Global Convergence of Hager–Zhang type Riemannian Conjugate Gradient Method

Hiroyuki Sakai, Hiroyuki Sato, and Hideaki Iiduka

Abstract

This paper presents the Hager–Zhang (HZ)-type Riemannian conjugate gradient method that uses the exponential retraction. We also present global convergence analyses of our proposed method under two kinds of assumptions. Moreover, we numerically compare our proposed methods with the existing methods by solving two kinds of Riemannian optimization problems on the unit sphere. The numerical results show that our proposed method has much better performance than the existing methods, i.e., the FR, DY, PRP and HS methods. In particular, they show that it has much higher performance than existing methods including the hybrid ones in computing the stability number of graphs problem.

1 Introduction

Riemannian optimization has been widely researched along with the developments of real-world applications in various fields, such as natural language processing [8, 11], signal processing [19], and computer vision [5, 6], in which large-scale problems can be expressed as certain optimization problems on Riemannian manifolds.

Many useful gradient methods [1, 16] have been developed for Riemannian optimization that can be obtained by extending the existing methods in Euclidean space to a Riemannian manifold. However, such extension is not always easy. For example, in the Euclidean space setting, the $(k+1)$-th approximation of optimal solutions is $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{\eta}_k$, where $\alpha_k > 0$, $\boldsymbol{x}_k \in \mathbb{R}^n$ is a point at the k-th iteration, and $\boldsymbol{\eta}_k$ is the search direction. However, such an update cannot be defined for general Riemannian manifolds because of nonlinear Riemannian geometric structure. We can generalize Riemannian gradient methods using *retractions* and *transports* that are divided into various types, as described below.

Smith [18] proposed using *exponential retraction* and *parallel transport* to generalize the optimization methods from Euclidean space to a Riemannian manifold. Absil, Mahony, and Sepulchre [1] proposed using a *general retraction* that approximates the exponential retraction and a *vector transport* which approximates the parallel transport. Note that a general retraction (resp. vector transport) is a generalization of the exponential retraction (resp. parallel

1

transport).

We focus on *Riemannian conjugate gradient (RCG) methods* as they offer both theoretical and practical benefits. A theoretical benefit of RCG methods is that we can show that they generate *sufficient descent search directions*, which decrease an objective function at every iteration, and converge *globally*, i.e., without depending on the choice of the initial point. A practical benefit of RCG methods is that they have efficient numerical performances, as shown in the previous studies [1, 16].

1.1   Previous results

The results for RCG methods that satisfy the sufficient descent condition and global convergence are summarized as in Table 1.

Ring and Wirth [9] presented a Fletcher–Reeves (FR) type of RCG method using a general retraction and vector transport, which is defined by the differentiated retraction, under the strong Wolfe conditions. The vector transport they used in [9] is assumed not to increase the norm of the search direction vector, which would be unnatural in both theory and practice. To overcome this limitations, Sato and Iwai [17] defined a *scaled vector transport* and showed convergence of the FR-type RCG method using a general retraction and scaled vector transport.

Sato [14] also investigated a Dai–Yuan (DY) type of RCG method using a general retraction and scaled vector transport and showed that it generates a sufficient descent direction and converges globally under the Wolfe conditions ("DY" row in Table 1). Comparison of the results in [17] with those in [14] reveals that the DY-type RCG method has a better global convergence than the FR-type one because it is based on the assumption of the Wolfe conditions, which are weaker than the strong Wolfe conditions.

A recently introduced hybrid RCG method [10] is defined by combining the good global convergence of the DY-type RCG method (see description above) with the efficient numerical performance of a Hestenes–Stiefel (HS) type of RCG method. This hybrid method generates a sufficient descent direction and converges globally under the strong Wolfe conditions ("HS-DY hybrid" row in Table 1). Another recently introduced hybrid method [12] combines the FR-tysspe RCG method with a Polak–Ribière–Polyak (PRP) type of RCG method. This hybrid method also generates a sufficient descent direction and converges globally under the strong Wolfe conditions ("FR-PRP hybrid" row in Table 1).

1.2   Goals

As described in Section 1.1, and shown in Table 1, existing RCG methods are capable for solving Riemannian optimization problems. Nevertheless, there are other powerful conjugate gradient methods in Euclidean space that could be generalized to Riemannian manifolds. A particularly interesting Euclidean con-

jugate gradient (ECG) method is the Hager–Zhang (HZ) type[1] [3] of conjugate gradient method, which is a very efficient conjugate gradient method for Euclidean optimization. Accordingly, the first goal of this paper is to clarify whether or not the HZ-type ECG method can be theoretically extended to a Riemannian manifold so as to guarantee its global convergence. Sakai and Iiduka [12] showed that the HZ-type RCG method using a general retraction and scaled vector transport generates a sufficient descent direction ("HZ" row in Table 1). This sufficient descent property does not depend on the line search conditions. However, the global convergence of the HZ-type RCG method has not been determined.

The second goal is to determine whether that the HZ-type RCG method performs better than the existing RCG methods listed in Table 1. The HZ-type ECG method tends to perform better in the Euclidean space setting than other ECG methods. Therefore, it would be useful to know whether the HZ-type RCG method has the same performance as the HZ-type ECG method.

1.3   Contributions

This paper makes two contributions. The first contribution is to show that the HZ-type RCG method using the *exponential retraction* and vector transport converges globally under the Wolfe conditions (Theorem 3.3). This contribution is an extension of Theorem 2.2 in [3] to a Riemannian manifold and shows theoretically for the first time the global convergence of the HZ-type RCG method. The second contribution is to provide numerical comparisons of the HZ-type RCG method with the existing RCG methods. The numerical results of this paper indicate that the HZ-type RCG method performs better than the existing ones in computing the stability number of graphs problem.

1.4   Difficulty to prove Theorem 3.3

A way to guarantee the global convergence property of the HZ-type ECG method in the Euclidean space is to assume the strong convexity of the objective function f. We thus assume that there exists a constant $\mu > 0$ such that

$$(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^{\top}(\boldsymbol{x} - \boldsymbol{y}) \geq \mu \left\| \boldsymbol{x} - \boldsymbol{y} \right\|^{2} \tag{1}$$

holds for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{n}$, which is equivalent to the condition that the smallest eigenvalue of the Hessian $\nabla^{2} f(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^{n}$ is not less than μ. Expression (1) is more useful in convergence analysis.

