Riemannian Accelerated Gradient Methods via Extrapolation

Andi Han
University of Sydney

Bamdev Mishra
Microsoft, India

Pratik Jawanpuria
Microsoft, India

Junbin Gao
University of Sydney

Abstract

In this paper, we propose a convergence acceleration scheme for general Riemannian optimization problems by extrapolating iterates on manifolds. We show that when the iterates are generated from the Riemannian gradient descent method, the scheme achieves the optimal convergence rate asymptotically and is computationally more favorable than the recently proposed Riemannian Nesterov accelerated gradient methods. A salient feature of our analysis is the convergence guarantees with respect to the use of general retraction and vector transport. Empirically, we verify the practical benefits of the proposed acceleration strategy, including robustness to the choice of different averaging schemes on manifolds.

1 INTRODUCTION

In this paper, we consider the optimization problem

$$\min_{x \in \mathcal{M}} f(x), \tag{1}$$

where $\mathcal{M}$ is a Riemannian manifold and $f : \mathcal{M} \to \mathbb{R}$ is a smooth, real-valued function. Optimization on a Riemannian manifold naturally appears in various fields of applications, including principal component analysis (Edelman et al., 1998; Zhang et al., 2016), tensor completion and factorization (Keshavan and Oh, 2009; Vandereycken, 2013; Boumal and Absil, 2015; Jawanpuria and Mishra, 2018; Nimishakavi et al., 2018), learning representations for hierarchical structures (Nickel and Kiela, 2017; Jawanpuria et al., 2019), dictionary learning (Cherian and Sra, 2016; Harandi et al., 2013), cross-lingual translation (Jawanpuria et al., 2020a,b, 2021), and optimal transport (Shi et al., 2021; Jawanpuria et al., 2021; Mishra et al., 2021; Han et al., 2022), to name a few. Riemannian optimization (Absil et al., 2009; Boumal, 2020) provides a universal and efficient framework for problem (1) that respects the intrinsic geometry of the constraint set.

The Riemannian gradient descent method (Udriste, 2013; Zhang and Sra, 2016; Absil et al., 2009; Boumal, 2020) generalizes the classical gradient descent method in the Euclidean space with intrinsic updates on manifolds. Various other approaches have also been explored for Riemannian optimization such as stochastic methods (Bonnabel, 2013; Zhang et al., 2016; Becigneul and Ganea, 2018; Kasai et al., 2018, 2019; Han and Gao, 2021a,b), second-order methods (Absil et al., 2007; Qi et al., 2010; Huang et al., 2015b; Agarwal et al., 2021), among others.

Existing works have also explored generalizing Nesterov acceleration (Nesterov, 1983) to Riemannian manifolds, including (Liu et al., 2017; Ahn and Sra, 2020; Zhang and Sra, 2018; Alimisis et al., 2020; Jin and Sra, 2022; Kim and Yang, 2022; Criscitiello and Boumal, 2022a). However, they primarily involve exponential map, inverse exponential map, and parallel transport (formally defined in Section 2), which are computationally expensive operations. In addition, the Nesterov acceleration based methods require the knowledge of smoothness and strong convexity constants, which are often unknown in practical settings. Furthermore, recent studies (Hamilton and Moitra, 2021; Criscitiello and Boumal, 2022b) show that global acceleration cannot be achieved on manifolds in general.

In this paper, we focus on an extrapolation based strategy to produce an accelerated sequence. The core idea is to compute extrapolation as a linear combination of the iterates where the weights depend nonlinearly on the iterates. Existing works (Aitken, 1927; Shanks, 1955; Brezinski et al., 2018; Wynn, 1956; Sidi et al., 1986; Walker and Ni, 2011; Scieur et al., 2020) have explored such strategy in the Euclidean setting. Recently, it has been shown in Scieur et al. (2020) that such nonlinear acceleration (Euclidean) scheme achieves optimal convergence rates asymptotically without knowing the function-specific constants.

A natural question is *can such extrapolation idea be generalized to Riemannian manifolds so that we achieve acceleration?* The nonlinear structure of manifolds imposes key technical challenges such as averaging on manifolds, distortion due to varying metric, computationally expensive

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

operations, like exponential map and parallel transport, to name a few. Nevertheless, we answer the above question affirmatively and our contributions are as follows.

- We propose an acceleration strategy for Riemannian optimization based on the idea of extrapolation, which we call the Riemannian nonlinear acceleration (RiemNA) strategy. We analyze several averaging schemes that generalize weighted averaging in the Euclidean space from various perspectives.
- When the iterates are generated by the Riemannian gradient descent method, we show RiemNA achieves the optimal asymptotic first-order convergence rate. We show the convergence is robust to the choice of different averaging schemes on manifolds.
- A salient feature is that convergence of RiemNA holds under general retraction and vector transport. This is in contrast to existing analyses for Riemannian accelerated gradient methods which employ exponential map and parallel transport (Liu et al., 2017; Zhang and Sra, 2018; Ahn and Sra, 2020; Kim and Yang, 2022).
- We empirically demonstrate the superiority of RiemNA over state-of-the-art methods both in terms of convergence speed and computational efficiency.

2 PRELIMINARIES AND RELATED WORKS

Basic Riemannian Geometry A Riemannian manifold $\mathcal{M}$ is a smooth manifold endowed with a smooth inner product structure (Riemannian metric) on the tangent space $T_x\mathcal{M}$, for all $x \in \mathcal{M}$. The Riemannian inner product between any $u, v \in T_x\mathcal{M}$ is written as $\langle u, v \rangle_x$ and the induced norm of a tangent vector u is $\|u\|_x = \sqrt{\langle u, u \rangle_x}$. A 'straight' line on manifold is called a geodesic $\gamma : [0, 1] \rightarrow \mathcal{M}$, which is a locally distance minimizing curve with zero acceleration. Riemannian distance between $x, y \in \mathcal{M}$ is $d(x, y) = \inf_\gamma \int_0^1 \|\gamma'(t)\|_{\gamma(t)} dt$ where $\gamma(0) = x, \gamma(1) = y$. Exponential map $\mathrm{Exp}_x : T_x\mathcal{M} \rightarrow \mathcal{M}$ maps a tangent vector $u \in T_x\mathcal{M}$ to $\gamma(1)$ with $\gamma(0) = x, \gamma'(0) = u$. If between $x, y \in \mathcal{M}$, there exists a unique geodesic connecting them, the exponential map has an inverse $\mathrm{Exp}_x^{-1}(y)$ and the distance can be computed as $d(x, y) = \|\mathrm{Exp}_x^{-1}(y)\|_x = \|\mathrm{Exp}_y^{-1}(x)\|_y$. In this work, we only consider a unique-geodesic subset $\mathcal{X}$ of the manifold, which we explicitly assume in Section 4. Parallel transport $\Gamma_x^y : T_x\mathcal{M} \rightarrow T_y\mathcal{M}$ allows tangent vectors to be transported along a geodesic that connects x to y such that the induced vector fields are in *parallel*. It is known that parallel transport is isometric, i.e., $\langle \Gamma_x^y u, \Gamma_x^y v \rangle_y = \langle u, v \rangle_x$ for any $u, v \in T_x\mathcal{M}$. In this paper, we also consider the more general retraction and vector transport that include exponential map and parallel transport as special cases. A retraction, $\mathrm{Retr}_x : T_x\mathcal{M} \rightarrow \mathcal{M}$ is a first-order approximation to the exponential map and a vector transport $\mathcal{T}_x^y : T_x\mathcal{M} \rightarrow T_y\mathcal{M}$ is a linear map between tangent spaces that approximates the parallel transport.

Function Classes on Riemannian Manifolds For a differentiable, real-valued function $f : \mathcal{M} \rightarrow \mathbb{R}$, its Riemannian gradient at x, $\mathrm{grad} f(x) \in T_x\mathcal{M}$ is the unique tangent vector that satisfies $\langle \mathrm{grad} f(x), u \rangle_x = \mathrm{D} f(x)[u] = \langle \nabla f(x), u \rangle_2$ for all $u \in T_x\mathcal{M}$, where $\mathrm{D} f(x)[u]$ represents the directional derivative of f along u and $\nabla f(x)$ is the Euclidean gradient and $\langle \cdot, \cdot \rangle_2$ represents the Euclidean inner product. The Riemannian Hessian at x, $\mathrm{Hess} f(x) : T_x\mathcal{M} \rightarrow T_x\mathcal{M}$ is a self-adjoint operator, defined as the covariant derivative of the Riemannian gradient. On Riemannian manifolds, one can extend the notion of gradient and Hessian Lipschitzness in the Euclidean space to *geodesic Lipschitz gradient and Hessian* associated with the exponential map and parallel transport. Some equivalent characterizations, including function smoothness and bounded Hessian norm also exist for the Riemannian counterparts. Furthermore, the notion of convexity can be similarly generalized to *geodesic convexity* where the convex combination is defined along the geodesics. Similar notions are also properly defined with respect to general retractions. The formal definitions are deferred to Appendix D. See also Boumal (2020) for a more thorough treatment.

Metric Distortion Due to the curved geometry of Riemannian manifolds, many of the metric properties in the linear space are lost. To perform convergence analysis, we require the following geometric lemmas on manifolds that provide bounds on the metric distortion.

Lemma 1 (Ahn and Sra (2020); Sun et al. (2019))**.** *Consider a compact subset $\mathcal{X} \subseteq \mathcal{M}$ with unique geodesic. Let $x, y = \mathrm{Exp}_x(u) \in \mathcal{X}$ for some $u \in T_x\mathcal{M}$. Then for any $v \in T_x\mathcal{M}$, we have $d(\mathrm{Exp}_x(u + v), \mathrm{Exp}_y(\Gamma_x^y v)) \leq \min\{\|u\|, \|v\|\} C_\kappa(\|u\| + \|v\|)$, where $\mathcal{X}$ has curvature upper bounded by κ in magnitude and $C_\kappa(r) := \cosh(\sqrt{\kappa} r) - \sinh(\sqrt{\kappa} r)/(\sqrt{\kappa} r)$.*

Lemma 2 (Ahn and Sra (2020); Karcher (1977); Mangoubi and Smith (2018); Sun et al. (2019))**.** *For a compact subset $\mathcal{X} \subseteq \mathcal{M}$ with unique geodesic, there exists constants $C_0 > 0$, $C_1, C_2 \geq 1$ that depend on the curvature and diameter of $\mathcal{X}$ such that for all $x, y, z \in \mathcal{X}$, $u \in T_x\mathcal{M}$ we have (1). $\|\Gamma_y^z \Gamma_x^y u - \Gamma_x^z u\|_z \leq C_0 d(x, y) d(y, z) \|u\|_x$. (2). $C_1^{-1} d(x, y) \leq \|\mathrm{Exp}_z^{-1}(x) - \mathrm{Exp}_z^{-1}(y)\|_z \leq C_2 d(x, y)$. (3). $d\big(\mathrm{Exp}_x(u), \mathrm{Exp}_y(\Gamma_x^y u)\big) \leq C_3 d(x, y)$.*

Related Works on Riemannian Acceleration Generalizing Nesterov acceleration strategy (Nesterov, 1983) from the Euclidean space to Riemannian manifolds for geodesic (strongly) convex functions has been explored in Liu et al. (2017); Zhang and Sra (2018); Ahn and Sra (2020); Kim and Yang (2022); Jin and Sra (2022). Works such as Alimisis et al. (2020); Duruisseaux and Leok (2022) have approached acceleration on manifolds inspired by the continuous dynamics formulation of the Nesterov acceleration in the Euclidean space (Su et al., 2014; Wibisono et al., 2016). Lastly, acceleration has also been studied for specific manifolds, including sphere and hyperbolic manifolds (Martínez-Rubio, 2022) and Stiefel manifold (Siegel, 2019).

In the next section, we explore Riemannian nonlinear acceleration based on an extrapolation strategy for iterates generated from a Riemannian solver. Since our convergence analysis is local, the contributions can benefit both geodesic (strongly) convex functions and many nonconvex functions. Further, our convergence rates hold beyond the use of exponential map and parallel transport, which are the primary focus of the aforementioned works.

3 RIEMANNIAN NONLINEAR ACCELERATION

We generalize the nonlinear acceleration strategy for Riemannian optimization via a weighted Riemannian averaging on the manifold. For a set of weights $\{c_i\}_{i=0}^k$ and points $\{x_i\}_{i=0}^k$ on the manifold, we define the weighted Riemannian average $\bar{x}_{c,x}$ as

$$\bar{x}_{c,x} = \tilde{x}_k, \quad \tilde{x}_i = \mathrm{Exp}_{\tilde{x}_{i-1}}\Big(\frac{c_i}{\sum_{j=0}^i c_j} \mathrm{Exp}^{-1}_{\tilde{x}_{i-1}}(x_i) \Big), \tag{Avg.1}$$

for $i = 0, ..., k$ and $\tilde{x}_{-1} = x_0$. When $\mathcal{M}$ is the Euclidean space, (Avg.1) recovers the weighted mean as $\bar{x}_{c,x} = \sum_{i=0}^k c_i x_i$ (see Lemma 11 in Appendix C).

The coefficients $\{c_i\}_{i=0}^k$ are determined by minimizing a weighted combination of the residuals $\mathrm{Exp}_{x_i}^{-1}(x_{i+1}) \in T_{x_i}\mathcal{M}, i = 0, ..., k$. Specifically, we consider the following optimization problem:

$$\min_{c \in \mathbb{R}^{k+1}: c^\top 1 = 1} \| \sum_{i=0}^k c_i r_i \|_{x_k}^2 + \lambda \|c\|_2^2, \tag{2}$$

which is a linear system of dimension $k+1$ and has a simple closed-form solution (see Proposition 2 in Appendix E). Here, $r_i = \Gamma_{x_i}^{x_k} \mathrm{Exp}_{x_i}^{-1}(x_{i+1}) \in T_{x_k}\mathcal{M}$ and $\Gamma_{x_i}^{x_k}$ is the parallel transport from x_i to x_k.

Our Riemannian nonlinear acceleration (RiemNA) strategy is presented in Algorithm 1, which takes a sequence of non-diverging iterates from any solver as input and constructs an extrapolation using coefficients $\{c_i\}_{i=0}^k$ that solve (2). The extrapolation is performed in parallel to the update of the iterate sequence. Note that when the manifold $\mathcal{M}$ is the Euclidean space, Algorithm 1 exactly recovers the nonlinear acceleration algorithm in Scieur et al. (2020).

Algorithm 1 Riemannian nonlinear acceleration (RiemNA)

```
Input: Iterate sequence $x_0, ..., x_{k+1}$. Regularization parameter $\lambda$.
Compute $r_i = \Gamma_{x_i}^{x_k} \mathrm{Exp}_{x_i}^{-1}(x_{i+1}), i = 0, ..., k$.
Solve $c^* = \arg\min_{c \in \mathbb{R}^{k+1}: c^\top 1 = 1} \| \sum_{i=0}^k c_i r_i \|_{x_k}^2 + \lambda \|c\|_2^2$.
Output: $\bar{x}_{c^*,x} = \tilde{x}_k$ computed from $\tilde{x}_i = \mathrm{Exp}_{\tilde{x}_{i-1}}\Big( \frac{c_i^*}{\sum_{j=0}^i c_j^*} \mathrm{Exp}^{-1}_{\tilde{x}_{i-1}}(x_i) \Big)$, with $\tilde{x}_{-1} = x_0$.
```

4 CONVERGENCE ACCELERATION FOR RIEMANNIAN GRADIENT DESCENT

This section analyzes the convergence acceleration of RiemNA (Algorithm 1) when the iterates are generated by the Riemannian gradient descent (RGD) method (Absil et al., 2009). In particular, we show that the extrapolated point (output of Algorithm 1) is a good estimate of the optimal solution and bound its distance to optimality. We start by making the following assumption throughout the paper.

Assumption 1. Let $x^* \in \mathcal{M}$ be a (strictly) local minimizer of f. The iterates generated, i.e., $x_0, x_1, \ldots$ stay within a neighbourhood $\mathcal{X}$ around x^* with unique geodesic. Furthermore, the sequence of iterates is non-divergent, i.e., $d(x_k, x^*) = O(d(x_0, x^*))$ for all $k \geq 0$.

The former condition in Assumption 1 ensures the exponential map is invertible and is standard for analyzing accelerated gradient methods on manifolds (Ahn and Sra, 2020; Jin and Sra, 2022; Kim and Yang, 2022). This condition is satisfied for any non-positively curved manifolds, such as symmetric positive definite (SPD) manifold with the affine-invariant metric (Bhatia, 2009). In addition, this also holds true for any sufficiently small subset $\mathcal{X}$ of any manifold.

Linear Iterates and Error Decomposition in the Euclidean Space First, we recall that the convergence analysis for *Euclidean* nonlinear acceleration (Scieur et al., 2020) relies critically on a sequence of linear fixed-point iterates that satisfy $\hat{x}_i - x^* = G(\hat{x}_{i-1} - x^*)$ for some positive semi-definite and contractive matrix G (with $\|G\|_2 < 1$). The main idea is to show that the algorithm converges optimally on $\hat{x}_i$ and then bound the deviation arising from the nonlinearity. In particular, let $\{x_i\}_{i=0}^k$ be the given iterates and $\{\hat{x}_i\}_{i=0}^k$ be the linear iterates. Consider $c^*, \hat{c}^*$ as the coefficients solving (2) in the Euclidean setup using $\{x_i\}_{i=0}^k$, $\{\hat{x}_i\}_{i=0}^k$ respectively. The convergence analysis in Scieur et al. (2020) aims to bound each term from the error decomposition:

$$\sum_{i=0}^{k} c_i^* x_i - x^* \tag{3}$$
$$= \underbrace{\sum_{i=0}^{k} \hat{c}_i^* \hat{x}_i - x^*}_{\text{Linear term}} + \underbrace{\sum_{i=0}^{k} (c_i^* - \hat{c}_i^*)\hat{x}_i}_{\text{Stability}} + \underbrace{\sum_{i=0}^{k} c_i^* (x_i - \hat{x}_i)}_{\text{Nonlinearity}} .$$

From Linearized Iterates to Iterates on Manifolds On general Riemannian manifolds, due to the curved geometry, it becomes nontrivial to generalize the error decomposition (3) to manifolds. Nevertheless, we start with identification of linearized iterates on manifolds in the tangent space of x^*. For notational convenience, we denote $\Delta_x := \mathrm{Exp}_{x^*}^{-1}(x)$ for any $x \in \mathcal{X}$. We now consider the linearized iterates $\hat{x}_i$ produced by the following progression as

$$\Delta_{\hat{x}_i} = G[\Delta_{\hat{x}_{i-1}}], \tag{4}$$

for some $G : T_{x^*}\mathcal{M} \to T_{x^*}\mathcal{M}$ as a self-adjoint, positive semi-definite operator with $\|G\|_{x^*} \le \sigma < 1$, where we denote $\|A\|_{x^*}$ as the operator norm for any linear operator A on the tangent space $T_{x^*}\mathcal{M}$. In fact, we show in Lemma 3 that the progression of iterates from the Riemannian gradient descent method is locally linear on the tangent space of the local minimizer x^*, thus satisfying (4) up to some error term. This requires the following regularity assumption on the objective function f.

