Accelerated gradient methods on the Grassmann and Stiefel manifolds

Xiaojing Zhu[1*]

[1] *School of Mathematics and Physics, Shanghai University of Electric Power, Yangpu, Shanghai 200090, China*

Abstract In this paper we extend a nonconvex Nesterov-type accelerated gradient (AG) method to optimization over the Grassmann and Stiefel manifolds. We propose an exponential-based AG algorithm for the Grassmann manifold and a retraction-based AG algorithm that exploits the Cayley transform for both of the Grassmann and Stiefel manifolds. Under some mild assumptions, we obtain the global rate of convergence of the exponential-based AG algorithm. With additional but reasonable assumptions on retraction and vector transport, the same global rate of convergence is obtained for the retraction-based AG algorithm. Details of computing the geometric objects as ingredients of our AG algorithms are also discussed. Preliminary numerical results demonstrate the potential effectiveness of our AG methods.

Keywords: Riemannian optimization, Grassmann manifold, Stiefel manifold, accelerated gradient, global rate of convergence

Mathematics Subject Classification 49Q99 · 65K05 · 90C30 · 90C48

1 Introduction

In this paper, we consider optimization on the Grassmann and Stiefel manifolds:

$$f^* = \min_{x\in\mathcal{M}} \; f(x), \tag{1}$$

where $\mathcal{M}$ is either the Grassmann manifold $\mathrm{Gr}(n,p)$ or the Stiefel manifold $\mathrm{St}(n,p)$ and f is a differentiable function over $\mathcal{M}$. The Grassmann manifold is defined as

$$\mathrm{Gr}(n,p) := \{\mathcal{X} \subset \mathbb{R}^n \mid \mathcal{X} \text{ is a subspace, } \dim(\mathcal{X}) = p\}.$$

The Stiefel manifold is defined as

$$\mathrm{St}(n,p) := \{X \in \mathbb{R}^{n\times p} \mid X^\top X = I_p\}.$$

Since any $\mathcal{X} \in \mathrm{Gr}(n,p)$ can be represented by $\mathcal{X} = \mathrm{span}(X)$ for some $X \in \mathrm{St}(n,p)$, problem (1) is also known as optimization with orthogonality constraints [14]:

$$\min_{X\in\mathbb{R}^{n\times p}} \; f(X) \;\; \text{s.t.} \;\; X^\top X = I_p,$$

where the underlying constrained manifold is $\mathrm{Gr}(n,p)$ if $f(XQ)$ is invariant for all $p \times p$ orthogonal matrices Q. Optimization with orthogonality constraints has broad applications in science and engineering, including linear and nonlinear eigenvalue problems, low-rank matrix optimization, principal component

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

analysis, electronic structures computations, machine learning, computer vision, image processing, model reduction, etc. The reader is referred to [2, 6, 8, 14, 22, 33, 38] and references therein for concrete examples of applications. Moreover, the Grassmann and Stiefel manifolds are fundamental in Riemannian optimization partly because they are also closely related to other manifolds such as the fixed-rank manifold [9, 16], the affine Grassmann manifold [30], the symplectic Stiefel manifold [18], and flag manifolds [40].

Edelman, Arias, and Smith's work [14] is a landmark achievement in the field of optimization on the Stiefel and Grassmann manifolds. They deeply studied the geometry of the two manifolds and developed conjugate gradient (CG) and Newton methods on them. Absil, Mahoney, and Sepulchre's monograph [2] lays the foundation for general Riemannian optimization and focuses on the Stiefel and Grassmann manifolds. Wen and Yin's efficient gradient method [38] is a high benchmark in the algorithmic aspect of this field. Key work on the Grassmann manifold also includes [1, 34]. In recent years, more and more advanced algorithms for solving problem (1) have been proposed, including gradient-type methods [28, 17, 37], CG methods [43, 45], second-order methods [24, 25, 23, 21], proximal gradient methods [11, 26, 27], stochastic variance reduced gradient methods [35, 29], etc.

In this paper, we focus on accelerated gradient methods on manifolds. Nesterov's accelerated gradient (AG) method [31] is extremely effective for convex optimization in Euclidean spaces. So in recent years, many researchers have tried to extend this method to Riemannian manifolds. For geodesically convex optimization on manifolds, the $\mathcal{O}(\frac{1}{k^2})$ global rate of convergence and local linear convergence of AG methods can be established [42, 4]. But for the Grassmann and Stiefel manifolds, which are compact manifolds, geodesically convex optimization is meaningless because there is no non-trivial convex function on a complete manifold with finite volume [39]. Although in [36], the simplest form of Nesterov's AG method is proposed to solve optimization problems on the Stiefel manifold, no convergence result is obtained. Accelerated proximal gradient methods for composite optimization on manifolds (i.e., Riemannian FISTA) have also been proposed (without result of global rate of convergence) [26]. So far, in the field of nonconvex optimization on manifolds, global rates of convergence of optimization methods such as gradient methods, trust region methods, and cubic regularization methods have been known [10, 3]. In particular, gradient methods can achieve $||\nabla f(x_k)|| \le \mathcal{O}(\frac{1}{\sqrt{k}})$, which is the best-known global rate of convergence for general nonconvex smooth problems by using only first-order information (without any assumption on the Hessian).

Ghadimi and Lan proposed an accelerated gradient method for solving general nonconvex smooth optimization in Euclidean spaces [19]:

$$\begin{cases} x_k = (1-\lambda_k)y_{k-1} + \lambda_k z_{k-1}, \\ y_k = x_k - \alpha_k \nabla f(x_k), \\ z_k = z_{k-1} - \beta_k \nabla f(x_k). \end{cases} \tag{2}$$

This is a variation of Nesterov's original AG method; they are equivalent to each other for special α_k, β_k, and λ_k. By a specific stepsize policy, Ghadimi and Lan proved that their method can achieve $||\nabla f(x_k)|| \le \mathcal{O}(\frac{1}{\sqrt{k}})$, the same global rate of convergence as that of gradient methods.

We aim to extending Ghadimi and Lan's AG method (2) to optimization on the Grassmann and Stiefel manifolds. The main tools used in our extension are retraction and vector transport [2], which are natural generalizations of the exponential map and parallel transport, respectively. There are plenty of choices for retraction and vector transport on the Grassmann and Stiefel manifolds. In this paper, we choose the Cayley transform retraction and vector transport [38, 43, 45]. The principal reason is that the Cayley transform is not only practical but also rich in theory. Note that our method radically differs from the recently proposed accelerated first-order method [12]. Although the latter can achieve a faster global rate of convergence $\mathcal{O}\big(\frac{1}{k^{4/7}}\big)$, it requires assumptions on the Hessian of the objective function. Moreover, that method is somewhat conceptual because it carries out acceleration in tangent spaces, i.e., acceleration for the pullback function (the composite function of the objective function and the exponential map). This compels that method to evaluate the gradient of the pullback function, which is impractical in general.

The contributions of this paper are mainly theoretical. In the aspect of algorithmic design, we propose two novel Riemannian versions of the nonconvex AG method (2). The first algorithm, designed specially for the Grassmann manifold, is implemented with the exponential map and the parallel transport. The second algorithm, designed for both of the Grassmann and Stiefel manifolds, is implemented with the

Cayley transform retraction and vector transport. In order to obtain a practical retraction-based AG algorithm, we derive simple low-rank formulas for the inverse maps of the Cayley transform retraction and vector transport. In the aspect of convergence analysis, we prove the $||\nabla f(x_k)|| \leq \mathcal{O}(\frac{1}{\sqrt{k}})$ global rate of convergence of our AG algorithms. The exponential-based AG algorithm possesses this global rate of convergence only under mild assumptions. The retraction-based AG algorithm also possesses this global rate of convergence with additional reasonable assumptions on retraction and vector transport. To our knowledge, this is the first result of global rate of convergence for nonconvex Riemannian Nesterov-type AG methods.

The rest of this paper is organized as follows. In Section 2, we review basic geometry and optimization tools on the Grassmann and Stiefel manifolds. Our new AG algorithms are proposed in Section 3. We prove the global rate of convergence of the proposed algorithms in Section 4. Implementation details of computing geometric objects and stepsize policy are discussed in Section 5. Preliminary numerical results are shown in Section 6 and conclusions are made in Section 7.

2 Preliminaries

2.1 Basic geometry of the Grassmann and Stiefel manifolds

We review some basic geometry of the Grassmann and Stiefel manifolds according to [14, 6]. The Grassmann manifold $\mathrm{Gr}(n,p)$ and the Stiefel manifold $\mathrm{St}(n,p)$ have the following quotient manifold structures:

$$\begin{aligned}\mathrm{Gr}(n,p) &\simeq O(n)/(O(p)\times O(n-p)),\\ \mathrm{St}(n,p) &\simeq O(n)/O(n-p),\end{aligned}$$

where $O(n) := \{Q \in \mathbb{R}^{n\times n} \mid Q^\top Q = I_n\}$ is the $n\times n$ orthogonal group. In this view, $\mathrm{Gr}(n,p)$ and $\mathrm{St}(n,p)$ are quotient manifolds of $O(n)$ and $O(n)$ is the total manifold of $\mathrm{Gr}(n,p)$ and $\mathrm{St}(n,p)$. These three manifolds can be connected together via the following maps:

$$\begin{aligned}&\pi^{\mathrm{SG}} : \mathrm{St}(n,p) \to \mathrm{Gr}(n,p) : X \mapsto \mathrm{span}(X),\\ &\pi^{\mathrm{OS}} : O(n) \to \mathrm{St}(n,p) : Q \mapsto QI_{n,p},\\ &\pi^{\mathrm{OG}} = \pi^{\mathrm{SG}}\circ\pi^{\mathrm{OS}} : O(n) \to \mathrm{Gr}(n,p) : Q \mapsto \mathrm{span}(QI_{n,p}).\end{aligned}$$

Let $Q \in O(n)$, $X = \pi^{\mathrm{OS}}(Q)$, and $\mathcal{X} = \pi^{\mathrm{OG}}(Q)$. Then $\mathcal{X} = \mathrm{span}(X)$ and $Q = [X, X_\perp]$ for some $X_\perp \in \mathbb{R}^{n\times(n-p)}$ such that $X^\top X_\perp = 0$ and $X_\perp^\top X_\perp = I_{n-p}$.

Let $\mathfrak{so}(n)$ be the Lie algebra of $O(n)$, i.e.,

$$\mathfrak{so}(n) := T_{I_n}O(n) = \{\Omega \in \mathbb{R}^{n\times n} \mid \Omega^\top = -\Omega\}.$$

The tangent space at an arbitrary $Q \in O(n)$ is given by

$$T_QO(n) = \{Q\Omega \mid \Omega \in \mathfrak{so}(n)\}.$$

Let $O(n)$ be endowed with the Riemannian metric

$$\langle Q\Omega, Q\tilde{\Omega}\rangle := \frac{1}{2}\mathrm{Tr}\big((Q\Omega)^\top(Q\tilde{\Omega})\big) = \frac{1}{2}\mathrm{Tr}(\Omega^\top\tilde{\Omega}) = -\frac{1}{2}\mathrm{Tr}(\Omega\tilde{\Omega}). \tag{3}$$

Under this metric, the tangent spaces $T_{\mathcal{X}}\mathrm{Gr}(n,p)$ and $T_X\mathrm{St}(n,p)$ can be represented as

$$T_{\mathcal{X}}\mathrm{Gr}(n,p) \simeq H_Q^{\pi^{\mathrm{OG}}}O(n) = \left\{Q\begin{pmatrix}0 & -A^\top\\ A & 0\end{pmatrix} \;\middle|\; A \in \mathbb{R}^{(n-p)\times p}\right\}, \tag{4}$$

$$T_X\mathrm{St}(n,p) \simeq H_Q^{\pi^{\mathrm{OS}}}O(n) = \left\{Q\begin{pmatrix}S & -A^\top\\ A & 0\end{pmatrix} \;\middle|\; S \in \mathfrak{so}(p),\ A \in \mathbb{R}^{(n-p)\times p}\right\}, \tag{5}$$

where $H_x^{\pi}\mathcal{M}$ denotes the horizontal (tangent) space to $\mathcal{M}$ at x with respect to the quotient map π. We can also represent $T_X\mathrm{St}(n,p)$ directly as

$$T_X\mathrm{St}(n,p) = \Big\{XS + X_{\perp}A \;\Big|\; S \in \mathfrak{so}(p),\; A \in \mathbb{R}^{(n-p)\times p}\Big\}. \tag{6}$$

Therefore

$$T_{\mathcal{X}}\mathrm{Gr}(n,p) \simeq H_X^{\pi^{\mathrm{SG}}}\mathrm{St}(n,p) = \Big\{X_{\perp}A \;\Big|\; A \in \mathbb{R}^{(n-p)\times p}\Big\}. \tag{7}$$

For a quotient manifold $\mathcal{M}/\sim$, the unique $\eta_x^h \in H_x^{\pi}\mathcal{M}$ such that $\eta_{\pi(x)} := d\pi_x(\eta_x^h) \in T_{\pi(x)}(\mathcal{M}/\sim)$ is called the horizontal lift of $\eta_{\pi(x)}$ to $T_x\mathcal{M}$ at x. If $\eta_{\mathcal{X}}$ and $\xi_{\mathcal{X}}$ are two arbitrary tangent vectors in $T_{\mathcal{X}}\mathrm{Gr}(n,p)$, then by (4) and (7) we have the representations: $\eta_X^h = X_{\perp}A$, $\xi_X^h = X_{\perp}B$, $\eta_Q^h = Q\mathfrak{A}$, and $\xi_Q^h = Q\mathfrak{B}$, where

$$\mathfrak{A} := \begin{pmatrix} 0 & -A^{\top} \\ A & 0 \end{pmatrix}, \quad \mathfrak{B} := \begin{pmatrix} 0 & -B^{\top} \\ B & 0 \end{pmatrix}. \tag{8}$$

If η_X and ξ_X are two arbitrary tangent vectors in $T_X\mathrm{St}(n,p)$, then by (5) we have the representations: $\eta_Q^h = Q\mathfrak{A}$ and $\xi_Q^h = Q\mathfrak{B}$, where

$$\mathfrak{A} := \begin{pmatrix} S_{\eta} & -A^{\top} \\ A & 0 \end{pmatrix}, \quad \mathfrak{B} := \begin{pmatrix} S_{\xi} & -B^{\top} \\ B & 0 \end{pmatrix}. \tag{9}$$

The Riemannian metric (3) on $O(n)$ induces naturally the following Riemannian metrics:

$$\langle \eta_{\mathcal{X}}, \xi_{\mathcal{X}} \rangle := \langle \eta_Q^h, \xi_Q^h \rangle = \frac{1}{2}\mathrm{Tr}(\mathfrak{A}^{\top}\mathfrak{B}) = \mathrm{Tr}(A^{\top}B) = \mathrm{Tr}((\eta_X^h)^{\top}\xi_X^h) \text{ on } T_{\mathcal{X}}\mathrm{Gr}(n,p),$$

$$\begin{aligned}\langle \eta_X, \xi_X \rangle := \langle \eta_Q^h, \xi_Q^h \rangle &= \frac{1}{2}\mathrm{Tr}(\mathfrak{A}^{\top}\mathfrak{B}) \\ &= \frac{1}{2}\mathrm{Tr}(S_{\eta}^{\top}S_{\xi}) + \mathrm{Tr}(A^{\top}B) = \mathrm{Tr}\Big(\eta_X^{\top}\Big(I_n - \frac{1}{2}XX^{\top}\Big)\xi_X\Big) \text{ on } T_X\mathrm{St}(n,p).\end{aligned}$$