However, (1) cannot be directly generalized to the Riemannian case. Instead, a natural definition of the strong convexity of f on a Riemannian manifold M is that there exists a constant $\mu > 0$ such that, for any $x \in M$, the smallest eigenvalue of the Riemannian Hessian Hess $f(x)$ is not less than μ. In Theorem 3.3, we have to start with this condition and without a Riemannian counterpart of (1).

---

[1]http://users.clas.ufl.edu/hager/papers/Software/

3

Noting that (1) is used to prove $(\nabla f(\boldsymbol{x}_{k+1}) - \nabla f(\boldsymbol{x}_k))^\top (\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) \geq \mu \alpha_k \|\boldsymbol{\eta}_k\|^2$ in Euclidean space, we need to show the Riemannian counterpart of this inequality, not that of (1). Fortunately, by imposing the Wolfe conditions on the step length, we can directly prove the desired inequality (27) from the assumption of the strong convexity of f, i.e., the condition that the eigenvalue of the Riemannian Hessian is uniformly lower bounded.

The remainder of this paper is organized as follows. Section 2 gives the mathematical preliminaries, including descriptions of retraction, vector transport, and existing RCG methods. Section 3 presents our results for the HZ-type RCG method. Section 4 provides numerical comparisons. Section 5 briefly summarizes the key points.

Table 1: RCG results of previous studies and our results

	Riemannian conjugate gradient methods			
	Exponential retraction and its differentiation (as vector transport)		General retraction and its (scaled) differentiation (as vector transport)	
	Sufficient descent condition	Global convergence	Sufficient descent condition	Global convergence
FR	Ring–Wirth (2012) [9] Sato–Iwai (2015) [17] (strong Wolfe conditions)	Ring–Wirth (2012) [9] Sato–Iwai (2015) [17] (strong Wolfe conditions)	Ring–Wirth (2012) [9] Sato–Iwai (2015) [17] (strong Wolfe conditions)	Ring–Wirth (2012) [9] Sato–Iwai (2015) [17] (strong Wolfe conditions)
DY	Sato (2016) [14] (Wolfe conditions)	Sato (2016) [14] (Wolfe conditions)	Sato (2016) [14] (Wolfe conditions)	Sato (2016) [14] (Wolfe conditions)
Hybrid (HS–DY)	Sakai–Iiduka (2020) [10] (strong Wolfe conditions)	Sakai–Iiduka (2020) [10] (strong Wolfe conditions)	Sakai–Iiduka (2020) [10] (strong Wolfe conditions)	Sakai–Iiduka (2020) [10] (strong Wolfe conditions)
Hybrid (FR–PRP)	Sakai–Iiduka (2021) [12] (strong Wolfe conditions)	Sakai–Iiduka (2021) [12] (strong Wolfe conditions)	Sakai–Iiduka (2021) [12] (strong Wolfe conditions)	Sakai–Iiduka (2021) [12] (strong Wolfe conditions)
HZ	Sakai–Iiduka (2021) [12] (without conditions)	**this work** (Wolfe conditions)	Sakai–Iiduka (2021) [12] (without conditions)	——

See Section 2 for definitions of retraction, vector transport, FR, DY, HS-DY hybrid, and FR-PRP hybrid, and (strong) Wolfe conditions and Section 3 for definition of HZ-type RCG method.

2 Mathematical Preliminaries

2.1 Notation, definitions, and lemma

Let $(M, \langle \cdot, \cdot \rangle)$ be a connected geodesically complete Riemannian manifold, where $\langle \cdot, \cdot \rangle_x : T_x M \times T_x M \to \mathbb{R}$ is a Riemannian metric at a point $x \in M$. Here, $T_x M$ is a tangent space at a point $x \in M$, and TM is a tangent bundle of M; i.e., $TM := \bigcup_{x \in M} T_x M$. Let $\exp_x : T_x M \to M$ be the exponential map at $x \in M$ and $\oplus$ be the Whitney sum defined as follows (see [13, Subchapter I.3 (p.16 (II))]):

$$TM \oplus TM := \{(\xi, \eta) : \xi, \eta \in T_x M, x \in M\}.$$

An unconstrained optimization problem on M is expressed as follows (see [1, 10, 12, 14, 17]):

Problem 2.1. *Let $f : M \to \mathbb{R}$ be smooth. Then, we would like to*

$$\text{minimize } f(x) \text{ subject to } x \in M.$$

To generalize line search optimization algorithms to Riemannian manifolds, the notions of a retraction and a vector transport are used.

Definition 2.1 (Retraction). *A retraction (see [1, Chapter 4, Definition 4.1.1]) is a smooth map $R : TM \to M$ that has the following properties.*

- $R_x(0_x) = x$;

- *With the canonical identification $T_{0_x} T_x M \simeq T_x M$, R_x satisfies*

$$(dR_x)_{0_x}(\xi) = \xi$$

  for all $\xi \in T_x M$,

where 0_x denotes the zero element of $T_x M$ and R_x denotes the restriction of R to $T_x M$.

Definition 2.2 (Vector transport). *A vector transport (see [1, Chapter 8, Definition 8.1.1]) is a smooth map $\mathcal{T} : TM \oplus TM \to TM$ that has the following properties.*

- *There exists a retraction R, called the retraction associated with $\mathcal{T}$, such that $\mathcal{T}_\eta(\xi) \in T_{R_x(\eta)} M$ for all $x \in M$ and for all $\eta, \xi \in T_x M$;*

- $\mathcal{T}_{0_x}(\xi) = \xi$ *for all $\xi \in T_x M$;*

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

Lemma 2.1 (The Gauss lemma [13]). *For any point $p \in M$, any $X \in T_p M$ and any $Y \in T_X(T_p M) \simeq T_p M$,*

$$\left\langle (d\exp_p)_X(X), (d\exp_p)_X(Y) \right\rangle_{\exp_p(X)} = \langle X, Y \rangle_p$$

