$,*

$$\left\|R_y^{-1}(x) - \exp_y^{-1}(x)\right\|_y \leq C_2\mathrm{dist}(x, y)^2 = C_2\left\|\exp_y^{-1}(x)\right\|_y^2. \tag{21}$$

Proof By Definition 1 and the inverse function theorem, whether $\mathcal{L}_y(x) = R_y^{-1}(x)$ or $\mathcal{L}_y(x) = \exp_y^{-1}(x)$ for $x, y \in \mathcal{N} \cap \tilde{\mathcal{N}}$, the map $\mathcal{L}_y$ satisfies the conditions $\mathcal{L}_y(y) = 0_y \in T_y\mathcal{M}$ and $\mathrm{D}\mathcal{L}_y(y) = \mathrm{id}_{T_y\mathcal{M}}$. Thus, $R_y^{-1}(x)$ is a first-order approximation of the logarithm $\exp_y^{-1}(x)$.

We next show the existence of $C_2 > 0$ such that the inequality in (21) holds. Let W be a (totally) normal neighborhood of p containing $\mathcal{N}$. Let φ be an exponential coordinate chart centered at p, the domain of which covers W, and let $\tilde{\varphi}$ be the natural coordinate chart on the tangent bundle $T\mathcal{M}$ generated by φ. Thus,

$$\tilde{\varphi}\left(v^i\left.\frac{\partial}{\partial x^i}\right|_q\right) = \left(x^1(q), \ldots, x^n(q), v^1, \ldots, v^n\right),$$

where $x^1, \ldots, x^n$ are the coordinate functions of φ, $\frac{\partial}{\partial x^1}, \ldots, \frac{\partial}{\partial x^n}$ are the corresponding coordinate fields, and n is the dimension of $\mathcal{M}$. Denote $\Omega = \mathcal{N} \cap \tilde{\mathcal{N}}$. Then, a coordinate expression of the map

$$F : \Omega \times \Omega \to T\mathcal{M},\ (y, x) \mapsto R_y^{-1}(x) - \exp_y^{-1}(x)$$

is

$$\hat{F} = \tilde{\varphi} \circ F \circ \left(\varphi^{-1} \times \varphi^{-1}\right) :\ \varphi(\Omega) \times \varphi(\Omega) \to \mathbb{R}^{2n},$$

where $\varphi^{-1} \times \varphi^{-1}$ is the Cartesian product of two φ^{-1} defined by $\left(\varphi^{-1} \times \varphi^{-1}\right)(\hat{y}, \hat{x}) = \left(\varphi^{-1}(\hat{y}), \varphi^{-1}(\hat{x})\right)$. We define $\hat{x} = \varphi(x)$ and $\hat{y} = \varphi(y)$ and express $\hat{F}$ component-wisely as

$$\hat{F}(\hat{y}, \hat{x}) = \left(\hat{y}^1, \ldots, \hat{y}^n, \hat{F}^1(\hat{y}, \hat{x}), \ldots, \hat{F}^n(\hat{y}, \hat{x})\right).$$

Note that all $\hat{F}^i$ are smooth functions as F is smooth. Additionally, $\varphi(\Omega)$ is a closed subset in $\mathbb{R}^n$ that may not be convex; therefore, if necessary, we extend $\hat{F}$ smoothly to $\varphi(\Omega) \times \mathrm{con}(\varphi(\Omega))$ in accordance with Lemma 2.26 in

[28]. Then, by Taylor's theorem and the knowledge that $\hat{F}(\hat{y},\hat{y}) = 0$ and $R_y^{-1}(x)$ agrees with $\exp_y^{-1}(x)$ up to first order, we have

$$\hat{F}^i(\hat{y},\hat{x}) = \frac{1}{2}(\hat{x}-\hat{y})^\top \nabla_{22}^2 \hat{F}^i(\hat{y},\hat{z}_i)(\hat{x}-\hat{y}), \quad i = 1,\dots,n,$$

where ∇_{22}^2 denotes the Hessian with respect to the second part of the variables and all $\hat{z}^i$ lie in the segment joining $\hat{x}$ and $\hat{y}$. Let $\hat{G}(\hat{y}) = \left(\hat{G}_{ij}(\hat{y})\right)$ be the matrix expression of the Riemannian metric on $T_y\mathcal{M}$ associated with the chart φ. Then, we have

$$\begin{aligned}
\left\|R_y^{-1}(x) - \exp_y^{-1}(x)\right\|_y &= \sqrt{\sum_{i=1}^n \sum_{j=1}^n \hat{G}_{ij}(\hat{y})\hat{F}^i(\hat{y},\hat{x})\hat{F}^j(\hat{y},\hat{x})} \\
&\le \frac{1}{2}\sqrt{\left\|\hat{G}(\hat{y})\right\|_2 \sum_{i=1}^n \left\|\nabla_{22}^2 \hat{F}^i(\hat{y},\hat{z}_i)\right\|_2^2}\, \|\hat{x}-\hat{y}\|_2^2 \\
&\le \frac{1}{2}\sqrt{n\left\|\hat{G}(\hat{y})\right\|_2}\max_{i\in\{1,\dots,n\}}\left\|\nabla_{22}^2 \hat{F}^i(\hat{y},\hat{z}_i)\right\|_2\, \|\hat{x}-\hat{y}\|_2^2.
\end{aligned}$$

Here, we use the fact that the 2-norm coincides with the spectral radius for real symmetric matrices. As $\varphi(\Omega)$ is compact, so too is $\mathrm{con}(\varphi(\Omega))$ by Corollary 2.30 in [32]. Then,

$$C_2' := \max_{\substack{\hat{y}\in\varphi(\Omega)\\ \hat{z}\in\mathrm{con}(\varphi(\Omega))\\ i\in\{1,\dots,n\}}} \frac{1}{2}\sqrt{n\left\|\hat{G}(\hat{y})\right\|_2}\, \left\|\nabla_{22}^2 \hat{F}^i(\hat{y},\hat{z})\right\|_2 < \infty.$$

Combining this relation with Lemma 1, we see that $C_2 := C_2'/m_1^2$ is the desired constant. Note that the equality in (21) comes directly from the basic properties of exp. □

Remark 1 If, in Lemma 3, $\tilde{\mathcal{N}}$ is assumed to be a compact subset of a totally retractive neighborhood of $p \in \mathcal{M}$ with respect to both R^{fw} and R^{bw}, we can find a constant C_2 that is independent of R^{fw} and R^{bw} and satisfies (21).

We can now give an upper bound for the deviation $\left\|-\alpha^{-1}R_y^{\mathrm{bw}^{-1}}(x) - \mathrm{D}R_x^{\mathrm{fw}}(\alpha\xi_x)[\xi_x]\right\|_y$.

Proposition 1 *Let $\mathcal{N}$ be a compact subset of a totally normal neighborhood of $p \in \mathcal{M}$, and let $\tilde{\mathcal{N}}$ be a compact subset of a totally retractive neighborhood of $p \in \mathcal{M}$ with respect to both R^{fw} and R^{bw}. Then, there exists a constant $C > 0$ such that, for all $x \in \mathcal{N}\cap\tilde{\mathcal{N}}$, $\xi_x \in T_x\mathcal{M}$, and $\alpha > 0$ satisfying both $y = R_x^{\mathrm{fw}}(\alpha\xi_x) \in \mathcal{N}\cap\tilde{\mathcal{N}}$ and $z = \exp_x(\alpha\xi_x) \in \mathcal{N}$, the following holds:*

$$\left\|-\alpha^{-1}R_y^{\mathrm{bw}^{-1}}(x) - \mathrm{D}R_x^{\mathrm{fw}}(\alpha\xi_x)[\xi_x]\right\|_y \le C(\alpha+\alpha^2)\|\xi_x\|_x^2.$$

Proof For simplicity of the proof, we emphasize that all variables x, ξ_x, α, $y = R_x^{\mathrm{fw}}(\alpha\xi_x)$, and $z = \exp_x(\alpha\xi_x)$ mentioned below are restricted by the proposition hypothesis. Thus, the constants $C_3,\dots,C_7$ and C introduced below are independent of these variables because of compactness, as in the previous lemmas. Note that not only x, y, and z remain in compact sets, but also $\|\alpha\xi_x\|_x \le \delta$ for some positive constant δ independent of x, according to Theorems 1 and 2.

We choose a basis for $T_p\mathcal{M}$ and let $\left(x^i\right)$ be the corresponding exponential coordinates and $\left(\frac{\partial}{\partial x^i}\right)$ be the corresponding coordinate vector fields. Suppose that $\hat{\exp}_x(t\xi_x) = (v_1(t),\dots,v_n(t))$ and $\hat{R}_x^{\mathrm{fw}}(t\xi_x) = (\omega_1(t),\dots,\omega_n(t))$, where n is the dimension of $\mathcal{M}$ and a hat continues to denote a coordinate expression. Then,

$$\mathrm{D}\exp_x(\alpha\xi_x)[\xi_x] = \frac{\mathrm{d}}{\mathrm{d}t}\exp_x(t\xi_x)\Big|_{t=\alpha} = \sum_{i=1}^n v_i'(\alpha)\left.\frac{\partial}{\partial x^i}\right|_z \tag{22}$$

and

$$\mathrm{D}R_x^{\mathrm{fw}}(\alpha\xi_x)[\xi_x] = \frac{\mathrm{d}}{\mathrm{d}t}R_x^{\mathrm{fw}}(t\xi_x)\Big|_{t=\alpha} = \sum_{i=1}^n \omega_i'(\alpha)\left.\frac{\partial}{\partial x^i}\right|_y. \tag{23}$$

As a retraction is a first-order approximation to the exponential, we have by Taylor's theorem that $\upsilon_i(\alpha) - \omega_i(\alpha) = \mathfrak{h}_{\alpha\xi_x}(\alpha\xi_x, \alpha\xi_x)$, where $\mathfrak{h}_\xi(\cdot,\cdot)$ is a quadratic form for any fixed ξ and varies smoothly in ξ. Thus, by the compactness of $\mathcal{N}$ or $\tilde{\mathcal{N}}$ and $\alpha\|\xi_x\|_x \le \delta$, there exists a constant $C_3 > 0$ such that

$$\left|\upsilon_i'(\alpha) - \omega_i'(\alpha)\right| \le C_3\alpha\|\xi_x\|_x^2. \tag{24}$$