With these metrics and the notation $G = \left(\frac{\partial f(X)}{\partial X_{ij}}\right)$ for the Euclidean gradient of f (the derivative of f with respect to X), the Riemannian gradients ∇f on $\mathrm{Gr}(n,p)$ and $\mathrm{St}(n,p)$ have the following unified formula (in the Stiefel manifold representation):

$$\nabla f(X) = G - XG^{\top}X.$$

Note that in the Grassmannian case it also holds $\nabla f(X) = G - XX^{\top}G$ because $G^{\top}X \equiv X^{\top}G$. The above metrics also induce the following canonical norms:

$$\|\eta_{\mathcal{X}}\|_{\mathrm{c}} := \sqrt{\langle \eta_{\mathcal{X}}, \eta_{\mathcal{X}} \rangle} = \frac{1}{\sqrt{2}}\sqrt{\mathrm{Tr}(\mathfrak{A}^{\top}\mathfrak{A})} = \frac{1}{\sqrt{2}}\|\mathfrak{A}\|_{\mathrm{F}} = \|A\|_{\mathrm{F}} = \|\eta_X^h\|_{\mathrm{F}} \text{ on } T_{\mathcal{X}}\mathrm{Gr}(n,p),$$

$$\|\eta_X\|_{\mathrm{c}} := \sqrt{\langle \eta_X, \eta_X \rangle} = \frac{1}{\sqrt{2}}\sqrt{\mathrm{Tr}(\mathfrak{A}^{\top}\mathfrak{A})} = \frac{1}{\sqrt{2}}\|\mathfrak{A}\|_{\mathrm{F}} = \frac{1}{\sqrt{2}}\|S_{\eta}\|_{\mathrm{F}} + \|A\|_{\mathrm{F}} \text{ on } T_X\mathrm{St}(n,p).$$

The exponential map $\overline{\exp}$ on $O(n)$ is given by

$$\overline{\exp}_Q(Q\Omega) = Q\mathrm{expm}(\Omega) = \mathrm{expm}(Q\Omega Q^{\top})Q, \tag{10}$$

where $\mathrm{expm}(A) := \sum_{i=0}^{\infty}\frac{1}{i!}A^i$ is the matrix exponential for any square matrix A. This formula implies that the exponential maps exp on $\mathrm{Gr}(n,p)$ and $\mathrm{St}(n,p)$ (in the orthogonal group representation) can be expressed uniformly as

$$\exp_x(\eta_x) \simeq \exp_Q(\eta_Q^h) = Q\mathrm{expm}(\mathfrak{A}) = \mathrm{expm}(Q\mathfrak{A}Q^{\top})Q, \tag{11}$$

where $x = \mathcal{X} \in \mathrm{Gr}(n,p)$ or $x = X \in \mathrm{St}(n,p)$. In the case of $\mathrm{Gr}(n,p)$,

$$Q\mathfrak{A}Q^{\top} = X_{\perp}AX^{\top} - XA^{\top}X_{\perp}^{\top} = \eta_X^h X^{\top} - X\big(\eta_X^h\big)^{\top}.$$

In the case of $\mathrm{St}(n,p)$,

$$\begin{aligned}
Q\mathfrak{A}Q^\top &= XS_\eta X^\top + X_\perp AX^\top - XA^\top X_\perp^\top \\
&= \Big(I_n - \frac{1}{2}XX^\top\Big)(XS_\eta + X_\perp A)X^\top - X(XS_\eta + X_\perp A)^\top\Big(I_n - \frac{1}{2}XX^\top\Big) \\
&= \Big(I_n - \frac{1}{2}XX^\top\Big)\eta_X X^\top - X\eta_X^\top\Big(I_n - \frac{1}{2}XX^\top\Big).
\end{aligned}$$

The parallel transport of $\xi_{\mathcal{X}}$ along the geodesic $\gamma(t) := \exp_{\mathcal{X}}(t\eta_{\mathcal{X}})$ on the Grassmann manifold $\mathrm{Gr}(n,p)$ is given by

$$P_\gamma^{t\leftarrow 0}\xi_{\mathcal{X}} = Q\mathrm{expm}(\mathfrak{A})\mathfrak{B} = \mathrm{expm}(Q\mathfrak{A}Q^\top)Q\mathfrak{B}. \tag{12}$$

Unfortunately, the parallel transport on the Stiefel manifold $\mathrm{St}(n,p)$ has no closed-form formula in general.

2.2 Retraction and vector transport

In practical Riemannian optimization algorithms, the exponential map and parallel transport are often replaced by a retraction and a vector transport. The definitions of retraction and vector transport are stated as follows [2]:

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

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

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

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

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

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

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

1. There is 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 $\eta_x, \xi_x, \zeta_x \in T_x\mathcal{M}$.

Now we introduce a quite useful retraction on the Grassmann and Stiefel manifolds, the Cayley transform, and its associated vector transport.

According to [2, 38, 45], this retraction is given by

$$\overline{R}_Q(Q\Omega) := Q\phi_{\mathrm{ct}}(\Omega) = \phi_{\mathrm{ct}}(Q\Omega Q^\top)Q, \tag{13}$$

$$R_x(\eta_x) \simeq R_Q(\eta_Q^h) := Q\phi_{\mathrm{ct}}(\mathfrak{A}) = \phi_{\mathrm{ct}}(Q\mathfrak{A}Q^\top)Q, \tag{14}$$

where

$$\phi_{\mathrm{ct}}(\mathfrak{A}) := \Big(I_n - \frac{1}{2}\mathfrak{A}\Big)^{-1}\Big(I_n + \frac{1}{2}\mathfrak{A}\Big) \tag{15}$$

is commonly known as the Cayley transform.

According to [43, 45], a vector transport $\mathcal{T}$ associated with the above retraction is

$$\mathcal{T}_{\eta_x}(\xi_x) \simeq \mathcal{T}_{\eta_Q^h}(\xi_Q^h) := Q\phi_{\mathrm{ct}}(\mathfrak{A})\mathfrak{B} = \phi_{\mathrm{ct}}(Q\mathfrak{A}Q^\top)Q\mathfrak{B}. \tag{16}$$

By (15) and the skew-symmetry of $\mathfrak{A}$, it is easy to see that $\phi_{\mathrm{ct}}(\mathfrak{A})$ is orthogonal. Then $\mathcal{T}_{\eta_x}(\cdot)$ is indeed isometric with respect to both of the canonical norm $||\cdot||_{\mathrm{c}}$ and the 2-norm $||\cdot||_2$, i.e., $||\mathcal{T}_{\eta_x}(\xi_x)||_{\mathrm{c}} = ||\xi_x||_{\mathrm{c}}$ and $||\mathcal{T}_{\eta_x}(\xi_x)||_2 = ||\xi_x||_2$. The isometry of $\mathcal{T}_{\eta_x}$ implies that the inverse vector transport $\mathcal{T}_{\eta_x}^{-1}$ exists for any η_x.

Algorithm 1: Exponential-based AG method on the Grassmann manifold

```
Input: $Y_0 = Z_0 \in \mathrm{Gr}(n,p)$, $\{\alpha_k\}$, $\{\beta_k\}$, $\{\lambda_k\}$: $0 < \alpha_k \leq \beta_k$, $\lambda_1 = 1$, $\lambda_k \in (0,1)$ for $k \geq 2$.
for $k = 1, 2, \dots$ do
    Compute
        $\eta_k = (1-\lambda_k)\exp^{-1}_{Z_{k-1}}(Y_{k-1}),$   (17)
        $X_k = \exp_{Z_{k-1}}(\eta_k),$   (18)
        $Y_k = \exp_{X_k}(-\alpha_k \nabla f(X_k)),$   (19)
        $Z_k = \exp_{Z_{k-1}}(-\beta_k P_{\gamma}^{Z_{k-1} \leftarrow X_k} \nabla f(X_k)).$   (20)
```

Algorithm 2: Retraction-based AG method on the Grassmann and Stiefel manifolds

```
Input: $Y_0 = Z_0 \in \mathcal{M}$ where $\mathcal{M} := \mathrm{Gr}(n,p)$ or $\mathcal{M} := \mathrm{St}(n,p)$, $\{\alpha_k\}$, $\{\beta_k\}$, $\{\lambda_k\}$: $0 < \alpha_k \leq \beta_k$,
       $\lambda_1 = 1$, $\lambda_k \in (0,1)$ for $k \geq 2$.
for $k = 1, 2, \dots$ do
    Compute
        $\eta_k = (1-\lambda_k) R^{-1}_{Z_{k-1}}(Y_{k-1}),$   (21)
        $X_k = R_{Z_{k-1}}(\eta_k),$   (22)
        $Y_k = R_{X_k}(-\alpha_k \nabla f(X_k)),$   (23)
        $Z_k = R_{Z_{k-1}}(-\beta_k \mathcal{T}^{-1}_{\eta_k} \nabla f(X_k)).$   (24)
```

3 Accelerated gradient algorithms

In this section, we present our Riemannian generalization of the accelerated gradient method (2) for optimization on the Grassmann and Stiefel manifolds.

We propose two versions of Riemannian accelerated gradient algorithms. Algorithm 1 is designed exclusively for the Grassmann manifold. This algorithm is implemented with the exponential map (11) and the parallel transport (12). Algorithm 2 is designed for both of the Grassmann and Stiefel manifolds. This algorithm is implemented with the Cayley transform retraction (14) and its isometric vector transport (16). In Section 5, we will discuss how to efficiently compute the exponential map with its inverse (the Riemannian logarithm) and the parallel transport on the Grassmann manifold, and the Cayley transform retraction and vector transport with their inverses on both of the Grassmann and Stiefel manifolds.

Both of Algorithms 1 and 2 belong to the class of three-point-type Riemannian accelerated gradient methods, because they generate three sequences $\{X_k\}_{k\geq 1}$, $\{Y_k\}_{k\geq 1}$, and $\{Z_k\}_{k\geq 1}$. Compared with traditional two-point-type Riemannian accelerated gradient methods such as (e.g., [26, 27, 36])

$$\begin{cases} X_k = R_{Y_{k-1}}(-\alpha_k \nabla f(Y_{k-1})), \\ t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2}, \\ Y_k = R_{X_k}\big(\frac{1-t_{k-1}}{t_k} R^{-1}_{X_k}(X_{k-1})\big), \end{cases}$$

our methods need additional computational effort during each iteration, i.e., computing an inverse vector transport (or a parallel translation) and one more retraction (or exponential). However, we will show that our methods have guaranteed global rate of convergence in the next section.

4 Convergence

In this section, we prove the global rates of convergence of Algorithms 1 and 2 based on the convergence analysis in Section 2 of [19]. Before our convergence analysis, we make the following important remark, which the reader need to keep in mind throughout this section.

Remark 1 Regarding Algorithm 1, we fix an orthogonal matrix $Q_{Z_0} \in O(n)$ such that $\pi^{\mathrm{OG}}(Q_{Z_0}) = Z_0$ for the initial point Z_0. For convenience, we identify Z_0 with the specified orthogonal group representation Q_{Z_0}, and recursively identify

$$X_k \overset{(18)}{=\!=} \exp_{Z_{k-1}}(\eta_k) \quad \text{with} \quad Q_{X_k} := \overline{\exp}_{Q_{Z_{k-1}}}(\eta^h_{Q_{Z_{k-1}}}),$$

$$Y_k \overset{(19)}{=\!=} \exp_{X_k}(-\alpha_k \nabla f(X_k)) \quad \text{with} \quad Q_{Y_k} := \overline{\exp}_{Q_{X_k}}(-\alpha_k \xi^h_{Q_{X_k}}),$$

and

$$Z_k \overset{(20)}{=\!=} \exp_{Z_{k-1}}(-\beta_k P_\gamma^{Z_{k-1}\leftarrow X_k} \nabla f(X_k)) \quad \text{with} \quad Q_{Z_k} := \overline{\exp}_{Q_{Z_{k-1}}}(-\beta_k \zeta^h_{Q_{Z_{k-1}}}),$$

where $\xi^h_{Q_{X_k}}$ is the horizontal lift of $\nabla f(X_k)$ at Q_{X_k} and $\zeta^h_{Q_{Z_{k-1}}}$ is the horizontal lift of $P_\gamma^{Z_{k-1}\leftarrow X_k} \nabla f(X_k)$ at $Q_{Z_{k-1}}$. We also identify η_k with $\eta^h_{Q_{Z_{k-1}}}$, $\nabla f(X_k)$ with $\xi^h_{Q_{X_k}}$, and $P_\gamma^{Z_{k-1}\leftarrow X_k} \nabla f(X_k)$ with $\zeta^h_{Q_{Z_{k-1}}}$. Then we make similar identifications for Algorithm 2. We will see that our convergence analysis will benefit a lot from the orthogonal group representations of the geometric objects on the Grassmann and Stiefel manifolds.

The following lemma will be used in our convergence theorems later.

Lemma 1 *Let $\{\tau_k\}_{k\geq 1}$ be the sequence of numbers defined by*

$$\tau_k := \begin{cases} 1, & k = 1 \\ \prod_{i=2}^{k} (1 - \lambda_i), & k \geq 2. \end{cases} \tag{25}$$

Then $\sum_{i=1}^{k} \tau_k \frac{\lambda_i}{\tau_i} = 1$.

Proof. Using $\lambda_1 = 1$ and $1 - \lambda_i = \frac{\tau_i}{\tau_{i-1}}$, we have

$$\sum_{i=1}^{k} \tau_k \frac{\lambda_i}{\tau_i} = \tau_k \Big(\frac{\lambda_1}{\tau_1} + \sum_{i=2}^{k} \frac{1}{\tau_i} (1 - \frac{\tau_i}{\tau_{i-1}}) \Big) = \tau_k \Big(\frac{1}{\tau_1} + \sum_{i=2}^{k} (\frac{1}{\tau_i} - \frac{1}{\tau_{i-1}}) \Big) = 1. \quad \square$$

4.1 Convergence of Algorithm 1

In this subsection, we focus on Algorithm 1. Keep in mind temporarily that the manifold $\mathcal{M}$ in question is the Grassmann manifold $\mathrm{Gr}(n, p)$.

To ensure that step (17) for computing η_k in Algorithm 1 is well defined, we need Assumption 1. This assumption is based the following important concepts. In differential geometry, if $\exp_x$ is a diffeomorphism of a neighborhood $\mathcal{V}$ of the origin in $T_x\mathcal{M}$, then $\mathcal{U} = \exp_x(\mathcal{V})$ is called a normal neighborhood of x (see, e.g., Section 3.3 in [13]). Furthermore, it is called a normal ball if $\mathcal{V}$ is an open ball of the origin in $T_x\mathcal{M}$.