2.2 Existing RCG Methods and Wolfe conditions

The RCG method $[1, 10, 12, 14, 17]$ is described as

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

$$\eta_k = \begin{cases} -g_k & k = 0, \\ -g_k + \beta_k \mathcal{T}_{\alpha_{k-1}\eta_{k-1}}(\eta_{k-1}) & k \geq 1, \end{cases} \tag{3}$$

where g_k is the Riemannian gradient of f at x_k, denoted by $\operatorname{grad} f(x_k)$, $\alpha_k > 0$ is the positive step size, and $\beta_{k+1} \in \mathbb{R}$ is a parameter chosen suitably. The $\beta_{k+1} \in \mathbb{R}$ parameters used in existing RCG methods are

$$\beta_{k+1}^{\mathrm{FR}} = \frac{\|g_{k+1}\|_{x_{k+1}}^2}{\|g_k\|_{x_k}^2}, \tag{4}$$

$$\beta_{k+1}^{\mathrm{PRP}} = \frac{\langle g_{k+1}, y_k \rangle_{x_{k+1}}}{\|g_k\|_{x_k}^2}, \tag{5}$$

$$\beta_{k+1}^{\mathrm{HS}} = \frac{\langle g_{k+1}, y_k \rangle_{x_{k+1}}}{\langle g_{k+1}, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}} - \langle g_k, \eta_k \rangle_{x_k}}, \tag{6}$$

$$\beta_{k+1}^{\mathrm{DY}} = \frac{\|g_{k+1}\|_{x_{k+1}}^2}{\langle g_{k+1}, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}} - \langle g_k, \eta_k \rangle_{x_k}}, \tag{7}$$

$$\beta_{k+1}^{\mathrm{Hybrid(HS-DY)}} = \max \left\{ 0, \min \left\{ \beta_{k+1}^{\mathrm{HS}}, \beta_{k+1}^{\mathrm{DY}} \right\} \right\}, \tag{8}$$

$$\beta_{k+1}^{\mathrm{Hybrid(FR-PRP)}} = \max \left\{ 0, \min \left\{ \beta_{k+1}^{\mathrm{FR}}, \beta_{k+1}^{\mathrm{PRP}} \right\} \right\}, \tag{9}$$

where $y_k := g_{k+1} - \mathcal{T}_{\alpha_k \eta_k}(g_k)$. To determine step size α_k in (2), we use line searches that satisfy the *Wolfe conditions* (see $[10, 12, 14, 17]$),

$$f(R_{x_k}(\alpha_k \eta_k)) \leq f(x_k) + c_1 \alpha_k \langle g_k, \eta_k \rangle_{x_k}, \tag{10}$$

$$\langle g_{k+1}, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}} \geq c_2 \langle g_k, \eta_k \rangle_{x_k}, \tag{11}$$

where $0 < c_1 < c_2 < 1$. When (11) is replaced with

$$\left| \langle g_{k+1}, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}} \right| \leq c_2 \left| \langle g_k, \eta_k \rangle_{x_k} \right|, \tag{12}$$

(10) and (12) are called *strong Wolfe conditions*.

Search direction η_k defined by (3) is said to be a *sufficient descent direction* if there exists $\kappa > 0$ such that, for all $k = 0, 1, \ldots$,

$$\langle g_k, \eta_k \rangle \leq -\kappa \|g_k\|_{x_k}^2.$$

Let us first consider the FR-type RCG method, i.e., the RCG method (2) and (3), using a general retraction and scaled vector transport, with (4). It is guaranteed to generate a sufficient descent direction and to converge globally under strong Wolfe conditions $[(10)$ and $(12)], [9, 17]$ (see also Table 1)

Next, let us consider the DY-type RCG method, i.e., the RCG method (2) and (3), using a general retraction and scaled vector transport, with (7). It is guaranteed to generate a sufficient descent direction and to converge globally under Wolfe conditions [(10) and (11)] [14] (see also Table 1). A hybrid method using either (8) or (9) also generates a sufficient descent direction and converges globally [10, 12] (see also Table 1).

3 HZ-type RCG Method

3.1 Assumptions

The parameter β_{k+1} used in the HZ-type RCG method [10, 12] is defined by

$$\beta_{k+1}^{\mathrm{HZ}} = \beta_{k+1}^{\mathrm{HS}} - \mu \frac{\|y_k\|_{x_{k+1}}^2 \langle g_{k+1}, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}}}{\left(\langle g_{k+1}, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}} - \langle g_k, \eta_k \rangle_{x_k} \right)^2}, \tag{13}$$

where $\mu > 1/4$ and $y_k := g_{k+1} - \mathcal{T}_{\alpha_k \eta_k}(g_k)$.

In this paper, we use the exponential map as a retraction, i.e., $R := \exp$. Moreover, we use the vector transport defined by the differential of the exponential retraction; i.e.,

$$\mathcal{T} : TM \oplus TM \to TM : (\eta, \xi) \mapsto \mathcal{T}_\eta(\xi) := (d \exp_x)_\eta(\xi),$$

for $\eta, \xi \in T_x M$. From the Gauss lemma (Lemma 2.1), we have

$$\langle g_k, \eta_k \rangle_{x_k} = \langle \mathcal{T}_{\alpha_k \eta_k}(g_k), \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}}. \tag{14}$$

This means that (13) and (6) can be written as

$$\beta_{k+1}^{\mathrm{HZ}} = \beta_{k+1}^{\mathrm{HS}} - \mu \frac{\|y_k\|_{x_{k+1}}^2 \langle g_{k+1}, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}}}{\langle y_k, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}}^2}, \tag{15}$$

$$\beta_{k+1}^{\mathrm{HS}} = \frac{\langle g_{k+1}, y_k \rangle_{x_{k+1}}}{\langle y_k, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}}}, \tag{16}$$

respectively. Therefore, the HZ-type RCG method with the exponential retraction can be described as Algorithm 3.1. In addition, we also consider the modified HZ method (see [4, (1.6)]), by replacing $\beta_{k+1}^{\mathrm{HZ}}$ in step 6 of Algorithm 3.1 by

$$\hat{\beta}_{k+1}^{\mathrm{HZ}} := \max\{\beta_{k+1}^{\mathrm{HZ}}, \zeta_{k+1}\}, \quad \zeta_{k+1} := -\frac{1}{\|\eta_{k+1}\|_{x_{k+1}} \min\{\zeta, \|g_{k+1}\|_{x_{k+1}}\}}, \tag{17}$$

where $\zeta > 0$ is a constant.