Assumption 2. The function f has geodesic Lipschitz gradient and Lipschitz Hessian.

Remark 1. Assumption 2 is used to ensure sufficient smoothness of the function such that the Riemannian gradient and Hessian are bounded at optimality.

Lemma 3. *Under Assumptions 1, 2, suppose the iterates generated by the Riemannian gradient descent method are* $x_{i+1} = \mathrm{Exp}_{x_i}(-\eta\, \mathrm{grad} f(x_i))$. *Then, we have*

$$\Delta_{x_i} = \big(\mathrm{id} - \eta\, \mathrm{Hess} f(x^*)\big)[\Delta_{x_{i-1}}] + \varepsilon_i$$

where id *denotes the identity operator and* $\|\varepsilon_i\|_{x^*} = O(d^2(x_i, x^*))$ *and* $\varepsilon_0 = 0$.

Lemma 3 suggests that it is reasonable to consider the linearized iterates $\{\hat{x}_k\}$ defined in (4) where $G = \mathrm{id} - \eta\, \mathrm{Hess} f(x^*)$. It is clear that for a strictly local minimizer x^*, there exists $\mu, L > 0$ such that $\mu\, \mathrm{id} \preceq \mathrm{Hess} f(x^*) \preceq L\, \mathrm{id}$. This is irrespective of whether the function f is geodesic strongly convex or has geodesic Lipschitz gradient. Thus, for proper choices of η, we can always ensure G is positive semi-definite and contractive.

In this paper, the convergence analysis focus on the case when $G = \mathrm{id} - \eta\, \mathrm{Hess} f(x^*)$ and $\{x_i\}$ are given by Riemannian gradient descent to simplify the bounds. However, we highlight that most of the analysis holds for more general and symmetric G.

Hence, the error in the manifold weighted average $\bar{x}_{c^*,x}$ computed from (Avg.1) leads to the decomposition (due to triangle inequality of Riemannian distance):

$$\begin{aligned} &d(\bar{x}_{c^*,x}, x^*) \\ &\le \underbrace{d(\bar{x}_{\hat{c}^*,\hat{x}}, x^*)}_{\text{Linear term}} + \underbrace{d(\bar{x}_{\hat{c}^*,\hat{x}}, \bar{x}_{c^*,\hat{x}})}_{\text{Stability}} + \underbrace{d(\bar{x}_{c^*,\hat{x}}, \bar{x}_{c^*,x})}_{\text{Nonlinearity}}, \end{aligned}$$

where we denote $\hat{c}^*$ as the coefficients solving (2) with the residuals $\hat{r}_i = \Delta_{\hat{x}_{i+1}} - \Delta_{\hat{x}_i}$ from the linearized iterates $\{\hat{x}_i\}$ in (4) and $\bar{x}_{\hat{c}^*,\hat{x}}$, $\bar{x}_{c^*,\hat{x}}$ as weighted average computed using pairs $\{(\hat{c}_i^*, \hat{x}_i)\}_{i=0}^k$ and $\{(c_i^*, \hat{x}_i)\}_{i=0}^k$ respectively. Before we bound each of the error term, we first present a lemma relating the averaging on manifolds to averaging on the tangent space.

Lemma 4. *Under Assumption 1, for some coefficients* $\{c_i\}_{i=0}^k$ *with* $\sum_{i=0}^k c_i = 1$ *and any iterate sequence* $\{x_i\}_{i=0}^k$, *consider* $\bar{x}_{c,x}$ *computed from* (Avg.1) *via the given coefficients and the iterates. Then, we have* $\Delta_{\bar{x}_{c,x}} = \sum_{i=0}^k c_i \Delta_{x_i} + e$, *where* $\|e\|_{x^*} = O(d^3(x_0, x^*))$.

Remark 2. Lemma 4 shows that the error between the averaging on the manifold and averaging on the tangent space is on the order of $O(d^3(x_0, x^*))$. This relies heavily on the metric distortion bound given in Lemma 1, 2, which only holds for the case of exponential map and parallel transport. Nevertheless, we highlight that when the general retraction and vector transport are used, we can follow the idea of (Tripuraneni et al., 2018, Lemma 12) to show the error is on the order of $O(d^2(x_0, x^*))$. Please see Proposition 3 in Appendix F and Section 6 for more details where we discuss convergence under a more general setup.

Error Bound From the Linear Term We show that extrapolation using the linearized iterates converges in a near-optimal rate, via the regularized Chebyshev polynomial. This generalizes the development of Scieur et al. (2020) (in the Euclidean setting) to manifolds.

Definition 1 (Regularized Chebyshev polynomial (Scieur et al., 2020))**.** The regularized Chebyshev polynomial of degree k, in the range of $[0, \sigma]$ with a regularization parameter α, denoted as $C_{k,\alpha}^{[0,\sigma]}(x)$ is defined as $C_{k,\alpha}^{[0,\sigma]}(x) = \arg\min_{p \in \mathcal{P}_k^1} \max_{x \in [0,\sigma]} p^2(x) + \alpha\|p\|_2^2$, where we denote $\mathcal{P}_k^1 := \{p \in \mathbb{R}[x] : \deg(p) = k, p(1) = 1\}$ as the set of polynomials of degree k with coefficients summing to 1 and $\|p\|_2$ is the Euclidean norm of the coefficients of the polynomial p. We write the maximum valued as $S_{k,\alpha}^{[0,\sigma]} := \sqrt{\max_{x \in [0,\sigma]} (C_{k,\alpha}^{[0,\sigma]}(x))^2 + \alpha\|C_{k,\alpha}^{[0,\sigma]}(x)\|_2^2}$.

In Lemma 5, we present the error bound coming from the linear term, which follows from the definition of regularized Chebyshev polynomial and Lemma 4. Due to the curvature of the manifold, we observe an additional error term $O(d^3(x_0, x^*))$ compared to the Euclidean counterpart, which becomes insignificant as approaching optimality.

Lemma 5 (Error from the linear term)**.** *Under Assumption 1, let $\bar{x}_{\hat{c}^*,\hat{x}}$ be computed from* (Avg.1) *using $\{(\hat{c}_i^*, \hat{x}_i)\}_{i=0}^k$. Then, $d(\bar{x}_{\hat{c}^*,\hat{x}}, x^*) \leq \frac{d(x_0,x^*)}{1-\sigma}\sqrt{(S_{k,\bar{\lambda}}^{[0,\sigma]})^2 - \frac{\lambda}{d^2(x_0,x^*)}\|\hat{c}^*\|_2^2} + \epsilon_1$, where $\bar{\lambda} = \lambda/d^2(x_0, x^*)$ and $\epsilon_1 = O(d^3(x_0, x^*))$.*

Error Bound From Coefficient Stability We now bound the deviation between the optimal coefficients computed via the Riemannian gradient descent iterates $\{x_i\}$ and the linearized iterates $\{\hat{x}_i\}$. To this end, we require the following result on the coefficients.

Lemma 6 (Bound on norm of coefficients)**.** *Under Assumptions 1, 2, let the coefficients $c^*, \hat{c}^*$ be solved from* (2) *using $\{x_i\}, \{\hat{x}_i\}$ respectively, where $\{x_i\}$ are given by the Riemannian gradient descent and $\{\hat{x}_i\}$ satisfy* (4)*. Then, we have $\|c^*\|_2 \leq \sqrt{\frac{\sum_{i=0}^k d^2(x_i,x_{i+1})+\lambda}{(k+1)\lambda}}$ and $\|c^* - \hat{c}^*\|_2 \leq \frac{1}{\lambda}\Big(\frac{2d(x_0,x^*)}{1-\sigma}\psi + (\psi)^2\Big)\|\hat{c}^*\|_2$ for some $\psi = O(d^2(x_0, x^*))$.*

It should be noted that in the Euclidean space, $\psi = \sum_{i=0}^k \|\Delta_{x_i} - \Delta_{\hat{x}_i}\|_2 = \|x_i - \hat{x}_i\|_2$ and also can be shown to have an order of $O(d^2(x_0, x^*))$ under certain Lipschitz conditions on the function (see Proposition 3.8 in Scieur et al. (2020)). On manifolds, the term ψ suffers from additional distortion coming from the metric, which is also on the order $O(d^2(x_0, x^*))$.

Based on Lemma 6, the error from coefficient stability can now be bounded as follows. The proof follows from linearizing the weighted average on the tangent space $T_{x^*}\mathcal{M}$ where we bound the deviation arising from the coefficients. Hence, an extra error ϵ_2 appears in the bound.

Lemma 7 (Error from coefficient estimation)**.** *Under the same settings as in Lemma 6, let $\bar{x}_{\hat{c}^*,\hat{x}}$, $\bar{x}_{c^*,\hat{x}}$ be computed from* (Avg.1) *using $\{(\hat{c}_i^*, \hat{x}_i)\}_{i=0}^k$ and $\{(c_i^*, \hat{x}_i)\}_{i=0}^k$ respectively. Then, $d(\bar{x}_{\hat{c}^*,\hat{x}}, \bar{x}_{c^*,\hat{x}}) \leq \frac{C_1}{\lambda(1-\sigma)}\Big(\frac{2d^2(x_0,x^*)}{1-\sigma}\psi + d(x_0, x^*)(\psi)^2\Big)\|\hat{c}^*\|_2 + \epsilon_2$, for some $\psi = O(d^2(x_0, x^*)), \epsilon_2 = O(d^3(x_0, x^*))$.*

Error Bound From Nonlinearity Next, we show that the nonlinearity term can be bounded in Lemma 8, which follows a similar idea of linearization on a fixed tangent space. Additional error ϵ_3 is again due to the curvature of the manifold, which vanishes when $\mathcal{M}$ is the Euclidean space.

Lemma 8 (Error from the nonlinearity)**.** *Under the same settings as in Lemma 6, we have $d(\bar{x}_{c^*,\hat{x}}, \bar{x}_{c^*,x}) \leq C_1\sqrt{\frac{\sum_{i=0}^k d^2(x_i,x_{i+1})+\lambda}{(k+1)\lambda}}\Big(\sum_{i=0}^k\sum_{j=0}^i \|\varepsilon_j\|_{x^*}\Big) + \epsilon_3$, where $\|\varepsilon_j\|_{x^*} = O(d^2(x_j, x^*))$ is defined in Lemma 3 and $\epsilon_3 = O(d^3(x_0, x^*))$.*

Finally, we combine Lemmas 5, 7, 8 to obtain the following convergence result for Algorithm 1 when the iterates are generated from the Riemannian gradient descent (RGD).

Theorem 1 (Convergence of RiemNA with RGD iterates)**.** *Under Assumptions 1, 2, let $\{x_i\}_{i=0}^k$ be given by the Riemannian gradient descent method, i.e., $x_{i+1} = \mathrm{Exp}_{x_i}(-\eta\,\mathrm{grad}f(x_i))$ and $\{\hat{x}_i\}_{i=0}^k$ be the linearized iterates satisfying $\Delta_{\hat{x}_i} = G[\Delta_{\hat{x}_{i-1}}]$ with $G = \mathrm{id} - \eta\,\mathrm{Hess}f(x^*)$, satisfying $\|G\|_{x^*} \leq \sigma < 1$. Then, Algorithm 1 with regularization parameter λ produces $\bar{x}_{c^*,x^*}$ that satisfies $d(\bar{x}_{c^*,x}, x^*) \leq d(x_0, x^*)\frac{S_{k,\bar{\lambda}}^{[0,\sigma]}}{1-\sigma}\sqrt{1 + \frac{C_1^2 d^2(x_0,x^*)\big(\frac{2d(x_0,x^*)}{1-\sigma}\psi+(\psi)^2\big)^2}{\lambda^3}} + C_1\sqrt{\frac{\sum_{i=0}^k d^2(x_i,x_{i+1})+\lambda}{(k+1)\lambda}}\Big(\sum_{i=0}^k\sum_{j=0}^i \|\varepsilon_j\|_{x^*}\Big) + \epsilon_1 + \epsilon_2 + \epsilon_3$, where $\psi = O(d^2(x_0, x^*)), \epsilon_1, \epsilon_2, \epsilon_3 = O(d^3(x_0, x^*))$ and $\varepsilon_i = O(d^2(x_i, x^*))$ is defined in Lemma 3.*

We prove that even with additional distortion from the curved geometry of the manifold, the asymptotic optimal convergence is still guaranteed. This is mainly due to the fact that all errors incurred by the metric distortion, i.e., $\epsilon_1, \epsilon_2, \epsilon_3$ are on the order of at least $O(d^2(x_0, x^*))$, which is primarily attributed to Lemma 4.

Proposition 1 (Asymptotic optimal convergence rate of RiemNA with RGD iterates)**.** *Under the same settings as in Theorem 1, set $\lambda = O(d^s(x_0, x^*))$ for $s \in (2, \frac{8}{3})$. Then, $\lim_{d(x_0,x^*)\to 0} \frac{d(\bar{x}_{c^*,x},x^*)}{d(x_0,x^*)} \leq \frac{1}{1-\sigma}\frac{2}{\beta^{-k}+\beta^k}$, where $\beta = \frac{1-\sqrt{1-\sigma}}{1+\sqrt{1-\sigma}}$.*

Remark 3. The asymptotic optimal convergence rate holds as long as $\epsilon_1, \epsilon_2, \epsilon_3$ are on the order of at least $O(d^2(x_0, x^*))$ such that $\lim_{d(x_0,x^*)\to 0}\frac{1}{d(x_0,x^*)}(\epsilon_1 + \epsilon_2 + \epsilon_3) = 0$.

Remark 4. Suppose at a (strictly) local minimizer, we have $0 \prec \mu\,\mathrm{id} \preceq \mathrm{Hess}f(x^*) \preceq L\,\mathrm{id}$. Then, by choosing $\eta = \frac{1}{L}$, we have $\sigma = 1 - \frac{\mu}{L}$. This corresponds to the optimal convergence rate obtained by Nesterov acceleration (Nesterov, 2003) and its Riemannian extensions such as Liu et al. (2017); Ahn and Sra (2020); Kim and Yang (2022) for geodesic strongly convex functions.

Implementation and Complexity Algorithm 2 presents an implementation for the proposed RiemNA strategy when the iterates are given by Riemannian gradient descent (RGD) method with fixed stepsize. Specifically, we run RGD to produce the iterate sequence $x_0, \ldots, x_{m-1}$, where m is the memory depth. Then, we compute $\bar{x}_{c^*,x}$ with these iterates by Algorithm 1. We then restart Riemannian gradient descent with $x_0 = \bar{x}_{c^*,x}$ for the next epoch. It should be noted that in this case, we do not require the inverse exponential map for computing the residuals.

RGD+RiemNA requires T RGD updates and $\lceil T/m \rceil$ calls to RiemNA. Overall, Algorithm 2 needs $T + \lceil T/m \rceil m$ calls to the exponential map and $\lceil T/m \rceil m$ calls each to the parallel transport and the inverse exponential map operations.

Algorithm 2 RGD+RiemNA

```
**Input:** Initialization $x_0$, stepsize $\eta$, regularization parameter $\lambda$, and memory depth $m$.
Set $t = 0$.
**while** $t \leq T$ **do**
    **for** $i = 1, ..., m$ **do**
        $x_i = \mathrm{Exp}_{x_{i-1}}(-\eta \, \mathrm{grad} f(x_{i-1}))$.
        $t = t + 1$.
    **end for**
    $r_i = -\eta \, \Gamma_{x_i}^{x_{m-1}} \mathrm{grad} f(x_i), \ i = 0, ..., m-1$.
    Solve $c^* = \arg\min_{c \in \mathbb{R}^m : c^\top 1 = 1} \| \sum_{i=0}^{k} c_i r_i \|^2_{x_{m-1}} + \lambda \|c\|_2^2$.
    Set $\bar{x}_{c^*,x} = \tilde{x}_{m-1}$ computed from $\tilde{x}_i = \mathrm{Exp}_{\tilde{x}_{i-1}}\big( \frac{c_i^*}{\sum_{j=0}^{i} c_j^*} \mathrm{Exp}^{-1}_{\tilde{x}_{i-1}}(x_i) \big)$, with $\tilde{x}_{-1} = x_0$.
    Restart with $x_0 = \bar{x}_{c^*,x}$.
**end while**
```

This is as efficient as the most practical implementation of the Riemannian Nesterov accelerated gradient methods (Zhang and Sra, 2018; Kim and Yang, 2022) (discussed in Appendix A.2) that require $2T$ calls each to the exponential and inverse exponential map operations.

5 ALTERNATIVE AVERAGING SCHEMES

In this section, we propose alternative averaging schemes on manifolds used for extrapolation. For the iterates obtained from the Riemannian gradient descent method, we show the schemes ensure the same asymptotically optimal convergence rate obtained in Proposition 1.

The first scheme we consider is based on the following equality in the Euclidean space for the weighted mean, i.e., $\sum_{i=0}^{k} c_i x_i = x_k - (\sum_{i=0}^{k-1} c_i)(x_k - x_{k-1}) - (\sum_{i=0}^{k-2} c_i)(x_{k-1} - x_{k-2}) - \cdots - c_0(x_1 - x_0)$. Accordingly, let $\theta_i = \sum_{j=0}^{i} c_j, i = 0, ..., k-1$. We define an alternative weighted averaging as

$$\bar{x}_{c,x} = \mathrm{Exp}_{x_k}\Big(- \sum_{i=0}^{k-1} \theta_i \Gamma_{x_i}^{x_k} \mathrm{Exp}_{x_i}^{-1}(x_{i+1}) \Big). \qquad \text{(Avg.2)}$$

Based on the earlier analysis, to show the convergence of $\bar{x}_{c,x}$ defined in (Avg.2), we only require to show that Lemma 4 holds for the new scheme, with an error of order at least $O(d^2(x_0, x^*))$. We formalize this claim in the next lemma and show the error is in fact on the order of $O(d^3(x_0, x^*))$.

Lemma 9. *Under Assumption 1, for some coefficients* $\{c_i\}_{i=0}^k$ *with* $\sum_{i=0}^k c_i = 1$ *and iterates* $\{x_i\}_{i=0}^k$, *consider* $\bar{x}_{c,x} = \mathrm{Exp}_{x_k}\big(- \sum_{i=0}^{k-1} \theta_i \Gamma_{x_i}^{x_k} \mathrm{Exp}_{x_i}^{-1}(x_{i+1}) \big)$, $\theta_i = \sum_{j=0}^{i} c_j$. *Then, we have* $\|\Delta_{\bar{x}_{c,x}} - \sum_{i=0}^k c_i \Delta_{x_i}\|_{x^*} = O(d^3(x_0, x^*))$.