Assumption 1 *The sequences $\{Y_k\}_{k\geq 1}$ and $\{Z_k\}_{k\geq 1}$ generated by Algorithm 1 satisfy that Y_k is in some normal ball of Z_k in $O(n)$.*

Owing to the geodesic formulas (10) and (11), the normal neighborhood (ball) in our case can be identified with the injectivity neighborhood (ball) of the matrix exponential expm : $\mathfrak{so}(n) \to O(n)$. By the injectivity neighborhood (ball) of expm we mean the following concept.

Definition 3 *An injectivity neighborhood (ball) of the matrix exponential* $\mathrm{expm} : \mathfrak{so}(n) \to O(n)$ *is* $\mathrm{expm}(\mathcal{V})$ *such that* expm *is a bijective of a neighborhood (an open ball)* $\mathcal{V}$ *of the origin in* $\mathfrak{so}(n)$ *onto its image.*

By Gantmacher's theorem (see, e.g., Theorem 1.27 in [20]), $\mathrm{expm}(X) = A$ for a nonsingular matrix A has a unique solution in the 2-norm ball $\{X \mid ||X||_2 < \pi\}$. So, $\mathcal{U}_0 := \{\exp_Q(\eta_Q) \mid \eta_Q \in T_Q O(n), ||\eta_Q||_2 < \pi\}$ is a normal neighborhood for all $Q \in O(n)$. But note that normal neighborhoods (much) larger than $\mathcal{U}_0$ may exist.

Besides Assumption 1, we need the following two assumptions.

Assumption 2 *f is differentiable and ∇f is L-Lipschitz continuous in the following sense:*

$$\left\|P_\gamma^{z\leftarrow x}\nabla f(x) - \nabla f(z)\right\|_{\mathrm{c}} \leq L\mathrm{dist}(x, z), \tag{26}$$

where dist *denotes the Riemannian distance.*

According to [42], (26) implies that f is also geodesically L-smooth, i.e.,

$$f(x) \leq f(z) + \left\langle \nabla f(z), \exp_z^{-1}(x)\right\rangle + \frac{L}{2}\mathrm{dist}(x, z)^2. \tag{27}$$

The above assumption is mild for the Grassmann manifold because of its compactness.

Assumption 3 *$Y_k(t)$ is in some normal ball of $Z_k(t)$ in $O(n)$ for all $t \in [0, \beta_k]$, where*

$$Y_k(t) := \exp_{X_k}(-t\nabla f(X_k)), \quad Z_k(t) := \exp_{Z_{k-1}}(-tP_\gamma^{Z_{k-1}\leftarrow X_k}\nabla f(X_k)). \tag{28}$$

The above assumption means that $Y_k(t)$ is not very far from $Z_k(t)$ so that $Y_k(t)$ is within domain of the inverse of the exponential map at $Z_k(t)$. Such kind of assumptions are for technical use and often occur in convergence analysis in Riemannian optimization methods.

The following lemma which shows that the distance between two parallel geodesics on the Grassmann manifold is non-increasing plays a crucial role in the main convergence theorem for Algorithm 1.

Lemma 2 *Suppose that Assumptions 1 and 3 hold. Let $Y_k(t)$ and $Z_k(t)$ be defined in* (28)*. Then*

$$\mathrm{dist}(Y_k(t), Z_k(t)) \leq \mathrm{dist}(X_k, Z_{k-1}), \quad \forall\ t \in [0, \beta_k].$$

Proof. According to the orthogonal group representation in the remark at the beginning of Section 4, we can denote $Z_{k-1} = Q_{Z_{k-1}}$, $\eta_k = Q_{Z_{k-1}}\mathfrak{A}$, and $P_\gamma^{Z_{k-1}\leftarrow X_k}\nabla f(X_k) = Q_{Z_{k-1}}\mathfrak{B}$, where $\mathfrak{A}$ and $\mathfrak{B}$ are of form (8). Then we have

$$X_k \xlongequal{(18)} \exp_{Z_{k-1}}(\eta_k) \xlongequal{(11)} Q_{Z_{k-1}}\mathrm{expm}(\mathfrak{A}) := Q_{X_k},$$

$$\nabla f(X_k) = P_\gamma^{X_k\leftarrow Z_{k-1}}P_\gamma^{Z_{k-1}\leftarrow X_k}\nabla f(X_k) \xlongequal{(12)} Q_{Z_{k-1}}\mathrm{expm}(\mathfrak{A})\mathfrak{B} = Q_{X_k}\mathfrak{B},$$

$$Z_k(t) \xlongequal{(28)} \exp_{Z_{k-1}}(-tP_\gamma^{Z_{k-1}\leftarrow X_k}\nabla f(X_k)) \xlongequal{(11)} Q_{Z_{k-1}}\mathrm{expm}(-t\mathfrak{B}) := Q_{Z_k(t)}, \tag{29}$$

and

$$Y_k(t) \xlongequal{(28)} \exp_{X_k}(-t\nabla f(X_k)) \xlongequal{(11)} Q_{X_k}\mathrm{expm}(-t\mathfrak{B}) = Q_{Z_{k-1}}\mathrm{expm}(\mathfrak{A})\mathrm{expm}(-t\mathfrak{B}). \tag{30}$$

By Assumption 3, there is a unique $\Omega(t) \in \mathfrak{so}(n)$ such that

$$Y_k(t) = \overline{\exp}_{Q_{Z_k(t)}}(Q_{Z_k(t)}\Omega(t)) \xlongequal{(10)} Q_{Z_k(t)}\mathrm{expm}(\Omega(t)), \tag{31}$$

where $\overline{\exp}$ is the exponential map on $O(n)$. So $\frac{1}{\sqrt{2}}||\Omega(t)||_{\mathrm{F}} = \overline{\mathrm{dist}}(Y_k(t), Z_k(t))$, where $\overline{\mathrm{dist}}$ is the distance on $O(n)$. Combining (29)–(31), we have

$$\mathrm{expm}(\Omega(t)) = \mathrm{expm}(t\mathfrak{B}) \cdot \mathrm{expm}(\mathfrak{A}) \cdot \mathrm{expm}(-t\mathfrak{B}) = \mathrm{expm}(\mathfrak{C}(t)), \tag{32}$$

where

$$\mathfrak{C}(t) := \mathrm{expm}(t\mathfrak{B}) \cdot \mathfrak{A} \cdot \mathrm{expm}(-t\mathfrak{B}).$$

Since

$$\begin{aligned}\frac{1}{\sqrt{2}}||\mathfrak{C}(t)||_{\mathrm{F}} \equiv \frac{1}{\sqrt{2}}||\mathfrak{A}||_{\mathrm{F}} = ||\eta_k||_{\mathrm{c}} &\overset{(17)}{=\!=\!=} (1-\lambda_k)||\exp^{-1}_{Z_{k-1}}(Y_{k-1})||_{\mathrm{c}} \\ &\leq ||\exp^{-1}_{Z_{k-1}}(Y_{k-1})||_{\mathrm{c}} \leq ||\overline{\exp}^{-1}_{Z_{k-1}}(Y_{k-1})||_{\mathrm{c}},\end{aligned}$$

where the last inequality follows from the property that a Riemannian submersion shortens distances (e.g., Proposition 2.109 in [15]), we have from Assumption 1 that $\overline{\exp}_{Q_{Z_{k-1}}}(Q_{Z_{k-1}}\mathfrak{C}(t)) = Q_{Z_{k-1}}\mathrm{expm}(\mathfrak{C}(t))$ is in some normal ball of Z_{k-1} in $O(n)$; therefore $\mathrm{expm}(\mathfrak{C}(t))$ is in some injectivity ball of expm. By Assumption 3, (29) and (31), we know that $\mathrm{expm}(\Omega(t))$ is also in some injectivity ball of expm. Thus, it follows from (32) that

$$\Omega(t) = \mathfrak{C}(t) = \mathrm{expm}(t\mathfrak{B}) \cdot \mathfrak{A} \cdot \mathrm{expm}(-t\mathfrak{B}).$$

Therefore $||\Omega(t)||_{\mathrm{F}} \equiv ||\mathfrak{A}||_{\mathrm{F}}$. Again, since a Riemannian submersion shortens distances, we obtain

$$\begin{aligned}\mathrm{dist}(Y_k(t), Z_k(t)) \leq \overline{\mathrm{dist}}(Y_k(t), Z_k(t)) &= \frac{1}{\sqrt{2}}||\Omega(t)||_{\mathrm{F}} \equiv \frac{1}{\sqrt{2}}||\mathfrak{A}||_{\mathrm{F}} \\ &= ||\eta_k||_{\mathrm{c}} \overset{(18)}{=\!=\!=} ||\exp^{-1}_{Z_{k-1}}(X_k)||_{\mathrm{c}} = \mathrm{dist}(X_k, Z_{k-1}).\end{aligned}$$

This completes the proof. □

Now we can give the main convergence result of Algorithm 1 as follows.

Theorem 1 *Suppose that Assumptions 1–3 hold. Let $\{\tau_k\}$ be the sequence of numbers defined by* (25)*. If $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\lambda_k\}$ are chosen such that*

$$c_k := 1 - L\beta_k - \frac{L(\beta_k-\alpha_k)^2}{2\beta_k\lambda_k\tau_k}\left(\sum_{i=k}^{N}\tau_i\right) > 0, \quad 1 \leq k \leq N, \tag{33}$$

then Algorithm 1 satisfies for all $N \geq 1$ that

$$\min_{k=1,\dots,N} ||\nabla f(X_k)||^2_{\mathrm{c}} \leq \frac{f(Z_0)-f^*}{\sum_{k=1}^{N}\beta_k c_k}.$$

Proof. Denote $\Delta_k := \nabla f(Z_{k-1}) - P_\gamma^{Z_{k-1}\leftarrow X_k}\nabla f(X_k)$. Using inequality (26), we have

$$\begin{aligned}||\Delta_k||_{\mathrm{c}} = \left\|\nabla f(Z_{k-1}) - P_\gamma^{Z_{k-1}\leftarrow X_k}\nabla f(X_k)\right\|_{\mathrm{c}} &\leq L\left\|\exp^{-1}_{Z_{k-1}}(X_k)\right\|_{\mathrm{c}} \\ \overset{(18)}{=\!=\!=} L||\eta_k||_{\mathrm{c}} &\overset{(17)}{=\!=\!=} L(1-\lambda_k)\mathrm{dist}(Y_{k-1}, Z_{k-1}).\end{aligned}$$

Using inequality (27), we also have

$$\begin{aligned}f(Z_k) &\leq f(Z_{k-1}) + \left\langle \nabla f(Z_{k-1}), \exp^{-1}_{Z_{k-1}}(Z_k)\right\rangle + \frac{L}{2}\left\|\exp^{-1}_{Z_{k-1}}(Z_k)\right\|^2_{\mathrm{c}} \\ &\overset{(20)}{=\!=\!=} f(Z_{k-1}) + \left\langle \nabla f(Z_{k-1}), -\beta_k P_\gamma^{Z_{k-1}\leftarrow X_k}\nabla f(X_k)\right\rangle + \frac{L}{2}\left\|\beta_k P_\gamma^{Z_{k-1}\leftarrow X_k}\nabla f(X_k)\right\|^2_{\mathrm{c}} \\ &= f(Z_{k-1}) + \left\langle \Delta_k + P_\gamma^{Z_{k-1}\leftarrow X_k}\nabla f(X_k), -\beta_k P_\gamma^{Z_{k-1}\leftarrow X_k}\nabla f(X_k)\right\rangle \\ &\quad + \frac{L}{2}\left\|\beta_k P_\gamma^{Z_{k-1}\leftarrow X_k}\nabla f(X_k)\right\|^2_{\mathrm{c}} \\ &= f(Z_{k-1}) - \beta_k\left(1-\frac{L\beta_k}{2}\right)\left\|P_\gamma^{Z_{k-1}\leftarrow X_k}\nabla f(X_k)\right\|^2_{\mathrm{c}} - \beta_k\left\langle\Delta_k, P_\gamma^{Z_{k-1}\leftarrow X_k}\nabla f(X_k)\right\rangle \\ &\leq f(Z_{k-1}) - \beta_k\left(1-\frac{L\beta_k}{2}\right)||\nabla f(X_k)||^2_{\mathrm{c}} + \beta_k||\Delta_k||_{\mathrm{c}}||\nabla f(X_k)||_{\mathrm{c}}.\end{aligned}$$

Combining the previous two inequalities, we obtain

$$
\begin{aligned}
f(Z_k) &\leq f(Z_{k-1}) - \beta_k \left(1 - \frac{L\beta_k}{2}\right) ||\nabla f(X_k)||_{\mathrm{c}}^2 + L(1-\lambda_k)\beta_k ||\nabla f(X_k)||_{\mathrm{c}} \mathrm{dist}(Y_{k-1}, Z_{k-1}) \\
&\leq f(Z_{k-1}) - \beta_k \left(1 - \frac{L\beta_k}{2}\right) ||\nabla f(X_k)||_{\mathrm{c}}^2 + \frac{L\beta_k^2}{2} ||\nabla f(X_k)||_{\mathrm{c}}^2 \\
&\quad + \frac{L(1-\lambda_k)^2}{2} \mathrm{dist}(Y_{k-1}, Z_{k-1})^2 \\
&= f(Z_{k-1}) - \beta_k (1 - L\beta_k) ||\nabla f(X_k)||_{\mathrm{c}}^2 + \frac{L(1-\lambda_k)^2}{2} \mathrm{dist}(Y_{k-1}, Z_{k-1})^2. \qquad (34)
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\mathrm{dist}(Y_k, Z_k) &\xlongequal{(19)(20)} \mathrm{dist}\Big(\exp_{X_k}(-\alpha_k \nabla f(X_k)),\ \exp_{Z_{k-1}}(-\beta_k P_\gamma^{Z_{k-1} \leftarrow X_k} \nabla f(X_k)) \Big) \\
&\leq \mathrm{dist}\Big(\exp_{X_k}(-\beta_k \nabla f(X_k)),\ \exp_{Z_{k-1}}(-\beta_k P_\gamma^{Z_{k-1} \leftarrow X_k} \nabla f(X_k)) \Big) \\
&\quad + \mathrm{dist}\Big(\exp_{X_k}(-\beta_k \nabla f(X_k)),\ \exp_{X_k}(-\alpha_k \nabla f(X_k)) \Big) \\
&\leq \mathrm{dist}(X_k, Z_{k-1}) + (\beta_k - \alpha_k) ||\nabla f(X_k)||_{\mathrm{c}} \\
&\xlongequal{(17)(18)} (1-\lambda_k) \mathrm{dist}(Y_{k-1}, Z_{k-1}) + (\beta_k - \alpha_k) ||\nabla f(X_k)||_{\mathrm{c}}, \qquad (35)
\end{aligned}
$$

where the first inequality follows from triangular inequality and the second inequality follows from (28) and Lemma 2. Dividing both sides of (35) inequality by τ_k and noting (25), we have

$$
\frac{\mathrm{dist}(Y_k, Z_k)}{\tau_k} \leq \frac{\mathrm{dist}(Y_{k-1}, Z_{k-1})}{\tau_{k-1}} + \frac{(\beta_k - \alpha_k) ||\nabla f(X_k)||_{\mathrm{c}}}{\tau_k}.
$$