Algorithm 3.1 HZ-type RCG method with exponential retraction for solving Problem 2.1 [1, 10, 12]

Input: Initial point $x_0 \in M$, convergence tolerance $\epsilon > 0$.
Output: Sequence $\{x_k\}_{k=0,1,\cdots} \subset M$.
1: Set $\eta_0 = -g_0 := -\operatorname{grad} f(x_0)$.
2: $k \leftarrow 0$.
3: **while** $\|g_k\|_{x_k} > \epsilon$ **do**
4:    Compute $\alpha_k > 0$ satisfying Wolfe conditions (10) and (11).
5:    Set

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

6:    Compute $g_{k+1} := -\operatorname{grad} f(x_{k+1})$ and β_{k+1} as (15) and set search direction

$$\eta_{k+1} = -g_{k+1} + \beta_{k+1}(d\exp_{x_k})_{\alpha_k \eta_k}(\eta_k).$$

7:    $k \leftarrow k+1$.
8: **end while**

We also consider the modified HZ method (see [4, (1.6)]) by replacing $\beta_{k+1}^{\mathrm{HZ}}$ in step 6 of Algorithm 3.1 with

$$\hat{\beta}_{k+1}^{\mathrm{HZ}} := \max\{\beta_{k+1}^{\mathrm{HZ}}, \zeta_{k+1}\}, \quad \zeta_{k+1} := -\frac{1}{\|\eta_{k+1}\|_{x_{k+1}} \min\{\zeta, \|g_{k+1}\|_{x_{k+1}}\}}, \quad (18)$$

where $\zeta > 0$ is a constant.

We consider Algorithm 3.1 under Assumption 3.1 (see [9, Theorem 2]) and Assumption 3.2 described below.

Assumption 3.1. *The objective function $f : M \to \mathbb{R}$ in Problem 2.1 is smooth and bounded below, and $f \circ \exp_{x_k} : T_{x_k} M \to \mathbb{R}$ is Lipschitz continuously differentiable on* $\operatorname{span}\{\eta_k\}$ *with uniform Lipschitz constant $L > 0$.*

The following is Zoutendijk's theorem (Theorem 3.1) for Riemannian manifolds under Assumption 3.1.

Theorem 3.1 (Zoutendijk). *Let $\{x_k\}_{k=0,1,\cdots} \subset M$ be a sequence generated by Algorithm 3.1. Suppose that Assumption 3.1 holds. If each step size $\alpha_k > 0$ satisfies Wolfe conditions (10) and (11), then*

$$\sum_{k=0}^{\infty} \frac{\langle g_k, \eta_k \rangle_{x_k}^2}{\|\eta_k\|_{x_k}^2} < \infty. \quad (19)$$

Assumption 3.2. *The objective function $f : M \to \mathbb{R}$ in Problem 2.1 is smooth, and there exists a constant $L > 0$ such that, for all $x, y \in M$,*

$$\|\operatorname{grad} f(x) - \mathcal{T}_X(\operatorname{grad} f(y))\|_x \leq L d(x, y), \tag{20}$$

where $X \in T_y M$ satisfies $x = \exp_y(X)$. Furthermore, f is strongly convex, i.e., there exists a constant $\mu > 0$ such that, for all $x \in M$, the smallest eigenvalue of the Riemannian Hessian $\operatorname{Hess} f(x)$ is not less than μ.

3.2   Convergence results

Our first result is that Algorithm 3.1 including the HZ-type RCG method generates a sufficient descent direction without depending on the line search conditions.

Theorem 3.2. *Let $\{x_k\}_{k=0,1,\dots} \subset M$ be a sequence generated by Algorithm 3.1 with $\beta_k \in [\beta_k^{\mathrm{HZ}}, \max\{\beta_k^{\mathrm{HZ}}, 0\}]$ [2]. If $\langle y_k, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}} \neq 0$, we have*

$$\langle g_k, \eta_k \rangle_{x_k} \leq - \left(1 - \frac{1}{4\mu}\right) \|g_k\|_{x_k}^2. \tag{21}$$

Proof. For $k = 0$, (21) clearly holds from $\langle g_0, \eta_0 \rangle_{x_0} = - \|g_0\|_{x_0}^2$. Subsequently, we assume $k \geq 1$. If $\beta_k = \beta_k^{\mathrm{HZ}}$, from [12, Theorem 3.4], (21) follows. On the other hand, if $\beta_k \neq \beta_k^{\mathrm{HZ}}$, then $\beta_k^{\mathrm{HZ}} \leq \beta_k \leq 0$. From (3), we have

$$\langle g_k, \eta_k \rangle_{x_k} = - \|g_k\|_{x_k}^2 + \beta_k \left\langle g_k, \mathcal{T}_{\alpha_{k-1} \eta_{k-1}}(\eta_{k-1}) \right\rangle_{x_k}.$$

If $\left\langle g_k, \mathcal{T}_{\alpha_{k-1} \eta_{k-1}}(\eta_{k-1}) \right\rangle_{x_k} \geq 0$, then (21) follows immediately since $\beta_k \leq 0$. If $\left\langle g_k, \mathcal{T}_{\alpha_{k-1} \eta_{k-1}}(\eta_{k-1}) \right\rangle_{x_k} < 0$, then

$$\begin{aligned}
\langle g_k, \eta_k \rangle_{x_k} &= - \|g_k\|_{x_k}^2 + \beta_k \left\langle g_k, \mathcal{T}_{\alpha_{k-1} \eta_{k-1}}(\eta_{k-1}) \right\rangle_{x_k} \\
&\leq - \|g_k\|_{x_k}^2 + \beta_k^{\mathrm{HZ}} \left\langle g_k, \mathcal{T}_{\alpha_{k-1} \eta_{k-1}}(\eta_{k-1}) \right\rangle_{x_k},
\end{aligned}$$

since $\beta_k^{\mathrm{HZ}} \leq \beta_k \leq 0$. Hence, (21) follows by analysis as in [12, Theorem 3.4]. $\qquad\square$

The following is the main theorem indicating that the HZ-type RCG method converges globally.