By the compactness of $\mathcal{N}$ or $\tilde{\mathcal{N}}$ and the fact

$$\alpha\left\|\sum_{i=1}^n \upsilon_i'(\alpha)\left.\frac{\partial}{\partial x^i}\right|_z\right\|_z = \alpha\left\|\sum_{i=1}^n \upsilon_i'(0)\left.\frac{\partial}{\partial x^i}\right|_x\right\|_x = \alpha\|\xi_x\|_x \le \delta,$$

we can find a constant $C_4 > 0$ such that

$$\max\left\{\alpha\left|\upsilon_i'(\alpha)\right|,\ \left\|\left.\frac{\partial}{\partial x^i}\right|_y\right\|_y\right\} \le C_4. \tag{25}$$

Let $\eta_x = \alpha^{-1}\exp_x^{-1}(y)$. By Lemma 1 with exponential coordinates centered at x and the compactness of $\mathcal{N}\cap\tilde{\mathcal{N}}$, there exists a constant $C_5 > 0$ such that

$$\operatorname{dist}(y,z) = \operatorname{dist}(\exp_x(\alpha\eta_x), \exp_x(\alpha\xi_x)) \le C_5\alpha\|\eta_x - \xi_x\|_x. \tag{26}$$

There also exists a constant $C_6 > 0$ such that

$$\|\eta_x - \xi_x\|_x = \left\|\alpha^{-1}\exp_x^{-1}(y) - \alpha^{-1}R_x^{\mathrm{fw}^{-1}}(y)\right\|_x \le C_2\alpha^{-1}\operatorname{dist}(x,y)^2 \le C_2C_6\alpha\|\xi_x\|_x^2. \tag{27}$$

The first inequality follows from (21) and the second follows from Lemma 1 with R^{fw}–retractive coordinates centered at x and the compactness of $\mathcal{N}\cap\tilde{\mathcal{N}}$. Let γ_e be the unique minimizing geodesic joining y with z. It follows from (22) that

$$P_{\gamma_e}^{y\leftarrow z}\mathrm{D}\exp_x(\alpha\xi_x)[\xi_x] = \sum_{i=1}^n \upsilon_i'(\alpha)P_{\gamma_e}^{y\leftarrow z}\left.\frac{\partial}{\partial x^i}\right|_z. \tag{28}$$

Combining (23)–(28) and using Lemma 2, we obtain

$$\begin{aligned}
&\left\|P_{\gamma_e}^{y\leftarrow z}\mathrm{D}\exp_x(\alpha\xi_x)[\xi_x] - \mathrm{D}R_x^{\mathrm{fw}}(\alpha\xi_x)[\xi_x]\right\|_y \le \sum_{i=1}^n\left\|\upsilon_i'(\alpha)P_{\gamma_e}^{y\leftarrow z}\left.\frac{\partial}{\partial x^i}\right|_z - \omega_i'(\alpha)\left.\frac{\partial}{\partial x^i}\right|_y\right\|_y\\
&\le \sum_{i=1}^n\left(\left|\upsilon_i'(\alpha)\right|\cdot\left\|P_{\gamma_e}^{y\leftarrow z}\left.\frac{\partial}{\partial x^i}\right|_z - \left.\frac{\partial}{\partial x^i}\right|_y\right\|_y + \left|\upsilon_i'(\alpha) - \omega_i'(\alpha)\right|\cdot\left\|\left.\frac{\partial}{\partial x^i}\right|_y\right\|_y\right)\\
&\le n\alpha^{-1}C_4C_1\operatorname{dist}(y,z) + nC_3C_4\alpha\|\xi_x\|_x^2 \le nC_4(C_1C_2C_5C_6 + C_3)\alpha\|\xi_x\|_x^2.
\end{aligned} \tag{29}$$

Let $\zeta_w = \frac{\mathrm{d}}{\mathrm{d}t}\exp_x\left(\exp_x^{-1}(w) + t\xi_x\right)\Big|_{t=0}$, which is a smooth local vector field. Then, $\mathrm{D}\exp_x(\alpha\eta_x)[\xi_x] = \zeta_y$ and $\mathrm{D}\exp_x(\alpha\xi_x)[\xi_x] = \zeta_z$. Using Lemma 2, (26), and (27), we obtain

$$\left\|\mathrm{D}\exp_x(\alpha\eta_x)[\xi_x] - P_{\gamma_e}^{y\leftarrow z}\mathrm{D}\exp_x(\alpha\xi_x)[\xi_x]\right\|_y = \left\|\zeta_y - P_{\gamma_e}^{y\leftarrow z}\zeta_z\right\|_y \le C_1\operatorname{dist}(y,z) \le C_1C_2C_5C_6\alpha^2\|\xi_x\|_x^2. \tag{30}$$

Because exp is smooth and $\{(x,\alpha\eta_x)\}_{x\in\mathcal{N}\cap\tilde{\mathcal{N}}}$ is contained in some compact subset of $T\mathcal{M}$, there exists a constant $C_7 > 0$ such that

$$\left\|\mathrm{D}\exp_x(\alpha\eta_x)[\xi_x] - \mathrm{D}\exp_x(\alpha\eta_x)[\eta_x]\right\|_y \le C_7\|\eta_x - \xi_x\|_x \le C_2C_6C_7\alpha\|\xi_x\|_x^2, \tag{31}$$

where the second inequality follows from (27). By (18), (21), and (27), we also have

$$\left\|-\alpha^{-1}R_y^{\mathrm{bw}^{-1}}(x) - \mathrm{D}\exp_x(\alpha\eta_x)[\eta_x]\right\|_y = \left\|\alpha^{-1}R_y^{\mathrm{bw}^{-1}}(x) - \alpha^{-1}\exp_y^{-1}(x)\right\|_y \le C_2C_6\alpha\|\xi_x\|_x^2. \tag{32}$$

Finally, combining (29)–(32), we obtain

$$\begin{aligned}
&\left\|-\alpha^{-1}R_y^{\mathrm{bw}^{-1}}(x) - \mathrm{D}R_x^{\mathrm{fw}}(\alpha\xi_x)[\xi_x]\right\|_y \\
\leq\ &\left\|P_{\gamma_e}^{y\leftarrow z}\mathrm{D}\exp_x(\alpha\xi_x)[\xi_x] - \mathrm{D}R_x^{\mathrm{fw}}(\alpha\xi_x)[\xi_x]\right\|_y + \left\|\mathrm{D}\exp_x(\alpha\eta_x)[\xi_x] - P_{\gamma_e}^{y\leftarrow z}\mathrm{D}\exp_x(\alpha\xi_x)[\xi_x]\right\|_y \\
&+ \left\|\mathrm{D}\exp_x(\alpha\eta_x)[\xi_x] - \mathrm{D}\exp_x(\alpha\eta_x)[\eta_x]\right\|_y + \left\|-\alpha^{-1}R_y^{\mathrm{bw}^{-1}}(x) - \mathrm{D}\exp_x(\alpha\eta_x)[\eta_x]\right\|_y \\
\leq\ &C(\alpha+\alpha^2)\|\xi_x\|_x^2,
\end{aligned}$$

where $C = \max\{nC_4(C_1C_2C_5C_6 + C_3) + C_2C_6C_7 + C_2C_6,\ C_1C_2C_5C_6\}$. □

4 Global convergence

In this section, we prove the global convergence of Algorithm 1. First, we make the following assumptions.

Assumption 1 *The objective function f is smooth and bounded below on $\mathcal{M}$.*

Assumption 2 *There exists a positive constant L_g such that $\|\nabla f(x_k)\|_{x_k} \leq L_g$ for all k.*

Assumption 3 *There exists a positive constant L_h such that $\sup_{t\in[0,\alpha_k]} \left\|\nabla^2(f\circ R_{x_k}^{\mathrm{fw}})(t\xi_k)\right\|_{x_k} \leq L_h$ for all k.*

Assumption 4 *There exists a positive constant C such that, for all k,*

$$\left\|-\alpha_k^{-1}R_{x_{k+1}}^{\mathrm{bw}^{-1}}(x_k) - \mathrm{D}R_{x_k}^{\mathrm{fw}}(\alpha_k\xi_k)[\xi_k]\right\|_{x_{k+1}} \leq C(\alpha_k+\alpha_k^2)\|\xi_k\|_{x_k}^2.$$

Remark 2 Assumption 4 is important for convergence. In Theorem 3, we prove Zotendijk's condition for inverse retraction using an argument similar to that for differentiated retraction. This assumption restricts the difference between the inverse and differentiated retractions and, thus, facilitates the argument in the proof of Theorem 3. From Proposition 1, a sufficient condition for Assumption 4 to hold is that there exists a compact subset $\mathcal{N}$ of a totally normal neighborhood of $p \in \mathcal{M}$, and a compact subset $\tilde{\mathcal{N}}$ of a totally retractive neighborhood of $p \in \mathcal{M}$ with respect to both R^{fw} and R^{bw}, such that $\{x_k\}$ is contained in $\mathcal{N}\cap\tilde{\mathcal{N}}$ and $\left\{\exp_{x_k}(\alpha_k\xi_k)\right\}$ is contained in $\mathcal{N}$. This sufficient condition is plausible if the iterates x_k are not excessively scattered.

4.1 Zoutendijk's theorem

Zoutendijk's theorem is a key result in global convergence theory for line search optimization algorithms, and can be established under the Wolfe conditions.

