A Riemannian conjugate gradient method for optimization on the Stiefel manifold *

Xiaojing Zhu †

Abstract In this paper we propose a new Riemannian conjugate gradient method for optimization on the Stiefel manifold. We introduce two novel vector transports associated with the retraction constructed by the Cayley transform. Both of them satisfy the Ring-Wirth nonexpansive condition, which is fundamental for convergence analysis of Riemannian conjugate gradient methods, and one of them is also isometric. It is known that the Ring-Wirth nonexpansive condition does not hold for traditional vector transports as the differentiated retractions of QR and polar decompositions. Practical formulae of the new vector transports for low-rank matrices are obtained. Dai's nonmonotone conjugate gradient method is generalized to the Riemannian case and global convergence of the new algorithm is established under standard assumptions. Numerical results on a variety of low-rank test problems demonstrate the effectiveness of the new method.

Keywords Riemannian optimization, Stiefel manifold, conjugate gradient method, retraction, vector transport, Cayley transform

Mathematics Subject Classification (2010) 49M37 49Q99 65K05 90C30

1 Introduction

In this paper, we focus on optimization on the Stiefel manifold

$$\begin{array}{ll} \min\limits_{X\in\mathbb{R}^{n\times p}} & f(X) \\ \text{s.t.} & X^\top X = I, \end{array} \tag{1}$$

where the feasible set

$$\mathrm{St}(n,p) = \{X \in \mathbb{R}^{n\times p} \mid X^\top X = I\} \tag{2}$$

with $p \le n$ is referred to as the Stiefel manifold [31], an $np - \frac{1}{2}p(p+1)$-dimensional embedded submanifold of the vector space $\mathbb{R}^{n\times p}$. Problem (1) has wide applications such as the linear eigenvalue problem [10, 27, 36], the orthogonal Procrustes problem [11], the joint diagonalization problem [32], the nearest low-rank correlation matrix problem [12, 19, 42], the Kohn-Sham total energy minimization [21, 34, 39], heterogeneous quadratics minimization [3], singular value decomposition [20, 29] and trace ratio optimization [22, 40, 41]. For more applications, see [1, 9, 14, 35] and references therein.

Problem (1) has attracted much attention in the optimization community due to its wide applications and computational difficulty. The recent developed technique Riemannian optimization is a mainstream approach to solving this problem. Riemannian optimization refers to optimization on Riemannian manifolds. It takes full advantage of the geometric structure of feasible sets to construct efficient constraint-preserving iterative schemes. Classical

*This research is supported by National Natural Science Foundation of China (Nos.11601317 and 11526135) and University Young Teachers' Training Scheme of Shanghai (No.ZZsdl15124).

†Corresponding author (X. Zhu)
College of Mathematics and Physics, Shanghai University of Electric Power, Shanghai 200090, China
E-mail address: 042563zxj@163.com

methods for unconstrained optimization in the Euclidean space [24], including the steepest descent method, Newton's method, conjugate gradient methods, quasi Newton methods and trust region methods, can be generalized to optimization on Riemannian manifolds [1, 14, 15, 17, 25, 28, 30]. A framework of Riemannian optimization has two stages in each iteration: (i) Find a tangent vector as the search direction or the trial step; (ii) Invoke a retraction that maps a tangent vector to a point on the manifold. A retraction can be viewed as a first-order approximation to the Riemannian exponential mapping [8]. For instance, the polar decomposition and the QR decomposition are two commonly used way to construct retractions on the Stiefel manifold [1]. The former is actually to project a tangent vector onto the manifold [2], while the latter is a practical approach preferred in many real applications such as the Kohn-Sham total energy minimization in electronic structure calculations [39]. For a comprehensive understanding of Riemannian optimization, see [1, 14] for example. Aiming at the Stiefel manifold (2), many practical Riemannian optimization algorithms have been developed. The algorithms proposed by Wen and Yin [35] and by Jiang and Dai [18] are two state-of-the-art algorithms. They can be viewed as Riemannian gradient-type methods on the Stiefel manifold. Wen and Yin used a low-rank refinement of the Cayley transform introduced in [23] as a retraction. Jiang and Dai proposed a framework of retraction-like constraint-preserving update schemes via decomposition in null and range spaces.

This paper is devoted to a Riemannian conjugate gradient method for solving problem (1). It is well known that in the Euclidean space, the conjugate gradient method generally outperforms the steepest descent method for its faster convergence and is more suitable than second-order methods such as Newton's method, quasi-Newton methods and trust region methods for large-scale problems. In Riemannian optimization, the conjugate gradient method still remains in the developmental stage. Absil, Mahony and Sepulchre introduced in their book [1] the concept of vector transports to implement the conjugate gradient method on Riemannian manifolds. This concept stems from the parallel translation of tangent fields [8]. A commonly used way to construct vector transports is the differentiated retraction operator. Ring and Wirth proved global convergence of the Riemannian Fletcher-Reeves conjugate gradient method under the assumption that the norm of a tangent vector does not increase after being transported. In what follows, we call this assumption the Ring-Wirth nonexpansive condition. To get rid of this assumption in global convergence analysis, Sato and Iwai introduced the notion of a scaled vector transport [30]. Recently, Sato has applied this new vector transport to the Riemannian Dai-Yuan method [6, 28]. However, this scaled vector transport violates the property of linearity. Later, an anonymous referee brought to our attention the simultaneous work [16] on intrinsic representation of vector transports on matrix manifolds including the Stiefel manifold. This new type of vector transport is isometric and cheap but not traditional because of no requirement of an associated retraction. Moreover, it has not a simple closed-form representation.

The proposed method for problem (1) has two main contributions. First, we implement the Cayley transform as a retraction and consequently introduce two novel associated vector transports. The first one is the differentiated retraction of the Cayley transform and the second one approximates the differentiated matrix exponential by the Cayley transform. We show that both of them satisfy the Ring-Wirth nonexpansive condition and the latter is also isometric. These properties are crucial for global convergence in Riemannian conjugate gradient methods. Note that most existing vector transports such as differentiated retractions of the QR decomposition and the polar decomposition do not satisfy the Ring-Wirth nonexpansive condition. Second, we generalize Dai's nonmonotone conjugate gradient method [5] from the Euclidean space to Riemannian manifolds. To the best of our knowledge, only monotone conjugate gradient methods that use the (strong) Wolfe conditions have been studied in Riemannian optimization so far. Numerical results have shown that nonmonotone techniques can greatly improve the efficiency of gradient-type methods such as [18, 35]. This motivates us to exploit the potentialities of nonmonotone techniques in Riemannian conjugate gradient methods.

This paper is organized as follows. Preliminaries on Riemannian conjugate gradient methods are given in Section 2. In Section 3, we propose two novel vector transports associated with the retraction constructed by the Cayley transform and then generalize Dai's monmonotone conjugate gradient method using the new vector transports. We establish global convergence for the new Riemannian conjugate gradient algorithm in a general setting in Section 4. Numerical results are shown in Section 5 and conclusions are made in the last section.

2 Preliminaries on Riemannian conjugate gradient methods

In this section, we give some important preliminaries on Riemannian conjugate gradient methods. Riemannian optimization refers to minimizing a function $f(x)$ on a Riemannian manifold $\mathcal{M}$. A general iteration of Riemannian

optimization has the form

$$x_{k+1} = R_{x_k}(\alpha_k \eta_k), \tag{3}$$

where $\eta_k \in T_{x_k}\mathcal{M}$ is a tangent vector and R is a retraction. The definition of a retraction is given as follows.

Definiton 1 [1] *A retraction on a manifold $\mathcal{M}$ is a smooth mapping R from the tangent bundle $T\mathcal{M}$ onto $\mathcal{M}$ with the following properties. Let R_x denote 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 $\mathrm{id}_{T_x\mathcal{M}}$ the identity mapping on $T_x\mathcal{M}$.

The construction of a conjugate gradient direction in the Euclidean space has the form

$$\eta_{k+1} = -\nabla f(x_{k+1}) + \beta_{k+1}\eta_k.$$

This formula is meaningless for Riemannian manifolds since vectors in different tangent spaces have no addition. So we have to find a way to transport a vector from one tangent space to another.

Definiton 2 [1] *A vector transport $\mathcal{T}$ on a manifold $\mathcal{M}$ is a smooth mapping*

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

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

$$T\mathcal{M} \oplus T\mathcal{M} = \{(\eta, \xi) | \eta, \xi \in T_x\mathcal{M}, x \in \mathcal{M}\}.$$

1. There exists a retraction R, called the retraction associated with $\mathcal{T}$, such that

$$\pi(\mathcal{T}_\eta(\xi)) = R_x(\eta), \quad \eta, \xi \in T_x\mathcal{M},$$

where $\pi(\mathcal{T}_\eta(\xi))$ denotes the foot of the tangent vector $\mathcal{T}_\eta(\xi)$.

2. $\mathcal{T}_{0_x}(\xi) = \xi$ for all $\xi \in T_x\mathcal{M}$.

3. $\mathcal{T}_\eta(a\xi + b\zeta) = a\mathcal{T}_\eta(\xi) + b\mathcal{T}_\eta(\zeta)$ for all $a, b \in \mathbb{R}$, $\eta, \xi, \zeta \in T_x\mathcal{M}$.

Definiton 3 [14] *A vector transport $\mathcal{T}$ is called isometric if it satisfies*

$$\langle \mathcal{T}_\eta(\xi), \mathcal{T}_\eta(\xi) \rangle_{R_x(\eta)} = \langle \xi, \xi \rangle_x$$

for all $\eta, \xi \in T_x\mathcal{M}$, where R is the retraction associated with $\mathcal{T}$.

Given a vector transport, we can define a Riemannian conjugate gradient direction as follows

$$\eta_{k+1} = -\nabla f(x_{k+1}) + \beta_{k+1}\mathcal{T}_{\alpha_k\eta_k}(\eta_k). \tag{4}$$

A commonly used vector transport is the differentiated retraction [1]

$$\mathcal{T}^R_\eta(\xi) = \mathrm{D}R_x(\eta)[\xi] = \left.\frac{\mathrm{d}}{\mathrm{d}t}R_x(\eta + t\xi)\right|_{t=0},$$

and consequently, we have

$$\mathcal{T}^R_{\alpha_k\eta_k}(\eta_k) = \mathrm{D}R_{x_k}(\alpha_k\eta_k)[\eta_k] = \left.\frac{\mathrm{d}}{\mathrm{d}t}R_{x_k}(t\eta_k)\right|_{t=\alpha_k}. \tag{5}$$

Traditional conjugate gradient methods such as the Fletcher-Reeves method use the strong Wolfe conditions, whose Riemannian version [25] is

$$f(R_{x_k}(\alpha_k\eta_k)) \le f(x_k) + c_1\alpha_k\langle \nabla f(x_k), \eta_k \rangle_{x_k}, \tag{6}$$

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

where $0 < c_1 < c_2 < 1$. Another type of conjugate gradient method, the Dai-Yuan method [6], only requires the weak Wolfe conditions, where (7a) is replaced by

$$\langle \nabla f(R_{x_k}(\alpha_k\eta_k)), \mathrm{D}R_{x_k}(\alpha_k\eta_k)[\eta_k]\rangle_{R_{x_k}(\alpha_k\eta_k)} \geq c_2\langle \nabla f(x_k), \eta_k\rangle_{x_k}. \tag{7b}$$

No matter the weak or strong Wolfe conditions are used, one need impose on vector transports the Ring-Wirth nonexpansive condition [25]

$$\langle \mathcal{T}_{\alpha_k\eta_k}(\eta_k), \mathcal{T}_{\alpha_k\eta_k}(\eta_k)\rangle_{R_{x_k}(\alpha_k\eta_k)} \leq \langle \eta_k, \eta_k\rangle_{x_k} \tag{8}$$

to establish global convergence for Riemannian conjugate gradient methods. To overcome this difficulty, Sato and Iwai [30] introduced the notion of a scaled vector transport

$$\mathcal{T}^{\text{scaled}}_{\alpha_k\eta_k}(\eta_k) = \begin{cases} \mathcal{T}^R_{\alpha_k\eta_k}(\eta_k), & \text{if } \|\mathcal{T}^R_{\alpha_k\eta_k}(\eta_k)\|_{R_{x_k}(\alpha_k\eta_k)} \leq \|\eta_k\|_{x_k}, \\ \frac{\|\eta_k\|_{x_k}}{\|\mathcal{T}^R_{\alpha_k\eta_k}(\eta_k)\|_{R_{x_k}(\alpha_k\eta_k)}}\mathcal{T}^R_{\alpha_k\eta_k}(\eta_k), & \text{otherwise.} \end{cases}$$

for the Riemannian Fletcher-Reeves method. Here we denote by $\|\cdot\|_x$ the induced norm of the metric $\langle\cdot,\cdot\rangle_x$. Recently, Sato [28] applied the scaled vector transport to the Riemannian Dai-Yuan method. However, the scaled vector transport does not possess linearity, the third condition in Definition 2. It is easy to see from Definition 3 that condition (8) must hold for every isometric vector transport. More recently, Huang, Absil and Gallivan [16] proposed some intrinsic representation of vector transports for matrix manifolds including the Stiefel manifold. Their intrinsic vector transports are not only isometric but also cheap to compute, requiring $O(np^2)$ flops.

3 A Riemannian conjugate gradient algorithm

In this section we describe in detail our new Riemannian conjugate gradient algorithm. The proposed algorithm has two creative ingredients, two new vector transports associated with the Cayley transform and a generalization of a nonmonotone conjugate gradient method.

3.1 Retraction via the Cayley transform

The tangent space $T_X\mathrm{St}(n,p)$ of the Stiefel manifold can be expressed as

$$T_X\mathrm{St}(n,p) = \left\{Z \in \mathbb{R}^{n\times p} : X^\top Z + Z^\top X = 0\right\} = \{X\Omega + X_\perp K : \Omega^\top = -\Omega, K \in \mathbb{R}^{(n-p)\times p}\}, \tag{9}$$

where $X_\perp \in \mathbb{R}^{n\times(n-p)}$ in the second expression is an orthogonal complement matrix of X. A natural metric for $T_X\mathrm{St}(n,p)$ is the Euclidean metric $\langle Y,Z\rangle_X = \mathrm{tr}(Y^\top Z)$, which induces a norm $\|Z\|_X = \|Z\|_F$, where $\|\cdot\|_F$ is the Frobenius norm of a matrix. Then the projection onto $T_X\mathrm{St}(n,p)$ is given by

$$\mathrm{Proj}_X M = (I - XX^\top)M + X\mathrm{skew}(X^\top M) = M - X\mathrm{sym}(X^\top M), \tag{10}$$

where $\mathrm{sym}(A) = \frac{1}{2}(A + A^\top)$ and $\mathrm{skew}(A) = \frac{1}{2}(A - A^\top)$. The Riemannian gradient $\nabla f(X)$ of the objective function can be computed according to the following projective property.

