Riemannian Conjugate Gradient Methods: General Framework and Specific Algorithms with Convergence Analyses

AUTHOR(S):

Sato, Hiroyuki

CITATION:

Sato, Hiroyuki. Riemannian Conjugate Gradient Methods: General Framework and Specific Algorithms with Convergence Analyses. SIAM Journal on Optimization 2022, 32(4): 2690-2717

ISSUE DATE:

2022-12

URL:

http://hdl.handle.net/2433/278343

SIAM J. OPTIM.
Vol. 32, No. 4, pp. 2690–2717

RIEMANNIAN CONJUGATE GRADIENT METHODS: GENERAL FRAMEWORK AND SPECIFIC ALGORITHMS WITH CONVERGENCE ANALYSES*

HIROYUKI SATO†

Abstract. Conjugate gradient methods are important first-order optimization algorithms both in Euclidean spaces and on Riemannian manifolds. However, while various types of conjugate gradient methods have been studied in Euclidean spaces, there are relatively fewer studies for those on Riemannian manifolds (i.e., Riemannian conjugate gradient methods). This paper proposes a novel general framework that unifies existing Riemannian conjugate gradient methods such as the ones that utilize a vector transport or inverse retraction. The proposed framework also develops other methods that have not been covered in previous studies. Furthermore, conditions for the convergence of a class of algorithms in the proposed framework are clarified. Moreover, the global convergence properties of several specific types of algorithms are extensively analyzed. The analysis provides the theoretical results for some algorithms in a more general setting than the existing studies and new developments for other algorithms. Numerical experiments are performed to confirm the validity of the theoretical results. The experimental results are used to compare the performances of several specific algorithms in the proposed framework.

Key words. conjugate gradient method, Riemannian optimization, Riemannian manifold, vector transport, inverse retraction

MSC codes. 65K05, 90C30, 90C48

DOI. 10.1137/21M1464178

1. Introduction.

Riemannian optimization (i.e., optimization on Riemannian manifolds) has recently attracted increasing attention owing to its vast variety of applications, including machine learning, control engineering, and numerical linear algebra [3, 17, 51]. While nonsmooth or constrained Riemannian optimization problems have also been studied (see [2, 19, 37, 61] and [13, 40, 56, 59] and references therein for nonsmooth and constrained Riemannian optimization, respectively), we focus on smooth unconstrained Riemannian optimization problems in this paper. The class of unconstrained Riemannian optimization problems is also important since it overlaps the class of constrained optimization problems in Euclidean space. This is because a constrained Euclidean optimization problem can be regarded as an unconstrained Riemannian optimization problem if the set of constraints forms a Riemannian manifold. An important example is the Stiefel manifold, which is defined as $\mathrm{St}(p,n) := \{X \in \mathbb{R}^{n\times p} \mid X^TX = I_p\}$, where $p \leq n$ [22, 62]. Any optimization problem in $\mathbb{R}^{n\times p}$ with the constraint $X^TX = I_p$ (and without any other constraint) on the decision variable matrix X can be considered as an unconstrained optimization problem on $\mathrm{St}(p,n)$. Another example is the manifold $\mathrm{SPD}(n)$, which comprises all $n \times n$ symmetric positive definite matrices [15, 42]. Furthermore, the class of unconstrained Riemannian optimization problems also covers problems that are not defined in Euclidean space, e.g., optimization problems on the Grassmann manifold $\mathrm{Grass}(p,n) := \{W \subset \mathbb{R}^n \mid W \text{ is a } p\text{-dimensional subspace of } \mathbb{R}^n\}$

*Received by the editors December 10, 2021; accepted for publication (in revised form) July 5, 2022; published electronically November 9, 2022.
https://doi.org/10.1137/21M1464178
Funding: This work was funded by JSPS KAKENHI grant JP20K14359.
†Department of Applied Mathematics and Physics, Kyoto University, Kyoto, 6068501, Japan (hsato@i.kyoto-u.ac.jp, http://www-optima.amp.i.kyoto-u.ac.jp/~hsato).

for $p \leq n$, whose decision variable W is a subspace of $\mathbb{R}^n$, not a vector or matrix [22, 64].

An important feature of unconstrained Riemannian optimization problems is that they may be solved using the generalized versions of unconstrained Euclidean optimization methods, which have been studied intensively, if we successfully generalize the Euclidean methods to those on Riemannian manifolds appropriately. Various Euclidean optimization methods have been generalized to the Riemannian case, such as cyclic proximal point algorithms [9, 14], stochastic optimization algorithms [16, 54, 60], and multiobjective optimization algorithms [11, 24]. One of the challenges in Riemannian optimization lies in the accomplishment of the generalization of Euclidean optimization algorithms to the Riemannian case with attention to geometric structures. Other challenges include linking the theory of Riemannian optimization with real-world applications [27, 55] and analyzing geometry of and providing geometric tools on various manifolds that can be exploited in optimization [28, 33].

In this paper, we address the CG methods on Riemannian manifolds, which we refer to as the Riemannian conjugate gradient (R-CG) methods. The CG methods are appealing when large-scale optimization problems are to be solved because each iteration computationally costs much less than the second-order methods; hence, the CG methods find an approximate solution in a moderately fast time. Several types of R-CG methods have been studied. In some of these studies, parallel translation along the geodesics is utilized [22, 23, 39, 57]. This type of approach is theoretically natural; nonetheless, there is room for computational improvement. Other studies use a more general map called vector transport [3, 47, 48, 49, 50, 53]. Using a vector transport typically enables the execution of each iteration of R-CG methods more easily than using parallel translation, whereas it may negatively affect or sometimes destroy the convergence property of the algorithm. Therefore, in this paper, we provide a general framework of the R-CG methods, which includes successful existing methods, and we clarify the conditions with which the R-CG methods have a good convergence property.

The contributions of this paper are twofold: (i) We provide a novel general framework of the R-CG methods, which unifies the existing R-CG methods and covers a wider class, and we clarify the assumptions that are naturally required to apply an R-CG method to a Riemannian optimization problem. (ii) We generalize several types of standard Euclidean CG methods to the Riemannian case in our proposed framework. We also provide global convergence analyses for some specific practical algorithms. Although we do not generalize all the Euclidean CG methods in this paper because there are various algorithms, this paper provides a basis for studies on R-CG methods.

This paper is organized as follows. In section 2, we introduce the notation used throughout the paper. In section 3, we review the Euclidean CG and some existing R-CG methods, and we clarify what should be further addressed for the existing methods using some examples as a motivation for this study. In section 4, we propose our new general framework of R-CG methods. Thereafter, we introduce several types of practical R-CG methods as examples of the proposed framework. Some conditions that are imposed on step lengths are also proposed. In section 5, we summarize some standard assumptions for a Riemannian optimization problem to be solved. We also generalize Zoutendijk's theorem to our framework. In section 6, we provide the convergence analyses of several types of R-CG methods and discuss their behavior in our framework. Section 7 provides the results of some numerical experiments of different types of R-CG methods, in which they are compared. Finally, we conclude the paper in section 8.

2. Preliminaries. In this section, we summarize the notation and problem setting used throughout the paper.

The tangent space of a finite-dimensional manifold $\mathcal{M}$ at $x \in \mathcal{M}$ is denoted as $T_x\mathcal{M}$, and the tangent bundle of $\mathcal{M}$ is denoted as $T\mathcal{M} := \{(x, \eta) \mid \eta \in T_x\mathcal{M},\ x \in \mathcal{M}\}$. For a map $F\colon \mathcal{M} \to \mathcal{N}$ between two manifolds $\mathcal{M}$ and $\mathcal{N}$, $\mathrm{D}F(x)\colon T_x\mathcal{M} \to T_{F(x)}\mathcal{N}$ denotes the derivative (pushforward) of F at $x \in \mathcal{M}$.

In what follows, $\mathcal{M}$ denotes a finite-dimensional Riemannian manifold with a Riemannian metric $\langle\cdot,\cdot\rangle$; therefore, the tangent space $T_x\mathcal{M}$ at any $x \in \mathcal{M}$ is an inner product space with the inner product $\langle\cdot,\cdot\rangle_x$, which is given by the Riemannian metric $\langle\cdot,\cdot\rangle$. In the tangent space $T_x\mathcal{M}$ with the inner product $\langle\cdot,\cdot\rangle_x$, the norm of $\eta \in T_x\mathcal{M}$ is defined as $\|\eta\|_x := \sqrt{\langle\eta,\eta\rangle_x}$. The Riemannian gradient $\operatorname{grad} f(x)$ of a function $f\colon \mathcal{M} \to \mathbb{R}$ at $x \in \mathcal{M}$ is defined as a unique tangent vector at x satisfying $\langle \operatorname{grad} f(x), \eta\rangle_x = \mathrm{D}f(x)[\eta]$ for any $\eta \in T_x\mathcal{M}$.

The Euclidean space $\mathbb{R}^n$ with the standard inner product can be regarded as a Riemannian manifold with the Riemannian metric $\langle\cdot,\cdot\rangle$ defined by $\langle\xi,\eta\rangle_x := \xi^T\eta$ for any $\xi, \eta \in T_x\mathbb{R}^n \simeq \mathbb{R}^n$ and $x \in \mathbb{R}^n$. In the Euclidean space $\mathbb{R}^n$ with this Riemannian metric, the Riemannian gradient $\operatorname{grad} f(x)$ of $f\colon \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$ is equal to $\nabla f(x) := (\partial f(x)/\partial x_i) \in \mathbb{R}^n$. We refer to this as the Euclidean gradient.[1] The Euclidean norm (i.e., the 2-norm) of $a \in \mathbb{R}^n$ is denoted by $\|a\|_2 := \sqrt{a^Ta}$. The orthogonal group is denoted as $\mathcal{O}(n) := \mathrm{St}(n,n) = \{X \in \mathbb{R}^{n\times n} \mid X^TX = I_n\}$.

We consider the following unconstrained optimization problem for minimizing a sufficiently smooth[2] objective function $f\colon \mathcal{M} \to \mathbb{R}$ on a Riemannian manifold $\mathcal{M}$.

Problem 2.1.

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && x \in \mathcal{M}. \end{aligned}$$

In most Riemannian optimization algorithms, we use a retraction $R\colon T\mathcal{M} \to \mathcal{M}$, which is a generalization of the exponential map on $\mathcal{M}$ [3, 6]. With the notation $R_x := R|_{T_x\mathcal{M}}\colon T_x\mathcal{M} \to \mathcal{M}$, a retraction R on $\mathcal{M}$ is defined to satisfy $R_x(0_x) = x$ and $\mathrm{D}R_x(0_x) = \mathrm{id}_{T_x\mathcal{M}}$ for all $x \in \mathcal{M}$, where 0_x and $\mathrm{id}_{T_x\mathcal{M}}$ are the zero vector of $T_x\mathcal{M}$ and identity map in $T_x\mathcal{M}$, respectively. While the exponential map is theoretically natural, a retraction is used to increase efficiency in numerical optimization. Regarding the manifold of fixed-rank matrices, for example, no closed-form expression of the exponential map is known for the embedded geometry in general [4, 5]. Therefore, using a retraction is essential to efficiently solve optimization problems on this manifold. Furthermore, studies have been conducted on retractions that can be computed efficiently. For example, a Cholesky-QR-based retraction on the generalized Stiefel manifold was proposed to improve the computational efficiency [52].

3. Review of and discussion on the existing Riemannian CG methods. In this section, we review the extension of the CG methods in Euclidean spaces to Riemannian optimization, i.e., R-CG methods. In particular, we discuss what has been done so far and what should be further studied in generalizing the Euclidean CG

[1] In Riemannian geometry, the symbol ∇ usually denotes an affine connection on Riemannian manifold $\mathcal{M}$. However, because we do not explicitly use affine connections in this paper and we sometimes compare the Euclidean and Riemannian CG methods, we distinguish between the Euclidean and Riemannian gradients of function f in $\mathbb{R}^n$ and on $\mathcal{M}$ by writing them as ∇f and $\operatorname{grad} f$, respectively.

[2] In section 5, we will clarify the condition on smoothness of f required in convergence analyses in section 6.

methods to Riemannian ones, providing some examples to clarify the issues resolved in this paper.

In the Euclidean case, the CG methods are line search algorithms to solve Problem 2.1 with $\mathcal{M} = \mathbb{R}^n$ and an objective function $f\colon \mathbb{R}^n \to \mathbb{R}$, where a sequence $\{x_k\}$ in $\mathbb{R}^n$ is generated by

$$x_{k+1} = x_k + t_k \eta_k \tag{3.1}$$

with $t_k > 0$ for $k \geq 0$ from an initial point $x_0 \in \mathbb{R}^n$, and the search directions $\eta_k \in \mathbb{R}^n$ are computed using the gradient $g_k := \nabla f(x_k)$ as $\eta_0 = -g_0$ and

$$\eta_{k+1} = -g_{k+1} + \beta_{k+1}\eta_k \tag{3.2}$$

for $k \geq 0$ [44, section 5.2]. The computation of real values β_{k+1} is crucial for the performance of CG methods. Defining $y_k := g_k - g_{k-1}$, the following six types of β_{k+1} are considered as standard ones:

$$\begin{aligned} \beta_{k+1}^{\mathrm{FR}} &= \frac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2}, & \beta_{k+1}^{\mathrm{DY}} &= \frac{\|g_{k+1}\|_2^2}{y_{k+1}^T \eta_k}, & \beta_{k+1}^{\mathrm{CD}} &= \frac{\|g_{k+1}\|_2^2}{-g_k^T \eta_k}, \\ \beta_{k+1}^{\mathrm{PRP}} &= \frac{g_{k+1}^T y_{k+1}}{\|g_k\|_2^2}, & \beta_{k+1}^{\mathrm{HS}} &= \frac{g_{k+1}^T y_{k+1}}{y_{k+1}^T \eta_k}, & \beta_{k+1}^{\mathrm{LS}} &= \frac{g_{k+1}^T y_{k+1}}{-g_k^T \eta_k}. \end{aligned} \tag{3.3}$$

They were proposed by Fletcher and Reeves [26], Dai and Yuan [20], Fletcher [25], Polak and Ribière [45] and Polyak [46], Hestenes and Stiefel [34], and Liu and Storey [41], respectively. Here, "CD" in $\beta_{k+1}^{\mathrm{CD}}$ represents "conjugate descent." Although we do not describe all the existing CG methods here, various other types of β_{k+1} have also been examined. An important example is $\beta_{k+1}^{\mathrm{HZ}} = \frac{1}{y_{k+1}^T \eta_k}(y_{k+1} - 2\eta_k \frac{\|y_{k+1}\|_2^2}{y_{k+1}^T \eta_k})^T g_{k+1}$ by Hager and Zhang [31], whose Riemannian version is studied in [49].

In the R-CG methods, like many other Riemannian optimization algorithms, a sequence $\{x_k\}$ on $\mathcal{M}$ is generated from an initial point $x_0 \in \mathcal{M}$ by

$$x_{k+1} = R_{x_k}(t_k \eta_k) \tag{3.4}$$

for $k \geq 0$, where $R\colon T\mathcal{M} \to \mathcal{M}$ is a retraction on $\mathcal{M}$ and the search direction η_k is in $T_{x_k}\mathcal{M}$ [3, Algorithm 13]. This is because $\mathcal{M}$ is not a linear space in general, and the line search strategy (3.1) is not applicable if the half-line $\{x_k + t\eta_k \mid t > 0\}$ does not lie on $\mathcal{M}$. The formula (3.4) indicates that, given $x_k \in \mathcal{M}$, we find the subsequent point x_{k+1} on the curve $\gamma_k(t) := R_{x_k}(t\eta_k)$ with $\gamma_k(0) = R_{x_k}(0) = x_k$ and $\dot{\gamma}_k(0) = \mathrm{D}R_{x_k}(0)[\eta_k] = \eta_k$, instead of on a line. This is one of the main differences between Euclidean and Riemannian optimization. In the R-CG methods, conditions for step lengths t_k and how to compute search directions η_k also differ from the Euclidean case, as described below.

In the remainder of this section, we define $g_k := \operatorname{grad} f(x_k)$ and assume that η_k is a descent direction, i.e., $\langle g_k, \eta_k \rangle_{x_k} < 0$. Then, step lengths t_k should be chosen to satisfy some conditions such as the Riemannian version of the Wolfe conditions

$$f(R_{x_k}(t_k\eta_k)) \leq f(x_k) + c_1 t_k \langle g_k, \eta_k \rangle_{x_k} \tag{3.5}$$

and

$$\langle \operatorname{grad} f(R_{x_k}(t_k\eta_k)), \mathrm{D}R_{x_k}(t_k\eta_k)[\eta_k] \rangle_{R_{x_k}(t_k\eta_k)} \geq c_2 \langle g_k, \eta_k \rangle_{x_k}, \tag{3.6}$$

or strong Wolfe conditions (3.5) and

$$|\langle \operatorname{grad} f(R_{x_k}(t_k\eta_k)), \mathrm{D}R_{x_k}(t_k\eta_k)[\eta_k]\rangle_{R_{x_k}(t_k\eta_k)}| \le c_2|\langle g_k, \eta_k\rangle_{x_k}|, \tag{3.7}$$

where c_1 and c_2 are constants satisfying $0 < c_1 < c_2 < 1$ [51, section 3.5]. An algorithm to find a step length satisfying the Riemannian (strong) Wolfe conditions is discussed in [50], which is a generalization of the Euclidean case discussed in [38, 44]. Furthermore, in this paper, we define the (Riemannian) generalized Wolfe conditions as (3.5) and

$$\begin{aligned} c_2\langle g_k, \eta_k\rangle_{x_k} &\le \langle \operatorname{grad} f(R_{x_k}(t_k\eta_k)), \mathrm{D}R_{x_k}(t_k\eta_k)[\eta_k]\rangle_{R_{x_k}(t_k\eta_k)} \\ &\le -c_3\langle g_k, \eta_k\rangle_{x_k}, \end{aligned} \tag{3.8}$$

where $c_3 \ge 0$ is a constant. Theoretically, if $c_3 = c_2$, then the generalized Wolfe conditions are equivalent to the strong Wolfe conditions. However, in some practical cases, c_3 can be much larger than c_2; thus, it is not so restrictive. If c_3 is sufficiently large, the generalized Wolfe conditions are close to the Wolfe conditions. On the other hand, if $c_3 = 0$, then the generalized Wolfe conditions are stricter than the strong Wolfe conditions.

