

RESEARCH ARTICLE

A new, globally convergent Riemannian conjugate gradient method

Hiroyuki Sato* and Toshihiro Iwai

*Department of Applied Mathematics and Physics, Graduate School of Informatics,
Kyoto University, Kyoto 606-8501, Japan*

()

This article deals with the conjugate gradient method on a Riemannian manifold with interest in global convergence analysis. The existing conjugate gradient algorithms on a manifold endowed with a vector transport need the assumption that the vector transport does not increase the norm of tangent vectors, in order to confirm that generated sequences have a global convergence property. In this article, the notion of a scaled vector transport is introduced to improve the algorithm so that the generated sequences may have a global convergence property under a relaxed assumption. In the proposed algorithm, the transported vector is rescaled in case its norm has increased during the transport. The global convergence is theoretically proved and numerically observed with examples. In fact, numerical experiments show that there exist minimization problems for which the existing algorithm generates divergent sequences, but the proposed algorithm generates convergent sequences.

Keywords: conjugate gradient method; Riemannian optimization; global convergence; “scaled” vector transport; Wolfe conditions

AMS Subject Classification: 65K05; 49M37; 90C30

1. Introduction

The conjugate gradient method was first developed by Hestenes and Stiefel as a tool for solving the linear equation $Ax = b$, where A is an $n \times n$ positive definite matrix [7]. The strategy of the linear conjugate gradient method is to minimize the quadratic function $x^T Ax/2 - b^T x$ of x in the successive search directions which are generated in such a manner that those directions are mutually conjugate with respect to A and eventually span the whole $\mathbb{R}^n$. As this method is generalized to be applicable to functions which are not restricted to those quadratic in x , the conjugate gradient method in its original form is particularly called the linear conjugate gradient method.

According to a nonlinear conjugate gradient method for minimizing a smooth function f which is not necessarily quadratic, the search direction η_k is determined by

$$\eta_k = -\text{grad } f(x_k) + \beta_k \eta_{k-1}, \quad (1.1)$$

*Corresponding author. Email: hsato@amp.i.kyoto-u.ac.jp

where β_k is a parameter to be defined suitably. Fletcher and Reeves [5] proposed to define β_k by $\beta_k := \|\text{grad } f(x_k)\|^2 / \|\text{grad } f(x_{k-1})\|^2$ (see [8] for another way to determine β_k).

On the other hand, iterative optimization methods on $\mathbb{R}^n$ have been developed so as to be applicable on Riemannian manifolds [1, 4]. Those generalized methods are called Riemannian optimization methods, which provide procedures for minimizing objective functions defined on a Riemannian manifold M . In a Riemannian optimization method, the usual line search should be replaced [1], as the concept of a line is generalized on a Riemannian manifold. Absil, Mahony, and Sepulchre proposed to use a retraction map to perform a search on a curve on M in place of the line search. As for the conjugate gradient method, Smith provided in [11] a conjugate gradient method on M along with other optimization algorithms on M . The difficulty we encounter in generalizing the conjugate gradient method to that on a manifold is that Eq. (1.1) makes no longer sense. This is because $\text{grad } f(x_k)$ and η_{k-1} belong to tangent spaces at different points on M in general, so that they cannot be added. Smith proposed to use the parallel translation along the geodesic at each iteration in order to make possible the addition of two tangent vectors and thereby to extend the iteration procedure (1.1). However, using the parallel translation on M is not computationally effective in general. A way to perform the conjugate gradient method on M in an efficient manner is to use a vector transport [1]. The global convergence in the conjugate gradient method with a vector transport on M has been recently discussed by Ring and Wirth [9]. They proved the global convergence under the condition that the vector transport in use does not increase the norm of the search direction vector. On the contrary, the present article provides numerical evidence to show that if the assumption is not satisfied, the conjugate gradient method with a general vector transport may fail to generate a globally converging series. In order to relax the assumption in [9], the notion of a “scaled” vector transport is introduced in this article and a new conjugate gradient algorithm is proposed with only a mild computational overhead per iteration.

The organization of this paper is as follows: The scaled vector transport is introduced in Section 2 after a brief review of some useful existing concepts. How to compute the step size is also discussed in this section. In Section 3, a brief review is made of the conjugate gradient method on a Riemannian manifold M , and then a new algorithm is proposed, in which the scaled vector transport is applied only if the vector transport increases the norm of the previous search direction. In Section 4, the global convergence for the proposed algorithm is proved in a manner similar to the usual one performed on $\mathbb{R}^n$, where the scaled vector transport used on a fitting occasion makes a generated sequence into a globally convergent one. Section 5 provides numerical experiments on simple problems which the existing algorithm cannot solve efficiently but the proposed algorithm can do. The numerical experiments show why the present algorithm can generate convergent sequences. Section 6 includes concluding remarks. It is shown in Appendix A that the Lipschitzian condition referred to in Subsection 4.1 is satisfied for some practical Riemannian optimization problems.

2. Setup for Riemannian optimization

2.1. Retraction

An unconstrained optimization problem on a Riemannian manifold M is described as follows:

Problem 2.1

$$\text{minimize } f(x), \quad (2.1)$$

$$\text{subject to } x \in M. \quad (2.2)$$

If M is the Euclidean space $\mathbb{R}^n$, the line search is performed with the updating formula

$$x_{k+1} = x_k + \alpha_k \eta_k, \quad (2.3)$$

where $x_k, x_{k+1} \in \mathbb{R}^n$ are a current point and an unknown next point, respectively, and where $\eta_k \in \mathbb{R}^n$ and $\alpha_k > 0$ are a search direction at x_k and a step size, respectively. However, the line search (2.3) does not make sense on a general manifold M . In order to generalize the line search (2.3) on $\mathbb{R}^n$ to that on M , the search direction η_k should be taken as a tangent vector in $T_{x_k}M$, and the addition in Eq. (2.3) should be replaced by another suitable operation. A natural alternative to the line search is a search along the geodesic emanating from x_k in the direction of η_k , but the geodesic will cause computational difficulty except for a few particular manifolds where the geodesics admit a tractable closed-form expression. A computationally efficient way is to use the following retraction map introduced in [1].

Definition 2.2 Let M and TM be a manifold and the tangent bundle of M , respectively. Let $R : TM \rightarrow M$ be a smooth map and R_x the restriction of R to T_xM . The R is called a retraction on M , if it has the following properties:

- (1) $R_x(0_x) = x$, where 0_x denotes the zero element of T_xM .
- (2) With the canonical identification $T_{0_x}T_xM \simeq T_xM$, R_x satisfies

$$DR_x(0_x) = \text{id}_{T_xM}, \quad (2.4)$$

where $DR_x(0_x)$ denotes the derivative of R_x at 0_x , and id_{T_xM} the identity map on T_xM .

As is easily seen, the exponential map on M is a typical example of a retraction. If we can find a computationally preferable retraction, we can perform an optimization procedure as follows:

Algorithm 2.1 The general framework of optimization methods for Problem 2.1 on a Riemannian manifold M

- 1: Choose an initial point $x_0 \in M$.
 - 2: **for** $k = 0, 1, 2, \dots$ **do**
 - 3: Compute the search direction $\eta_k \in T_{x_k}M$ and the step size $\alpha_k > 0$.
 - 4: Compute the next iterate by $x_{k+1} := R_{x_k}(\alpha_k \eta_k)$, where R is a retraction on M .
 - 5: **end for**
-

The choice of a search direction and a step size characterizes the individual optimization method. We proceed to the vector transport in search for computationally efficient conjugate gradient methods.

2.2. Vector transport and scaled vector transport

In a (nonlinear) conjugate gradient method on the Euclidean space $\mathbb{R}^n$, the search directions η_k are chosen to be

$$\eta_k = -\text{grad } f(x_k) + \beta_k \eta_{k-1}, \quad k \geq 0, \quad (2.5)$$

where $\beta_0 = 0$, and where β_k with $k \geq 1$ are determined in several possible manners. For example, β_k are determined by

$$\beta_k^{\text{FR}} = \frac{\text{grad } f(x_k)^T \text{grad } f(x_k)}{\text{grad } f(x_{k-1})^T \text{grad } f(x_{k-1})}, \quad (2.6)$$

or

$$\beta_k^{\text{PR}} = \frac{\text{grad } f(x_k)^T (\text{grad } f(x_k) - \text{grad } f(x_{k-1}))}{\text{grad } f(x_{k-1})^T \text{grad } f(x_{k-1})}, \quad (2.7)$$

where FR and PR are abbreviations of Fletcher-Reeves and Polak-Ribière, respectively [8].

However, if $\mathbb{R}^n$ is replaced by a Riemannian manifold M , $\text{grad } f(x_k) \in T_{x_k}M$ and $\eta_{k-1} \in T_{x_{k-1}}M$ belong to different tangent spaces, so that $-\text{grad } f(x_k) + \beta_k \eta_{k-1}$ in Eq. (2.5) does not make sense. The quantity $\text{grad } f(x_k) - \text{grad } f(x_{k-1})$ in Eq. (2.7) makes no sense on M either. In order to modify the vector addition in Eqs. (2.5) and (2.7) into a suitable operation on M , Smith proposed to use the parallel translation of tangent vectors along a geodesic [11]. However, no computationally efficient formula is known for the parallel translation along a geodesic even for the Stiefel manifold except when it reduces to the sphere or the orthogonal group. Absil *et al.* [1] proposed the notion of a vector transport as an alternative to the parallel translation. The vector transport is a generalization of the parallel translation and can enhance computational efficiency of algorithms, if defined suitably.