Theorem 3.3. *Let $\{x_k\}_{k=0,1,\dots} \subset M$ be a sequence generated by Algorithm 3.1 with $\beta_{k+1} = \beta_{k+1}^{\mathrm{HZ}}$ under Assumptions 3.1 and 3.2. Suppose that each step size $\alpha_k > 0$ satisfies Wolfe conditions (10) and (11). Then either $\|g_{k_0}\|_{x_{k_0}} = 0$ for some $k_0 \in \mathbb{N}$, or*

$$\lim_{k \to \infty} \|g_k\|_{x_k} = 0. \tag{22}$$

---

[2] The modified HZ (18) satisfies $\hat{\beta}_k^{\mathrm{HZ}} \in [\beta_k^{\mathrm{HZ}}, \max\{\beta_k^{\mathrm{HZ}}, 0\}]$.

Proof. If $g_{k_0} = 0$ for some $k_0 \in \mathbb{N}$, then (22) obviously follows. Assume that $g_k \neq 0$ for all $k \in \mathbb{N}$. From (14) and (11), we have

$$(c_2 - 1) \langle g_k, \eta_k \rangle_{x_k} \leq \langle g_{k+1} - \mathcal{T}_{\alpha_k \eta_k}(g_k), \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}}.$$

Moreover, (20) and the Cauchy–Schwarz inequality imply

$$
\begin{aligned}
\langle g_{k+1} - \mathcal{T}_{\alpha_k \eta_k}(g_k), \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}} &\leq \|g_{k+1} - \mathcal{T}_{\alpha_k \eta_k}(g_k)\|_{x_{k+1}} \|\mathcal{T}_{\alpha_k \eta_k}(\eta_k)\|_{x_{k+1}} \\
&= \|g_{k+1} - \mathcal{T}_{\alpha_k \eta_k}(g_k)\|_{x_{k+1}} \|\eta_k\|_{x_k} \\
&\leq L \alpha_k \|\eta_k\|_{x_k}^2.
\end{aligned}
$$

Therefore, we obtain

$$(c_2 - 1) \langle g_k, \eta_k \rangle_{x_k} \leq L \alpha_k \|\eta_k\|_{x_k}^2,$$

which with Theorem 3.2 implies

$$\alpha_k \geq \frac{1 - c_2}{L} \frac{|\langle g_k, \eta_k \rangle_{x_k}|}{\|\eta_k\|_{x_k}^2}. \tag{23}$$

Moreover,

$$
\begin{aligned}
\langle y_k, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}} &= \langle g_{k+1} - \mathcal{T}_{\alpha_k \eta_k}(g_k), \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}} \\
&= \langle g_{k+1}, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}} - \langle \mathcal{T}_{\alpha_k \eta_k}(g_k), \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}} \\
&= \langle g_{k+1}, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}} - \langle g_k, \eta_k \rangle_{x_k} \\
&\geq c_2 \langle g_k, \eta_k \rangle_{x_k} - \langle g_k, \eta_k \rangle_{x_k} \\
&= (c_2 - 1) \langle g_k, \eta_k \rangle_{x_k}, \tag{24}
\end{aligned}
$$

where the third equation comes from (14), and the inequality comes from (11). Furthermore, we define $\phi_{x,\mu}(t) := f(\exp_x(t\mu))$. From Taylor's theorem, we have

$$
\begin{aligned}
f(x_{k+1}) - f(x_k) &= f(\exp_{x_k}(\alpha_k \eta_k)) - f(x_k) \\
&= \phi_{x_k, \frac{\eta_k}{\|\eta_k\|_{x_k}}}(\alpha_k \|\eta_k\|_{x_k}) - \phi_{x_k, \frac{\mu_k}{\|\eta_k\|_{x_k}}}(0) \\
&= \phi'_{x_k, \frac{\eta_k}{\|\eta_k\|_{x_k}}}(0) \alpha_k \|\eta_k\|_{x_k} + \frac{1}{2} \phi''_{x_k, \frac{\eta_k}{\|\eta_k\|_{x_k}}}(\theta)(\alpha_k \|\eta_k\|_{x_k})^2,
\end{aligned}
$$

11

for some $\theta \in [0, \alpha_k \|\eta_k\|_{x_k}]$. Moreover, by defining $c_{x,\mu}(t) := \exp_x(t\mu)$, we have

$$\phi''_{x_k, \frac{\eta_k}{\|\eta_k\|_{x_k}}}(\theta) = \left\langle \mathrm{Hess}\, f(x_k) \left[c'_{x_k, \frac{\eta_k}{\|\eta_k\|_{x_k}}}(\theta) \right], c'_{x_k, \frac{\eta_k}{\|\eta_k\|_{x_k}}}(\theta) \right\rangle_{c_{x_k, \frac{\eta_k}{\|\eta_k\|_{x_k}}}(\theta)}$$

$$\geq \mu \left\| c'_{x_k, \frac{\eta_k}{\|\eta_k\|_{x_k}}}(\theta) \right\|^2_{c_{x_k, \frac{\eta_k}{\|\eta_k\|_{x_k}}}(\theta)}$$

$$= \mu \left\| \frac{\eta_k}{\|\eta_k\|_{x_k}} \right\|^2_{x_k}$$

$$= \mu$$

from the strong convexity of f. Using this evaluation of ϕ'', we obtain

$$f(x_{k+1}) - f(x_k) \geq \left\langle g_k, \frac{\eta_k}{\|\eta_k\|_{x_k}} \right\rangle_{x_k} \alpha_k \|\eta_k\|_{x_k} + \frac{\mu}{2}(\alpha_k \|\eta_k\|_{x_k})^2$$

$$= \alpha_k \langle g_k, \eta_k \rangle_{x_k} + \frac{\mu}{2}(\alpha_k \|\eta_k\|_{x_k})^2$$

Therefore, we obtain

$$f(x_{k+1}) - f(x_k) \geq \alpha_k \langle g_k, \eta_k \rangle_{x_k} + \frac{\mu}{2}(\alpha_k \|\eta_k\|_{x_k})^2. \tag{25}$$