Summing them up and noting $Y_0 = Z_0$, we obtain

$$
\mathrm{dist}(Y_k, Z_k) \leq \tau_k \sum_{i=1}^k \frac{\beta_i - \alpha_i}{\tau_i} ||\nabla f(X_i)||_{\mathrm{c}} = \sum_{i=1}^k \tau_k \frac{\lambda_i}{\tau_i} \cdot \frac{\beta_i - \alpha_i}{\lambda_i} ||\nabla f(X_i)||_{\mathrm{c}}.
$$

Using the above inequality, Lemma 1, and Jensen's inequality, we have

$$
\begin{aligned}
\mathrm{dist}(Y_k, Z_k)^2 &\leq \left(\sum_{i=1}^k \tau_k \frac{\lambda_i}{\tau_i} \cdot \frac{\beta_i - \alpha_i}{\lambda_i} ||\nabla f(X_i)||_{\mathrm{c}} \right)^2 \leq \sum_{i=1}^k \tau_k \frac{\lambda_i}{\tau_i} \cdot \frac{(\beta_i - \alpha_i)^2}{\lambda_i^2} ||\nabla f(X_i)||_{\mathrm{c}}^2 \\
&= \tau_k \sum_{i=1}^k \frac{(\beta_i - \alpha_i)^2}{\lambda_i \tau_i} ||\nabla f(X_i)||_{\mathrm{c}}^2.
\end{aligned}
$$

Replacing the above bound in (34) and using (25), we obtain

$$
\begin{aligned}
f(Z_k) &\leq f(Z_{k-1}) - \beta_k (1 - L\beta_k) ||\nabla f(X_k)||_{\mathrm{c}}^2 + \frac{L(1-\lambda_k)^2 \tau_{k-1}}{2} \sum_{i=1}^{k-1} \frac{(\beta_i - \alpha_i)^2}{\lambda_i \tau_i} ||\nabla f(X_i)||_{\mathrm{c}}^2 \\
&\leq f(Z_{k-1}) - \beta_k (1 - L\beta_k) ||\nabla f(X_k)||_{\mathrm{c}}^2 + \frac{L\tau_k}{2} \sum_{i=1}^{k-1} \frac{(\beta_i - \alpha_i)^2}{\lambda_i \tau_i} ||\nabla f(X_i)||_{\mathrm{c}}^2.
\end{aligned}
$$

Summing up the above inequalities and using the definition of c_k in (33), we have

$$
\begin{aligned}
f(Z_N) &\leq f(Z_0) - \sum_{k=1}^N \beta_k (1 - L\beta_k) ||\nabla f(X_k)||_{\mathrm{c}}^2 + \frac{L}{2} \sum_{k=1}^N \tau_k \sum_{i=1}^k \frac{(\beta_i - \alpha_i)^2}{\lambda_i \tau_i} ||\nabla f(X_i)||_{\mathrm{c}}^2 \\
&= f(Z_0) - \sum_{k=1}^N \beta_k (1 - L\beta_k) ||\nabla f(X_k)||_{\mathrm{c}}^2 + \frac{L}{2} \sum_{k=1}^N \frac{(\beta_k - \alpha_k)^2}{\lambda_k \tau_k} \left(\sum_{i=k}^N \tau_i \right) ||\nabla f(X_k)||_{\mathrm{c}}^2 \\
&= f(Z_0) - \sum_{k=1}^N \beta_k c_k ||\nabla f(X_k)||_{\mathrm{c}}^2.
\end{aligned}
$$

Re-arranging the terms in the above inequality and noting that $f(Z_N) \geq f^*$ we obtain

$$\min_{k=1,\ldots,N} ||\nabla f(X_k)||_{\mathrm{c}}^2 \left(\sum_{k=1}^{N} \beta_k c_k \right) \leq \sum_{k=1}^{N} \beta_k c_k ||\nabla f(X_k)||_{\mathrm{c}}^2 \leq f(Z_0) - f^*.$$

This completes the proof. □

Corollary 1 *Suppose that Assumptions 1–3 hold and that* $\alpha_k \equiv \frac{1}{2L}$ *and* $\lambda_k = \frac{2}{k+1}$*. If*

$$\beta_k \in \left[\alpha_k, \left(1 + \frac{\lambda_k}{4}\right) \alpha_k \right],$$

then Algorithm 1 satisfies for all $N \geq 1$ *that*

$$\min_{k=1,\ldots,N} ||\nabla f(X_k)||_{\mathrm{c}}^2 \leq \frac{6L(f(Z_0) - f^*)}{N}.$$

Proof. This is just a copy of Corollary 1 in [19]. □

4.2 Convergence of Algorithm 2

In this subsection, we focus on Algorithm 2. Now the manifold $\mathcal{M}$ in question is either the Grassmann manifold $\mathrm{Gr}(n, p)$ or the Stiefel manifold $\mathrm{St}(n, p)$.

For Algorithm 2, we do not need to introduce the concept of retractive neighborhood (ball) [25] because of the injectivity of the Cayley transform (15).

Lemma 3 *The Cayley transform*

$$\phi_{\mathrm{ct}}(\Omega) : \mathfrak{so}(n) \to O(n) : \Omega \mapsto \left(I_n - \frac{1}{2}\Omega\right)^{-1} \left(I_n + \frac{1}{2}\Omega\right)$$

is injective, and

$$\phi_{\mathrm{ct}}^{-1}(Q) = 2(Q - I_n)(Q + I_n)^{-1}.$$

Proof. Let $Q = \phi_{\mathrm{ct}}(\Omega) = \left(I_n - \frac{1}{2}\Omega\right)^{-1}\left(I_n + \frac{1}{2}\Omega\right) \in O(n)$. Then we have $\Omega(Q + I_n) = 2(Q - I_n)$. Since $Q + I_n = 2\left(I_n - \frac{1}{2}\Omega\right)^{-1}$ is invertible, $\Omega = 2(Q - I_n)(Q + I_n)^{-1}$ is uniquely determined. □

Now we present an assumption to ensure that step (21) for computing η_k in Algorithm 2 is well defined and that $\{||\eta_k||_{\mathrm{c}}\}_{k \geq 1}$ is bounded.

Assumption 4 *The sequences* $\{Y_k\}_{k \geq 1}$ *and* $\{Z_k\}_{k \geq 1}$ *generated by Algorithm 2 satisfy that* Y_k *is in the image of* $R_{Z_k}(\cdot)$ *in* $\mathcal{M}$ *and that* $\left\{||R_{Z_k}^{-1}(Y_k)||_{\mathrm{c}}\right\}_{k \geq 1}$ *is bounded.*

The following assumption is a modification of Assumption 2, which is also mild because the Grassmann and Stiefel manifolds are both compact.

Assumption 5 f *is differentiable and* ∇f *is* L*-Lipschitz continuous in the following sense:*

$$\left\| \mathcal{T}_{\eta_z}^{-1} \nabla f(x) - \nabla f(z) \right\|_{\mathrm{c}} \leq L \mathrm{dist}(x, z), \tag{36}$$

where $x = R_z(\eta_z)$*. Moreover,* f *is* L*-retraction-smooth, i.e.,*

$$f(R_z(\eta_z)) \leq f(z) + \langle \nabla f(z), \eta_z \rangle + \frac{L}{2} ||\eta_z||_{\mathrm{c}}^2. \tag{37}$$

Note that in the above assumption (37) can not be implied by (36) for a general retraction R other than the exponential map exp.

The following assumption is a weak analog to Assumption 3.

Assumption 6 *$Y_k(t)$ is in the image of $\overline{R}_{Z_k(t)}(\cdot)$ in $O(n)$ for all $t \in [0, \beta_k]$, where*

$$Y_k(t) := R_{X_k}(-t\nabla f(X_k)), \quad Z_k(t) := R_{Z_{k-1}}(-t\mathcal{T}_{\eta_k}^{-1}\nabla f(X_k)). \tag{38}$$

To obtain results similar to Lemma 2, we need an additional assumption.

Assumption 7 *The following two inequalities hold:*

$$\left\|R_{Z_k(t)}^{-1}(Y_k(t))\right\|_{\mathrm{c}} \leq \left\|\overline{R}_{Z_k(t)}^{-1}(Y_k(t))\right\|_{\mathrm{c}} \tag{39}$$

and

$$\left\|R_{Z_k(t)}^{-1}(Y_k(t+\Delta t)) - R_{Z_k(t)}^{-1}(Y_k(t))\right\|_{\mathrm{c}} \leq \Upsilon\left\|\overline{R}_{Z_k(t)}^{-1}(Y_k(t+\Delta t)) - \overline{R}_{Z_k(t)}^{-1}(Y_k(t))\right\|_{\mathrm{c}} \tag{40}$$

for some constant $\Upsilon > 0$.

Assumption (39) means the inverse retraction in the quotient manifold has no larger magnitude than the inverse retraction in the total manifold. It holds naturally for the exponential map because distance in the quotient manifold is no longer than distance in the total manifold; see the end of the proof of Lemma 2. Assumption (40) is reasonable if Υ is a sufficiently large constant.

Lemma 4 *Let $z(t) = R_z(t\eta_z)$. Then*

$$\mathrm{dist}(z(t_0), z(t_1)) \leq (t_1 - t_0)||\eta_z||_{\mathrm{c}}$$

for all $t_1 > t_0$. In particular,

$$\mathrm{dist}(z, R_z(\eta_z)) = \mathrm{dist}(z, z(1)) \leq ||\eta_z||_{\mathrm{c}}.$$

Proof. According to the remark at the beginning of Section 4, we can denote $z = Q$ and $\eta_z = Q\mathfrak{A}$, where $\mathfrak{A}$ is of form (8) if $\mathcal{M} = \mathrm{Gr}(n,p)$ and of form (9) if $\mathcal{M} = \mathrm{St}(n,p)$. Using (14)–(15) and differentiating $R_z(t\eta_z)$ with respect to t gives

$$\frac{d}{dt}R_z(t\eta_z) = Q\Big(I_n - \frac{t}{2}\mathfrak{A}\Big)^{-2}\mathfrak{A}.$$

Then

$$\Big\|\frac{d}{dt}R_z(t\eta_z)\Big\|_{\mathrm{c}}^2 = \frac{1}{2}\mathrm{Tr}\Big(\mathfrak{A}^\top\Big(I_n - \frac{t^2}{4}\mathfrak{A}^2\Big)^{-2}\mathfrak{A}\Big) \leq \frac{1}{2}\mathrm{Tr}(\mathfrak{A}^\top\mathfrak{A}) = ||\eta_z||_{\mathrm{c}}^2,$$

where the inequality follows from the skew-symmetry of $\mathfrak{A}$. Thus we obtain

$$\mathrm{dist}(z(t_0), z(t_1)) \leq \int_{t_0}^{t_1}\Big\|\frac{d}{dt}R_z(t\eta_z)\Big\|_{\mathrm{c}}dt \leq \int_{t_0}^{t_1}||\eta_z||_{\mathrm{c}}dt = (t_1 - t_0)||\eta_z||_{\mathrm{c}}.$$

This completes the proof. □

The following lemma is a retraction version of Lemma 2. Its proof follows from the same idea of the proof of Lemma 2, but additionally utilizes Lemma 4 and Assumption 7.

Lemma 5 *Suppose that Assumptions 4, 6 and 7 hold. Let $Y_k(t)$ and $Z_k(t)$ be defined in (38). Then*

$$\left\|R_{Z_k(t)}^{-1}(Y_k(t))\right\|_{\mathrm{c}} \leq \left\|R_{Z_{k-1}}^{-1}(X_k)\right\|_{\mathrm{c}} = ||\eta_k||_{\mathrm{c}}, \quad \forall\ t \in [0, \beta_k].$$

Proof. According to the remark at the beginning of Section 4 again, we can denote $Z_{k-1} = Q_{Z_{k-1}}$, $\eta_k = Q_{Z_{k-1}}\mathfrak{A}$, and $\mathcal{T}_{\eta_k}^{-1}\nabla f(X_k) = Q_{Z_{k-1}}\mathfrak{B}$, where $\mathfrak{A}$ and $\mathfrak{B}$ are of form (8) if $\mathcal{M} = \mathrm{Gr}(n,p)$ and of form (9) if $\mathcal{M} = \mathrm{St}(n,p)$. Then we have

$$X_k \overset{(22)}{=\!=\!=} R_{Z_{k-1}}(\eta_k) \overset{(14)}{=\!=\!=} Q_{Z_{k-1}}\phi_{\mathrm{ct}}(\mathfrak{A}) := Q_{X_k},$$

$$\nabla f(X_k) = \mathcal{T}_{\eta_k}\mathcal{T}_{\eta_k}^{-1}\nabla f(X_k) \overset{(16)}{=\!=\!=} Q_{Z_{k-1}}\phi_{\mathrm{ct}}(\mathfrak{A})\mathfrak{B} = Q_{X_k}\mathfrak{B},$$

$$Z_k(t) \overset{(38)}{=\!=\!=} R_{Z_{k-1}}(-t\mathcal{T}_{\eta_k}^{-1}\nabla f(X_k)) \overset{(14)}{=\!=\!=} Q_{Z_{k-1}}\phi_{\mathrm{ct}}(-t\mathfrak{B}) := Q_{Z_k(t)}, \tag{41}$$

and

$$Y_k(t) \overset{(38)}{=\!=\!=} R_{X_k}(-t\nabla f(X_k)) \overset{(14)}{=\!=\!=} Q_{X_k}\phi_{\mathrm{ct}}(-t\mathfrak{B}) = Q_{Z_{k-1}}\phi_{\mathrm{ct}}(\mathfrak{A})\phi_{\mathrm{ct}}(-t\mathfrak{B}). \tag{42}$$

By Assumption 6 and Lemma 3, there is a unique $\Omega(t) \in \mathfrak{so}(n)$ such that

$$Y_k(t) = \overline{R}_{Q_{Z_k(t)}}(Q_{Z_k(t)}\Omega(t)) \overset{(13)}{=\!=\!=} Q_{Z_k(t)}\phi_{\mathrm{ct}}(\Omega(t)), \tag{43}$$

where $\overline{R}$ is the Cayley transform retraction on $O(n)$. So $\frac{1}{\sqrt{2}}||\Omega(t)||_{\mathrm{F}} = \left\|\overline{R}_{Z_k(t)}^{-1}(Y_k(t))\right\|_{\mathrm{c}}$ is the length of the inverse retraction from $Z_k(t)$ to $Y_k(t)$ on $O(n)$. Combining (41)–(43), we have

$$\phi_{\mathrm{ct}}(\Omega(t)) = \phi_{\mathrm{ct}}(t\mathfrak{B}) \cdot \phi_{\mathrm{ct}}(\mathfrak{A}) \cdot \phi_{\mathrm{ct}}(-t\mathfrak{B}) = \phi_{\mathrm{ct}}(\phi_{\mathrm{ct}}(t\mathfrak{B}) \cdot \mathfrak{A} \cdot \phi_{\mathrm{ct}}(-t\mathfrak{B})). \tag{44}$$

This together with Lemma 3 implies

$$\Omega(t) = \phi_{\mathrm{ct}}(t\mathfrak{B}) \cdot \mathfrak{A} \cdot \phi_{\mathrm{ct}}(-t\mathfrak{B}).$$

Therefore $||\Omega(t)||_{\mathrm{F}} \equiv ||\mathfrak{A}||_{\mathrm{F}}$. By Lemma 4 and (39) in Assumption 7, we obtain

$$\begin{aligned}\mathrm{dist}(Y_k(t), Z_k(t)) &\le \left\|R_{Z_k(t)}^{-1}(Y_k(t))\right\|_{\mathrm{c}} \le \left\|\overline{R}_{Z_k(t)}^{-1}(Y_k(t))\right\|_{\mathrm{c}} \\ &= \frac{1}{\sqrt{2}}||\Omega(t)||_{\mathrm{F}} \equiv \frac{1}{\sqrt{2}}||\mathfrak{A}||_{\mathrm{F}} = ||\eta_k||_{\mathrm{c}} \overset{(22)}{=\!=\!=} \left\|R_{Z_{k-1}}^{-1}(X_k)\right\|_{\mathrm{c}}.\end{aligned}$$

This completes the proof. □

Lemma 5 alone is not enough for the convergence of Algorithm 2 because we can not completely copy the key inequality (35) in the proof of Theorem 1. So, we need the following technical result.