Lemma 9 allows convergence under the averaging scheme (Avg.2) to be established exactly following the same steps as before. This is sufficient to show that the same convergence bounds hold, i.e., Theorem 1 and Proposition 1.

Weighted Fréchet Mean In addition, we discuss the weighted Fréchet mean in Appendix B, which can also be used in place of the two aforementioned averaging schemes. We have provided similar error bounds as in Lemma 9 that lead to similar convergence guarantees.

6 CONVERGENCE UNDER GENERAL RETRACTION AND VECTOR TRANSPORT

In this section, we generalize our convergence results for RiemNA with general retraction and vector transport operations. To the best of our knowledge, Riemannian acceleration has not been studied under general retraction and vector transport. To this end, we make the following standard assumptions, which include bounding the deviation between retraction and exponential map as well as between vector transport and parallel transport. In addition, we require the Lipschitz gradient and Hessian to be compatible with retraction and vector transport.

Assumption 3. The neighbourhood $\mathcal{X}$ is totally retractive where retraction has a smooth inverse. Function f has retraction Lipschitz gradient and Lipschitz Hessian.

Assumption 4. There exists constants $a_0, a_1, a_2, \delta_{a_0,a_1} > 0$ such that for all $x, y, z \in \mathcal{X}$, $\|\mathrm{Retr}_x^{-1}(y)\|_x \leq \delta_{a_0,a_1}$, we have (1). $a_0 d(x,y) \leq \|\mathrm{Retr}_x^{-1}(y)\|_x \leq a_1 d(x,y)$ and (2). $\|\mathrm{Exp}_x^{-1}(z) - \mathrm{Retr}_x^{-1}(z)\|_x \leq a_2 \|\mathrm{Retr}_x^{-1}(z)\|_x^2$.

Assumption 5. The vector transport $\mathcal{T}_x^y$ is isometric and there exists a constant $a_3 > 0$ such that for all $x, y \in \mathcal{X}$, $\|\mathcal{T}_x^y u - \Gamma_x^y u\|_y \leq a_3 \|\mathrm{Retr}_x^{-1}(y)\|_x \|u\|_x$.

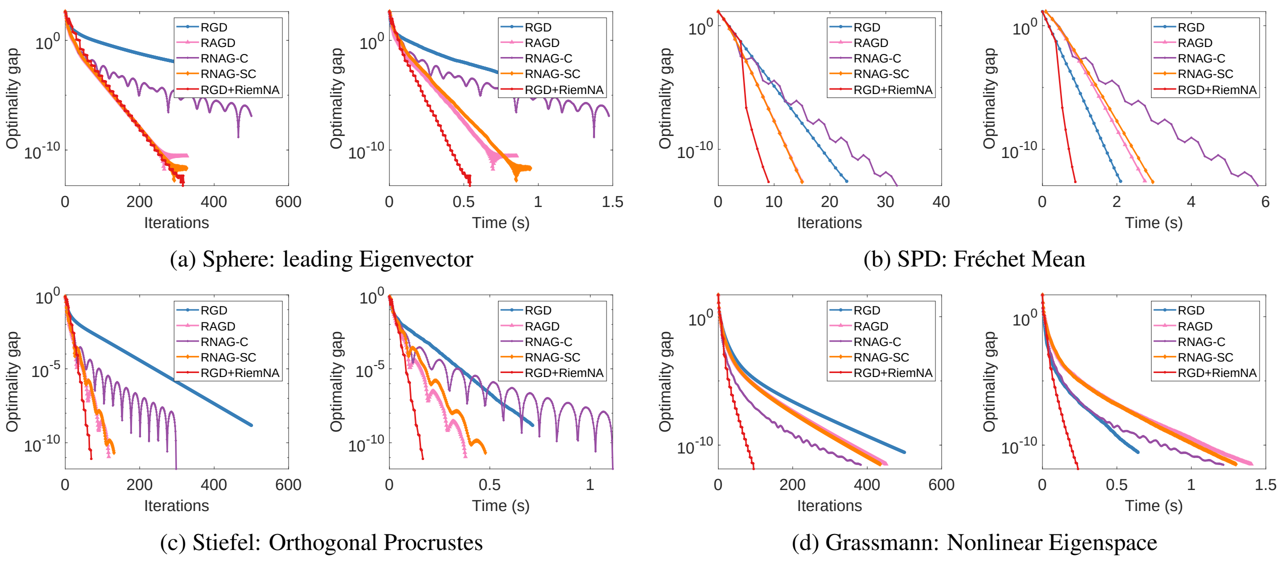
Assumptions 3-5 are commonly used for analyzing Riemannian first-order algorithms with retraction and vector transport (Ring and Wirth, 2012; Huang et al., 2015b; Sato et al., 2019; Kasai et al., 2018; Han and Gao, 2021a).

In this section, we only show convergence under the recursive weighted average computation for extrapolation, i.e.,

$$\bar{x}_{c,x} = \tilde{x}_k, \qquad \tilde{x}_i = \mathrm{Retr}_{\tilde{x}_{i-1}}\Big(\frac{c_i}{\sum_{j=0}^{i} c_j} \mathrm{Retr}_{\tilde{x}_{i-1}}^{-1}(x_i) \Big). \tag{5}$$

Similar analysis can be also performed on the alternative two averaging schemes discussed in Section 5.

The next theorem shows that asymptotic optimal convergence rate can also be achieved using retraction and vector transport. The proof is similar to the case for exponential map and parallel transport and employs the Assumptions 4, 5. In particular, both these two assumptions ensure the devi-



(a) Sphere: leading Eigenvector

(b) SPD: Fréchet Mean

(c) Stiefel: Orthogonal Procrustes

(d) Grassmann: Nonlinear Eigenspace

Figure 1: Comparing proposed RGD+RiemNA with existing approaches: RGD, RAGD, RNAG-C, and RNAG-SC. We observe that RGD+RiemNA outperforms all the baselines.

ations between retraction and exponential map, vector transport and parallel transport are on the order of $O(d^2(x_0, x^*))$. Thus, the additional error terms $\epsilon_1, \epsilon_2, \epsilon_3 = O(d^2(x_0, x^*))$.

Theorem 2 (Convergence under general retraction and vector transport)**.** *Under Assumptions 1, 3, 4, and 5, let* $\{x_i\}_{i=0}^k$ *be given by Riemannian gradient descent via retraction, i.e.,* $x_i = \mathrm{Retr}_{x_{i-1}}(-\eta\,\mathrm{grad} f(x_{i-1}))$ *and* $\{\hat{x}_i\}_{i=0}^k$ *be the linearized iterates satisfying* $\mathrm{Retr}_{x^*}^{-1}(\hat{x}_i) = G[\mathrm{Retr}_{x^*}^{-1}(\hat{x}_{i-1})]$ *with* $G = \mathrm{id} - \eta\,\mathrm{Hess} f(x^*)$*, satisfying* $\|G\|_{x^*} \leq \sigma < 1$*. Then, using retraction and vector transport in Algorithm 1 and letting* $\bar{x}_{c,x}$ *be computed from* (5)*, the same asymptotic optimal convergence rate (Proposition 1) holds under the same choice of* $\lambda = O(d^s(x_0, x^*))$*,* $s \in (2, \frac{8}{3})$.

Theorem 2 allows Algorithm 2 to be implemented with general retraction and vector transport without affecting the optimal convergence rate achieved asymptotically.

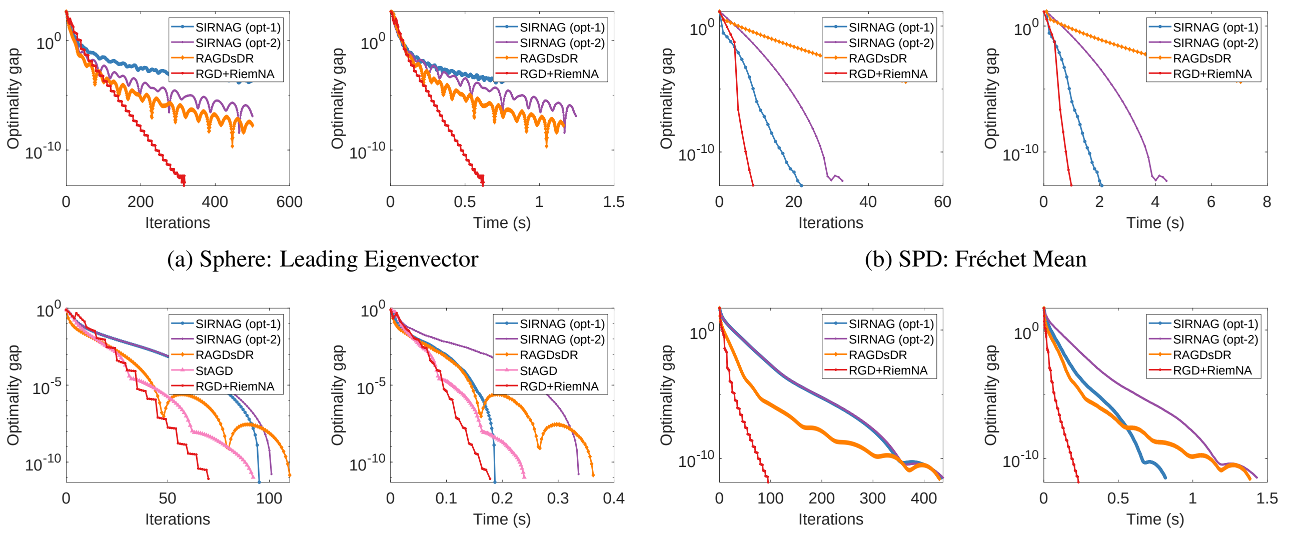
7 EXPERIMENTS

In this section, we evaluate the performance of our Riemannian nonlinear acceleration (RiemNA) strategy on various applications. For RiemNA, we only consider the recursive weighted average in (Avg.1) for the main experiments. The codes can be found on `https://github.com/andyjm3/RiemNA`.

Baselines We compare the proposed RGD+RiemNA (Algorithm 2) with state-of-the-art Riemannian Nesterov accelerated gradient (RNAG) methods (Kim and Yang, 2022). We also include RAGD, a variant of Nesterov acceleration on manifolds proposed in (Zhang and Sra, 2018), and RGD as baselines. In particular, we compare with RNAG-C (Kim and Yang, 2022) (designed for geodesic convex functions) and RNAG-SC (Kim and Yang, 2022) and RAGD (Zhang and Sra, 2018) (designed for geodesic strongly convex functions) regardless of whether the objective is of the particular class. More details of the algorithms are in Appendix A.2.

Parameters RNAG-C, RNAG-SC, and RAGD require the knowledge of geodesic Lipschitz constant L (Kim and Yang, 2022). Further, RNAG-SC and RAGD require the geodesic strong convexity parameter μ. In particular, the stepsize of RNAG-C, RNAG-SC and RAGD should be set as $1/L$. If such constants are available, we set them accordingly. Otherwise, we tune over the parameters L, μ for RNAG-C, RNAG-SC to obtain the best results and set the same parameters for RAGD for comparability. Following Kim and Yang (2022), the additional parameters ξ, ζ are fixed to be 1 for RNAG-C, RNAG-SC and $\beta = \sqrt{\mu/L}/5$ for RAGD. We set stepsize of RGD to be $1/L$ if available and tune the stepsize otherwise. For the proposed RGD+RiemNA, we fix $\lambda = 10^{-8}$ and choose memory depth $m \in \{5, 10\}$. It should be emphasized that RGD+RiemNA is agnostic to function specific constants.

For fair comparisons, we use exponential map, inverse exponential map, and parallel transport for all the algorithms whenever such operations are properly defined. For other cases, we use retraction, inverse retraction, and vector transport even though the baseline acceleration methods are not analyzed under such general operations. We emphasize that we maintain consistency in the use of these operations across all the algorithms. The experiments are coded in Matlab using Manopt (Boumal et al., 2014). The stopping criterion for all the algorithms is gradient norm reaching below 10^{-6}.



(a) Sphere: Leading Eigenvector

(b) SPD: Fréchet Mean

(c) Stiefel: Orthogonal Procrustes

(d) Grassmann: Nonlinear Eigenspace

Figure 2: Comparing RGD+RiemNA with additional approaches: SIRNAG, RAGDsDR, and StAGD. SIRNAG (opt-1) and (opt-2) represent SIRNAG with two update options. We observe that RGD-RiemNA maintains its superior performance.

Applications We consider four applications: leading eigenvector computation (Absil et al., 2007), Fréchet mean of symmetric positive definite (SPD) matrices with the affine-invariant metric (Bhatia, 2009), orthogonal Procrustes problem (Eldén and Park, 1999), and the nonlinear eigenspace problem (Zhao et al., 2015). These applications solve problems on sphere, SPD, Stiefel, and Grassmann manifolds respectively. See Appendix A.1 for detailed introduction of the manifolds, along with the relevant operations required for the experiments. We highlight that except for the task of Fréchet mean which is geodesic strongly convex, other problems are in general nonconvex.

Leading Eigenvector Computation The problem computes the leading eigenvector of a symmetric matrix A of size $d \times d$, by solving $\min_{x \in \mathcal{S}^{d-1}} \{f(x) := -\frac{1}{2} x^\top A x\}$, where $\mathcal{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ denotes the sphere manifold of intrinsic dimension $d-1$. For the experiment, we generate a positive definite matrix A with condition number 10^3 and exponentially decaying eigenvalues in dimension $d = 10^3$. As shown in (Kim and Yang, 2022, Proposition 7.1), the problem has geodesic L-Lipschitz gradient with L to be the eigengap of matrix A, i.e., the difference between maximum and minimum eigenvalues of A. The optimal solution of the problem is given by $-\frac{1}{2}\lambda_{\max}(A)$, where $\lambda_{\max}$ extracts the largest eigenvalue of A.

The stepsize is thus set as $1/L$ for all methods. For RNAG-SC and RAGD, we set $\mu = 10$. For RiemNA, we set memory depth to be $m = 10$. We use exponential and inverse exponential map as well as projection-type vector transport for all algorithms including RGD+RiemNA.

Fréchet Mean of SPD Matrices We consider the problem of computing the Fréchet mean of symmetric positive definite (SPD) matrices $\{A_i\}_{i=1}^N$ of size $d \times d$ under the affine-invariant metric (Bhatia, 2009), i.e., $\min_{X \in \mathbb{S}^d_{++}} \frac{1}{2N} \sum_{i=1}^N \|\mathrm{logm}(X^{-1/2} A_i X^{-1/2})\|_{\mathrm{F}}^2$. Here, $\mathbb{S}^d_{++}$ is the set of SPD matrices of size $d \times d$, $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm, and $\mathrm{logm}(\cdot)$ is the matrix logarithm. To trace the optimality gap, we compute the optimal solution of the problem by running R-LBFGS method (Huang et al., 2016) until the gradient norm falls below 10^{-10}.

For the experiments, we use exponential map and its inverse as well as the parallel transport for all the algorithms. As commented previously, the geometry is negatively curved, and thus, the Fréchet mean problem is geodesic 1-strongly convex ($\mu = 1$). For this problem, we generate random $N = 100$ SPD matrices of dimension $d = 10$. The stepsize for all methods are tuned and set to be 0.5. For RiemNA, we set memory depth $m = 5$.

Orthogonal Procrustes Problem We also consider the orthogonal Procrustes problem on the Stiefel manifold (Eldén and Park, 1999). Suppose we are given $A \in \mathbb{R}^{r \times r}, B \in \mathbb{R}^{p \times r}$, the objective is $\min_{X \in \mathrm{St}(p,r)} \|XA - B\|_{\mathrm{F}}^2$ where $\mathrm{St}(p,r) := \{X \in \mathbb{R}^{p \times r} : X^\top X = I\}$ is the set of column orthonormal matrices, which forms the so-called Stiefel manifold with the canonical metric. The optimal solution is similarly computed by running R-LBFGS.

To implement the algorithms, we use QR-based retraction and inverse retraction as well as projection-type vector transport. We generate random matrices A, B where the entries are normal distributed. We set $p = 100, r = 5$. For this problem, both L and μ are unknown. Hence, we tune and set stepsize to be 1 for all methods. For RNAG-SC and RAGD, we select $\mu = 0.005$ and for RiemNA we set memory depth $m = 5$.

Nonlinear Eigenspace Problem Finally, the problem of computing nonlinear eigenspace arises as the total energy minimization on the Grassmann manifold (Zhao et al., 2015), i.e., $\min_{X \in \mathrm{Gr}(p,r)} \frac{1}{2}\mathrm{tr}(X^\top L X) + \frac{1}{4}\rho(X)^\top L^{-1}\rho(X)$ where $\rho(X) := \mathrm{diag}(XX^\top)$ and L is a discrete Laplacian operator. The optimal solution is similarly computed by running R-LBFGS.

For experiment, we implement the algorithms with QR-based retraction and inverse retraction as well as projection-based vector transport similar to Stiefel manifold. We generate L as a tridiagonal matrix with main diagonal entries to be 2 and sub- and super-diagonal entries to be -1. The stepsize is tuned and set to be 0.1 for all methods and for RNAG-SC, RAGD, $\mu = 5$ and for RiemNA, $m = 5$.

Results In Figure 1, we plot optimality gap, $f(x_t) - f(x^*)$, against both iteration number and runtime for all the algorithms. We make the following observations:

- Proposed RGD+RiemNA consistently outperforms the baselines in runtime across all the applications.
- In iteration counts as well, RGD+RiemNA is consistently better than others in all the applications except in the leading eigevector problem, where RGD+RiemNA matches the performance of RAGD and RNAG-SC.
- In Figure 1a, RGD+RiemNA is faster than RAGD and RNAG-SC even though the number of iterations needed are similar. This implies that RGD+RiemNA is computationally more efficient. This is in accordance with RGD+RiemNA requiring fewer number of calls to manifold operations like exponential map (or retraction) and parallel transport (or vector transport).
- For the SPD Fréchet mean problem, which is geodesic strongly convex, RGD+RiemNA consistently exhibits faster convergence than others where the extrapolation step leads to significant convergence acceleration.
- RGD+RiemNA does not necessarily ensure descent in the objective for the initial iterations. Only in the later phase the acceleration takes place. This is in accordance with our local convergence analysis.

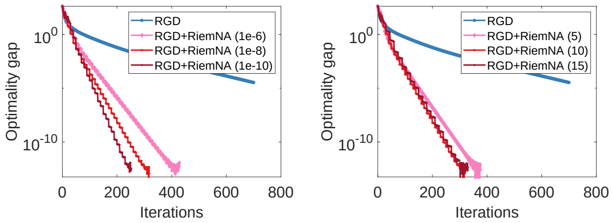


Figure 3: Parameter sensitivity on the leading eigenvector problem. Left: we vary λ by fixing $m = 10$. Right: we vary m by fixing $\lambda = 10^{-8}$. Our proposed RGD+RiemNA is robust to parameter changes.