From (10) and (25), we have

$$\langle g_k, \eta_k \rangle_{x_k} \leq \frac{\mu}{2(c_1 - 1)} \alpha_k \|\eta_k\|^2_{x_k}. \tag{26}$$

Therefore, from (24) and (26), we obtain

$$\langle y_k, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}} \geq \gamma \alpha_k \|\eta_k\|^2_{x_k}, \tag{27}$$

where

$$\gamma := \frac{\mu(1 - c_2)}{2(1 - c_1)}.$$

The assumption $g_k \neq 0$ implies that $\eta_k \neq 0$, which together with $\alpha_k > 0$ yield $\langle y_k, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}} \neq 0$. From (10) and the lower boundedness of f,

$$\sum_{k=0}^{\infty} c_1 \alpha_k \langle g_k, \eta_k \rangle_{x_k} \geq \sum_{k=0}^{\infty} (f(x_{k+1}) - f(x_k))$$

$$= \lim_{j \to +\infty} f(x_j) - f(x_0) > -\infty,$$

12

which implies

$$\sum_{k=0}^{\infty} \alpha_k \langle g_k, \eta_k \rangle_{x_k} > -\infty.$$

Combining this with the lower bound for α_k given in (23) and the sufficient descent property in Theorem 3.2 gives

$$\sum_{k=0}^{\infty} \frac{\|g_k\|_{x_k}^4}{\|\eta_k\|_{x_k}^2} < \infty. \tag{28}$$

From (20), we obtain

$$\|y_k\|_{x_{k+1}} = \|g_{k+1} - \mathcal{T}_{\alpha_k \eta_k}(g_k)\|_{x_{k+1}} \leq L\alpha_k \|\eta_k\|_{x_k}. \tag{29}$$

From (15), (16), (27) and (29), we have

$$
\begin{aligned}
\left|\beta_{k+1}^{\text{HZ}}\right| &= \left| \frac{\langle g_{k+1}, y_k \rangle_{x_{k+1}}}{\langle y_k, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}}} - \mu \frac{\|y_k\|_{x_{k+1}}^2 \langle g_{k+1}, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}}}{\langle y_k, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}}^2} \right| \\
&\leq \frac{\|g_{k+1}\|_{x_{k+1}} \|y_k\|_{x_{k+1}}}{\gamma \alpha_k \|\eta_k\|_{x_k}^2} + \mu \frac{\|y_k\|_{x_{k+1}}^2 \|g_{k+1}\|_{x_{k+1}} \|\eta_k\|_{x_k}}{\gamma^2 \alpha_k^2 \|\eta_k\|_{x_k}^4} \\
&\leq \frac{L\alpha_k \|\eta_k\|_{x_k} \|g_{k+1}\|_{x_{k+1}}}{\gamma \alpha_k \|\eta_k\|_{x_k}^2} + \mu \frac{L^2 \alpha_k^2 \|\eta_k\|_{x_k}^3 \|g_{k+1}\|_{x_{k+1}}}{\gamma^2 \alpha_k^2 \|\eta_k\|_{x_k}^4} \\
&= \left(\frac{L}{\gamma} + \frac{\mu L^2}{\gamma^2} \right) \frac{\|g_{k+1}\|_{x_{k+1}}}{\|\eta_k\|_{x_k}}.
\end{aligned}
$$

Hence, we have

$$
\begin{aligned}
\|\eta_{k+1}\|_{x_{k+1}} &= \left\| -g_{k+1} + \beta_{k+1}^{\text{HZ}} \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \right\|_{x_{k+1}} \\
&\leq \|g_{k+1}\| + \left|\beta_{k+1}^{\text{HZ}}\right| \|\eta_k\|_{x_k} \\
&\leq \left(1 + \frac{L}{\gamma} + \frac{\mu L^2}{\gamma^2} \right) \|g_{k+1}\|_{x_{k+1}}.
\end{aligned}
$$

Combining this upper bound with (28), we obtain

$$\sum_{k=0}^{\infty} \|g_k\|_{x_k}^2 < \infty,$$

which completes the proof. $\qquad \square$

13

3.3 Comparison of HZ-type ECG method with HZ-type RCG method

Let us consider $M = \mathbb{R}^n$. Then, $\beta_{k+1}^{\mathrm{HZ}}$ defined by (15) can be expressed

$$
\begin{aligned}
\beta_{k+1}^{\mathrm{HZ}} &= \beta_{k+1}^{\mathrm{HS}} - \mu \frac{\|y_k\|_{x_{k+1}}^2 \langle g_{k+1}, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}}}{\langle y_k, \mathcal{T}_{\alpha_k \eta_k}(\eta_k) \rangle_{x_{k+1}}^2} \\
&= \frac{y_k^\top g_{k+1}}{\eta_k^\top y_k} - \mu \frac{\|y_k\|^2 g_{k+1}^\top \eta_k}{(y_k^\top \eta_k)^2} \\
&= \frac{1}{\eta_k^\top y_k} \left(y_k - \mu \frac{\|y_k\|^2}{\eta_k^\top y_k} \eta_k \right)^\top g_{k+1},
\end{aligned}
$$

which implies that $\beta_{k+1}^{\mathrm{HZ}}$ defined by (15) with $\mu = 2$ coincides with (1.3) in [4] used in the HZ-type ECG method. Inequality (21) with $\mu = 2$ (see Theorem 3.2) is the sufficient descent property of the HZ-type RCG method; i.e.,

$$
\langle g_k, \eta_k \rangle_{x_k} \le -\frac{7}{8} \|g_k\|_{x_k}^2,
$$

which, together with $M = \mathbb{R}^n$, implies (1.9) in Theorem 1.1 of [4]:

$$
g_k^\top \eta_k \le -\frac{7}{8} \|g_k\|^2.
$$

Accordingly, Theorem 3.2 is a natural extended result of Theorem 1.1 in [4] to a Riemannian manifold.

Theorem 2.2 in [4] implies that the HZ-type ECG method converges globally if the Wolfe conditions hold and if

- $f \colon \mathbb{R}^n \to \mathbb{R}$ is strongly convex with a constant $c > 0$ and $\nabla f \colon \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with Lipschitz constant $L > 0$ on the level set $\mathcal{L} := \{x \in \mathbb{R}^n \colon f(x_0) \le f(x)\}$.