Lemma 6 *Suppose that Assumptions 4, 6 and 7 hold. Let $Y_k(t)$ and $Z_k(t)$ be defined in* (38)*. If $\{t, t+\Delta t\} \subset [0, \beta_k]$ and*

$$|\Delta t| \le \frac{1}{3\sqrt{1+\frac{1}{4}||\eta_k||_2^2} \cdot ||\nabla f(X_k)||_2},$$

then

$$\left\|R_{Z_k(t)}^{-1}(Y_k(t+\Delta t))\right\|_{\mathrm{c}} \le \left\|R_{Z_k(t)}^{-1}(Y_k(t))\right\|_{\mathrm{c}} + \frac{\sqrt{2n}\Upsilon}{2}(4+||\eta_k||_2^2)||\nabla f(X_k)||_2|\Delta t|,$$

where $||\cdot||_2$ is in the sense of viewing a tangent vector as its horizontal lift to $T_QO(n)$.

Proof. We follow the notations in Lemma 5. By Assumption 6 and Lemma 3, there exists $\Xi(t, \Delta t) \in \mathfrak{so}(n)$ such that

$$\overline{R}_{Q_{Z_k(t)}}(Q_{Z_k(t)}\Xi(t, \Delta t)) \overset{(13)}{=\!=\!=} Q_{Z_k(t)}\phi_{\mathrm{ct}}(\Xi(t, \Delta t)) = Y_k(t+\Delta t).$$

Using $Q_{X_k} = Q_{Z_{k-1}}\phi_{\mathrm{ct}}(\mathfrak{A})$ in the proof of Lemma 5 and the above equation, we have

$$\begin{aligned}Q_{Z_{k-1}}\phi_{\mathrm{ct}}(-t\mathfrak{B})\phi_{\mathrm{ct}}(\Xi(t, \Delta t)) &\overset{(41)}{=\!=\!=} Q_{Z_k(t)}\phi_{\mathrm{ct}}(\Xi(t, \Delta t)) = Y_k(t+\Delta t) \\ &\overset{(38)(14)}{=\!=\!=\!=\!=} Q_{X_k}\phi_{\mathrm{ct}}(-(t+\Delta t)\mathfrak{B}) \\ &= Q_{Z_{k-1}}\phi_{\mathrm{ct}}(\mathfrak{A})\phi_{\mathrm{ct}}(-(t+\Delta t)\mathfrak{B}).\end{aligned}$$

This implies

$$\phi_{\mathrm{ct}}(\Xi(t, \Delta t)) = \phi_{\mathrm{ct}}(t\mathfrak{B}) \cdot \phi_{\mathrm{ct}}(\mathfrak{A}) \cdot \phi_{\mathrm{ct}}(-(t+\Delta t)\mathfrak{B}). \tag{45}$$

To obtain the result, we will give a perturbation analysis for $\Xi(t, \Delta t)$. Denote $F_{\Delta t} := \phi_{\mathrm{ct}}(\Xi(t, \Delta t))$ and $\Delta F := F_{\Delta t} - F_0$. By (44) we have

$$F_0 = \phi_{\mathrm{ct}}(\Xi(t, 0)) = \phi_{\mathrm{ct}}(\Omega(t)) = \phi_{\mathrm{ct}}(t\mathfrak{B}) \cdot \phi_{\mathrm{ct}}(\mathfrak{A}) \cdot \phi_{\mathrm{ct}}(-t\mathfrak{B}). \tag{46}$$

This implies that F_0 and $\phi_{\mathrm{ct}}(\mathfrak{A})$ have the same eigenvalues. Then a simple spectral calculation with noticing (15) and the skew-symmetry of $\mathfrak{A}$ reveals that $F_0 + I_n$ is invertible and

$$||(F_0 + I_n)^{-1}||_2 \leq \frac{1}{2}\sqrt{1 + \frac{1}{4}||\mathfrak{A}||_2^2} = \frac{1}{2}\sqrt{1 + \frac{1}{4}||\eta_k||_2^2}. \tag{47}$$

Combining (45) and (46) yields

$$\Delta F = \phi_{\mathrm{ct}}(t\mathfrak{B}) \cdot \phi_{\mathrm{ct}}(\mathfrak{A}) \cdot (\phi_{\mathrm{ct}}(-(t + \Delta t)\mathfrak{B}) - \phi_{\mathrm{ct}}(-t\mathfrak{B})). \tag{48}$$

By Lemma 3 we have

$$\begin{aligned} \Xi(t, \Delta t) &= \phi_{\mathrm{ct}}^{-1}(F_{\Delta t}) = 2(F_{\Delta t} - I_n)(F_{\Delta t} + I_n)^{-1} \\ &= 2I_n - 4(F_{\Delta t} + I_n)^{-1} = 2I_n - 4(\Delta F + F_0 + I_n)^{-1} \\ &= 2I_n - 4(F_0 + I_n)^{-1}(\Delta F(F_0 + I_n)^{-1} + I_n)^{-1}. \end{aligned}$$

Using Taylor's theorem of matrix functions (e.g., Theorem 4.8 in [20]), we obtain

$$||\Xi(t, \Delta t) - \Xi(t, 0)|| \leq 4||\Delta F|| \cdot ||(F_0 + I_n)^{-1}||^2 \max_{0 \leq \theta \leq 1} \left\|(\theta \Delta F(F_0 + I_n)^{-1} + I_n)^{-2}\right\|, \tag{49}$$

where the norm $||\cdot||$ is arbitrary. Combining (47) and (49) yields

$$||\Xi(t, \Delta t) - \Xi(t, 0)||_2 \leq \frac{(1 + \frac{1}{4}||\eta_k||_2^2)||\Delta F||_2}{\left(1 - \frac{1}{2}\sqrt{1 + \frac{1}{4}||\eta_k||_2^2} \cdot ||\Delta F||_2\right)^2}. \tag{50}$$

It follows from (15) that

$$\begin{aligned} \phi_{\mathrm{ct}}(-(t + \Delta t)\mathfrak{B}) &= 2\Big(I_n + \frac{t}{2}\mathfrak{B} + \frac{\Delta t}{2}\mathfrak{B}\Big)^{-1} - I_n \\ &= 2\Big(I_n + \Big(I + \frac{t}{2}\mathfrak{B}\Big)^{-1}\frac{\Delta t}{2}\mathfrak{B}\Big)^{-1}\Big(I_n + \frac{t}{2}\mathfrak{B}\Big)^{-1} - I_n. \end{aligned}$$

Then using Taylor's theorem again, we obtain

$$\begin{aligned} &||\phi_{\mathrm{ct}}(-(t + \Delta t)\mathfrak{B}) - \phi_{\mathrm{ct}}(-t\mathfrak{B})|| \\ \leq\; &|\Delta t| \cdot ||\mathfrak{B}|| \cdot \Big\|\Big(I_n + \frac{t}{2}\mathfrak{B}\Big)^{-1}\Big\|^2 \max_{0 \leq \theta \leq 1}\Big\|\Big(I_n + \theta\Big(I + \frac{t}{2}\mathfrak{B}\Big)^{-1}\frac{\Delta t}{2}\mathfrak{B}\Big)^{-2}\Big\|, \end{aligned} \tag{51}$$

where the norm $||\cdot||$ is arbitrary. Since $\mathfrak{B}$ is skew-symmetric, $\left\|\left(I_n + \frac{t}{2}\mathfrak{B}\right)^{-1}\right\|_2 \leq 1$. Then (51) implies

$$||\phi_{\mathrm{ct}}(-(t + \Delta t)\mathfrak{B}) - \phi_{\mathrm{ct}}(-t\mathfrak{B})||_2 \leq \frac{||\mathfrak{B}||_2|\Delta t|}{\left(1 - \frac{1}{2}||\mathfrak{B}||_2|\Delta t|\right)^2}. \tag{52}$$

Since $\phi_{\mathrm{ct}}(\mathfrak{A})$ and $\phi_{\mathrm{ct}}(t\mathfrak{B})$ are orthogonal, we have from (48) that

$$||\Delta F||_2 = ||\phi_{\mathrm{ct}}(-(t + \Delta t)\mathfrak{B}) - \phi_{\mathrm{ct}}(-t\mathfrak{B})||_2.$$

Combining the above equality with (50) and (52), we obtain

$$||\Xi(t, \Delta t) - \Xi(t, 0)||_2 \leq \frac{\left(1 - \frac{1}{2}||\mathfrak{B}||_2|\Delta t|\right)^2(1 + \frac{1}{4}||\eta_k||_2^2)||\mathfrak{B}||_2|\Delta t|}{\left(\left(1 - \frac{1}{2}||\mathfrak{B}||_2|\Delta t|\right)^2 - \frac{1}{2}\sqrt{1 + \frac{1}{4}||\eta_k||_2^2} \cdot ||\mathfrak{B}||_2|\Delta t|\right)^2}. \tag{53}$$

If

$$|\Delta t| \leq \frac{1}{3\sqrt{1 + \frac{1}{4}||\eta_k||_2^2} \cdot ||\mathfrak{B}||_2} = \frac{1}{3\sqrt{1 + \frac{1}{4}||\eta_k||_2^2} \cdot ||\nabla f(X_k)||_2},$$

then (53) implies

$$
\begin{aligned}
||\Xi(t,\Delta t) - \Xi(t,0)||_2 \le \frac{(1+\frac{1}{4}||\eta_k||_2^2)||\mathfrak{B}||_2|\Delta t|}{((1-\frac{1}{6})^2 - \frac{1}{6})^2} &\le (4+||\eta_k||_2^2)||\mathfrak{B}||_2|\Delta t| \\
&= (4+||\eta_k||_2^2)||\nabla f(X_k)||_2|\Delta t|.
\end{aligned}
$$

Hence, by (40) in Assumption 7 and the above inequality, we obtain

$$
\begin{aligned}
\big\|R^{-1}_{Z_k(t)}(Y_k(t+\Delta t))\big\|_{\mathrm{c}} &\le \big\|R^{-1}_{Z_k(t)}(Y_k(t))\big\|_{\mathrm{c}} + \big\|R^{-1}_{Z_k(t)}(Y_k(t+\Delta t)) - R^{-1}_{Z_k(t)}(Y_k(t))\big\|_{\mathrm{c}} \\
&\le \big\|R^{-1}_{Z_k(t)}(Y_k(t))\big\|_{\mathrm{c}} + \Upsilon\big\|\overline{R}^{-1}_{Z_k(t)}(Y_k(t+\Delta t)) - \overline{R}^{-1}_{Z_k(t)}(Y_k(t))\big\|_{\mathrm{c}} \\
&= \big\|R^{-1}_{Z_k(t)}(Y_k(t))\big\|_{\mathrm{c}} + \frac{\Upsilon}{\sqrt{2}}\big\|\Xi(t,\Delta t) - \Xi(t,0)\big\|_{\mathrm{F}} \\
&\le \big\|R^{-1}_{Z_k(t)}(Y_k(t))\big\|_{\mathrm{c}} + \frac{\sqrt{2n}\Upsilon}{2}\big\|\Xi(t,\Delta t) - \Xi(t,0)\big\|_2 \\
&\le \big\|R^{-1}_{Z_k(t)}(Y_k(t))\big\|_{\mathrm{c}} + \frac{\sqrt{2n}\Upsilon}{2}(4+||\eta_k||_2^2)||\nabla f(X_k)||_2|\Delta t|.
\end{aligned}
$$

This completes the proof. □

Now we can give the main convergence result of Algorithm 2 as follows.