Lemma 1 [1] *Let $\overline{f}$ be a smooth function defined on a Riemannian manifold $\overline{\mathcal{M}}$ and let f denote the restriction of $\overline{f}$ to a Riemannian submanifold $\mathcal{M}$. The gradient of f is equal to the projection of the gradient of $\overline{f}$ onto $T_x\mathcal{M}$, that is, $\nabla f(x) = \mathrm{Proj}_x\nabla\overline{f}(x)$.*

Applying Lemma 1 with (10), we have

$$\nabla f(X) = G - X\mathrm{sym}(X^\top G), \tag{11}$$

where $G = \frac{\partial f(X)}{\partial X}$ denotes the gradient of $f(X)$ in the Euclidean space.

The well-known Cayley transform [23, 35] is adopted for our retraction. We introduce briefly this retraction as follows. For all $Z \in T_X\mathrm{St}(n,p)$, it holds

$$Z = W_Z X, \tag{12}$$

where

$$W_Z = P_X Z X^\top - X Z^\top P_X \quad \text{and} \quad P_X = I - \frac{1}{2} X X^\top. \tag{13}$$

This leads to a retraction on the Stiefel manifold

$$R_X(tZ) = X(t) = \left(I - \frac{t}{2} W_Z\right)^{-1} \left(I + \frac{t}{2} W_Z\right) X, \tag{14}$$

which is the so-called Cayley transform. It can be shown that the curve $\Gamma(t) = R_X(tZ)$ is contained in $\mathrm{St}(n, p)$ and satisfies $\Gamma(0) = X$ and $\Gamma'(0) = W_Z X = Z$.

If $Z = \nabla f(X)$, it follows from (11) and (13) that

$$W_Z = P_X \nabla f(X) X^\top - X \nabla f(X)^\top P_X = P_X G X^\top - X G^\top P_X.$$

Wen and Yin [35] proposed a refinement for (14) in the case of low-rank matrices. Rewriting W_{Z_k} as $W_{Z_k} = U_k V_k^\top$, where $U_k = [P_{X_k} Z_k, X_k]$ and $V_k = [X_k, -P_{X_k} Z_k]$ in general and $U_k = [-P_{X_k} G_k, X_k]$ and $V_k = [X_k, P_{X_k} G_k]$ when $Z_k = -\nabla f(X_k)$, they applied the Sherman-Morrison-Woodbury formula

$$\left(I - \frac{\alpha_k}{2} W_{Z_k}\right)^{-1} = \left(I - \frac{\alpha_k}{2} U_k V_k^\top\right)^{-1} = I + \frac{\alpha_k}{2} U_k \left(I - \frac{\alpha_k}{2} V_k^\top U_k\right)^{-1} V_k^\top$$

to obtain the following update scheme

$$X_k(\alpha_k) = X_k + \alpha_k U_k \left(I - \frac{\alpha_k}{2} V_k^\top U_k\right)^{-1} V_k^\top X_k. \tag{15}$$

Now we consider the complexity of (15) after Z_k is obtained. Forming

$$V_k^\top U_k = \begin{bmatrix} \frac{1}{2} X_k^\top Z_k & I \\ -Z_k^\top Z_k + \frac{1}{2}(Z_k^\top X_k)(X_k^\top Z_k) & -\frac{1}{2} Z_k^\top X_k \end{bmatrix}$$

needs $2np^2 + O(p^3)$ flops, and $V_k^\top X_k$ comes for free since it is the right part of $V_k^\top U_k$. The inversion $(I + V_k^\top U_k)^{-1}$ takes $O(p^3)$ flops, and the final assembly of $X_k(\alpha_k)$ takes another $4np^2 + O(np)$ flops. Hence, the complexity of (15) is $6np^2 + O(np) + O(p^3)$ flops. Updating $X_k(\alpha_k)$ for a different α_k (during backtracking line search) has a lower cost at $4np^2 + O(p^3)$ flops.

3.2 Vector transport as differentiated retraction

Now we derive computational formulae for the vector transport as the differentiated retraction of (14). Differentiating both sides of

$$\left(I - \frac{1}{2} W_Z - \frac{t}{2} W_Y\right) R_X(Z + tY) = \left(I + \frac{1}{2} W_Z + \frac{t}{2} W_Y\right) X$$

with respect to t, we have

$$-\frac{1}{2} W_Y R_X(Z + tY) + \left(I - \frac{1}{2} W_Z - \frac{t}{2} W_Y\right) \frac{\mathrm{d}}{\mathrm{d}t} R_X(Z + tY) = \frac{1}{2} W_Y X.$$

Then

$$\begin{aligned} \mathcal{T}_Z^R(Y) = \left.\frac{\mathrm{d}}{\mathrm{d}t} R_X(Z + tY)\right|_{t=0} &= \frac{1}{2} \left(I - \frac{1}{2} W_Z\right)^{-1} W_Y (X + R_X(Z)) \\ &= \frac{1}{2} \left(I - \frac{1}{2} W_Z\right)^{-1} W_Y \left(X + \left(I - \frac{1}{2} W_Z\right)^{-1} \left(I + \frac{1}{2} W_Z\right) X\right) \\ &= \left(I - \frac{1}{2} W_Z\right)^{-1} W_Y \left(I - \frac{1}{2} W_Z\right)^{-1} X, \end{aligned} \tag{16}$$

and therefore

$$\mathcal{T}^R_{\alpha_k Z_k}(Z_k) = \left(I - \frac{\alpha_k}{2} W_{Z_k}\right)^{-2} W_{Z_k} X_k = \left(I - \frac{\alpha_k}{2} W_{Z_k}\right)^{-2} Z_k, \tag{17}$$

where the second equality follows from (12). The following lemma shows that (17) satisfies the Ring-Wirth nonexpansive property (8).

Lemma 2 *The vector transport formula given by (17) satisfies the Ring-Wirth nonexpansive condition (8) with respect to the Euclidean metric.*

Proof Since W_{Z_k} is skew-symmetric, its eigenvalues are zeros or pure imaginary numbers. Therefore $I - \frac{\alpha_k^2}{4} W_{Z_k}^2$ is a symmetric matrix with all eigenvalues not less than 1. Then we have from (17) that

$$\begin{aligned}\langle \mathcal{T}^R_{\alpha_k Z_k}(Z_k), \mathcal{T}^R_{\alpha_k Z_k}(Z_k)\rangle_{R_{X_k}(\alpha_k Z_k)} &= \mathrm{tr}\left(Z_k^\top \left(I + \frac{\alpha_k}{2} W_{Z_k}\right)^{-2} \left(I - \frac{\alpha_k}{2} W_{Z_k}\right)^{-2} Z_k\right) \\ &= \mathrm{tr}\left(Z_k^\top \left(I - \frac{\alpha_k^2}{4} W_{Z_k}^2\right)^{-2} Z_k\right) \\ &\leq \langle Z_k, Z_k\rangle_{X_k},\end{aligned}$$

which completes the proof. □

In general, by the definition of vector transports, the mapping $\mathcal{T}_{\alpha\eta} : T_x\mathcal{M} \to T_{R_x(\alpha\eta)}\mathcal{M}$ is an isomorphism and therefore preserves linear independence for all sufficiently small α. This implies that, for all sufficiently small α, if $\xi_1, \cdots, \xi_d$, where d is the dimension of $\mathcal{M}$, form a basis of $T_x\mathcal{M}$, then $\mathcal{T}_{R_x(\alpha\eta)}(\xi_1), \cdots, \mathcal{T}_{R_x(\alpha\eta)}(\xi_d)$ form a basis of $T_{R_x(\alpha\eta)}\mathcal{M}$. To see this local isomorphic property more clearly for our special case of vector transport (16), we do the following calculations.

The isomorphism of $\mathcal{T}^R_{\alpha Z} : T_X\mathrm{St}(n,p) \to T_{R_X(\alpha Z)}\mathrm{St}(n,p) : Y \mapsto \left(I - \frac{\alpha}{2} W_Z\right)^{-1} W_Y \left(I - \frac{\alpha}{2} W_Z\right)^{-1} X$ means this linear mapping has full rank. Taking the vectorization operator $\mathrm{vec}(\cdot)$ and plugging the second expression of (9) in (13), we have

$$\begin{aligned}\mathrm{vec}(\mathcal{T}^R_{\alpha Z}(Y)) &= [X^\top \otimes I] \cdot \left[\left(I + \frac{\alpha}{2} W_Z\right)^{-1} \otimes \left(I - \frac{\alpha}{2} W_Z\right)^{-1}\right] \cdot [X \otimes X,\ \ X \otimes X_\perp - (X_\perp \otimes X)\Pi] \cdot \begin{bmatrix} \mathrm{vec}(\Omega) \\ \mathrm{vec}(K) \end{bmatrix} \\ &= B_1 \cdot L(\alpha) \cdot B_2 \cdot \begin{bmatrix} \mathrm{vec}(\Omega) \\ \mathrm{vec}(K) \end{bmatrix},\end{aligned}$$

where $\otimes$ is the Kronecker product and Π is the symmetric permutation matrix that satisfies $\mathrm{vec}(A^\top) = \Pi\mathrm{vec}(A)$ [33]. Obviously, the full rankness of $B_1 L(\alpha) B_2$ implies the isomorphism of $\mathcal{T}^R_{\alpha Z}$. By $T_0^R(Y) = Y$ in the definition of vector transports or plugging $\alpha = 0$ in $L(\alpha)$ directly, we see that $B_1 L(0) B_2 = B_1 B_2 = [I \otimes X,\ \ I \otimes X_\perp]$ is definitely full-rank. But in general $B_1 L(\alpha) B_2$ may be rank deficient although $L(\alpha)$ is invertible and B_1 and B_2 are both full-rank. Thus $\mathcal{T}^R_{\alpha Z}$ is isomorphic for those αs which make $B_1 L(\alpha) B_2$ a full-rank matrix. Continuity guarantees the existence of a safe interval around $\alpha = 0$ for the full rankness of $B_1 L(\alpha) B_2$.

Now we consider the computation of (17) for low-rank matrices. Applying again the Sherman-Morrison-Woodbury formula

$$\left(I - \frac{\alpha_k}{2} W_{Z_k}\right)^{-1} = \left(I - \frac{\alpha_k}{2} U_k V_k^\top\right)^{-1} = I + \frac{\alpha_k}{2} U_k \left(I - \frac{\alpha_k}{2} V_k^\top U_k\right)^{-1} V_k^\top,$$

we have

$$\begin{aligned}\mathcal{T}^R_{\alpha_k Z_k}(Z_k) &= W_{Z_k} \left(I - \frac{\alpha_k}{2} W_{Z_k}\right)^{-2} X_k. \\ &= U_k V_k^\top \left[I + \frac{\alpha_k}{2} U_k \left(I - \frac{\alpha_k}{2} V_k^\top U_k\right)^{-1} V_k^\top\right]^2 X_k \\ &= U_k \left[V_k^\top X_k + \alpha_k V_k^\top U_k \left(I - \frac{\alpha_k}{2} V_k^\top U_k\right)^{-1} V_k^\top X_k + \frac{\alpha_k^2}{4} \left(V_k^\top U_k\right)^2 \left(I - \frac{\alpha_k}{2} V_k^\top U_k\right)^{-2} V_k^\top X_k\right] \\ &= U_k \left[V_k^\top X_k + \alpha_k V_k^\top U_k \left(I - \frac{\alpha_k}{2} V_k^\top U_k\right)^{-1} V_k^\top X_k + \frac{\alpha_k}{2} \left(\left(I - \frac{\alpha_k}{2} V_k^\top U_k\right)^{-1} - I\right) V_k^\top U_k \left(I - \frac{\alpha_k}{2} V_k^\top U_k\right)^{-1} V_k^\top X_k\right] \\ &= U_k \left[V_k^\top X_k + \frac{\alpha_k}{2} V_k^\top U_k \left(I - \frac{\alpha_k}{2} V_k^\top U_k\right)^{-1} V_k^\top X_k + \frac{\alpha_k}{2} \left(I - \frac{\alpha_k}{2} V_k^\top U_k\right)^{-1} V_k^\top U_k \left(I - \frac{\alpha_k}{2} V_k^\top U_k\right)^{-1} V_k^\top X_k\right].\end{aligned}$$

Then

$$\mathcal{T}^R_{\alpha_k Z_k}(Z_k) = U_k \left[M_{k,1} + \frac{\alpha_k}{2} M_{k,2} M_{k,3} + \frac{\alpha_k}{2} \left(I - \frac{\alpha_k}{2} M_{k,2}\right)^{-1} M_{k,2} M_{k,3}\right], \tag{18}$$

where

$$M_{k,1} = V_k^\top X_k, \quad M_{k,2} = V_k^\top U_k, \quad M_{k,3} = \left(I - \frac{\alpha_k}{2} V_k^\top U_k\right)^{-1} V_k^\top X_k. \tag{19}$$

With the memories of $M_{k,1}$, $M_{k,2}$ and $M_{k,3}$ after computing (15), (18) takes only $4np^2 + O(p^3)$ flops.

3.3 An isometric vector transport

Now we introduce a new isometric vector transport motivated by the retraction using the matrix exponential

$$X(t) = R_X^{\exp}(tZ) = e^{tW_Z} X,$$

which is also the Riemannian exponential under the canonical metric $\langle Y, Z\rangle_X = \operatorname{tr}(Y^\top(I - \frac{1}{2}XX^\top)Z)$ [23, 43]. By direct calculations, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} R_{X_k}^{\exp}(tZ_k)\Big|_{t=\alpha_k} = e^{\alpha_k W_{Z_k}} W_{Z_k} X_k = e^{\alpha_k W_{Z_k}} Z_k, \tag{20}$$

where the second equality still follows from (12).

The differentiated matrix exponential (20) indicates to us a new formula for $\mathcal{T}_{\alpha_k Z_k}(Z_k)$, that is,

$$\mathcal{T}_{\alpha_k Z_k}(Z_k) = \left(I - \frac{\alpha_k}{2} W_{Z_k}\right)^{-1} \left(I + \frac{\alpha_k}{2} W_{Z_k}\right) Z_k = \left(I - \frac{\alpha_k}{2} W_{Z_k}\right)^{-1} \left(I + \frac{\alpha_k}{2} W_{Z_k}\right) W_{Z_k} X_k = W_{Z_k} X_{k+1}. \tag{21}$$

The intrinsic idea of (21) is to approximate the exponential part $e^{\alpha_k W_{Z_k}}$ in (20) by the corresponding Cayley transform

$$\left(I - \frac{\alpha_k}{2} W_{Z_k}\right)^{-1} \left(I + \frac{\alpha_k}{2} W_{Z_k}\right).$$

In fact, $\left(I - \frac{t}{2}A\right)^{-1}\left(I + \frac{t}{2}A\right)$ is also the first-order diagonal Padé approximant to the matrix exponential e^{tA} [13]. Making further efforts, we find a new isometric vector transport formula

$$\mathcal{T}_Z(Y) = \left(I - \frac{1}{2} W_Z\right)^{-1} \left(I + \frac{1}{2} W_Z\right) Y, \tag{22}$$

which satisfies (21). The following lemma shows that (22) really defines an isometric vector transport.