In this paper, we focus on the differentiated retraction $\mathcal{T}^R$ as a vector transport, which is defined to be

$$\mathcal{T}_{\eta_x}^R(\xi_x) := \text{DR}_x(\eta_x)[\xi_x], \quad \eta_x, \xi_x \in T_x M, \quad (2.8)$$

where R is a retraction on M . We here note that $\mathcal{T}^R$ satisfies the conditions in the definition of a vector transport, as is easily verified [1].

In what follows, we assume that M is a Riemannian manifold and denote the Riemannian metric evaluated at $x \in M$ by $\langle \cdot, \cdot \rangle_x$. The norm of a tangent vector $\xi_x \in T_x M$ evaluated at $x \in M$ is defined to be $\|\xi_x\|_x = \sqrt{\langle \xi_x, \xi_x \rangle_x}$. We here have to note that though the parallel translation is an isometry, a vector transport is not required to preserve the norm of vectors in general. The differentiated retraction $\mathcal{T}^R$ is not always an isometry either. In analysing the convergence for the conjugate gradient method later, it will be crucial whether the vector transport $\mathcal{T}^R$ increases the norm of vectors or not. In order to prevent the vector transport $\mathcal{T}^R$ from increasing the norm of vectors, we define the scaled vector transport $\mathcal{T}^0 : TM \oplus TM \rightarrow TM$ associated with $\mathcal{T}^R$ as follows:

Definition 2.3 Let R be a retraction on a Riemannian manifold M . Let $\mathcal{T}^R$ be a vector transport defined by (2.8) with respect to R . The scaled vector transport

$\mathcal{T}^0$ associated with $\mathcal{T}^R$ is defined as

$$\mathcal{T}_{\eta_x}^0(\xi_x) = \frac{\|\xi_x\|_x}{\|\mathcal{T}_{\eta_x}^R(\xi_x)\|_{R_x(\eta_x)}} \mathcal{T}_{\eta_x}^R(\xi_x), \quad \eta_x, \xi_x \in T_x M. \quad (2.9)$$

The scaled vector transport $\mathcal{T}^0$ thus defined is no longer a vector transport since it is not linear. However, $\mathcal{T}^0$ satisfies

$$\|\mathcal{T}_{\eta_x}^0(\xi_x)\|_{R_x(\eta_x)} = \|\xi_x\|_x, \quad \eta_x, \xi_x \in T_x M, \quad (2.10)$$

which is a key property for the global convergence of the algorithm we will propose.

2.3. Strong Wolfe conditions

In computing the step size α_k in the conjugate gradient method on $\mathbb{R}^n$, the strong Wolfe conditions are often used [8], which require α_k to satisfy

$$f(x_k + \alpha_k \eta_k) \leq f(x_k) + c_1 \alpha_k \text{grad } f(x_k)^T \eta_k, \quad (2.11)$$

$$|\text{grad } f(x_k + \alpha_k \eta_k)^T \eta_k| \leq c_2 |\text{grad } f(x_k)^T \eta_k|, \quad (2.12)$$

with $0 < c_1 < c_2 < 1$. In particular, c_1 and c_2 are often taken so as to satisfy $0 < c_1 < c_2 < 1/2$ in the conjugate gradient method. In order to extend the strong Wolfe conditions on $\mathbb{R}^n$ to those on M , we start by reviewing the strong Wolfe conditions (2.11) and (2.12). For a current point x_k and a search direction η_k , one performs a line search for the function defined by

$$\phi(\alpha) = f(x_k + \alpha \eta_k), \quad \alpha > 0. \quad (2.13)$$

Requiring α_k to give a sufficient decrease in the value of f , one imposes the condition

$$\phi(\alpha_k) \leq \phi(0) + c_1 \alpha_k \phi'(0), \quad (2.14)$$

which yields (2.11). In order to prevent α_k from being excessively short, the α_k is required to satisfy

$$|\phi'(\alpha_k)| \leq c_2 |\phi'(0)|, \quad (2.15)$$

which implies (2.12).

In order to generalize the strong Wolfe conditions to those on M , we define a function ϕ on M , in an analogous manner to (2.13), to be

$$\phi(\alpha) = f(R_{x_k}(\alpha \eta_k)), \quad \alpha > 0, \quad (2.16)$$

where R is a retraction on M . The conditions (2.14) and (2.15) applied to (2.16) give rise to

$$f(R_{x_k}(\alpha_k \eta_k)) \leq f(x_k) + c_1 \alpha_k \langle \text{grad } f(x_k), \eta_k \rangle_{x_k}, \quad (2.17)$$

$$|\langle \text{grad } f(R_{x_k}(\alpha_k \eta_k)), DR_{x_k}(\alpha_k \eta_k)[\eta_k] \rangle_{R_{x_k}(\alpha_k \eta_k)}| \leq c_2 |\langle \text{grad } f(x_k), \eta_k \rangle_{x_k}|, \quad (2.18)$$

respectively, where $0 < c_1 < c_2 < 1$. We call the conditions (2.17) and (2.18) the strong Wolfe conditions. The existence of a step size satisfying (2.17) and (2.18) can be shown by an almost verbatim repetition of that for the strong Wolfe conditions on $\mathbb{R}^n$ (see [8]).

Proposition 2.4 Let M be a Riemannian manifold with a retraction R . If a smooth objective function f on M is bounded below on $\{R_{x_k}(\alpha\eta_k) | \alpha > 0\}$ for $x_k \in M$ and for a descent direction $\eta_k \in T_{x_k}M$, and if constants c_1 and c_2 satisfy $0 < c_1 < c_2 < 1$, then there exists a step size α_k which satisfies the strong Wolfe conditions (2.17) and (2.18).

We note that the strong Wolfe conditions (2.17) and (2.18) together with the existence of a step size satisfying them are also discussed in [9].

We now look into the second condition (2.18). If we introduce a vector transport $\mathcal{T}^R$ as the differentiated retraction given by (2.8), then Eq. (2.18) can be expressed as

$$|\langle \text{grad } f(R_{x_k}(\alpha_k \eta_k)), \mathcal{T}_{\alpha_k \eta_k}^R(\eta_k) \rangle_{R_{x_k}(\alpha_k \eta_k)}| \leq c_2 |\langle \text{grad } f(x_k), \eta_k \rangle_{x_k}|. \quad (2.19)$$

An idea for further generalization of this condition to that in an algorithm with a general vector transport $\mathcal{T}$ is to replace (2.19) by

$$|\langle \text{grad } f(R_{x_k}(\alpha_k \eta_k)), \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{R_{x_k}(\alpha_k \eta_k)}| \leq c_2 |\langle \text{grad } f(x_k), \eta_k \rangle_{x_k}|. \quad (2.20)$$

However, if $\mathcal{T} \neq \mathcal{T}^R$, the existence of a step size satisfying both (2.17) and (2.20) is unclear in general. In view of this, the differentiated retraction $\mathcal{T}^R$ is considered to be a natural choice of a vector transport $\mathcal{T}$, for which a step size satisfying (2.17) and (2.20) is shown to exist. In what follows, we use the differentiated retraction $\mathcal{T}^R$ and the scaled one $\mathcal{T}^0$.

3. A new conjugate gradient method on a Riemannian manifold

If a Riemannian manifold M is given a retraction R and the corresponding vector transport $\mathcal{T}^R$, a standard Fletcher-Reeves type conjugate gradient method on M is described as follows [1, 9]:

Algorithm 3.1 A standard Fletcher-Reeves type conjugate gradient method for Problem 2.1 on a Riemannian manifold M

- 1: Choose an initial point $x_0 \in M$.
- 2: Set $\eta_0 = -\text{grad } f(x_0)$.
- 3: **for** $k = 0, 1, 2, \dots$ **do**
- 4: Compute the step size $\alpha_k > 0$ satisfying the strong Wolfe conditions (2.17) and (2.18) with $0 < c_1 < c_2 < 1/2$. Set

$$x_{k+1} = R_{x_k}(\alpha_k \eta_k), \quad (3.1)$$

where R is a retraction on M .

- 5: Set

$$\beta_{k+1} = \frac{\|\text{grad } f(x_{k+1})\|_{x_{k+1}}^2}{\|\text{grad } f(x_k)\|_{x_k}^2}, \quad (3.2)$$

$$\eta_{k+1} = -\text{grad } f(x_{k+1}) + \beta_{k+1} \mathcal{T}_{\alpha_k \eta_k}^R(\eta_k), \quad (3.3)$$

where $\mathcal{T}^R$ is the differentiated retraction defined by (2.8).

- 6: **end for**
-

In [9], the convergence property of Algorithm 3.1 is verified under the assumption that the inequality

$$\|\mathcal{T}_{\alpha_k \eta_k}^R(\eta_k)\|_{x_{k+1}} \leq \|\eta_k\|_{x_k} \quad (3.4)$$

holds for all $k \in \mathbb{N}$. However, the assumption does not always hold in general. For example, the assumption does not hold on the sphere endowed with the orthographic retraction [2]. In Section 5, we will numerically treat such a case.