Theorem 3 *Let $\{x_k\} \subset \mathcal{M}$ be a sequence generated by a Riemannian optimization scheme of form (2). Suppose Assumptions 1–4 hold, with $\langle\nabla f(x_k), \xi_k\rangle_{x_k} < 0$, $\|\nabla f(x_k)\|_{x_k} \leq \mu\|\xi_k\|_{x_k}$ for some constant $\mu > 0$, and α_k satisfying the Wolfe conditions (15) and (16). Then,*

$$\sum_{k=0}^{\infty} \frac{\langle\nabla f(x_k), \xi_k\rangle_{x_k}^2}{\|\xi_k\|_{x_k}^2} < \infty. \tag{33}$$

Proof The proof is similar to that for Zoutendijk's theorem in Euclidean spaces, e.g., Theorem 3.2 in [29]. The main difference is that our modified Wolfe conditions (15) and (16) yield a different lower bound for α_k. From Assumption 3, we have the following, for all k:

$$\begin{aligned}
&\left\langle\nabla f(x_{k+1}), \mathrm{D}R_{x_k}^{\mathrm{fw}}(\alpha_k\xi_k)[\xi_k]\right\rangle_{x_{k+1}} - \langle\nabla f(x_k), \xi_k\rangle_{x_k} \\
=\ &\mathrm{D}(f\circ R_{x_k}^{\mathrm{fw}})(\alpha_k\xi_k)[\xi_k] - \mathrm{D}(f\circ R_{x_k}^{\mathrm{fw}})(0)[\xi_k] = \left\langle\nabla(f\circ R_{x_k}^{\mathrm{fw}})(\alpha_k\xi_k), \xi_k\right\rangle_{x_k} - \left\langle\nabla(f\circ R_{x_k}^{\mathrm{fw}})(0), \xi_k\right\rangle_{x_k} \\
=\ &\int_0^1 \left\langle\nabla^2(f\circ R_{x_k}^{\mathrm{fw}})(t\alpha_k\xi_k)[\xi_k], \xi_k\right\rangle_{x_k} \alpha_k \mathrm{d}t \leq L_h\alpha_k\|\xi_k\|_{x_k}^2.
\end{aligned} \tag{34}$$

Assumptions 2 and 4 yield

$$\begin{aligned}\left\langle \nabla f(x_{k+1}), \mathrm{D}R^{\mathrm{fw}}_{x_k}(\alpha_k\xi_k)[\xi_k]\right\rangle_{x_{k+1}} &\geq -\left\langle \nabla f(x_{k+1}), \alpha_k^{-1} {R^{\mathrm{bw}}_{x_{k+1}}}^{-1}(x_k)\right\rangle_{x_{k+1}} - C\left(\alpha_k+\alpha_k^2\right)\|\nabla f(x_{k+1})\|_{x_{k+1}}\|\xi_k\|^2_{x_k}\\ &\geq -\left\langle \nabla f(x_{k+1}), \alpha_k^{-1} {R^{\mathrm{bw}}_{x_{k+1}}}^{-1}(x_k)\right\rangle_{x_{k+1}} - CL_g\left(\alpha_k+\alpha_k^2\right)\|\xi_k\|^2_{x_k}. \end{aligned} \tag{35}$$

Combining (16), (34), and (35), we have

$$\begin{aligned} L_h\alpha_k\|\xi_k\|^2_{x_k} &\geq -\left\langle \nabla f(x_{k+1}), \alpha_k^{-1} {R^{\mathrm{bw}}_{x_{k+1}}}^{-1}(x_k)\right\rangle_{x_{k+1}} - CL_g\left(\alpha_k+\alpha_k^2\right)\|\xi_k\|^2_{x_k} - \langle \nabla f(x_k), \xi_k\rangle_{x_k}\\ &\geq (c_2-1)\langle \nabla f(x_k), \xi_k\rangle_{x_k} - CL_g\left(\alpha_k+\alpha_k^2\right)\|\xi_k\|^2_{x_k}, \end{aligned}$$

which implies

$$\alpha_k \geq \frac{2(c_2-1)\langle \nabla f(x_k), \xi_k\rangle_{x_k}}{(L_h+CL_g)\|\xi_k\|^2_{x_k} + \sqrt{(L_h+CL_g)^2\|\xi_k\|^4_{x_k} + 4(c_2-1)CL_g\|\xi_k\|^2_{x_k}\langle \nabla f(x_k), \xi_k\rangle_{x_k}}}.$$

This expression, together with $\|\nabla f(x_k)\|_{x_k} \leq \mu\|\xi_k\|_{x_k}$, yields

$$\alpha_k \geq \frac{2(c_2-1)\langle \nabla f(x_k), \xi_k\rangle_{x_k}}{\left(L_h+CL_g+\sqrt{(L_h+CL_g)^2+4(1-c_2)CL_g\mu}\right)\|\xi_k\|^2_{x_k}}.$$

By substituting the above inequality into (15) and summing the resultant expression, we have

$$f(x_{k+1}) \leq f(x_0) - \frac{2(1-c_2)}{L_h+CL_g+\sqrt{(L_h+CL_g)^2+4(1-c_2)CL_g\mu}}\sum_{i=0}^{k}\frac{\langle \nabla f(x_i), \xi_i\rangle^2_{x_i}}{\|\xi_i\|^2_{x_i}}. \tag{36}$$

As f is bounded below from Assumption 1, we obtain (33) by taking the limit of (36). □

We next discuss the existence of step lengths satisfying the modified Wolfe conditions. A key step is to represent the quantity $\left\langle \nabla f(R^{\mathrm{fw}}_x(\alpha\xi_x)), -\alpha^{-1}{R^{\mathrm{bw}}_{y(\alpha)}}^{-1}(x)\right\rangle_{R^{\mathrm{fw}}_x(\alpha\xi_x)}$ in the second Wolfe condition as the (partial) derivative of some function. To this end, we first require two technical lemmas.

Lemma 4 [28] *Let*

$$g(t) = \begin{cases} e^{-1/t}, & t>0,\\ 0, & t\leq 0.\end{cases}$$

Given any two real numbers r_1 and r_2 such that $r_1 < r_2$, the cutoff function $h(r_1, r_2, t)$ defined by

$$h(r_1, r_2, t) = \frac{g(t-r_1)}{g(t-r_1)+g(r_2-t)}$$

is smooth and satisfies $h(r_1, r_2, t) \equiv 0$ for $t \leq r_1$, $0 < h(r_1, r_2, t) < 1$ for $r_1 < t < r_2$, and $h(r_1, r_2, t) \equiv 1$ for $t \geq r_2$.

Lemma 5 *Denote $y(\alpha) = R^{\mathrm{fw}}_x(\alpha\xi_x)$ and $\eta_{y(\alpha)} = -\alpha^{-1}{R^{\mathrm{bw}}_{y(\alpha)}}^{-1}(x)$ for all $\alpha > 0$ such that ${R^{\mathrm{bw}}_{y(\alpha)}}^{-1}(x)$ is well-defined. Let U be a normal neighborhood of x and $V = \exp_x^{-1}(U)$. If U and V are sufficiently large, there exists a smooth map $\varsigma : \mathcal{D} \subset \mathbb{R}^2 \to \mathcal{M} : (\alpha, t) \mapsto \varsigma(\alpha, t)$ satisfying $\varsigma(\alpha, 0) = x$, $\varsigma(\alpha, \alpha) = y(\alpha)$, $\frac{\partial}{\partial t}\varsigma(\alpha, 0) = \xi_x$, and $\frac{\partial}{\partial t}\varsigma(\alpha, \alpha) = \eta_{y(\alpha)}$.*

Proof Intuitively, there are infinitely many maps satisfying the claim of the lemma. For rigorousness, we here construct a specific example of such a map ς. If U is sufficiently large to ensure $y(\alpha) \in U$, we can define

$$\delta_x(\alpha) = \alpha^{-1}\exp_x^{-1}(y(\alpha)) \text{ and } \zeta_x(\alpha) = \mathrm{D}\exp_x(\alpha\delta_x(\alpha))^{-1}[\eta_{y(\alpha)}],$$

where the second term is well-defined because the differential of a diffeomorphism is an isomorphism. We set a constant $\epsilon \in \left(0, \frac{1}{6}\right)$ and define

$$u_x(\alpha) = \frac{\delta_x(\alpha) - 2\epsilon\zeta_x(\alpha) - 2\epsilon\xi_x}{1-4\epsilon},$$

$$v_x(\alpha, t) = t\xi_x + h(\epsilon\alpha, 3\epsilon\alpha, t)[(t - 2\epsilon\alpha)(u_x(\alpha) - \xi_x)],$$

and

$$w_x(\alpha, t) = v_x(\alpha, t) + h(\alpha - 3\epsilon\alpha, \alpha - \epsilon\alpha, t)[(t - \alpha + 2\epsilon\alpha)\zeta_x(\alpha) + v_x(\alpha, \alpha - 2\epsilon\alpha) - v_x(\alpha, t)],$$

where h is the cutoff function defined in Lemma 4. It follows from Lemma 4 that $v_x(\alpha, t)$ is smooth and satisfies

$$v_x(\alpha, 0) = 0_x, \;\; v_x(\alpha, \alpha - 2\epsilon\alpha) = \alpha\delta_x(\alpha) - 2\epsilon\alpha\zeta_x(\alpha), \;\; \text{and} \;\; \frac{\partial}{\partial t} v_x(\alpha, 0) = \xi_x.$$

Using these conditions and Lemma 4, we see that $w_x(\alpha, t)$ is smooth and satisfies

$$w_x(\alpha, 0) = v_x(\alpha, 0) = 0_x, \;\; w_x(\alpha, \alpha) = 2\epsilon\alpha\zeta_x(\alpha) + v_x(\alpha, \alpha - 2\epsilon\alpha) = \alpha\delta_x(\alpha),$$

$$\frac{\partial}{\partial t} w_x(\alpha, 0) = \frac{\partial}{\partial t} v_x(\alpha, 0) = \xi_x, \;\; \text{and} \;\; \frac{\partial}{\partial t} w_x(\alpha, \alpha) = \zeta_x(\alpha).$$

Now, if $V = \exp_x^{-1}(U)$ is sufficiently large to ensure $w_x(\alpha, t) \in V$, we can define the map $\varsigma(\alpha, t) = \exp_x(w_x(\alpha, t))$. This ς satisfies all requirements. □

Corollary 2 *Let $y(\alpha)$, $\eta_{y(\alpha)}$, and ς be defined as in Lemma 5 and let $\{\gamma_\alpha\}$ be the family of smooth curves $\gamma_\alpha(t) = \varsigma(\alpha, t)$. Then, $\gamma_\alpha(0) = x$, $\gamma_\alpha(\alpha) = y(\alpha)$, $\dot\gamma_\alpha(0) = \xi_x$, and $\dot\gamma_\alpha(\alpha) = \eta_{y(\alpha)}$.*

Under the hypothesis of Lemma 5, i.e., that U and $V = \exp_x^{-1}(U)$ are sufficiently large, we define the following composite function for any smooth f:

$$\varphi(\alpha, t) = f(\varsigma(\alpha, t)) = f(\gamma_\alpha(t)).$$

Then, φ is smooth and satisfies

$$\varphi(\alpha, 0) = f(\gamma_\alpha(0)) = f(x),$$

$$\varphi(\alpha, \alpha) = f(\gamma_\alpha(\alpha)) = f(y(\alpha)) = f(R_x^{\text{fw}}(\alpha\xi_x)),$$

$$\frac{\partial}{\partial t}\varphi(\alpha, 0) = \langle \nabla f(\gamma_\alpha(t)), \dot\gamma_\alpha(t)\rangle_{\gamma_\alpha(t)}\Big|_{t=0} = \langle \nabla f(x), \xi_x\rangle_x,$$

and

$$\frac{\partial}{\partial t}\varphi(\alpha, \alpha) = \langle \nabla f(\gamma_\alpha(t)), \dot\gamma_\alpha(t)\rangle_{\gamma_\alpha(t)}\Big|_{t=\alpha} = \Big\langle \nabla f(y(\alpha)), \eta_{y(\alpha)}\Big\rangle_{y(\alpha)} = \Big\langle \nabla f(R_x^{\text{fw}}(\alpha\xi_x)), -\alpha^{-1} {R_{y(\alpha)}^{\text{bw}}}^{-1}(x)\Big\rangle_{R_x^{\text{fw}}(\alpha\xi_x)}.$$