Lemma 3 *The mapping $\mathcal{T}$ defined by*

$$\mathcal{T}_Z(Y) : T_X\mathrm{St}(n,p) \to T_{R_X(Z)}\mathrm{St}(n,p) : Y \mapsto \left(I - \frac{1}{2} W_Z\right)^{-1} \left(I + \frac{1}{2} W_Z\right) Y$$

is a vector transport on the Stiefel manifold. Moreover, it is isometric with respect to the Euclidean metric.

Proof We first show that $\mathcal{T}$ satisfies all the conditions in Definition 2 and then prove its isometry. Consider the Cayley transform

$$R_X(Z) = \left(I - \frac{1}{2} W_Z\right)^{-1} \left(I + \frac{1}{2} W_Z\right) X$$

for any $X \in \mathrm{St}(n,p)$ and any $Y, Z \in T_X\mathrm{St}(n,p)$. Using the skew-symmetry of W_Z and the first expression in (9), we have

$$\mathcal{T}_Z(Y)^\top R_X(Z) + R_X(Z)^\top \mathcal{T}_Z(Y) = Y^\top X + X^\top Y = 0,$$

which means $\mathcal{T}_Z(Y) \in T_{R_X(Z)}\mathrm{St}(n,p)$ and therefore $R_X(Z)$ is the foot of $\mathcal{T}_Z(Y)$. Thus we have proved that the Cayley transform $R_X(Z)$ is a retraction associated with $\mathcal{T}$. The second condition $\mathcal{T}_{0_X}(Y) = Y$ and the third condition

$$\mathcal{T}_Z(aY_1 + bY_2) = a\mathcal{T}_Z(Y_1) + b\mathcal{T}_Z(Y_2)$$

are trivial according to the definition (22) of $\mathcal{T}$. The smoothness of (22) follows immediately from (13) and (22) itself. Hence $\mathcal{T}$ is indeed a vector transport on the Stiefel manifold. The isometry

$$\langle \mathcal{T}_Z(Y), \mathcal{T}_Z(Y)\rangle_{R_X(Z)} = \langle Y, Y\rangle_X$$

of $\mathcal{T}$ follows immediately from the skew-symmetry of W_Z. □

Remark 1 *Since $\mathcal{T}$ in (22) is isometric, it also satisfies the Ring-Wirth nonexpansive condition (8).*

Isometric vector transports $\mathcal{T}$ can preserve orthogonality and therefore make $\mathcal{T}_\eta : T_x\mathcal{M} \to T_{R_x(\eta)}\mathcal{M}$ an isomorphism. This means that if $\xi_1, \cdots, \xi_d$, where d is the dimension of $\mathcal{M}$, form an orthonormal basis of $T_x\mathcal{M}$, then $\mathcal{T}_{R_x(\eta)}(\xi_1), \cdots, \mathcal{T}_{R_x(\eta)}(\xi_d)$ form an orthonormal basis of $T_{R_x(\eta)}\mathcal{M}$. Moreover, isometry is fundamental for vector transports in theory of Riemannnian quasi-Newton methods [14, 15, 16, 17].

Now we consider the computation of (21) for low-rank matrices. One can also use (15) to obtain a refinement of (21) as follows

$$\mathcal{T}_{\alpha Z_k}(Z_k) = U_k V_k^\top X_{k+1} = U_k \left[V_k^\top X_k + \alpha_k V_k^\top U_k \left(I - \frac{\alpha_k}{2} V_k^\top U_k \right)^{-1} V_k^\top X_k \right].$$

Then

$$\mathcal{T}_{\alpha Z_k}(Z_k) = U_k \left(M_{k,1} + \alpha_k M_{k,2} M_{k,3} \right), \tag{23}$$

where $M_{k,1}$, $M_{k,2}$ and $M_{k,3}$ are the same as those in (19). With the memories of $M_{k,1}$, $M_{k,2}$ and $M_{k,3}$ after computing (15), (23) also takes $4np^2 + O(p^3)$ flops.

Now we discuss briefly an extension of vector transport (22). In fact, (22) can be generalized to a new class of isometric vector transports. Applying the mth-order diagonal Padé approximant $q_m(A)^{-1} p_m(A)$ to the matrix exponential e^A, where

$$p_m(x) = \sum_{i=0}^{m} \frac{(2m-i)!m!}{(2m)!(m-i)!} \frac{x^i}{i!} \quad \text{and} \quad q_m(x) = \sum_{i=0}^{m} \frac{(2m-i)!m!}{(2m)!(m-i)!} \frac{(-x)^i}{i!}$$

are the pair of Padé polynomials, one can use the same way as Lemma 3 to prove that

$$\mathcal{T}_Z(Y) = q_m(W_Z)^{-1} p_m(W_Z) Y \tag{24}$$

is an isometric vector transport on the Stiefel manifold, whose associated retraction is

$$R_X(Z) = q_m(W_Z)^{-1} p_m(W_Z) X, \tag{25}$$

which has been proved to be a retraction on the Stiefel manifold [43]. From (24) we have

$$\mathcal{T}_{\alpha_k Z_k}(Z_k) = q_m(\alpha_k W_{Z_k})^{-1} p_m(\alpha_k W_{Z_k}) Z_k = q_m(\alpha_k W_{Z_k})^{-1} p_m(\alpha_k W_{Z_k}) W_{Z_k} X_k = W_{Z_k} X_{k+1}. \tag{26}$$

It is commonly known that the mth-order diagonal Padé approximant is a $2m$th-order approximation to the origin function. Specifically, for the matrix exponential, we have [13]

$$e^A - q_m(A)^{-1} p_m(A) = \frac{(-1)^m}{(2m)!} A^{2m+1} q_m(A)^{-1} \int_0^1 e^{tA} (1-t)^m t^m \mathrm{d}t.$$

Computing the couple of (25) and (26) directly is impractical if m is large. A possible way to implement this approach is to exploit subspace techniques [26, 37]. Further study on this Padé-based approach is beyond the scope of the paper, and we leave it to the future.

3.4 Relation between the new vector transports

Now we investigate the relation between the two vector transports proposed in Sections 3.2 and 3.3. The first one (17) follows a natural idea that is to generate a vector transport by the differentiated retraction (5). This idea stems from the parallel translation along a geodesic [8]

$$P_\Gamma^{\alpha \leftarrow 0} \eta_x = \left. \frac{\mathrm{d}}{\mathrm{d}t} \mathrm{Exp}(t\eta_x) \right|_{t=\alpha},$$

where $\Gamma(t) = \mathrm{Exp}(t\eta_x)$ with Exp being the Riemannian exponential is the geodesic starting from x with the initial velocity $\dot{\Gamma}(0) = \eta_x \in T_x\mathcal{M}$. However, there is some doubt whether the second vector transport (21) provides a quality direction corresponding to the previous search direction although it is associated with the same retraction as (17).

Let θ_k be the angle between $\mathcal{T}_{\alpha_k Z_k}(Z_k)$ and $\mathcal{T}^R_{\alpha_k Z_k}(Z_k)$. We compare the two vector transports by estimating this angle. Although we know immediately that $\theta_k \to 0$ as $\alpha_k \to 0$ from the fact

$$\lim_{\alpha_k \to 0} \left\| \mathcal{T}_{\alpha_k Z_k}(Z_k) - \mathcal{T}^R_{\alpha_k Z_k}(Z_k) \right\|_{R_{X_k}(\alpha_k Z_k)} = 0$$

implied by Definition 2, it is necessary to inspect how $\mathcal{T}_{\alpha_k Z_k}(Z_k)$ deviates from $\mathcal{T}^R_{\alpha_k Z_k}(Z_k)$ when α_k gets large. By (17) and (21), $\mathcal{T}_{\alpha_k Z_k}(Z_k)$ and $\mathcal{T}^R_{\alpha_k Z_k}(Z_k)$ can be related by

$$\mathcal{T}_{\alpha_k Z_k}(Z_k) = \left(I - \frac{\alpha_k^2}{4} W_{Z_k}^2 \right) \mathcal{T}^R_{\alpha_k Z_k}(Z_k). \tag{27}$$

We can derive a lower bound for $\cos\theta_k$ by the following lemma.

Lemma 4 *For any symmetric positive definite matrix B and vector x, we have*

$$\frac{x^\top B x}{\|x\|_2 \|Bx\|_2} \geq \frac{1}{\|B\|_2^{1/2} \|B^{-1}\|_2^{1/2}}.$$

Proof Because B is symmetric positive definite, we have that the matrices $B^{1/2}$ and $B^{-1/2}$ exist and that $\|B^{1/2}\|_2 = \|B\|_2^{1/2}$ and $\|B^{-1/2}\|_2 = \|B^{-1}\|_2^{1/2}$. Let $\tilde{x} = B^{1/2} x$. Then we have

$$\begin{aligned} \frac{x^\top B x}{\|x\|_2 \|Bx\|_2} &= \frac{\tilde{x}^\top \tilde{x}}{\|B^{-1/2}\tilde{x}\|_2 \|B^{1/2}\tilde{x}\|_2} \\ &\geq \frac{\|\tilde{x}\|_2^2}{\|B^{-1/2}\|_2 \|\tilde{x}\|_2 \|B^{1/2}\|_2 \|\tilde{x}\|_2} \\ &= \frac{1}{\|B\|_2^{1/2} \|B^{-1}\|_2^{1/2}}, \end{aligned}$$

which completes the proof. □

It follows from (27) and Lemma 4 that

$$\begin{aligned} \cos\theta_k &= \frac{\operatorname{tr}\left(\mathcal{T}^R_{\alpha_k Z_k}(Z_k)^\top \left(I - \frac{\alpha_k^2}{4} W_{Z_k}^2 \right) \mathcal{T}^R_{\alpha_k Z_k}(Z_k) \right)}{\left\| \mathcal{T}^R_{\alpha_k Z_k}(Z_k) \right\|_F \left\| \left(I - \frac{\alpha_k^2}{4} W_{Z_k}^2 \right) \mathcal{T}^R_{\alpha_k Z_k}(Z_k) \right\|_F} \\ &\geq \frac{1}{\left\| I - \frac{\alpha_k^2}{4} W_{Z_k}^2 \right\|_2^{1/2} \left\| \left(I - \frac{\alpha_k^2}{4} W_{Z_k}^2 \right)^{-1} \right\|_2^{1/2}} \\ &= \sqrt{\frac{4 + \alpha_k^2 \beta_{k,1}^2}{4 + \alpha_k^2 \beta_{k,n}^2}}, \end{aligned}$$

where $\beta_{k,1}, \beta_{k,2}, \cdots, \beta_{k,n}$ with $|\beta_{k,1}| \leq |\beta_{k,2}| \leq \cdots \leq |\beta_{k,n}|$ are the imaginary parts of the eigenvalues of W_{Z_k}. This means θ_k is an acute angle between 0 and $\arccos \sqrt{\frac{4+\alpha_k^2\beta_{k,1}^2}{4+\alpha_k^2\beta_{k,n}^2}}$. So θ_k must be small if $|\beta_{k,n}|$ is close to $|\beta_{k,1}|$. From this observation, we see that $\mathcal{T}_{\alpha_k Z_k}(Z_k)$ is at least not a bad direction compared with the differentiated retraction $\mathcal{T}^R_{\alpha_k Z_k}(Z_k)$.

Now we consider the constructions of the two vector transports. In the point of view that the Cayley transform is an approximation to the matrix exponential, $\mathcal{T}^R_{\alpha_k Z_k}(Z_k)$ can be viewed as the differentiation of an approximate exponential, while $\mathcal{T}_{\alpha_k Z_k}(Z_k)$ can be viewed as an approximation to the differentiated exponential. The difference between them is the order of approximation and differentiation operators.

3.5 Comparison with existing vector transports

Most existing Riemannian conjugate gradient methods use the QR decomposition or the polar decomposition to construct a retraction, and naturally use the corresponding differentiated retraction as a vector transport. Retractions via the QR decomposition and the polar decomposition are given by [1]

$$R_X^{\mathrm{qr}}(Z) = \mathrm{qf}(X+Z), \tag{28}$$

where qf denotes the Q factor of the QR decomposition, and

$$R_X^{\mathrm{pol}}(Z) = (X+Z)(I+Z^\top Z)^{-\frac{1}{2}}. \tag{29}$$

According to [1], the differentiated retraction of (28) is

$$\mathcal{T}_Z^{\mathrm{qr}}(Y) = \mathrm{D}R_X^{\mathrm{qr}}(Z)[Y] = R_X^{\mathrm{qr}}(Z)\rho_{\mathrm{skew}}(R_X^{\mathrm{qr}}(Z)^\top Y \mathrm{rf}(X+Z)^{-1}) + (I - R_X^{\mathrm{qr}}(Z)R_X^{\mathrm{qr}}(Z)^\top)Y\mathrm{rf}(X+Z)^{-1},$$

where rf denotes the R factor of the QR decomposition and

$$(\rho_{\mathrm{skew}}(A))_{ij} = \begin{cases} A_{ij}, & \text{if } i > j, \\ 0, & \text{if } i = j, \\ -A_{ji}, & \text{if } i < j. \end{cases}$$

Letting $X_{k+1} = R_{X_k}^{\mathrm{qr}}(\alpha_k Z_k)$, we immediately have

$$\mathcal{T}_{\alpha_k Z_k}^{\mathrm{qr}}(Z_k) = X_{k+1}\rho_{\mathrm{skew}}(X_{k+1}^\top Z_k \mathrm{rf}(X_k + \alpha_k Z_k)^{-1}) + (I - X_{k+1}X_{k+1}^\top)Z_k \mathrm{rf}(X_k+\alpha_k Z_k)^{-1}. \tag{30}$$

According to [17], the differentiated retraction of (29) is

$$\mathcal{T}_Z^{\mathrm{pol}}(Y) = \mathrm{D}R_X^{\mathrm{pol}}(Z)[Y] = R_X^{\mathrm{pol}}(Z)\Lambda + (I - R_X^{\mathrm{pol}}(Z)R_X^{\mathrm{pol}}(Z)^\top)Y(I+Z^\top Z)^{-\frac{1}{2}},$$

where Λ is the unique solution of the Sylvester equation

$$\Lambda(I+Z^\top Z)^{\frac{1}{2}} + (I+Z^\top Z)^{\frac{1}{2}}\Lambda = R_X^{\mathrm{pol}}(Z)^\top Y - Y^\top R_X^{\mathrm{pol}}(Z).$$

When $Y = Z$, the solution of the above equation has a closed form $\Lambda = (I+Z^\top Z)^{-\frac{1}{2}}X^\top Z(I+Z^\top Z)^{-\frac{1}{2}}$, yielding

$$\mathcal{T}_{\alpha_k Z_k}^{\mathrm{pol}}(Z_k) = (I - \alpha_k X_{k+1}(I+\alpha_k^2 Z_k^\top Z_k)^{-\frac{1}{2}}Z_k^\top)Z_k(I+\alpha_k^2 Z_k^\top Z_k)^{-\frac{1}{2}}, \tag{31}$$

where $X_{k+1} = R_{X_k}^{\mathrm{pol}}(\alpha_k Z_k)$.