We proceed to generalizing the computation of the search directions (3.2) in the CG methods to the Riemannian case. The search direction at the initial point $x_0 \in \mathcal{M}$ is naturally determined, i.e., $\eta_0 = -g_0 = -\operatorname{grad} f(x_0)$. At the right-hand side of (3.2), $-\nabla f(x_{k+1}) \in \mathbb{R}^n$ is generalized to the negative Riemannian gradient $-g_{k+1} = -\operatorname{grad} f(x_{k+1}) \in T_{x_{k+1}}\mathcal{M}$ on $\mathcal{M}$, whereas $\beta_{k+1} \in \mathbb{R}$ and $\eta_k \in T_{x_k}\mathcal{M}$ on $\mathcal{M}$. Consequently, two vectors $-g_{k+1}$ and $\beta_{k+1}\eta_k$ belong to the distinct tangent spaces $T_{x_{k+1}}\mathcal{M}$ and $T_{x_k}\mathcal{M}$, respectively; therefore, they cannot be added together. To resolve this issue, several approaches were considered in the literature which we outline below.

In [22, 39, 57], R-CG methods were discussed in which $\eta_k \in T_{x_k}\mathcal{M}$ is parallel translated to $T_{x_{k+1}}\mathcal{M}$ to compute η_{k+1}, i.e., η_{k+1} is computed as

$$\eta_{k+1} = -g_{k+1} + \beta_{k+1}\mathrm{P}_{\gamma_k}^{1\leftarrow 0}(\eta_k), \tag{3.9}$$

where $\mathrm{P}_{\gamma_k}^{1\leftarrow 0}\colon T_{x_k}\mathcal{M} \to T_{x_{k+1}}\mathcal{M}$ is the parallel translation along the geodesic γ_k connecting x_k and x_{k+1} as $\gamma_k(0) = x_k$ and $\gamma_k(1) = x_{k+1}$. However, in some cases, the parallel translation is numerically impractical. For example, no closed form for the parallel translation along the geodesic on the Stiefel manifold is known [3, Example 8.1.2].

In [3, Definition 8.1.1], the concept of a more general map, called a vector transport, was proposed. Let $T\mathcal{M} \oplus T\mathcal{M} := \{(\xi, \eta) \mid \xi, \eta \in T_x\mathcal{M},\ x \in \mathcal{M}\}$ denote the Whitney sum. A map $\mathcal{T}\colon T\mathcal{M} \oplus T\mathcal{M} \to T\mathcal{M}$ is called a vector transport on $\mathcal{M}$ if it satisfies the following conditions: (1) There exists a retraction R on $\mathcal{M}$ such that $\mathcal{T}_\eta(\xi) \in T_{R_x(\eta)}\mathcal{M}$ for all $x \in \mathcal{M}$ and $\xi, \eta \in T_x\mathcal{M}$. (2) For any $x \in \mathcal{M}$ and $\xi \in T_x\mathcal{M}$, $\mathcal{T}_{0_x}(\xi) = \xi$ holds, where 0_x is the zero vector of $T_x\mathcal{M}$, i.e., $\mathcal{T}_{0_x}$ is the identity map. (3) For any $a, b \in \mathbb{R}$, $x \in \mathcal{M}$, and $\xi, \eta, \zeta \in T_x\mathcal{M}$, $\mathcal{T}_\eta(a\xi + b\zeta) = a\mathcal{T}_\eta(\xi) + b\mathcal{T}_\eta(\zeta)$ holds, i.e., $\mathcal{T}_\eta$ is a linear map from $T_x\mathcal{M}$ to $T_{R_x(\eta)}\mathcal{M}$. Note that a map $\mathcal{T}$ defined by $\mathcal{T}_\eta(\xi) := \mathrm{P}_{\gamma_{x,\eta}}^{1\leftarrow 0}(\xi)$ is a vector transport, where $\mathrm{P}_{\gamma_{x,\eta}}^{1\leftarrow 0}$ is the parallel translation along the geodesic $\gamma_{x,\eta}(t) := \operatorname{Exp}_x(t\eta)$ connecting $\gamma_{x,\eta}(0) = x$ and $\gamma_{x,\eta}(1) = \operatorname{Exp}_x(\eta)$ with the exponential map Exp as a retraction.

Using a general vector transport $\mathcal{T}$ on $\mathcal{M}$, the formula (3.9) is generalized to

$$\eta_{k+1} = -g_{k+1} + \beta_{k+1}\mathcal{T}_{t_k\eta_k}(\eta_k). \tag{3.10}$$

Note that the right-hand side is well defined because the first condition in the definition of vector transport ensures $\mathcal{T}_{t_k\eta_k}(\eta_k) \in T_{R_{x_k}(t_k\eta_k)}\mathcal{M} = T_{x_{k+1}}\mathcal{M}$. The formula (3.10) is more general than (3.9) since (3.10) includes (3.9) as a special case. However, careful consideration is required to make the resultant R-CG method work appropriately. A specific vector transport and its modified version (called a scaled vector transport) have been studied to be utilized in R-CG methods as follows.

In [47], Ring and Wirth analyzed the R-CG method (3.4) and (3.10) with a specific type of β_{k+1} defined by $\beta_{k+1} = \|g_{k+1}\|^2_{x_{k+1}}/\|g_k\|^2_{x_k}$, which is a natural generalization of $\beta^{\mathrm{FR}}_{k+1}$ in (3.3). They proved the global convergence property of this type of R-CG method with the differentiated retraction $\mathcal{T}^R$ as a vector transport $\mathcal{T}$ in (3.10), i.e.,

$$\eta_{k+1} = -g_{k+1} + \beta_{k+1}\mathcal{T}^R_{t_k\eta_k}(\eta_k) \tag{3.11}$$

with

$$\mathcal{T}^R_\eta(\xi) := \mathrm{D}R_x(\eta)[\xi], \qquad \xi,\, \eta \in T_x\mathcal{M}, \quad x \in \mathcal{M}, \tag{3.12}$$

assuming the inequality

$$\|\mathcal{T}^R_{t_k\eta_k}(\eta_k)\|_{x_{k+1}} \le \|\eta_k\|_{x_k} \quad \text{for all } k \ge 0. \tag{3.13}$$

However, this inequality does not necessarily hold, even in very natural situations, as shown by the following example.

Example 3.1. We consider the following QR-based retraction[3] R on the Stiefel manifold $\mathrm{St}(p,n)$ with $p \le n$:

$$R_X(\eta) := \mathrm{qf}(X+\eta), \qquad \eta \in T_X\,\mathrm{St}(p,n), \quad X \in \mathrm{St}(p,n), \tag{3.14}$$

where $T_X\,\mathrm{St}(p,n) = \{\xi \in \mathbb{R}^{n\times p} \mid X^T\xi + \xi^T X = 0\}$ and $\mathrm{qf}(\cdot)$ returns the Q-factor of the QR decomposition of the full-rank matrix in parentheses, i.e., if $A \in \mathbb{R}^{n\times p}$ is of full rank[4] and is uniquely decomposed as $A = QR$ with $Q \in \mathrm{St}(p,n)$ and R being an upper triangular matrix with positive diagonal elements, then $\mathrm{qf}(A) = Q$. For this retraction R, the differentiated retraction $\mathcal{T}^R$ is computed as [3, Example 8.1.5]

$$\mathcal{T}^R_\eta(\xi) := \mathrm{D}R_X(\eta)[\xi] = X_+\rho_{\mathrm{skew}}(X_+^T\xi R_+^{-1}) + (I_n - X_+X_+^T)\xi R_+^{-1} \tag{3.15}$$

for $X \in \mathrm{St}(p,n)$ and $\xi, \eta \in T_X\,\mathrm{St}(p,n)$, where $X+\eta = X_+R_+$ is the QR decomposition of $X+\eta$ (i.e., $X_+ = \mathrm{qf}(X+\eta)$ and $R_+ = X_+^T(X+\eta)$) and $\rho_{\mathrm{skew}}(\cdot)$ returns the skew-symmetric matrix that has the same size and strict lower part as those of the matrix in parentheses.

When $n \ge p \ge 2$, inequality (3.13) does not necessarily hold. As an example, we consider the Stiefel manifold $\mathrm{St}(p,n)$ with $n = p = 3$, which is reduced to the orthogonal group $\mathcal{O}(p) = \mathcal{O}(3)$, as a Riemannian submanifold of the Euclidean space $\mathbb{R}^{p\times p}$, i.e., the Riemannian metric on $\mathrm{St}(p,n)$ is defined as $\langle\xi,\eta\rangle_X = \mathrm{tr}(\xi^T\eta)$ for $X \in \mathrm{St}(p,n)$ and $\xi,\eta \in T_X\,\mathrm{St}(p,n)$. Assume that $X_k = I_3 \in \mathcal{O}(3)$, the search direction is

$$\eta_k = \begin{pmatrix} 0 & -1 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{pmatrix} \in T_{X_k}\mathcal{O}(3) = T_{I_3}\mathcal{O}(3) = \{Y \in \mathbb{R}^{3\times 3} \mid Y + Y^T = 0\},$$

[3] Another popular retraction is the one based on the polar decomposition [3, Example 4.1.3]. Furthermore, a recent study [10] implies that retractions can be efficient and still close to geodesics.

[4] In (3.14), $X+\eta$ is always of full rank because $\mathrm{rank}(X+\eta) = \mathrm{rank}((X+\eta)^T(X+\eta)) = \mathrm{rank}(I_p + \eta^T\eta) = p$. The last equality follows from the fact that $I_p + \eta^T\eta$ is invertible because it is positive definite.

and the step length is $t_k = 0.1$. Then, from (3.15), a straightforward calculation yields

$$\|\mathcal{T}^R_{t_k\eta_k}(\eta_k)\|_{R_{X_k}(t_k\eta_k)} \approx 2.47 > \sqrt{6} = \|\eta_k\|_{X_k},$$

which violates (3.13).[5]

While inequality (3.13) is important for the global convergence property, the linearity of a vector transport is not crucial in R-CG methods. Based on this important observation, Sato and Iwai [53] proposed the notion of a scaled vector transport $\mathcal{T}^{(0)}$ associated with a vector transport $\mathcal{T}$ defined as

$$\mathcal{T}^{(0)}_\eta(\xi) := \frac{\|\xi\|_x}{\|\mathcal{T}_\eta(\xi)\|_{R_x(\eta)}}\mathcal{T}_\eta(\xi)$$

for $x \in \mathcal{M}$ and $\xi, \eta \in T_x\mathcal{M}$. They also proposed a strategy where, in (3.11), the scaled vector transport $\mathcal{T}^{(0)}$ associated with the differentiated retraction $\mathcal{T}^R$ is used instead of $\mathcal{T}^R$ only when inequality (3.13) is violated, and $\mathcal{T}^R$ itself without scaling is used if otherwise. This approach can be regarded as

$$\eta_{k+1} = -g_{k+1} + \beta_{k+1}s_k\mathcal{T}^R_{t_k\eta_k}(\eta_k) \tag{3.16}$$

with the scaling parameter $s_k := \min\{1, \|\eta_k\|_{x_k}/\|\mathcal{T}^R_{t_k\eta_k}(\eta_k)\|_{R_{x_k}(t_k\eta_k)}\} > 0$, where $t_k\eta_k \neq 0$. Note that $\mathcal{T}^{(0)}$ is not a vector transport because $\mathcal{T}^{(0)}_\eta\colon T_x\mathcal{M} \to T_{R_x(\eta)}\mathcal{M}$ for $x \in \mathcal{M}$ and $\eta \in T_x\mathcal{M}$ is not a linear map. Considering the same framework, Sato [50], Sakai and Iiduka [48], and Sakai and Iiduka [49] analyzed the Dai–Yuan-type, some hybrid β_{k+1}-based, and Hager–Zhang-type of R-CG methods, respectively.

Recently, Zhu and Sato [63] proposed a completely different approach from (3.10) or (3.11). Their algorithm uses an additional retraction R^{bw}, where "bw" represents "backward" and R^{bw} can be the same as or different from R in (3.4). Specifically, given $\eta_k \in T_{x_k}\mathcal{M}$ and $x_{k+1} \in \mathcal{M}$, the search direction η_{k+1} at x_{k+1} is computed as

$$\eta_{k+1} = -g_{k+1} - \beta_{k+1}s_kt_k^{-1}(R^{\mathrm{bw}}_{x_{k+1}})^{-1}(x_k) \tag{3.17}$$

with the scaling parameter $s_k := \min\{1, \|\eta_k\|_{x_k}/\|t_k^{-1}(R^{\mathrm{bw}}_{x_{k+1}})^{-1}(x_k)\|_{x_{k+1}}\}$. This means that the quantity $-t_k^{-1}(R^{\mathrm{bw}}_{x_{k+1}})^{-1}(x_k)$ is used in (3.17) instead of $\mathcal{T}^R_{t_k\eta_k}(\eta_k)$ in (3.16). In [63], the FR- and DY-types of R-CG methods with inverse retraction are analyzed. Furthermore, the inverse retraction is easily computed in some specific cases (e.g., when R^{bw} is the orthographic retraction). An example of inverse retraction $(R^{\mathrm{bw}})^{-1}$ on the manifold of symmetric positive definite matrices, without the necessity of knowing the explicit expression of R^{bw}, is found in [30]. Since (3.17) is vector transport free, other approaches can also be possible for the R-CG methods, leading to the idea of the general framework proposed in the subsequent section.

4. New general framework of Riemannian CG methods. In this section, we propose a new framework of R-CG methods, which contains the existing R-CG methods as special cases. Furthermore, we generalize standard formulas for β_{k+1} in (3.3) to the Riemannian case. Some conditions for step lengths in the proposed framework are also introduced.

[5] An exact calculation shows that $\|\mathcal{T}^R_{t_k\eta_k}(\eta_k)\|_{R_{X_k}(t_k\eta_k)} = 200\sqrt{42849907}/530553$.

4.1. Algorithm. We propose a new general framework of R-CG methods in which we use a general map $\mathscr{T}^{(k)}\colon T_{x_k}\mathcal{M} \to T_{x_{k+1}}\mathcal{M}$ to transport $\eta_k \in T_{x_k}\mathcal{M}$ to $T_{x_{k+1}}\mathcal{M}$, i.e., the search direction η_{k+1} is computed as

$$\eta_{k+1} = -g_{k+1} + \beta_{k+1} s_k \mathscr{T}^{(k)}(\eta_k), \tag{4.1}$$

where $g_{k+1} := \operatorname{grad} f(x_{k+1})$ and s_k is a scaling parameter satisfying

$$0 < s_k \leq \min\left\{1, \frac{\|\eta_k\|_{x_k}}{\|\mathscr{T}^{(k)}(\eta_k)\|_{x_{k+1}}}\right\}, \tag{4.2}$$

which stems from the same idea as a scaled vector transport (see subsection 4.2 for more details). We summarize the proposed framework of the R-CG methods as Algorithm 4.1.

Algorithm 4.1. General framework of the R-CG methods for Problem 2.1 on Riemannian manifold $\mathcal{M}$ with retraction R.

1: Choose an initial point $x_0 \in \mathcal{M}$ and set $\eta_0 := -\operatorname{grad} f(x_0)$.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Compute a step length $t_k > 0$ and $x_{k+1} := R_{x_k}(t_k\eta_k)$.
4: Compute $g_{k+1} := \operatorname{grad} f(x_{k+1})$ and $\beta_{k+1} \in \mathbb{R}$.
5: Compute a search direction as $\eta_{k+1} := -g_{k+1} + \beta_{k+1} s_k \mathscr{T}^{(k)}(\eta_k)$, where $\mathscr{T}^{(k)}\colon T_{x_k}\mathcal{M} \to T_{x_{k+1}}\mathcal{M}$ and $0 < s_k \leq \min\{1, \|\eta_k\|_{x_k}/\|\mathscr{T}^{(k)}(\eta_k)\|_{x_{k+1}}\}$.
6: $k := k + 1$.
7: **end for**

We need to clarify what conditions $\mathscr{T}^{(k)}$ should satisfy, how to compute β_k, and how step lengths t_k should be chosen. We address these in the subsequent subsections.

4.2. Map $\mathscr{T}^{(k)}$ and scaling parameter s_k. The map $\mathscr{T}^{(k)}$ in Algorithm 4.1 can be any map such that it appropriately transports $\eta_k \in T_{x_k}\mathcal{M}$ to $T_{x_{k+1}}\mathcal{M}$. Several conditions used in convergence analyses in section 6 are discussed at the end of this subsection. An important feature of Algorithm 4.1 is that we do not necessarily require $\mathscr{T}^{(k)}$ to be based on a vector transport. Furthermore, $\mathscr{T}^{(k)}$ is not necessarily a linear map. Therefore, the R-CG method with inverse retraction introduced in (3.17) is also contained in this framework as specifically explained in Example 4.1.

Furthermore, any inequality corresponding to (3.13) is not required in terms of $\mathscr{T}^{(k)}$. Instead, the scaling parameter $s_k \in (0, \min\{1, \|\eta_k\|_{x_k}/\|\mathscr{T}^{(k)}(\eta_k)\|_{x_{k+1}}\}]$ plays a role to ensure a similar inequality $\|s_k\mathscr{T}^{(k)}(\eta_k)\|_{x_{k+1}} \leq \|\eta_k\|_{x_k}$ for all $k \geq 0$.