We wish to relax the assumption (3.4) by using a scaled vector transport. An idea for improving Algorithm 3.1 is to replace $\mathcal{T}^R$ by the scaled vector transport $\mathcal{T}^0$ defined by (2.9). However, this causes difficulty in computing effectively a step size α_k satisfying (2.20) with $\mathcal{T} = \mathcal{T}^0$.

A simple but effective idea for improving Algorithm 3.1 is that each step size is always computed so as to satisfy the strong Wolfe conditions (2.17) and (2.18), but the scaled vector transport $\mathcal{T}^0$ is adopted if it is necessary for the purpose of convergence. More specifically, we use the scaled vector transport $\mathcal{T}^0$ only if the vector transport $\mathcal{T}^R$ increases the norm of the previous search direction vector, that is, we introduce $\mathcal{T}^{(k)}$ defined by

$$\mathcal{T}_{\alpha_k \eta_k}^{(k)}(\eta_k) = \begin{cases} \mathcal{T}_{\alpha_k \eta_k}^R(\eta_k), & \text{if } \|\mathcal{T}_{\alpha_k \eta_k}^R(\eta_k)\|_{x_{k+1}} \leq \|\eta_k\|_{x_k}, \\ \mathcal{T}_{\alpha_k \eta_k}^0(\eta_k), & \text{otherwise,} \end{cases} \quad (3.5)$$

as a substitute for $\mathcal{T}^R$ in Step 5 of Algorithm 3.1. This idea is realized in the following algorithm.

Algorithm 3.2 A scaled Fletcher-Reeves type conjugate gradient method for Problem 2.1 on a Riemannian manifold M

- 1: Choose an initial point $x_0 \in M$.
- 2: Set $\eta_0 = -\text{grad } f(x_0)$.
- 3: **for** $k = 0, 1, 2, \dots$ **do**
- 4: Compute the step size $\alpha_k > 0$ satisfying the strong Wolfe conditions (2.17) and (2.18) with $0 < c_1 < c_2 < 1/2$. Set

$$x_{k+1} = R_{x_k}(\alpha_k \eta_k), \quad (3.6)$$

where R is a retraction on M .

- 5: Set

$$\beta_{k+1} = \frac{\|\text{grad } f(x_{k+1})\|_{x_{k+1}}^2}{\|\text{grad } f(x_k)\|_{x_k}^2}, \quad (3.7)$$

$$\eta_{k+1} = -\text{grad } f(x_{k+1}) + \beta_{k+1} \mathcal{T}_{\alpha_k \eta_k}^{(k)}(\eta_k), \quad (3.8)$$

where $\mathcal{T}^{(k)}$ is defined by (3.5), and where $\mathcal{T}^R$ and $\mathcal{T}^0$ are the differentiated retraction and the associated scaled vector transport defined by (2.8) and (2.9), respectively.

- 6: **end for**
-

We will prove in Section 4 the global convergence property of the proposed algorithm, and give in Section 5 numerical examples in which the inequality (3.4) does not hold for all $k \in \mathbb{N}$ but our Algorithm 3.2 indeed has an advantage in generating convergent sequences.

4. Convergence analysis of the new algorithm

In this section, we verify the convergence property of Algorithm 3.2.

4.1. Zoutendijk's theorem

Zoutendijk's theorem about a series associated with search directions on $\mathbb{R}^n$ is not only valid for the conjugate gradient method but also valid for general descent algorithms [8]. This theorem can be generalized so as to be applicable to a general descent algorithm (Algorithm 2.1) on a Riemannian manifold M . In the same manner as in $\mathbb{R}^n$, we define on a Riemannian manifold M the angle θ_k between the steepest descent direction $-\text{grad } f(x_k)$ and the search direction η_k through

$$\cos \theta_k = -\frac{\langle \text{grad } f(x_k), \eta_k \rangle_{x_k}}{\|\text{grad } f(x_k)\|_{x_k} \|\eta_k\|_{x_k}}. \quad (4.1)$$

Then, Zoutendijk's theorem on M is stated as follows:

Theorem 4.1 Suppose that in Algorithm 2.1 on a Riemannian manifold M , a descent direction η_k and a step size α_k satisfy the strong Wolfe conditions (2.17) and (2.18). If the objective function f is bounded below and of C^1 -class, and if there

exists a Lipschitzian constant $L > 0$ such that

$$|D(f \circ R_x)(t\eta)[\eta] - D(f \circ R_x)(0)[\eta]| \leq Lt, \quad \eta \in T_x M \text{ with } \|\eta\|_x = 1, x \in M, t \geq 0, \quad (4.2)$$

then the following series converges;

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \|\text{grad } f(x_k)\|_{x_k}^2 < \infty. \quad (4.3)$$

The proof of this theorem can be performed in the same manner as that for Zoutendijk's theorem on $\mathbb{R}^n$. See [9] for more detail.

Remark 4.1 *We remark that the inequality (4.2) is a weaker condition than the Lipschitz continuous differentiability of $f \circ R_x$. We will show in Appendix A that Eq. (4.2) holds for objective functions in practical Riemannian optimization problems. A further discussion on the relation with the standard Lipschitz continuous differentiability will be also made in the same appendix.*

4.2. Global convergence

We first extend a lemma in [3] so as to be applicable to Algorithm 3.2 as follows:

Lemma 4.2 The search direction η_k determined in Algorithm 3.2 is a descent direction satisfying

$$-\frac{1}{1 - c_2} \leq \frac{\langle \text{grad } f(x_k), \eta_k \rangle_{x_k}}{\|\text{grad } f(x_k)\|_{x_k}^2} \leq \frac{2c_2 - 1}{1 - c_2}. \quad (4.4)$$

Proof: The proof runs by induction. For $k = 0$, the inequality (4.4) clearly holds on account of

$$\frac{\langle \text{grad } f(x_0), \eta_0 \rangle_{x_0}}{\|\text{grad } f(x_0)\|_{x_0}^2} = \frac{\langle \text{grad } f(x_0), -\text{grad } f(x_0) \rangle_{x_0}}{\|\text{grad } f(x_0)\|_{x_0}^2} = -1. \quad (4.5)$$

We here note that $0 < c_1 < c_2 < 1/2$. Suppose that η_k is a descent direction satisfying (4.4) for some k . Note that on account of Eq. (3.8) with Eq. (3.5), $\mathcal{T}^R$ and $\mathcal{T}^{(k)}$ are related by $\|\mathcal{T}_{\alpha_k \eta_k}^{(k)}(\eta_k)\|_{x_{k+1}} \leq \|\mathcal{T}_{\alpha_k \eta_k}^R(\eta_k)\|_{x_{k+1}}$ in each case. Since $\mathcal{T}_{\alpha_k \eta_k}^{(k)}(\eta_k)$ and $\mathcal{T}_{\alpha_k \eta_k}^R(\eta_k)$ are in the same direction with the inequality $\|\mathcal{T}_{\alpha_k \eta_k}^{(k)}(\eta_k)\|_{x_{k+1}} \leq \|\mathcal{T}_{\alpha_k \eta_k}^R(\eta_k)\|_{x_{k+1}}$ in norm, we have

$$|\langle \text{grad } f(x_{k+1}), \mathcal{T}_{\alpha_k \eta_k}^{(k)}(\eta_k) \rangle_{x_{k+1}}| \leq |\langle \text{grad } f(x_{k+1}), \mathcal{T}_{\alpha_k \eta_k}^R(\eta_k) \rangle_{x_{k+1}}|. \quad (4.6)$$

We also note that the vector transport $\mathcal{T}^R$ is defined to be $\mathcal{T}_{\eta_x}^R(\xi_x) = DR_x(\eta_x)[\xi_x]$ in the algorithm. It then follows from (2.18) and (4.6) that

$$c_2 \langle \text{grad } f(x_k), \eta_k \rangle_{x_k} \leq \langle \text{grad } f(x_{k+1}), \mathcal{T}_{\alpha_k \eta_k}^{(k)}(\eta_k) \rangle_{x_{k+1}} \leq -c_2 \langle \text{grad } f(x_k), \eta_k \rangle_{x_k}, \quad (4.7)$$

where it is to be noted that η_k is in a descent direction. The middle term in (4.4) with $k + 1$ for k is computed as

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

$$= -1 + \frac{\langle \text{grad } f(x_{k+1}), \mathcal{T}_{\alpha_k \eta_k}^{(k)}(\eta_k) \rangle_{x_{k+1}}}{\|\text{grad } f(x_k)\|_{x_k}^2}, \quad (4.8)$$

where the definition (3.7) of β_{k+1} has been used. Therefore, we obtain from (4.7) and (4.8)

$$-1 + c_2 \frac{\langle \text{grad } f(x_k), \eta_k \rangle_{x_k}}{\|\text{grad } f(x_k)\|_{x_k}^2} \leq \frac{\langle \text{grad } f(x_{k+1}), \eta_{k+1} \rangle_{x_{k+1}}}{\|\text{grad } f(x_{k+1})\|_{x_{k+1}}^2} \leq -1 - c_2 \frac{\langle \text{grad } f(x_k), \eta_k \rangle_{x_k}}{\|\text{grad } f(x_k)\|_{x_k}^2}. \quad (4.9)$$

The inequality (4.4) for $k+1$ immediately follows from the induction hypothesis. $\square$

We proceed to the global convergence property of Algorithm 3.2. The convergence of the conjugate gradient method has been already proved on $\mathbb{R}^n$ by Al-Baali [3]. Exploiting the idea of the proof used in [3], we show that Algorithm 3.2 generates converging sequences on a Riemannian manifold.