Theorem 3.3 in this paper is satisfied under the Wolfe conditions and Assumptions 3.1 and 3.2, i.e.,

(i) $f \colon M \to \mathbb{R}$ is smooth and bounded below; $f \circ \exp_{x_k} \colon T_{x_k} M \to \mathbb{R}$ is Lipschitz continuously differentiable on $\mathrm{span}\{\eta_k\}$ with uniform Lipschitz constant $L > 0$;

(ii) There exists a constant $L > 0$ such that, for all $x, y \in M$,

$$
\|\mathrm{grad}\, f(x) - \mathcal{T}_X(\mathrm{grad}\, f(y))\|_x \le L d(x, y), \tag{30}
$$

where $X \in T_y M$ satisfies $x = \exp_y(X)$. Furthermore, f is strongly convex; i.e., there exists constant $\mu > 0$ such that, for all $x \in M$, the smallest eigenvalue of the Riemannian Hessian $\mathrm{Hess}\, f(x)$ is not less than μ.

Under the Euclidean space setting, f is strongly convex with a constant c if and only if the smallest eigenvalue of $\nabla^2 f(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^n$ is not less than c. Moreover, (30) in the Euclidean space setting is the same as the existence of $L > 0$ such that, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \le L\|\boldsymbol{x} - \boldsymbol{y}\|, \tag{31}$$

that is, ∇f is Lipschitz continuous. Obviously, the strong convexity of f implies that f is bounded below. Following Assumption 4.1 and Remark 4.1 in [15], we can see that, in the Euclidean space setting, the Lipschitz continuity of $f \circ \exp_{x_k}$ (see (i)) is equivalent to (31). Therefore, Theorem 3.3 is a natural extended result of Theorem 2.2 in [3] to a Riemannian manifold.

4 Numerical Experiments

We compared the performances of the HZ method with existing RCG methods, i.e., the FR, DY, PRP, HS, HS–DY hybrid, and FR–PRP hybrid methods. We solved two Riemannian optimization problems (Problem 4.1 and 4.2) on the unit sphere on a MacBook Air laptop computer (2020) with a 1.1-GHz Intel Core i3 CPU, 8-GB 3733-MHz LPDDR4X memory, and the Catalina 10.15.7 OS. The algorithms were written in Python 3.9.12. Each problem was solved 100 times with each algorithm, that is, 200 times in total.

We used a line search algorithm [12, Algorithm 3] for the strong Wolfe conditions (10) and (12) with $c_1 = 10^{-4}$ and $c_2 = 0.9$. If

$$\|\operatorname{grad} f(x_k)\|_{x_k} < 10^{-6},$$

was satisfied, we determined that the sequence had converged to an optimal solution.

For comparison, we calculated performance profile $P_s : \mathbb{R} \to [0, 1]$ [2], defined as follows. Let $\mathcal{P}$ and $\mathcal{S}$ be the set of problems and solvers, respectively. For each $p \in \mathcal{P}$ and $s \in \mathcal{S}$, we define

$$t_{p,s} := (\text{iterations or time required to solve problem } p \text{ by solver } s).$$

We define performance ratio $r_{p,s}$ as

$$r_{p,s} := \frac{t_{p,s}}{\min_{s' \in \mathcal{S}} t_{p,s'}}$$

and define the performance profile for all $\tau \in \mathbb{R}$ as

$$P_s(\tau) := \frac{|\{p \in \mathcal{P} : r_{p,s} \le \tau\}|}{|\mathcal{P}|},$$

where $|A|$ denotes the number of elements in set A.

4.1 The Rayleigh quotient minimization problem on the unit sphere

Problem 4.1 is the Rayleigh-quotient minimization problem on the unit sphere. The optimal solutions are the unit eigenvectors of A associated with the smallest eigenvalue (see [1, Chapter 4.6]).

Problem 4.1. *For a symmetric positive-definite matrix A,*

$$minimize \quad f(x) := x^\top A x,$$
$$subject\ to \quad x \in \mathbb{S}^{n-1} := \{x \in \mathbb{R}^n : \|x\| = 1\},$$

where $\|\cdot\|$ denotes the Euclidean norm.

In the experiments, we generated a matrix A randomly with $n = 100$ by using `sklearn.datasets.make_spd_matrix`.

Figure 1 plots the performance profile of each algorithm versus the number of iterations. It shows that the HZ method had much better performance than the FR, DY, PRP, and HS methods. Figure 2 plots the performance profile of each algorithm versus the elapsed time. It also shows that the performance of the HZ method was much better than those of the FR, DY, PRP, and HS methods. The two hybrid methods had even better performance. In particular, the figures show that they are suitable for solving Problem 4.1.
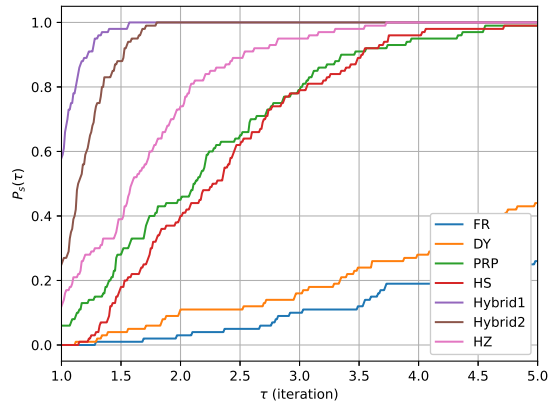

Figure 1: Performance profile versus number of iterations for Problem 4.1.

4.2 Computation of stability number

We define the stability number $S(G)$ of an undirected graph $G = (E, V)$ as the size of the maximum stable set in G. Motzkin and Straus showed that solving the stability number of graphs problem is equivalent to solving Problem 4.2 [7].
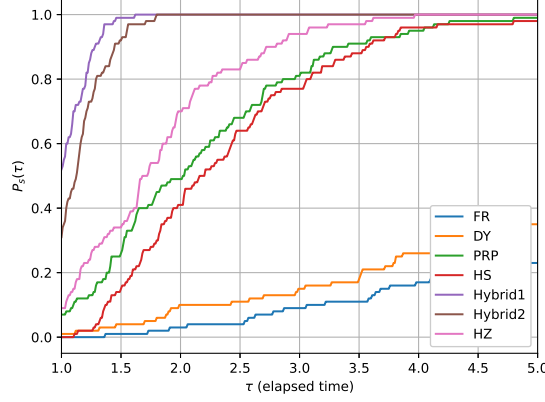
Figure 2: Performance profile versus elapsed time for Problem 4.1.