Although vector transports (30) and (31) are often used in practice, they are not theoretically perfect, namely has a hidden danger of divergence, since the Ring-Wirth nonexpansive condition (8) can not be guaranteed for them. The recent proposed intrinsic vector transport in [16] is isometric and has a competitive complexity. Procedures for this new vector transport are too complicated to describe here and we just mention that this vector transport does not require an associated retraction, different from the original definition (Definition 2). The reader is referred to Algorithms 1 to 5 in [16] for details.

Complexity of aforementioned vector transports and their associated retractions (if exist) are summarized in Table 1. Note that the complexity of the vector transport in [16] can be reduced to $8np^2 + O(np) + O(p^3)$ if the QR retraction computed by Householder reflections is used.

3.6 A generalization of Dai's nonmonotone method

We consider Dai's nonmonotone conjugate gradient method [5]. His formula of β_{k+1} is

$$\beta_{k+1}^{\mathrm{D}} = \frac{\|\nabla f(x_{k+1})\|^2}{\max\{y_k^\top \eta_k, -\nabla f(x_k)^\top \eta_k\}},$$

where $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$. We generalize this formula to the Riemannian case as

$$\beta_{k+1}^{\mathrm{D}} = \frac{\|\nabla f(x_{k+1})\|_{x_{k+1}}^2}{\max\{\langle \nabla f(x_{k+1}), \mathcal{T}_{\alpha_k\eta_k}(\eta_k)\rangle_{x_{k+1}} - \langle \nabla f(x_k), \eta_k\rangle_{x_k}, -\langle \nabla f(x_k), \eta_k\rangle_{x_k}\}}. \tag{32}$$

Table 1: Complexity of various retractions and their associated vector transports

Computational scheme	Complexity of retraction		Complexity of vector transport
	First α_k	New α_k	
Cayley transform: (15) and (18)	$6np^2 + O(np) + O(p^3)$	$4np^2 + O(np) + O(p^3)$	$4np^2 + O(p^3)$
Cayley transform: (15) and (23)	$6np^2 + O(np) + O(p^3)$	$4np^2 + O(np) + O(p^3)$	$4np^2 + O(p^3)$
QR decomposition: (28) and (30)	$2np^2 + O(np) + O(p^3)$	$2np^2 + O(np) + O(p^3)$	$8np^2 + O(np) + O(p^3)$
polar decomposition: (29) and (31)	$3np^2 + O(np) + O(p^3)$	$2np^2 + O(np) + O(p^3)$	$4np^2 + O(np) + O(p^3)$
Intrinsic representation in [16]	no associated retraction		$10np^2 + O(np) + O(p^3)$

Then we choose $\beta_{k+1} \in [0, \beta_{k+1}^{\mathrm{D}}]$, which is well defined since $\beta_{k+1}^{\mathrm{D}} > 0$ due to $\langle \nabla f(x_k), \eta_k \rangle_{x_k} < 0$ shown by Lemma 5 in the next section. Dai's algorithm is globally convergent with a nonmonotone line search condition instead of the Wolfe conditions. The Riemannian generalization of this nonmonotone condition is

$$f(R_{x_k}(\alpha_k \eta_k)) \leq \max\{f(x_k), \cdots, f(x_{k-\min\{m-1,k\}})\} + \delta \alpha_k \langle \nabla f(x_k), \eta_k \rangle_{x_k}, \tag{33}$$

where m is some positive integer. An advantage of this Armijo-type condition over the Wolfe conditions is that we do not need to compute any vector transport for (33) when updating α_k, while equation (7a) or (7b) in the Wolfe conditions involves a vector transport.

Like unconstrained optimization in the Euclidean space, $\|\nabla f(x_k)\|_{x_k} \leq \epsilon$ for some constant $\epsilon > 0$ or something like that is a reasonable stopping criterion for Riemannian optimization. In fact, the Lagrangian of problem (1) is

$$\mathcal{L}(X, \Lambda) = f(X) - \frac{1}{2}\mathrm{tr}(\Lambda(X^\top X - I)),$$

where Λ is a symmetric matrix representing the Lagrange multiplier. Then the first-order optimality conditions in the Euclidean sense [35] are $X^\top X = I$ and $\frac{\partial}{\partial X}\mathcal{L}(X, \Lambda) = G - XG^\top X = 0$ with $\Lambda = G^\top X$. Under the condition $X^\top X = I$, $\frac{\partial}{\partial X}\mathcal{L}(X, G^\top X) = 0$ is equivalent to $\nabla f(X) = 0$ since it follows from (11) that

$$\nabla f(X) = \left(I - \frac{1}{2}XX^\top\right)(G - XG^\top X) = \left(I - \frac{1}{2}XX^\top\right)\frac{\partial}{\partial X}\mathcal{L}(X, G^\top X). \tag{34}$$

Thus, first-order critical points in the Euclidean sense can be interpreted as stationary points in the Riemannian sense.

A formal presentation of the new algorithm is given as follows.

Algorithm 1: A Riemannian nonmonotone CG algorithm on the Stiefel manifold

Data: $\epsilon, \delta, \lambda \in (0, 1)$, $m \in \mathbb{N}^+$, $\alpha_{\max} > \alpha_0 > \alpha_{\min} > 0$, $X_0 \in \mathrm{St}(n, p)$, $Z_0 = -\nabla f(X_0)$, $k = 0$.

1 **while** $\|\nabla f(X_k)\|_{X_k} > \epsilon$ **do**

2 **if** $f(R_{X_k}(\alpha_k Z_k)) \leq \max\{f(X_k), \cdots, f(X_{k-\min\{m-1,k\}})\} + \delta \alpha_k \langle \nabla f(X_k), Z_k \rangle_{X_k}$ **then**

3 Set $X_{k+1} = R_{X_k}(\alpha_k Z_k)$, where R is defined by (14), which is computed via (15) if $p \ll n$;

4 **else**

5 Set $\alpha_k \leftarrow \lambda \alpha_k$ and go to line 2;

6 Compute $Z_{k+1} = -\nabla f(X_{k+1}) + \beta_{k+1} \mathcal{T}_{\alpha_k Z_k}(Z_k)$, where $\beta_{k+1} \in [0, \beta_{k+1}^{\mathrm{D}}]$ with

$$\beta_{k+1}^{\mathrm{D}} = \frac{\|\nabla f(X_{k+1})\|_{X_{k+1}}^2}{\max\{\langle \nabla f(X_{k+1}), \mathcal{T}_{\alpha_k Z_k}(Z_k) \rangle_{X_{k+1}} - \langle \nabla f(X_k), Z_k \rangle_{X_k}, -\langle \nabla f(X_k), Z_k \rangle_{X_k}\}},$$

and $\mathcal{T}_{\alpha_k Z_k}(Z_k)$ is given by (17) or (21), which is computed via (18) or (23) if $p \ll n$;

7 Update $\alpha_{k+1} \in [\alpha_{\min}, \alpha_{\max}]$ and set $k \leftarrow k + 1$;

Remark 2 *To achieve high efficiency in implementation, the actual stopping criterion is much more sophisticated than the simple one $\|\nabla f(X_k)\|_{X_k} \leq \epsilon$. Detailed algorithmic issues such as the stopping criterion, strategies for choosing β_{k+1} and the initial steplength, input parameters, and feasibility restoration are presented in Section 5.*

In fact, Algorithm 1 can be put in a more general framework, a Riemannian nonmonotone CG algorithm which is not restricted to the case of the Stiefel manifold. We describe this general framework in Algorithm 2 as follows. In the next section we will prove global convergence of Algorithm 2.

Algorithm 2: A Riemannian nonmonotone CG algorithm on general manifolds

Data: $\epsilon, \delta, \lambda \in (0,1)$, $m \in \mathbb{N}^+$, $\alpha_{\max} > \alpha_0 > \alpha_{\min} > 0$, $x_0 \in \mathcal{M}$, $\eta_0 = -\nabla f(x_0)$, $k = 0$.

1 **while** $\|\nabla f(x_k)\|_{x_k} > \epsilon$ **do**

2 **if** $f(R_{x_k}(\alpha_k \eta_k)) \leq \max\{f(x_k), \cdots, f(x_{k-\min\{m-1,k\}})\} + \delta \alpha_k \langle \nabla f(x_k), \eta_k \rangle_{x_k}$ **then**

3 Set $x_{k+1} = R_{x_k}(\alpha_k \eta_k)$, where R is a retraction;

4 **else**

5 Set $\alpha_k \leftarrow \lambda \alpha_k$ and go to line 2;

6 Compute $\eta_{k+1} = -\nabla f(x_{k+1}) + \beta_{k+1} \mathcal{T}_{\alpha_k \eta_k}(\eta_k)$, where $\beta_{k+1} \in [0, \beta_{k+1}^{\mathrm{D}}]$ with

$$\beta_{k+1}^{\mathrm{D}} = \frac{\|\nabla f(x_{k+1})\|_{x_{k+1}}^2}{\max\{\langle \nabla f(x_{k+1}), \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}} - \langle \nabla f(x_k), \eta_k \rangle_{x_k}, -\langle \nabla f(x_k), \eta_k \rangle_{x_k}\}},$$

and $\mathcal{T}$ is a vector transport;

7 Update $\alpha_{k+1} \in [\alpha_{\min}, \alpha_{\max}]$ and set $k \leftarrow k+1$;

4 Global convergence in a general setting

In this section, we show global convergence for Algorithm 2, where the considered manifold $\mathcal{M}$ is an arbitrary compact Riemannian manifold and the retraction R and the vector transport $\mathcal{T}$ are unspecified. Therefore the results are also valid for Algorithm 1 on the Stiefel manifold. The strategy generalizes that for the Euclidean case in [5]. Throughout this section, we make the following assumptions.

Assumption 1 *The objective function f is continuously differentiable on the compact Riemannian manifold $\mathcal{M}$, and there exists a Lipschitzian constant $L > 0$ such that*

$$|\mathrm{D}(f \circ R_x)(t\eta)[\eta] - \mathrm{D}(f \circ R_x)(0_x)[\eta]| \leq Lt$$

for all $\eta \in T_x\mathcal{M}$, $\|\eta\|_x = 1$, $x \in \mathcal{M}$, and $t \geq 0$.

Assumption 2 *The vector transport $\mathcal{T}$ satisfies the Ring-Wirth nonexpansive condition*

$$\langle \mathcal{T}_{\alpha_k \eta_k}(\eta_k), \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{R_{x_k}(\alpha_k \eta_k)} \leq \langle \eta_k, \eta_k \rangle_{x_k} \quad \text{for all } k.$$

Remark 3 *Note that we have shown in Lemmas 2 and 3 that for the Stiefel manifold both of the vector transports (16) and (22) satisfy Assumption 2.*

The following two lemmas are crucial for the final convergence theorem.

Lemma 5 *Suppose Algorithm 2 does not terminate in finitely many iterations. Then we have, for all k,*

$$\langle \nabla f(x_k), \eta_k \rangle_{x_k} < 0. \tag{35}$$

Therefore $\beta_{k+1}^{\mathrm{D}} > 0$ and $\beta_{k+1} \in [0, \beta_{k+1}^{\mathrm{D}}]$ is well defined.

Proof We proceed by induction. Since $\eta_0 = -\nabla f(x_0)$, (35) holds immediately for $k = 0$. Suppose (35) holds for some k. Then $\beta^{\mathrm{D}}_{k+1} > 0$ according to (32) and therefore β_{k+1} is well defined and the ratio $r_{k+1} = \frac{\beta_{k+1}}{\beta^{\mathrm{D}}_{k+1}}$ lies in $[0, 1]$. By (4) and (32), we have

$$\begin{aligned}
&\langle \nabla f(x_{k+1}), \eta_{k+1}\rangle_{x_{k+1}} \\
=&\frac{r_{k+1}\langle \nabla f(x_{k+1}), \mathcal{T}_{\alpha_k\eta_k}(\eta_k)\rangle_{x_{k+1}} - \max\{\langle \nabla f(x_{k+1}), \mathcal{T}_{\alpha_k\eta_k}(\eta_k)\rangle_{x_{k+1}} - \langle \nabla f(x_k), \eta_k\rangle_{x_k}, -\langle \nabla f(x_k), \eta_k\rangle_{x_k}\}}{\max\{\langle \nabla f(x_{k+1}), \mathcal{T}_{\alpha_k\eta_k}(\eta_k)\rangle_{x_{k+1}} - \langle \nabla f(x_k), \eta_k\rangle_{x_k}, -\langle \nabla f(x_k), \eta_k\rangle_{x_k}\}} \cdot \\
&\|\nabla f(x_{k+1})\|^2_{x_{k+1}}.
\end{aligned} \tag{36}$$