Theorem 4.3 Consider Algorithm 3.2. If (4.2) and hence (4.3) hold, then

$$\liminf_{k \rightarrow \infty} \|\text{grad } f(x_k)\|_{x_k} = 0. \quad (4.10)$$

Proof: If $\text{grad } f(x_k) = 0$ for some k , let k_0 be the smallest integer among such k . Then, we have $\beta_{k_0} = 0$ and $\eta_{k_0} = 0$ from (3.7) and (3.8) with $k_0 = k+1$, so that $x_{k_0+1} = R_{x_{k_0}}(\alpha_{k_0} \eta_{k_0}) = R_{x_{k_0}}(0) = x_{k_0}$. It then follows that $\text{grad } f(x_k) = 0$ for all $k \geq k_0$. Eq. (4.10) clearly holds in such a case.

We shall consider the case in which $\text{grad } f(x_k) \neq 0$ for all k and prove (4.10) by contradiction. Assume that (4.10) does not hold, that is, there exists a constant $\gamma > 0$ such that

$$\|\text{grad } f(x_k)\|_{x_k} \geq \gamma > 0, \quad \forall k \geq 0. \quad (4.11)$$

Now from (4.1) and (4.4), we obtain

$$\cos \theta_k \geq \frac{1 - 2c_2}{1 - c_2} \frac{\|\text{grad } f(x_k)\|_{x_k}}{\|\eta_k\|_{x_k}}. \quad (4.12)$$

On account of Thm. 4.1, Eqs. (4.3) and (4.12) are put together to provide

$$\sum_{k=0}^{\infty} \frac{\|\text{grad } f(x_k)\|_{x_k}^4}{\|\eta_k\|_{x_k}^2} < \infty. \quad (4.13)$$

On the other hand, Eqs. (4.6), (4.4), and the strong Wolfe condition (2.18) are put together to give

$$\begin{aligned} |\langle \text{grad } f(x_k), \mathcal{T}_{\alpha_{k-1} \eta_{k-1}}^{(k-1)}(\eta_{k-1}) \rangle_{x_k}| &\leq |\langle \text{grad } f(x_k), \mathcal{T}_{\alpha_{k-1} \eta_{k-1}}^R(\eta_{k-1}) \rangle_{x_k}| \\ &\leq -c_2 \langle \text{grad } f(x_{k-1}), \eta_{k-1} \rangle_{x_{k-1}} \\ &\leq \frac{c_2}{1 - c_2} \|\text{grad } f(x_{k-1})\|_{x_{k-1}}^2. \end{aligned} \quad (4.14)$$

Using this inequality and the definition of β_k , we obtain the recurrence inequality

for $\|\eta_k\|_{x_k}^2$ as follows:

$$\begin{aligned}
& \|\eta_k\|_{x_k}^2 \\
&= \|- \text{grad } f(x_k) + \beta_k \mathcal{T}_{\alpha_{k-1}\eta_{k-1}}^{(k-1)}(\eta_{k-1})\|_{x_k}^2 \\
&\leq \|\text{grad } f(x_k)\|_{x_k}^2 + 2\beta_k |\langle \text{grad } f(x_k), \mathcal{T}_{\alpha_{k-1}\eta_{k-1}}^{(k-1)}(\eta_{k-1}) \rangle_{x_k}| + \beta_k^2 \|\mathcal{T}_{\alpha_{k-1}\eta_{k-1}}^{(k-1)}(\eta_{k-1})\|_{x_k}^2 \\
&\leq \|\text{grad } f(x_k)\|_{x_k}^2 + \frac{2c_2}{1-c_2} \beta_k \|\text{grad } f(x_{k-1})\|_{x_{k-1}}^2 + \beta_k^2 \|\eta_{k-1}\|_{x_{k-1}}^2 \\
&= c \|\text{grad } f(x_k)\|_{x_k}^2 + \beta_k^2 \|\eta_{k-1}\|_{x_{k-1}}^2,
\end{aligned} \tag{4.15}$$

where we have used the fact that $\|\mathcal{T}_{\alpha_{k-1}\eta_{k-1}}^{(k-1)}(\eta_{k-1})\|_{x_k} \leq \|\eta_{k-1}\|_{x_{k-1}}$ and put $c := (1 + c_2)/(1 - c_2) > 1$. The successive use of this inequality together with the definition of β_k results in

$$\begin{aligned}
& \|\eta_k\|_{x_k}^2 \\
&\leq c (\|\text{grad } f(x_k)\|_{x_k}^2 + \beta_k^2 \|\text{grad } f(x_{k-1})\|_{x_{k-1}}^2 + \cdots + \beta_k^2 \beta_{k-1}^2 \cdots \beta_2^2 \|\text{grad } f(x_1)\|_{x_1}^2) \\
&\quad + \beta_k^2 \beta_{k-1}^2 \cdots \beta_1^2 \|\eta_0\|_{x_0}^2 \\
&= c \|\text{grad } f(x_k)\|_{x_k}^4 (\|\text{grad } f(x_k)\|_{x_k}^{-2} + \|\text{grad } f(x_{k-1})\|_{x_{k-1}}^{-2} + \cdots + \|\text{grad } f(x_1)\|_{x_1}^{-2}) \\
&\quad + \|\text{grad } f(x_k)\|_{x_k}^4 \|\text{grad } f(x_0)\|_{x_0}^{-2} \\
&< c \|\text{grad } f(x_k)\|_{x_k}^4 \sum_{j=0}^k \|\text{grad } f(x_j)\|_{x_j}^{-2} \leq \frac{c}{\gamma^2} \|\text{grad } f(x_k)\|_{x_k}^4 (k+1),
\end{aligned} \tag{4.16}$$

where use has been made of (4.11) in the last inequality. The inequality (4.16) gives rise to

$$\sum_{k=0}^{\infty} \frac{\|\text{grad } f(x_k)\|_{x_k}^4}{\|\eta_k\|_{x_k}^2} \geq \sum_{k=0}^{\infty} \frac{\gamma^2}{c} \frac{1}{k+1} = \infty. \tag{4.17}$$

This contradicts (4.13) and the proof is completed. $\square$

5. Numerical experiments

In this section, we compare Algorithm 3.2 with Algorithm 3.1 by numerical experiments. As is shown in [9], if the vector transport $\mathcal{T}^R$ as the differentiated retraction satisfies the inequality (3.4), the convergence property of Algorithm 3.1 is proved. However, if (3.4) does not hold, it is not always ensured that sequences generated by Algorithm 3.1 converge. In contrast with this, Algorithm 3.2 indeed works well even if (3.4) fails to hold, as is verified in Thm. 4.3. In the following, we give two examples which show that Algorithm 3.2 works better than Algorithm 3.1. One of the examples is somewhat artificial but well illustrates the situation in which a sequence generated by Algorithm 3.1 is unlikely to converge. The other is a more natural example encountered in a practical problem.

In both of two examples, we consider the following Rayleigh quotient minimization problem on the sphere $S^{n-1} := \{x \in \mathbb{R}^n \mid x^T x = 1\}$ [1, 6]:

Problem 5.1

$$\text{minimize } f(x) = x^T A x, \quad (5.1)$$

$$\text{subject to } x \in S^{n-1}, \quad (5.2)$$

where $A := \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ with $\lambda_1 < \lambda_2 < \dots < \lambda_n$. The optimal solutions of this problem are $\pm(1, 0, 0, \dots, 0)^T$, which are the unit eigenvectors of A associated with the smallest eigenvalue λ_1 .

5.1. A sphere endowed with a peculiar metric

Consider Problem 5.1 with $n = 20$ and $A = \text{diag}(1, 2, \dots, 20)$. A Riemannian metric $g(\cdot, \cdot)$ on S^{n-1} is here defined by

$$g_x(\xi_x, \eta_x) := \xi_x^T G_x \eta_x, \quad \xi_x, \eta_x \in T_x S^{n-1}, \quad (5.3)$$

where $G_x := \text{diag}(10000(x^{(1)})^2 + 1, 1, 1, \dots, 1)$, and where $x^{(1)}$ denotes the first component of the column vector x . It is to be noted that this metric is not the standard one on S^{n-1} . The norm $\|\xi_x\|_x$ of $\xi_x \in T_x S^{n-1}$ is then defined to be $\|\xi_x\|_x = \sqrt{g_x(\xi_x, \xi_x)}$. If x is close to the optimal solutions $\pm(1, 0, 0, \dots, 0)$, then $(x^{(1)})^2$ is nearly 1. Since the first diagonal element of G_x is large because of the coefficient 10000, the closer x is to $\pm(1, 0, 0, \dots, 0)$, the larger the norm $\|\xi_x\|_x$ tends to be.