Comparison with Additional Baselines We also compare with additional Riemannian acceleration methods in Figure 2, including an ODE-based acceleration method SIRNAG (Alimisis et al., 2020), an adaptive momentum-based acceleration method RAGDsDR (Alimisis et al., 2021), and an acceleration method for the Stiefel manifold StAGD (Siegel, 2019). We notice that the curvature parameter $\zeta \geq 1$ is required for both SIRNAG and RAGDsDR, which should be set as 1 if the manifold is positively curved and $\zeta > 1$ when the minimum curvature is negative. For the case of leading eigenvector problem, which is on sphere, manifold of positive curvature, we fix $\zeta = 1$. Otherwise, we first tune ζ for SIRNAG and RAGDsDR. Then the stepsize is tuned accordingly. For StAGD, only the stepsize is tuned. In Figure 2, we observe that RGD+RiemNA outperforms the above baselines as well. Even in the Stiefel case, our general RGD+RiemNA is faster than the specialized acceleration method StAGD.

Ablation Studies In Figure 3, we test the sensitivity of RGD+RiemNA to the choices of regularization parameter λ (on the left with $m = 10$ fixed) and memory depth m (on the right with $\lambda = 10^{-8}$ fixed). The results demonstrate robustness of RiemNA under various choices of regularization parameter λ and memory depth m. Additionally, in Appendix A.3, we also test on alternative averaging schemes where we show that RGD+RiemNA with (Avg.2) performs very similar to the strategy (Avg.1).

8 CONCLUSION

In this paper, we introduce a scheme for accelerating first-order Riemannian optimization algorithms, based on the idea of iterate extrapolation on the manifolds. The extrapolation step is performed via novel intrinsic weighted averaging schemes on manifolds. We show that Riemannian acceleration achieves convergence with asymptotically optimal rates irrespective of function classes. We also show our analysis holds with computationally cheap retraction and vector transport operations. Empirically, we see superior performance of the proposed algorithm RGD+RiemNA against many state-of-the-art Riemannian acceleration algorithms.

Even though the convergence analysis of our proposed acceleration scheme is asymptotic, we empirically observe its good performance against the baselines. It thus raises the question whether non-asymptotic convergence rates can be established. While we have focused on analyzing the RGD, it is also interesting to see whether such an acceleration scheme can be applied to other algorithm classes, such as momentum-based and stochastic algorithms.

References

Absil, P.-A., Baker, C. G., and Gallivan, K. A. (2007). Trust-region methods on Riemannian manifolds. *Foundations of Computational Mathematics*, 7(3):303–330.

Absil, P.-A., Mahony, R., and Sepulchre, R. (2009). Optimization algorithms on matrix manifolds. In *Optimization Algorithms on Matrix Manifolds*. Princeton University Press.

Agarwal, N., Boumal, N., Bullins, B., and Cartis, C. (2021). Adaptive regularization with cubics on manifolds. *Mathematical Programming*, 188(1):85–134.

Ahn, K. and Sra, S. (2020). From Nesterov's estimate sequence to Riemannian acceleration. In *Conference on Learning Theory*, pages 84–118. PMLR.

Aitken, A. C. (1927). On Bernoulli's numerical solution of algebraic equations. *Proceedings of the Royal Society of Edinburgh*, 46:289–305.

Alimisis, F., Orvieto, A., Bécigneul, G., and Lucchi, A. (2020). A continuous-time perspective for modeling acceleration in Riemannian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1297–1307. PMLR.

Alimisis, F., Orvieto, A., Becigneul, G., and Lucchi, A. (2021). Momentum improves optimization on Riemannian manifolds. In *International Conference on Artificial Intelligence and Statistics*, pages 1351–1359. PMLR.

Andrews, B. and Hopper, C. (2010). *The Ricci flow in Riemannian geometry: a complete proof of the differentiable* $1/4$*-pinching sphere theorem*. springer.

Becigneul, G. and Ganea, O.-E. (2018). Riemannian adaptive optimization methods. In *International Conference on Learning Representations*.

Bhatia, R. (2009). Positive definite matrices. In *Positive Definite Matrices*. Princeton university press.

Bollapragada, R., Scieur, D., and d'Aspremont, A. (2022). Nonlinear acceleration of momentum and primal-dual algorithms. *Mathematical Programming*, pages 1–38.

Bonnabel, S. (2013). Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229.

Boumal, N. (2020). An introduction to optimization on smooth manifolds. *Available online, May*, 3.

Boumal, N. and Absil, P.-A. (2015). Low-rank matrix completion via preconditioned optimization on the Grassmann manifold. *Linear Algebra and its Applications*, 475:200–239.

Boumal, N., Absil, P.-A., and Cartis, C. (2019). Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33.

Boumal, N., Mishra, B., Absil, P.-A., and Sepulchre, R. (2014). Manopt, a matlab toolbox for optimization on manifolds. *The Journal of Machine Learning Research*, 15(1):1455–1459.

Brezinski, C., Redivo-Zaglia, M., and Saad, Y. (2018). Shanks sequence transformations and Anderson acceleration. *SIAM Review*, 60(3):646–669.

Cherian, A. and Sra, S. (2016). Riemannian dictionary learning and sparse coding for positive definite matrices. *IEEE Transactions on Neural Networks and Learning Systems*, 28(12):2859–2871.

Criscitiello, C. and Boumal, N. (2022a). An accelerated first-order method for non-convex optimization on manifolds. *Foundations of Computational Mathematics*, pages 1–77.

Criscitiello, C. and Boumal, N. (2022b). Negative curvature obstructs acceleration for strongly geodesically convex optimization, even with exact first-order oracles. In *Conference on Learning Theory*, pages 496–542. PMLR.

Duruisseaux, V. and Leok, M. (2022). A variational formulation of accelerated optimization on Riemannian manifolds. *SIAM Journal on Mathematics of Data Science*, 4(2):649–674.

d'Aspremont, A., Scieur, D., Taylor, A., et al. (2021). Acceleration methods. *Foundations and Trends® in Optimization*, 5(1-2):1–245.

Edelman, A., Arias, T. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353.

Eldén, L. and Park, H. (1999). A procrustes problem on the stiefel manifold. *Numerische Mathematik*, 82(4):599–619.

Hamilton, L. and Moitra, A. (2021). No-go theorem for acceleration in the hyperbolic plane. *arXiv:2101.05657*.

Han, A. and Gao, J. (2021a). Improved variance reduction methods for Riemannian non-convex optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Han, A. and Gao, J. (2021b). Riemannian stochastic recursive momentum method for non-convex optimization. In *International Joint Conference on Artificial Intelligence*, pages 2505–2511.

Han, A., Mishra, B., Jawanpuria, P., and Gao, J. (2022). Riemannian block SPD coupling manifold and its application to optimal transport. *arXiv:2201.12933*.

Harandi, M., Sanderson, C., Shen, C., and Lovell, B. C. (2013). Dictionary learning and sparse coding on Grassmann manifolds: An extrinsic solution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3120–3127.

Huang, W. (2013). *Optimization algorithms on Riemannian manifolds with applications*. PhD thesis, The Florida State University.

Huang, W., Absil, P.-A., and Gallivan, K. A. (2015a). A riemannian symmetric rank-one trust-region method. *Mathematical Programming*, 150(2):179–216.

Huang, W., Absil, P.-A., and Gallivan, K. A. (2016). A riemannian BFGS method for nonconvex optimization problems. In *Numerical Mathematics and Advanced Applications ENUMATH 2015*, pages 627–634. Springer.

Huang, W., Gallivan, K. A., and Absil, P.-A. (2015b). A Broyden class of quasi-Newton methods for Riemannian optimization. *SIAM Journal on Optimization*, 25(3):1660–1685.

Jawanpuria, P., Meghwanshi, M., and Mishra, B. (2019). Low-rank approximations of hyperbolic embeddings. In *IEEE Conference on Decision and Control*.

Jawanpuria, P., Meghwanshi, M., and Mishra, B. (2020a). Geometry-aware domain adaptation for unsupervised alignment of word embeddings. In *Annual Meeting of the Association for Computational Linguistics*.

Jawanpuria, P., Meghwanshi, M., and Mishra, B. (2020b). A simple approach to learning unsupervised multilingual embeddings. In *Conference on Empirical Methods in Natural Language Processing*.

Jawanpuria, P. and Mishra, B. (2018). A unified framework for structured low-rank matrix learning. In *International Conference on Machine Learning*.

Jawanpuria, P., Satya Dev, N. T. V., and Mishra, B. (2021). Efficient robust optimal transport: formulations and algorithms. In *IEEE Conference on Decision and Control*.

Jin, J. and Sra, S. (2022). Understanding Riemannian acceleration via a proximal extragradient framework. In *Conference on Learning Theory*, pages 2924–2962. PMLR.

Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30(5):509–541.

Kasai, H., Jawanpuria, P., and Mishra, B. (2019). Riemannian adaptive stochastic gradient algorithms on matrix manifolds. In *International Conference on Machine Learning*, pages 3262–3271. PMLR.

Kasai, H., Sato, H., and Mishra, B. (2018). Riemannian stochastic recursive gradient algorithm. In *International Conference on Machine Learning*, pages 2516–2524. PMLR.

Keshavan, R. H. and Oh, S. (2009). A gradient descent algorithm on the Grassman manifold for matrix completion. *arXiv:0910.5260*.

Kim, J. and Yang, I. (2022). Accelerated gradient methods for geodesically convex optimization: Tractable algorithms and convergence analysis. In *International Conference on Machine Learning*, pages 11255–11282. PMLR.

Liu, Y., Shang, F., Cheng, J., Cheng, H., and Jiao, L. (2017). Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, volume 30.

Mangoubi, O. and Smith, A. (2018). Rapid mixing of geodesic walks on manifolds with positive curvature. *The Annals of Applied Probability*, 28(4):2501–2543.

Martínez-Rubio, D. (2022). Global Riemannian acceleration in hyperbolic and spherical spaces. In *International Conference on Algorithmic Learning Theory*, pages 768–826. PMLR.

Mishra, B., Satyadev, N., Kasai, H., and Jawanpuria, P. (2021). Manifold optimization for non-linear optimal transport problems. *arXiv:2103.00902*.

Nesterov, Y. (2003). *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.

Nesterov, Y. E. (1983). A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547.

Nickel, M. and Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. In *NeurIPS*.

Nimishakavi, M., Jawanpuria, P., and Mishra, B. (2018). A dual framework for low-rank tensor completion. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Qi, C., Gallivan, K. A., and Absil, P.-A. (2010). Riemannian BFGS algorithm with applications. In *Recent Advances in Optimization and its Applications in Engineering*, pages 183–192. Springer.

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

Sato, H., Kasai, H., and Mishra, B. (2019). Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport. *SIAM Journal on Optimization*, 29(2):1444–1472.

Scieur, D., d'Aspremont, A., and Bach, F. (2020). Regularized nonlinear acceleration. *Mathematical Programming*, 179(1):47–83.

Scieur, D., Oyallon, E., d'Aspremont, A., and Bach, F. (2018). Online regularized nonlinear acceleration. *arXiv:1805.09639*.

Shanks, D. (1955). Non-linear transformations of divergent and slowly convergent sequences. *Journal of Mathematics and Physics*, 34(1-4):1–42.

Shi, D., Gao, J., Hong, X., Boris Choy, S., and Wang, Z. (2021). Coupling matrix manifolds assisted optimization for optimal transport problems. *Machine Learning*, 110(3):533–558.

Sidi, A., Ford, W. F., and Smith, D. A. (1986). Acceleration of convergence of vector sequences. *SIAM Journal on Numerical Analysis*, 23(1):178–196.

Siegel, J. W. (2019). Accelerated optimization with orthogonality constraints. *arXiv:1903.05204*.

Su, W., Boyd, S., and Candes, E. (2014). A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. In *Advances in Neural Information Processing Systems*, volume 27.

Sun, Y., Flammarion, N., and Fazel, M. (2019). Escaping from saddle points on Riemannian manifolds. *Advances in Neural Information Processing Systems*, 32.

Tripuraneni, N., Flammarion, N., Bach, F., and Jordan, M. I. (2018). Averaging stochastic gradient descent on Riemannian manifolds. In *Conference On Learning Theory*, pages 650–687. PMLR.

Udriste, C. (2013). *Convex functions and optimization methods on Riemannian manifolds*, volume 297. Springer Science & Business Media.

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

Waldmann, S. (2012). Geometric wave equations. *arXiv:1208.4706*.

Walker, H. F. and Ni, P. (2011). Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 49(4):1715–1735.

Wibisono, A., Wilson, A. C., and Jordan, M. I. (2016). A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358.

Wynn, P. (1956). On a device for computing the e m (s n) transformation. *Mathematical Tables and Other Aids to Computation*, pages 91–96.

Zhang, H., J Reddi, S., and Sra, S. (2016). Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, volume 29.

Zhang, H. and Sra, S. (2016). First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pages 1617–1638. PMLR.

Zhang, H. and Sra, S. (2018). An estimate sequence for geodesically convex optimization. In *Conference On Learning Theory*, pages 1703–1723. PMLR.

Zhao, Z., Bai, Z.-J., and Jin, X.-Q. (2015). A riemannian newton algorithm for nonlinear eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 36(2):752–774.

A EXPERIMENT DETAILS AND ADDITIONAL EXPERIMENTS

A.1 Geometry of Specific Riemannian Manifolds

Sphere Manifold The sphere manifold $\mathcal{S}^{d-1}$ is an embedded submanifold of $\mathbb{R}^d$ with the tangent space identified as $T_x\mathcal{S}^{d-1} = \{u \in \mathbb{R}^d : x^\top u = 0\}$. The Riemannian metric is given by $\langle u, v\rangle = \langle u, v\rangle_2$ for $u, v \in T_x\mathcal{S}^{d-1}$. We use the exponential map derived as $\mathrm{Exp}_x(u) = \cos(\|u\|_2)x + \sin(\|u\|_2)\frac{u}{\|u\|_2}$ and the inverse exponential map as $\mathrm{Exp}_x^{-1}(y) = \arccos(x^\top y)\frac{\mathrm{Proj}_x(y-x)}{\|\mathrm{Proj}_x(y-x)\|_2}$ where $\mathrm{Proj}_x(v) = v - (x^\top v)x$ is the orthogonal projection of any $v \in \mathbb{R}^d$ to the tangent space $T_x\mathcal{S}^{d-1}$. The vector transport is given by the projection operation, i.e., $\mathcal{T}_x^y u = \mathrm{Proj}_y(u)$.

Symmetric Positive Definite (SPD) Manifold The SPD manifold of dimension d is denoted as $\mathbb{S}^d_{++} := \{X \in \mathbb{R}^{d\times d} : X^\top = X, X \succ 0\}$. The tangent space $T_X\mathcal{M}$ is the set of symmetric matrices. The affine-invariant Riemannian metric is given by $\langle U, V\rangle_X = \mathrm{tr}(X^{-1}UX^{-1}V)$ for any $U, V \in T_X\mathbb{S}^d_{++}$. We make use of the exponential map, which is $\mathrm{Exp}_X(U) = X\mathrm{expm}(X^{-1}U)$ where $\mathrm{expm}(\cdot)$ is the matrix exponential. The inverse exponential map is derived as $\mathrm{Exp}_X^{-1}(Y) = X\mathrm{logm}(X^{-1}Y)$ for any $X, Y \in \mathbb{S}^d_{++}$. We consider the parallel transport given by $\Gamma_X^Y U = EUE^\top$ with $E = (YX^{-1})^{1/2}$.

Stiefel Manifold The Stiefel manifold of dimension $p \times r$ is written as $\mathrm{St}(p, r) := \{X \in \mathbb{R}^{p\times r} : X^\top X = I\}$. The Riemannian metric is the Euclidean inner product defined as $\langle U, V\rangle_X = \langle U, V\rangle_2$. We consider the QR-based retraction $\mathrm{Retr}_X(U) = \mathrm{qf}(X + U)$ where $\mathrm{qf}(\cdot)$ returns the Q-factor from the QR decomposition. The inverse retraction is derived as for $X, Y \in \mathcal{O}(d)$ $\mathrm{Retr}_X^{-1}(Y) = YR - X$, where R is solved from the system $X^\top YR + R^\top Y^\top X = 2I$. The vector transport is given by the orthogonal projection, which is $\mathcal{T}_X^Y = U - Y\{Y^\top U\}_{\mathrm{S}}$ where $\{A\}_{\mathrm{S}} := (A + A^\top)/2$.

Grassmann Manifold The Grassmann manifold of dimension $p \times r$, denoted as $\mathrm{Gr}(p, r)$, is the set of all r dimensional subspaces in $\mathbb{R}^p$ $(p \geq r)$. Each point on the Grassmann manifold can be identified as a column orthonormal matrices $X \in \mathbb{R}^{p\times r}$, $X^\top X = I$ and two points $X, Y \in \mathrm{Gr}(p, r)$ are equivalent if $X = YO$ for some $O \in \mathcal{O}(r)$, the $r \times r$ orthogonal matrix. Hence Grassmann manifold is a quotient manifold of the Stiefel manifold. We consider the popular QR-based retraction, i.e. $R_X(U) = \mathrm{qf}(X + U)$ where for simplicity, we let X to represent the equivalence class and U represents the horizontal lift of the tangent vector. The inverse retraction is also based on QR factorization, i.e. $R_X^{-1}(Y) = Y(X^\top Y)^{-1} - X$. Vector transport is $\mathcal{T}_X^Y U = U - XX^\top U$.

A.2 Baseline Riemannian Acceleration Methods

Here, we include the implementation details of the Riemannian Nesterov accelerated gradient methods presented in (Zhang and Sra, 2018; Kim and Yang, 2022; Alimisis et al., 2020, 2021; Siegel, 2019). It is worth noting that those algorithms have been analyzed under the exponential map, inverse exponential map, and parallel transport. In contrast, the proposed RGD+RiemNA works with general retraction and vector transport.

We first present the (constant-stepsize) RAGD method in (Zhang and Sra, 2018, Algorithm 2), which is included in Algorithm 3. We see the algorithm requires three times evaluation of the exponential map and two times the inverse exponential map at every iteration.