If, in addition, f is bounded below and ξ_x is a descent direction, we can find the smallest value $t_1(\alpha)$ of t such that

$$\varphi(\alpha, t_1(\alpha)) = \varphi(\alpha, 0) + c_1 t_1(\alpha)\frac{\partial}{\partial t}\varphi(\alpha, 0),$$

where the left-hand side is well-defined if $V = \exp_x^{-1}(U)$ is sufficiently large. Clearly,

$$\varphi(\alpha, t) < \varphi(\alpha, 0) + c_1 t\frac{\partial}{\partial t}\varphi(\alpha, 0), \;\; \forall\, t < t_1(\alpha).$$

By the mean value theorem, we can also find the smallest value $t_2(\alpha)$ of $t \in (0, t_1(\alpha))$ such that

$$\frac{\partial}{\partial t}\varphi(\alpha, t_2(\alpha)) = c_1\frac{\partial}{\partial t}\varphi(\alpha, 0).$$

As φ is smooth and $\frac{\partial}{\partial t}\varphi(\alpha, 0) = \langle \nabla f(x), \xi_x\rangle_x < 0$, we have, for all sufficiently small α,

$$\frac{\partial}{\partial t}\varphi(\alpha, \alpha) < c_1\frac{\partial}{\partial t}\varphi(\alpha, 0).$$

Therefore, $t_2(\alpha) > \alpha$ for all sufficiently small $\alpha > 0$. Furthermore, if there exists $\bar\alpha > 0$ satisfying $\frac{\partial^2}{\partial t^2}\varphi(\bar\alpha, t_2(\bar\alpha)) \neq 0$, $t_2(\alpha)$ is a smooth function in some open interval around $\bar\alpha$, by the implicit function theorem and the smoothness of φ. Let $I_{\bar\alpha}$ be the largest such interval. We can then prove the existence of step lengths that satisfy the strong Wolfe conditions under reasonable regularity assumptions.

Proposition 2 *Suppose that f is smooth and bounded below, $\langle \nabla f(x), \xi_x \rangle_x < 0$, and that U and $V = \exp_x^{-1}(U)$ are sufficiently large, where U is a normal neighborhood of x. Suppose also that there is a number $\bar{\alpha} > 0$ that satisfies $t_2(\bar{\alpha}) > \bar{\alpha}$ and $\frac{\partial^2}{\partial t^2}\varphi(\bar{\alpha}, t_2(\bar{\alpha})) \neq 0$, ensuring $t_2(\alpha)$ is locally a smooth function of α. Furthermore, suppose that $\sup\{t_2(\alpha) : \alpha \in I_{\bar{\alpha}}\} \in I_{\bar{\alpha}}$, where φ, t_2, and $I_{\bar{\alpha}}$ are defined as above. Then, there is an interval for the value of α such that*

$$f(R_x^{\mathrm{fw}}(\alpha\xi_x)) \leq f(x) + c_1\alpha \langle \nabla f(x), \xi_x \rangle_x$$

and

$$\left| \left\langle \nabla f(R_x^{\mathrm{fw}}(\alpha\xi_x)), -\alpha^{-1} {R_{y(\alpha)}^{\mathrm{bw}}}^{-1}(x) \right\rangle_{R_x^{\mathrm{fw}}(\alpha\xi_x)} \right| \leq -c_2 \langle \nabla f(x), \xi_x \rangle_x .$$

Proof By the proposition hypotheses, by increasing α from $\bar{\alpha}$ to the right endpoint of $I_{\bar{\alpha}}$ (which may be $+\infty$), we can find an intersecting value α^* of α such that $\alpha^* = t_2(\alpha^*) < t_1(\alpha^*)$, which implies

$$\varphi(\alpha^*, \alpha^*) < \varphi(\alpha, 0) + c_1\alpha^* \frac{\partial}{\partial t}\varphi(\alpha^*, 0)$$

and

$$\frac{\partial}{\partial t}\varphi(\alpha^*, \alpha^*) = c_1 \frac{\partial}{\partial t}\varphi(\alpha^*, 0) > c_2 \frac{\partial}{\partial t}\varphi(\alpha^*, 0). \tag{37}$$

As the left-hand side of (37) is negative, we also have $\left|\frac{\partial}{\partial t}\varphi(\alpha^*, \alpha^*)\right| < -c_2 \frac{\partial}{\partial t}\varphi(\alpha^*, 0)$. Then, by continuity, there exists an interval around α^* for the value of α such that

$$\varphi(\alpha, \alpha) \leq \varphi(\alpha, 0) + c_1\alpha \frac{\partial}{\partial t}\varphi(\alpha, 0) \text{ and } \left|\frac{\partial}{\partial t}\varphi(\alpha, \alpha)\right| \leq -c_2 \frac{\partial}{\partial t}\varphi(\alpha, 0).$$

The above two inequalities are identical to the desired expressions. □

4.2 Global convergence of the Fletcher–Reeves method

In this subsection, we prove the global convergence property of Algorithm 1 when implemented with the Fletcher–Reeves formula (12). The final convergence theorem relies on the following crucial lemma.

Lemma 6 *Let $\{x_k\}$ be generated by Algorithm 1 implemented with the Fletcher–Reeves formula (12) and the strong Wolfe conditions (15) and (17) with $c_2 < \frac{1}{2}$. Then, for all k,*

$$-\frac{1}{1-c_2} \leq \frac{\langle \nabla f(x_k), \xi_k \rangle_{x_k}}{\|\nabla f(x_k)\|_{x_k}^2} \leq \frac{2c_2 - 1}{1 - c_2} \tag{38}$$

and

$$\|\nabla f(x_k)\|_{x_k} \leq \frac{1 - c_2}{1 - 2c_2} \|\xi_k\|_{x_k}. \tag{39}$$

Proof The proof is similar to that of Lemma 14 in [31], which imitates the original proof of Theorem 1 in [5] for Euclidean spaces. The result is obvious for $k = 0$ as $\xi_0 = -\nabla f(x_0)$. We perform an induction by noting from (9) and (12) that

$$\begin{aligned} \frac{\langle \nabla f(x_{k+1}), \xi_{k+1} \rangle_{x_{k+1}}}{\|\nabla f(x_{k+1})\|_{x_{k+1}}^2} &= -1 - \beta_k s_k \frac{\left\langle \nabla f(x_{k+1}), \alpha_k^{-1} {R_{x_{k+1}}^{\mathrm{bw}}}^{-1}(x_k) \right\rangle_{x_{k+1}}}{\|\nabla f(x_{k+1})\|_{x_{k+1}}^2} \\ &= -1 - s_k \frac{\left\langle \nabla f(x_{k+1}), \alpha_k^{-1} {R_{x_{k+1}}^{\mathrm{bw}}}^{-1}(x_k) \right\rangle_{x_{k+1}}}{\|\nabla f(x_k)\|_{x_k}^2}. \end{aligned}$$

This relation, together with (10) and (17), yields

$$-1 + c_2 \frac{\langle \nabla f(x_k), \xi_k \rangle_{x_k}}{\|\nabla f(x_k)\|_{x_k}^2} \leq \frac{\langle \nabla f(x_{k+1}), \xi_{k+1} \rangle_{x_{k+1}}}{\|\nabla f(x_{k+1})\|_{x_{k+1}}^2} \leq -1 - c_2 \frac{\langle \nabla f(x_k), \xi_k \rangle_{x_k}}{\|\nabla f(x_k)\|_{x_k}^2},$$

from which (38) follows by induction. Inequality (39) is a straightforward consequence of (38) and the Cauchy–Schwarz inequality. □

The global convergence theorem is established as follows.

Theorem 4 *Let $\{x_k\}$ be generated by Algorithm 1 implemented with the Fletcher–Reeves formula (12) and the strong Wolfe conditions (15) and (17) with $c_2 < \frac{1}{2}$. Suppose Assumptions 1–4 hold. Then,* $\liminf_{k\to\infty} \|\nabla f(x_k)\|_{x_k} = 0$.

Proof The proof is similar to that of Proposition 15 in [31], which imitates the original proof of Theorem 2 in [5] for Euclidean spaces. Multiplying (38) by $\frac{\|\nabla f(x_k)\|_{x_k}}{\|\xi_k\|_{x_k}}$, we obtain

$$-\frac{1}{1-c_2}\frac{\|\nabla f(x_k)\|_{x_k}}{\|\xi_k\|_{x_k}} \leq \frac{\langle \nabla f(x_k), \xi_k\rangle_{x_k}}{\|\nabla f(x_k)\|_{x_k}\|\xi_k\|_{x_k}} \leq \frac{2c_2-1}{1-c_2}\frac{\|\nabla f(x_k)\|_{x_k}}{\|\xi_k\|_{x_k}}.$$

This, together with Lemma 6 and Theorem 3, yields

$$\sum_{k=0}^{\infty} \frac{\|\nabla f(x_k)\|_{x_k}^4}{\|\xi_k\|_{x_k}^2} < \infty. \tag{40}$$

Furthermore, by (17) and the first inequality in (38), we have

$$\left|\left\langle \nabla f(x_{k+1}), \alpha_k^{-1} R_{x_{k+1}}^{\mathrm{bw}\ -1}(x_k)\right\rangle_{x_{k+1}}\right| \leq \frac{c_2}{1-c_2}\|\nabla f(x_k)\|_{x_k}^2. \tag{41}$$