Example 4.1. In Algorithm 4.1, we know several choices of $\mathscr{T}^{(k)}$ since this algorithm includes all the R-CG methods introduced in section 3. For example, if we set $\mathscr{T}^{(k)}(\eta_k) := \mathrm{P}_{\gamma_k}^{1\leftarrow 0}(\eta_k)$ and $s_k := 1$ in Algorithm 4.1 with parallel translation P_{γ_k} along the geodesic γ_k connecting x_k and x_{k+1}, then (4.1) reduces to (3.9). Here, we can take $s_k = 1$ because the parallel translation is isometric, i.e., $\|\mathrm{P}_{\gamma_k}^{1\leftarrow 0}(\eta_k)\|_{x_{k+1}} = \|\eta_k\|_{x_k}$. If we set $\mathscr{T}^{(k)}(\eta_k) := \mathcal{T}^R_{t_k\eta_k}(\eta_k)$ with the differentiated retraction $\mathcal{T}^R$ defined by (3.12) and $s_k := \min\{1, \|\eta_k\|_{x_k}/\|\mathcal{T}^R_{t_k\eta_k}(\eta_k)\|_{x_{k+1}}\}$, then (4.1) reduces to (3.16). Furthermore, if we set $\mathscr{T}^{(k)}(\eta_k) := -t_k^{-1}(R^{\mathrm{bw}}_{R_{x_k}(t_k\eta_k)})^{-1}(x_k)$ with the inverse of a retraction R^{bw} on $\mathcal{M}$ and $s_k := \min\{1, \|\eta_k\|_{x_k}/\|t_k^{-1}(R^{\mathrm{bw}}_{x_{k+1}})^{-1}(x_k)\|_{x_{k+1}}\}$, then (4.1) reduces to (3.17). For these three examples, the chosen s_k can be uniformly written as $s_k = \min\{1, \|\eta_k\|_{x_k}/\|\mathscr{T}^{(k)}(\eta_k)\|_{x_{k+1}}\}$.

In [51, Algorithm 4.2], Sato proposed a prototype of Algorithm 4.1, where s_k satisfies (4.2) and $\mathscr{T}^{(k)}(\eta_k) := \mathcal{T}^R_{t_k\eta_k}(\eta_k)$. However, Algorithm 4.1 is more general than the algorithm in [51] because $\mathscr{T}^{(k)}$ is not restricted to the differentiated retraction-based map.

Algorithm 4.1 covers a wider class of R-CG methods than the existing ones partly because we do not limit the scaling parameter $s_k > 0$ to be a specific form such as $s_k = \min\{1, \|\eta_k\|_{x_k}/\|\mathscr{T}^{(k)}(\eta_k)\|_{x_{k+1}}\}$. Moreover, it contains R-CG methods with $\mathscr{T}^{(k)}$ that have not been discussed in the literature. An example is to use a vector transport based on the orthogonal projection to the tangent spaces, which we will detail for the sphere and Grassmann manifold cases in Examples 4.5 and 4.6, respectively.

Subsequently, we introduce two conditions (4.3) and (4.4) on $\mathscr{T}^{(k)}$, which will be used in global convergence analyses of R-CG methods in section 6.

Assumption 4.2. For maps $\mathscr{T}^{(k)}$ in Algorithm 4.1, there exist $C \geq 0$ and index sets $K_1 \subset \mathbb{N}$ and $K_2 = \mathbb{N}\backslash K_1$ (the complement of K_1) such that

$$\|\mathscr{T}^{(k)}(\eta_k) - \mathrm{D}R_{x_k}(t_k\eta_k)[\eta_k]\|_{x_{k+1}} \leq Ct_k\|\eta_k\|^2_{x_k} \quad \text{for all } k \in K_1 \tag{4.3}$$

and

$$\|\mathscr{T}^{(k)}(\eta_k) - \mathrm{D}R_{x_k}(t_k\eta_k)[\eta_k]\|_{x_{k+1}} \leq C(t_k + t_k^2)\|\eta_k\|^2_{x_k} \quad \text{for all } k \in K_2 \tag{4.4}$$

hold, where $\mathbb{N}$ is the set of nonnegative integers.

Remark 4.3. In fact, for any $k \geq 0$, the inequality in (4.4) is weaker than that in (4.3). Therefore, Assumption 4.2 is equivalent to the condition that there exists $C \geq 0$ such that $\|\mathscr{T}^{(k)}(\eta_k) - \mathrm{D}R_{x_k}(t_k\eta_k)[\eta_k]\|_{x_{k+1}} \leq C(t_k + t_k^2)\|\eta_k\|^2_{x_k}$ holds for all $k \geq 0$. We write Assumption 4.2 with a stricter inequality (4.3) because we can weaken the condition imposed on t_k when $\mathscr{T}^{(k)}$ satisfies the inequality in (4.3) in forthcoming Theorem 5.3.

Assumption 4.2 requires that $\mathscr{T}^{(k)}$ is an approximation of the differentiated retraction $\mathcal{T}^R$. The R-CG methods with the differentiated retraction trivially satisfies the assumption, especially (4.3) with $C = 0$ and $K_1 = \mathbb{N}$. Furthermore, this assumption, especially (4.4) with $K_2 = \mathbb{N}$, is also natural for the R-CG methods with inverse retraction, as discussed in [63].

The proposed R-CG methods (Algorithm 4.1) will be analyzed in section 6 with Assumption 4.2. We realize that R-CG methods with some vector transports $\mathcal{T}$ that have not been analyzed yet also have convergence properties if $\mathscr{T}^{(k)}$ defined by $\mathcal{T}$ satisfies Assumption 4.2. Examples of such vector transports are shown as follows.

First, we prepare the following lemma used in Examples 4.5 and 4.6.

LEMMA 4.4. *Let h be a one-variable function defined for $t > 0$ as*

$$h(t) = \frac{1 - 1/\sqrt{1+t^2}}{t\sqrt{1+t^2}}.$$

Then, $0 < h(t) \leq C_0$ holds for all $t > 0$, where $C_0 = 4\sqrt{2/(349 + 85\sqrt{17})} \approx 0.2139$.

This lemma can be straightforwardly proved by differentiating h to obtain the maximum value of h, which is equal to C_0.

Example 4.5. Consider the sphere $S^{n-1} := \{x \in \mathbb{R}^n \mid x^Tx = 1\}$ with a retraction R and Riemannian metric $\langle\cdot,\cdot\rangle$ defined as $R_x(\eta) := (x+\eta)/\|x+\eta\|_2$ and $\langle\xi,\eta\rangle_x := \xi^T\eta$ for $x \in S^{n-1}$ and $\xi, \eta \in T_xS^{n-1}$, respectively. Thus, we regard S^{n-1} as a Riemannian

submanifold of $\mathbb{R}^n$. Generally, for a Riemannian submanifold, the orthogonal projections to the tangent spaces define a vector transport [3, section 8.1.3]. We can define such a vector transport $\mathcal{T}^P$ on S^{n-1} based on the orthogonal projection as

$$\mathcal{T}^P_\eta(\xi) := P_{R_x(\eta)}(\xi) = (I_n - R_x(\eta)R_x(\eta)^T)\xi = \left(I_n - \frac{(x+\eta)(x+\eta)^T}{\|x+\eta\|_2^2}\right)\xi$$

for $x \in S^{n-1}$ and $\eta, \xi \in T_xS^{n-1}$, where $P_y(d) = (I_n - yy^T)d$ is the orthogonal projection of $d \in \mathbb{R}^n$ to the tangent space $T_yS^{n-1} = \{z \in \mathbb{R}^n \mid y^Tz = 0\}$ at $y \in S^{n-1}$. This vector transport is typically used in R-CG methods practically (e.g., implemented in Manopt [18], Pymanopt [58], and Manopt.jl [12]); nevertheless, to the author's knowledge, it has not been theoretically discussed in detail in terms of the convergence properties of the R-CG methods. Therefore, it is meaningful to verify that $\mathscr{T}^{(k)}$ defined by $\mathcal{T}^P$ satisfies Assumption 4.2.

The differentiated retraction $\mathcal{T}^R$ is written as

$$\mathcal{T}^R_\eta(\xi) := \mathrm{D}R_x(\eta)[\xi] = \frac{1}{\|x+\eta\|_2}\left(I_n - \frac{(x+\eta)(x+\eta)^T}{\|x+\eta\|_2^2}\right)\xi = \frac{1}{\|x+\eta\|_2}P_{R_x(\eta)}(\xi).$$

Therefore, in Algorithm 4.1 with $\mathcal{M} = S^{n-1}$, we can evaluate the difference of the two vector transports, with $\eta = t_k\eta_k$ and $\xi = \eta_k \in T_{x_k}S^{n-1}$, as

$$\|\mathcal{T}^P_{t_k\eta_k}(\eta_k) - \mathcal{T}^R_{t_k\eta_k}(\eta_k)\|_{R_{x_k}(t_k\eta_k)} = \left|1 - \frac{1}{\|x_k + t_k\eta_k\|_2}\right| \|P_{R_{x_k}(t_k\eta_k)}(\eta_k)\|_{R_{x_k}(t_k\eta_k)}. \tag{4.5}$$

Considering $x_k^Tx_k = 1$ and $x_k^T\eta_k = 0$, we get $\|x_k + t_k\eta_k\|_2 = \sqrt{1 + t_k^2\|\eta_k\|_2^2} > 0$ and

$$\|P_{R_{x_k}(t_k\eta_k)}(\eta_k)\|^2_{R_{x_k}(t_k\eta_k)} = \left\|\eta_k - \frac{t_k\|\eta_k\|_2^2}{1 + t_k^2\|\eta_k\|_2^2}(x_k + t_k\eta_k)\right\|_2^2 = \frac{\|\eta_k\|_2^2}{1 + t_k^2\|\eta_k\|_2^2}. \tag{4.6}$$

If $t_k\|\eta_k\|_2 > 0$, using h and C_0 in Lemma 4.4 and combining (4.5) and (4.6), we obtain

$$\begin{aligned}\|\mathcal{T}^P_{t_k\eta_k}(\eta_k) - \mathcal{T}^R_{t_k\eta_k}(\eta_k)\|_{R_{x_k}(t_k\eta_k)} &= \left(1 - \frac{1}{\sqrt{1 + (t_k\|\eta_k\|_2)^2}}\right)\frac{\|\eta_k\|_2}{\sqrt{1 + (t_k\|\eta_k\|_2)^2}} \\ &= h(t_k\|\eta_k\|_2)\cdot t_k\|\eta_k\|_2^2 \le C_0t_k\|\eta_k\|_2^2 = C_0t_k\|\eta_k\|_{x_k}^2,\end{aligned}$$

whereas if $t_k\eta_k = 0$, we have $\|\mathcal{T}^P_0(\eta_k) - \mathcal{T}^R_0(\eta_k)\|_{x_k} = \|\eta_k - \eta_k\|_{x_k} = 0$. Therefore, for the sphere S^{n-1}, $\mathscr{T}^{(k)}(\eta_k) := \mathcal{T}^P_{t_k\eta_k}(\eta_k)$ satisfies the condition in Assumption 4.2 with $C = C_0$ and $K_1 = \mathbb{N}$.

Example 4.6. Consider the Grassmann manifold $\mathrm{Grass}(p,n) \simeq \mathrm{St}(p,n)/\mathcal{O}(p)$ with $p \le n$. For $X \in \mathrm{Grass}(p,n)$, let $\bar{X} \in \mathrm{St}(p,n)$ denote a representative of X and let $\bar{\eta}$ denote the horizontal lift of any $\eta \in T_X\,\mathrm{Grass}(p,n)$ at $\bar{X}$. We endow $\mathrm{Grass}(p,n)$ with the Riemannian metric $\langle\xi,\eta\rangle_X := \mathrm{tr}(\bar{\xi}^T\bar{\eta})$ for $\xi,\eta \in T_X\,\mathrm{Grass}(p,n)$ and the retraction R based on the polar decomposition defined through $\overline{R_X(\eta)} := (\bar{X}+\bar{\eta})(I_p + \bar{\eta}^T\bar{\eta})^{-1/2}$. For this retraction R, similarly to the previous example, the projection-based vector transport $\mathcal{T}^P$ and differentiated retraction $\mathcal{T}^R$ are written as [3, 36]

$$\overline{\mathcal{T}^P_\eta(\xi)} = (I_n - YY^T)\bar{\xi}, \qquad \overline{\mathcal{T}^R_\eta(\xi)} = (I_n - YY^T)\bar{\xi}(Y^T(\bar{X}+\bar{\eta}))^{-1},$$

where $Y = \overline{R_X(\eta)}$. We can generalize the discussion in Example 4.5 to this case.

We consider Algorithm 4.1 with $\mathcal{M} = \mathrm{Grass}(p, n)$. Here, we omit the subscript k for simplicity. Since $\bar{\eta}^T\bar{\eta}$ is symmetric positive semidefinite, it is decomposed as $\bar{\eta}^T\bar{\eta} =: Q\,\mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)Q^T$ with $Q \in \mathcal{O}(p)$ and $\lambda_1, \lambda_2, \ldots, \lambda_p \geq 0$. Defining $Z := \overline{R_X(t\eta)} = (\bar{X} + t\bar{\eta})(I_p + t^2\bar{\eta}^T\bar{\eta})^{-1/2}$, noting $\bar{X}^T\bar{\eta} = 0$, and using h and C_0 in Lemma 4.4, we obtain

$$\begin{aligned}
\|\mathcal{T}^P_{t\eta}(\eta) - \mathcal{T}^R_{t\eta}(\eta)\|^2_{R_X(t\eta)} &= \|(I_n - ZZ^T)\bar{\eta}(I_p - (Z^T(\bar{X} + t\bar{\eta}))^{-1})\|_F^2 \\
&= \mathrm{tr}(\bar{\eta}^T\bar{\eta}(I_p + t^2\bar{\eta}^T\bar{\eta})^{-1}(I_p - (I_p + t^2\bar{\eta}^T\bar{\eta})^{-1/2})^2) \\
&= \sum_{i=1}^{p}\left(1 - \frac{1}{\sqrt{1 + t^2\lambda_i}}\right)^2 \frac{\lambda_i}{1 + t^2\lambda_i} \\
&= \sum_{i=1}^{p} h\big(t\sqrt{\lambda_i}\big)^2 t^2\lambda_i^2 \\
&\leq {C_0}^2 t^2\left(\sum_{i=1}^{p}\lambda_i\right)^2 = (C_0 t\|\bar{\eta}\|_F^2)^2 = (C_0 t\|\eta\|_X^2)^2,
\end{aligned}$$

implying that $\|\mathcal{T}^P_{t\eta}(\eta) - \mathcal{T}^R_{t\eta}(\eta)\|_{R_X(t\eta)} \leq C_0 t\|\eta\|_X^2$. Therefore, $\mathscr{T}^{(k)}(\eta_k) := \mathcal{T}^P_{t_k\eta_k}(\eta_k)$ satisfies the condition in Assumption 4.2 with $C = C_0$ and $K_1 = \mathbb{N}$.

4.3. Computation of $\boldsymbol{\beta_{k+1}}$ in R-CG methods. In Algorithm 4.1, the computation of β_{k+1} in each iteration is crucial, and it affects the performance of the R-CG methods. Some of the six types of β_{k+1} in Euclidean CG methods shown in (3.3) have been generalized to the Riemannian case in each R-CG algorithm with a specific choice of $\mathscr{T}^{(k)}$ in the literature. For example, Smith [57] and Edelman, Arias, and Smith [22] proposed the generalization of $\beta^{\mathrm{LS}}_{k+1}$ and $\beta^{\mathrm{PRP}}_{k+1}$ with parallel translation along the geodesic, respectively. Ring and Wirth [47] and Sato and Iwai [53] analyzed the generalization of $\beta^{\mathrm{FR}}_{k+1}$ with the (scaled) vector transport defined through the differentiated retraction. Sato [50] proposed and analyzed the generalization of $\beta^{\mathrm{DY}}_{k+1}$ in the same framework as in [53]. Sakai and Iiduka [48] recently discussed a class of β_{k+1} containing a combination of the generalizations of $\beta^{\mathrm{DY}}_{k+1}$ and $\beta^{\mathrm{HS}}_{k+1}$ with the same (scaled) vector transports. Furthermore, Zhu and Sato [63] proposed and analyzed the generalizations of $\beta^{\mathrm{FR}}_{k+1}$ and $\beta^{\mathrm{DY}}_{k+1}$ with inverse retraction.

Here, we propose the Riemannian versions of the six types of β_{k+1}, generalized from (3.3) in the Euclidean CG methods. We put $g_k := \mathrm{grad}\, f(x_k) \in T_{x_k}\mathcal{M}$. From (3.3), we observe that $\beta^{\mathrm{FR}}_{k+1}$, $\beta^{\mathrm{DY}}_{k+1}$, and $\beta^{\mathrm{CD}}_{k+1}$ have a common numerator $\|g_{k+1}\|_2^2$ in the Euclidean case. This quantity can be easily and naturally generalized to the Riemannian case as $\|g_{k+1}\|^2_{x_{k+1}}$, i.e., the Euclidean gradient is replaced with the Riemannian gradient and the Euclidean norm is generalized to the norm in $T_{x_{k+1}}\mathcal{M}$ defined by the Riemannian metric. On the other hand, $\beta^{\mathrm{PRP}}_{k+1}$, $\beta^{\mathrm{HS}}_{k+1}$, and $\beta^{\mathrm{LS}}_{k+1}$ have the common numerator $g^T_{k+1}y_{k+1}$, where $y_{k+1} := g_{k+1} - g_k$. This is generalized to the Riemannian case on $\mathcal{M}$ by transporting $g_k \in T_{x_k}\mathcal{M}$ to $T_{x_{k+1}}\mathcal{M}$ using some map $\mathscr{S}^{(k)}\colon T_{x_k}\mathcal{M} \to T_{x_{k+1}}\mathcal{M}$ (which is possibly equal to $\mathscr{T}^{(k)}$) and some scaling parameter $l_k > 0$, and taking the inner product $\langle g_{k+1}, g_{k+1} - l_k\mathscr{S}^{(k)}(g_k)\rangle_{x_k}$ in $T_{x_k}\mathcal{M}$. Note that the map $\mathscr{S}^{(k)}$ is used to transport g_k, while $\mathscr{T}^{(k)}$ is used to transport η_k. Since the two maps play similar but different roles, we do not necessarily require $\mathscr{S}^{(k)}$ to be equal to $\mathscr{T}^{(k)}$.