Algorithm 3 RAGD (Zhang and Sra, 2018)

```
Input: Initialization $x_0$, parameter $\beta > 0$, stepsize $h \leq \frac{1}{L}$, strong convexity parameter $\mu > 0$.
Initialize $v_0 = x_0$.
Set $\alpha = \frac{\sqrt{\beta^2 + 4(1+\beta)\mu h} - \beta}{2}$, $\gamma = \frac{\sqrt{\beta^2 + 4(1+\beta)\mu h} - \beta}{\sqrt{\beta^2 + 4(1+\beta)\mu h} + \beta}\mu$, $\bar{\gamma} = (1+\beta)\gamma$.
for $k = 0, ..., K-1$ do
    Compute $\alpha_k \in (0, 1)$ from the equation $\alpha_k^2 = h_k((1-\alpha_k)\gamma_k + \alpha_k\mu)$.
    $y_k = \mathrm{Exp}_{x_k}\big(\frac{\alpha\gamma}{\gamma + \alpha\mu}\mathrm{Exp}_{x_k}^{-1}(v_k)\big)$
    $x_{k+1} = \mathrm{Exp}_{y_k}(-h\,\mathrm{grad}f(y_k))$
    $v_{k+1} = \mathrm{Exp}_{y_k}\big(\frac{(1-\alpha)\gamma}{\bar{\gamma}}\mathrm{Exp}_{y_k}^{-1}(v_k) - \frac{\alpha}{\bar{\gamma}}\mathrm{grad}f(y_k)\big)$
end for
Output: $x_K$
```

Below we present RNAG-C (Algorithm 4), which is designed for geodesic convex functions and RNAG-SC (Algorithm 5) which is for geodesic strongly convex functions in (Kim and Yang, 2022). We observe the algorithms require two times evaluation of the exponential map, inverse exponential map as well as parallel transport.

Algorithm 4 RNAG-C (Kim and Yang, 2022)

```
**Input:** Initialization $x_0$, parameters $\xi, T > 0$, stepsize $s \leq \frac{1}{L}$.
Initialize $\bar{v}_0 = 0 \in T_{x_0}\mathcal{M}$.
Set $\lambda_k = \frac{k+2\xi+T}{2}$.
**for** $k = 0, ..., K-1$ **do**
    $y_k = \mathrm{Exp}_{x_k}\big(\frac{\xi}{\lambda_k+\xi-1}\bar{v}_k\big)$
    $x_{k+1} = \mathrm{Exp}_{y_k}(-s\,\mathrm{grad}f(y_k))$
    $v_k = \Gamma_{x_k}^{y_k}\big(\bar{v}_k - \mathrm{Exp}_{x_k}^{-1}(y_k)\big)$
    $\bar{\bar{v}}_{k+1} = v_k - \frac{s\lambda_k}{\xi}\mathrm{grad}f(y_k)$
    $\bar{v}_{k+1} = \Gamma_{y_k}^{x_{k+1}}\big(\bar{\bar{v}}_{k+1} - \mathrm{Exp}_{y_k}^{-1}(x_{k+1})\big)$
**end for**
**Output:** $x_K$
```

Algorithm 5 RNAG-SC (Kim and Yang, 2022)

```
**Input:** Initialization $x_0$, parameter $\xi$, stepsize $s \leq \frac{1}{L}$, strong convexity parameter $\mu$.
Set $q = \mu s$.
**for** $k = 0, ..., K-1$ **do**
    $y_k = \mathrm{Exp}_{x_k}\big(\frac{\sqrt{\xi q}}{1+\sqrt{\xi q}}\bar{v}_k\big)$
    $x_{k+1} = \mathrm{Exp}_{y_k}\big(-s\,\mathrm{grad}f(y_k)\big)$
    $v_k = \Gamma_{x_k}^{y_k}\big(\bar{v}_k - \mathrm{Exp}_{x_k}^{-1}(y_k)\big)$
    $\bar{\bar{v}}_{k+1} = \big(1 - \sqrt{\frac{q}{\xi}}\big)v_k + \sqrt{\frac{q}{\xi}}\big(-\frac{1}{\mu}\mathrm{grad}f(y_k)\big)$
    $\bar{v}_{k+1} = \Gamma_{y_k}^{x_{k+1}}\big(\bar{\bar{v}}_{k+1} - \mathrm{Exp}_{y_k}^{-1}(x_{k+1})\big)$
**end for**
**Output:** $x_K$
```

We also include SIRNAG (Alimisis et al., 2020), RAGDsDR (Alimisis et al., 2021) and StAGD (Siegel, 2019). We have included the detailed steps in Algorithm 6 and 7 respectively. Specifically, SIRNAG is the discretization of an ODE on manifolds that achieves acceleration. For the purpose of experiments, we only consider the version for geodesic convex functions. This is because the version for geodesic strongly convex functions only differs in one parameter setting. SIRNAG involves two update options, SIRNAG (opt-1) and SIRNAG (opt-2), which correspond to two strategies of discretization.

Algorithm 6 SIRNAG

```
**Input:** Initialization $x_0$. Integration stepsize $h$. curvature parameter $\zeta$.
**for** $k = 0, ..., K-1$ **do**
    $\beta_k = \frac{k-1}{k+2\zeta}$.
    **Option I**: $a_k = \beta_k v_k - h\,\mathrm{grad}f(x_k)$.
    **Option II**: $a_k = \beta_k v_k - h\,\mathrm{grad}f\big(\mathrm{Exp}_{x_k}(h\beta_k v_k)\big)$.
    $x_{k+1} = \mathrm{Exp}_{x_k}(h\,a_k)$.
    $v_{k+1} = \Gamma_{x_k}^{x_{k+1}}a_k$.
**end for**
**Output:** $x_K$.
```

RAGDsDR, accelerates the convergence for both geodesic convex and weakly-quasi-convex functions by exploiting momentum. For experiments, we only consider the convex version. We follow the empirical choice of β_k suggested in the paper.

Finally, specifically for the orthogonal Procrustes problem, we include the acceleration method (Siegel, 2019) designed for the Stiefel manifold as another baseline, which we call StAGD. In particular, we implement the version with function restart

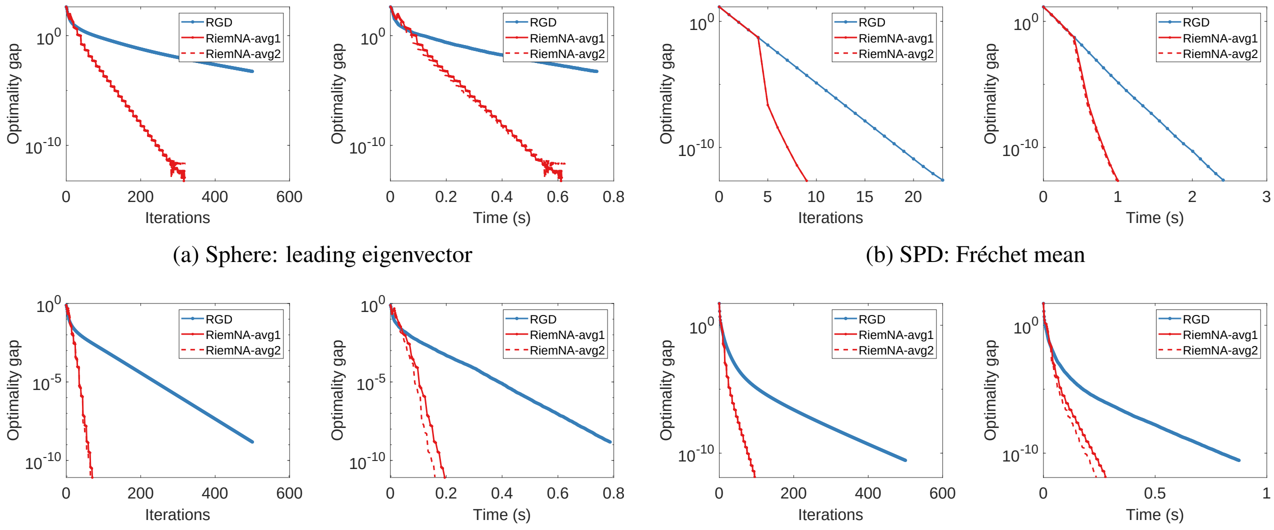
Algorithm 7 RAGDsDR

```
Input: Initialization $x_0$. Smoothness parameter $L$. curvature parameter $\zeta$.
$v_0 = x_0$, $A_0 = 0$.
for $k = 0, ..., K-1$ do
    $\beta_k = \frac{k}{k+2}$.
    $y_k = \mathrm{Exp}_{v_k}\big(\beta_k \mathrm{Exp}_{v_k}^{-1}(x_k)\big)$
    $x_{k+1} = \mathrm{Exp}_{y_k}(-\frac{1}{L}\mathrm{grad} f(y_k))$
    Solve $a_{k+1} > 0$ from the equation $\frac{\zeta a_{k+1}^2}{A_k + a_{k+1}} = \frac{1}{L}$.
    $A_{k+1} = A_k + a_{k+1}$.
    $v_{k+1} = \mathrm{Exp}_{v_k}(-a_{k+1}\Gamma_{y_k}^{v_k}\mathrm{grad} f(y_k))$.
end for
Output: $x_K$.
```

(Siegel, 2019, Algorithm 4.1) and without using linesearch for comparability. It is worth noticing that (Siegel, 2019) applies the Cayley-based retraction and canonical Riemannian metric (Edelman et al., 1998) for the implementation.

A.3 Ablation Study: Use of Alternative Averaging Schemes

We next evaluate the numerical performance of RiemNA when using alternative averaging scheme, i.e. (Avg.2). Specifically, the average is given by $\bar{x}_{c,x} = \mathrm{Retr}_{x_k}\big(-\sum_{i=0}^{k-1}\theta_i \Gamma_{x_i}^{x_k}\mathrm{Retr}_{x_i}^{-1}(x_{i+1})\big) = \mathrm{Retr}_{x_k}\big(-\sum_{i=0}^{k-1}\theta_i r_i\big)$ where we use the general retraction. It is worth mentioning that (Avg.2) is more efficient by avoiding k times evaluation of inverse retraction map. We compare the use of two averaging schemes in Figure 4 where we observe almost identical convergence behaviour when measured against the iteration. For runtime, (Avg.2) can further reduce computational cost compared to (Avg.1), especially for the Stiefel manifold and Grassmann manifold where the inverse retraction is expensive. Even though for SPD manifold, the inverse exponential map is expensive, because the number of iteration to convergence is small, we do not observe a significant reduction in runtime.



(a) Sphere: leading eigenvector

(b) SPD: Fréchet mean

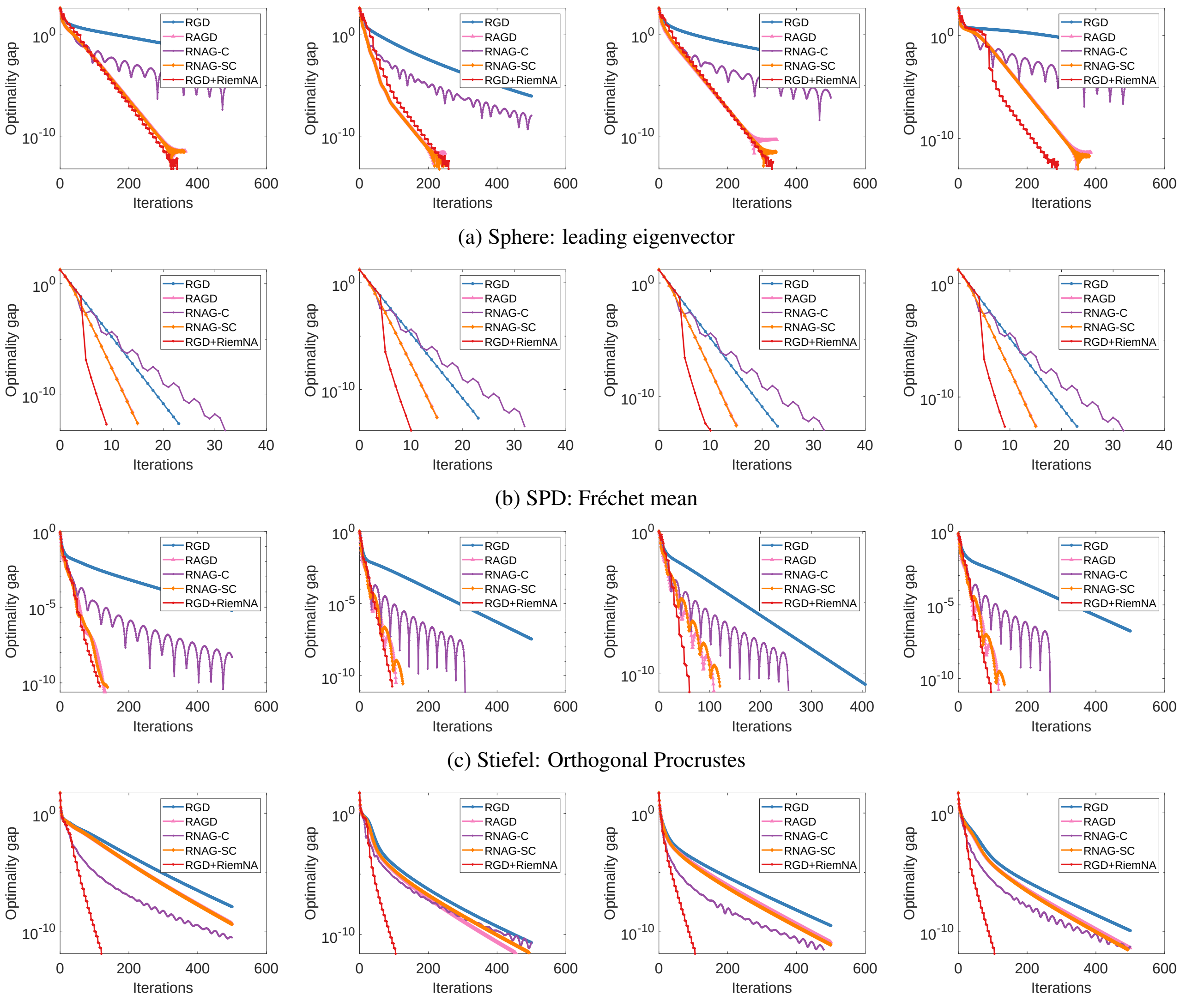
(c) Stiefel: Orthogonal Procrustes

(d) Grassmann: Nonlinear eigenspace

Figure 4: Comparison of different averaging schemes, i.e., (Avg.1) (used in the main text) and (Avg.2). We observe almost identical convergence in terms of iterations. (Avg.2) is more efficient, particularly for the case Stiefel and Grassmann manifold where the inverse retraction is costly.

A.4 Sensitivity to Data Generation and Initialization

Here, we provide additional independent experiment runs to test the model sensitivity to randomness in data generation and initialization. Each column in Figure 5 corresponds to a run with a fixed random seed. From Figure 5, we observe the proposed RGD+RiemNA maintains its outperformance against all baselines with good stability.



(a) Sphere: leading eigenvector

(b) SPD: Fréchet mean

(c) Stiefel: Orthogonal Procrustes

(d) Grassmann: Nonlinear eigenspace

Figure 5: Additional experiment runs with different data and initialization. Each column corresponds to an independent run. We observe the better performance of RGD+RiemNA in all the runs.

B ALTERNATIVE AVERAGING SCHEME VIA WEIGHTED FRÉCHET MEAN

We also consider the weighted Fréchet mean for computing the weighted average on manifolds, defined as

$$\bar{x}_{c,x} = \arg\min_{x \in \mathcal{X}} \sum_{i=0}^{k} c_i d^2(x, x_i). \tag{Avg.3}$$

Nevertheless, for general manifolds, it is not guaranteed the existence and uniqueness of the solution. In fact, one can ensure the uniqueness of the solution when the function $\frac{1}{2}d^2(x, x')$ is geodesic τ-strongly convex in x. From (Alimisis et al., 2020, Lemma 2), we see that the geodesic strong convexity of problem (Avg.3) holds for sufficiently small $\mathcal{X}$ on any manifold as well as for any non-positively curved manifold. Specifically, when $\mathcal{M}$ is non-positively curved, we have $\tau = 1$. While for other manifolds, let D be the diameter of $\mathcal{X}$ and $\kappa^+ > 0$ be the upper curvature bound. Then, geodesic strong convexity is satisfied with $\tau < 1$ when $D < \frac{\pi}{2\sqrt{\kappa^+}}$.

Lemma 10. *Under Assumption 1, suppose $x \mapsto \frac{1}{2}d^2(x, x')$ is geodesic τ-strongly convex in x for any $x' \in \mathcal{X}$. Consider $\bar{x}_{c,x} = \arg\min_{x \in \mathcal{X}} \sum_{i=0}^k c_i d^2(x, x_i)$. Then $d(\bar{x}_{c,x}, x^*) \leq \tau \| \sum_{i=0}^k c_i \Delta_{x_i} \|_{x^*}$ and $\| \Delta_{\bar{x}_{c,x}} - \sum_{i=0}^k c_i \Delta_{x_i} \|_{x^*} = O(d^3(x_0, x^*))$.*

Under the additional assumption of geodesic strong convexity, Lemma 10 shows an extra tighter bound on $d(\bar{x}_{c,x}, x^*)$, i.e., $d(\bar{x}_{c,x}, x^*) \leq \tau \| \sum_{i=0}^k c_i \Delta_{x_i} \|_{x^*}$. Thus, we see the error from the linear term does not suffer from metric distortion ($\epsilon_1 = 0$). The error bound from coefficient stability and nonlinearity terms however, still incur additional errors as the previous two averaging schemes. Lemma 10 allows convergence under the two averaging schemes to be established by exactly following the same steps as before. This is sufficient to show the same convergence bound holds (i.e., Theorem 1 and Proposition 1).

C FROM EUCLIDEAN AVERAGING TO RIEMANNIAN AVERAGING

To extend the idea of weighted average to manifolds, we first rewrite the weighted average on the Euclidean space as follows.

Lemma 11 (Weighted average recursion)**.** *Given a set of coefficients $\{c_i\}_{i=0}^k$ with $\sum_{i=0}^k c_i = 1$ and a set of iterates $\{x_i\}_{i=0}^k$. Let the streaming weighted average be defined as $\tilde{x}_i = \tilde{x}_{i-1} + \gamma_i(x_i - \tilde{x}_{i-1})$ where $\gamma_i = \frac{c_i}{\sum_{j=0}^i c_j}$ for $i = 0, ..., k$ and $\tilde{x}_{-1} = x_0$. Then $\tilde{x}_k = \sum_{i=0}^k c_i x_i$.*

Proof. For some $\gamma_1, ..., \gamma_k$, the streaming weighted average is defined as $\tilde{x}_i = \tilde{x}_{i-1} + \gamma_i(x_i - \tilde{x}_{i-1})$ for $i \in [k]$. We first show the streaming weighted average has the form

$$\tilde{x}_i = \prod_{j=1}^{i}(1 - \gamma_j)x_0 + \gamma_1 \prod_{j=2}^{i}(1 - \gamma_j)x_1 + \cdots + \gamma_i x_i, \quad \forall i \in [k].$$

We prove such argument by induction. For $i = 1$, it is clear that $\tilde{x}_1 = (1 - \gamma_1)x_0 + \gamma_1 x_1$ and satisfies the form. Suppose at $i = k'$, the equality is satisfied, then for $i = k' + 1$, we have

$$\tilde{x}_{k'+1} = \tilde{x}_{k'} + \gamma_{k'+1}(x_{k'+1} - \tilde{x}_{k'}) = (1 - \gamma_{k'+1})\tilde{x}_{k'} + \gamma_{k'+1}x_{k'+1}$$

which satisfies the equality. Hence this argument holds for all $i \in [k]$. Finally, at $i = k$, we see that the choice that $\gamma_i = \frac{c_i}{\sum_{j=0}^i c_i}$ leads to the matching coefficients. □

D FUNCTION CLASSES ON RIEMANNIAN MANIFOLDS

This section briefly reviews various functions classes on Riemannian manifolds.