Taking the squared norm of (9) and using (11), (12), and (41), we obtain

$$\begin{aligned}
\|\xi_{k+1}\|_{x_{k+1}}^2 &= \|\nabla f(x_{k+1})\|_{x_{k+1}}^2 + 2\beta_k s_k \left\langle \nabla f(x_{k+1}), \alpha_k^{-1} R_{x_{k+1}}^{\mathrm{bw}\ -1}(x_k)\right\rangle_{x_{k+1}} + \beta_k^2 s_k^2 \left\|\alpha_k^{-1} R_{x_{k+1}}^{\mathrm{bw}\ -1}(x_k)\right\|_{x_{k+1}}^2 \\
&\leq \|\nabla f(x_{k+1})\|_{x_{k+1}}^2 + \frac{2c_2}{1-c_2}\beta_k\|\nabla f(x_k)\|_{x_k}^2 + \beta_k^2\|\xi_k\|_{x_k}^2 \\
&= \frac{1+c_2}{1-c_2}\|\nabla f(x_{k+1})\|_{x_{k+1}}^2 + \frac{\|\nabla f(x_{k+1})\|_{x_{k+1}}^4 \|\xi_k\|_{x_k}^2}{\|\nabla f(x_k)\|_{x_k}^4}.
\end{aligned}$$

This implies the relation

$$\|\xi_k\|_{x_k}^2 \leq \frac{1+c_2}{1-c_2}\|\nabla f(x_k)\|_{x_k}^4 \sum_{i=0}^{k} \|\nabla f(x_i)\|_{x_i}^{-2}. \tag{42}$$

We now complete the proof by contradiction. Suppose there exists a constant $c > 0$ such that $\|\nabla f(x_k)\|_{x_k} \geq c$ for all k. It follows from (42) that $\|\xi_k\|_{x_k}^2 \leq \frac{1+c_2}{1-c_2}\|\nabla f(x_k)\|_{x_k}^4 \frac{k+1}{c^2}$, and therefore,

$$\sum_{k=0}^{\infty} \frac{\|\nabla f(x_k)\|_{x_k}^4}{\|\xi_k\|_{x_k}^2} \geq \frac{1-c_2}{1+c_2}\sum_{k=0}^{\infty}\frac{c^2}{k+1} = \infty,$$

which contradicts (40). Hence, $\liminf_{k\to\infty} \|\nabla f(x_k)\|_{x_k} = 0$. □

4.3 Global convergence of the Dai–Yuan method

In this subsection, we prove the global convergence property of Algorithm 1 implemented with the Dai–Yuan formula (14). The final convergence theorem relies on the following crucial lemma.

Lemma 7 *Let $\{x_k\}$ be generated by Algorithm 1 implemented with the Dai–Yuan formula (14) and the Wolfe conditions (15) and (16). Then, for all k,*

$$-\frac{1}{1-c_2}\|\nabla f(x_k)\|_{x_k}^2 \leq \langle \nabla f(x_k), \xi_k\rangle_{x_k} \leq -\frac{1}{1+c_3}\|\nabla f(x_k)\|_{x_k}^2 \tag{43}$$

and

$$\|\nabla f(x_k)\|_{x_k} \leq (1+c_3)\|\xi_k\|_{x_k}. \tag{44}$$

Proof We prove this lemma by induction. For $k = 0$, (43) holds obviously from $\xi_0 = -\nabla f(x_0)$. Assume (43) holds for k. Then, from (9) and (14), we have

$$\begin{aligned}\langle \nabla f(x_{k+1}), \xi_{k+1}\rangle_{x_{k+1}} &= \left\langle \nabla f(x_{k+1}), -\nabla f(x_{k+1}) - \beta_k s_k \alpha_k^{-1} R_{x_{k+1}}^{\mathrm{bw}\ -1}(x_k)\right\rangle_{x_{k+1}} \\ &= -\|\nabla f(x_{k+1})\|_{x_{k+1}}^2 + \frac{\|\nabla f(x_{k+1})\|_{x_{k+1}}^2 \left\langle \nabla f(x_{k+1}), s_k \alpha_k^{-1} R_{x_{k+1}}^{\mathrm{bw}\ -1}(x_k)\right\rangle_{x_{k+1}}}{\left\langle \nabla f(x_{k+1}), s_k \alpha_k^{-1} R_{x_{k+1}}^{\mathrm{bw}\ -1}(x_k)\right\rangle_{x_{k+1}} + \langle \nabla f(x_k), \xi_k\rangle_{x_k}} \\ &= -\frac{\|\nabla f(x_{k+1})\|_{x_{k+1}}^2 \langle \nabla f(x_k), \xi_k\rangle_{x_k}}{\left\langle \nabla f(x_{k+1}), s_k \alpha_k^{-1} R_{x_{k+1}}^{\mathrm{bw}\ -1}(x_k)\right\rangle_{x_{k+1}} + \langle \nabla f(x_k), \xi_k\rangle_{x_k}}.\end{aligned}$$

This, together with (10) and (16), indicates that (43) holds for $k+1$. Inequality (44) is a straightforward consequence of (43) and the Cauchy–Schwarz inequality. □

The global convergence theorem is established as follows.

Theorem 5 *Let $\{x_k\}$ be generated by Algorithm 1 implemented with the Dai–Yuan formula (14) and the Wolfe conditions (15) and (16). Suppose Assumptions 1–4 hold. Then,* $\liminf_{k\to\infty} \|\nabla f(x_k)\|_{x_k} = 0$.

Proof The proof is similar to that of Theorem 4.2 in [33], which imitates the original proof of Theorem 3.3 in [9] for Euclidean spaces. By Lemma 7 and Theorem 3, we have (33). It follows from (9) that

$$\xi_{k+1} + \nabla f(x_{k+1}) = -\beta_k s_k \alpha_k^{-1} R_{x_{k+1}}^{\mathrm{bw}\ -1}(x_k).$$

Taking the squared norm of the equation above, we have

$$\|\xi_{k+1}\|_{x_{k+1}}^2 = \beta_k^2 s_k^2 \left\|\alpha_k^{-1} R_{x_{k+1}}^{\mathrm{bw}\ -1}(x_k)\right\|_{x_{k+1}}^2 - 2\langle \nabla f(x_{k+1}), \xi_{k+1}\rangle - \|\nabla f(x_{k+1})\|_{x_{k+1}}^2.$$

This relation, together with (11), yields

$$\|\xi_{k+1}\|_{x_{k+1}}^2 \le \beta_k^2 \|\xi_k\|_{x_k}^2 - 2\langle \nabla f(x_{k+1}), \xi_{k+1}\rangle - \|\nabla f(x_{k+1})\|_{x_{k+1}}^2.$$

Dividing both sides by $\langle \nabla f(x_{k+1}), \xi_{k+1}\rangle_{x_{k+1}}^2$ and using (13), we obtain

$$\begin{aligned}\frac{\|\xi_{k+1}\|_{x_{k+1}}^2}{\langle \nabla f(x_{k+1}), \xi_{k+1}\rangle_{x_{k+1}}^2} &\le \frac{\|\xi_k\|_{x_k}^2}{\langle \nabla f(x_k), \xi_k\rangle_{x_k}^2} - \frac{2}{\langle \nabla f(x_{k+1}), \xi_{k+1}\rangle_{x_{k+1}}} - \frac{\|\nabla f(x_{k+1})\|_{x_{k+1}}^2}{\langle \nabla f(x_{k+1}), \xi_{k+1}\rangle_{x_{k+1}}^2} \\ &= \frac{\|\xi_k\|_{x_k}^2}{\langle \nabla f(x_k), \xi_k\rangle_{x_k}^2} - \left(\frac{1}{\|\nabla f(x_{k+1})\|_{k+1}} + \frac{\|\nabla f(x_{k+1})\|_{k+1}}{\langle \nabla f(x_{k+1}), \xi_{k+1}\rangle_{x_{k+1}}}\right)^2 + \frac{1}{\|\nabla f(x_{k+1})\|_{x_{k+1}}^2} \\ &\le \frac{\|\xi_k\|_{x_k}^2}{\langle \nabla f(x_k), \xi_k\rangle_{x_k}^2} + \frac{1}{\|\nabla f(x_{k+1})\|_{x_{k+1}}^2}.\end{aligned}$$

This implies the relation

$$\frac{\|\xi_k\|_{x_k}^2}{\langle \nabla f(x_k), \xi_k\rangle_{x_k}^2} \le \frac{\|\xi_0\|_{x_0}^2}{\langle \nabla f(x_0), \xi_0\rangle_{x_0}^2} + \sum_{i=1}^{k} \frac{1}{\|\nabla f(x_i)\|_{x_i}^2} = \sum_{i=0}^{k} \frac{1}{\|\nabla f(x_i)\|_{x_i}^2}. \tag{45}$$

We next complete the proof by contradiction. Suppose there exists a constant $c > 0$ such that $\|\nabla f(x_k)\|_{x_k} \ge c$ for all $k \ge 0$. It follows from (45) that $\frac{\|\xi_k\|_{x_k}^2}{\langle \nabla f(x_k), \xi_k\rangle_{x_k}^2} \le \frac{k+1}{c^2}$, and therefore,

$$\sum_{k=0}^{\infty} \frac{\langle \nabla f(x_k), \xi_k\rangle_{x_k}^2}{\|\xi_k\|_{x_k}^2} \ge \sum_{k=0}^{\infty} \frac{c^2}{k+1} = \infty,$$

which contradicts (33). Hence, $\liminf_{k\to\infty} \|\nabla f(x_k)\|_{x_k} = 0$. □

5 Implementation details of inverse retractions

In this section, we discuss implementation details of several practical inverse retractions on two specific matrix manifolds, i.e., the Stiefel and fixed-rank manifolds. The content of this section is almost a review of the relevant literature.

For general submanifolds of Euclidean spaces, one can always use the simple inverse orthographic retraction [4]

$$R_x^{\mathrm{or}-1}(y) = \mathrm{Proj}_{T_x\mathcal{M}}(y - x), \tag{46}$$

where R_x^{or} is the orthographic retraction [3]

$$R_x^{\mathrm{or}}(\xi_x) = \mathrm{Proj}_{(x+\xi_x+T_x^{\perp}\mathcal{M})\cap\mathcal{M}}(x + \xi_x).$$

Note that (46) is obviously different from the projective vector transport $\mathcal{T}_{\eta_x}^{\mathrm{pr}}(\xi_x) = \mathrm{Proj}_{T_{R_x(\eta_x)}\mathcal{M}}\xi_x$ introduced in [2].