Further, $\beta^{\mathrm{FR}}_{k+1}$ and $\beta^{\mathrm{PRP}}_{k+1}$ have the common denominator $\|g_k\|_2^2$, which is generalized to $\|g_k\|^2_{x_k}$, and $\beta^{\mathrm{CD}}_{k+1}$ and $\beta^{\mathrm{LS}}_{k+1}$ have the common denominator $-g^T_k\eta_k$, which is generalized to $-\langle g_k, \eta_k\rangle_{x_k}$. Finally, $\beta^{\mathrm{DY}}_{k+1}$ and $\beta^{\mathrm{HS}}_{k+1}$ have the common denominator $y^T_{k+1}\eta_k$.

In [50], where $\mathscr{T}^{(k)}(\eta_k) := \mathcal{T}^R_{t_k\eta_k}(\eta_k)$ and $s_k := \min\{1, \|\eta_k\|_{x_k}/\|\mathcal{T}^R_{t_k\eta_k}(\eta_k)\|_{x_{k+1}}\}$, the quantity $y_{k+1}^T\eta_k = g_{k+1}^T\eta_k - g_k^T\eta_k$ in the Euclidean case is generalized to the quantity $\langle g_{k+1}, s_k\mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}} - \langle g_k, \eta_k\rangle_{x_k}$. We follow this approach in (4.8) and (4.11) below.

In summary, we obtain the following formulas for the Riemannian version of β_{k+1} in (4.1), some of which depend on maps $\mathscr{T}^{(k)}$ and $\mathscr{S}^{(k)}$:

$$\beta^{\text{R-FR}}_{k+1} = \frac{\|g_{k+1}\|^2_{x_{k+1}}}{\|g_k\|^2_{x_k}}, \tag{4.7}$$

$$\beta^{\text{R-DY}}_{k+1} = \frac{\|g_{k+1}\|^2_{x_{k+1}}}{\langle g_{k+1}, s_k\mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}} - \langle g_k, \eta_k\rangle_{x_k}}, \tag{4.8}$$

$$\beta^{\text{R-CD}}_{k+1} = \frac{\|g_{k+1}\|^2_{x_{k+1}}}{-\langle g_k, \eta_k\rangle_{x_k}}, \tag{4.9}$$

$$\beta^{\text{R-PRP}}_{k+1} = \frac{\|g_{k+1}\|^2_{x_{k+1}} - \langle g_{k+1}, l_k\mathscr{S}^{(k)}(g_k)\rangle_{x_{k+1}}}{\|g_k\|^2_{x_k}}, \tag{4.10}$$

$$\beta^{\text{R-HS}}_{k+1} = \frac{\|g_{k+1}\|^2_{x_{k+1}} - \langle g_{k+1}, l_k\mathscr{S}^{(k)}(g_k)\rangle_{x_{k+1}}}{\langle g_{k+1}, s_k\mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}} - \langle g_k, \eta_k\rangle_{x_k}}, \tag{4.11}$$

$$\beta^{\text{R-LS}}_{k+1} = \frac{\|g_{k+1}\|^2_{x_{k+1}} - \langle g_{k+1}, l_k\mathscr{S}^{(k)}(g_k)\rangle_{x_{k+1}}}{-\langle g_k, \eta_k\rangle_{x_k}}. \tag{4.12}$$

Here, $l_k > 0$ and $\mathscr{S}^{(k)}\colon T_{x_k}\mathcal{M} \to T_{x_{k+1}}\mathcal{M}$ in (4.10)–(4.12) play similar roles to those of s_k and $\mathscr{T}^{(k)}$, respectively. However, we do not impose any specific conditions on l_k and $\mathscr{S}^{(k)}$ at this stage. Practically, it may be desirable that $l_k\mathscr{S}^{(k)}(g_k) \approx g_k$ holds when $t_k\eta_k \approx 0$, indicating when $x_{k+1} \approx x_k$. The R-CG methods with modified $\beta^{\text{R-PRP}}_{k+1}$, $\beta^{\text{R-HS}}_{k+1}$, and $\beta^{\text{R-LS}}_{k+1}$ will be discussed in detail in subsection 6.2.

We can verify that they all reduce to the corresponding existing β_k (if the literature exists, e.g., [50, 53, 63]) by specifying maps $\mathscr{T}^{(k)}$ and $\mathscr{S}^{(k)}$, such as a vector transport or inverse retraction. These discussions on generalization of several types of β_{k+1} will be justified through the convergence analyses in section 6.

4.4. Step length t_k. In the R-CG methods, the (strong) Wolfe conditions are especially important to guarantee their convergence properties. Because we introduce $\mathscr{T}^{(k)}$ in Algorithm 4.1, we need to slightly modify the conditions. In this subsection, we assume that the current iteration $x_k \in \mathcal{M}$ and search direction $\eta_k \in T_{x_k}\mathcal{M}$ are given. Further, we assume that η_k is a descent direction, i.e., $\langle g_k, \eta_k\rangle_{x_k} < 0$.

We revisit conditions (3.6)–(3.8), which appear in the (strong/generalized) Wolfe conditions. In these three conditions, the quantity $\mathrm{D}R_{x_k}(t_k\eta_k)[\eta_k]$ is commonly used. This is written as $\mathrm{D}R_{x_k}(t_k\eta_k)[\eta_k] = \mathcal{T}^R_{t_k\eta_k}(\eta_k)$ for the differentiated retraction $\mathcal{T}^R$ defined as (3.12). We generalize the (strong/generalized) Wolfe conditions by replacing $\mathcal{T}^R_{t_k\eta_k}(\eta_k)$ with $\mathscr{T}^{(k)}(\eta_k)$. Specifically, (3.6)–(3.8) are generalized as

$$\langle \operatorname{grad} f(R_{x_k}(t_k\eta_k)), \mathscr{T}^{(k)}(\eta_k)\rangle_{R_{x_k}(t_k\eta_k)} \geq c_2\langle g_k, \eta_k\rangle_{x_k}, \tag{4.13}$$

$$|\langle \operatorname{grad} f(R_{x_k}(t_k\eta_k)), \mathscr{T}^{(k)}(\eta_k)\rangle_{R_{x_k}(t_k\eta_k)}| \leq c_2|\langle g_k, \eta_k\rangle_{x_k}|, \tag{4.14}$$

and

$$c_2\langle g_k, \eta_k\rangle_{x_k} \leq \langle \operatorname{grad} f(R_{x_k}(t_k\eta_k)), \mathscr{T}^{(k)}(\eta_k)\rangle_{R_{x_k}(t_k\eta_k)} \leq -c_3\langle g_k, \eta_k\rangle_{x_k}, \tag{4.15}$$

respectively. We define the $\mathscr{T}^{(k)}$-*Wolfe conditions* as (3.5) and (4.13), *strong* $\mathscr{T}^{(k)}$-*Wolfe conditions* as (3.5) and (4.14), and *generalized* $\mathscr{T}^{(k)}$-*Wolfe conditions* as (3.5) and (4.15), where $0 < c_1 < c_2 < 1$ and $c_3 \geq 0$. Note that the scaling parameter s_k in Algorithm 4.1 does not appear in these conditions.

Subsequently, we discuss whether t_k satisfying the (strong/generalized) $\mathscr{T}^{(k)}$-Wolfe conditions exists. It is sufficient to show that t_k satisfying the generalized $\mathscr{T}^{(k)}$-Wolfe conditions (3.5) and (4.15) with $c_3 = 0$ exists because such t_k also satisfies the $\mathscr{T}^{(k)}$-Wolfe conditions (3.5) and (4.13), strong $\mathscr{T}^{(k)}$-Wolfe conditions (3.5) and (4.14), and generalized $\mathscr{T}^{(k)}$-Wolfe conditions (3.5) and (4.15) with any $c_3 \geq 0$.

If $\mathscr{T}^{(k)}(\eta_k) := \mathrm{D}R_{x_k}(t_k\eta_k)[\eta_k]$, then the (strong/generalized) $\mathscr{T}^{(k)}$-Wolfe conditions are reduced to the Riemannian (strong/generalized) Wolfe conditions, i.e., (4.13)–(4.15) simplify to (3.6)–(3.8), respectively. In particular, in this case, by defining $\phi_k(t) := f(R_{x_k}(t\eta_k))$, the generalized $\mathscr{T}^{(k)}$-Wolfe conditions (3.5) and (4.15) with $c_3 = 0$ are rewritten as $\phi_k(t_k) \leq \phi_k(0) + c_1 t_k \phi_k'(0)$ and $c_2\phi_k'(0) \leq \phi_k'(t_k) \leq 0$, respectively. Then, we can prove that $t_k > 0$ satisfying the two inequalities exists, similar to those in the Euclidean case. A complete proof for the Riemannian case is found in [51, Proposition 3.5]. If $\mathscr{T}^{(k)}(\eta_k) := -t_k^{-1}(R^{\mathrm{bw}}_{R_{x_k}(t_k\eta_k)})^{-1}(x_k)$, then the study on R-CG methods with inverse retraction [63] reveals that there exists t_k satisfying the generalized $\mathscr{T}^{(k)}$-Wolfe conditions (3.5) and (4.15) with $c_3 = 0$.

5. Assumptions and Zoutendijk's theorem. In this section, we discuss and summarize assumptions required for guaranteeing the global convergence of the R-CG methods. Although the proposed framework (Algorithm 4.1) is quite general, it is important to clarify the conditions with which the R-CG methods work appropriately. We have already discussed the conditions for $\mathscr{T}^{(k)}$ and t_k in Algorithm 4.1 in section 4. In subsection 5.1, we state the conditions imposed on Problem 2.1. Furthermore, we extend (the Riemannian version of) Zoutendijk's theorem to a theorem (Theorem 5.3) in the framework of Algorithm 4.1.

5.1. Assumptions for retraction and objective function. We assume the following condition on the objective function f.

Assumption 5.1. The Riemannian manifold $\mathcal{M}$ in Problem 2.1 is endowed with a retraction $R\colon T\mathcal{M} \to \mathcal{M}$. The objective function f in Problem 2.1 is of class C^1, bounded below on $\mathcal{M}$, i.e., there exists a constant $f_* \in \mathbb{R}$ such that $f(x) \geq f_*$ for all $x \in \mathcal{M}$, and satisfies the following condition:

(5.1) There exists a constant $L > 0$ such that, for all $x \in \mathcal{M}$, $\eta \in T_x\mathcal{M}$ with $\|\eta\|_x = 1$, and $t \geq 0$, it holds $|\mathrm{D}(f \circ R_x)(t\eta)[\eta] - \mathrm{D}(f \circ R_x)(0)[\eta]| \leq Lt$.

Furthermore, the norm of the gradient of f is upper bounded on the sublevel set $\{x \in \mathcal{M} \mid f(x) \leq f(x_0)\}$ for the initial point x_0 of Algorithm 4.1. This implies that there exists $L_g > 0$ such that $\|g_k\|_{x_k} \leq L_g$ if t_k in Algorithm 4.1 satisfies the Armijo condition (3.5) because (3.5) guarantees that $\{f(x_k)\}$ is monotonically nonincreasing.

Remark 5.2. Condition (5.1) is weaker than the condition that $\mathrm{grad}(f \circ R_x)$ is Lipschitz continuous for all $x \in \mathcal{M}$ with the same Lipschitz constant $L > 0$, i.e.,

$$\|\mathrm{grad}(f \circ R_x)(\xi) - \mathrm{grad}(f \circ R_x)(\eta)\|_x \leq L\|\xi - \eta\|_x \quad \text{for all } \xi, \eta \in T_x\mathcal{M} \tag{5.2}$$

holds for all $x \in \mathcal{M}$. Indeed, if (5.2) holds for all $x \in \mathcal{M}$, then (5.1) holds because we have, for any $\eta \in T_x\mathcal{M}$ with $\|\eta\|_x = 1$ and any $t \geq 0$,

$$\begin{aligned}|\mathrm{D}(f\circ R_x)(t\eta)[\eta]-\mathrm{D}(f\circ R_x)(0)[\eta]| &= |\langle \operatorname{grad}(f\circ R_x)(t\eta)-\operatorname{grad}(f\circ R_x)(0),\eta\rangle_x|\\ &\le \|\operatorname{grad}(f\circ R_x)(t\eta)-\operatorname{grad}(f\circ R_x)(0)\|_x\|\eta\|_x\\ &\le L\|t\eta\|_x\|\eta\|_x = Lt.\end{aligned}$$

Condition (5.1) is also closely related to the condition that $f\circ R$ is radially Lipschitz continuously differentiable [3, Definition 7.4.1], i.e., there exist real values $L>0$ and $\delta>0$ such that, for all $x\in\mathcal{M}$, $\eta\in T_x\mathcal{M}$ with $\|\eta\|_x=1$, and $t<\delta$, it holds that

$$\left|\frac{d}{d\tau}(f\circ R_x)(\tau\eta)|_{\tau=t}-\frac{d}{d\tau}(f\circ R_x)(\tau\eta)|_{\tau=0}\right|\le Lt.$$

Indeed, when $\delta=\infty$, this condition is equivalent to (5.1).

5.2. Riemannian version of Zoutendijk's theorem with $\mathscr{T}^{(k)}$. In Euclidean optimization, Zoutendijk's theorem plays an important role in analyzing various optimization algorithms (see, e.g., [44, Theorem 3.2]). Its Riemannian version is also discussed in [47, 51, 53]. They normally state a property for a sequence generated with step lengths that satisfy the Wolfe conditions. Here, to analyze the proposed R-CG methods with $\mathscr{T}^{(k)}$, we provide a similar theorem about the sequences generated by step lengths $t_k>0$, each of which satisfies the $\mathscr{T}^{(k)}$-Wolfe conditions. Note that the following result is not limited to the case of CG methods.

THEOREM 5.3. *Consider Problem* 2.1 *on a Riemannian manifold* $\mathcal{M}$ *with a retraction* R *and suppose Assumption* 5.1. *We also assume that* $\mathscr{T}^{(k)}$ *satisfies Assumption* 4.2 *and* t_k *satisfies the* $\mathscr{T}^{(k)}$*-Wolfe conditions* (3.5) *and* (4.13) *with* $0<c_1<c_2<1$ *for all* $k\ge 0$. *Let* $g_k:=\operatorname{grad} f(x_k)$. *If* $\langle g_k,\eta_k\rangle_{x_k}<0$ *for all* $k\ge 0$ *and there exists* $\mu>0$ *such that* $\|g_k\|_{x_k}\le\mu\|\eta_k\|_{x_k}$ *for all* $k\in K_2$, *then we have*

$$\sum_{k=0}^{\infty}\frac{\langle g_k,\eta_k\rangle_{x_k}^2}{\|\eta_k\|_{x_k}^2}<\infty, \tag{5.3}$$

where K_2 *is the subset of* $\mathbb{N}$ *in Assumption* 4.2.

Proof. The proof is done by combining the discussions in [51, Theorem 3.2] and [63, Theorem 3].

From the assumption as well as the triangle and Cauchy–Schwarz inequalities, we obtain

$$\begin{aligned}&(c_2-1)\langle g_k,\eta_k\rangle_{x_k}\overset{(4.13)}{\le}\langle g_{k+1},\mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}}-\langle g_k,\eta_k\rangle_{x_k}\\ &\le|\langle g_{k+1},\mathscr{T}^{(k)}(\eta_k)-\mathrm{D}R_{x_k}(t_k\eta_k)[\eta_k]\rangle_{x_{k+1}}|\\ &\quad+|\langle g_{k+1},\mathrm{D}R_{x_k}(t_k\eta_k)[\eta_k]\rangle_{x_{k+1}}-\langle g_k,\eta_k\rangle_{x_k}|\\ &=|\langle g_{k+1},\mathscr{T}^{(k)}(\eta_k)-\mathrm{D}R_{x_k}(t_k\eta_k)[\eta_k]\rangle_{x_{k+1}}|\\ &\quad+\|\eta_k\|_{x_k}\left|\mathrm{D}(f\circ R_{x_k})\left(t_k\|\eta_k\|_{x_k}\frac{\eta_k}{\|\eta_k\|_{x_k}}\right)\left[\frac{\eta_k}{\|\eta_k\|_{x_k}}\right]-\mathrm{D}(f\circ R_{x_k})(0)\left[\frac{\eta_k}{\|\eta_k\|_{x_k}}\right]\right|\\ &\overset{(5.1)}{\le}\|g_{k+1}\|_{x_{k+1}}\|\mathscr{T}^{(k)}(\eta_k)-\mathrm{D}R_{x_k}(t_k\eta_k)[\eta_k]\|_{x_{k+1}}+Lt_k\|\eta_k\|_{x_k}^2\\ &\overset{(4.3),(4.4)}{\le}\begin{cases}C\|g_{k+1}\|_{x_{k+1}}t_k\|\eta_k\|_{x_k}^2+Lt_k\|\eta_k\|_{x_k}^2, & k\in K_1,\\ C\|g_{k+1}\|_{x_{k+1}}(t_k+t_k^2)\|\eta_k\|_{x_k}^2+Lt_k\|\eta_k\|_{x_k}^2, & k\in K_2.\end{cases}\end{aligned}$$

Therefore, we obtain

$$t_k \geq -\frac{1-c_2}{C\|g_{k+1}\|_{x_{k+1}}+L}\frac{\langle g_k,\eta_k\rangle_{x_k}}{\|\eta_k\|_{x_k}^2} \geq -\frac{1-c_2}{CL_g+L}\frac{\langle g_k,\eta_k\rangle_{x_k}}{\|\eta_k\|_{x_k}^2}, \qquad k\in K_1. \tag{5.4}$$