If $\langle \nabla f(x_{k+1}), \mathcal{T}_{\alpha_k\eta_k}(\eta_k)\rangle_{x_{k+1}} \geq 0$, it follows from the induction hypothesis and (36) that

$$\langle \nabla f(x_{k+1}), \mathcal{T}_{\alpha_k\eta_k}(\eta_k)\rangle_{x_{k+1}} - \langle \nabla f(x_k), \eta_k\rangle_{x_k} > 0$$

and

$$\langle \nabla f(x_{k+1}), \eta_{k+1}\rangle_{x_{k+1}} = \frac{(r_{k+1}-1)\langle \nabla f(x_{k+1}), \mathcal{T}_{\alpha_k\eta_k}(\eta_k)\rangle_{x_{k+1}} + \langle \nabla f(x_k), \eta_k\rangle_{x_k}}{\langle \nabla f(x_{k+1}), \mathcal{T}_{\alpha_k\eta_k}(\eta_k)\rangle_{x_{k+1}} - \langle \nabla f(x_k), \eta_k\rangle_{x_k}} \|\nabla f(x_{k+1})\|^2_{x_{k+1}},$$

which show that $\langle \nabla f(x_{k+1}), \eta_{k+1}\rangle_{x_{k+1}} < 0$. If $\langle \nabla f(x_{k+1}), \mathcal{T}_{\alpha_k\eta_k}(\eta_k)\rangle_{x_{k+1}} < 0$, it follows from (36) that

$$\langle \nabla f(x_{k+1}), \eta_{k+1}\rangle_{x_{k+1}} = -\frac{\langle \nabla f(x_k), \eta_k\rangle_{x_k} + r_{k+1}\langle \nabla f(x_{k+1}), \mathcal{T}_{\alpha_k\eta_k}(\eta_k)\rangle_{x_{k+1}}}{\langle \nabla f(x_k), \eta_k\rangle_{x_k}} \|\nabla f(x_{k+1})\|^2_{x_{k+1}},$$

which also gives $\langle \nabla f(x_{k+1}), \eta_{k+1}\rangle_{x_{k+1}} < 0$. Therefore (35) holds for $k + 1$ in both cases. By induction, we conclude that (35) is true for all k. □

Lemma 6 *Suppose that f satisfies Assumption 1. Then there exists a positive constant $\mu > 0$ for Algorithm 2 such that*

$$\alpha_k \geq \min\left\{\alpha_k^{\mathrm{init}}, -\mu\frac{\langle \nabla f(x_k), \eta_k\rangle_{x_k}}{\|\eta_k\|^2_{x_k}}\right\} \quad \textit{for all } k, \tag{37}$$

where α_k^{init} denotes the initial steplength at iteration k. Furthermore, we have

$$\sum_{j\geq 1} \min_{i=1,\cdots,m}\{-\alpha_{mj+i-2}\langle \nabla f(x_{mj+i-2}), \eta_{mj+i-2}\rangle_{x_{mj+i-2}}\} < +\infty. \tag{38}$$

Proof Since R is a retraction, we have from Definition 1 that

$$\mathrm{D}(f\circ R_{x_k})(0_{x_k})[\eta_k] = \frac{\mathrm{d}}{\mathrm{d}t} f(R_{x_k}(t\eta_k))\Big|_{t=0} = \mathrm{D}f(x_k)[\mathrm{D}R_{x_k}(0_{x_k})[\eta_k]] = \mathrm{D}f(x_k)[\eta_k] = \langle \nabla f(x_k), \eta_k\rangle_{x_k},$$

whose value is negative by (35). It then follows from Taylor's theorem that

$$\begin{aligned}
f(R_{x_k}(\alpha_k\eta_k)) - f(x_k) &= f(R_{x_k}(\alpha_k\eta_k)) - f(R_{x_k}(0_{x_k})) \\
&= \alpha_k\langle \nabla f(x_k), \eta_k\rangle_{x_k} + \int_0^{\alpha_k} (\mathrm{D}(f\circ R_{x_k})(t\eta_k)[\eta_k] - \mathrm{D}(f\circ R_{x_k})(0_{x_k})[\eta_k])\,\mathrm{d}t \\
&\leq \alpha_k\langle \nabla f(x_k), \eta_k\rangle_{x_k} + \int_0^{\alpha_k} \left|\mathrm{D}(f\circ R_{x_k})(t\eta_k)[\eta_k] - \mathrm{D}(f\circ R_{x_k})(0_{x_k})[\eta_k]\right| \mathrm{d}t \\
&\leq \alpha_k\langle \nabla f(x_k), \eta_k\rangle_{x_k} + \frac{1}{2}L\alpha_k^2\|\eta_k\|^2_{x_k},
\end{aligned}$$

where the last inequality uses the Lipschitzian property in Assumption 1. Thus

$$\alpha_k \leq \frac{2(\delta-1)\langle \nabla f(x_k), \eta_k\rangle_{x_k}}{L\|\eta_k\|^2_{x_k}}$$

implies

$$f(R_{x_k}(\alpha_k\eta_k)) \leq f(x_k) + \delta\alpha_k\langle \nabla f(x_k), \eta_k\rangle_{x_k}.$$

It therefore follows from (33) that (37) holds for $\mu = \frac{2(1-\delta)\lambda}{L}$.

The proof of the second part is the same as that of Theorem 3.2 in [5]. To make this lemma self-contained, we prove it as follows. For $j \geq 0$, define

$$F_j = \max\{f(x_{mj}), f(x_{mj+1}), \cdots, f(x_{mj+m-1})\}.$$

We show by induction that

$$f(x_{mj+i-1}) \leq F_{j-1} + \delta\alpha_{mj+i-2}\langle \nabla f(x_{mj+i-2}), \eta_{mj+i-2}\rangle_{x_{mj+i-2}} \text{ for all } i = 1, \cdots, m. \tag{39}$$

By (33), it is easy to see that (39) holds for $i = 1$. Assume (39) holds for all $i = 1, \cdots, l$ for some $1 \leq l \leq m-1$. Then it follows from (35) that

$$f(x_{mj+i-1}) \leq F_{j-1} \text{ for all } i = 1, \cdots, l.$$

This, together with (33), implies that

$$\begin{aligned} f(x_{mj+l}) &\leq \max\{f(x_{mj+l-1}), f(x_{mj+l-2}), \cdots, f(x_{mj+l-m})\} + \delta\alpha_{mj+l-1}\langle \nabla f(x_{mj+l-1}), \eta_{mj+l-1}\rangle_{x_{mj+l-1}} \\ &\leq F_{j-1} + \delta\alpha_{mj+l-1}\langle \nabla f(x_{mj+l-1}), \eta_{mj+l-1}\rangle_{x_{mj+l-1}}, \end{aligned}$$

which means (39) holds for $l+1$ and therefore holds for $i = 1, \cdots, m$. By (39) and the definition of F_j, we have

$$F_j \leq F_{j-1} + \delta \max_{i=1,\cdots,m} \{\alpha_{mj+i-2}\langle \nabla f(x_{mj+i-2}), \eta_{mj+i-2}\rangle_{x_{mj+i-2}}\}. \tag{40}$$

Since Assumption 1 implies $\{F_j\}$ is bounded below, we deduce by summing (40) over j that (38) is true. □

Now we are able to state our main global convergence result.

Theorem 1 *Suppose Assumptions 1 and 2 hold and Algorithm 2 does not terminate in finitely many iterations. Then the sequence $\{x_k\}$ generated by Algorithm 2 converges in the sense that*

$$\liminf_{k\to\infty} \|\nabla f(x_k)\|_{x_k} = 0. \tag{41}$$

Hence there exists at least one accumulation point which is a first-order critical point.

Proof We prove this theorem by contradiction. Suppose $\liminf_{k\to\infty} \|\nabla f(x_k)\|_{x_k} \neq 0$. This means there exists a constant $\gamma \in (0, 1)$ such that

$$\|\nabla f(x_k)\|_{x_k} \geq \gamma \quad \text{for all } k. \tag{42}$$

It can be seen from (36) that the formula (32) of β^{D}_{k+1} can be rewritten as

$$\begin{aligned} \beta^{\mathrm{D}}_{k+1} &= \frac{\langle \nabla f(x_{k+1}), \eta_{k+1}\rangle_{x_{k+1}}}{r_{k+1}\langle \nabla f(x_{k+1}), \mathcal{T}_{\alpha_k\eta_k}(\eta_k)\rangle_{x_{k+1}} - \max\{\langle \nabla f(x_{k+1}), \mathcal{T}_{\alpha_k\eta_k}(\eta_k)\rangle_{x_{k+1}} - \langle \nabla f(x_k), \eta_k\rangle_{x_k}, -\langle \nabla f(x_k), \eta_k\rangle_{x_k}\}} \\ &= \frac{\langle \nabla f(x_{k+1}), \eta_{k+1}\rangle_{x_{k+1}}}{\langle \nabla f(x_k), \eta_k\rangle_{x_k} - \max\{(1-r_{k+1})\langle \nabla f(x_{k+1}), \mathcal{T}_{\alpha_k\eta_k}(\eta_k)\rangle_{x_{k+1}}, -r_{k+1}\langle \nabla f(x_{k+1}), \mathcal{T}_{\alpha_k\eta_k}(\eta_k)\rangle_{x_{k+1}}\}}, \end{aligned}$$

which, together with the results of Lemma 5, yields

$$0 \leq \beta_{k+1} \leq \beta^{\mathrm{D}}_{k+1} \leq \frac{\langle \nabla f(x_{k+1}), \eta_{k+1}\rangle_{x_{k+1}}}{\langle \nabla f(x_k), \eta_k\rangle_{x_k}}. \tag{43}$$

Using (4), we have

$$\|\eta_{k+1}\|^2_{x_{k+1}} = -2\langle \nabla f(x_{k+1}), \eta_{k+1}\rangle_{x_{k+1}} - \|\nabla f(x_{k+1})\|^2_{x_{k+1}} + \beta^2_{k+1}\|\mathcal{T}_{\alpha_k\eta_k}(\eta_k)\|^2_{x_{k+1}}. \tag{44}$$

Dividing (44) by $\langle \nabla f(x_{k+1}), \eta_{k+1}\rangle^2_{x_{k+1}}$ and using (43), we obtain

$$\begin{aligned} \frac{\|\eta_{k+1}\|^2_{x_{k+1}}}{\langle \nabla f(x_{k+1}), \eta_{k+1}\rangle^2_{x_{k+1}}} &\leq -\frac{2}{\langle \nabla f(x_{k+1}), \eta_{k+1}\rangle_{x_{k+1}}} - \frac{\|\nabla f(x_{k+1})\|^2_{x_{k+1}}}{\langle \nabla f(x_{k+1}), \eta_{k+1}\rangle^2_{x_{k+1}}} + \frac{\|\mathcal{T}_{\alpha_k\eta_k}(\eta_k)\|^2_{x_{k+1}}}{\langle \nabla f(x_k), \eta_k\rangle^2_{x_k}} \\ &\leq -\frac{2}{\langle \nabla f(x_{k+1}), \eta_{k+1}\rangle_{x_{k+1}}} - \frac{\|\nabla f(x_{k+1})\|^2_{x_{k+1}}}{\langle \nabla f(x_{k+1}), \eta_{k+1}\rangle^2_{x_{k+1}}} + \frac{\|\eta_k\|^2_{x_k}}{\langle \nabla f(x_k), \eta_k\rangle^2_{x_k}} \\ &= -\left(\frac{\|\nabla f(x_{k+1})\|_{x_{k+1}}}{\langle \nabla f(x_{k+1}), \eta_{k+1}\rangle_{x_{k+1}}} + \frac{1}{\|\nabla f(x_{k+1})\|_{x_{k+1}}}\right)^2 + \frac{1}{\|\nabla f(x_{k+1})\|^2_{x_{k+1}}} + \frac{\|\eta_k\|^2_{x_k}}{\langle \nabla f(x_k), \eta_k\rangle^2_{x_k}} \\ &\leq \frac{1}{\|\nabla f(x_{k+1})\|^2_{x_{k+1}}} + \frac{\|\eta_k\|^2_{x_k}}{\langle \nabla f(x_k), \eta_k\rangle^2_{x_k}}, \end{aligned} \tag{45}$$

where the second inequality is guaranteed by Assumption 2.

Combining the recursion (45) and the assumption (42), we obtain

$$\frac{\|\eta_k\|_{x_k}^2}{\langle \nabla f(x_k), \eta_k\rangle_{x_k}^2} \leq \sum_{i=1}^{k} \frac{1}{\|\nabla f(x_i)\|_{x_i}^2} + \frac{\|\eta_0\|_{x_0}^2}{\langle \nabla f(x_0), \eta_0\rangle_{x_0}^2} = \sum_{i=1}^{k} \frac{1}{\|\nabla f(x_i)\|_{x_i}^2} + 1 \leq \frac{k+1}{\gamma^2}. \tag{46}$$

Then

$$\frac{\langle \nabla f(x_k), \eta_k\rangle_{x_k}^2}{\|\eta_k\|_{x_k}^2} \geq \frac{\gamma^2}{k+1}. \tag{47}$$

On the other hand, from the second inequality in (45) and (46), we have

$$\frac{\|\nabla f(x_k)\|_{x_k}^2}{\langle \nabla f(x_k), \eta_k\rangle_{x_k}^2} + \frac{2}{\langle \nabla f(x_k), \eta_k\rangle_{x_k}} \leq \frac{\|\eta_{k-1}\|_{x_{k-1}}^2}{\langle \nabla f(x_{k-1}), \eta_{k-1}\rangle_{x_{k-1}}^2} \leq \frac{k}{\gamma^2},$$

which yields

$$\frac{k}{\gamma^2}\langle \nabla f(x_k), \eta_k\rangle_{x_k}^2 \geq 2\langle \nabla f(x_k), \eta_k\rangle_{x_k} + \|\nabla f(x_k)\|_{x_k}^2. \tag{48}$$

If

$$-\langle \nabla f(x_k), \eta_k\rangle_{x_k} \leq \frac{3}{8}\|\nabla f(x_k)\|_{x_k}^2,$$

it follows from (48) that

$$\langle \nabla f(x_k), \eta_k\rangle_{x_k}^2 \geq \frac{\gamma^2}{4k}\|\nabla f(x_k)\|_{x_k}^2.$$

Therefore, we have from (35) and (42) that

$$-\langle \nabla f(x_k), \eta_k\rangle_{x_k} \geq \min\left\{\frac{3\gamma^2}{8}, \frac{\gamma^2}{2\sqrt{k}}\right\}. \tag{49}$$

Using $\alpha_k^{\text{init}} \geq \alpha_{\min}$, (37), (47) and (49), we have

$$\begin{aligned}
&\sum_{j\geq 1} \min_{i=1,\cdots,m} \{-\alpha_{mj+i-2}\langle \nabla f(x_{mj+i-2}), \eta_{mj+i-2}\rangle_{x_{mj+i-2}}\} \\
&\geq \sum_{j\geq 1} \min_{i=1,\cdots,m} \min\left\{-\alpha_{mj+i-2}^{\text{init}}\langle \nabla f(x_{mj+i-2}), \eta_{mj+i-2}\rangle_{x_{mj+i-2}}, \mu\frac{\langle \nabla f(x_{mj+i-2}), \eta_{mj+i-2}\rangle_{x_{mj+i-2}}^2}{\|\eta_{mj+i-2}\|_{x_{mj+i-2}}^2}\right\} \\
&\geq \sum_{j\geq 1} \min_{i=1,\cdots,m} \min\left\{\frac{3\gamma^2\alpha_{\min}}{8}, \frac{\gamma^2\alpha_{\min}}{2\sqrt{mj+i-2}}, \frac{\mu\gamma^2}{mj+i-1}\right\} \\
&\geq \sum_{j\geq 1} \min\left\{\frac{3\gamma^2\alpha_{\min}}{8}, \frac{\gamma^2\alpha_{\min}}{2\sqrt{m(j+1)}}, \frac{\mu\gamma^2}{m(j+1)}\right\} \\
&= +\infty,
\end{aligned}$$

which contradicts (38). Therefore (42) is impossible and (41) follows.