5.1 Inverse retraction on the Stiefel manifold

For the Stiefel manifold

$$\mathrm{St}(n, r) = \{X \in \mathbb{R}^{n\times r} : X^\top X = I_r\},$$

using (46), the orthogonality of X, and the tangent projection representation [2]

$$\mathrm{Proj}_{T_X\mathrm{St}(n,r)}Z = Z - \frac{1}{2}X(X^\top Z + Z^\top X),$$

we can express the inverse orthographic retraction as

$$R_X^{\mathrm{or}-1}(Y) = \mathrm{Proj}_{T_X\mathrm{St}(n,r)}(Y - X) = \mathrm{Proj}_{T_X\mathrm{St}(n,r)}Y = Y - \frac{1}{2}X(X^\top Y + Y^\top X). \tag{47}$$

We next consider the inverse QR retraction. The QR retraction [2] is

$$R_X^{\mathrm{qr}}(\xi_X) = \mathrm{qf}(X + \xi_X),$$

where qf denotes the Q-factor of the QR factorization with positive diagonal elements in the R-factor. Referring to [26], we briefly introduce the computation of the inverse QR retraction. Suppose there exists a tangent vector $\xi_X \in T_X\mathrm{St}(n, r)$ such that $Y = \mathrm{qf}(X+\xi_X)$. We define $U = \mathrm{rf}(X+\xi_X)$, where rf is the R-factor of the QR factorization corresponding to qf. Then,

$$R_X^{\mathrm{qr}-1}(Y) = \xi_X = YU - X, \tag{48}$$

where the upper triangular U is unknown. Using (48), the orthogonality of X, and the tangent space representation

$$T_X\mathrm{St}(n, r) = \left\{\xi_X \in \mathbb{R}^{n\times r} : X^\top\xi_X + \xi_X^\top X = 0\right\}, \tag{49}$$

we have $X^\top YU + U^\top Y^\top X = 2I_r$. This equation can be solved efficiently via a simple recursive procedure.

We also consider the inverse Cayley-transform retraction. The Cayley-transform retraction [41] has the form

$$R_X^{\mathrm{ct}}(\xi_X) = \left(I_n - \frac{1}{2}(P_X\xi_X X^\top - X\xi_X^\top P_X)\right)^{-1}\left(I_n + \frac{1}{2}(P_X\xi_X X^\top - X\xi_X^\top P_X)\right)X, \tag{50}$$

where $P_X = I_n - \frac{1}{2}XX^\top$. Let $V_X = P_X\xi_X$. According to [36], if $I_r + X^\top Y$ is invertible, for any r-by-r symmetric matrix S, V_X of the form

$$V_X = 2Y(I_r + X^\top Y)^{-1} + XS \tag{51}$$

solves the equation

$$Y = R_X^{\mathrm{ct}}(P_X^{-1}V_X) = R_X^{\mathrm{ct}}(\xi_X). \tag{52}$$

Conversely, we discover that if $I_r + X^\top Y$ is invertible, any solution to (52) must have the form of (51). Let V_X and $\tilde{V}_X$ be two solutions to (52) and let $\Delta = V_X - \tilde{V}_X$. It follows from (50) and (52) that

$$\Delta(I_r + X^\top Y) - X\Delta^\top(X + Y) = 0. \tag{53}$$

We decompose Δ as $\Delta = XS + X_\perp K$, where $X_\perp \in \mathbb{R}^{n\times(n-r)}$ is an orthogonal complement to X, and substitute it into (53). Hence, we obtain

$$X(S - S^\top)(I_r + X^\top Y) + XK^\top X_\perp^\top Y + X_\perp K(I_r + X^\top Y) = 0.$$

As $I_r + X^\top Y$ is invertible, we deduce that $K = 0$ and S is symmetric. Next, using the orthogonality of X, $P_X^{-1} = I_n + XX^\top$, (49), and (51), we determine that, if $P_X^{-1}V_x \in T_X\mathrm{St}(n,r)$, then

$$S = (I_r + X^\top Y)^{-1} + (I_r + Y^\top X)^{-1} - 2I_r.$$

Therefore, we obtain the following closed-form expression for the inverse retraction:

$$R_X^{\mathrm{ct}-1}(Y) = \xi_X = 2Y(I_r + X^\top Y)^{-1} + 2X(I_r + Y^\top X)^{-1} - 2X. \tag{54}$$

One final point to note is that $I_r + X^\top Y$ with $Y = R_X^{\mathrm{ct}}(\xi_X)$ is invertible if ξ_X is sufficiently small.

5.2 Inverse retraction on the fixed-rank manifold

We next consider the fixed-rank manifold

$$\mathcal{M}_r^{m\times n} = \{X \in \mathbb{R}^{m\times n} : \mathrm{rank}(X) = r\}.$$

Suppose $X = USV^\top$, where $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$ are orthonormal and $S \in \mathbb{R}^{r\times r}$ is nonsingular. Then, according to [27] and [40], $T_X\mathcal{M}_r^{m\times n}$ and the projection onto it are given, respectively, by

$$T_X\mathcal{M}_r^{m\times n} = \left\{U\dot{S}V^\top + U_pV^\top + UV_p^\top : U^\top U_p = 0, V^\top V_p = 0, U_p \in \mathbb{R}^{m\times r}, V_p \in \mathbb{R}^{n\times r}, \dot{S} \in \mathbb{R}^{r\times r}\right\}, \tag{55}$$

and

$$\mathrm{Proj}_{T_X\mathcal{M}_r^{m\times n}}Z = ZVV^\top + UU^\top Z - UU^\top ZVV^\top.$$

From [4], if $\xi_X = \mathrm{Proj}_{T_X\mathcal{M}_r^{m\times n}}Z$ is in the form of (55), then

$$\dot{S} = U^\top ZV, \quad U_p = (I_m - UU^\top)ZV, \quad V_p = (I_n - VV^\top)Z^\top U.$$

It follows from [4] that the orthographic retraction has the form

$$R_X^{\mathrm{or}}(\xi_X) = \left(U(S + \dot{S}) + U_p\right)\left(S + \dot{S}\right)^{-1}\left((S + \dot{S})V^\top + V_p^\top\right) = U_+S_+V_+^\top, \tag{56}$$

where $U(S + \dot{S}) + U_p = U_+S_U$ and $V(S + \dot{S})^\top + V_p = V_+S_V$ are orthonormalizations and $S_+ = S_U(S + \dot{S})^{-1}S_V^\top$. Additionally, the inverse of (56) can be expressed as

$$R_X^{\mathrm{or}-1}(Y) = YVV^\top + UU^\top Y - UU^\top YVV^\top - X,$$

which is equivalent to

$$R_X^{\mathrm{or}-1}(Y) = U\left(U^\top U_Y S_Y V_Y^\top V - S\right)V^\top + \left((I_m - UU^\top)U_Y S_Y V_Y^\top V\right)V^\top + U\left(U^\top U_Y S_Y V_Y^\top (I_n - VV^\top)\right) \tag{57}$$

if $Y = U_Y S_Y V_Y^\top$. Note that (57) is in the form of (55).

6 Numerical experiments

In this section, we report preliminary numerical results that show the potential usefulness of Riemannian CG with inverse retraction. The following test problems were considered: the Brockett cost function minimization problem and the joint diagonalization problem over the Stiefel manifold and the matrix completion problem over the fixed-rank manifold. All the experiments were performed in MATLAB and our code is available at http://www.optimization-online.org/DB_HTML/2020/05/7798.html.

6.1 The Brockett cost function minimization problem

Our first test problem for the Stiefel manifold was the Brockett cost function minimization problem [2]

$$\min_{X\in\mathbb{R}^{n\times r}} \operatorname{tr}(X^\top AXN) \ \text{ s.t. } \ X^\top X = I_r, \tag{58}$$

where $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix and $N \in \mathbb{R}^{r\times r}$ is a positive diagonal matrix. In this problem, the objective function is $f(X) = \operatorname{tr}(X^\top AXN)$ and its Euclidean gradient is $\mathcal{G}(X) = 2AXN$. In our experiments, the data matrix A and the initial iterate X_0 were randomly generated by the commands

$$\texttt{A} = \texttt{randn(n)},\ \texttt{A} = \texttt{A} + \texttt{A}',\ \texttt{N} = \texttt{diag(1 : r)},\ \texttt{X0} = \texttt{orth(randn(n, r))},$$

where $n = 100$ and $r = 5$.

We compared nine Riemannian CG algorithms: `Alg1.InvOR`, `Alg1.InvQR`, `Alg1.InvCT`, `RCGstd1`, `RCGstd2`, `Manopt.InvOR`, `Manopt.InvQR`, and `Manopt.InvCT`. The first three algorithms were implementations of Algorithm 1. `RCGstd1` and `RCGstd2` were standard Riemannian CG algorithms, which were identical to the first three algorithms in all respects except that inverse retraction was replaced by vector transport. `Manopt` was the default Riemannian CG algorithm in Manopt version 5.0 [7]. The last three algorithms were modifications of `Manopt`, which were identical to `Manopt` in all respects except that vector transport was replaced by inverse retraction. Detailed description of these algorithms is shown in Table 1.

Table 1: Description of algorithms applied to problem (58)

Algorithm	Retraction	Inverse retraction	Vector transport	Line search	β_k
`Alg1.InvOR`	QR	inverse orthographic	–	strong Wolfe	DY
`Alg1.InvQR`	QR	inverse QR	–	strong Wolfe	DY
`Alg1.InvCT`	QR	inverse Cayley	–	strong Wolfe	DY
`RCGstd1`	QR	–	differentiated QR	strong Wolfe	DY
`RCGstd2`	projective	–	projective	strong Wolfe	DY
`Manopt`	default	–	default	default	default
`Manopt.InvOR`	default	inverse orthographic	–	default	default
`Manopt.InvQR`	default	inverse QR	–	default	default
`Manopt.InvCT`	default	inverse Cayley	–	default	default

For the first five algorithms, the stopping criterion was

$$\frac{\|\nabla f(X_k)\|_{\mathrm{F}}}{\|\nabla f(X_0)\|_{\mathrm{F}}} \le 10^{-10} \ \text{ or } \ \frac{|f(X_k) - f(X_{k-1})|}{|f(X_k)|} \le 10^{-20}.$$

Furthermore, the initial step length of each iteration was set according to Section 3.5 in [29] as

$$\alpha_k^{\text{initial}} = \min\left\{\max\left\{\alpha_{k-1}\frac{\operatorname{tr}(\nabla f(X_{k-1})^\top \xi_{k-1})}{\operatorname{tr}(\nabla f(X_k)^\top \xi_k)},\ 10^{-8}\right\},\ 10^4\right\},$$

the line search procedure was that indicated in Figure 2.5.3 of [37], and the parameters were $c_1 = 10^{-8}$ and $c_2 = 0.75$.

Numerical results corresponding to the averages of 20 random runs are reported in Table 2 and the corresponding average gradient norm histories are shown in Figure 2. One can observe that `Alg1.InvOR`, `Alg1.InvQR`, and `Alg1.InvCT` were slightly superior to their counterparts `RCGstd1` and `RCGstd2` and that `Manopt.InvOR`, `Manopt.InvQR`, and `Manopt.InvCT` exhibited similar performance to their counterpart `Manopt`.