For $k\in K_2$, we have

$$C\|g_{k+1}\|_{x_{k+1}}\|\eta_k\|_{x_k}^2 t_k^2 + (L+C\|g_{k+1}\|_{x_{k+1}})\|\eta_k\|_{x_k}^2 t_k + (1-c_2)\langle g_k,\eta_k\rangle_{x_k} \geq 0.$$

It follows from $t_k>0$ and $(1-c_2)\langle g_k,\eta_k\rangle_{x_k}<0$ that

$$t_k \geq -\frac{2(1-c_2)}{u_k}\frac{\langle g_k,\eta_k\rangle_{x_k}}{\|\eta_k\|_{x_k}^2}, \qquad k\in K_2,$$

where

$$\begin{aligned} u_k &:= L+C\|g_{k+1}\|_{x_{k+1}} + \sqrt{(L+C\|g_{k+1}\|_{x_{k+1}})^2 - 4C(1-c_2)\|g_{k+1}\|_{x_{k+1}}\frac{\langle g_k,\eta_k\rangle_{x_k}}{\|\eta_k\|_{x_k}^2}} \\ &\leq L+CL_g+\sqrt{(L+CL_g)^2+4C(1-c_2)L_g\mu} =: u. \end{aligned}$$

Here, we used $\|g_{k+1}\|_{x_{k+1}}\leq L_g$ and $-\langle g_k,\eta_k\rangle_{x_k}\leq\|g_k\|_{x_k}\|\eta_k\|_{x_k}\leq\mu\|\eta_k\|_{x_k}^2$ from the assumption $\|g_k\|_{x_k}\leq\mu\|\eta_k\|_{x_k}$ for $k\in K_2$. Using the constant $u>0$, we obtain

$$t_k \geq -\frac{2(1-c_2)}{u}\frac{\langle g_k,\eta_k\rangle_{x_k}}{\|\eta_k\|_{x_k}^2}, \qquad k\in K_2. \tag{5.5}$$

With the constant $U:=\min\{(1-c_2)/(CL_g+L),2(1-c_2)/u\}>0$, (5.4) and (5.5) yield $t_k\geq -U\langle g_k,\eta_k\rangle_{x_k}/\|\eta_k\|_{x_k}^2$ for all $k\geq 0$. This and (3.5) yield

$$f(x_{k+1}) \leq f(x_k) - c_1U\frac{\langle g_k,\eta_k\rangle_{x_k}^2}{\|\eta_k\|_{x_k}^2} \leq f(x_0) - c_1U\sum_{j=0}^{k}\frac{\langle g_j,\eta_j\rangle_{x_j}^2}{\|\eta_j\|_{x_j}^2}.$$

It follows from Assumption 5.1 that $f(x)\geq f_*$ for all $x\in\mathcal{M}$. Therefore, we have

$$\sum_{j=0}^{k}\frac{\langle g_j,\eta_j\rangle_{x_j}^2}{\|\eta_j\|_{x_j}^2} \leq \frac{f(x_0)-f(x_{k+1})}{c_1U} \leq \frac{f(x_0)-f_*}{c_1U}\ (=\text{const.}).$$

Taking the limit $k\to\infty$, we obtain the desired result. This completes the proof. ∎

Remark 5.4. In Theorem 5.3, we assume that there exists $\mu>0$ such that $\|g_k\|_{x_k}\leq\mu\|\eta_k\|_{x_k}$ for all $k\in K_2$. This assumption does not appear when we discuss Zoutendijk's theorem with the standard Wolfe conditions (3.5) and (3.6) [47, 51], i.e., the case of $\mathscr{T}^{(k)}(\eta_k)=\mathcal{T}^R_{t_k\eta_k}(\eta_k)$. This is because we can take $K_1=\mathbb{N}$ and $K_2=\emptyset$ for this case. Thus, this assumption may seem restrictive. However, fortunately, in the subsequent analyses of our R-CG methods, we realize that this assumption automatically holds (requiring the generalized $\mathscr{T}^{(k)}$-Wolfe conditions with $c_3=0$ for the CD-type in subsection 6.1.3). Therefore, we do not need to explicitly suppose this assumption in the R-CG methods, and Theorem 5.3 is still a powerful tool, even for general $\mathscr{T}^{(k)}$.

Practically, we consider taking K_1 as large as possible in Assumption 4.2. Therefore, K_2 is expected to consist of only those values of k for which (4.3) does not (or is not shown to) hold. Taking K_2 as small as possible, we require the condition $\|g_k\|_{x_k}\leq\mu\|\eta_k\|_{x_k}$ for as small number of k as possible in Theorem 5.3.

6. Convergence analyses of the R-CG methods. Considering the theoretical convergence properties, the quantity in the numerator of the formulas for β_{k+1} is influential. In what follows, we divide the six types of β_{k+1} into two categories, depending on their numerators. One consists of $\beta_{k+1}^{\text{R-FR}}$, $\beta_{k+1}^{\text{R-DY}}$, and $\beta_{k+1}^{\text{R-CD}}$ with the numerator $\|g_{k+1}\|_{x_{k+1}}^2$, and the other consists of $\beta_{k+1}^{\text{R-PRP}}$, $\beta_{k+1}^{\text{R-HS}}$, and $\beta_{k+1}^{\text{R-LS}}$ with the numerator $\|g_{k+1}\|_{x_{k+1}}^2 - \langle g_{k+1}, s_k \mathscr{S}^{(k)}(g_k)\rangle_{x_{k+1}}$.

6.1. Global convergence analyses of the R-CG methods with $\boldsymbol{\beta}^{\text{R-FR}}$, $\boldsymbol{\beta}^{\text{R-DY}}$, and $\boldsymbol{\beta}^{\text{R-CD}}$. We prove the global convergence properties of the R-CG methods with $\beta_{k+1}^{\text{R-FR}}$, $\beta_{k+1}^{\text{R-DY}}$, and $\beta_{k+1}^{\text{R-CD}}$, defined in (4.7)–(4.9), and with some other related β_{k+1}. In the subsequent analyses, a key property is that the algorithms with appropriately chosen step lengths satisfy the sufficient descent condition, i.e., there exists a constant $c > 0$ such that $\langle g_k, \eta_k\rangle_{x_k} \le -c\|g_k\|_{x_k}^2$ holds for all $k \ge 0$.

6.1.1. R-CG methods based on $\boldsymbol{\beta}^{\text{R-FR}}$. We recall that $\beta_{k+1}^{\text{R-FR}}$ is defined in (4.7) as $\beta_{k+1}^{\text{R-FR}} := \|g_{k+1}\|_{x_{k+1}}^2/\|g_k\|_{x_k}^2$. The convergence analysis of the R-CG method with $\beta_{k+1} = \beta_{k+1}^{\text{R-FR}}$ can be completed following the standard discussion in the existing ones (e.g., [51, section 4.4]). Here, we provide an analysis for the more general class of β_{k+1}, i.e., β_{k+1} satisfying $|\beta_{k+1}| \le \beta_{k+1}^{\text{R-FR}}$. This generalization can be used to develop other β_{k+1} based on $\beta_{k+1}^{\text{R-FR}}$ (see, e.g., subsection 6.2.1). We first show that β_{k+1} satisfying $|\beta_{k+1}| \le \beta_{k+1}^{\text{R-FR}}$ guarantees sufficient descent directions and that the ratio $\|g_k\|_{x_k}/\|\eta_k\|_{x_k}$ is bounded above.

PROPOSITION 6.1. *Let sequence $\{x_k\}$ be generated by Algorithm 4.1 with β_{k+1} satisfying $|\beta_{k+1}| \le \beta_{k+1}^{\text{R-FR}}$, where $\beta_{k+1}^{\text{R-FR}}$ is defined as (4.7). If, for all $k \ge 0$, $g_k := \operatorname{grad} f(x_k) \ne 0$ and step lengths t_k satisfy the strong $\mathscr{T}^{(k)}$-Wolfe conditions (3.5) and (4.14) with $0 < c_1 < c_2 < 1/2$, then we have*

$$-\frac{1}{1-c_2} \le \frac{\langle g_k, \eta_k\rangle_{x_k}}{\|g_k\|_{x_k}^2} \le -\frac{1-2c_2}{1-c_2} \tag{6.1}$$

and

$$\|g_k\|_{x_k} \le \frac{1-c_2}{1-2c_2}\|\eta_k\|_{x_k}. \tag{6.2}$$

Proof. We define $a_k := \langle g_k, \eta_k\rangle_{x_k}/\|g_k\|_{x_k}^2$ and prove (6.1) by induction. For $k = 0$, (6.1) clearly holds since $a_0 = -1$ from $\eta_0 = -g_0$. Subsequently, assume that (6.1) is true for some $k \ge 0$. Then, (4.1) yields

$$a_{k+1} = \frac{\langle g_{k+1}, -g_{k+1} + \beta_{k+1} s_k \mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}}}{\|g_{k+1}\|_{x_{k+1}}^2} = -1 + \beta_{k+1} s_k \frac{\langle g_{k+1}, \mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}}}{\|g_{k+1}\|_{x_{k+1}}^2}.$$

Here, (4.14) implies $|\langle g_{k+1}, \mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}}| \le -c_2\langle g_k, \eta_k\rangle_{x_k}$. Considering $0 < s_k \le 1$ from (4.2) and $|\beta_{k+1}| \le \beta_{k+1}^{\text{R-FR}} = \|g_{k+1}\|_{x_{k+1}}^2/\|g_k\|_{x_k}^2$, we obtain

$$|a_{k+1} + 1| \le \frac{|\beta_{k+1}|}{\|g_{k+1}\|_{x_{k+1}}^2}|\langle g_{k+1}, \mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}}| \le -c_2\frac{\langle g_k, \eta_k\rangle_{x_k}}{\|g_k\|_{x_k}^2} = -c_2 a_k,$$

indicating $-1 + c_2 a_k \le a_{k+1} \le -1 - c_2 a_k$. Since (6.1) yields $a_k \ge -1/(1-c_2)$, we have $-1/(1-c_2) \le a_{k+1} \le -(1-2c_2)/(1-c_2)$, i.e., (6.1) also holds if k is replaced with $k+1$. This ends the proof of (6.1) for all $k \ge 0$.

It follows from the Cauchy–Schwarz inequality $\langle g_k, \eta_k\rangle_{x_k} \ge -\|g_k\|_{x_k}\|\eta_k\|_{x_k}$ that $a_k \ge -\|\eta_k\|_{x_k}/\|g_k\|_{x_k}$, which, together with (6.1), yields (6.2). ∎

Using this theorem, we show the global convergence property of R-CG methods with β_{k+1} satisfying $\beta_{k+1} \leq |\beta_{k+1}^{\text{R-FR}}|$.

THEOREM 6.2. *Under Assumption* 5.1, *let sequence* $\{x_k\}$ *be generated by Algorithm* 4.1 *with* β_{k+1} *satisfying* $|\beta_{k+1}| \leq \beta_{k+1}^{\text{R-FR}}$, *where* $\beta_{k+1}^{\text{R-FR}}$ *is defined as* (4.7). *If* $\mathscr{T}^{(k)}$ *satisfies Assumption* 4.2 *and the step lengths satisfy the strong* $\mathscr{T}^{(k)}$*-Wolfe conditions* (3.5) *and* (4.14) *with* $0 < c_1 < c_2 < 1/2$, *then for* $g_k := \operatorname{grad} f(x_k)$, *we have*

$$\liminf_{k\to\infty} \|g_k\|_{x_k} = 0. \tag{6.3}$$

Proof. If $g_{k_0} = 0$ holds for some $k_0 \geq 0$, then (4.1) and (4.7) imply that $g_k = 0$ for all $k \geq k_0$; thus, (6.3) holds.

Subsequently, we assume $g_k \neq 0$ for all $k \geq 0$ and prove (6.3) by contradiction. To this end, we assume that (6.3) does not hold, indicating that there exists $\varepsilon > 0$ such that $\|g_k\|_{x_k} \geq \varepsilon$ for all $k \geq 0$. Furthermore, since the assumption in Proposition 6.1 holds, we have (6.1) and (6.2). Hence, the assumption in Theorem 5.3 is also ensured to imply (5.3). With $c := (1+c_2)/(1-c_2)$, we can evaluate $\|\eta_{k+1}\|_{x_{k+1}}^2$ as

$$\begin{aligned}
\|\eta_{k+1}\|_{x_{k+1}}^2 &\overset{(4.1)}{=} \|g_{k+1}\|_{x_{k+1}}^2 - 2\beta_{k+1}s_k\langle g_{k+1}, \mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}} + \beta_{k+1}^2 s_k^2 \|\mathscr{T}^{(k)}(\eta_k)\|_{x_{k+1}}^2 \\
&\overset{(4.2)}{\leq} \|g_{k+1}\|_{x_{k+1}}^2 + 2|\beta_{k+1}||\langle g_{k+1}, \mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}}| + |\beta_{k+1}|^2\|\eta_k\|_{x_k}^2 \\
&\overset{(4.14)}{\leq} \|g_{k+1}\|_{x_{k+1}}^2 - 2c_2\beta_{k+1}^{\text{R-FR}}\langle g_k, \eta_k\rangle_{x_k} + (\beta_{k+1}^{\text{R-FR}})^2\|\eta_k\|_{x_k}^2 \\
&\overset{(6.1)}{\leq} \|g_{k+1}\|_{x_{k+1}}^2 + \frac{2c_2}{1-c_2}\beta_{k+1}^{\text{R-FR}}\|g_k\|_{x_k}^2 + (\beta_{k+1}^{\text{R-FR}})^2\|\eta_k\|_{x_k}^2 \\
&\overset{(4.7)}{=} c\|g_{k+1}\|_{x_{k+1}}^2 + (\beta_{k+1}^{\text{R-FR}})^2\|\eta_k\|_{x_k}^2.
\end{aligned}$$

This recurrence relation together with $\|\eta_0\|_{x_0} = \|g_0\|_{x_0}$ and $c > 1$ gives

$$\begin{aligned}
\|\eta_k\|_{x_k}^2 &\leq c\left(\|g_k\|_{x_k}^2 + \sum_{j=1}^{k-1}(\beta_k^{\text{R-FR}})^2(\beta_{k-1}^{\text{R-FR}})^2\cdots(\beta_{j+1}^{\text{R-FR}})^2\|g_j\|_{x_j}^2\right) \\
&\quad + (\beta_k^{\text{R-FR}})^2(\beta_{k-1}^{\text{R-FR}})^2\cdots(\beta_1^{\text{R-FR}})^2\|\eta_0\|_{x_0}^2 \\
&\overset{(4.7)}{<} c\|g_k\|_{x_k}^4 \sum_{j=0}^{k}\|g_j\|_{x_j}^{-2} \leq \frac{c}{\varepsilon^2}\|g_k\|_{x_k}^4(k+1).
\end{aligned}$$

Therefore, using (6.1) again, we obtain

$$\sum_{j=0}^{k}\frac{\langle g_j, \eta_j\rangle_{x_j}^2}{\|\eta_j\|_{x_j}^2} > \frac{\varepsilon^2}{c}\sum_{j=0}^{k}\frac{\langle g_j, \eta_j\rangle_{x_j}^2}{\|g_j\|_{x_j}^4}\frac{1}{j+1} \geq \frac{\varepsilon^2(1-2c_2)^2}{c(1-c_2)^2}\sum_{j=1}^{k+1}\frac{1}{j} \to \infty \quad (k\to\infty).$$

Thus, the left-hand side also diverges to infinity. This contradicts (5.3), ending the proof. ∎

The global convergence property is ensured for Algorithm 4.1 with $\beta_{k+1} = \beta_{k+1}^{\text{R-FR}}$ as a corollary of Theorem 6.2. Because we have analyzed a class of β_{k+1}, rather than the specific $\beta_{k+1}^{\text{R-FR}}$ only, we can apply the results here to other R-CG methods such as one with $\beta_{k+1}^{\text{R-CD}}$ (see subsection 6.1.3).

6.1.2. R-CG methods based on $\beta^{\text{R-DY}}$. In this subsection, we give a global convergence analysis of the R-CG methods with a class of β_{k+1} containing $\beta_{k+1}^{\text{R-DY}} := \|g_{k+1}\|_{x_{k+1}}^2/(\langle g_{k+1}, s_k\mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}} - \langle g_k, \eta_k\rangle_{x_k})$ in (4.8). We first show that such R-CG methods generate descent search directions and that, with additional assumptions, the search directions are sufficient descent directions. Thereafter, we build a global convergence result. Proposition 6.3 and Theorem 6.4 are inspired by, but are more general than, the results in [51, section 4.5] and [63, section 4.3]. Therefore, their proofs are not verbatim.

PROPOSITION 6.3. *Let sequence $\{x_k\}$ be generated by Algorithm 4.1 with β_{k+1} satisfying $0 \le \beta_{k+1} \le \beta_{k+1}^{\text{R-DY}}$, where $\beta_{k+1}^{\text{R-DY}}$ is defined as (4.8). Assume that, for all $k \ge 0$, $g_k := \operatorname{grad} f(x_k) \ne 0$ and the step lengths t_k satisfy the $\mathscr{T}^{(k)}$-Wolfe conditions (3.5) and (4.13) with $0 < c_1 < c_2 < 1$. Then, the algorithm is well defined, i.e., $\beta_{k+1}^{\text{R-DY}} > 0$ holds, and β_{k+1} satisfying $0 \le \beta_{k+1} \le \beta_{k+1}^{\text{R-DY}}$ exists for every $k \ge 0$, and we have*

$$\langle g_k, \eta_k\rangle_{x_k} < \min\{0, \langle g_{k+1}, s_k\mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}}\}. \tag{6.4}$$

Furthermore, if, for an arbitrary $k \ge 1$[6], t_{k-1} satisfies the generalized $\mathscr{T}^{(k-1)}$-Wolfe conditions (3.5) and (4.15) with $0 < c_1 < c_2 < 1$ and $c_3 \ge 0$, then for this k, it holds that

$$-\frac{1}{1-c_2} \le \frac{\langle g_k, \eta_k\rangle_{x_k}}{\|g_k\|_{x_k}^2} \le -\frac{1}{1+c_3} \tag{6.5}$$

and

$$\|g_k\|_{x_k} \le (1+c_3)\|\eta_k\|_{x_k}. \tag{6.6}$$

Proof. We first prove $\beta_{k+1}^{\text{R-DY}} > 0$ and (6.4) by induction.