D.1 Geodesic Gradient Lipschitzness and Hessian Lipschitzness

First, we provide several equivalent characterizations for the gradient and Hessian Lipschitzness. For proof and more detailed discussions, see (Boumal, 2020, Section 10.4).

Lemma 12 (Geodesic gradient Lipschitzness and function smoothness)**.** *A function $f : \mathcal{M} \to \mathbb{R}$ has geodesic L-Lipschitz gradient in $\mathcal{X} \subseteq$ if for all $x, y = \mathrm{Exp}_x(u) \in \mathcal{X}$ in the domain of the exponential map, we have*

$$\|\Gamma_{\gamma(t)}^x \mathrm{grad} f(\gamma(t)) - \mathrm{grad} f(x)\|_x \leq L\|tu\|_x,$$

for all $t \in [0, 1]$ and $\gamma(t) \coloneqq \mathrm{Exp}_x(tu)$. This is equivalent to function having bounded Hessian as $\|\mathrm{Hess} f(x)\|_x \coloneqq \max_{u \in T_x\mathcal{M}: \|u\|_x = 1} \|\mathrm{Hess} f(x)[u]\|_x \leq L$, where $\|\mathrm{Hess} f(x)\|_x$ denotes the operator norm of Riemannian Hessian. If function f has geodesic L-Lipschitz gradient, then function f is geodesic L-smooth, which satisfies

$$|f(y) - f(x) - \langle \mathrm{grad} f(x), u \rangle_x| \leq \frac{L}{2}\|u\|_x^2.$$

Lemma 13 (Geodesic Hessian Lipschitzness). *A function f has geodesic ρ-Lipschitz Hessian in $\mathcal{X} \subseteq \mathcal{M}$ if for all $x, y = \mathrm{Exp}_x(u) \in \mathcal{X}$ in the domain of the exponential map, we have*

$$\|\Gamma_{\gamma(t)}^x \circ \mathrm{Hess}f(\gamma(t)) \circ \Gamma_x^{\gamma(t)} - \mathrm{Hess}f(x)\|_x \leq \rho\|tu\|_x^3,$$

for all $t \in [0,1]$ and $\gamma(t) := \mathrm{Exp}_x(tu)$. If function f has geodesic ρ-Lipschitz Hessian, then function f satisfies

$$|f(y) - f(x) - \langle \mathrm{grad}f(x), u\rangle_x - \frac{1}{2}\langle u, \mathrm{Hess}f(x)[u]\rangle_x| \leq \frac{\rho}{6}\|u\|_x^3$$
$$\|\Gamma_y^x \mathrm{grad}f(y) - \mathrm{grad}f(x) - \mathrm{Hess}f(x)[u]\|_x \leq \frac{\rho}{2}\|u\|^2.$$

D.2 Retraction Gradient Lipschitzness and Hessian Lipschitzness

In this section, we define the gradient and Hessian Lipschitzness with respect to a retraction, which generalizes the definitions in Section D.1.

Definition 2 (Retraction gradient Lipschitzness). A function $f : \mathcal{M} \to \mathbb{R}$ has retraction L-Lipschitz gradient in $\mathcal{X} \subseteq \mathcal{M}$ if for all $x, y = R_x(u) \in \mathcal{X}$, we have

$$\|\Gamma_{c(t)}^x \mathrm{grad}f(c(t)) - \mathrm{grad}f(x)\|_x \leq L\|tu\|_x$$

where we denote $c(t) = R_x(tu)$.

Definition 3 (Retraction Hessian Lipschitzness). A function $f : \mathcal{M} \to \mathbb{R}$ has retraction ρ-Lipschitz Hessian in $\mathcal{X} \subseteq \mathcal{M}$ if for all $x, y \in R_x(u) \in \mathcal{X}$ in the domain of the retraction, we have

$$\|\Gamma_{c(t)}^x \circ \mathrm{Hess}f(c(t)) \circ \Gamma_x^{c(t)} - \mathrm{Hess}f(x)\|_x \leq \rho\|tu\|_x^3,$$

where we denote $c(t) = R_x(tu)$.

D.3 Geodesic Convexity and Strong Convexity

We start with the notion of *geodesic convex set*. A subset $\mathcal{X} \subseteq \mathcal{M}$ is called geodesic convex if for any two points in the set, there exists a geodesic joining them that lies entirely within the set.

Definition 4 (Geodesic (strong) convexity). A function $f : \mathcal{X} \to \mathbb{R}$ is geodesic convex in a geodesic convex set $\mathcal{X}$ if for any geodesic $\gamma : [0,1] \to \mathcal{X}$, we have $f(\gamma(t)) \leq (1-t)f(x) + tf(y)$ where we let $x = \gamma(0), y = \gamma(1)$. The function is geodesic μ-strongly convex if $(f \circ \gamma)''(t) \geq \mu d^2(x,y)$ for all $t \in [0,1]$. This is equivalent to $\mathrm{Hess}f(x) \succeq \mu\,\mathrm{id}$ for all $x \in \mathcal{X}$.

A similar notion of convexity with respect to retraction also exists by replacing the geodesic curve $\gamma(t)$ with retraction curve $c(t) = \mathrm{Retr}_x(tu)$. See Huang et al. (2015b) for more details.

E MAIN PROOFS

Before we proceed with the proofs of the results in the paper, we introduce a lemma that is used often in the course of the proof.

Lemma 14. *Under Assumption 1, for any $w, x, y, z \in \mathcal{X}$, we have $\|\Gamma_w^x \Gamma_y^w \mathrm{Exp}_y^{-1}(z) - \big(\mathrm{Exp}_x^{-1}(z) - \mathrm{Exp}_x^{-1}(y)\big)\|_x \leq C_0 d(y,w)d(w,x)d(y,z) + C_2 \min\{d(y,z), d(x,y)\}C_\kappa\big(d(y,z) + d(x,y)\big)$.*

Proof of Lemma 14.

$$\begin{aligned}
&\|\Gamma_w^x \Gamma_y^w \mathrm{Exp}_y^{-1}(z) - \big(\mathrm{Exp}_x^{-1}(z) - \mathrm{Exp}_x^{-1}(y)\big)\|_x \\
&\leq \|\Gamma_w^x \Gamma_y^w \mathrm{Exp}_y^{-1}(z) - \Gamma_y^x \mathrm{Exp}_y^{-1}(z)\|_x + \|\Gamma_y^x \mathrm{Exp}_y^{-1}(z) - \big(\mathrm{Exp}_x^{-1}(z) - \mathrm{Exp}_x^{-1}(y)\big)\|_x \\
&\leq C_0 d(y,w)d(w,x)d(y,z) + C_2 d\Big(\mathrm{Exp}_x\big(\Gamma_y^x \mathrm{Exp}_y^{-1}(z) + \mathrm{Exp}_x^{-1}(y)\big), z\Big) \\
&\leq C_0 d(y,w)d(w,x)d(y,z) + C_2 d\Big(\mathrm{Exp}_x\big(\Gamma_y^x \mathrm{Exp}_y^{-1}(z) + \mathrm{Exp}_x^{-1}(y)\big), \mathrm{Exp}_y(\mathrm{Exp}_y^{-1}(z))\Big) \\
&\leq C_0 d(y,w)d(w,x)d(y,z) + C_2 \min\{d(y,z), d(x,y)\}C_\kappa\big(d(y,z) + d(x,y)\big).
\end{aligned}$$

where we apply Lemma 1 and 2. □

E.1 Proof of Proposition 2

We show in Proposition 2 that the optimal coefficients c^* has a closed-form solution.

Proposition 2. *Let $R = [\langle r_i, r_j \rangle_{x_k}]_{i,j} \in \mathbb{R}^{(k+1)\times(k+1)}$ collects all pairwise inner products. Then the solution $c^* = \arg\min_{c \in \mathbb{R}^{k+1}: c^\top 1 = 1} \| \sum_{i=0}^k c_i r_i \|^2_{x_k} + \lambda \|c\|_2^2$ is explicitly derived as $c^* = \frac{(R+\lambda I)^{-1} 1}{1^\top (R+\lambda I)^{-1} 1}$.*

Proof of Proposition 2. Let $\mu \in \mathbb{R}$ be the dual variable. Then we have c^*, μ^* satisfy the KKT system:

$$\begin{bmatrix} 2(R + \lambda I) & 1 \\ 1^\top & 0 \end{bmatrix} \begin{bmatrix} c^* \\ \mu^* \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Solving the system yields the desired result. □

E.2 Proof of Lemma 3

Proof of Lemma 3. First, we consider the pushforward operator $\mathrm{Exp}_x^y : T_x\mathcal{M} \to T_y\mathcal{M}$ for any $x, y \in \mathcal{M}$, defined as $\mathrm{Exp}_x^y(u) := \mathrm{Exp}_y^{-1}(\mathrm{Exp}_x(u))$ for any $u \in T_x\mathcal{M}$. The differential of Exp_x^y at 0 along $u \in T_x\mathcal{M}$ is derived as

$$\begin{aligned} \mathrm{DExp}_x^y(0)[u] = \mathrm{DExp}_y^{-1}(\mathrm{Exp}_x(0))[\mathrm{DExp}_x(0)[u]] = \mathrm{DExp}_y^{-1}(x)[u] &= [\mathrm{DExp}_y(\mathrm{Exp}_y^{-1}(x))]^{-1}[u] \\ &= (T_y^x)^{-1}[u] \end{aligned}$$

where we denote $T_x^y(v) = \mathrm{DExp}_x(\mathrm{Exp}_x^{-1}(y))[v] \in T_y\mathcal{M}$ for $v \in T_x\mathcal{M}$. The second equality is due to $\mathrm{Exp}_x(0) = 0, \mathrm{DExp}_x(0) = \mathrm{id}$ and the third equality follows from the inverse function theorem. Then by Taylor's theorem for $\mathrm{Exp}_{x_i}^{x^*}$ around 0, we have

$$\begin{aligned} \mathrm{Exp}_{x_*}^{-1}(x_{i+1}) &= \mathrm{Exp}_{x_i}^{x^*}(\mathrm{Exp}_{x_i}^{-1}(x_{i+1})) \\ &= \mathrm{Exp}_{x_i}^{x^*}(0) + \mathrm{DExp}_{x_i}^{x^*}(0)[\mathrm{Exp}_{x_i}^{-1}(x_{i+1})] + \frac{1}{2}\mathrm{D}^2\mathrm{Exp}_{x_i}^{x^*}(\zeta_i)[\mathrm{Exp}_{x_i}^{-1}(x_{i+1}), \mathrm{Exp}_{x_i}^{-1}(x_{i+1})] \\ &= \mathrm{Exp}_{x^*}^{-1}(x_i) - \eta (T_{x^*}^{x_i})^{-1}[\mathrm{grad} f(x_i)] + \frac{\eta^2}{2}\mathrm{D}^2\mathrm{Exp}_{x_i}^{x^*}(\zeta_i)[\mathrm{grad} f(x_i), \mathrm{grad} f(x_i)] \\ &= \mathrm{Exp}_{x^*}^{-1}(x_i) - \eta (T_{x^*}^{x_i})^{-1}[\mathrm{grad} f(x_i)] + \frac{\eta^2}{2}\epsilon_i \end{aligned} \tag{6}$$

for some $\zeta_i = s\mathrm{Exp}_{x_i}^{-1}(x_{i+1}),\ s \in (0,1)$. We let $\epsilon_i := \mathrm{D}^2\mathrm{Exp}_{x_i}^{x^*}(\zeta_i)[\mathrm{grad} f(x_i), \mathrm{grad} f(x_i)]$ with $\|\epsilon_i\|_{x^*} = O\big(\|\mathrm{grad} f(x_i)\|^2_{x_i}\big)$. Then by Hessian Lipschitzness (Lemma 13), we have around x^*

$$e_i := \Gamma_{x_i}^{x^*}\mathrm{grad} f(x_i) - \mathrm{Hess} f(x^*)[\mathrm{Exp}_{x^*}^{-1}(x_i)] \leq \frac{\rho}{2}\|\mathrm{Exp}_{x^*}^{-1}(x_i)\|^2_{x^*}. \tag{7}$$

Combining (6) with (7) yields

$$\begin{aligned} \mathrm{Exp}_{x_*}^{-1}(x_{i+1}) - \mathrm{Exp}_{x^*}^{-1}(x_i) &= -\eta(\Gamma_{x_i}^{x^*} T_{x^*}^{x_i})^{-1}[\Gamma_{x_i}^{x^*}\mathrm{grad} f(x_i)] + \frac{\eta^2}{2}\epsilon_i \\ &= -\eta(\Gamma_{x_i}^{x^*} T_{x^*}^{x_i})^{-1}[\mathrm{Hess} f(x^*)[\mathrm{Exp}_{x^*}^{-1}(x_i)] + e_i] + \frac{\eta^2}{2}\epsilon_i. \end{aligned} \tag{8}$$

To show the desired result, it remains to show the operator $(\Gamma_{x_i}^{x^*} T_{x^*}^{x_i})^{-1}$ is locally identity. This is verified in (Tripuraneni et al., 2018, Lemma 6) for general retraction. We restate here and adapt to the case of exponential map.

Consider the function $H(u) := (\Gamma_x^{\mathrm{Exp}_x(u)})^{-1} T_x^{\mathrm{Exp}_x(u)} : T_x\mathcal{M} \to L(T_x\mathcal{M})$, where $L(T_x\mathcal{M})$ denotes the set of linear maps on $T_x\mathcal{M}$. Let $\gamma(t) = \mathrm{Exp}_x(tu)$. Then we have

$$\begin{aligned} \frac{d}{dt}H(tu)|_{t=0} = \frac{d}{dt}(\Gamma_x^{\gamma(t)})^{-1} T_x^{\gamma(t)}|_{t=0} = \Big((\Gamma_x^{\gamma(t)})^{-1}\frac{D}{dt}T_x^{\gamma(t)}\Big)|_{t=0} &= \big(\frac{D}{dt}\mathrm{DExp}_x(tu)\big)|_{t=0} \\ &= \frac{D^2}{dt^2}\mathrm{Exp}_x(tu)|_{t=0} = 0. \end{aligned}$$

where the second equality is due to the property of parallel transport (see for example (Boumal, 2020, Proposition 10.37)). In addition, from (Waldmann, 2012, Theorem A.2.9), we see the second order derivative of H is given by $\frac{d^2}{dt^2}H(tu)|_{t=0} = \frac{1}{6}\mathrm{Riem}_x(u,\cdot)u$ where we denote Riem_x as the Riemann curvature tensor evaluated at x. We notice that $\mathrm{Riem}_x(u,\cdot)u : T_x\mathcal{M} \to T_x\mathcal{M}$ is symmetric with respect to the Riemannian metric (see for example (Andrews and Hopper, 2010)).

For any $v \in T_x\mathcal{M}$, $H(u)[v] \in T_x\mathcal{M}$, we apply the Taylor's theorem for H up to second order, which yields

$$H(u)[v] = v + \frac{1}{6}\mathrm{Riem}_x(u,v)u + O(\|u\|^3),$$

Let $x = x^*$ and $u = \mathrm{Exp}_{x^*}^{-1}(x_i) = \Delta_{x_i}$. Then we obtain for any $v \in T_{x^*}\mathcal{M}$, $H(u)[v] \in T_{x^*}\mathcal{M}$

$$\Gamma_{x_i}^{x^*} T_{x^*}^{x_i}[v] = v + \frac{1}{6}\mathrm{Riem}_{x^*}\big(\Delta_{x_i}, v\big)\Delta_{x_i} + O(\|\Delta_{x_i}\|^3). \tag{9}$$

It satisfies that $(\Gamma_{x_i}^{x^*} T_{x^*}^{x_i})^{-1} = \mathrm{id} - \frac{1}{6}\mathrm{Riem}_{x^*}\big(\Delta_{x_i}, \cdot\big)\Delta_{x_i} + O(\|\Delta_{x_i}\|^3)$. Substituting this result into (8), we obtain

$$\begin{aligned}
&\Delta_{x_{i+1}} - \Delta_{x_i} \\
&= -\eta\Big(\mathrm{id} - \frac{1}{6}\mathrm{Riem}_{x^*}\big(\Delta_{x_i}, \cdot\big)\Delta_{x_i} + O(\|\Delta_{x_i}\|^3)\Big)\big[\mathrm{Hess}f(x^*)[\Delta_{x_i}] + e_i\big] + \frac{\eta^2}{2}\epsilon_i \\
&= -\eta\,\mathrm{Hess}f(x^*)[\Delta_{x_i}] - \eta e_i + \frac{\eta}{6}\mathrm{Riem}_{x^*}(\Delta_{x_i}, \mathrm{Hess}f(x^*)[\Delta_{x_i}] + e_i)\Delta_{x_i} + \frac{\eta^2}{2}\epsilon_i + O(\|\Delta_{x_i}\|^3).
\end{aligned}$$

Let $\varepsilon_i = -\eta e_i + \frac{\eta}{6}\mathrm{Riem}_{x^*}(\Delta_{x_i}, \mathrm{Hess}f(x^*)[\Delta_{x_i}] + e_i)\Delta_{x_i} + \frac{\eta^2}{2}\epsilon_i + O(\|\Delta_{x_i}\|^3)$. We can bound the error term as follows.

$$\|\varepsilon_i\|_{x^*}^2 = O(\|e_i\|_{x^*}^2 + \|\Delta_{x_i}\|_{x^*}^4\|\mathrm{grad}f(x_i)\|_{x_i}^2 + \|\epsilon_i\|_{x^*}^2 + \|\Delta_{x_i}\|_{x^*}^6) = O(\|\Delta_{x_i}\|^4),$$

where we use the bounds on $\|e_i\|_{x^*}, \|\epsilon_i\|_{x^*}$ as well as $\mathrm{Hess}f(x^*)[\Delta_{x_i}] + e_i = \Gamma_{x_i}^{x^*}\mathrm{grad}f(x_i)$ and geodesic gradient Lipschitzness (Lemma 12) such that $\|\mathrm{grad}f(x_i)\|^2 \leq L\|\Delta_i\|_{x^*}^2$. □

E.3 Proof of Lemma 4

Proof of Lemma 4. The proof is by induction. Let $\gamma_i = \frac{c_i}{\sum_{j=0}^i c_j}$ and first we rewrite the averaging on tangent space as following the recursion defined as $\widetilde{\Delta}_{x_i} = \widetilde{\Delta}_{x_{i-1}} + \gamma_i(\Delta_{x_i} - \widetilde{\Delta}_{x_{i-1}})$. As we have shown in Lemma 11, $\sum_{i=0}^k c_i\Delta_{x_i} = \widetilde{\Delta}_{x_k}$. To show the difference between $\Delta_{\bar{x}_{c,x}}$, based on Lemma 2, it suffices to show the distance between $\tilde{x}_k = \bar{x}_{c,x}$ and $\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_k})$ is bounded.