With respect to the metric (5.3), the gradient of f is described as

$$\text{grad } f(x) = 2 \left(I - \frac{G_x^{-1} x x^T}{x^T G_x^{-1} x} \right) G_x^{-1} A x. \quad (5.4)$$

Indeed, the right-hand side of (5.4) belongs to $T_x S^{n-1} = \{\xi \in \mathbb{R}^n \mid x^T \xi = 0\}$ and it holds that

$$g_x \left(2 \left(I - \frac{G_x^{-1} x x^T}{x^T G_x^{-1} x} \right) G_x^{-1} A x, \xi \right) = 2x^T A \xi = Df(x)[\xi] \quad (5.5)$$

for any $\xi \in T_x S^{n-1}$. Let R be the retraction on S^{n-1} defined by

$$R_x(\xi) = \frac{x + \xi}{\sqrt{(x + \xi)^T (x + \xi)}}, \quad \xi \in T_x S^{n-1}, \quad x \in S^{n-1}, \quad (5.6)$$

which is the special case of the QR retraction (A5) on the Stiefel manifold defined in Appendix A. For this R , the differentiated retraction $\mathcal{T}^R$ defined by (2.8) is written out as

$$\mathcal{T}_\eta^R(\xi) = \frac{1}{\sqrt{(x + \eta)^T (x + \eta)}} \left(I - \frac{(x + \eta)(x + \eta)^T}{(x + \eta)^T (x + \eta)} \right) \xi, \quad \eta, \xi \in T_x S^{n-1}, \quad x \in S^{n-1}. \quad (5.7)$$

We note that though the metric endowed with is not the standard one, the Lipschitzian condition (4.2) holds, as is mentioned in Rem. A.2 in Appendix A. Hence from Thm. 4.3, Algorithm 3.2 works well in theory.

Figs. 5.1, 5.2, and 5.3 show numerical results from applying Algorithm 3.1 to Problem 5.1 with the initial point $x_0 = (1, 1, \dots, 1)^T / 2\sqrt{5} \in S^{n-1}$ with $n = 20$.

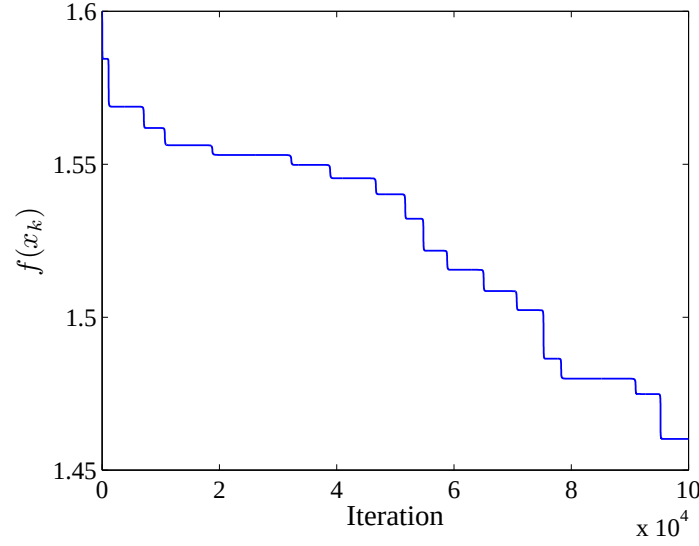


Figure 5.1. The sequence of the values $f(x_k)$ of the objective function f evaluated on the sequence $\{x_k\}$ generated by Algorithm 3.1.

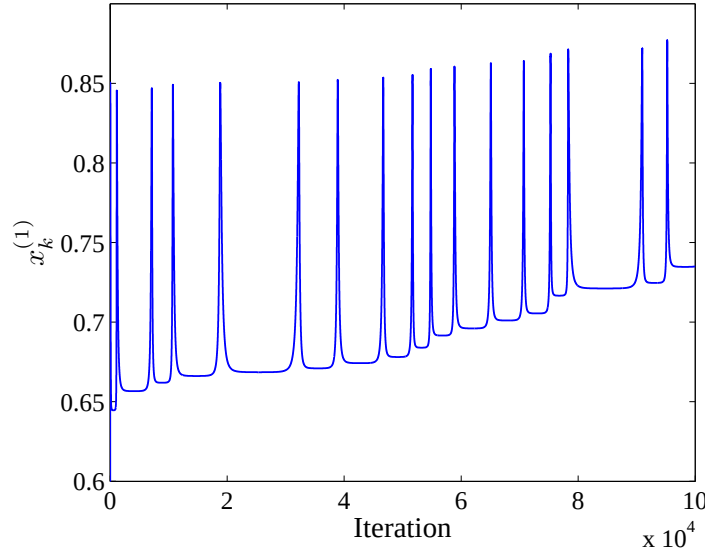


Figure 5.2. The sequence of the first components $x_k^{(1)}$ from the sequence $\{x_k\}$ generated by Algorithm 3.1.

The vertical axes of Figs. 5.1, 5.2, and 5.3 carry values of $f(x_k)$ at x_k , values of the first components $x_k^{(1)}$ of x_k , and values of the ratios $\|\mathcal{T}_{\alpha_k \eta_k}^R(\eta_k)\|_{x_{k+1}}/\|\eta_k\|_{x_k}$, respectively. Note that for the optimal solution $x_* = (1, 0, 0, \dots, 0)^T \in S^{n-1}$ which the current generated sequence $\{x_k\}$ is expected to approach, the target value is $f(x_*) = x_*^{(1)} = 1$ in both Figs. 5.1 and 5.2. Though the $\{x_k\}$ seems to come close to x_* bit by bit, the convergence is not observed even after 10^5 iterations. At the iteration number 10^5 , $f(x_k)$ is far from $f(x_*) = 1$, as is seen from Fig. 5.1. Fig. 5.2 shows that the sequence is intermittently repelled from the target point, when approaching it. If more iterations, say 10^7 , are performed, the graph of $\{x_k^{(1)}\}$ has almost the same shape, that is, sharp peaks repeatedly appear in Fig. 5.2 with

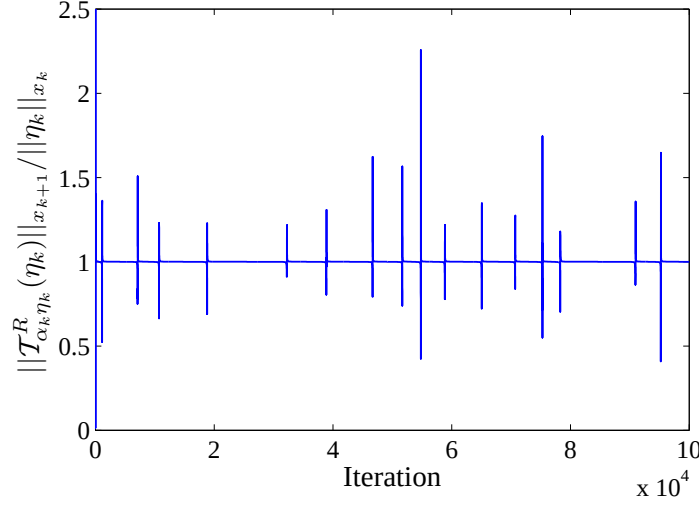


Figure 5.3. Ratios $\|\mathcal{T}_{\alpha_k \eta_k}^R(\eta_k)\|_{x_{k+1}} / \|\eta_k\|_{x_k}$ evaluated on the sequences $\{x_k\}$ and $\{\eta_k\}$ generated by Algorithm 3.1.

extended iterations. If $\|\mathcal{T}_{\alpha_k \eta_k}^R(\eta_k)\|_{x_{k+1}} / \|\eta_k\|_{x_k} \leq 1$ for all $k \in \mathbb{N}$, the sequence $\{x_k\}$ would converge. However, as is shown in Fig. 5.3, the ratio $\|\mathcal{T}_{\alpha_k \eta_k}^R(\eta_k)\|_{x_{k+1}} / \|\eta_k\|_{x_k}$ intermittently exceeds the value 1. This fact seems to prevent the sequence from converging, as long as numerical experiments suggest. To gain insight into the non-convergence problem, we put Figs. 5.2 and 5.3 together into Fig. 5.4, which shows that the peaks of two graphs synchronize. This suggests that the violation

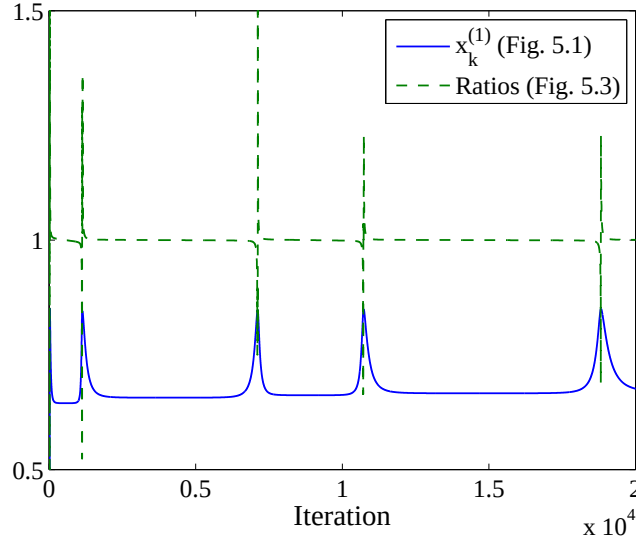


Figure 5.4. $x_k^{(1)}$ and $\|\mathcal{T}_{\alpha_k \eta_k}^R(\eta_k)\|_{x_{k+1}} / \|\eta_k\|_{x_k}$ by Algorithm 3.1.

of the inequality (3.4) makes the sequence fail to approach the optimal solution x_* . This phenomenon is caused by the large first diagonal element of G_x in the neighbourhood of x_* .