For $k = 0$, $\langle g_0, \eta_0\rangle_{x_0} = -\|g_0\|_{x_0}^2 < 0$. If $\langle g_1, \mathscr{T}^{(0)}(\eta_0)\rangle_{x_1} \ge 0$, then (6.4) holds from $s_0 > 0$. Otherwise, from $\langle g_1, \mathscr{T}^{(0)}(\eta_0)\rangle_{x_1}, \langle g_0, \eta_0\rangle_{x_0} < 0$, $s_0 \le 1$, $c_2 < 1$, and (4.13), we have $\langle g_1, s_0\mathscr{T}^{(0)}(\eta_0)\rangle_{x_1} \ge \langle g_1, \mathscr{T}^{(0)}(\eta_0)\rangle_{x_1} \ge c_2\langle g_0, \eta_0\rangle_{x_0} > \langle g_0, \eta_0\rangle_{x_0}$, indicating that (6.4) holds. Moreover, (6.4) directly ensures $\beta_1^{\text{R-DY}} > 0$.

Now assume that, for some $k \ge 0$, $\beta_{k+1}^{\text{R-DY}} > 0$ and (6.4) hold. Then, β_{k+1} satisfying $0 \le \beta_{k+1} \le \beta_{k+1}^{\text{R-DY}}$ exists. If $0 < \beta_{k+1} \le \beta_{k+1}^{\text{R-DY}}$, we obtain

$$\begin{aligned}
\langle g_{k+1}, \eta_{k+1}\rangle_{x_{k+1}} &\overset{(4.1)}{=} -\|g_{k+1}\|_{x_{k+1}}^2 + \beta_{k+1}\langle g_{k+1}, s_k\mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}} \\
&= -\|g_{k+1}\|_{x_{k+1}}^2 + \beta_{k+1}(\langle g_{k+1}, s_k\mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}} - \langle g_k, \eta_k\rangle_{x_k}) + \beta_{k+1}\langle g_k, \eta_k\rangle_{x_k} \\
&\overset{(6.4)}{\le} -\|g_{k+1}\|_{x_{k+1}}^2 + \beta_{k+1}^{\text{R-DY}}(\langle g_{k+1}, s_k\mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}} - \langle g_k, \eta_k\rangle_{x_k}) + \beta_{k+1}\langle g_k, \eta_k\rangle_{x_k} \\
&\overset{(4.8)}{=} \beta_{k+1}\langle g_k, \eta_k\rangle_{x_k} < 0.
\end{aligned}$$

If $\beta_{k+1} = 0$, then we have $\langle g_{k+1}, \eta_{k+1}\rangle_{x_{k+1}} = -\|g_{k+1}\|_{x_{k+1}}^2 < 0$. We can also prove $\langle g_{k+1}, \eta_{k+1}\rangle_{x_{k+1}} < \langle g_{k+2}, s_{k+1}\mathscr{T}^{(k+1)}(\eta_{k+1})\rangle_{x_{k+2}}$ as in the previous paragraph. Therefore, (6.4) holds if k is replaced with $k+1$. Hence, (6.4) is proved for all $k \ge 0$.

We proceed to prove (6.5) and (6.6) for any $k \ge 1$ with which t_{k-1} satisfies the generalized $\mathscr{T}^{(k-1)}$-Wolfe conditions (3.5) and (4.15). It follows from (4.1) that

$$\langle g_k, \eta_k\rangle_{x_k} = -\|g_k\|_{x_k}^2 + \beta_k\langle g_k, s_{k-1}\mathscr{T}^{(k-1)}(\eta_{k-1})\rangle_{x_k}. \tag{6.7}$$

[6] For $k = 0$, (6.5) and (6.6) clearly hold without any assumption since $\eta_0 = -g_0$.

We now prove the first inequality in (6.5). Here, from (4.15) and $s_k \leq 1$, we observe that $\langle g_k, s_{k-1}\mathscr{T}^{(k-1)}(\eta_{k-1})\rangle_{x_k} \geq s_{k-1}c_2\langle g_{k-1}, \eta_{k-1}\rangle_{x_{k-1}} \geq c_2\langle g_{k-1}, \eta_{k-1}\rangle_{x_{k-1}}$. Thus, it follows from (6.7), $0 \leq \beta_k \leq \beta_k^{\text{R-DY}}$, (4.8), (4.15), and $c_2\langle g_{k-1}, \eta_{k-1}\rangle_{x_{k-1}} < 0$ that

$$\begin{aligned}
\langle g_k, \eta_k\rangle_{x_k} &\geq -\|g_k\|_{x_k}^2 + c_2\beta_k\langle g_{k-1}, \eta_{k-1}\rangle_{x_{k-1}} \\
&\geq -\|g_k\|_{x_k}^2 + c_2\beta_k^{\text{R-DY}}\langle g_{k-1}, \eta_{k-1}\rangle_{x_{k-1}} \\
&= \|g_k\|_{x_k}^2\left(-1 + \frac{c_2\langle g_{k-1}, \eta_{k-1}\rangle_{x_{k-1}}}{\langle g_k, s_{k-1}\mathscr{T}^{(k-1)}(\eta_{k-1})\rangle_{x_k} - \langle g_{k-1}, \eta_{k-1}\rangle_{x_{k-1}}}\right) \\
&\geq \|g_k\|_{x_k}^2\left(-1 + \frac{c_2\langle g_{k-1}, \eta_{k-1}\rangle_{x_{k-1}}}{c_2\langle g_{k-1}, \eta_{k-1}\rangle_{x_{k-1}} - \langle g_{k-1}, \eta_{k-1}\rangle_{x_{k-1}}}\right) = \frac{\|g_k\|_{x_k}^2}{c_2 - 1}.
\end{aligned}$$

To prove the second inequality in (6.5), we consider the following two cases. If $\langle g_k, s_{k-1}\mathscr{T}^{(k-1)}(\eta_{k-1})\rangle_{x_k} \leq 0$, then (6.7) and $\beta_k \geq 0$ yield

$$\langle g_k, \eta_k\rangle_{x_k} \leq -\|g_k\|_{x_k}^2 \leq -\frac{\|g_k\|_{x_k}^2}{1 + c_3}.$$

Otherwise (i.e., if $\langle g_k, s_{k-1}\mathscr{T}^{(k-1)}(\eta_{k-1})\rangle_{x_k} > 0$), (6.7), $\beta_k \leq \beta_k^{\text{R-DY}}$, and (4.8) give

$$\begin{aligned}
\langle g_k, \eta_k\rangle_{x_k} &\leq -\|g_k\|_{x_k}^2 + \beta_k^{\text{R-DY}}\langle g_k, s_{k-1}\mathscr{T}^{(k-1)}(\eta_{k-1})\rangle_{x_k} \\
&= \frac{\|g_k\|_{x_k}^2\langle g_{k-1}, \eta_{k-1}\rangle_{x_{k-1}}}{\langle g_k, s_{k-1}\mathscr{T}^{(k-1)}(\eta_{k-1})\rangle_{x_k} - \langle g_{k-1}, \eta_{k-1}\rangle_{x_{k-1}}}.
\end{aligned}$$

Noting $\langle g_{k-1}, \eta_{k-1}\rangle_{x_{k-1}} < 0$ and $\langle g_k, s_{k-1}\mathscr{T}^{(k-1)}(\eta_{k-1})\rangle_{x_k} \leq -c_3\langle g_{k-1}, \eta_{k-1}\rangle_{x_{k-1}}$ from (4.15), we obtain

$$\langle g_k, \eta_k\rangle_{x_k} \leq \frac{\|g_k\|_{x_k}^2}{-c_3 - 1},$$

indicating that the second inequality in (6.5) always holds. Finally, (6.6) is a direct consequence from (6.5) and the Cauchy–Schwarz inequality, completing the proof. ∎

THEOREM 6.4. *Under Assumption* 5.1, *let sequence* $\{x_k\}$ *be generated by Algorithm* 4.1 *with* $\mathscr{T}^{(k)}$ *satisfying Assumption* 4.2. *We assume that* β_{k+1} *in the algorithm satisfies* $0 \leq \beta_{k+1} \leq \beta_{k+1}^{\text{R-DY}}$ *for all* $k \geq 0$ *and* t_k *satisfies the* $\mathscr{T}^{(k)}$*-Wolfe conditions* (3.5) *and* (4.13) *with* $0 < c_1 < c_2 < 1$. *Furthermore, assume that for* $k \in K_2\backslash\{0\}$, t_{k-1} *satisfies the generalized* $\mathscr{T}^{(k-1)}$*-Wolfe conditions* (3.5) *and* (4.15) *with* $0 < c_1 < c_2 < 1$ *and* $c_3 \geq 0$, *where* $\beta_{k+1}^{\text{R-DY}}$ *is defined as* (4.8), *and* K_2 *is the index set in Assumption* 4.2. *Then, for* $g_k := \operatorname{grad} f(x_k)$, *we have*

$$\liminf_{k\to\infty} \|g_k\|_{x_k} = 0. \tag{6.8}$$

Proof. It is sufficient to show (6.8) for the case $g_k \neq 0$ for all $k \geq 0$. From (4.1), we have $\eta_{k+1} + g_{k+1} = \beta_{k+1}s_k\mathscr{T}^{(k)}(\eta_k)$. Taking the norm and squaring, we obtain

$$\|\eta_{k+1}\|_{x_{k+1}}^2 = \beta_{k+1}^2 s_k^2\|\mathscr{T}^{(k)}(\eta_k)\|_{x_{k+1}}^2 - 2\langle g_{k+1}, \eta_{k+1}\rangle_{x_{k+1}} - \|g_{k+1}\|_{x_{k+1}}^2. \tag{6.9}$$

From the proof of Proposition 6.3, for all $k \geq 0$, we can confirm the inequality $\langle g_{k+1}, \eta_{k+1}\rangle_{x_{k+1}} \leq \beta_{k+1}\langle g_k, \eta_k\rangle_{x_k} \leq 0$ (the two equal signs do not hold simultaneously), leading to $\beta_{k+1}^2\langle g_k, \eta_k\rangle_{x_k}^2 \leq \langle g_{k+1}, \eta_{k+1}\rangle_{x_{k+1}}^2$. Dividing both sides of (6.9) by $\langle g_{k+1}, \eta_{k+1}\rangle_{x_{k+1}}^2 > 0$, we obtain

$$\begin{aligned}\frac{\|\eta_{k+1}\|_{x_{k+1}}^2}{\langle g_{k+1},\eta_{k+1}\rangle_{x_{k+1}}^2} &\le \frac{s_k^2\|\mathscr{T}^{(k)}(\eta_k)\|_{x_{k+1}}^2}{\langle g_k,\eta_k\rangle_{x_k}^2} - \frac{2}{\langle g_{k+1},\eta_{k+1}\rangle_{x_{k+1}}} - \frac{\|g_{k+1}\|_{x_{k+1}}^2}{\langle g_{k+1},\eta_{k+1}\rangle_{x_{k+1}}^2}\\ &\overset{(4.2)}{\le} \frac{\|\eta_k\|_{x_k}^2}{\langle g_k,\eta_k\rangle_{x_k}^2} + \frac{1}{\|g_{k+1}\|_{x_{k+1}}^2} - \left(\frac{1}{\|g_{k+1}\|_{x_{k+1}}} + \frac{\|g_{k+1}\|_{x_{k+1}}}{\langle g_{k+1},\eta_{k+1}\rangle_{x_{k+1}}}\right)^2\\ &\le \frac{\|\eta_k\|_{x_k}^2}{\langle g_k,\eta_k\rangle_{x_k}^2} + \frac{1}{\|g_{k+1}\|_{x_{k+1}}^2}. \qquad (6.10)\end{aligned}$$

To accomplish the proof by contradiction, we assume $\liminf_{k\to\infty}\|g_k\|_{x_k} > 0$, which, together with $g_k \neq 0$ for all $k \ge 0$, implies that there exists $\varepsilon > 0$ such that $\|g_k\|_{x_k} \ge \varepsilon$ for all $k \ge 0$. Therefore, from (6.10), we obtain

$$\frac{\|\eta_k\|_{x_k}^2}{\langle g_k,\eta_k\rangle_{x_k}^2} \le \frac{\|\eta_0\|_{x_0}^2}{\langle g_0,\eta_0\rangle_{x_0}^2} + \sum_{j=1}^{k}\frac{1}{\|g_j\|_{x_j}^2} = \sum_{j=0}^{k}\frac{1}{\|g_j\|_{x_j}^2} \le \frac{k+1}{\varepsilon^2},$$

which gives

$$\sum_{k=0}^{N}\frac{\langle g_k,\eta_k\rangle_{x_k}^2}{\|\eta_k\|_{x_k}^2} \ge \varepsilon^2\sum_{k=1}^{N+1}\frac{1}{k} \to \infty \quad (N\to\infty). \qquad (6.11)$$

On the other hand, Proposition 6.3 indicates that the assumption in Theorem 5.3 holds, and we have (5.3), contradicting (6.11). Therefore, (6.8) must hold. ∎

The DY-type R-CG methods have an advantage over the FR-type in that they do not require the strong $\mathscr{T}^{(k)}$-Wolfe conditions. Although the generalized $\mathscr{T}^{(k)}$-Wolfe conditions are required for k such that (4.3) does not hold, the constant $c_3 \ge 0$ can be taken as any large constant. Therefore, the conditions are not too restrictive and are much weaker than the strong $\mathscr{T}^{(k)}$-Wolfe conditions.

6.1.3. R-CG methods based on $\beta^{\text{R-CD}}$. The following proposition provides important properties of $\beta_{k+1}^{\text{R-CD}} := \|g_{k+1}\|_{x_{k+1}}^2/(-\langle g_k,\eta_k\rangle_{x_k})$ defined as (4.9).

PROPOSITION 6.5. *Let sequence $\{x_k\}$ be generated by Algorithm 4.1 with β_{k+1} satisfying $0 \le \beta_{k+1} \le \beta_{k+1}^{\text{R-CD}}$, where $\beta_{k+1}^{\text{R-CD}}$ is defined as (4.9). Assume that, for all $k \ge 0$, $g_k := \operatorname{grad} f(x_k) \neq 0$ and step lengths t_k satisfy the $\mathscr{T}^{(k)}$-generalized Wolfe conditions (3.5) and (4.15) with $0 < c_1 < c_2 < 1$ and $c_3 = 0$. Then, the algorithm is well defined, i.e., $\beta_{k+1}^{\text{R-CD}} > 0$ holds, and β_{k+1} satisfying $0 \le \beta_{k+1} \le \beta_{k+1}^{\text{R-CD}}$ exists for every $k \ge 0$. Furthermore, we have the sufficient descent condition on η_k as $\langle g_k,\eta_k\rangle_{x_k} \le -\|g_k\|_{x_k}^2$ and $0 \le \beta_{k+1} \le \beta_{k+1}^{\text{R-FR}}$ for all $k \ge 0$, where $\beta_{k+1}^{\text{R-FR}}$ is defined as (4.7). In particular, $0 < \beta_{k+1}^{\text{R-CD}} \le \beta_{k+1}^{\text{R-FR}}$ holds for all $k \ge 0$.*

Proof. We first show $\langle g_k,\eta_k\rangle_{x_k} \le -\|g_k\|_{x_k}^2$ for all $k \ge 0$ by induction. For any $k \ge 0$, if this inequality holds, then $\beta_{k+1}^{\text{R-CD}} > 0$ directly follows from $\langle g_k,\eta_k\rangle_{x_k} < 0$, and $\beta_{k+1} \in [0, \beta_{k+1}^{\text{R-CD}}]$ actually exists. For $k = 0$, it is clear that $\langle g_0,\eta_0\rangle_{x_0} = -\|g_0\|_{x_0}^2$. Subsequently, we assume that $\langle g_k,\eta_k\rangle_{x_k} \le -\|g_k\|_{x_k}^2$ (< 0) for some $k \ge 0$. Then, considering $\beta_{k+1}^{\text{R-CD}} > 0$ and the inequality $\langle g_{k+1}, \mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}} \le 0$ from (4.15) with $c_3 = 0$, we use (4.1) and (4.9) to obtain

$$\begin{aligned}\frac{\langle g_{k+1},\eta_{k+1}\rangle_{x_{k+1}}}{\|g_{k+1}\|_{x_{k+1}}^2} &= \frac{\langle g_{k+1}, -g_{k+1} + \beta_{k+1}s_k\mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}}}{\|g_{k+1}\|_{x_{k+1}}^2}\\ &= -1 - s_k\frac{\beta_{k+1}}{\beta_{k+1}^{\text{R-CD}}}\frac{\langle g_{k+1}, \mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}}}{\langle g_k,\eta_k\rangle_{x_k}} \le -1.\end{aligned}$$

Thus, $\langle g_{k+1}, \eta_{k+1}\rangle_{x_{k+1}} \leq -\|g_{k+1}\|^2_{x_{k+1}}$ holds as desired, and by induction, we have $\langle g_k, \eta_k\rangle_{x_k} \leq -\|g_k\|^2_{x_k}$ for all $k \geq 0$.