The existence of a first-order critical accumulation point follows immediately from the compactness of the manifold in Assumption 1. □

5 Numerical results

In this section, we demonstrate the effectiveness of Algorithm 1 on a variety of test problems. Because we focus on low-rank problems in this paper, only problems where $p \ll n$ are considered in our numerical experiments. We implemented Algorithm 1 in MATLAB (Release 2015a) and all the experiments were performed on a ThinkPad X240s laptop with an Intel Core i7-4500U processor and 8GB of RAM.

5.1 Algorithmic issues

According to Algorithm 1, the parameter β_{k+1} is chosen from the range $[0, \beta_{k+1}^{\mathrm{D}}]$. In our implementation, we choose β_{k+1} by means of

$$\beta_{k+1} = \min\left\{\beta_{k+1}^{\mathrm{D}}, \beta_{k+1}^{\mathrm{FR}}\right\},$$

where

$$\beta_{k+1}^{\mathrm{FR}} = \frac{\|\nabla f(X_{k+1})\|_F^2}{\|\nabla f(X_k)\|_F^2}$$

is the Fletcher-Reeves parameter. The initial steplength $\alpha_{k+1}^{\mathrm{init}}$ at iteration $k+1$ is set to be

$$\alpha_{k+1}^{\mathrm{init}} = \max\left\{\min\left\{\alpha_{k+1}^{\mathrm{BB}}, \alpha_{\max}\right\}, \alpha_{\min}\right\},$$

where

$$\alpha_{k+1}^{\mathrm{BB}} = \frac{\mathrm{tr}(S_k^\top S_k)}{\left|\mathrm{tr}(Y_k^\top S_k)\right|}$$

with $S_k = \alpha_k Z_k$ being the displacement of the variable in the tangent space and $Y_k = \nabla f(X_{k+1}) - \nabla f(X_k)$ being the difference of gradients, is a Riemannian generalization of the Barzilai-Borwein steplength [4]. Here we do not use a vector transport for Y_k and just compute Y_k as subtraction of two gradients in different tangent spaces because using a vector transport might increase CPU time.

We use the same stopping criterion as that of Wen and Yin's algorithm called `OptStiefelGBB` [35]: we let Algorithm 1 run up to K iterations and stop it at iteration $k < K$ if $\left\|\frac{\partial}{\partial X}\mathcal{L}(X_k, G_k^\top X_k)\right\|_F \leq \epsilon$, or $\mathrm{tol}_k^x \leq \epsilon_x$ and $\mathrm{tol}_k^f \leq \epsilon_f$, or

$$\mathrm{mean}\left\{\mathrm{tol}_{k-\min\{k,T\}+1}^x, \cdots, \mathrm{tol}_k^x\right\} \leq 10\epsilon_x \quad \text{and} \quad \mathrm{mean}\left\{\mathrm{tol}_{k-\min\{k,T\}+1}^f, \cdots, \mathrm{tol}_k^f\right\} \leq 10\epsilon_f$$

for some constants $\epsilon, \epsilon_x, \epsilon_f \in (0,1)$ and $T, K \in \mathbb{N}^+$, where

$$\mathrm{tol}_k^x = \frac{\|X_k - X_{k-1}\|_F}{\sqrt{n}} \quad \text{and} \quad \mathrm{tol}_k^f = \frac{|f(X_k) - f(X_{k-1})|}{|f(X_{k-1})| + 1}.$$

Note that $\frac{\partial}{\partial X}\mathcal{L}(X, G^\top X) = G - XG^\top X$ is actually the Riemannian gradient of $f(X)$ under the canonical metric. So we also denote this quantity by $\nabla_c f(X)$. From (34), $\nabla_c f(X)$ and $\nabla f(X)$ have the relation

$$\nabla_c f(X) = \left(I + XX^\top\right)\nabla f(X).$$

Let us now say a few words about the two metrics. The Euclidean metric is a standard metric in the view of the embedded submanifold geometry of the Stiefel manifold. The canonical metric is precisely the metric derived from the quotient space structure of the Stiefel manifold. The former is more extrinsic since it is Euclidean and the latter is more intrinsic since it weighs the independent degrees of freedom of the tangent vector equally. The reader is referred to [9] for more details on the two metrics.

Considering numerical instability of the update scheme (15) caused by the Sherman-Morrison-Woodbury formula, we restore the computed solutions X_k using a modified Gram-Schmidt process if infeasibility in the sense of $\|X_k^\top X_k - I\|_F > \epsilon_c$ for some constant $\epsilon_c > 0$ is detected after the final iteration.

Values of the input parameters for our algorithm are summarized in Table 2 as follows. All starting points X_0 were feasible and generated randomly by means of $X_0 = \mathtt{orth}(\mathtt{randn}(n, p))$.

5.2 Tests for vector transports

We compared effectiveness of Riemannian CG algorithms with various vector transports by two simple tests. The considered vector transports were the differentiated QR retraction, the scaled differentiated QR retraction, the differentiated polar retraction, the intrinsic vector transport and the proposed vector transports. For the intrinsic vector transport, we use the retraction via the Cayley transform. These algorithms are different from each other only in ways of constructing retractions and vector transports.

Table 2: Summary of input parameters for Algorithm 1

Parameter	Value	Description
ϵ	10^{-6}	tolerance for the norm of gradients
ϵ_x	10^{-6}	tolerance for the difference of variables
ϵ_f	10^{-12}	tolerance for the difference of function values
ϵ_c	10^{-13}	tolerance for constraint violation
δ	10^{-4}	nonmonotone line search Armijo-Wolfe constant
m	2	nonmonotone line search backward integer for consecutive previous function values
λ	0.2	steplength shrinkage factor
$\alpha_{\max}$	1	upper threshold for steplength
$\alpha_{\min}$	10^{-20}	lower threshold for steplength
α_0	10^{-3}	initial steplength
T	5	backward integer for consecutive previous differences of variables and function values
K	1000	maximal number of iterations

Our first test problem was the linear eigenvalue problem formulated as

$$\max_{X\in\mathbb{R}^{n\times p}} \operatorname{tr}(X^\top AX) \quad \text{s.t.} \quad X^\top X = I,$$

where A is a symmetric n-by-n matrix. The corresponding objective function and its gradient are

$$f(X) = -\operatorname{tr}(X^\top AX) \quad \text{and} \quad G = \frac{\partial f(X)}{\partial X} = -2AX.$$

In our experiments, the problem size is $n = 1000$ and $p = 5$. We considered two types of A. One was $A = \texttt{diag}(1, 2, \cdots, 1000)$ and the other was $A = M^\top M$, where M is generated randomly by $\texttt{randn}(1000)$.

Our second test problem was the orthogonal Procrustes problem [11] formulated as

$$\min_{X\in\mathbb{R}^{n\times p}} \|AX - B\|_F^2 \quad \text{s.t.} \quad X^\top X = I,$$

where $A \in \mathbb{R}^{l\times n}$ and $B \in \mathbb{R}^{l\times p}$. The corresponding objective function and its gradient are

$$f(X) = \operatorname{tr}(X^\top A^\top AX - 2B^\top AX) \quad \text{and} \quad G = \frac{\partial f(X)}{\partial X} = 2A^\top AX - 2A^\top B.$$

We set $n = 1000$ and $p = 5$ and considered two types of A and B. One is $A = I$ and $B = \texttt{ones}(1000, 5)/\texttt{sqrt}(1000)$ and the other is $A = \texttt{randn}(1000)$ and $B = \texttt{randn}(1000, 5)$.

An average of results of 10 random tests of these algorithms is reported in Tables 3 and 4, where "`Algor.1a`" and "`Algor.1b`" represent Algorithm 1 using vector transports (18) and (23) respectively, "`CGQR`", "`CGQRs`", "`CGPol`", "`CGInt`" represent the counterparts of Algorithm 1 using QR, scaled QR, polar decomposition, and the intrinsic vector transport respectively, "fval" means the objective function value, "feasi" means the feasibility $\|X^\top X - I\|_F$, "nrmg" and "rnrmg" mean the norm of the gradient $\|\nabla_c f(X)\|_F = \|\frac{\partial}{\partial X}\mathcal{L}(X, G^\top X)\|_F$ and the relative norm of the gradient respectively, "itr" means the number of iterations, "nfe" means the number of function evaluations, and "time" means the CPU time.

We observe that `CGQR` and `CGQRs` were efficient for the first problem but powerless for the second problem, while `CGPol` performed quite the opposite. The intrinsic vector transport of `CGInt` was found to be more time consuming than its theoretical prediction (see Table 1). The reason may be that the many loops in the procedure of the intrinsic vector transport made `CGInt` numerically ineffective in MATLAB. Overall, the two implementations of Algorithm 1, `Algor.1a` and `Algor.1b`, were the best ones among all these algorithms. In addition, we present in Figures 1 and 2 the plots of optimality (relative norm of the gradient) vs iteration of one particular test to compare further the efficiency of these algorithms. We see that `Algor.1a` and `Algor.1b` outperformed the other four competitors in general.

Failure of the Ring-Wirth condition is a potential cause of divergence for Riemannian CG methods, we also show in Figure 3 the sequence of ρ_k^{100}, where $\rho_k = \frac{\|\mathcal{T}_{\alpha_k Z_k}(Z_k)\|_F}{\|Z_k\|_F}$ is the expansive ratio, by an extra linear eigenvalue

Table 3: Performance of various vector transports on the linear eigenvalue problem

Algorithm	fval	feasi	nrmg	rnrmg	itr	nfe	time
fixed data matrix A							
`CGQR`	-4.9900e+03	6.86e-16	2.01e-03	1.56e-06	267.8	516.7	0.21
`CGQRs`	-4.9900e+03	6.61e-16	3.76e-03	2.93e-06	274.1	532.1	0.23
`CGPol`	-4.9855e+03	4.22e-16	3.75e+01	2.91e-02	1000	4726.3	0.96
`CGInt`	-4.9900e+03	5.72e-15	2.59e-01	2.02e-04	1000	1498.6	2.17
`Algor.1a`	-4.9900e+03	3.97e-15	7.90e-03	6.13e-06	227.9	351.5	0.18
`Algor.1b`	-4.9900e+03	3.44e-15	5.87e-03	4.56e-06	238.3	364.2	0.18
random data matrix A							
`CGQR`	-1.9397e+04	1.85e-15	2.40e-02	5.38e-06	76.3	99.4	0.53
`CGQRs`	-1.9397e+04	1.93e-15	2.57e-02	5.75e-06	76.3	97.6	0.53
`CGPol`	-1.9397e+04	2.26e-15	3.47e+00	7.83e-04	657.6	3098.7	15.34
`CGInt`	-1.9397e+04	3.51e-15	1.74e-02	3.92e-06	106.0	151.3	0.94
`Algor.1a`	-1.9397e+04	4.05e-15	1.67e-02	3.77e-06	74.2	94.5	0.51
`Algor.1b`	-1.9397e+04	4.12e-15	2.15e-02	4.85e-06	73.8	95.3	0.50

Table 4: Performance of various vector transports on the orthogonal Procrustes problem

Algorithm	fval	feasi	nrmg	rnrmg	itr	nfe	time
fixed data matrices A and B							
`CGQR`	1.4449	2.09e-15	3.44e+00	7.69e-01	1000	2972.9	0.869
`CGQRs`	1.7141	2.04e-15	3.75e+00	8.38e-01	1000	2943.4	0.924
`CGPol`	0.5279	2.09e-15	3.76e-05	8.41e-06	18.0	28.0	0.015
`CGInt`	0.5279	5.06e-14	4.13e-07	9.23e-08	37.8	41.7	0.077
`Algor.1a`	0.5279	1.20e-14	8.38e-07	1.87e-07	18.7	19.7	0.015
`Algor.1b`	0.5279	1.14e-14	1.74e-06	3.89e-07	17.9	18.9	0.013
random data matrices A and B							
`CGQR`	-3.1723e+03	2.16e-15	1.74e+03	2.74e-01	1000	2711.6	4.27
`CGQRs`	-3.0483e+03	2.04e-15	1.40e+03	2.20e-01	1000	2710.0	4.32
`CGPol`	-4.1687e+03	1.62e-15	8.31e-03	1.31e-06	53.2	62.8	0.12
`CGInt`	-4.1687e+03	2.55e-15	1.71e-02	2.71e-06	59.5	76.3	0.24
`Algor.1a`	-4.1687e+03	2.48e-15	5.19e-03	8.18e-07	54.7	66.2	0.13
`Algor.1b`	-4.1687e+03	2.49e-15	6.98e-03	1.10e-06	55.1	66.4	0.13

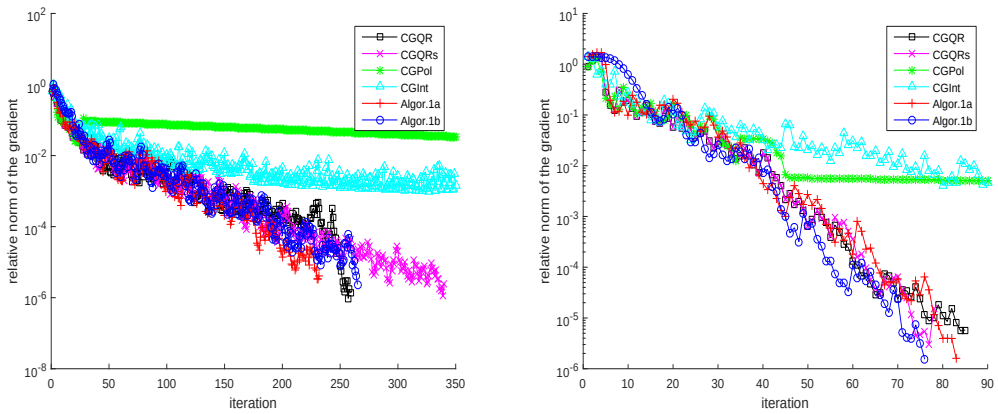

Figure 1: Relative norm of the gradient on the linear eigenvalue problem (The left is the fixed case and the right is the random case)