6.2 The joint diagonalization problem

The second test problem involving the Stiefel manifold considered in this work was the joint diagonalization problem [8, 38]

$$\max_{X\in\mathbb{R}^{n\times r}} \sum_{j=1}^{m} \left\|\operatorname{diag}(X^\top A_j X)\right\|_{\mathrm{F}}^2 \ \text{ s.t. } \ X^\top X = I_r, \tag{59}$$

Table 2: Numerical results of various algorithms applied to problem (58)

Algorithm	# Iterations	Function value	Norm of gradient	Time [s]
`Alg1.InvOR`	357.55	−390.235681739689	1.454967×10^{-5}	0.0718
`Alg1.InvQR`	374.70	−390.235681739691	1.167126×10^{-5}	0.1233
`Alg1.InvCT`	362.90	−390.235681739686	1.984884×10^{-5}	0.0980
`RCGstd1`	530.75	−390.235681739688	1.795625×10^{-5}	0.1025
`RCGstd2`	553.50	−390.235681739685	2.274411×10^{-5}	0.1204
`Manopt`	347.85	−390.235681739687	5.154200×10^{-6}	1.0458
`Manopt.InvOR`	342.20	−390.235681739686	5.160849×10^{-6}	1.0173
`Manopt.InvQR`	352.30	−390.235681739686	6.349105×10^{-6}	1.1029
`Manopt.InvCT`	349.15	−390.235681739685	5.848011×10^{-6}	1.0454

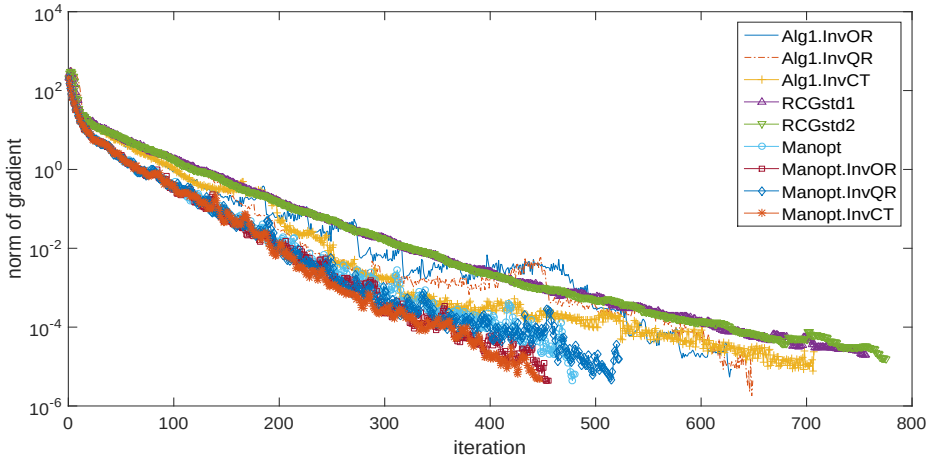

Figure 2: Average gradient norm histories of various algorithms applied to problem (58)

where all $A_j \in \mathbb{R}^{n\times n}$ are symmetric. In this problem, the objective function is $f(X) = -\sum_{j=1}^{m} \left\|\mathrm{diag}(X^\top A_j X)\right\|_{\mathrm{F}}^2$ and its Euclidean gradient is $\mathcal{G}(X) = -4\sum_{j=1}^{m} A_j X \mathrm{diag}(X^\top A_j X)$. In our experiments, the data matrices A_j and the initial iterate X_0 were randomly generated by the commands

$$\mathtt{B_j = randn(n),\ A_j = diag(randn(1,n).\hat{}2) + 0.1 * (B_j + B_j'),\ X0 = orth(randn(n,r))},$$

where $m = 100$ and $n = r = 20$.

The test algorithms for problem (59) were identical to those for problem (58), with parameters unchanged. Numerical results corresponding to the averages of 20 random runs are reported in Table 3 and the corresponding average gradient norm histories are shown in Figure 3. One can observe that the algorithms using inverse retraction exhibited comparable performance to those of their counterparts using traditional vector transports.

6.3 The low-rank matrix completion problem

The test problem for the fixed-rank manifold considered in this work was the low-rank matrix completion problem formulated as [40]

$$\min_{X\in\mathbb{R}^{m\times n}} \frac{1}{2}\|\mathrm{P}_\Omega(X - A)\|_{\mathrm{F}}^2 \ \text{ s.t. } \ \mathrm{rank}(X) = r, \tag{60}$$

Table 3: Numerical results of various algorithms applied to problem (59)

Algorithm	# Iterations	Function value	Norm of gradient	Time [s]
`Alg1.InvOR`	30.00	−6091.50260201438	1.923563×10^{-5}	0.2217
`Alg1.InvQR`	32.00	−6091.50260201438	1.464818×10^{-5}	0.2833
`Alg1.InvCT`	31.85	−6091.50260201437	2.441115×10^{-5}	0.2455
`RCGstd1`	32.65	−6091.50260201437	2.980332×10^{-5}	0.2525
`RCGstd2`	32.50	−6091.50260201437	3.308454×10^{-5}	0.2789
`Manopt`	31.45	−6091.50260201438	3.232603×10^{-5}	0.2890
`Manopt.InvOR`	30.75	−6091.50260201438	3.675166×10^{-5}	0.2845
`Manopt.InvQR`	31.25	−6091.50260201438	3.149593×10^{-5}	0.3115
`Manopt.InvCT`	31.25	−6091.50260201438	3.226247×10^{-5}	0.2892

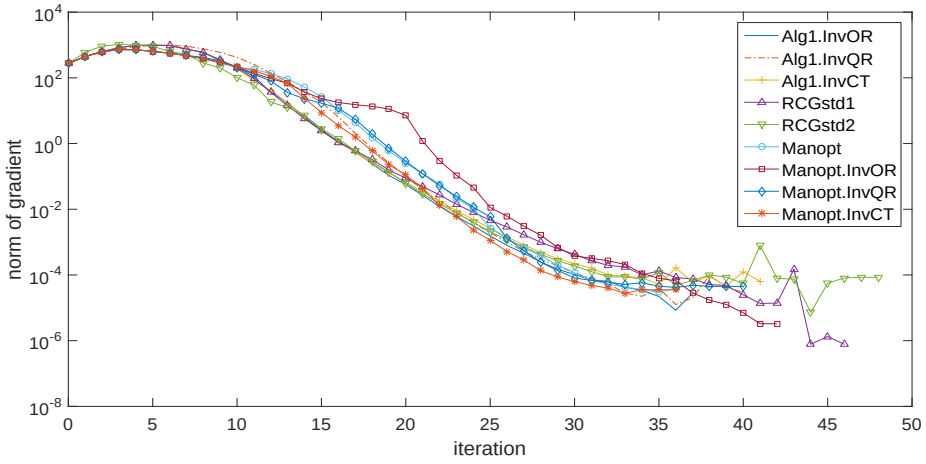

Figure 3: Average gradient norm histories of various algorithms applied to problem (59)

where $A \in \mathbb{R}^{m\times n}$ is a real matrix, Ω is a subset of $\{1,\dots,m\}\times\{1,\dots,n\}$, and

$$\mathrm{P}_\Omega : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n} : M_{ij} \mapsto \begin{cases} M_{ij}, & (i,j) \in \Omega, \\ 0, & (i,j) \notin \Omega. \end{cases}$$

In this problem, the objective function is $f(X) = \frac{1}{2}\|\mathrm{P}_\Omega(X-A)\|_{\mathrm{F}}^2$ and its Euclidean gradient is $\mathcal{G}(X) = \mathrm{P}_\Omega(X-A)$. In this work, the data matrix A and the initial iterate (X_0, U_0, S_0, V_0) were randomly generated by the commands

$$\texttt{A} = \texttt{randn(m,r)} * \texttt{randn(n,r)}', \ \texttt{[U0,S0,V0]} = \texttt{svds(PA,r)}, \ \texttt{X0} = \texttt{U0} * \texttt{S0} * \texttt{V0}',$$

where $\texttt{PA} = \mathrm{P}_\Omega(A)$, $m = n = 2000$, and $r = 20$. The sampling ratio $\rho = \mathrm{Prob}\{(i,j) \in \Omega\} = \mathrm{E}\left(\frac{|\Omega|}{mn}\right)$ was set to be identical to that of Manopt, i.e., $\rho = \frac{4r(m+n-r)}{mn}$, which is said to ensure unique recovery of A [40].

We compared four Riemannian CG algorithms: `Alg1`, `RCGstd`, `Manopt`, and `Manopt.InvRetr`. `Alg1` was an implementation of Algorithm 1. `RCGstd` was a standard Riemannian CG algorithm, which was identical to `Alg1` in all respects except that inverse retraction was replaced by vector transport. `Manopt` was the default Riemannian CG algorithm in Manopt version 5.0. `Manopt.InvRetr` was a modification of `Manopt`, which was identical to `Manopt` in all respects except that vector transport was replaced by inverse retraction. Detailed description of these algorithms is shown in Table 4.

For `Alg1` and `RCGstd`, the stopping criterion was $\|\nabla f(X_k)\|_{\mathrm{F}} \le 10^{-6}$, the parameters were $c_1 = 10^{-8}$ and $c_2 = 0.5$, and the initial step length of each iteration was set according to Section 3 in [40], as

$$\alpha_k^{\text{initial}} = \min\left\{\max\left\{-\frac{\mathrm{tr}(\nabla f(X_k)^\top \xi_k)}{\mathrm{tr}(\mathrm{P}_\Omega(\xi_k)^\top \mathrm{P}_\Omega(\xi_k))},\ 10^{-8}\right\},\ 10^4\right\}.$$

Table 4: Description of algorithms applied to problem (60)

Algorithm	Retraction	Inverse retraction	Vector transport	Line search	β_k
`Alg1`	orthographic	inverse orthographic	–	strong Wolfe	FR
`RCGstd`	orthographic	–	projective	strong Wolfe	FR
`Manopt`	default	–	default	default	default
`Manopt.InvRetr`	default	inverse orthographic	–	default	default

Numerical results corresponding to the averages of 10 random runs are reported in Table 5 and the corresponding average gradient norm histories are shown in Figure 4. One can observe that `Alg1` was superior to its counterpart `RCGstd` and that `Manopt.InvRetr` exhibited almost identical performance to its counterpart `Manopt`.