Theorem 2 *Suppose that Assumptions 4–7 hold. Let $\{\tau_k\}$ be the sequence of numbers defined by* (25)*. If*

$$
0 < \beta_k - \alpha_k \le \frac{1}{3\sqrt{1+\frac{1}{4}||\eta_k||_2^2}\cdot||\nabla f(X_k)||_2}
$$

and if $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\lambda_k\}$ are chosen such that

$$
c_k := 1 - L\beta_k - \frac{L\varrho^2(\beta_k-\alpha_k)^2}{\beta_k\lambda_k\tau_k}\left(\sum_{i=k}^{N}\tau_i\right) > 0, \quad 1 \le k \le N, \tag{54}
$$

where

$$
\varrho := \sup_{k\ge 1}\frac{\sqrt{2n}\Upsilon(4+||\eta_k||_2^2)||\nabla f(X_k)||_2}{2||\nabla f(X_k)||_{\mathrm{c}}} \in [4\Upsilon, +\infty), \tag{55}
$$

then Algorithm 2 satisfies for all $N \ge 1$ that

$$
\min_{k=1,\dots,N}||\nabla f(X_k)||_{\mathrm{c}}^2 \le \frac{f(Z_0)-f^*}{\sum_{k=1}^{N}\beta_k c_k}. \tag{56}
$$

Proof. Denote $\Delta_k := \nabla f(Z_{k-1}) - \mathcal{T}^{-1}_{\eta_k}\nabla f(X_k)$. Then we have

$$
\begin{aligned}
||\Delta_k||_{\mathrm{c}} = ||\nabla f(Z_{k-1}) - \mathcal{T}^{-1}_{\eta_k}\nabla f(X_k)||_{\mathrm{c}} &\le L\mathrm{dist}(X_k, Z_{k-1}) \\
&\le L||\eta_k||_{\mathrm{c}} \overset{(21)}{=\!=} L(1-\lambda_k)\big\|R^{-1}_{Z_{k-1}}(Y_{k-1})\big\|_{\mathrm{c}},
\end{aligned}
$$

where the first inequality follows from inequality (36) and the second inequality follows from (22) and Lemma 4. Using inequality (37), we also have

$$
\begin{aligned}
f(Z_k) &\le f(Z_{k-1}) + \big\langle\nabla f(Z_{k-1}), R^{-1}_{Z_{k-1}}(Z_k)\big\rangle + \frac{L}{2}\big\|R^{-1}_{Z_{k-1}}(Z_k)\big\|_{\mathrm{c}}^2 \\
&\overset{(24)}{=\!=} f(Z_{k-1}) + \big\langle\nabla f(Z_{k-1}), -\beta_k\mathcal{T}^{-1}_{\eta_k}\nabla f(X_k)\big\rangle + \frac{L}{2}\big\|\beta_k\mathcal{T}^{-1}_{\eta_k}\nabla f(X_k)\big\|_{\mathrm{c}}^2 \\
&= f(Z_{k-1}) + \big\langle\Delta_k + \mathcal{T}^{-1}_{\eta_k}\nabla f(X_k), -\beta_k\mathcal{T}^{-1}_{\eta_k}\nabla f(X_k)\big\rangle + \frac{L}{2}\big\|\beta_k\mathcal{T}^{-1}_{\eta_k}\nabla f(X_k)\big\|_{\mathrm{c}}^2 \\
&= f(Z_{k-1}) - \beta_k\left(1-\frac{L\beta_k}{2}\right)\big\|\mathcal{T}^{-1}_{\eta_k}\nabla f(X_k)\big\|_{\mathrm{c}}^2 - \beta_k\big\langle\Delta_k, \mathcal{T}^{-1}_{\eta_k}\nabla f(X_k)\big\rangle \\
&\le f(Z_{k-1}) - \beta_k\left(1-\frac{L\beta_k}{2}\right)||\nabla f(X_k)||_{\mathrm{c}}^2 + \beta_k||\Delta_k||_{\mathrm{c}}||\nabla f(X_k)||_{\mathrm{c}}.
\end{aligned}
$$

Combining the previous two inequalities, we obtain

$$\begin{aligned}
f(Z_k) &\leq f(Z_{k-1}) - \beta_k \left(1 - \frac{L\beta_k}{2}\right) ||\nabla f(X_k)||_{\mathrm{c}}^2 + L(1-\lambda_k)\beta_k ||\nabla f(X_k)||_{\mathrm{c}} \left\| R_{Z_{k-1}}^{-1}(Y_{k-1}) \right\|_{\mathrm{c}} \\
&\leq f(Z_{k-1}) - \beta_k \left(1 - \frac{L\beta_k}{2}\right) ||\nabla f(X_k)||_{\mathrm{c}}^2 + \frac{L\beta_k^2}{2} ||\nabla f(X_k)||_{\mathrm{c}}^2 \\
&\quad + \frac{L(1-\lambda_k)^2}{2} \left\| R_{Z_{k-1}}^{-1}(Y_{k-1}) \right\|_{\mathrm{c}}^2 \\
&= f(Z_{k-1}) - \beta_k (1 - L\beta_k) ||\nabla f(X_k)||_{\mathrm{c}}^2 + \frac{L(1-\lambda_k)^2}{2} \left\| R_{Z_{k-1}}^{-1}(Y_{k-1}) \right\|_{\mathrm{c}}^2. \qquad (57)
\end{aligned}$$

Then we have

$$\begin{aligned}
\left\| R_{Z_k}^{-1}(Y_k) \right\|_{\mathrm{c}} &\overset{(23)(24)}{=\!=\!=} \left\| R^{-1}_{R_{Z_{k-1}}(-\beta_k \mathcal{T}_{\eta_k}^{-1} \nabla f(X_k))} \Big(R_{X_k}(-\alpha_k \nabla f(X_k)) \Big) \right\|_{\mathrm{c}} \\
&\leq \left\| R^{-1}_{R_{Z_{k-1}}(-\beta_k \mathcal{T}_{\eta_k}^{-1} \nabla f(X_k))} \Big(R_{X_k}(-\beta_k \nabla f(X_k)) \Big) \right\|_{\mathrm{c}} \\
&\quad + \frac{\sqrt{2n}\Upsilon}{2} (\beta_k - \alpha_k)(4 + ||\eta_k||_2^2) ||\nabla f(X_k)||_2 \\
&\leq ||\eta_k||_{\mathrm{c}} + \varrho(\beta_k - \alpha_k) ||\nabla f(X_k)||_{\mathrm{c}} \\
&= (1-\lambda_k) \left\| R_{Z_{k-1}}^{-1}(Y_{k-1}) \right\|_{\mathrm{c}} + \varrho(\beta_k - \alpha_k) ||\nabla f(X_k)||_{\mathrm{c}},
\end{aligned}$$

where the first inequality follows from Lemma 6 and the second inequality follows from Lemma 5, (38) and (55). Dividing both sides of the above equality by τ_k and noting (25), we have

$$\frac{\left\| R_{Z_k}^{-1}(Y_k) \right\|_{\mathrm{c}}}{\tau_k} \leq \frac{\left\| R_{Z_{k-1}}^{-1}(Y_{k-1}) \right\|_{\mathrm{c}}}{\tau_{k-1}} + \frac{\varrho(\beta_k - \alpha_k) ||\nabla f(X_k)||_{\mathrm{c}}}{\tau_k}.$$

Summing them up and noting $Y_0 = Z_0$, we obtain

$$\left\| R_{Z_k}^{-1}(Y_k) \right\|_{\mathrm{c}} \leq \varrho \tau_k \sum_{i=1}^{k} \frac{\beta_i - \alpha_i}{\tau_i} ||\nabla f(X_i)||_{\mathrm{c}} = \varrho \sum_{i=1}^{k} \tau_k \frac{\lambda_i}{\tau_i} \cdot \frac{\beta_i - \alpha_i}{\lambda_i} ||\nabla f(X_i)||_{\mathrm{c}}.$$

Using the above inequality, Lemma 1, and Jensen's inequality, we have

$$\left\| R_{Z_k}^{-1}(Y_k) \right\|_{\mathrm{c}}^2 \leq \varrho^2 \sum_{i=1}^{k} \tau_k \frac{\lambda_i}{\tau_i} \cdot \frac{(\beta_i - \alpha_i)^2}{\lambda_i^2} ||\nabla f(X_i)||_{\mathrm{c}}^2 = \varrho^2 \tau_k \sum_{i=1}^{k} \frac{(\beta_i - \alpha_i)^2}{\lambda_i \tau_i} ||\nabla f(X_i)||_{\mathrm{c}}^2.$$

Replacing the above bound in (57) and using (25), we obtain

$$\begin{aligned}
f(Z_k) &\leq f(Z_{k-1}) - \beta_k(1 - L\beta_k) ||\nabla f(X_k)||_{\mathrm{c}}^2 + L(1-\lambda_k)^2 \varrho^2 \tau_{k-1} \sum_{i=1}^{k-1} \frac{(\beta_i - \alpha_i)^2}{\lambda_i \tau_i} ||\nabla f(X_i)||_{\mathrm{c}}^2 \\
&\leq f(Z_{k-1}) - \beta_k(1 - L\beta_k) ||\nabla f(X_k)||_{\mathrm{c}}^2 + L\varrho^2 \tau_k \sum_{i=1}^{k-1} \frac{(\beta_i - \alpha_i)^2}{\lambda_i \tau_i} ||\nabla f(X_i)||_{\mathrm{c}}^2.
\end{aligned}$$

Summing up the above inequalities and using the definition of c_k in (54), we have

$$\begin{aligned}
f(Z_N) &\leq f(Z_0) - \sum_{k=1}^{N} \beta_k (1 - L\beta_k) ||\nabla f(X_k)||_{\mathrm{c}}^2 + L\varrho^2 \sum_{k=1}^{N} \tau_k \sum_{i=1}^{k} \frac{(\beta_i - \alpha_i)^2}{\lambda_i \tau_i} ||\nabla f(X_i)||_{\mathrm{c}}^2 \\
&= f(Z_0) - \sum_{k=1}^{N} \beta_k (1 - L\beta_k) ||\nabla f(X_k)||_{\mathrm{c}}^2 + L\varrho^2 \sum_{k=1}^{N} \frac{(\beta_k - \alpha_k)^2}{\lambda_k \tau_k} \left(\sum_{i=k}^{N} \tau_i \right) ||\nabla f(X_k)||_{\mathrm{c}}^2 \\
&= f(Z_0) - \sum_{k=1}^{N} \beta_k c_k ||\nabla f(X_k)||_{\mathrm{c}}^2.
\end{aligned}$$

Re-arranging the terms in the above inequality and noting that $f(Z_N) \geq f^*$ we obtain

$$\min_{k=1,\ldots,N} ||\nabla f(X_k)||_{\mathrm{c}}^2 \left(\sum_{k=1}^{N} \beta_k c_k \right) \leq \sum_{k=1}^{N} \beta_k c_k ||\nabla f(X_k)||_{\mathrm{c}}^2 \leq f(Z_0) - f^*.$$

This completes the proof. □

Corollary 2 *Suppose that Assumptions 4–7 hold and that* $\alpha_k \equiv \frac{1}{2L}$ *and* $\lambda_k = \frac{2}{k+1}$*. If*

$$\alpha_k \leq \beta_k \leq \left(1 + \min\left\{ \frac{\lambda_k}{4\varrho},\ \frac{1}{3\sqrt{1+\frac{1}{4}||\eta_k||_2^2} \cdot ||\nabla f(X_k)||_2} \right\}\right) \alpha_k, \tag{58}$$

where ϱ *is defined in* (55)*, then Algorithm 2 satisfies for all* $N \geq 1$ *that*

$$\min_{k=1,\ldots,N} ||\nabla f(X_k)||_{\mathrm{c}}^2 \leq \frac{5L(f(Z_0) - f^*)}{N}. \tag{59}$$

Proof. This follows from the proof of Corollary 1 in [19], but for self-containedness, we prove the result as follows. By (25) we have

$$\tau_k = \frac{2}{k(k+1)} = \frac{\lambda_k}{k}, \tag{60}$$

which implies

$$\sum_{i=1}^{N} \tau_i = \sum_{i=1}^{N} \frac{2}{i(i+1)} = 2\sum_{i=1}^{N} \left(\frac{1}{i} - \frac{1}{i+1} \right) \leq \frac{2}{k}. \tag{61}$$

It follows from (54), (58), (60), (61), $\lambda_k \leq 1$, and $\varrho \geq 4$ that

$$\begin{aligned} c_k &= 1 - L\left[\beta_k + \frac{\varrho^2(\beta_k - \alpha_k)^2}{\beta_k \lambda_k \tau_k} \left(\sum_{i=k}^{N} \tau_i \right)\right] \\ &\geq 1 - L\left[\left(1 + \frac{\lambda_k}{4\varrho}\right)\alpha_k + \frac{\lambda_k^2 \alpha_k^2}{16} \cdot \frac{1}{\alpha_k \lambda_k \tau_k} \cdot \frac{2}{k}\right] \\ &\geq 1 - L\left[\left(1 + \frac{1}{16}\right)\alpha_k + \frac{1}{8}\alpha_k\right] \\ &= 1 - L\alpha_k\left(1 + \frac{1}{16} + \frac{1}{8}\right) = \frac{13}{32}. \end{aligned}$$

Thus,

$$\beta_k c_k \geq \alpha_k c_k \geq \frac{13}{32}\alpha_k = \frac{13}{64L} \geq \frac{1}{5L}.$$

Combining this with (56), we obtain (59). □

5 Implementation details

5.1 Computing geometric objects

In this subsection we discuss practical ways to computing the geometric objects involved in our AG algorithms. Although the $n \times n$ orthogonal group representation simplifies our theoretical analysis, the $n \times p$ Stiefel manifold representation with efficient implementation is appealing in numerical computation. In the rest of this section, let $X \in \mathrm{St}(n,p)$, $Y \in \mathrm{St}(n,p)$, $\mathcal{X} = \mathrm{span}(X) \in \mathrm{Gr}(n,p)$, and $\mathcal{Y} = \mathrm{span}(Y) \in \mathrm{Gr}(n,p)$.

5.1.1 Geometric objects on the Stiefel manifold

According to [38], the Cayley tranform retraction (14) on the Stiefel manifold has the following low-rank expression:

$$R_X^{\mathrm{St}}(\eta_X) = X + U\Big(I_{2p} - \frac{1}{2}V^\top U\Big)^{-1}V^\top X, \tag{62}$$

where

$$U = \Big[\eta_X - \frac{1}{2}XX^\top\eta_X,\ X\Big],\quad V = \Big[X,\ \frac{1}{2}XX^\top\eta_X - \eta_X\Big].$$

This formula follows from $R_X^{\mathrm{St}}(\eta_X) = \phi_{\mathrm{ct}}(UV^\top)X$ and

$$\phi_{\mathrm{ct}}(UV^\top) = I_n + U\Big(I_{2p} - \frac{1}{2}V^\top U\Big)^{-1}V^\top. \tag{63}$$

The inverse of this retraction is given in [44] as follows:

$$(R_X^{\mathrm{St}})^{-1}(Y) = 2Y(I_p + X^\top Y)^{-1} + 2X(I_p + Y^\top X)^{-1} - 2X. \tag{64}$$

By (63), the vector transport (16) on the Stiefel manifold has the following low-rank expression:

$$\mathcal{T}_{\eta_X}^{\mathrm{St}}(\xi_X) = \phi_{\mathrm{ct}}(UV^\top)\xi_X = \xi_X + U\Big(I_{2p} - \frac{1}{2}V^\top U\Big)^{-1}V^\top\xi_X. \tag{65}$$

Combining (63) and (65), we obtain the inverse of this vector transport:

$$\big(\mathcal{T}_{\eta_X}^{\mathrm{St}}\big)^{-1}(\zeta_Y) = \phi_{\mathrm{ct}}(-UV^\top)\zeta_Y = \zeta_Y - U\Big(I_{2p} + \frac{1}{2}V^\top U\Big)^{-1}V^\top\zeta_Y, \tag{66}$$

where $Y = R_X^{\mathrm{St}}(\eta_X)$. To our knowledge, (66) is new although it is straightforward from (65).

5.1.2 Geometric objects on the Grassmann manifold

According to [14], the exponential map (11) on the Grassmann manifold has the following low-rank expression:

$$\exp_{\mathcal{X}}^{\mathrm{Gr}}(\eta_{\mathcal{X}}) = (XV\cos\Sigma + U\sin\Sigma)V^\top, \tag{67}$$

where $U\Sigma V^\top$ is a thin singular value decomposition (SVD) of η_X^h. The Riemannian logarithm $\log_{\mathcal{X}}^{\mathrm{Gr}}(\mathcal{Y}) = \big(\exp_{\mathcal{X}}^{\mathrm{Gr}}\big)^{-1}(\mathcal{Y})$ on the Grassmann manifold can be computed by Algorithm 5.3 in [6]. If $X^\top Y$ is invertible, an equivalent approach for computing $\log_{\mathcal{X}}^{\mathrm{Gr}}(\mathcal{Y})$ is given in [1]:

$$\log_{\mathcal{X}}^{\mathrm{Gr}}(\mathcal{Y}) = \tilde{U}\arctan(\tilde{\Sigma})\tilde{V}^\top, \tag{68}$$

where $\tilde{U}\tilde{\Sigma}\tilde{V}^\top$ is a thin SVD of $(Y - XX^\top Y)(X^\top Y)^{-1}$.