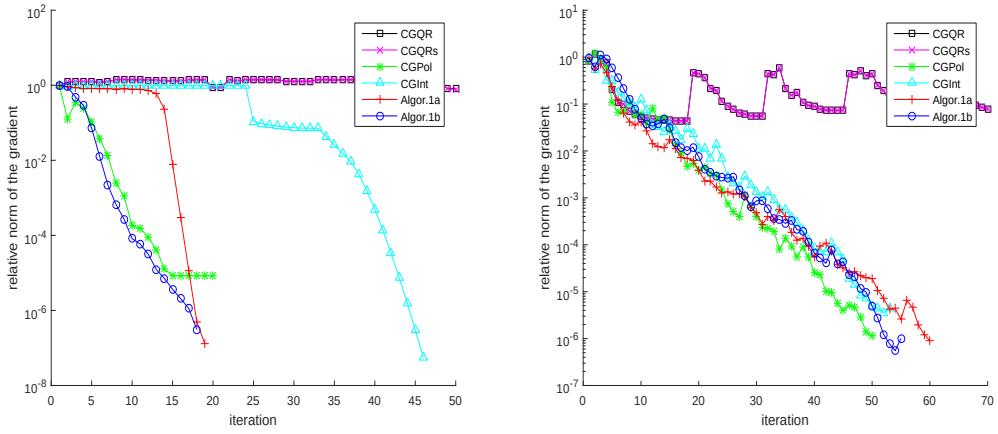

Figure 2: Relative norm of the gradient on the orthogonal Procrustes problem (The left is the fixed case and the right is the random case)

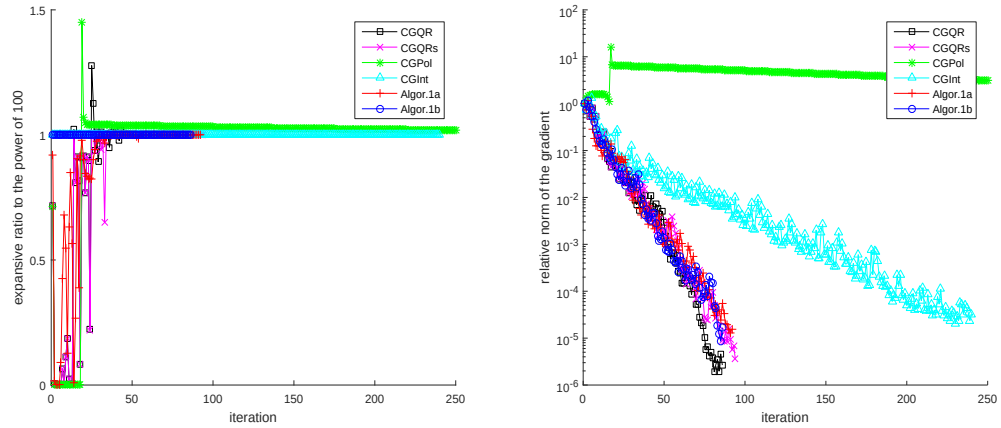

Figure 3: Expansive ratio of vector transport (left) and optimality vs iteration (right)

problem, where $n = 100$, $p = 5$ and $A = \texttt{diag}(901, 902, \cdots, 1000)$. This expansive ratio of `CGQR` and `CGPol` exceeded 1 for several times and the ratios of another four algorithms agreed with theory. The corresponding optimality vs iteration plot is also presented.

Since `CGQRs` is expected to converge even for the orthogonal Procrustes problem, we now check for `CGQRs` whether the Wolfe conditions (6) and (7b) with $c_1 = 10^{-4}$ and $c_2 = 1 - 10^{-4}$ are satisfied or not in the case of fixed A and B. The average proportion of the iterations that satisfies the Wolfe conditions is 73.95%. So, we continue to check whether `CGQRs` converges if we enforce it to satisfy the Wolfe conditions (6) and (7b). We denote this modification of `CGQRs` by `CGQRsw`. The initial stepsize of the successive iteration of `CGQRsw` was reset to 10^{-1}, a constant, because the BB stepsize had been shown by numerical experiments to be not suitable for the Wolfe conditions. An average of the results of `CGQRsw` is as follows: feval = 2.3426, feasi = 1.97×10^{-15}, nrmg = 4.01, rnrmg = 0.896, itr = 12.2, nfe = 114.1 and time = 0.046. We see that `CGQRsw` still did not succeed although it terminated within 1000 iterations. A major reason may be the Wolfe stepsizes are too small to make a significant progress. Therefore we recommend a more relaxed step acceptance criterion such as the nonmonotone condition (33). Recall that another advantage of (33) over the Wolfe conditions is that no vector transports need to be computed, while (7a) or (7b) in the Wolfe conditions needs to compute a vector transport.

5.3 Further numerical performance

Now we show more numerical results of Algorithm 1 in comparison with `OptStiefelGBB`, a state-of-art algorithm proposed by Wen and Yin [35]. `OptStiefelGBB` is a Riemannian gradient-type algorithm that also uses the Cayley transform as a retraction on the Stiefel manifold. One thing we emphasize here is that `OptStiefelGBB` provides an option of using the canonical metric or the Euclidean metric. We implemented `OptStiefelGBB` under the Euclidean metric for equity in comparison. Because Zhang and Hager's nonmonotone technique [38] used in `OptStiefelGBB` is extremely powerful, we also show the results of a modification of `Algor.1b` where the standard monotone technique is replaced with Zhang and Hager's technique. We denote this modification by `Algor.1b+ZH`.

Whether a more sophisticated line search procedure such as the polynomial interpolation line search algorithm A6.3.1mod in [7] would improve the Riemannian CG method is not very clear, but experience tells us that the backtracking line search procedure may be preferred in large-scale optimization for its simple calculations. Furthermore, the results below show that backtracking is good enough since the average number of times of calculating α_k at each iteration is less than 2.

Three types of problems were considered:

(i) The orthogonal Procrustes problem, where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times p}$ were random matrices generated by $A = \texttt{rand}(n)/\texttt{sqrt}(n)$ and $B = \texttt{rand}(n, p)$.

(ii) The heterogeneous quadratics minimization problem [3] formulated as

$$\min_{X\in\mathbb{R}^{n\times p}} \sum_{i=1}^{p} X_{[i]}^\top A_i X_{[i]} \quad \text{s.t.} \quad X^\top X = I, \tag{50}$$

where A_is are $n\times n$ symmetric matrices and $X_{[i]}$ denotes the ith column of X. The corresponding objective function and its gradient are

$$f(X) = \sum_{i=1}^{p} X_{[i]}^\top A_i X_{[i]} \quad \text{and} \quad G = \frac{\partial f(X)}{\partial X} = [G_{[i]}] = [2A_i X_{[i]}],$$

where $G_{[i]}$ denotes the ith column of G. The data matrices A_is were

$$A_i = \texttt{diag}\left(\frac{(i-1)n+1}{p} : \frac{1}{p} : \frac{in}{p}\right) \tag{51}$$

and

$$A_i = \texttt{diag}\left(\frac{(i-1)n+1}{p} : \frac{1}{p} : \frac{in}{p}\right) + B_i + B_i^\top, \tag{52}$$

where B_is were random matrices generated by $B_i = 0.1\texttt{randn}(n)$.

(iii) The joint diagonalization problem [32] formulated as

$$\max_{X\in\mathbb{R}^{n\times p}} \sum_{j=1}^{N} \|\mathrm{diag}(X^\top A_j X)\|_2^2 \quad \text{s.t.} \quad X^\top X = I,$$

where A_js are $n\times n$ symmetric matrices. The corresponding objective function and its gradient are

$$f(X) = -\sum_{j=1}^{N} \|\mathrm{diag}(X^\top A_j X)\|_2^2 \quad \text{and} \quad G = \frac{\partial f(X)}{\partial X} = [G_{[i]}] = \left[-4\sum_{j=1}^{N} (X_{[i]}^\top A_j X_{[i]}) A_j X_{[i]}\right].$$

The data matrices A_js were

$$A_j = \texttt{diag}\left(\sqrt{n+1}, \sqrt{n+2}, \cdots, \sqrt{2n}\right) + B_j + B_j^\top,$$

where B_js were random matrices generated by $B_j = \texttt{randn}(n)$. □

Note that the optimal value and solutions of (50) with matrices (51) can be characterized by the following proposition.

Proposition 1 *The set of optimal solutions of problem (50) with coefficient matrices (51) is exactly the set of matrices X^* having the form $X^* = \begin{bmatrix} Q \\ 0 \end{bmatrix}$, where Q is a $p \times p$ orthogonal matrix. Moreover, the corresponding optimal value is $f(X^*) = \frac{n(p-1)+p+1}{2}$.*

Proof It is easy to see from (51) that $A_{i+1} = A_i + \left(\frac{n}{p}\right) I$ for all $i = 1, \cdots, p-1$. Then (50) is equivalent to the linear eigenvalue problem

$$\min_{X \in \mathbb{R}^{n \times p}} \operatorname{tr}(X^\top A_1 X) \quad \text{s.t.} \quad X^\top X = I.$$

Hence, X is a minimum point of (50) if and only if X forms an orthonormal basis of the eigenspace associated with the p smallest eigenvalues of A_1. Since $\{e_1, e_2, \cdots, e_p\}$, where e_i denotes the ith column of the identity matrix, is a set of orthonormal eigenvectors associated with the p smallest eigenvalues of A_1, the set of optimal solutions of (50) consists of all matrices X^* having the form $X^* = \begin{bmatrix} Q \\ 0 \end{bmatrix}$, where Q is a $p \times p$ orthogonal matrix. The corresponding optimal value is obtained immediately by direct calculations. □

An average of results of 10 random tests of the considered algorithms is reported in Tables 5, 6 and 7, where the items are the same as those of Tables 3 and 4. From Table 5 we see that the two implementations of Algorithm 1 outperformed `OptStiefelGBB` and that Zhang and Hager's nonmonotone technique weakened the efficiency of our method considerably. From Table 6 we see that the two implementations of Algorithm 1 were slightly inferior to `OptStiefelGBB` and that Zhang and Hager's nonmonotone technique did not help improve the efficiency of our method. From Table 7 we see that the two implementations of Algorithm 1 were competitive with `OptStiefelGBB` and that Zhang and Hager's nonmonotone technique improved slightly the efficiency of our method. A common observation from these tables is that Algorithm 1 had less iterations but more function evaluations per iteration than `OptStiefelGBB` in general. Moreover, the average nfe/itr ratio of `OptStiefelGBB` was close to 1 and the average nfe/itr ratio of Algorithm 1 was about $1.5 \sim 1.8$. Zhang and Hager's nonmonotone technique helped reduce this ratio of our method to a certain extent.

Table 5: Numerical results on the orthogonal Procrustes problem

Algorithm	fval	feasi	nrmg	rnrmg	itr	nfe	time
			$n = 5000$ and $p = 5$				
`OptStiefelGBB`	-4.3833e+03	4.53e-15	5.90e-03	1.05e-06	273.4	276.2	8.79
`Algor.1a`	-4.3833e+03	2.92e-14	2.62e-03	4.70e-07	69.4	121.4	3.80
`Algor.1b`	-4.3833e+03	2.33e-14	4.87e-03	8.73e-07	77.0	132.1	4.09
`Algor.1b+ZH`	-4.3833e+03	2.51e-14	5.46e-03	9.74e-07	190.8	303.3	9.50
			$n = 5000$ and $p = 10$				
`OptStiefelGBB`	-6.7495e+03	6.74e-15	7.82e-03	9.88e-07	180.9	188.7	6.73
`Algor.1a`	-6.7495e+03	1.85e-14	4.24e-03	5.36e-07	71.3	126.7	4.57
`Algor.1b`	-6.7495e+03	1.24e-14	5.55e-03	7.02e-07	76.9	133.2	4.74
`Algor.1b+ZH`	-6.7495e+03	3.56e-14	1.28e-02	1.61e-06	178.7	285.7	10.19
			$n = 10000$ and $p = 5$				
`OptStiefelGBB`	-8.7501e+03	4.96e-15	1.18e-02	1.06e-06	149.1	153.2	19.78
`Algor.1a`	-8.7501e+03	2.76e-14	3.51e-03	3.16e-07	66.4	116.8	15.18
`Algor.1b`	-8.7501e+03	3.27e-14	5.11e-03	4.58e-07	66.4	116.0	14.86
`Algor.1b+ZH`	-8.7501e+03	1.09e-14	2.66e-02	2.39e-06	204.6	329.8	42.67
			$n = 10000$ and $p = 10$				
`OptStiefelGBB`	-1.3444e+04	9.98e-15	1.82e-02	1.15e-06	189.1	193.8	27.24
`Algor.1a`	-1.3444e+04	4.65e-14	9.08e-03	5.73e-07	80.7	145.4	20.10
`Algor.1b`	-1.3444e+04	3.01e-14	7.68e-03	4.86e-07	83.1	149.4	20.53
`Algor.1b+ZH`	-1.3444e+04	2.06e-14	2.75e-02	1.74e-06	193.5	320.7	46.05

To compare the efficiency of algorithms more fairly, we also present in Figures 4, 5 and 6 the corresponding plots of optimality vs iteration of one particular test. From Figure 4 we see that `Algor.1a` and `Algor.1b` went down more steeply than the other two. From Figures 5 and 6 we see that the four competitors were basically in the same rate of reduction.