To this end, we first notice that $\tilde{x}_0 = x_0 = \mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_0})$ and we consider bounding the difference between $\tilde{x}_1$ and $\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_1})$. To derive the bound, we first observe that by Lemma 1,

$$\begin{aligned}
&d\Big(\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_1}), \mathrm{Exp}_{x_0}\big(\Gamma_{x^*}^{x_0}\gamma_1(\Delta_{x_1} - \widetilde{\Delta}_{x_0})\big)\Big) \\
&= d\Big(\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_0} + \gamma_1(\Delta_{x_1} - \widetilde{\Delta}_{x_0})), \mathrm{Exp}_{x_0}\big(\Gamma_{x^*}^{x_0}\gamma_1(\Delta_{x_1} - \widetilde{\Delta}_{x_0})\big)\Big) \\
&\leq d(x_0, x^*)C_\kappa\Big(\|\widetilde{\Delta}_{x_0}\|_{x^*} + \gamma_1\|\Delta_{x_1} - \widetilde{\Delta}_{x_0}\|_{x^*}\Big),
\end{aligned} \tag{10}$$

where we see $x_0 = \mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_0})$ with $\widetilde{\Delta}_{x_0} = \Delta_{x_0}$. In addition,

$$\begin{aligned}
d\Big(\tilde{x}_1, \mathrm{Exp}_{x_0}\big(\Gamma_{x^*}^{x_0}\gamma_1(\Delta_{x_1} - \widetilde{\Delta}_{x_0})\big)\Big) &= d\Big(\mathrm{Exp}_{x_0}\big(\gamma_1\mathrm{Exp}_{x_0}^{-1}(x_1)\big), \mathrm{Exp}_{x_0}\big(\Gamma_{x^*}^{x_0}\gamma_1(\Delta_{x_1} - \widetilde{\Delta}_{x_0})\big)\Big) \\
&\leq \gamma_1 C_1\|\mathrm{Exp}_{x_0}^{-1}(x_1) - \Gamma_{x^*}^{x_0}(\Delta_{x_1} - \Delta_{x_0})\|_{x_0} \\
&\leq \gamma_1 C_1 C_2 d(x_0, x^*)C_\kappa\big(d(x_0, x_1) + d(x_0, x^*)\big).
\end{aligned} \tag{11}$$

where the last inequality is from the proof of Lemma 14. Thus combining (10), (11) leads to

$$\begin{aligned}
&d\big(\tilde{x}_1, \mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_1})\big) \\
&\leq d(x_0, x^*)C_\kappa\Big(\|\widetilde{\Delta}_{x_0}\|_{x^*} + \gamma_1\|\Delta_{x_1} - \widetilde{\Delta}_{x_0}\|_{x^*}\Big) + \gamma_1 C_1 C_2 d(x_0, x^*)C_\kappa\big(d(x_0, x_1) + d(x_0, x^*)\big).
\end{aligned}$$

By noticing $C_\kappa(x) = O(x^2)$, we see $d(\tilde{x}_1, \mathrm{Exp}_{x^*}(\tilde{\Delta}_{x_1})) = O(d^3(x_0, x^*))$.

Now suppose at $i \leq k-1$, we have $d(\tilde{x}_i, \mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})) = O(d^3(x_0, x^*))$ and we wish to show $d(\tilde{x}_{i+1}, \mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_{i+1}})) = O(d^3(x_0, x^*))$. To this end, we first see $\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_{i+1}}) = \mathrm{Exp}_{x^*}\big(\widetilde{\Delta}_{x_i} + \gamma_{i+1}(\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i})\big)$ and by Lemma 1

$$\begin{aligned}
& d\Big(\mathrm{Exp}_{x^*}\big(\widetilde{\Delta}_{x_i} + \gamma_{i+1}(\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i})\big), \mathrm{Exp}_{\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})}\big(\Gamma_{x^*}^{\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})}\gamma_{i+1}(\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i})\big)\Big) \\
& \leq \|\widetilde{\Delta}_{x_i}\|_{x^*} C_\kappa\big(\|\widetilde{\Delta}_{x_i}\|_{x^*} + \gamma_{i+1}\|\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i}\|_{x^*}\big) \\
& = O(d^3(x_0, x^*)),
\end{aligned}$$

where the order of $O(d^3(x_0, x^*))$ is due to $C_\kappa(x) = O(x^2)$ and $\|\widetilde{\Delta}_{x_i}\|_{x^*} = O(d(x_0, x^*))$, which can be shown by induction.

Further, noticing $\tilde{x}_{i+1} = \mathrm{Exp}_{\tilde{x}_i}\big(\gamma_{i+1}\mathrm{Exp}_{\tilde{x}_i}^{-1}(x_{i+1})\big)$, we can show

$$\begin{aligned}
& d\Big(\tilde{x}_{i+1}, \mathrm{Exp}_{\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})}\big(\Gamma_{x^*}^{\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})}\gamma_{i+1}(\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i})\big)\Big) \\
& \leq d\Big(\mathrm{Exp}_{\tilde{x}_i}\big(\gamma_{i+1}\mathrm{Exp}_{\tilde{x}_i}^{-1}(x_{i+1})\big), \mathrm{Exp}_{\tilde{x}_i}\big(\Gamma_{x^*}^{\tilde{x}_i}\gamma_{i+1}(\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i})\big)\Big) \\
& \quad + d\Big(\mathrm{Exp}_{\tilde{x}_i}\big(\Gamma_{x^*}^{\tilde{x}_i}\gamma_{i+1}(\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i})\big), \mathrm{Exp}_{\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})}\big(\Gamma_{x^*}^{\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})}\gamma_{i+1}(\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i})\big)\Big).
\end{aligned} \tag{12}$$

The first term on the right of (12) can be bounded as

$$\begin{aligned}
& d\Big(\mathrm{Exp}_{\tilde{x}_i}\big(\gamma_{i+1}\mathrm{Exp}_{\tilde{x}_i}^{-1}(x_{i+1})\big), \mathrm{Exp}_{\tilde{x}_i}\big(\Gamma_{x^*}^{\tilde{x}_i}\gamma_{i+1}(\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i})\big)\Big) \\
& \leq \gamma_{i+1}\|\mathrm{Exp}_{\tilde{x}_i}^{-1}(x_{i+1}) - \Gamma_{x^*}^{\tilde{x}_i}(\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i})\|_{\tilde{x}_i} \\
& = \gamma_{i+1}\|\Gamma_{\tilde{x}_i}^{x^*}\mathrm{Exp}_{\tilde{x}_i}^{-1}(x_{i+1}) - (\Delta_{x_{i+1}} - \Delta_{\tilde{x}_i}) + (\widetilde{\Delta}_{x_i} - \Delta_{\tilde{x}_i})\|_{x^*} \\
& \leq \gamma_{i+1}\|\Gamma_{\tilde{x}_i}^{x^*}\mathrm{Exp}_{\tilde{x}_i}^{-1}(x_{i+1}) - (\Delta_{x_{i+1}} - \Delta_{\tilde{x}_i})\|_{x^*} + \gamma_{i+1}\|\widetilde{\Delta}_{x_i} - \Delta_{\tilde{x}_i}\|_{x^*} \\
& \leq \gamma_{i+1}C_2 d(\tilde{x}_i, x^*)C_\kappa\big(d(x_{i+1}, \tilde{x}_i) + d(\tilde{x}_i, x^*)\big) + \gamma_{i+1}C_2 d(\tilde{x}_i, \mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})),
\end{aligned} \tag{13}$$

where we again use the result from the proof of Lemma 14. To see (13) is on the order of $O(d^3(x_0, x^*))$, we only need to show $\|\Delta_{\tilde{x}_i}\|_{x^*}^2 = d^2(\tilde{x}_i, x^*) = O(d^2(x_0, x^*))$, which can be seen by a simple induction argument. First, it is clear that $\|\Delta_{\tilde{x}_0}\|_{x^*}^2 = d^2(x_0, x^*)$. Then suppose for any $i < k$, we have $d(\tilde{x}_i, x^*) = O(d(x_0, x^*))$. Then from Lemma 2, we have

$$\begin{aligned}
d(\tilde{x}_{i+1}, x^*) \leq C_1\|\mathrm{Exp}_{\tilde{x}_i}^{-1}(\tilde{x}_{i+1}) - \mathrm{Exp}_{\tilde{x}_i}^{-1}(x^*)\|_{\tilde{x}_i} & \leq \frac{C_1 c_{i+1}}{\sum_{j=0}^{i+1} c_j} d(\tilde{x}_i, x_{i+1}) + d(\tilde{x}_i, x^*) \\
& \leq \big(\frac{C_1 c_{i+1}}{\sum_{j=0}^{i+1} c_j} + 1\big) d(\tilde{x}_i, x^*) + d(x_{i+1}, x^*) = O(d(x_0, x^*)).
\end{aligned}$$

Thus, using $d(\tilde{x}_i, \mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})) = O(d^3(x_0, x^*))$, we see (13) is on the order of $O(d^3(x_0, x^*))$.

Now we bound the second term on the right of (12). Particularly,

$$\begin{aligned}
& d\Big(\mathrm{Exp}_{\tilde{x}_i}\big(\Gamma_{x^*}^{\tilde{x}_i}\gamma_{i+1}(\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i})\big), \mathrm{Exp}_{\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})}\big(\Gamma_{x^*}^{\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})}\gamma_{i+1}(\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i})\big)\Big) \\
& \leq d\Big(\mathrm{Exp}_{\tilde{x}_i}\big(\Gamma_{x^*}^{\tilde{x}_i}\gamma_{i+1}(\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i})\big), \mathrm{Exp}_{\tilde{x}_i}\big(\Gamma_{\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})}^{\tilde{x}_i}\Gamma_{x^*}^{\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})}\gamma_{i+1}(\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i})\big)\Big) \\
& \quad + d\Big(\mathrm{Exp}_{\tilde{x}_i}\big(\Gamma_{\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})}^{\tilde{x}_i}\Gamma_{x^*}^{\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})}\gamma_{i+1}(\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i}), \mathrm{Exp}_{\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})}\big(\Gamma_{x^*}^{\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})}\gamma_{i+1}(\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i})\big)\Big) \\
& \leq \gamma_{i+1}C_1C_0\|\widetilde{\Delta}_{x_i}\|_{x^*} d\big(\tilde{x}_i, \mathrm{Exp}_{x^*}\big(\widetilde{\Delta}_{x_i}\big)\big)\|\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i}\|_{x^*} + C_3 d(\tilde{x}_i, \mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})) \\
& = O(d^3(x_0, x^*)),
\end{aligned}$$

where we apply Lemma 2 multiple times. Combining the previous results, we see

$$\begin{aligned}
&d(\tilde{x}_{i+1}, \mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_{i+1}})) \\
&\le d\Big(\tilde{x}_{i+1}, \mathrm{Exp}_{\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})}\big(\Gamma_{x^*}^{\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})}\gamma_{i+1}(\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i})\big)\Big) \\
&\qquad + d\Big(\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i} + \gamma_{i+1}(\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i})), \mathrm{Exp}_{\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})}\big(\Gamma_{x^*}^{\mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_i})}\gamma_{i+1}(\Delta_{x_{i+1}} - \widetilde{\Delta}_{x_i})\big)\Big) \\
&= O(d^3(x_0, x^*))
\end{aligned}$$

Now applying Lemma 2, we obtain

$$\|\Delta_{\tilde{x}_{i+1}} - \widetilde{\Delta}_{x_{i+1}}\|_{x^*} \le C_2 d(\tilde{x}_{i+1}, \mathrm{Exp}_{x^*}(\widetilde{\Delta}_{x_{i+1}})) = O(d^3(x_0, x^*))$$

for all $i \le k-1$. Let $i = k-1$ we have $\|\Delta_{\tilde{x}_k} - \widetilde{\Delta}_{x_k}\|_{x^*} = \|\Delta_{\bar{x}_{c,x}} - \sum_{i=0}^k c_i \Delta_{x_i}\|_{x^*} = O(d^3(x_0, x^*))$. Thus the proof is complete. □

E.4 Proof of Lemma 5

Proof of Lemma 5. Directly combining Lemma 15 and Lemma 4 gives the result. □

Lemma 15 (Convergence of the linearized iterates)**.** *Consider the linearized iterates* $\{\hat{x}_i\}_{i=0}^k$ *satisfying* (4) *for some* $G \succeq 0$ *with* $\|G\|_{x^*} \le \sigma < 1$. *Let* $\hat{r}_i = \Delta_{\hat{x}_{i+1}} - \Delta_{\hat{x}_i}$, $\hat{c}^* = \arg\min_{c^\top 1 = 1} \|\sum_{i=0}^k c_i \hat{r}_i\|_{x^*}^2 + \lambda \|c\|_2^2$. *Then*

$$\|\sum_{i=0}^k \hat{c}_i^* \Delta_{\hat{x}_i}\|_{x^*} \le \frac{d(x_0, x^*)}{1-\sigma}\sqrt{(S_{k,\bar{\lambda}}^{[0,\sigma]})^2 - \frac{\lambda}{d^2(x_0, x^*)}\|\hat{c}^*\|_2^2}$$

Proof of Lemma 15. The proof follows from (Scieur et al., 2020, Proposition 3.4) and we include it here for completeness. Denote $\mathcal{P}_k^1 := \{p \in \mathbb{R}[x] : \deg(p) = k, p(1) = 1\}$ as the set of polynomials of degree k with coefficients summing to 1. Noticing that $\hat{r}_i = \Delta_{\hat{x}_{i+1}} - \Delta_{\hat{x}_i} = \big(G - \mathrm{id}\big)[\Delta_{\hat{x}_i}] = \big(G - \mathrm{id}\big)G^i[\Delta_{x_0}]$, we have $\|\sum_{i=0}^k c_i \hat{r}_i\|_{x^*}^2 = \|(G - \mathrm{id})p(G)[\Delta_{x_0}]\|_{x^*}^2$ where $p \in \mathcal{P}_k^1$ and $\{c_i\}_{i=0}^k$ are the corresponding coefficients. Then we obtain

$$\begin{aligned}
\min_{p \in \mathcal{P}_k^1}\Big\{\|(G - \mathrm{id})p(G)[\Delta_{x_0}]\|_{x^*}^2 + \lambda\|c\|_2^2\Big\} &\le d^2(x_0, x^*) \min_{p \in \mathcal{P}_k^1}\Big\{\|p(G)\|_{x^*}^2 + \frac{\lambda}{d^2(x_0, x^*)}\|p\|_2^2\Big\} \\
&\le d^2(x_0, x^*) \min_{p \in \mathcal{P}_k^1} \max_{M : 0 \preceq M \preceq \sigma \mathrm{id}}\Big\{\|p(M)\|_{x^*}^2 + \frac{\lambda}{d^2(x_0, x^*)}\|p\|_2^2\Big\} \\
&= d^2(x_0, x^*) \min_{p \in \mathcal{P}_k^1} \max_{x \in [0, \sigma]}\Big\{p^2(x) + \frac{\lambda}{d^2(x_0, x^*)}\|p\|_2^2\Big\} \\
&= (S_{k,\bar{\lambda}}^{[0,\sigma]})^2 d^2(x_0, x^*),
\end{aligned}$$

where $\bar{\lambda} = \lambda / d^2(x_0, x^*)$ and we use the fact that $\|G - \mathrm{id}\|_{x^*} \le 1$. Then

$$\begin{aligned}
\|\sum_{i=0}^k \hat{c}_i^* \Delta_{\hat{x}_i}\|_{x^*}^2 &= \|\sum_{i=0}^k \hat{c}_i^* (G - \mathrm{id})^{-1}\hat{r}_i\|_{x^*}^2 \\
&\le \|(G - \mathrm{id})^{-1}\|_{x^*}^2 \Big(\|\sum_{i=0}^k \hat{c}_i^* \hat{r}_i\|_{x^*}^2 + \lambda\|\hat{c}^*\|_2^2 - \lambda\|\hat{c}^*\|_2^2\Big) \\
&\le \frac{d^2(x_0, x^*)}{(1-\sigma)^2}\Big((S_{k,\bar{\lambda}}^{[0,\sigma]})^2 - \frac{\lambda}{d^2(x_0, x^*)}\|\hat{c}^*\|_2^2\Big),
\end{aligned}$$

where we see that $\|(G - \mathrm{id})^{-1}\|_{x^*} \le \frac{1}{1-\sigma}$. □

E.5 Proof of Lemma 6

Proof of Lemma 6. From Proposition 2 and following (Scieur et al., 2020, Proposition 3.2), we obtain

$$\|c^*\| \leq \sqrt{\frac{\|R\|_2 + \lambda}{(k+1)\lambda}}.$$

Now we bound $\|R\|_2$. First we see R can be rewritten as $\mathcal{R}^\top \mathcal{G}_{x_k} \mathcal{R}$, where $\mathcal{G}_{x_k} \in \mathbb{R}^{r \times r}$ is the positive definite metric tensor at x_k and $\mathcal{R} = [\vec{r_i}] \in \mathbb{R}^{r \times k}$ is the collection of tangent vector in an orthonormal basis and r is the intrinsic dimension of the manifold. Thus we can write Riemannian inner product as $\langle r_i, r_j \rangle_{x_k} = \vec{r_i}^\top \mathcal{G}_{x_k} \vec{r_j}$ and

$$\|R\|_2 = \|\mathcal{G}_{x_k}^{1/2} \mathcal{R}\|_2^2 \leq \|\mathcal{G}_{x_k}^{1/2} R\|_{\mathrm{F}}^2 = \sum_{i=0}^{k} \vec{r_i}^\top \mathcal{G}_{x_k} \vec{r_i} = \sum_{i=0}^{k} \|r_i\|_{x_k}^2 = \sum_{i=0}^{k} d^2(x_i, x_{i+1}).$$