A low-rank expression for the parallel transport of $\xi_{\mathcal{X}}$ along the geodesic $\gamma(t) = \exp_{\mathcal{X}}^{\mathrm{Gr}}(t\eta_{\mathcal{X}})$ is also given in [14]:

$$P_\gamma^{t\leftarrow 0}\xi_{\mathcal{X}} = \xi_X^h - (XV\sin\Sigma t + U(I_p - \cos\Sigma t))U^\top\xi_X^h. \tag{69}$$

Let $\zeta_{\mathcal{Y}} = P_\gamma^{1\leftarrow 0}\xi_{\mathcal{X}}$ where $\mathcal{Y} = \exp_{\mathcal{X}}^{\mathrm{Gr}}(\eta_{\mathcal{X}})$. Combining (69) with $X^\top\xi_X^h = X^\top U = 0$, we have

$$\cos\Sigma\cdot U^\top\xi_X^h = U^\top\zeta_Y^h,\quad \sin\Sigma\cdot U^\top\xi_X^h = -V^\top X^\top\zeta_Y^h.$$

Then

$$U^\top\xi_X^h = \cos\Sigma\cdot U^\top\zeta_Y^h - \sin\Sigma\cdot V^\top X^\top\zeta_Y^h.$$

Substituting this in (69) yields

$$P_\gamma^{0\leftarrow 1}\zeta_{\mathcal{Y}} = \xi_X^h = \zeta_Y^h - XX^\top\zeta_Y^h + U(I_p - \cos\Sigma)(\cos\Sigma\cdot U^\top\zeta_Y^h - \sin\Sigma\cdot V^\top X^\top\zeta_Y^h). \tag{70}$$

To our knowledge, (70) is not found in the literature although it is not hard to derive it from (69).

According to [45], the Cayley transform retraction (14) on the Grassmann manifold has the following low-rank expression:

$$R^{\mathrm{Gr}}_{\mathcal{X}}(\eta_{\mathcal{X}}) = X + \eta^h_X - \Big(\frac{1}{2}X + \frac{1}{4}\eta^h_X\Big)\Big(I_p + \frac{1}{4}(\eta^h_X)^\top \eta^h_X\Big)^{-1}(\eta^h_X)^\top \eta^h_X. \tag{71}$$

Now we derive a formula for the inverse of this retraction. Let $\eta_{\mathcal{X}} = (R^{\mathrm{Gr}}_{\mathcal{X}})^{-1}(\mathcal{Y})$. This implies $\eta^h_X = (R^{\mathrm{St}}_X)^{-1}(Y\hat{Q})$ for some $\hat{Q} \in O(p)$. Then we have from (64) that

$$\eta^h_X = 2Y\hat{Q}(I_p + X^\top Y\hat{Q})^{-1} + 2X(I_p + \hat{Q}^\top Y^\top X)^{-1} - 2X. \tag{72}$$

Using $X^\top \eta^h_X = 0$, we have

$$\begin{aligned} I_p &= X^\top Y\hat{Q}(I_p + X^\top Y\hat{Q})^{-1} + (I_p + \hat{Q}^\top Y^\top X)^{-1} \\ &= (I_p + X^\top Y\hat{Q} - I_p)(I_p + X^\top Y\hat{Q})^{-1} + (I_p + \hat{Q}^\top Y^\top X)^{-1} \\ &= I_p - (I_p + X^\top Y\hat{Q})^{-1} + (I_p + \hat{Q}^\top Y^\top X)^{-1}. \end{aligned} \tag{73}$$

This implies $X^\top Y\hat{Q} = \hat{Q}^\top Y^\top X$, i.e., $X^\top Y\hat{Q}$ is symmetric. Let $X^\top Y = \hat{U}\hat{\Sigma}\hat{V}^\top$ be an SVD. It is easy to see that $\hat{Q} = \hat{V}\hat{U}^\top$. Then we have from (72) that

$$\begin{aligned} \eta^h_X &= 2Y\hat{V}\hat{U}^\top(I_p + \hat{U}\hat{\Sigma}\hat{U}^\top)^{-1} + 2X(I_p + \hat{U}\hat{\Sigma}\hat{U}^\top)^{-1} - 2X \\ &= 2Y\hat{V}(I_p + \hat{\Sigma})^{-1}\hat{U}^\top + 2X\hat{U}(I_p + \hat{\Sigma})^{-1}\hat{U}^\top - 2X \\ &= 2Y\hat{V}(I_p + \hat{\Sigma})^{-1}\hat{U}^\top - 2X\hat{U}\hat{\Sigma}(I_p + \hat{\Sigma})^{-1}\hat{U}^\top \\ &= 2(Y\hat{V} - X\hat{U}\hat{\Sigma})(I_p + \hat{\Sigma})^{-1}\hat{U}^\top. \end{aligned}$$

Thus we conclude that

$$(R^{\mathrm{Gr}}_{\mathcal{X}})^{-1}(\mathcal{Y}) = 2(Y\hat{V} - X\hat{U}\hat{\Sigma})(I_p + \hat{\Sigma})^{-1}\hat{U}^\top, \tag{74}$$

where $\hat{U}\hat{\Sigma}\hat{V}^\top$ forms an SVD of $X^\top Y$.

A low-rank expression for the vector transport (16) on the Grassmann manifold is also given in [45]:

$$\mathcal{T}^{\mathrm{Gr}}_{\eta_{\mathcal{X}}}(\xi_{\mathcal{X}}) = \xi^h_X - \Big(X + \frac{1}{2}\eta^h_X\Big)\Big(I_p + \frac{1}{4}(\eta^h_X)^\top \eta^h_X\Big)^{-1}(\eta^h_X)^\top \xi^h_X. \tag{75}$$

Now we derive a formula for the inverse of this vector transport. Let $\zeta_{\mathcal{Y}} = \mathcal{T}^{\mathrm{Gr}}_{\eta_{\mathcal{X}}}(\xi_{\mathcal{X}})$ where $\mathcal{Y} = R^{\mathrm{Gr}}_{\mathcal{X}}(\eta_{\mathcal{X}})$. Combining (75) and $X^\top \xi^h_X = X^\top \eta^h_X = 0$, we have

$$X^\top \zeta^h_Y = -\Big(I_p + \frac{1}{4}(\eta^h_X)^\top \eta^h_X\Big)^{-1}(\eta^h_X)^\top \xi^h_X.$$

Substituting the above formula in (75) yields

$$\big(\mathcal{T}^{\mathrm{Gr}}_{\eta_{\mathcal{X}}}\big)^{-1}(\zeta_{\mathcal{Y}}) = \xi^h_X = \zeta^h_Y - \Big(X + \frac{1}{2}\eta^h_X\Big)X^\top \zeta^h_Y. \tag{76}$$

To our knowledge, (74) and (76) are new.

5.2 Other issues

Corollaries 1 and 2 in Section 4 suggest us to choose the stepsizes α_k and β_k as $\alpha_k = \frac{1}{2L}$ and $\beta_k = (1 + \mathcal{O}(\lambda_k))\alpha_k$. However, in real applications, a Lipschitz constant L is not easy to obtain. Even if L is known, this stepsize policy for α_k and β_k is usually not efficient in practice. So, in our implementation, we set $\alpha_1^{\mathrm{ini}} = \frac{1}{L}$ and set α_k^{ini} as the Barzilai–Borwein (BB) stepsize [5] if $k \geq 2$ for the initial guess of α_k. Specifically, let $S_{k-1} = \alpha_{k-1}\nabla f(X_{k-1})$ and $W_{k-1} = \nabla f(X_{k-1}) - \nabla f(Y_{k-1})$ (the subtraction is in the Euclidean sense). We set

$$\alpha_k^{\mathrm{ini}} = \max\left(\min\left(\alpha_k^{\mathrm{BB}}, 10^{20}\right), 10^{-20}\right),$$

where

$$\alpha_k^{\mathrm{BB}} = \frac{\mathrm{Tr}(S_{k-1}^\top S_{k-1})}{|\mathrm{Tr}(S_{k-1}^\top W_{k-1})|} \quad \text{or} \quad \alpha_k^{\mathrm{BB}} = \frac{|\mathrm{Tr}(S_{k-1}^\top W_{k-1})|}{\mathrm{Tr}(W_{k-1}^\top W_{k-1})}.$$

Then we choose $\alpha_k = \alpha_k^{\mathrm{ini}}\mu^{-i_k}$ with $\mu > 1$, where i_k is the smallest nonnegative integer such that

$$f(Y_k) \leq \max\{f(X_k), f(Y_{k-1})\} - \nu\alpha_k||\nabla f(X_k)||_{\mathrm{F}}^2$$

for some constant $\nu \in (0, 1)$. For the stepsize β_k, we simply set $\beta_k = \omega\alpha_k$ for some constant $\omega > 1$.

For numerical stability, we will re-orthogonalize a newly obtained iterate X_+ by the polar decomposition $X_+ \leftarrow X_+(X_+^\top X_+)^{-\frac{1}{2}}$ if infeasibility is detected, i.e., $\left\|X_+^\top X_+ - I_p\right\|_{\mathrm{F}} > 10^{-13}$, where X_+ stands for any of X_k, Y_k, and Z_k. This is exactly the orthogonal projection of X_+ onto the Stiefel manifold $\mathrm{St}(n, p)$ and can also be realized by $X_+ \leftarrow U_+V_+^\top$, where $U_+\Sigma_+V_+^\top$ is an SVD of X_+.

6 Numerical results

In this section, we present numerical results on three test problems to illustrate the efficiency of our AG methods. All experiments were performed in Matlab R2015a on a Thinkpad T480s Laptop with Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz 1.80 GHz and 8GB of RAM.

We compare Algorithm 1 (Alg1) and Algorithm 2 (Alg2) with a basic gradient descent algorithm (Grad) and the state-of-the-art algorithm proposed by Wen and Yin (OptM beta 1.0[1]) [38]. In Grad we use a backtracking strategy for choosing the stepsize α_k. Specifically, $\alpha_k = \alpha_k^{\mathrm{ini}}\mu^{-i_k}$ with $\mu > 1$, where i_k is the smallest nonnegative integer such that

$$f(X_k) \leq f(X_{k-1}) - \nu\alpha_k||\nabla f(X_{k-1})||_{\mathrm{F}}^2.$$

But we choose $\alpha_k^{\mathrm{ini}} = \frac{1}{L}$ as the initial guess for α_k, where the constant L will be reduced to $\frac{1}{2}L$ if

$$f(X_k) \leq f(X_{k-1}) - \frac{1}{4}\alpha_k||\nabla f(X_{k-1})||_{\mathrm{F}}^2.$$

OptM is BB step gradient method with the nonmonotone line search technique proposed by Zhang and Hager [41]. Both of Grad and OptM use the Cayley transform retraction. The stopping criterion is $||\nabla f(X_k)||_{\mathrm{F}} \leq \epsilon$.

6.1 The linear eigenvalue problem

Our first test problem is the linear eigenvalue problem:

$$\min_{X\in\mathbb{R}^{n\times p}} f(X) := \frac{1}{2}\mathrm{Tr}(X^\top AX) \quad \text{s.t.} \quad X^\top X = I_p, \tag{77}$$

where $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix. This is an optimization problem over the Grassmann manifold. In our test, we set $n = 10000$ and $p \in \{25, 50\}$. The data matrix A is constructed as

$$A = \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix}_{n\times n} \quad \text{where} \quad B = \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}_{n/2\times n/2}. \tag{78}$$

The initial point X_0 is generated randomly by $X_0 = \mathrm{orth}(\mathrm{randn}(n, p))$. Since $||B||_2 \leq 4$, we set $L = 4$. The other parameters are chosen as $\mu = 4$, $\nu = 10^{-4}$, $\omega = 5$, and $\epsilon = 10^{-4}$. It is easy to see that any p-dimensional eigenspace corresponding to the zero eigenvalue is an optimal solution to this problem.

We compare Alg1 and Alg2 with Grad and OptM. The results are summarized in Table 1, where "niter" denotes the total number of iterations, "time (s)" denotes the running time in seconds, "fval",

[1] It can be downloaded from https://github.com/optsuite/OptM.

Table 1: Numerical results of problem (77)

$p = 25$	niter	time (s)	fval	nrmg	feasi
Alg1	182	11.1	7.12239977e–06	9.3207e–05	8.6648e–14
Alg2	192	7.2	6.60477142e–06	9.1401e–05	1.6335e–14
Grad	1944	29.7	9.69315450e–06	9.9999e–05	1.4802e–14
OptM	247	4.0	9.73284867e–06	9.8145e–05	3.2344e–15
$p = 50$	niter	time (s)	fval	nrmg	feasi
Alg1	227	27.8	1.07312761e–05	9.8214e–05	6.0155e–14
Alg2	221	16.4	1.12086597e–05	9.8982e–05	2.4271e–14
Grad	2581	76.6	1.36170699e–05	9.9768e–05	3.0983e–14
OptM	290	12.4	1.11601969e–05	9.9853e–05	3.9068e–15

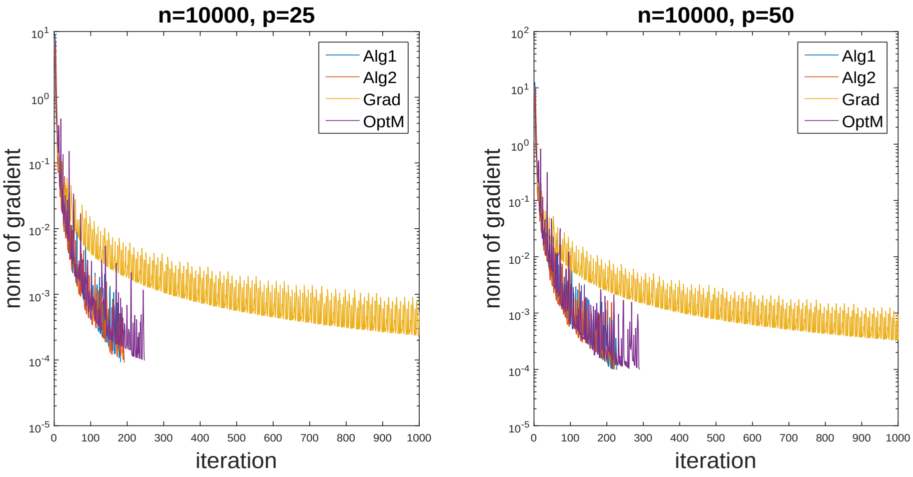

Figure 1: Record of norms of gradients in the results of problem (77)

“nrmg”, and “feasi” denote the final values of the objective function $f(X_k)$, the Frobenius norm of the gradient $||\nabla f(X_k)||_{\mathrm{F}}$, and the feasibility measure $||X_k^\top X_k - I_p||_{\mathrm{F}}$, respectively. We have the following observations. First, Alg1, Alg2, and OptM are significantly better than Grad. Second, Alg2 is basically competitive with OptM; Alg1 or Alg2 has the least number of iterations and OptM takes the least running time. The cause of Alg2 taking more time per iteration than OptM is that AG methods need additional computation for retraction and vector transport compared with gradient-type methods. Third, Alg2 outperforms Alg1 apparently in running time. The reason is that the computational advantage of the Cayley transform retraction and vector transport over the exponential map and parallel transport will be highlighted when the computational effort for the objective function and gradient is not dominant in each iteration; in this problem, since A is sparse, the computational effort for the objective function and gradient is little. We also show the record of norms of gradients generated by the four algorithms (within 1000 iterations) in Figure 1.