In contrast with this, in Algorithm 3.2, the vector transport $\mathcal{T}^R$ is scaled if necessary, and thereby generated sequences converge to solve Problem 5.1. In comparison with Fig. 5.2, Fig. 5.5 shows that the present algorithm generates a converging sequence, resolving the difficulty of being repelled from the optimal solution. We

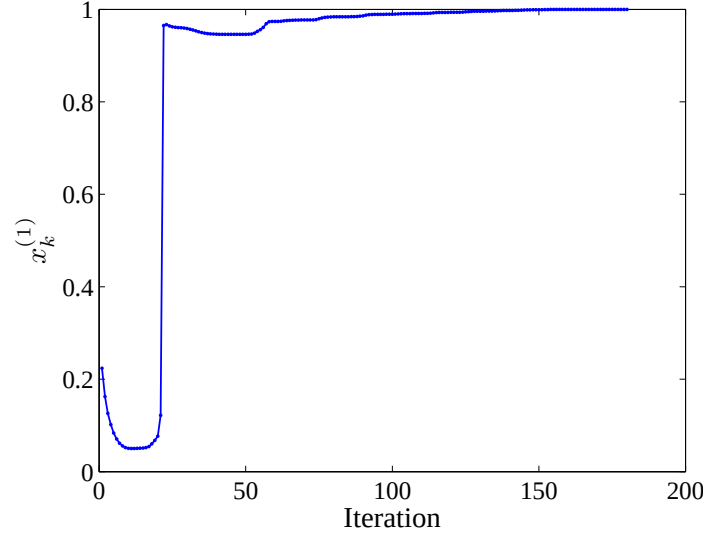


Figure 5.5. The sequence of the first components $x_k^{(1)}$ from the sequence $\{x_k\}$ generated by Algorithm 3.2.

here note that the inequality $\|\mathcal{T}_{\alpha_k \eta_k}^{(k)}(\eta_k)\|_{x_{k+1}} \leq \|\eta_k\|_{x_k}$ is never violated in this algorithm.

We now investigate the performance of Algorithm 3.2 in more detail with interest in comparison with a restart strategy in the conjugate gradient method. As is

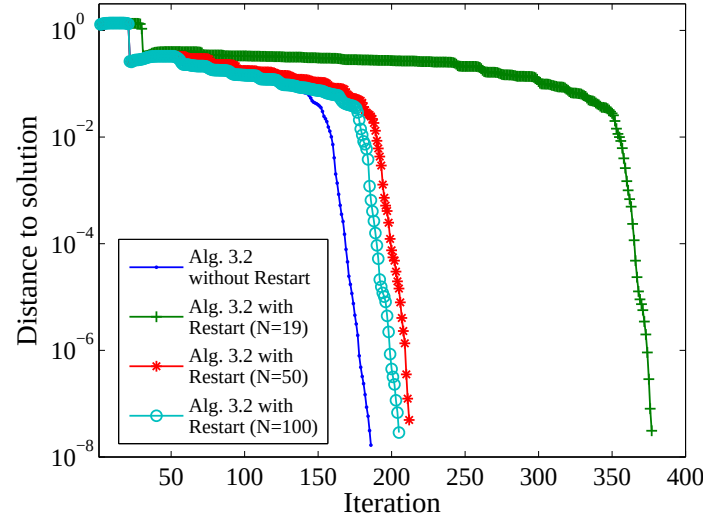


Figure 5.6. The sequences of the distances between x_k and x_* with respect to the sequences $\{x_k\}$ generated by Algorithm 3.2 with several restarting strategies.

well known, in a nonlinear conjugate gradient method on the Euclidean space, the iteration is often restarted at every N steps by taking a steepest descent search direction, where N is usually chosen to be the dimension of the search space in the problem. To gain a sight of the performance of the restart method on a Riemannian manifold, we introduce a similar restart strategy into Algorithms 3.1 and 3.2, that is, we set $\beta_{k+1} = 0$ in Step 5 of each algorithm at every N steps. A choice for N is 19, which is the dimension of S^{n-1} with $n = 20$. For comparison, the both algorithms with restarts are also performed for $N = 50$ and $N = 100$. The

results from Algorithm 3.2 with and without restart are shown in Fig. 5.6. The vertical axis of Fig. 5.6 carries $\sqrt{(x_k - x_*)^T(x_k - x_*)}$, which is an approximation of the distance between x_k and x_* on S^{n-1} . We can observe from the graphs in Fig. 5.6 that Algorithm 3.2 with and without restart has a superlinear convergence property. Fig. 5.6 shows further that Algorithm 3.2 without restart exhibits better performance than Algorithm 3.2 with a few variants of restarts, which means that the restart strategy fails to improve the performance of Algorithm 3.2.

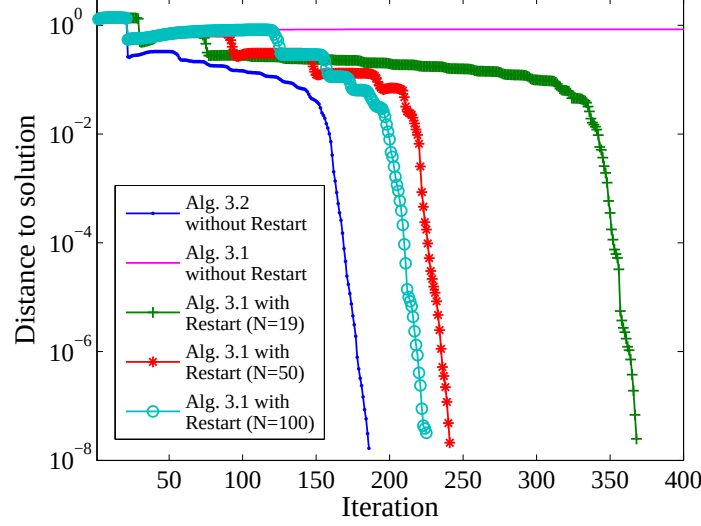


Figure 5.7. The sequences of the distances between x_k and x_* with respect to the sequences $\{x_k\}$ generated by Algorithm 3.2 and Algorithm 3.1 with several restarting strategies.

On the contrary, the restart strategy improves the performance of Algorithm 3.1, but the resultant performance is not comparable to Algorithm 3.2 without restart yet. A numerical evidence is shown in Fig. 5.7.

5.2. The sphere endowed with the orthographic retraction

We give a more natural example, in which the inequality (3.4) is never satisfied. Consider Problem 5.1 with $n = 100$ and $A = \text{diag}(1, 2, \dots, 100)/100$. The difference from the example in Subsection 5.1 is the choice of a Riemannian metric and a retraction. We in turn endow the sphere S^{n-1} with the induced metric $\langle \cdot, \cdot \rangle$ from the natural inner product on $\mathbb{R}^n$:

$$\langle \xi_x, \eta_x \rangle_x := \xi_x^T \eta_x, \quad \xi_x, \eta_x \in T_x S^{n-1}. \quad (5.8)$$

The norm of $\xi_x \in T_x S^{n-1}$ is then defined to be $\|\xi_x\|_x = \sqrt{\xi_x^T \xi_x}$ as usual. With the natural metric $\langle \cdot, \cdot \rangle$, the gradient of f is written out as

$$\text{grad } f(x) = 2(I - xx^T)Ax. \quad (5.9)$$

We consider the orthographic retraction R on S^{n-1} [2], which is defined to be

$$R_x(\xi) = \sqrt{1 - \xi^T \xi} x + \xi, \quad \xi \in T_x S^{n-1} \text{ with } \|\xi\|_x < 1. \quad (5.10)$$

Associated with this R , the vector transport $\mathcal{T}^R$ is written out as

$$\mathcal{T}_\eta^R(\xi) = \xi - \frac{\eta^T \xi}{\sqrt{1 - \eta^T \eta}} x, \quad \eta, \xi \in T_x S^{n-1} \text{ with } \|\eta\|_x, \|\xi\|_x < 1, \quad x \in S^{n-1}. \quad (5.11)$$

For this $\mathcal{T}^R$, the norm $\|\mathcal{T}_\eta^R(\xi)\|_{R_x(\eta)}$ is evaluated as

$$\|\mathcal{T}_\eta^R(\xi)\|_{R_x(\eta)}^2 = \|\xi\|_x^2 + \frac{(\eta^T \xi)^2}{1 - \|\eta\|_x^2} \geq \|\xi\|_x^2, \quad (5.12)$$

where use has been made of $x^T x = 1$ and $x^T \xi = 0$. Thus, the inequality (3.4), which is the key condition for the proof of the global convergence property of Algorithm 3.1, is violated unless $\eta_k = 0$. In spite of this fact, we may try to perform Algorithm 3.1 for this problem. If the generated sequence does not diverge, we can compare the result with that obtained by Algorithm 3.2. We performed Algorithms 3.1 and 3.2 and obtained Fig. 5.8, whose vertical axis carries $\sqrt{(x_k - x_*)^T (x_k - x_*)}$. The figure shows the superiority of the proposed algorithm.