Furthermore, this result directly leads to

$$\beta^{\text{R-CD}}_{k+1} = \frac{\|g_{k+1}\|^2_{x_{k+1}}}{-\langle g_k, \eta_k\rangle_{x_k}} \leq \frac{\|g_{k+1}\|^2_{x_{k+1}}}{\|g_k\|^2_{x_k}} = \beta^{\text{R-FR}}_{k+1}.$$

Combining this relationship with $0 \leq \beta_{k+1} \leq \beta^{\text{R-CD}}_{k+1}$ yields $0 \leq \beta_{k+1} \leq \beta^{\text{R-FR}}_{k+1}$. ∎

As a special case of the result in Proposition 6.5, we have $|\beta^{\text{R-CD}}_{k+1}| \leq \beta^{\text{R-FR}}_{k+1}$ when we choose $\beta_{k+1} \equiv \beta^{\text{R-CD}}_{k+1}$ for all $k \geq 0$. Therefore, as a corollary of Theorem 6.2, we obtain the convergence result for $\beta^{\text{R-CD}}_{k+1}$, requiring t_k to satisfy the generalized $\mathscr{T}^{(k)}$-Wolfe conditions with $0 < c_1 < c_2 < 1/2$ and $c_3 = 0$. Furthermore, in fact, as the following theorem states, the condition on c_2 can be weaken as $0 < c_1 < c_2 < 1$, and β_{k+1} can be any value satisfying $0 \leq \beta_{k+1} \leq \beta^{\text{R-CD}}_{k+1}$.

THEOREM 6.6. *Under Assumption* 5.1*, let sequence* $\{x_k\}$ *be generated by Algorithm* 4.1 *with* β_{k+1} *satisfying* $0 \leq \beta_{k+1} \leq \beta^{\text{R-CD}}_{k+1}$*, where* $\beta^{\text{R-CD}}_{k+1}$ *is defined as* (4.9)*. If* $\mathscr{T}^{(k)}$ *satisfies Assumption* 4.2 *and the step lengths satisfy the generalized* $\mathscr{T}^{(k)}$*-Wolfe conditions* (3.5) *and* (4.15) *with* $0 < c_1 < c_2 < 1$ *and* $c_3 = 0$*, then for* $g_k := \operatorname{grad} f(x_k)$*, we have*

$$\liminf_{k\to\infty} \|g_k\|_{x_k} = 0.$$

Proof. Following the discussion in the proof of Proposition 6.1, we can prove $-1 - c_2 \leq \langle g_k, \eta_k\rangle_{x_k}/\|g_k\|^2_{x_k} \leq -1$ and $\|g_k\|_{x_k} \leq \|\eta_k\|_{x_k}$ instead of (6.1) and (6.2), respectively. Therefore, the assumption in Zoutendijk's theorem (Theorem 5.3) is satisfied, and the subsequent proof is the same as that of Theorem 6.2, where we note $|\beta_{k+1}| \leq |\beta^{\text{R-FR}}_{k+1}|$ from Proposition 6.5. ∎

The discussion on FR- and DY-types of R-CG methods here is partly similar to that in previous studies, where some specific choices of s_k and $\mathscr{T}^{(k)}$ are utilized. However, the analyses provided in this section are meaningful and not trivial since they address our general framework of R-CG methods (i.e., Algorithm 4.1) and more general classes of β_{k+1}. Furthermore, to the author's knowledge, no discussion on the CD-type of the R-CG method has been conducted before, even for a specific $\mathscr{T}^{(k)}$ such as those based on parallel translation or vector transport.

6.2. Global convergence analyses of R-CG methods with variants of $\boldsymbol{\beta}^{\text{R-PRP}}$, $\boldsymbol{\beta}^{\text{R-HS}}$, and $\boldsymbol{\beta}^{\text{R-LS}}$. While some existing studies discuss the FR- and DY-types of R-CG methods, the theoretical properties of the PRP-, HS-, and LS-types of R-CG methods were not well known until now. Furthermore, even in Euclidean spaces, these three types of CG methods with the (strong) Wolfe conditions are not generally guaranteed to converge [8, section 4.2]. Therefore, various variants are proposed and analyzed. For example, $\beta^{\text{PRP+}}_{k+1} := \max\{\beta^{\text{PRP}}_{k+1}, 0\}$ is known to generate convergent sequences under some assumptions in the Euclidean case [29]. Some comprehensive surveys on the Euclidean CG methods are found in [7, 8, 32, 43]. In this subsection, we generalize some examples of such variants by exploiting the theoretical results in subsection 6.1.

Considering the Euclidean CG methods, the FR-, DY-, and CD-types may suffer from jamming, i.e., if the search direction is nearly orthogonal to the gradient at some iteration and the corresponding step is small, the subsequent sequences are likely to make little progress [44, section 5.2]. An important feature of the PRP-, HS-, and LS-types is that they can avoid jamming. This is because the quantity in the numerator of

$\beta_{k+1}^{\mathrm{PRP}}$, $\beta_{k+1}^{\mathrm{HS}}$, and $\beta_{k+1}^{\mathrm{LS}}$ becomes close to 0 when $x_{k+1} \approx x_k$, and the search direction is almost the steepest descent (SD) direction $-\nabla f(x_{k+1})$. This phenomenon can also be explained in the R-CG methods, assuming that $l_k\mathscr{S}^{(k)}(g_k) \approx g_k$ when $x_{k+1} \approx x_k$. For example, consider $\beta_{k+1}^{\text{R-PRP}} := (\|g_{k+1}\|_{x_{k+1}}^2 - \langle g_{k+1}, l_k\mathscr{S}^{(k)}(g_k)\rangle_{x_{k+1}})/\|g_k\|_{x_k}^2$ as defined in (4.10). If $x_{k+1} \approx x_k$, then $g_{k+1} \approx g_k$ when grad f is continuous, and the numerator is approximated as $\|g_{k+1}\|_{x_{k+1}}^2 - \langle g_{k+1}, l_k\mathscr{S}^{(k)}(g_k)\rangle_{x_{k+1}} \approx \|g_k\|_{x_k}^2 - \langle g_k, g_k\rangle_{x_k} = 0$. Therefore, the subsequent search direction in Algorithm 4.1 is $\eta_{k+1} \approx -g_{k+1}$, which is the negative gradient of f at x_{k+1}. Therefore, the PRP-type of R-CG methods can be considered to be equipped with an automatic restart strategy, indicating that the search direction is almost reset as the SD direction when $x_{k+1} \approx x_k$. The same is also applied to the HS- and LS-types of R-CG methods.

As mentioned, although the PRP-, HS-, and LS-types of (R-)CG methods may be practically superior to the FR-, DY-, and CD-types, they do not necessarily generate convergent sequences. Here, we observe that $\beta_{k+1}^{\mathrm{FR}}$ and $\beta_{k+1}^{\mathrm{PRP}}$ have the same form of denominator. Therefore, the PRP-type R-CG methods can be regarded as practically modified versions of the FR-type R-CG methods so that they have the aforementioned restart mechanism, while the FR-types have theoretically better convergence properties than the PRP-types. Based on this discussion, a natural modification of $\beta_{k+1}^{\text{R-PRP}}$ is $\beta_{k+1} = \max\{0, \min\{\beta_{k+1}^{\text{R-PRP}}, \beta_{k+1}^{\text{R-FR}}\}\}$, which ensures the condition $0 \le \beta_{k+1} \le \beta_{k+1}^{\text{R-FR}}$ in Theorem 6.2. We can also develop this discussion for HS–DY- and CD–LS-types.

6.2.1. R-CG methods with modified $\boldsymbol{\beta}^{\text{R-PRP}}$. Based on the above discussion, a practical implementation of the PRP-type β_{k+1} defined in (4.10), which is $\beta_{k+1}^{\text{R-PRP}} := (\|g_{k+1}\|_{x_{k+1}}^2 - \langle g_{k+1}, l_k\mathscr{S}^{(k)}(g_k)\rangle_{x_{k+1}})/\|g_k\|_{x_k}^2$, may be to combine it with $\beta_{k+1}^{\text{R-FR}}$ as $\beta_{k+1} = \max\{0, \min\{\beta_{k+1}^{\text{R-PRP}}, \beta_{k+1}^{\text{R-FR}}\}\}$. The Euclidean version of this approach (i.e., $\beta_{k+1} = \max\{0, \min\{\beta_{k+1}^{\mathrm{PRP}}, \beta_{k+1}^{\mathrm{FR}}\}\}$) is mentioned in [35]. Because $0 \le \beta_{k+1} \le \beta_{k+1}^{\text{R-FR}}$ holds, Theorem 6.2 implies the following result.

THEOREM 6.7. *Let $\{x_k\}$ be a sequence generated by Algorithm* 4.1 *under the assumption in Theorem* 6.2 *and $\beta_{k+1} = \beta_{k+1}^{\text{R-PRP–FR}} := \max\{0, \min\{\beta_{k+1}^{\text{R-PRP}}, \beta_{k+1}^{\text{R-FR}}\}\}$. Then, for $g_k := \operatorname{grad} f(x_k)$, we have $\liminf_{k\to\infty} \|g_k\|_{x_k} = 0$.*

6.2.2. R-CG methods with modified $\boldsymbol{\beta}^{\text{R-HS}}$. We here discuss the HS-type $\beta_{k+1}^{\text{R-HS}} := (\|g_{k+1}\|_{x_{k+1}}^2 - \langle g_{k+1}, l_k\mathscr{S}^{(k)}(g_k)\rangle_{x_{k+1}})/(\langle g_{k+1}, s_k\mathscr{T}^{(k)}(\eta_k)\rangle_{x_{k+1}} - \langle g_k, \eta_k\rangle_{x_k})$ defined as in (4.11). We develop the idea discussed in subsection 6.2.1 and use Theorem 6.4 to propose and analyze the following algorithm, whose Euclidean counterpart $\beta_{k+1} = \max\{0, \min\{\beta_{k+1}^{\mathrm{HS}}, \beta_{k+1}^{\mathrm{DY}}\}\}$ is proposed in [21].

THEOREM 6.8. *Let $\{x_k\}$ be a sequence generated by Algorithm* 4.1 *under the assumption in Theorem* 6.4 *and $\beta_{k+1} = \beta_{k+1}^{\text{R-HS–DY}} := \max\{0, \min\{\beta_{k+1}^{\text{R-HS}}, \beta_{k+1}^{\text{R-DY}}\}\}$. Then, for $g_k := \operatorname{grad} f(x_k)$, we have $\liminf_{k\to\infty} \|g_k\|_{x_k} = 0$.*

6.2.3. R-CG methods with modified $\boldsymbol{\beta}^{\text{R-LS}}$. Finally, we discuss the LS-type $\beta_{k+1}^{\text{R-LS}} := (\|g_{k+1}\|_{x_{k+1}}^2 - \langle g_{k+1}, l_k\mathscr{S}^{(k)}(g_k)\rangle_{x_{k+1}})/(-\langle g_k, \eta_k\rangle_{x_k})$ defined in (4.12). From Theorem 6.6, we can similarly propose and analyze the following algorithm, which is a generalization of $\beta_{k+1} = \max\{0, \min\{\beta_{k+1}^{\mathrm{LS}}, \beta_{k+1}^{\mathrm{CD}}\}\}$ for the Euclidean case [7].

THEOREM 6.9. *Let $\{x_k\}$ be a sequence generated by Algorithm* 4.1 *under the assumption in Theorem* 6.6 *and $\beta_{k+1} = \beta_{k+1}^{\text{R-LS–CD}} := \max\{0, \min\{\beta_{k+1}^{\text{R-LS}}, \beta_{k+1}^{\text{R-CD}}\}\}$. Then, for $g_k := \operatorname{grad} f(x_k)$, we have $\liminf_{k\to\infty} \|g_k\|_{x_k} = 0$.*

7. Numerical experiments. In this section, we compare Algorithm 4.1 with several choices of β_{k+1}. Specifically, we use $\beta_{k+1}^{\text{R-FR}}$, $\beta_{k+1}^{\text{R-DY}}$, $\beta_{k+1}^{\text{R-CD}}$, $\beta_{k+1}^{\text{R-PRP}}$, $\beta_{k+1}^{\text{R-HS}}$, and

$\beta_{k+1}^{\text{R-LS}}$ in (4.7)–(4.12) and $\beta_{k+1}^{\text{R-PRP–FR}}$, $\beta_{k+1}^{\text{R-HS–DY}}$, and $\beta_{k+1}^{\text{R-LS–CD}}$ in subsection 6.2 as hybrid methods. Since one of our contributions is that the proposed class of R-CG methods (Algorithm 4.1) offers the use of a user-selected map $\mathscr{T}^{(k)}$, we deal with two optimization problems using different choices of $\mathscr{T}^{(k)}$.

The experiments were carried out in double-precision floating-point arithmetic on a PC (Intel Xeon CPU E5-2620 v4, 128 GB RAM) equipped with MATLAB R2021b. In all the experiments below, we implemented the R-CG methods based on `conjugategradient` in Manopt [18], which is a MATLAB toolbox for Riemannian optimization. The step length computed in each iteration satisfies the Armijo condition (3.5) by default. The iterations of the R-CG methods were terminated when $\|g_k\|_{x_k}/\|g_0\|_{x_0} < 10^{-6}$ was attained.

7.1. R-CG methods on the product of Grassmann manifolds with the projection-based vector transport for singular value decomposition. We consider Problem 2.1 of large size with the manifold $\mathcal{M} := \mathrm{Grass}(p,m) \times \mathrm{Grass}(p,n)$, where $m = 50{,}000$, $n = 3{,}000$, and $p = 100$, implying that the dimension of the search space is $\dim \mathcal{M} = p(m-p) + p(n-p) = 5{,}280{,}000$. Each point W on the Grassmann manifold $\mathrm{Grass}(p,n) \simeq \mathrm{St}(p,n)/\mathcal{O}(p)$ is expressed as an equivalence class $[U] := \{UQ \mid Q \in \mathcal{O}(p)\}$ with a representative $U \in \mathrm{St}(p,n)$. We define the objective function f on $\mathcal{M}$ as $f([U],[V]) := -\|U^T A V\|_F^2/2$, where A is a randomly generated $m \times n$ matrix and $\|\cdot\|_F$ is the Frobenius norm. This problem is for the singular value decomposition of A.

In the experiments, we used the polar-based retraction and the projection-based vector transport $\mathcal{T}^P$, which are the default settings in Manopt's `grassmannfactory`, and set $\mathscr{T}^{(k)}(\eta_k) := \mathcal{T}^P_{t_k\eta_k}(\eta_k)$ and $s_k := \min\{1, \|\eta_k\|_{x_k}/\|\mathscr{T}^{(k)}(\eta_k)\|_{x_{k+1}}\}$ in Algorithm 4.1. Note that this $\mathscr{T}^{(k)}$ satisfies Assumption 4.2 as discussed in Example 4.6. Similarly, we set $\mathscr{S}^{(k)}(g_k) := \mathcal{T}^P_{t_k\eta_k}(g_k)$ and $l_k := \min\{1, \|g_k\|_{x_k}/\|\mathscr{S}^{(k)}(g_k)\|_{x_{k+1}}\}$ in (4.10)–(4.12). We compared $\beta_{k+1}^{\text{R-SD}} := 0$ (SD method) and the nine types of β_{k+1} mentioned above (six standard types and three hybrid ones) with the same initial point, which was randomly generated.

Figure 1 shows the convergence histories of the 10 methods. In this figure, we observe several clusters of the graphs: SD and FR; DY, CD, and LS–CD; PRP, HS, LS, PRP–FR, and HS–DY. We observe that SD is the slowest, as expected, and FR, DY, and CD are not so fast either. These three types of R-CG methods are considered similar as discussed in section 6. The other three types, PRP, HS, and LS, are faster than FR, DY, and CD. Regarding the hybrid methods, PRP–FR and HS–DY methods showed similar performance to PRP and HS. On the other hand, LS–CD is slower than LS. In this case, LS–CD seems to be slowed down by the effect of CD.

We further applied the Riemannian trust-region (TR) method [1] for the same problem with the same initial point and compared computational time. The time (in seconds) taken for the relative gradient norm to become less than 10^{-6} is summarized in Table 1. As Figure 1 implies, SD and FR did not achieve the stopping criterion within 30 minutes. Furthermore, although the convergence of HS was fast, it failed to find a step length satisfying the Armijo condition at $k = 656$. The minimum value of the relative gradient norm that HS attained was $1.47 \times 10^{-6} > 10^{-6}$ (at $k = 651$ in 827.7 seconds). It is worth noting that the TR method took much longer time than most CG methods. Of course, the TR method has the advantage of superlinear convergence once a point sufficiently close to an optimal solution is obtained. However, Table 1 shows that the CG methods are competitive enough for the purpose of obtaining a reasonably good solution.

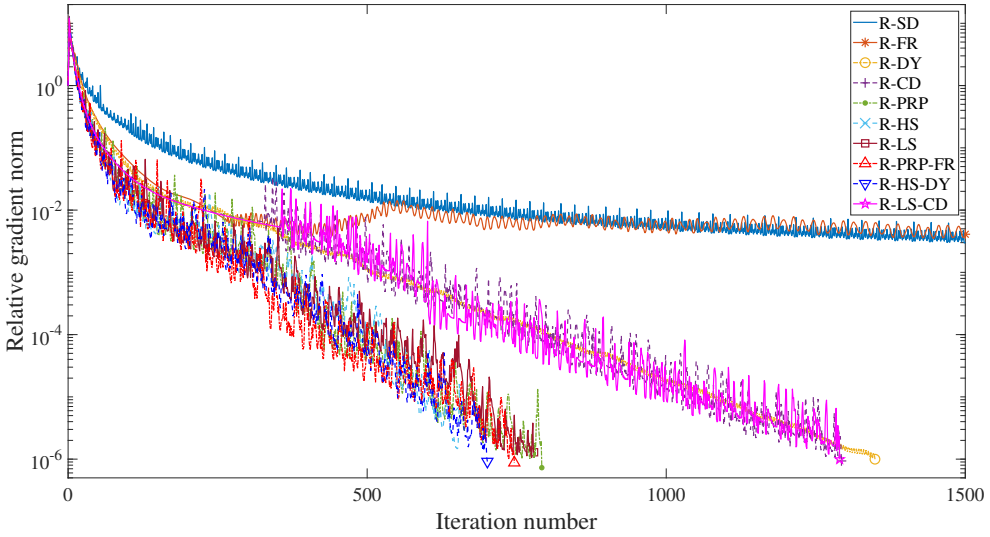

FIG. 1. *Numerical results for Problem* 2.1 *on* $\mathcal{M} = \mathrm{Grass}(p,m) \times \mathrm{Grass}(p,n)$. *The horizontal axis represents the iteration number* k, *and the vertical axis represents the relative norm of the gradient of the objective function* $\|g_k\|_{x_k}/\|g_0\|_{x_0}$. *A marker is put on each graph at the last iteration for visibility.*

TABLE 1
Computation time [s] required for $\|g_k\|_{x_k}/\|g_0\|_{x_0} < 10^{-6}$ *to be satisfied.*

	SD	FR	DY	CD	PRP	HS	LS	PRP–FR	HS–DY	LS–CD	TR
Time [s]	–	–	1594.9	1590.6	1026.5	–	974.6	951.3	890.2	1610.0	25961.6

7.2. R-CG methods on the manifold of symmetric positive definite matrices for solving Lyapunov equation. Subsequently, we consider Problem 2.1 with $\mathcal{M} := \mathrm{SPD}(n)$ endowed with the Bures–Wasserstein geometry [42], where $n = 50$, implying that $\dim \mathcal{M} = n(n+1)/2 = 1{,}275$. We define the objective function f on $\mathcal{M}$ as $f(X) := \mathrm{tr}(XAX) - \mathrm{tr}(XC)$, where $A, C \in \mathrm{Sym}(n)$ are generated in the same way as in (`Ex2`) of [33]. The resultant optimization problem is for solving the Lyapunov equation $AX + XA = C$ for $X \in \mathcal{M}$.