Problem 4.2. *For an undirected graph* $G = (E, V)$,

$$minimize \quad f(x) := \sum_{i=1}^{n} x_i^4 + \sum_{(i,j) \in E} x_i^2 x_j^2,$$

$$subject\ to \quad x \in \mathbb{S}^{n-1} := \{x \in \mathbb{R}^n : \|x\| = 1\},$$

where $n := |V|$ *and* $\|\cdot\|$ *denotes the Euclidean norm.*

In the experiments, we generated a graph $G = (E, V)$ randomly with $n = 100$ by using `networkx.fast_gnp_random_graph`. We set the probability for edge creation to 0.1.

Figure 3 plots the performance profile of each algorithm versus the number of iterations. It shows that the HZ method had much better performance than the existing methods. Figure 4 plots the performance profile of each algorithm versus the elapsed time. It also shows that the performance of the HZ method was much better than those of the existing methods. In particular, the figures show that the HZ method is suitable for solving Problem 4.2.

5   Conclusion

We have presented a Hager–Zhang (HZ)-type Riemannian conjugate gradient method that uses exponential retraction and presented two global convergence properties under different assumptions. We numerically compared the performance of the proposed method with those of existing Riemannian conjugate gradient methods for two Riemannian optimization problems on the unit sphere. The results show that the HZ method has much better performance than the FR, DY, PRP, and HS methods. In particular, we showed that the HZ method
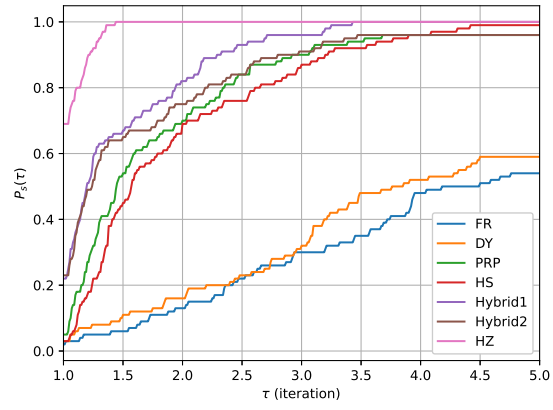
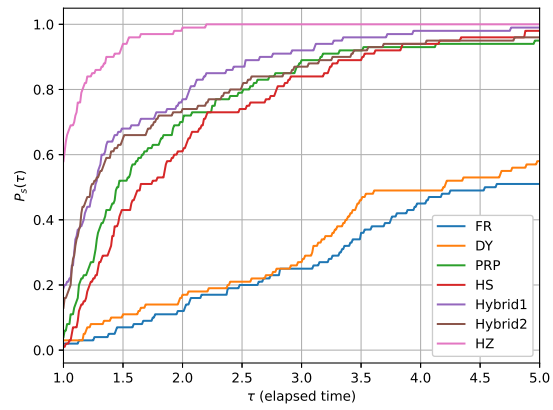Figure 3: Performance profile versus number of iterations for Problem 4.2.

Figure 4: Performance profile versus elapsed time for Problem 4.2.

is suitable for the stability number of graphs problem. It had much better performance than existing methods, including hybrid methods, for computing the stability number.

In a future paper, we will present the HZ method using a general retraction and its convergence analyses.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[2] E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91(2):201–213, 2002.

[3] W. W. Hager and H. Zhang. A new conjugate gradient method with guaranteed descent and an efficient line search. *SIAM Journal on Optimization*, 16(1):170–192, 2005.

[4] W. W. Hager and H. Zhang. A new conjugate gradient method with guaranteed descent and an efficient line search. *SIAM Journal on Optimization*, 16(1):170–192, 2005.

[5] R. Hosseini and S. Sra. Matrix manifold optimization for Gaussian mixtures. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[6] H. Kasai and B. Mishra. Low-rank tensor completion: a riemannian manifold preconditioning approach. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1012–1021, New York, New York, USA, 20–22 Jun 2016. PMLR.

[7] T. S. Motzkin and E. G. Straus. Maxima for graphs and a new proof of a theorem of turán. *Canadian Journal of Mathematics*, 17:533–540, 1965.

[8] M. Nickel and D. Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[9] W. Ring and B. Wirth. Optimization methods on Riemannian manifolds and their application to shape space. *SIAM Journal on Optimization*, 22(2):596–627, 2012.

[10] H. Sakai and H. Iiduka. Hybrid Riemannian conjugate gradient methods with global convergence properties. *Computational Optimization and Applications*, 77(3):811–830, 2020.

[11] H. Sakai and H. Iiduka. Riemannian adaptive optimization algorithm and its application to natural language processing. *IEEE Transactions on Cybernetics*, pages 1–12, 2021.

[12] H. Sakai and H. Iiduka. Sufficient descent Riemannian conjugate gradient methods. *Journal of Optimization Theory and Applications*, 190(1):130–150, 2021.

[13] T. Sakai. *Riemannian Geometry*, volume 149. American Mathematical Society, 1996.

[14] H. Sato. A Dai-Yuan-type Riemannian conjugate gradient method with the weak Wolfe conditions. *Computational Optimization and Applications*, 64(1):101–118, 2016.

[15] H. Sato. Riemannian conjugate gradient methods: General framework and specific algorithms with convergence analyses, 2021.

[16] H. Sato. *Riemannian Optimization and Its Applications*. Springer Briefs in Control, Automation and Robotics. Springer International Publishing, 2021.

[17] H. Sato and T. Iwai. A new, globally convergent Riemannian conjugate gradient method. *Optimization*, 64(4):1011–1031, 2015.

[18] S. T. Smith. Optimization techniques on Riemannian manifolds. *Fields Institute Communications*, 3:113–136, 1994.

[19] B. Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.