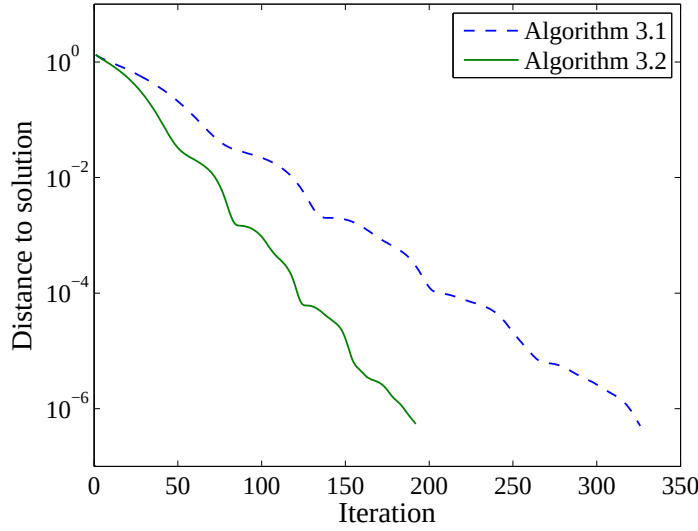


Figure 5.8. The sequences of distances between x_k and x_* for the sequences $\{x_k\}$ generated by Algorithms 3.1 and 3.2 with the orthographic retraction.

6. Concluding Remarks

We have dealt with the global convergence of the conjugate gradient method with the Fletcher-Reeves β . Though the conjugate gradient method generates globally converging sequences in the Euclidean space, the conjugate gradient method on a Riemannian manifold M has not been shown to have a convergence property in general, but under the assumption that the vector transport $\mathcal{T}^R$ as the differentiated retraction does not increase the norm of the tangent vector, the convergence is proved in [9]. If the parallel translation is adopted as a vector transport, the conjugate gradient method is shown to generate converging sequences, as is given in [11]. However, the parallel translation is not convenient for computational effectiveness. For computational efficiency, we have introduced a vector transport, in

place of the parallel translation, with a modification that the vector transport $\mathcal{T}^R$ is replaced by the scaled vector transport $\mathcal{T}^0$ only when $\mathcal{T}^R$ increases the norm of the search direction vector. The idea is simple but effective. We have achieved a balance between computational efficiency and the global convergence by proposing Algorithm 3.2. We have shown the convergence of the present algorithm both in the theoretical and the numerical viewpoints. In particular, we have performed numerical experiments to show that the present algorithm can solve problems for which the existing algorithm cannot work well because of the violation of the assumption about the vector transport.

Appendix A. Examples in which the condition (4.2) holds

In Thm. 4.1, we assume that the condition (4.2) holds. We here compare (4.2) with the condition that $f \circ R_x$ is Lipschitz continuously differentiable uniformly for x , that is, there exists a Lipschitz constant $L > 0$ such that

$$\|D(f \circ R_x)(\xi) - D(f \circ R_x)(\zeta)\| \leq L\|\xi - \zeta\|_x, \quad \xi, \zeta \in T_x M, x \in M, \quad (\text{A1})$$

where the $\|\cdot\|$ of the left-hand side means the operator norm (see [9] for detail). The condition (A1) is equivalent to

$$\sup_{\|\eta\|_x=1} |(D(f \circ R_x)(\xi) - D(f \circ R_x)(\zeta))[\eta]| \leq L\|\xi - \zeta\|_x, \quad \xi, \zeta \in T_x M, x \in M. \quad (\text{A2})$$

In particular, setting $\zeta = 0$ and $\xi = t\eta$ in (A2) yields (4.2). In this sense, the condition (4.2) is a weaker form of (A1). The assumption (4.2) is of practical use. For example, the problem of minimizing the Brockett cost function on the Stiefel manifold $\text{St}(p, n)$ with the natural induced metric [1] has this property, as is shown below.

Let n, p be positive integers with $n \geq p$. The Stiefel manifold $\text{St}(p, n)$ is defined to be $\text{St}(p, n) := \{X \in \mathbb{R}^{n \times p} \mid X^T X = I_p\}$. We consider $\text{St}(p, n)$ as a Riemannian submanifold of $\mathbb{R}^{n \times p}$ endowed with the natural induced metric

$$\langle \xi, \eta \rangle_X := \text{tr}(\xi^T \eta), \quad \xi, \eta \in T_X \text{St}(p, n). \quad (\text{A3})$$

Let A be an $n \times n$ symmetric matrix and $N := \text{diag}(\mu_1, \mu_2, \dots, \mu_p)$ with $0 < \mu_1 < \mu_2 < \dots < \mu_p$. The Brockett cost function f is defined on $\text{St}(p, n)$ to be

$$f(X) = \text{tr}(X^T A X N). \quad (\text{A4})$$

Further, the QR decomposition-based retraction (which we call the QR retraction) R is defined to be

$$R_X(\xi) := \text{qf}(X + \xi), \quad \xi \in T_X \text{St}(p, n), X \in \text{St}(p, n), \quad (\text{A5})$$

where $\text{qf}(B)$ denotes the Q-factor of the QR decomposition of a full rank matrix $B \in \mathbb{R}^{n \times p}$. That is, if B is decomposed into $B = QR$, where $Q \in \text{St}(p, n)$ and R is an upper triangular $p \times p$ matrix with positive diagonal elements, then $\text{qf}(B) = Q$.

Proposition A.1 The inequality (4.2) holds for the Brockett cost function (A4) on $M = \text{St}(p, n)$, where $\text{St}(p, n)$ is endowed with the natural induced metric (A3), and where the QR retraction (A5) is adopted.

Proof: Since the function (A4) is smooth, we have only to show that

$$\left| \frac{d^2}{dt^2} (f \circ R_X)(t\eta) \right| \leq L, \quad \eta \in T_X \text{St}(p, n) \text{ with } \|\eta\|_X = 1, \quad X \in \text{St}(p, n), \quad t \geq 0. \quad (\text{A6})$$

In fact, Eq. (4.2) is a straightforward consequence of this inequality. Let $Q(t)$ be a curve defined by $R_X(t\eta) = \text{qf}(X + t\eta)$, and $x_k, \eta_k, q_k(t)$ denote the k -th column vectors of $X, \eta, Q(t)$, respectively. Then, through the Gram-Schmidt orthonormalization process, we obtain

$$q_k(t) = \frac{x_k + t\eta_k - \sum_{i=1}^{k-1} (q_i(t), x_k + t\eta_k) q_i(t)}{\|x_k + t\eta_k - \sum_{i=1}^{k-1} (q_i(t), x_k + t\eta_k) q_i(t)\|}, \quad (\text{A7})$$

where $(a, b) := a^T b$ and $\|a\| := \sqrt{(a, a)}$ for n -dimensional vectors a, b . By induction on k , we can take vector-valued polynomials $g_k(t)$ in t satisfying

$$q_k(t) = \frac{g_k(t)}{\|g_k(t)\|}, \quad t \geq 0. \quad (\text{A8})$$

Indeed, for $k = 1$, (A8) holds with $g_1(t) = x_1 + t\eta_1$. Suppose that (A8) holds for $1, \dots, k-1$. Then we can write out $q_k(t)$ as

$$q_k(t) = \frac{\prod_{j=1}^{k-1} \|g_j(t)\|^2 (x_k + t\eta_k) - \sum_{i=1}^{k-1} \prod_{j \neq i} \|g_j(t)\|^2 (g_i(t), x_k + t\eta_k) g_i(t)}{\prod_{j=1}^{k-1} \|g_j(t)\|^2 (x_k + t\eta_k) - \sum_{i=1}^{k-1} \prod_{j \neq i} \|g_j(t)\|^2 (g_i(t), x_k + t\eta_k) g_i(t)}. \quad (\text{A9})$$

Denoting by $g_k(t)$ the numerator of the right-hand side of (A9), which is a polynomial in t , we obtain (A8).

Let

$$h(X, \eta, t) = \frac{d^2}{dt^2} (f \circ R_X)(t\eta). \quad (\text{A10})$$

Then, the $h(X, \eta, t)$ is written out as

$$h(X, \eta, t) = \sum_{k=1}^p \mu_k \frac{d^2}{dt^2} (q_k(t)^T A q_k(t)). \quad (\text{A11})$$

Since $q_k(t)^T A q_k(t) = g_k(t)^T A g_k(t) / \|g_k(t)\|^2$, and since the degree of the numerator polynomial in t is not more than that of the denominator polynomial, the degree of the numerator polynomial from the right-hand side of (A11) is less than that of the denominator polynomial, so that one has, as $t \rightarrow \infty$,

$$\lim_{t \rightarrow \infty} h(X, \eta, t) = 0. \quad (\text{A12})$$

This implies that $h(X, \eta, t)$ is bounded with respect to $t \geq 0$. Moreover, the $h(X, \eta, t)$ is continuous with respect to X and η on the compact set $\{(X, \eta) \in T \text{St}(p, n) \mid \|\eta\|_X = 1\}$. It then turns out that $h(X, \eta, t)$ is bounded on the whole domain, which implies that there exists $L > 0$ such that (A6) holds. This completes the proof. $\square$

Remark A.1 Reviewing the proof, we observe that since the QR retraction is ir-respective of the metric with which the $\text{St}(p, n)$ is endowed, and since the set

$\{(X, \eta) \in T\text{St}(p, n) \mid \|\eta\|_X = 1\}$ is compact with respect to any metric on $\text{St}(p, n)$, the inequality (4.2) with R being the QR retraction (A5) holds for the Brockett cost function (A4) independently of the choice of a metric.