We used the exponential retraction and set $\mathscr{T}^{(k)}(\eta_k) := \eta_k$ and $s_k := 1$ in Algorithm 4.1, and $\mathscr{S}^{(k)}(g_k) := g_k$ and $l_k := 1$ for (4.10)–(4.12). We again compared the R-SD and nine types of R-CG methods, where the initial point was $X_0 = I_n$.

Overall, the results of Figure 2 can be explained similarly to those in the previous subsection. It is observed that SD is the slowest, as expected. Among the R-CG methods, CD is faster than FR, but slower than DY, PRP, HS, LS, and the hybrid methods (except for LS–CD). LS–CD is slower than PRP–FR and HS–DY, possibly because CD negatively affected the performance of LS.

8. Concluding remarks. In this paper, to address unconstrained Riemannian optimization problems (Problem 2.1), we proposed a general framework of R-CG methods (Algorithm 4.1) with maps $\mathscr{T}^{(k)}$ and scaling parameters s_k for $k \geq 0$. Several conditions on $\mathscr{T}^{(k)}$, s_k, and the step lengths t_k were developed to provide convergence analyses for the types of Algorithm 4.1.

An important parameter characterizing the (R-)CG methods is β_{k+1} in (3.2) (Euclidean case) and (4.1) (Riemannian case). As the six standard types of R-CG

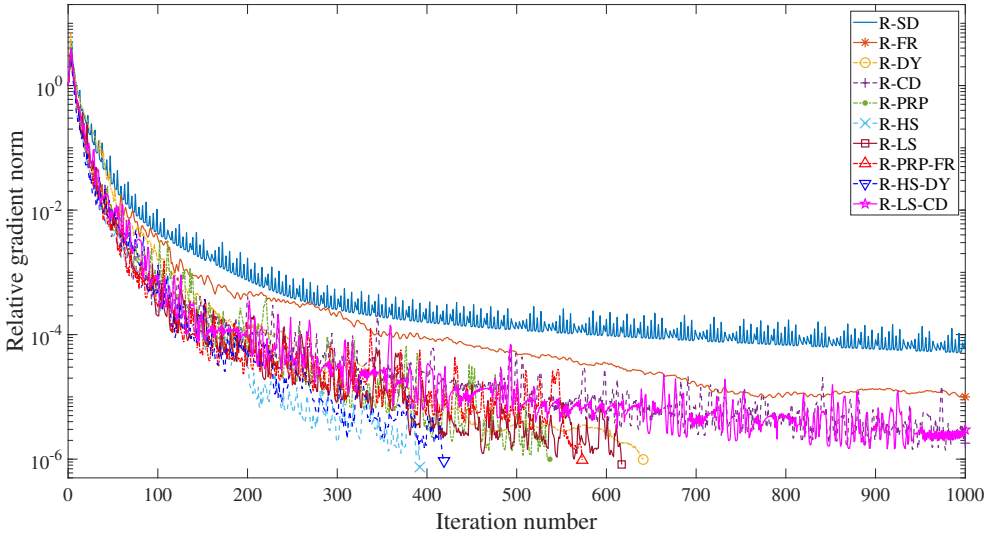

FIG. 2. *Numerical results for Problem* 2.1 *on* $\mathcal{M} = \mathrm{SPD}(n)$. *The horizontal axis represents the iteration number* k, *and the vertical axis represents the relative norm of the gradient of the objective function* $\|g_k\|_{X_k}/\|g_0\|_{X_0}$. *A marker is put on each graph at the last iteration for visibility.*

methods, we generalized (omitting the subscript $k+1$) β^{FR}, β^{DY}, β^{CD}, β^{PRP}, β^{HS}, and β^{LS} in Euclidean spaces to the Riemannian counterparts $\beta^{\text{R-FR}}$, $\beta^{\text{R-DY}}$, $\beta^{\text{R-CD}}$, $\beta^{\text{R-PRP}}$, $\beta^{\text{R-HS}}$, and $\beta^{\text{R-LS}}$, respectively. We extended Zoutendijk's theorem to our proposed framework of the R-CG methods and extensively analyzed the FR-, DY-, and CD-types of R-CG methods to guarantee the global convergence properties. The analyses also claim that any choice of nonnegative β that is smaller or equal to $\beta^{\text{R-FR}}$, $\beta^{\text{R-DY}}$, or $\beta^{\text{R-CD}}$ with appropriate assumptions ensures the global convergence of R-CG methods. For the PRP-, HS-, and LS-types of R-CG methods, even whose Euclidean versions are not necessarily globally convergent, we discussed why they can outperform the other three types of R-CG methods by explaining that they are considered to be equipped with a restart mechanism when jamming occurs. Furthermore, modifying them and exploiting the analyses of the FR-, DY-, and CD types of R-CG methods, we proposed several practically and theoretically appropriate hybrid methods.

We demonstrated numerical experiments to observe the performances of several R-CG methods in the framework of Algorithm 4.1. The CD-type of R-CG methods have been rarely used in the literature; however, they are possibly superior to the FR- or DY-type methods. On the other hand, considering our experiments, the PRP-, HS-, and LS-types of R-CG methods behaved much better than the FR-, DY-, and CD-types. If guaranteeing the global convergence is important, it is also nice to use the hybrid types, i.e., PRP–FR, HS–DY, and LS–CD types of methods.

Since there are various types of CG methods even for the Euclidean case, this paper does not cover all the existing methods in detail. However, we believe that the proposed general framework will be the foundation for future studies on R-CG methods.

Acknowledgments. The author would like to thank the editor and anonymous referees for their insightful comments that helped improve the paper significantly.

Conflict of interest. The author declares that he has no conflict of interest.

REFERENCES

[1] P.-A. Absil, C. G. Baker, and K. A. Gallivan, *Trust-region methods on Riemannian manifolds*, Found. Comput. Math., 7 (2007), pp. 303–330.
[2] P.-A. Absil and S. Hosseini, *A collection of nonsmooth Riemannian optimization problems*, in Nonsmooth Optimization and Its Applications, Springer, New York, 2019, pp. 1–15.
[3] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2008.
[4] P.-A. Absil and J. Malick, *Projection-like retractions on matrix manifolds*, SIAM J. Optim., 22 (2012), pp. 135–158.
[5] P.-A. Absil and I. V. Oseledets, *Low-rank retractions: A survey and new results*, Comput. Optim. Appl., 62 (2015), pp. 5–29.
[6] R. L. Adler, J.-P. Dedieu, J. Y. Margulies, M. Martens, and M. Shub, *Newton's method on Riemannian manifolds and a geometric model for the human spine*, IMA J. Numer. Anal., 22 (2002), pp. 359–390.
[7] N. Andrei, 40 *Conjugate Gradient Algorithms for Unconstrained Optimization. A Survey on Their Definition*, Tech. report, ICI, 2008.
[8] N. Andrei, *Nonlinear Conjugate Gradient Methods for Unconstrained Optimization*, Springer, 2020.
[9] M. Bacák, R. Bergmann, G. Steidl, and A. Weinmann, *A second order nonsmooth variational model for restoring manifold-valued images*, SIAM J. Sci. Comput., 38 (2016), pp. A567–A597.
[10] T. Bendokat and R. Zimmermann, *Efficient quasi-geodesics on the Stiefel manifold*, in International Conference on Geometric Science of Information, Springer, New York, 2021, pp. 763–771.
[11] G. d. C. Bento, J. X. da Cruz Neto, and L. V. de Meireles, *Proximal point method for locally Lipschitz functions in multiobjective optimization of Hadamard manifolds*, J. Optim. Theory Appl., 179 (2018), pp. 37–52.
[12] R. Bergmann, *Manopt.jl: Optimization on Manifolds in Julia*, J. Open Source Software, 7 (2022), 3866.
[13] R. Bergmann, R. Herzog, M. Silva Louzeiro, D. Tenbrinck, and J. Vidal-Núñez, *Fenchel duality theory and a primal-dual algorithm on Riemannian manifolds*, Found. Comput. Math., 21 (2021), pp. 1465–1504.
[14] R. Bergmann and A. Weinmann, *Inpainting of cyclic data using first and second order differences*, in International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, Springer, New York, 2015, pp. 155–168.
[15] R. Bhatia, *Positive Definite Matrices*, Princeton Ser. Appl. Math. 24, Princeton University Press, Princeton, NJ, 2009.
[16] S. Bonnabel, *Stochastic gradient descent on Riemannian manifolds*, IEEE Trans. Automat. Control, 58 (2013), pp. 2217–2229.
[17] N. Boumal, *An Introduction to Optimization on Smooth Manifolds*, Cambridge University Press, to appear, http://www.nicolasboumal.net/book.
[18] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, *Manopt, a Matlab toolbox for optimization on manifolds*, J. Mach. Learn Res., 15 (2014), pp. 1455–1459.
[19] S. Chen, S. Ma, A. M. C. So, and T. Zhang, *Proximal gradient method for nonsmooth optimization over the Stiefel manifold*, SIAM J. Optim., 30 (2020), pp. 210–239.
[20] Y.-H. Dai and Y. Yuan, *A nonlinear conjugate gradient method with a strong global convergence property*, SIAM J. Optim., 10 (1999), pp. 177–182.
[21] Y.-h. Dai and Y. Yuan, *An efficient hybrid conjugate gradient method for unconstrained optimization*, Ann. Oper. Res., 103 (2001), pp. 33–47.
[22] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 303–353.
[23] A. Edelman and S. T. Smith, *On conjugate gradient-like methods for eigen-like problems*, BIT, 36 (1996), pp. 494–508.
[24] O. P. Ferreira, M. S. Louzeiro, and L. F. Prudente, *Iteration-complexity and asymptotic analysis of steepest descent method for multiobjective optimization on Riemannian manifolds*, J. Optim. Theory Appl., 184 (2020), pp. 507–533.
[25] R. Fletcher, *Practical Methods of Optimization*, John Wiley & Sons, New York, 2013.
[26] R. Fletcher and C. M. Reeves, *Function minimization by conjugate gradients*, Comput. J., 7 (1964), pp. 149–154.
[27] B. Gao and P.-A. Absil, *A Riemannian rank-adaptive method for low-rank matrix completion*, Comput. Optim. Appl., 81 (2022), pp. 67–90.

[28] B. Gao, N. T. Son, P.-A. Absil, and T. Stykel, *Riemannian optimization on the symplectic Stiefel manifold*, SIAM J. Optim., 31 (2021), pp. 1546–1575.
[29] J. C. Gilbert and J. Nocedal, *Global convergence properties of conjugate gradient methods for optimization*, SIAM J. Optim., 2 (1992), pp. 21–42.
[30] J. Goto and H. Sato, *Approximated logarithmic maps on Riemannian manifolds and their applications*, JSIAM Lett., 13 (2021), pp. 17–20.
[31] W. W. Hager and H. Zhang, *A new conjugate gradient method with guaranteed descent and an efficient line search*, SIAM J. Optim., 16 (2005), pp. 170–192.
[32] W. W. Hager and H. Zhang, *A survey of nonlinear conjugate gradient methods*, Pac. J. Optim., 2 (2006), pp. 35–58.
[33] A. Han, B. Mishra, P. K. Jawanpuria, and J. Gao, *On Riemannian optimization over positive definite matrices with the Bures-Wasserstein geometry*, Adv. Neural Inf. Process. Syst., 34 (2021).
[34] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Natl. Bur. Stand., 49 (1952), pp. 409–436.
[35] Y. Hu and C. Storey, *Global convergence result for conjugate gradient methods*, J. Optim. Theory Appl., 71 (1991), pp. 399–405.
[36] W. Huang, *Optimization Algorithms on Riemannian Manifolds with Applications*, Ph.D. thesis, Florida State University, 2013.
[37] A. Kovnatsky, K. Glashoff, and M. M. Bronstein, *MADMM: A generic algorithm for non-smooth optimization on manifolds*, in European Conference on Computer Vision, Springer, 2016, pp. 680–696.
[38] C. Lemaréchal, *A view of line-searches*, in Optimization and Optimal Control, Springer, New York, 1981, pp. 59–78.
[39] A. Lichnewsky, *Une methode de gradient conjugue sur des varietes application a certains problemes de valeurs propres non lineaires*, Numer. Funct. Anal. Optim., 1 (1979), pp. 515–560.
[40] C. Liu and N. Boumal, *Simple algorithms for optimization on Riemannian manifolds with constraints*, Appl. Math. Optim., 82 (2020), pp. 949–981, https://doi.org/10.1007/s00245-019-09564-3.
[41] Y. Liu and C. Storey, *Efficient generalized conjugate gradient algorithms, part* 1: *Theory*, J. Optim. Theory Appl., 69 (1991), pp. 129–137.
[42] L. Malagò, L. Montrucchio, and G. Pistone, *Wasserstein Riemannian geometry of Gaussian densities*, Inform. Geom., 1 (2018), pp. 137–179.
[43] Y. Narushima and H. Yabe, *A survey of sufficient descent conjugate gradient methods for unconstrained optimization*, SUT J. Math., 50 (2014), pp. 167–203.
[44] J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed., Springer, New York, 2006.
[45] E. Polak and G. Ribière, *Note sur la convergence de méthodes de directions conjuguées*, ESAIM Math. Model. Numer. Anal., 3 (1969), pp. 35–43.
[46] B. T. Polyak, *The conjugate gradient method in extremal problems*, USSR Comput. Math. Math. Phys., 9 (1969), pp. 94–112.
[47] W. Ring and B. Wirth, *Optimization methods on Riemannian manifolds and their application to shape space*, SIAM J. Optim., 22 (2012), pp. 596–627.
[48] H. Sakai and H. Iiduka, *Hybrid Riemannian conjugate gradient methods with global convergence properties*, Comput. Optim. Appl., 77 (2020), pp. 811–830.
[49] H. Sakai and H. Iiduka, *Sufficient descent Riemannian conjugate gradient methods*, J. Optim. Theory Appl., 190 (2021), pp. 130–150.
[50] H. Sato, *A Dai–Yuan-type Riemannian conjugate gradient method with the weak Wolfe conditions*, Comput. Optim. Appl., 64 (2016), pp. 101–118.
[51] H. Sato, *Riemannian Optimization and Its Applications*, Springer, New York, 2021.
[52] H. Sato and K. Aihara, *Cholesky QR-based retraction on the generalized Stiefel manifold*, Comput. Optim. Appl., 72 (2019), pp. 293–308.
[53] H. Sato and T. Iwai, *A new, globally convergent Riemannian conjugate gradient method*, Optimization, 64 (2015), pp. 1011–1031.
[54] H. Sato, H. Kasai, and M. Bamdev, *Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport*, SIAM J. Optim., 29 (2019), pp. 1444–1472.
[55] K. Sato, H. Sato, and T. Damm, *Riemannian optimal identification method for linear systems with symmetric positive-definite matrix*, IEEE Trans. Automat. Control, 65 (2020), pp. 4493–4508.
[56] A. Schiela and J. Ortiz, *An SQP method for equality constrained optimization on Hilbert manifolds*, SIAM J. Optim., 31 (2021), pp. 2255–2284.
[57] S. T. Smith, *Optimization techniques on Riemannian manifolds*, in Hamiltonian and Gradient Flows, Algorithms and Control, AMS, Providence, RI, 1994, pp. 113–135.

[58] J. Townsend, N. Koep, and S. Weichwald, *Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation*, J. Mach. Learn Res., 17 (2016), pp. 4755–4759.
[59] Y. Yamakawa and H. Sato, *Sequential optimality conditions for nonlinear optimization on Riemannian manifolds and a globally convergent augmented lagrangian method*, Comput. Optim. Appl., (2022), pp. 1–25.
[60] H. Zhang, S. J. Reddi, and S. Sra, *Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds*, in Adv. Neural Inf. Process. Syst., 29 (2016), pp. 4592–4600.
[61] J. Zhang, S. Ma, and S. Zhang, *Primal-dual optimization algorithms over Riemannian manifolds: An iteration complexity analysis*, Math. Program., 184 (2020), pp. 445–490.
[62] X. Zhu, *A Riemannian conjugate gradient method for optimization on the Stiefel manifold*, Comput. Optim. Appl., 67 (2017), pp. 73–110.
[63] X. Zhu and H. Sato, *Riemannian conjugate gradient methods with inverse retraction*, Comput. Optim. Appl., 77 (2020), pp. 779–810.
[64] X. Zhu and H. Sato, *Cayley-transform-based gradient and conjugate gradient algorithms on Grassmann manifolds*, Adv. Comput. Math., 47 (2021), pp. 1–28.