Table 6: Numerical results on the heterogeneous quadratics minimization problem

Algorithm	fval	feasi	nrmg	rnrmg	itr	nfe	time
Fixed matrices (51): $n = 5000$ and $p = 5$							
OptStiefelGBB	1.0003e+04	4.04e-15	1.05e-02	8.15e-06	609.6	639.0	1.41
Algor.1a	1.0003e+04	1.18e-14	1.17e-02	9.03e-06	500.7	834.3	1.79
Algor.1b	1.0003e+04	1.09e-14	1.26e-02	9.76e-06	504.3	832.3	1.77
Algor.1b+ZH	1.0003e+04	1.11e-14	1.65e-02	1.28e-05	609.5	875.4	2.00
Fixed matrices (51): $n = 10000$ and $p = 10$							
OptStiefelGBB	4.5006e+04	1.23e-14	2.81e-01	1.49e-04	626.3	665.0	6.33
Algor.1a	4.5006e+04	6.75e-14	3.56e-02	1.95e-05	524.4	877.8	8.39
Algor.1b	4.5006e+04	6.73e-14	2.67e-02	1.46e-05	521.3	871.2	8.29
Algor.1b+ZH	4.5006e+04	6.75e-14	3.90e-02	2.14e-05	552.4	790.9	8.19
Random matrices (52): $n = 500$ and $p = 5$							
OptStiefelGBB	288.6093	2.18e-15	4.40e-04	4.36e-06	356.6	379.0	3.81
Algor.1a	288.6093	3.38e-15	5.36e-04	5.37e-06	289.8	456.4	4.59
Algor.1b	288.6093	3.05e-15	5.51e-04	5.52e-06	274.7	428.0	4.31
Algor.1b+ZH	288.6093	3.10e-15	5.32e-04	5.30e-06	312.7	416.4	4.19
Random matrices (52): $n = 1000$ and $p = 5$							
OptStiefelGBB	583.3786	2.83e-15	2.72e-03	1.36e-05	787.0	821.9	35.00
Algor.1a	583.3786	4.13e-15	2.18e-03	1.08e-05	623.8	1039.2	44.35
Algor.1b	583.3786	4.43e-15	1.80e-03	9.04e-06	637.4	1068.0	45.62
Algor.1b+ZH	583.3786	4.32e-15	2.42e-03	1.20e-05	745.4	1076.3	45.90

Table 7: Numerical results on the joint diagonalization problem

Algorithm	fval	feasi	nrmg	rnrmg	itr	nfe	time
$n = 500$, $p = 3$ and $N = 3$							
OptStiefelGBB	-3.6150e+04	1.08e-15	1.74e-01	1.67e-05	593.7	626.7	7.07
Algor.1a	-3.6150e+04	2.62e-15	2.18e-01	2.09e-05	466.1	760.4	8.54
Algor.1b	-3.6150e+04	2.98e-15	1.42e-01	1.36e-05	407.6	662.3	7.39
Algor.1b+ZH	-3.6150e+04	2.71e-15	1.14e-01	1.09e-05	473.1	657.0	7.35
$n = 500$, $p = 3$ and $N = 5$							
OptStiefelGBB	-4.6418e+04	8.26e-16	1.78e-01	1.30e-05	374.9	402.9	7.54
Algor.1a	-4.6418e+04	2.82e-15	1.23e-01	8.94e-06	266.1	413.3	7.76
Algor.1b	-4.6418e+04	2.93e-15	1.38e-01	1.01e-05	272.3	414.9	7.73
Algor.1b+ZH	-4.6418e+04	3.15e-15	2.08e-01	1.53e-05	280.3	362.6	6.88
$n = 1000$, $p = 3$ and $N = 3$							
OptStiefelGBB	-7.3059e+04	1.48e-15	3.04e-01	1.43e-05	446.1	475.2	22.19
Algor.1a	-7.3059e+04	3.60e-15	2.20e-01	1.05e-05	316.7	498.3	23.26
Algor.1b	-7.3059e+04	4.18e-15	2.48e-01	1.17e-05	302.5	471.5	22.07
Algor.1b+ZH	-7.3059e+04	2.85e-15	2.13e-01	1.00e-05	331.5	440.6	20.60
$n = 1000$, $p = 3$ and $N = 5$							
OptStiefelGBB	-9.2325e+04	1.43e-15	3.66e-01	1.36e-05	378.1	399.4	31.70
Algor.1a	-9.2325e+04	2.48e-15	2.36e-01	8.82e-06	275.6	431.2	33.93
Algor.1b	-9.2325e+04	2.80e-15	3.20e-01	1.19e-05	261.2	405.5	31.96
Algor.1b+ZH	-9.2325e+04	2.44e-15	1.82e-01	6.81e-06	283.9	373.5	29.41

6 Conclusions

In this paper, we present a new Riemannian conjugate gradient method for optimization on the Stiefel manifold. The main contributions are twofold. First, we propose two novel vector transports associated with the Cayley transform. The first one is the differentiated retraction of the associated retraction and the second one is an approximation to the differentiated matrix exponential. Both of them satisfy the Ring-Wirth nonexpansive condition and

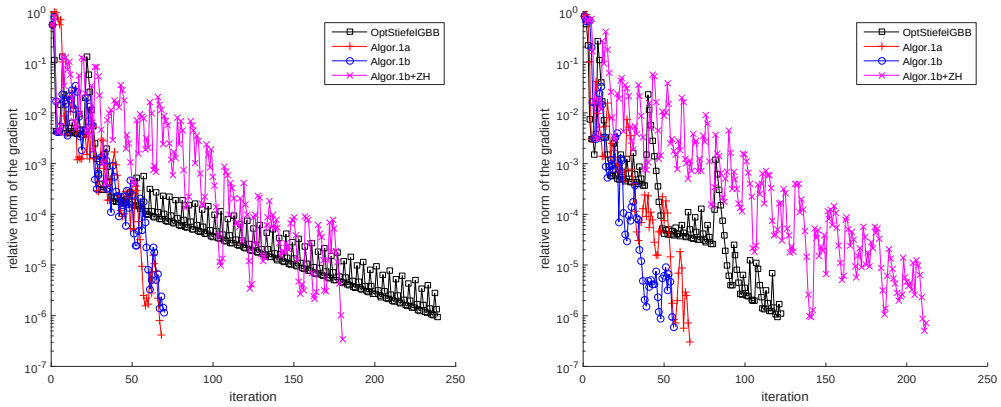

Figure 4: Relative norm of the gradient on the orthogonal Procrustes problem (The left is for $n = 5000$ and $p = 5$. The right is for $n = 10000$ and $p = 5$.)

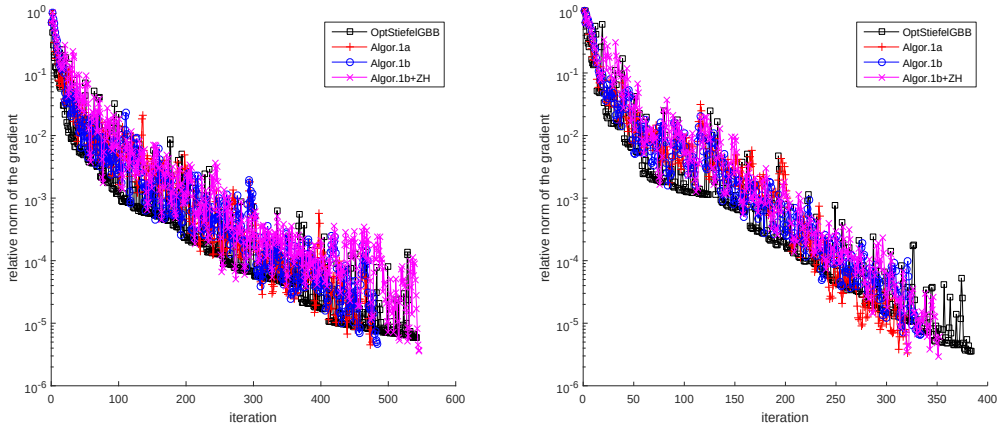

Figure 5: Relative norm of the gradient on the heterogeneous quadratics minimization problem (The left is for the fixed case with $n = 5000$. The right is for the random case with $n = 500$.)

the latter is also isometric. This enables the new algorithm to be globally convergent. Practical schemes of the new vector transports in low-rank cases are obtained. Second, we propose a Riemannian version of Dai's nonmonotone conjugate gradient method and prove its global convergence in a general Riemannian CG framework. To our knowledge, this is the first global convergence result for a nonmonotone Riemannian conjugate gradient method.

We compared our new vector transports with existing ones on some numerical tests. The results show the new vector transports are more efficient in general. We also tested the new algorithm in comparison with Wen and Yin's state-of-art Riemannian gradient algorithm on a variety of problems. The results show our algorithm is competitive with theirs. Nevertheless, current numerical performance of the proposed method is far from being satisfactory. We still pursue faster speed and higher stability for large-scale problems. One possible way to speed up the algorithm is to exploit subspace techniques. This remains under research.

A preprint of the paper with its MATLAB codes is publicly available on the web page http://www.optimization-online.org/DB_HTML/2016/09/5617.html.

Acknowledgments

The author is very grateful to the coordinating editor and two anonymous referees for their detailed and valuable comments and suggestions which helped improve the quality of this paper. The author would also like to thank Prof. Zaiwen Wen and Prof. Wotao Yin for sharing their MATLAB codes of `OptStiefelGBB` online.

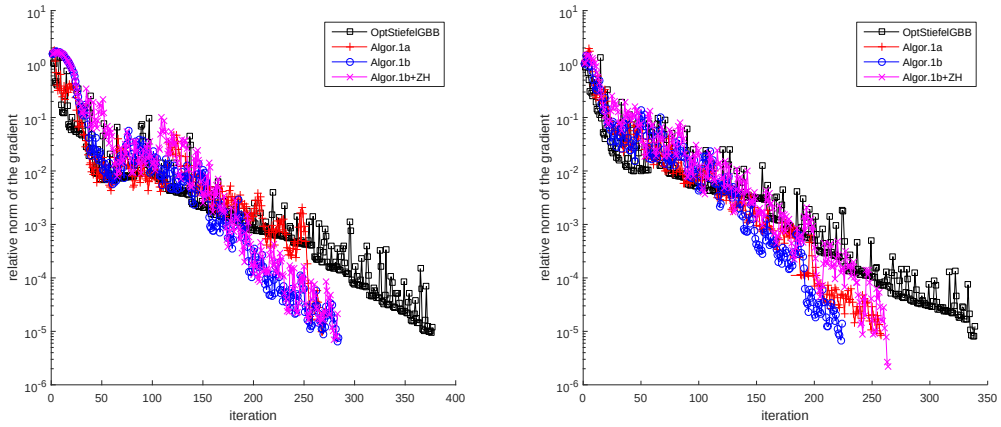

Figure 6: Relative norm of the gradient on the joint diagonalization problem (The left is for $n = 500$, $p = 3$ and $N = 5$. The right is for $n = 1000$, $p = 3$ and $N = 5$.)

References

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

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

[3] Balogh, J., Csendes, T., Rapcsák, T.: Some global optimization problems on Stiefel manifolds. J. Global Optim. 30, 91-101 (2004)

[4] Barzilai, J., Borwein, J.M.: Two-point step size gradient methods. IMA J. Numer. Anal. 8, 141-148 (1988)

[5] Dai, Y.: A nonmonotone conjugate gradient algorithm for unconstrained optimization. J. Sys. Sci. Complexity. 15, 139-145 (2002)

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

[7] Dennis, J.E., Schnabel, R.B.: Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Prentice-Hall, Englewood Cliffs, NJ, 1983. Reprinted by SIAM Publications, 1993.

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

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

[10] Golub, G.H., Van Loan, C.F.: Matrix Computations, 4th edn. The Johns Hopkins University Press, Baltimore (2013)

[11] Gower, J.C., Dijksterhuis, G.B.: Procrustes Problems, volume 30 of Oxford Statistical Science Series. Oxford University Press, Oxford (2004)

[12] I. Grubišić., R. Pietersz, Efficient rank reduction of correlation matrices, Linear Algebra Appl. 422, 629-653 (2007)

[13] Higham, N.J.: Functions of Matrices: Theory and Computation. SIAM, Philadelphia, PA, USA (2008)

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

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

[16] Huang, W., Absil, P.-A., Gallivan, K.A.: Intrinsic representation of tangent vectors and vector transports on matrix manifolds. Tech. report UCL-INMA-2016.08

[17] 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)

[18] Jiang, B., Dai, Y.: A framework of constraint preserving update schemes for optimization on Stiefel manifold. Math. Program. 153, 535-575 (2015)

[19] Li, Q., Qi, H.: A sequential semismooth Newton method for the nearest low-rank correlation matrix problem. SIAM J. Optim. 21, 1641-1666 (2011)

[20] Liu, X., Wen, Z., Zhang, Y.: Limited memory block Krylov subspace optimization for computing dominant singular value decompositions. SIAM J. Sci. Comput. 35, 1641-1668 (2013)

[21] Liu, X., Wen, Z., Wang, X., Ulbrich, M., Yuan, Y.: On the analysis of the discretized Kohn-Sham density functional theory. SIAM J. Numer. Anal. 53, 1758-1785 (2015)

[22] Ngo, T.T., Bellalij, M., Saad, Y.: The trace ratio optimization problem. SIAM Rev. 54, 545-569 (2012)

[23] Nishimori, Y., Akaho, S.: Learning algorithms utilizing quasi-geodesic flows on the Stiefel manifold. Neurocomputing 67, 106-135 (2005)

[24] Nocedal, J., Wright, S.J.: Numerical Optimization, 2nd edn. Springer, New York (2006)

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

[26] Saad, Y.: Iterative Methods for Sparse Linear Systems, second edn. SIAM, Philadelphia, PA, USA (2003)

[27] Saad, Y.: Numerical Methods for Large Eigenvalue Problems, revised edn. SIAM, Philadelphia, PA, USA (2011)

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

[29] Sato, H., Iwai, T.: A Riemannian optimization approach to the matrix singular value decompositon. SIAM J. Optim. 23, 188-212 (2013)

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

[31] Stiefel, E.: Richtungsfelder und fernparallelismus in n-dimensionalen mannigfaltigkeiten. Comment. Math. Helv. 8, 305-353 (1935)

[32] Theis, F.J., Cason, T.P., Absil, P.-A.: Soft dimension reduction for ICA by joint diagonalization on the Stiefel manifold. In: Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation, vol. 5441, pp. 354-361 (2009)

[33] Van Loan, C.F.: The ubiquitous Kronecker product. J. Comput. Appl. Math. 123, 85-100 (2000)

[34] Wen, Z., Milzarek, A., Ulbrich, M., Zhang, H.: Adaptive regularized self-consistent field iteration with exact Hessian for electronic structure calculations. SIAM J. Sci. Comput. 35, A1299-A1324 (2013)

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

[36] Wen, Z., Yang, C., Liu, X., Zhang, Y.: Trace-penalty minimization for large-scale eigenspace computation. J. Sci. Comput. 66, 1175-1203 (2016)

[37] Yuan, Y.: Subspace techniques for nonlinear optimization, in: R. Jeltsch, D.Q. Li and I. H. Sloan, eds., Some Topics in Industrial and Applied Mathematics (Series in Contemporary Applied Mathematics CAM 8) (Higher Education Press. Beijing, 2007), pp. 206-218.

[38] Zhang, H., Hager, W.W.: A nonmonotone line search technique and its application to unconstrained optimization. SIAM J. Optim. 14, 1043-1056 (2004)

[39] Zhang, X., Zhu, J., Wen, Z., Zhou, A.: Gradient type optimization methods for electronic structure calculations. SIAM J. Sci. Comput. 36, C265-C289 (2014)

[40] Zhang, L., Li, R.: Maximization of the sum of the trace ratio on the Stiefel manifold, I: Theory. Sci. China Math. 57, 2495-2508 (2014)

[41] Zhang, L., Li, R.: Maximization of the sum of the trace ratio on the Stiefel manifold, II: Computation. Sci. China Math. 58, 1549-1566 (2015)

[42] Zhu, X.: A feasible filter method for the nearest low-rank correlation matrix problem. Numer. Algor. 69, 763-784 (2015)

[43] Zhu, X., Duan, C: Gradient methods with approximate exponential retractions for optimization on the Stiefel manifold. A manuscript submitted to Optimization