Remark A.2 We also note that Prop. A.1 and Rem. A.1 cover both the Rayleigh quotient on the sphere S^{n-1} as $p = 1$ and the Brockett cost function on the orthogonal group as $p = n$. In particular, the inequality (4.2) holds for the function (5.1), though the sphere S^{n-1} is endowed with the non-standard metric (5.3).

Another example for (4.2) comes from the problem of minimizing the function

$$F(U, V) = \text{tr}(U^T A V N) \quad (\text{A13})$$

on $\text{St}(p, m) \times \text{St}(p, n)$, where A is an $m \times n$ matrix and $N = \text{diag}(\mu_1, \dots, \mu_p)$ with $\mu_1 > \dots > \mu_p > 0$. An optimal solution to this problem gives the singular value decomposition of A [10]. Let m, n, p be positive integers with $m \geq n \geq p$. We consider $\text{St}(p, m) \times \text{St}(p, n)$ as a Riemannian submanifold of $\mathbb{R}^{m \times p} \times \mathbb{R}^{n \times p}$ endowed with the natural induced metric;

$$\begin{aligned} \langle (\xi_1, \eta_1), (\xi_2, \eta_2) \rangle_{(U, V)} &:= \text{tr}(\xi_1^T \xi_2) + \text{tr}(\eta_1^T \eta_2), \\ (\xi_1, \eta_1), (\xi_2, \eta_2) &\in T_{(U, V)}(\text{St}(p, m) \times \text{St}(p, n)). \end{aligned} \quad (\text{A14})$$

As in the previous example on $\text{St}(p, n)$, the QR retraction on $\text{St}(p, m) \times \text{St}(p, n)$ is defined by

$$R_{(U, V)}(\xi, \eta) := (\text{qf}(U + \xi), \text{qf}(V + \eta)), \quad (\xi, \eta) \in T_{(U, V)}(\text{St}(p, m) \times \text{St}(p, n)) \quad (\text{A15})$$

for $(U, V) \in \text{St}(p, m) \times \text{St}(p, n)$.

Proposition A.2 The inequality (4.2) holds for the objective function (A13) on $M = \text{St}(p, m) \times \text{St}(p, n)$, where M is endowed with the natural induced metric (A14) and with the QR retraction (A15).

Proof: We shall show that

$$\left| \frac{d^2}{dt^2} (F \circ R_{(U, V)}) (t(\xi, \eta)) \right| \leq L \quad (\text{A16})$$

for $(\xi, \eta) \in T_{(U, V)}(\text{St}(p, m) \times \text{St}(p, n))$ with $\|(\xi, \eta)\|_{(U, V)} = 1$, $(U, V) \in \text{St}(p, m) \times \text{St}(p, n)$, $t \geq 0$. Put $Q(t) = \text{qf}(U + t\xi)$, $S(t) = \text{qf}(V + t\eta)$. Let $q_k(t)$ and $s_k(t)$ denote the k -th column vectors of $Q(t)$ and $S(t)$, respectively. From Prop. A.1 and its course of the proof, there exist vector-valued polynomials $g_k(t)$ and $h_k(t)$ such that

$$q_k(t) = \frac{g_k(t)}{\|g_k(t)\|}, \quad s_k(t) = \frac{h_k(t)}{\|h_k(t)\|}. \quad (\text{A17})$$

Let

$$H(U, V, \xi, \eta, t) = \frac{d^2}{dt^2} (F \circ R_{(U, V)}) (t(\xi, \eta)). \quad (\text{A18})$$

Then we have

$$H(U, V, \xi, \eta, t) = \sum_{k=1}^p \mu_k \frac{d^2}{dt^2} (q_k(t)^T A s_k(t)). \quad (\text{A19})$$

Since $q_k(t)^T A s_k(t) = g_k(t)^T A h_k(t) / (\|g_k(t)\| \|h_k(t)\|)$, by the same reasoning as that for $h(X, \xi, t)$ in Prop. A.1, we have

$$\lim_{t \rightarrow \infty} H(U, V, \xi, \eta, t) = 0, \quad (\text{A20})$$

so that $H(U, V, \xi, \eta, t)$ is bounded with respect to $t \geq 0$. Further, $H(U, V, \xi, \eta, t)$ is continuous with respect to (U, V, ξ, η) on the compact set $\{(U, V, \xi, \eta) \in T(\text{St}(p, m) \times \text{St}(p, n)) \mid \|(\xi, \eta)\|_{(U, V)} = 1\}$. Hence $H(U, V, \xi, \eta, t)$ is bounded on the whole domain. This completes the proof. $\square$

A remark similar to Rem. A.1 can be made on the metric to be endowed with on $\text{St}(p, m) \times \text{St}(p, n)$. The validity of (4.2) is independent of the choice of a metric.

Returning to the case of a general Riemannian manifold M , we make a further comment on (4.2). We are interested in the range of $t \geq 0$. Assume that M is compact and f is smooth. A smooth function on a compact set is Lipschitz continuously differentiable. However, the set $\{(x, \eta, t) \in TM \times \mathbb{R} \mid \|\eta\|_x = 1, t \geq 0\}$ is not compact even though M is compact. Therefore, it is not so clear that the inequality (4.2) holds in general. We here note that the inequality (4.2) is used in the form

$$D(f \circ R_{x_k})(\alpha_k \eta_k)[\eta_k] - D(f \circ R_{x_k})(0)[\eta_k] \leq \alpha_k L \|\eta_k\|_{x_k}^2 \quad (\text{A21})$$

for the proof of Thm. 4.1. A question then arises as to under what condition the inequality (A21) holds. If it is ensured that there exists a constant $m > 0$ such that $\alpha_k \|\eta_k\|_{x_k} \leq m$ for all k , then we can prove (A21). Indeed, in order to prove (A21) in such a case, the range of t in (4.2) can be restricted to $0 \leq t \leq m$, and the inequality we need to prove as a counterpart to (4.2) is written as

$$|D(f \circ R_x)(t\eta)[\eta] - D(f \circ R_x)(0)[\eta]| \leq Lt, \quad \eta \in T_x M \text{ with } \|\eta\|_x = 1, x \in M, 0 \leq t \leq m. \quad (\text{A22})$$

In order that (A22) hold, it is sufficient that there exists a constant $L > 0$ satisfying

$$\left| \frac{d^2}{dt^2} (f \circ R_x)(t\eta) \right| \leq L, \quad \eta \in T_x M \text{ with } \|\eta\|_x = 1, x \in M, 0 \leq t \leq m. \quad (\text{A23})$$

Since the left-hand side of the inequality (A23) is continuous with respect to t on a compact set $\{t \in \mathbb{R} \mid 0 \leq t \leq m\}$, there exists $L_{x, \eta}$ for each $(x, \eta) \in \mathcal{M}$ such that (A23) with $L = L_{x, \eta}$ holds, where $\mathcal{M} = \{(x, \eta) \in TM \mid \|\eta\|_x = 1\}$. The compactness of the set $\mathcal{M}$ ensures the existence of $L := \sup_{(x, \eta) \in \mathcal{M}} L_{x, \eta}$ and the L thus defined satisfies (A23).

Acknowledgements

The authors would like to thank the anonymous referees for providing them with valuable comments that helped them to significantly brush up the paper. The first author appreciates the JSPS Research Fellowship for Young Scientists.

References

- [1] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.
- [2] P.-A. Absil and J. Malick, Projection-like retractions on matrix manifolds, *SIAM J. Optim.*, **22**(1):135–158, 2012.
- [3] M. Al-Baali, Descent property and global convergence of the Fletcher-Reeves method with inexact line search, *I.M.A. Journal on Numerical Analysis*, **5**:121–124, 1985.
- [4] Alan Edelman, Tomás A. Arias, and Steven T. Smith, The geometry of algorithms with orthogonality constraints, *SIAM J. Matrix Anal. Appl.*, **20**(2):303–353, 1998.
- [5] R. Fletcher and C. M. Reeves, Function minimization by conjugate gradients, *Comput. J.*, **7**:149–154, 1964.
- [6] U. Helmke and J. B. Moore, *Optimization and Dynamical Systems*. Comm. Control Engrg. Ser., Springer, Berlin, 1994.
- [7] M. R. Hestenes and E. Stiefel, Methods of conjugate gradients for solving linear systems, *J. Res. Nat. Bur. Stand.*, **49**:409–436 (1953), 1952.
- [8] J. Nocedal and S.J. Wright, *Numerical Optimization*, 2nd ed., Springer, New York, 2006.
- [9] W. Ring and B. Wirth, Optimization methods on Riemannian manifolds and their application to shape space, *SIAM J. Optim.*, **22**(2):596–627, 2012.
- [10] H. Sato and T. Iwai, A Riemannian optimization approach to the matrix singular value decomposition, *SIAM J. Optim.*, **23**(1):188–212, 2013.
- [11] S. T. Smith, Optimization techniques on Riemannian manifolds. In *Hamiltonian and gradient flows, algorithms and control*, *Fields Inst. Commun.*, **3**:113–136. Amerl Math. Soc., Providence, RI, 1994.