6.2 The Karcher mean of subspaces

Our second test problem is the Karcher mean of subspaces [1]:

$$\min_{\mathcal{X}} \; f(\mathcal{X}) := \frac{1}{2m}\sum_{i=1}^{m} \mathrm{dist}^2(\mathcal{X}, \mathcal{Y}_i) \quad \text{s.t.} \quad \mathcal{X} \in \mathrm{Gr}(n, p), \tag{79}$$

Table 2: Numerical results of problem (79)

$n = 500$	niter	time (s)	fval	nrmg	feasi
Alg2	287	13.1	1.53409480e01	9.9231e–05	1.6689e–14
Grad	822	26.5	1.53445078e01	9.8585e–05	3.9953e–15
OptM	346	9.1	1.53413737e01	9.8882e–05	2.8199e–15
$n = 1000$	niter	time (s)	fval	nrmg	feasi
Alg2	484	68.5	3.06673847e01	9.8918e–05	1.8319e–14
Grad	1983	203.5	3.06642887e01	9.8155e–05	7.4222e–15
OptM	822	63.6	3.06621452e01	9.8843e–05	4.1647e–15

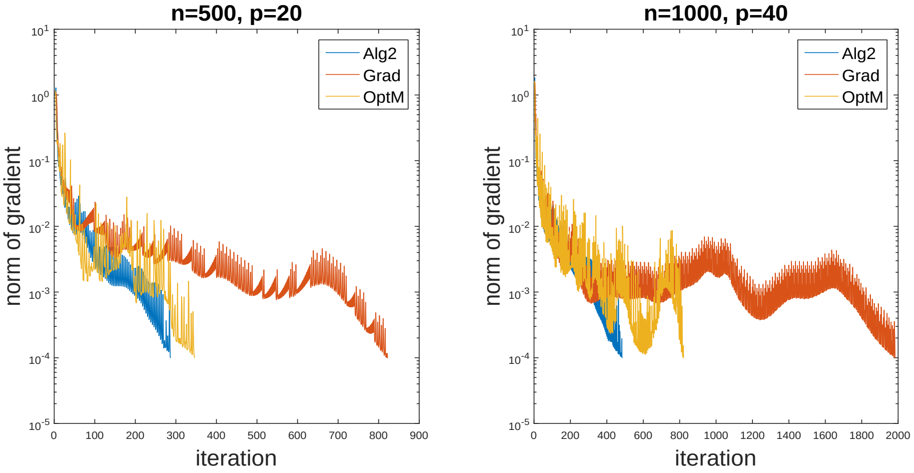

Figure 2: Record of norms of gradients in the results of problem (79)

where $\mathcal{Y}_i \in \mathrm{Gr}(n, p)$, $i = 1, \ldots, m$. This is also an optimization problem over the Grassmann manifold. Equivalently, we can reformulate (79) as

$$\min_{\mathcal{X}} \; f(\mathcal{X}) := \frac{1}{2m} \sum_{i=1}^{m} \left\| \log_{\mathcal{X}}^{\mathrm{Gr}}(\mathcal{Y}_i) \right\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \mathcal{X} \in \mathrm{Gr}(n, p).$$

By the Gauss lemma (see, e.g., Lemma 5.5.5 in [32]), the (Riemannian) gradient of f is

$$\nabla f(\mathcal{X}) = -\frac{1}{m} \sum_{i=1}^{m} \left(\exp_{\mathcal{X}}^{\mathrm{Gr}} \right)^{-1}(\mathcal{Y}_i) = -\frac{1}{m} \sum_{i=1}^{m} \log_{\mathcal{X}}^{\mathrm{Gr}}(\mathcal{Y}_i),$$

where the Riemannian logarithm $\log_{\mathcal{X}}^{\mathrm{Gr}}(\cdot)$ is computed by the methods described in Section 5.1.1. In our test, we set $m = 30$ and $(n, p) \in \{(500, 20), (1000, 40)\}$. The data matrices Y_i and the initial point X_0 are generated randomly by $Y_i = \mathrm{orth}(\mathrm{randn}(n, p))$ and $X_0 = \mathrm{orth}(\mathrm{randn}(n, p))$. The constant L is set as $L = 20$ for the $n = 500$ case and $L = 30$ for the $n = 1000$ case. The other parameters μ, ν, ω, and ϵ are the same as those in Section 6.1.

In this test, we only compare Alg2 with Grad and OptM due to the computational superiority of Alg2 to Alg1. The results are summarized in Table 2 and the record of norms of gradients generated by the three algorithms are shown in Figure 2. We can still observe that Alg2 and OptM are far better than Grad and that Alg2 is basically competitive with OptM.

Table 3: Numerical results of problem (80)

$n = 1000$	niter	time (s)	fval	nrmg	feasi
Alg2	177	14.2	3.39843639e–06	9.7353e–05	5.0400e–15
Grad	2153	117.3	4.66722062e–06	9.9349e–05	4.4932e–15
OptM	229	10.2	4.96948811e–06	9.9683e–05	2.7726e–15
$n = 2000$	niter	time (s)	fval	nrmg	feasi
Alg2	207	63.2	-9.89064099e–06	9.7405e–05	4.7429e–15
Grad	2383	519.2	-2.80447993e–06	9.9775e–05	4.1585e–15
OptM	289	47.4	-1.37668293e–05	9.8654e–05	2.9845e–15

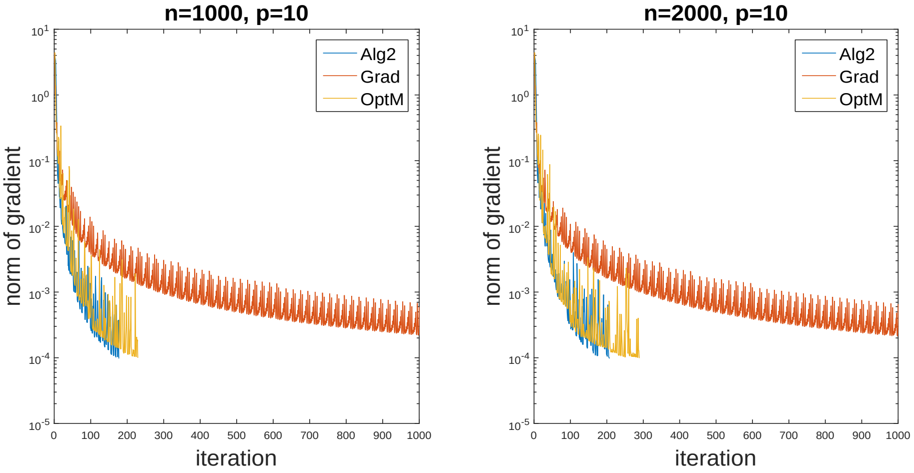

Figure 3: Record of norms of gradients in the results of problem (80)

6.3 Minimization of sums of heterogeneous quadratic functions

Our third test problem is minimization of sums of heterogeneous quadratic functions [7]:

$$\min_{X \in \mathbb{R}^{n \times p}} \ f(X) := \frac{1}{2} \sum_{i=1}^{p} X_{(i)}^{\top} A_i X_{(i)} \quad \text{s.t.} \quad X^{\top} X = I_p, \tag{80}$$

where $A_i \in \mathbb{R}^{n \times n}$, $i = 1, \ldots, p$ are symmetric matrices. This is an optimization problem over the Stiefel manifold. There are no efficient numerical methods to solve this problem yet even if the data matrices A_i have very simple structures, say, A_i are diagonal. In our test, we set $n \in \{1000, 2000\}$ and $p = 10$. The data matrices A_i are constructed as $A_i = A + \frac{1}{2}(E_i + E_i^{\top})$, where A has the same structure as in (78) and E_i are random perturbation matrices generated by $E_i = 10^{-6}\text{randn}(n)$. The initial point X_0 is generated randomly by $X_0 = \text{orth}(\text{randn}(n, p))$. The constant L is chosen as $L = 5$ and the other parameters μ, ν, ω, and ϵ are the same as those in Section 6.1.

We compare Alg2 with Grad and OptM. The results are summarized in Table 3 and the record of norms of gradients generated by the three algorithms (within 1000 iterations) are shown in Figure 3. The numerical behaviors of the three algorithms in this test problem are similar to those in the previous two test problems.

7 Conclusions

In this paper we extend a nonconvex Nesterov-type accelerated gradient method to optimization over the Grassmann and Stiefel manifolds. We have made two main contributions. On the one hand, we have proposed two implementable Riemannian accelerated gradient algorithms. The first one, designed specially for the Grassmann manifold, is based on the exponential map and parallel transport. The second one, designed for both of the Grassmann and Stiefel manifolds, is based on the Cayley transform retraction and vector transport. Moreover, efficient formulas for the inverse maps of the Cayley transform retraction and vector transport are obtained. On the other hand, we have obtained the global rate of convergence of the proposed algorithms under some reasonable assumptions. To our knowledge, this is the first result of global convergence rate of Riemannian Nesterov-type accelerated gradient methods for non-geodesically convex optimization (on the Grassmann and Stiefel manifolds). Preliminary numerical results on three test problems illustrate the potential effectiveness of the proposed algorithms. Our future work will focus on efficient implementation of our accelerated gradient methods and their extension to other specific or even general manifolds.

Funding and Conflicts of Interests

This work is supported by the National Natural Science Foundation of China (Grant Nos. 12271342, 12172210 and 11601317).

References

[1] Absil, P.-A., Mahony, R., Sepulchre, R.: Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. Acta Appl. Math. 80, 199–220 (2004)

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

[3] Agarwal, N., Boumal, N., Bullins, B., Cartis, C.: Adaptive regularization with cubics on manifolds. Math. Program. 188, 85–134 (2021)

[4] Ahn, K., Sra, S.: From Nesterov's estimate sequence to Riemannian acceleration. arXiv:2001.08876v1 (2020)

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

[6] Bendokat, T., Zimmermann, R., Absil, P.-A.: A Grassmann manifold handbook: basic geometry and computational aspects. arXiv:2011.13699v2 (2020)

[7] Bolla, M., Michaletzky, G., Tusnády, G., Ziermann, M.: Extrema of sums of heterogeneous quadratic forms. Linear Algebra Appl. 269, 331–365 (1998)

[8] Boumal, N.: An Introduction to Optimization on Smooth Manifolds. Cambridge University Press (2023)

[9] Boumal, N., Absil, P.-A.: Low-rank matrix completion via preconditioned optimization on the Grassmann manifold. Linear Algebra Appl. 475, 200–239 (2015)

[10] Boumal, N., Absil, P.-A., Cartis, C.: Global rates of convergence for nonconvex optimization on manifolds. IMA J. Numer. Anal. 39, 1–33 (2018)

[11] Chen, S., Ma, S., So, A. M.-C., Zhang, T.: Proximal gradient method for nonsmooth optimization over the Stiefel manifold. SIAM J. Optim. 30, 210–239 (2020)

[12] Criscitiello, C., Boumal, N.: An accelerated first-order method for non-convex optimization on manifolds. Found. Comput. Math. 23, 1433–1509 (2023)

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

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

[15] Gallot, S., Hulin, D., Lafontaine, J. Riemannian Geometry, 3rd edn. Springer, Berlin, Heidelberg (2004)

[16] Gao, B., Absil, P.-A.: A Riemannian rank-adaptive method for low-rank matrix completion. Comput. Optim. Appl. 81, 67–90 (2022)

[17] Gao, B., Liu, X., Chen, X., Yuan, Y.: A new first-order algorithmic framework for optimization problems with orthogonality constraints. SIAM J. Optim. 28, 302–332 (2018)

[18] Gao, B., Thanh Son, N., Absil, P.-A., Stykel, T.: Riemannian optimization on the symplectic Stiefel manifold. SIAM J. Optim. 31, 1546–1575 (2021)

[19] Ghadimi, S., Lan, G.: Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Math. Program. 156, 59–99 (2016)

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

[21] Hu, J., Jiang, B., Lin, L., Wen, Z., Yuan, Y.: Structured quasi-Newton methods for optimization with orthogonality constraints. SIAM J. Sci. Comput. 41, A2230–A2269 (2019)

[22] Hu, J., Liu, X., Wen, Z., Yuan, Y.: A brief introduction to manifold optimization. J. Oper. Res. Soc. China 8, 199–248 (2020)

[23] Hu, J., Milzarek, A., Wen, Z., Yuan, Y.: Adaptive quadratically regularized Newton method for Riemannian optimization. SIAM J. Matrix Anal. Appl. 39, 1181–1207 (2018)

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

[25] 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)

[26] Huang, W., Wei, K.: Riemannian proximal gradient methods. Math. Program. 194, 371–413 (2022)

[27] Huang, W., Wei, K.: An inexact Riemannian proximal gradient method. Comput. Optim. Appl. 85, 1–32 (2023)

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

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

[30] Lim, L.-H., Wong, K. S.-W., Ye, K.: Numerical algorithms on the affine Grassmannian. SIAM J. Matrix Anal. Appl. 40, 371–393 (2019)

[31] Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Mathematics Doklady 27(2), 372–376 (1983)

[32] Petersen, P. Riemannian Geometry, 3rd edn. Springer International Publishing (2016)

[33] Sato, H.: Riemannian Optimization and Its Applications. Springer Nature, Switzerland (2021)

[34] Sato, H., Iwai, T.: Optimization algorithms on the Grassmann manifold with application to matrix eigenvalue problems. Japan J. Indust. Appl. Math. 31, 355–400 (2014)

[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. J. Comp. Math. 39, 207–226 (2021)

[37] Wang, L., Gao, B., Liu, X.: Multipliers correction methods for optimization problems over the Stiefel manifold. CSIAM Trans. Appl. Math. 2, 508–531 (2021)

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

[39] Yau, S.-T.: Non-existence of continuous convex functions on certain Riemannian manifolds. Math. Ann. 207, 269–270 (1974)

[40] Ye, K., Wong, K. S.-W., Lim, L.-H.: Optimization on flag manifolds. Math. Program. 194, 621–660 (2022)

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

[42] Zhang, H., Sra, S.: An estimate sequence for geodesically convex optimization. Proceedings of Machine Learning Research, vol. 75, 1–21 (2018)

[43] Zhu, X.: A Riemannian conjugate gradient method for optimization on the Stiefel manifold. Comput. Optim. Appl. 67, 73–110 (2017)

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

[45] Zhu, X., Sato, H.: Cayley-transform-based gradient and conjugate gradient algorithms on Grassmann manifolds. Adv. Comput. Math. 47:56 (2021)