Table 5: Numerical results of various algorithms applied to problem (60)

Algorithm	# Iterations	Function value	Norm of gradient	Time [s]
`Alg1`	37.1	$6.06736638223860 \times 10^{-12}$	8.286083×10^{-7}	12.4800
`RCGstd`	46.8	$4.91138125836903 \times 10^{-12}$	8.375996×10^{-7}	15.1866
`Manopt`	34.5	$6.08686781817038 \times 10^{-12}$	8.026180×10^{-7}	7.1001
`Manopt.InvRetr`	34.5	$6.08916568422615 \times 10^{-12}$	8.027833×10^{-7}	7.0993

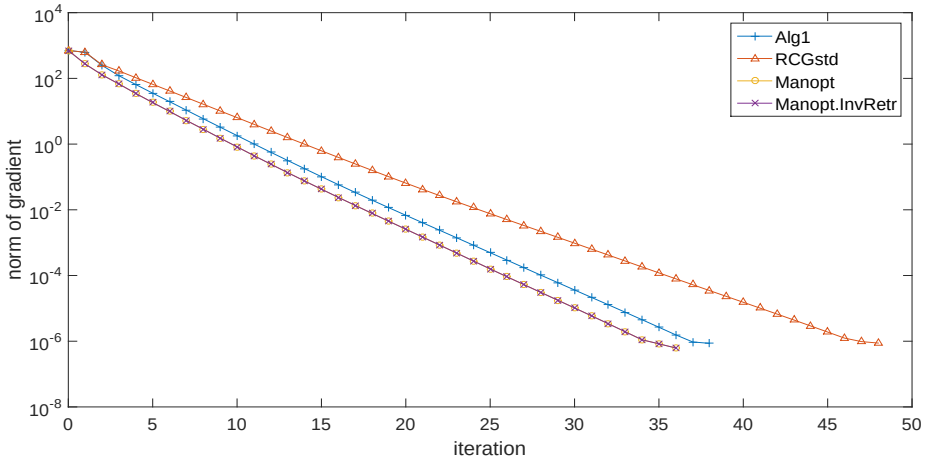

Figure 4: Average gradient norm histories of various algorithms applied to problem (60)

7 Conclusions

In this study, we developed a new class of Riemannian CG methods. Our purpose was to establish inverse retraction as a competitive alternative to the vector transport used in existing methods. The advantages of inverse retraction are its variety and easy implementation. In addition, various choices for inverse retraction exist, as the backward retraction associated with an inverse retraction can differ from the forward retraction.

The main contribution of this work is search direction construction by employing inverse retraction instead of vector transport; this was accomplished via the newly proposed Riemannian CG approach. We demonstrated both theoretically and experimentally that inverse retraction can constitute a competitive alternative to classical vector transports such as the differentiated retraction and orthogonal projection. In the theoretical context, we

proposed modified Riemannian Wolfe conditions associated with inverse retraction and proved the global convergence properties of the new methods. In numerical experiments, we compared the new and classical methods and implemented several inverse retractions in Manopt. Numerical results show that the new methods have comparable performance to their classical counterparts.

As inverse retraction is nonlinear, further research on numerical stability of inverse retraction for large-scale problems is needed to improve the robustness of the new methods. We will also focus on more applications of the new methods in future.

Acknowledgements

X. Zhu is supported by National Natural Science Foundation of China Grant Number 11601317 and H. Sato is supported by JSPS KAKENHI Grant Number JP20K14359. The authors are grateful to the coordinating editor and two anonymous referees for their valuable comments and suggestions.

References

[1] Absil, P.-A., Baker, C. G., Gallivan, K. A.: Trust-region methods on Riemannian manifolds. Found. Comput. Math. 7, 303–330 (2007)

[2] Absil, P.-A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton, NJ (2008)

[3] Absil, P.-A., Malick, J.: Projection-like retractions on matrix manifolds. SIAM J. Optim. 22, 135–158 (2012)

[4] Absil, P.-A., Oseledets, I. V.: Low-rank retractions: a survey and new results, Comput. Optim. Appl. 62, 5–29 (2015)

[5] Al-Baali, M.: Descent property and global convergence of the Fletcher–Reeves method with inexact line search, IMA J. Numer. Anal. 5, 121–124 (1985)

[6] Bonnabel, S.: Stochastic gradient descent on Riemannian manifolds. IEEE Trans. Autom. Control 58, 2217–2229 (2013)

[7] 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)

[8] Cardoso, J. F., Souloumiac, A.: Blind beamforming for non-Gaussian signals. In IEEE Proceedings F-Radar and Signal Processing, 140, 362–370, (1993)

[9] Dai, Y., Yuan, Y.: A nonlinear conjugate gradient method with a strong global convergence property. SIAM J, Optim. 10, 177–182 (1999)

[10] do Carmo, M. P.: Riemannian Geometry. Translated from the second Portuguese edition by Francis Flaherty. Mathematics: Theory & Applications. Birkhäuser Boston Inc., Boston, MA (1992)

[11] Edelman, A., Arias, T. A., Smith, S. T.: The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. 20, 303–353 (1998)

[12] Fletcher, R., Reeves, C. M.: Function minimization by conjugate gradients. Comput. J. 7, 149–154 (1964)

[13] 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, Springer, London, pp. 281–293 (2012)

[14] Grohs, P., Hosseini, S.: ε-subgradient algorithms for locally Lipschitz functions on Riemannian manifolds. Adv. Comput. Math. 42, 333–360 (2016)

[15] Hosseini, S., Huang, W., Yousefpour, R.: Line search algorithms for locally Lipschitz functions on Riemannian manifolds. SIAM J. Optim. 28, 596–619 (2018)

[16] Hosseini, S., Uschmajew, A.: A Riemannian gradient sampling algorithm for nonsmooth optimization on manifolds. SIAM J. Optim. 27, 173–189 (2017)

[17] Hosseini, S., Uschmajew, A.: A gradient sampling method on algebraic varieties and application to nonsmooth low-rank optimization. SIAM J. Optim. 29, 2853–2880 (2019)

[18] Hu, J., Liu, X., Wen, Z., Yuan, Y.: A brief introduction to manifold optimization. J. Oper. Res. Soc. China, published online, DOI: 10.1007/s40305-020-00295-9

[19] Huang, W.: Optimization algorithms on Riemannian manifolds with applications. Ph.D. thesis, Department of Mathematics, Florida State University (2013)

[20] Huang, W., Absil, P.-A., Gallivan, K. A.: A symmetric rank-one trust-region method. Math. Program. 150, 179–216 (2015)

[21] Huang, W., Absil, P.-A., Gallivan, K. A.: A Riemannian BFGS method without differentiated retraction for nonconvex optimization problems. SIAM J. Optim. 28, 470–495 (2018)

[22] Huang, W., Gallivan, K. A., Absil, P.-A.: A Broyden class of quasi-Newton methods for Riemannian optimization. SIAM J. Optim. 25, 1660–1685 (2015)

[23] Huang, W., Wei, K.: Extending FISTA to Riemannian optimization for sparse PCA. arXiv:1909.05485v1 (2019)

[24] Huang, W., Wei, K.: Riemannnian proximal gradient methods. arXiv:1909.06065v1 (2019)

[25] Jiang, B., Ma, S., Man-Cho So, A., Zhang, S.: Vector transport-free SVRG with general retraction for Riemannian optimization: complexity analysis and practical implementation. arXiv:1705.09059v1 (2017)

[26] Kaneko, T., Fiori, S., Tanaka, T.: Empirical arithmetic averaging over the compact Stiefel manifold, IEEE Transactions on Signal Processing 61, 883–894 (2013)

[27] Koch, O., Lubich, C.: Dynamical low-rank approximation. SIAM J. Matrix Anal. Appl. 29, 434–454 (2007)

[28] Lee, J. M.: Introduction to Smooth Manifolds, 2nd ed., Springer, New York, NY (2012)

[29] Nocedal, J., Wright, S. J.: Numerical Optimization, 2nd ed., Springer, New York, NY (2006)

[30] Qi, C., Gallivan, K. A., Absil, P.-A.: Riemannian BFGS algorithm with applications. In: Recent Advances in Optimization and its Applications in Engineering, Springer-Verlag Berlin, Heidelberg, pp. 183–192 (2010)

[31] Ring, W., Wirth, B.: Optimization methods on Riemannian manifolds and their application to shape space. SIAM J. Optim. 22, 596–627 (2012)

[32] Rockafellar, R. T., Wets, J.-B.: Variational Analysis. Springer, Berlin, Heidelberg (1998)

[33] Sato, H.: A Dai–Yuan-type Riemannian conjugate gradient method with the weak Wolfe conditions. Comput. Optim. Appl. 64, 101–118 (2016)

[34] Sato, H., Iwai, T.: A new, globally convergent Riemannian conjugate gradient method. Optimization 64, 1011–1031 (2015)

[35] Sato, H., Kasai, H., Mishra, B.: Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport. SIAM J. Optim. 29, 1444–1472 (2019)

[36] Siegel, J. W.: Accelerated optimization with orthogonality constraints. arXiv:1903.05204v3 (2019)

[37] Sun, W., Yuan, Y.: Optimization Theory and Methods: Nonlinear Programming. Springer, Boston, MA (2006)

[38] Theis, F. J., Cason, T. P., Absil, P.-A.: Soft dimension reduction for ICA by joint diagonalization on the Stiefel manifold. In: Adali, T., Jutten, C., Romano, J.M.T., Barros, A. (eds.) Independent Component Analysis and Signal Separation. Lecture Notes in Computer Science, vol. 5441, pp. 354–361. Springer, Berlin (2009)

[39] Tripuraneni, N., Flammarion, N., Bach, F., Jordan, M. I.: Averaging stochastic gradient descent on Riemannian manifolds. PMLR 75, 1–38 (2018)

[40] Vandereycken, B.: Low-rank matrix completion by Riemannian optimization. SIAM J. Optim. 23, 1214–1236 (2013)

[41] Wen, Z., Yin, W.: A feasible method for optimization with orthogonality constraints. Math. Program. 142, 397–434 (2013)

[42] Zhang, H., Reddi, S. J., Sra, S.: Fast stochastic optimization on Riemannian manifolds. arXiv:1605.07147v1 (2016)

[43] Zhang, H., Sra, S.: First-order methods for geodesically convex optimization. JMLR: Workshop and Conference Proceedings 49, 1–22 (2016)

[44] Zhang, H., Sra, S.: An estimate sequence for geodesically convex optimization. PMLR 75, 1–21 (2018)