On the other hand, denote the perturbation matrix $P = R - \hat{R}$. Then from Proposition 2 and following (Scieur et al., 2020, Proposition 3.2), we have

$$\|\delta^c\|_2 \leq \frac{\|P\|_2}{\lambda} \|\hat{c}^*\|_2.$$

Now we need to bound $\|P\|_2$. Let $\mathcal{E}_i = \Delta_{x_i} - \Delta_{\hat{x}_i}$. Then we have

$$\begin{aligned} \|\Gamma_{x_k}^{x^*} r_i - \hat{r}_i\|_{x^*} &= \|\Gamma_{x_k}^{x^*} r_i - (\Delta_{x_{i+1}} - \Delta_{x_i}) + (\Delta_{x_{i+1}} - \Delta_{x_i}) - \hat{r}_i\|_{x^*} \\ &\leq \|\Gamma_{x_k}^{x^*} r_i - (\Delta_{x_{i+1}} - \Delta_{x_i})\|_{x^*} + \|(\Delta_{x_{i+1}} - \Delta_{x_i}) - \hat{r}_i\|_{x^*} \\ &= \|\Gamma_{x_k}^{x^*} r_i - (\Delta_{x_{i+1}} - \Delta_{x_i})\|_{x^*} + \|\mathcal{E}_{i+1} - \mathcal{E}_i\|_{x^*} \end{aligned} \tag{14}$$

where we use Lemma 2. Now we respectively bound each of the two terms on the right. First we see from Lemma 14,

$$\|\Gamma_{x_k}^{x^*} r_i - (\Delta_{x_{i+1}} - \Delta_{x_i})\|_{x^*} \leq C_0 d(x_i, x_k) d(x_k, x^*) d(x_i, x_{i+1}) + C_2 d(x_i, x^*) C_\kappa\big(d(x_i, x^*) + d(x_i, x_{i+1})\big) \tag{15}$$

Further, we bound $\|\mathcal{E}_{i+1} - \mathcal{E}_i\|_{x^*}$. From Lemma 3, we have $\mathcal{E}_i = G[\mathcal{E}_{i-1}] + \varepsilon_i, \mathcal{E}_0 = 0$ and

$$\|\mathcal{E}_{i+1} - \mathcal{E}_i\|_{x^*} = \|(G - \mathrm{id})\mathcal{E}_i + \varepsilon_{i+1}\|_{x^*} = \|(G - \mathrm{id}) \sum_{j=1}^{i} G^{i-j} \varepsilon_j + \varepsilon_{i+1}\|_{x^*} \leq \sum_{j=1}^{i+1} \|\varepsilon_j\|_{x^*}. \tag{16}$$

Combining (16), (15), (14) leads to

$$\|\Gamma_{x_k}^{x^*} r_i - \hat{r}_i\|_{x^*} \leq C_0 d(x_i, x_k) d(x_k, x^*) d(x_i, x_{i+1}) + C_2 d(x_i, x^*) C_\kappa\big(d(x_i, x^*) + d(x_i, x_{i+1})\big) + \sum_{j=1}^{i+1} \|\varepsilon_j\|_{x^*}.$$

Finally, recall we can write $R = \mathcal{R}^\top \mathcal{G}_{x_k} \mathcal{R}$ and similarly for $\hat{R} = \hat{\mathcal{R}}^\top \mathcal{G}_{x^*} \hat{\mathcal{R}}$ where $\hat{\mathcal{R}} = [\vec{\hat{r}_i}]$. By isometry of parallel transport, we have $R = \mathcal{R}_{x^*}^\top \mathcal{G}_{x^*} \mathcal{R}_{x^*}$ where $\mathcal{R}_{x^*} = [\overrightarrow{\Gamma_{x_k}^{x^*} r_i}]$. Let $E = \mathcal{G}_{x^*}^{1/2}(\mathcal{R}_{x^*} - \hat{\mathcal{R}})$. Then

$$\|P\|_2 = \|\mathcal{R}_{x^*}^\top \mathcal{G}_{x^*} \mathcal{R}_{x^*} - \hat{\mathcal{R}}^\top \mathcal{G}_{x^*} \hat{\mathcal{R}}\|_2 \leq 2\|E\|_2 \|\mathcal{G}_{x^*}^{1/2} \hat{R}\|_2 + \|E\|_2^2.$$

Notice that

$$\|\mathcal{G}_{x^*}^{1/2} \hat{R}\|_2 \leq \|\mathcal{G}_{x^*}^{1/2} \hat{R}\|_{\mathrm{F}} \leq \sum_{i=0}^{k} \|\hat{r}_i\|_{x^*} \leq \sum_{i=0}^{k} \|(G - \mathrm{id}) G^i \hat{r}_0\|_{x^*} \leq \sum_{i=0}^{k} \sigma^i \|\hat{r}_0\|_{x^*} \leq \frac{1 - \sigma^{k+1}}{1 - \sigma} d(x_0, x^*),$$

Also

$$\begin{aligned} \|E\|_2 = \|\mathcal{G}_{x^*}^{1/2}(\mathcal{R}_{x^*} - \hat{\mathcal{R}})\|_2 &\leq \sum_{i=0}^{k} \|\Gamma_{x_k}^{x^*} r_i - \hat{r}_i\|_{x^*} \\ &\leq d(x_k, x^*) C_0 \sum_{i=0}^{k} d(x_i, x_k) d(x_i, x_{i+1}) + C_2 \sum_{i=0}^{k} d(x_i, x^*) C_\kappa\big(d(x_i, x^*) + d(x_i, x_{i+1})\big) + \sum_{i=0}^{k} \sum_{j=1}^{i+1} \|\varepsilon_j\|_{x^*} \\ &= O(d^2(x_0, x^*)), \end{aligned}$$

where we notice that $C_\kappa(d(x_i, x^*) + d(x_i, x_{i+1})) = O(d^2(x_i, x^*))$ and recall that $\|\varepsilon_j\|_{x^*} = O(d^2(x_j, x^*)) = O(d^2(x_0, x^*))$. Thus $\|P\|_2 \leq 2\psi \frac{1-\sigma^{k+1}}{1-\sigma} d(x_0, x^*) + (\psi)^2$ where $\psi = O(d^2(x_0, x^*))$. □

E.6 Proof of Lemma 7

Proof of Lemma 7. From Lemma 2, we first observe that $d(\bar{x}_{\hat{c}^*,\hat{x}}, \bar{x}_{c^*,\hat{x}}) \leq C_1 \|\Delta_{\bar{x}_{\hat{c}^*,\hat{x}}} - \Delta_{\bar{x}_{c^*,\hat{x}}}\|_{x^*}$. Now we derive a bound on the term $\|\Delta_{\bar{x}_{\hat{c}^*,\hat{x}}} - \Delta_{\bar{x}_{c^*,\hat{x}}}\|_{x^*}$. Notice that from Lemma 4, we have

$$
\begin{aligned}
\|\Delta_{\bar{x}_{\hat{c}^*,\hat{x}}} - \Delta_{\bar{x}_{c^*,\hat{x}}}\|_{x^*} &= \| \sum_{i=0}^{k} (\hat{c}_i^* - c_i^*) \Delta_{\hat{x}_i} + \hat{\epsilon} \|_{x^*} \leq \|\delta^c\|_2 \big(\sum_{i=0}^{k} \|\Delta_{\hat{x}_i}\|_{x^*}^2 \big)^{1/2} + \|\hat{\epsilon}\|_{x^*} \\
&\leq \|\delta^c\|_2 (\sum_{i=0}^{k} \|\Delta_{\hat{x}_i}\|_{x^*}) + \|\hat{\epsilon}\|_{x^*} \\
&\leq \|\delta^c\|_2 (\sum_{i=0}^{k} \|G\|^i \|\Delta_{x_0}\|_{x^*}) + \|\hat{\epsilon}\|_{x^*} \\
&\leq \frac{1-\sigma^{k+1}}{1-\sigma} d(x_0, x^*) \|\delta^c\|_2 + \|\hat{\epsilon}\|_{x^*} \\
&\leq \frac{1}{1-\sigma} \frac{d(x_0, x^*)}{\lambda} \Big(\frac{1}{1-\sigma} 2\psi d(x_0, x^*) + (\psi)^2 \Big) \|\hat{c}^*\|_2 + \|\hat{\epsilon}\|_{x^*}
\end{aligned}
$$

for some $\|\hat{\epsilon}\|_{x^*} = O(d^3(x_0, x^*))$ and we denote $\delta^c = c^* - \hat{c}^*$. The bound on $\|\delta^c\|_2$ is from Lemma 6. □

E.7 Proof of Lemma 8

Proof of Lemma 8. Similarly to Lemma 7, we first see $d(\bar{x}_{c^*,\hat{x}}, \bar{x}_{c^*,x}) \leq C_1 \|\Delta_{\bar{x}_{c^*,\hat{x}}} - \Delta_{\bar{x}_{c^*,x}}\|_{x^*}$ due to Lemma 2. Again using Lemma 4, we see

$$
\begin{aligned}
\|\Delta_{\bar{x}_{c^*,\hat{x}}} - \Delta_{\bar{x}_{c^*,x}}\|_{x^*} = \| \sum_{i=0}^{k} c_i^* (\Delta_{x_i} - \Delta_{\hat{x}_i}) + \hat{\epsilon} \|_{x^*} &\leq \|c^*\|_2 (\sum_{i=0}^{k} \|\mathcal{E}_i\|_{x^*}^2)^{1/2} + \|\hat{\epsilon}\|_{x^*} \\
&\leq \|c^*\|_2 (\sum_{i=0}^{k} \|\mathcal{E}_i\|_{x^*}) + \|\hat{\epsilon}\|_{x^*}
\end{aligned}
$$

where $\|\hat{\epsilon}\|_{x^*} = O(d^3(x_0, x^*))$ and $\mathcal{E}_i = \Delta_{x_i} - \Delta_{\hat{x}_i}$. From Lemma 3, we have $\mathcal{E}_i = G[\mathcal{E}_{i-1}] + \varepsilon_i, \mathcal{E}_0 = 0$. Thus we can bound

$$
\|\mathcal{E}_i\|_{x^*} = \| \sum_{j=1}^{i} G^{i-j} \varepsilon_j \|_{x^*} \leq \sum_{j=1}^{i} \|\varepsilon_j\|_{x^*}.
$$

Then using Lemma 6 to bound $\|c^*\|_2$, we obtain

$$
\|\Delta_{\bar{x}_{c^*,\hat{x}}} - \Delta_{\bar{x}_{c^*,x}}\|_{x^*} \leq \sqrt{\frac{\sum_{i=0}^{k} d^2(x_i, x_{i+1}) + \lambda}{(k+1)\lambda}} \Big(\sum_{i=0}^{k} \sum_{j=0}^{i} \|\varepsilon_j\|_{x^*} \Big) + \epsilon_3,
$$

where $\epsilon_3 = O(d^3(x_0, x^*))$. □

E.8 Proof of Theorem 1

Proof of Theorem 1. Following the decomposition of error, we show

$$
\begin{aligned}
& d(\bar{x}_{c^*,x}, x^*) \\
& \leq d(\bar{x}_{\hat{c}^*,\hat{x}}, x^*) + d(\bar{x}_{\hat{c}^*,\hat{x}}, \bar{x}_{c^*,\hat{x}}) + d(\bar{x}_{c^*,\hat{x}}, \bar{x}_{c^*,x}) \\
& \leq \frac{d(x_0, x^*)}{1-\sigma} \sqrt{(S_{k,\bar{\lambda}}^{[0,\sigma]})^2 - \frac{\lambda}{d^2(x_0, x^*)} \|\hat{c}^*\|_2^2} + \frac{C_1 d(x_0, x^*)}{\lambda(1-\sigma)} \Big(\frac{2d(x_0, x^*)}{1-\sigma} \psi + (\psi)^2 \Big) \|\hat{c}^*\|_2 \\
& \qquad + C_1 \sqrt{\frac{\sum_{i=0}^{k} d^2(x_i, x_{i+1}) + \lambda}{(k+1)\lambda}} \Big(\sum_{i=0}^{k} \sum_{j=0}^{i} \|\varepsilon_j\|_{x^*} \Big) + \epsilon_1 + \epsilon_2 + \epsilon_3.
\end{aligned}
$$

Now we maximize the bound over $\|\hat{c}^*\|$. From (Scieur et al., 2020, Proposition A.1), we see the maximum of a function $g(x) = c\sqrt{a - \bar{\lambda} x^2} + bx$ is $\sqrt{a}\sqrt{c^2 + \frac{b^2}{\bar{\lambda}}}$ where $\bar{\lambda} = \lambda / d^2(x_0, x^*)$. Let $a = (S^{[0,\sigma]}_{k,\bar{\lambda}})^2$, $b = \frac{C_1 d(x_0,x^*)}{\lambda(1-\sigma)}\Big(\frac{2d(x_0,x^*)}{1-\sigma}\psi + (\psi)^2\Big)$, $c = \frac{d(x_0,x^*)}{1-\sigma}$. We then obtain

$$
\begin{aligned}
d(\bar{x}_{c^*,x}, x^*) \leq{}& S^{[0,\sigma]}_{k,\bar{\lambda}} \sqrt{\frac{d^2(x_0,x^*)}{(1-\sigma)^2} + \frac{C_1^2 d^4(x_0,x^*)\Big(\frac{2d(x_0,x^*)}{1-\sigma}\psi + (\psi)^2\Big)^2}{\lambda^3 (1-\sigma)^2}} \\
&+ C_1 \sqrt{\frac{\sum_{i=0}^k d^2(x_i, x_{i+1}) + \lambda}{(k+1)\lambda}} \Big(\sum_{i=0}^k \sum_{j=0}^i \|\varepsilon_j\|_{x^*} \Big) + \epsilon_1 + \epsilon_2 + \epsilon_3,
\end{aligned}
$$

which completes the proof. □

E.9 Proof of Proposition 1

Proof of Proposition 1. Dividing the bound from Theorem 1 by $d(x_0, x^*)$ gives

$$
\begin{aligned}
\frac{d(\bar{x}_{c^*,x}, x^*)}{d(x_0,x^*)} \leq{}& \frac{S^{[0,\sigma]}_{k,\bar{\lambda}}}{1-\sigma} \sqrt{1 + O(d^{(2-3s)}(x_0,x^*)\Big(\frac{2d(x_0,x^*)}{1-\sigma}\psi + (\psi)^2\Big)^2} \\
&+ C_1 \sqrt{\frac{\sum_{i=0}^k d^2(x_i, x_{i+1})}{(k+1)O(d^s(x_0,x^*))} + \frac{1}{k+1}} \Big(\sum_{i=0}^k \sum_{j=0}^i \|\varepsilon_j\|_{x^*} \Big) + \frac{1}{d(x_0,x^*)}\big(\epsilon_1 + \epsilon_2 + \epsilon_3\big).
\end{aligned}
$$

By $\psi = O(d^2(x_0,x^*))$, the first term of the bound simplifies to $\frac{S^{[0,\sigma]}_{k,\bar{\lambda}}}{1-\sigma}\sqrt{1 + O(d^{(8-3s)}(x_0,x^*))}$, and similarly because $d(x_i, x_{i+1}) = O(d(x_0,x^*))$, $\|\varepsilon_j\|_{x^*} = O(d^2(x_0,x^*))$ under Assumption 1, the second term simplifies to $O(\sqrt{d^2(x_0,x^*) + d^{(4-s)}(x_0,x^*)})$ and the last term reduces to $O(d^2(x_0,x^*))$ as $\epsilon_1, \epsilon_2, \epsilon_3 = O(d^3(x_0,x^*))$. Hence we obtain

$$
\frac{d(\bar{x}_{c^*,x}, x^*)}{d(x_0,x^*)} \leq \frac{S^{[0,\sigma]}_{k,\bar{\lambda}}}{1-\sigma}\sqrt{1 + O(d^{(8-3s)}(x_0,x^*))} + O(\sqrt{d^2(x_0,x^*) + d^{(4-s)}(x_0,x^*)}) + O(d^2(x_0,x^*)).
$$

Finally we notice that the last two terms vanishes when $d(x_0,x^*) \to 0$ for the choice of s. For the first term, given that when $d(x_0,x^*) \to 0$, $\bar{\lambda} = O(d^{(s-2)}(x_0,x^*)) \to 0$ and $O(d^{(8-3s)}(x_0,x^*)) \to 0$ for $s \in (2, \frac{8}{3})$, then

$$
\lim_{d(x_0,x^*) \to 0} \frac{S^{[0,\sigma]}_{k,\bar{\lambda}}}{1-\sigma}\sqrt{1 + O(d^{(2-3s)}(x_0,x^*))} = \frac{S^{[0,\sigma]}_{k,0}}{1-\sigma} = \frac{1}{1-\sigma}\frac{2}{\beta^{-k} + \beta^k}
$$

where $\beta = \frac{1-\sqrt{1-\sigma}}{1+\sqrt{1-\sigma}}$. This follows because without regularization, $S^{[0,\sigma]}_{k,0}$ reduces to the rescaled and shifted Chebyshev polynomial. See for example (d'Aspremont et al., 2021). □

E.10 Proof of Lemma 9

Proof of Lemma 9. First, we write

$$
\sum_{i=0}^k c_i \Delta_{x_i} = \Delta_{x_k} - \sum_{i=0}^{k-1} \theta_i (\Delta_{x_{i+1}} - \Delta_{x_i}).
$$

By Lemma 1, we obtain

$$
\begin{aligned}
d\Big(\mathrm{Exp}_{x^*}\big(\sum_{i=0}^k c_i \Delta_{x_i}\big), \mathrm{Exp}_{x_k}\big(-\Gamma_{x^*}^{x_k} \sum_{i=0}^{k-1} \theta_i (\Delta_{x_{i+1}} - \Delta_{x_i})\big)\Big) \\
&\leq d(x_k, x^*) C_\kappa \Big(d(x_k, x^*) + \|\sum_{i=0}^{k-1} \theta_i (\Delta_{x_{i+1}} - \Delta_{x_i})\|_{x^*} \Big) \\
&\leq d(x_k, x^*) C_\kappa \Big(d(x_k, x^*) + \sum_{i=0}^{k-1} \theta_i (d(x_{i+1}, x^*) + d(x_i, x^*)) \Big),
\end{aligned}
$$

where we use the fact that $C_\kappa(x)$ is increasing for $x > 0$. In addition, from Lemma 2,

$$\begin{aligned} d\Big(\bar{x}_{c,x}, \mathrm{Exp}_{x_k}\big(-\Gamma_{x^*}^{x_k}\sum_{i=0}^{k-1}\theta_i(\Delta_{x_{i+1}} - \Delta_{x_i})\big)\Big) &\leq C_1 \| \sum_{i=0}^{k-1} \theta_i \big(\Gamma_{x^*}^{x_k}(\Delta_{x_{i+1}} - \Delta_{x_i}) - \Gamma_{x_i}^{x_k}\mathrm{Exp}_{x_i}^{-1}(x_{i+1})\big)\|_{x_k} \\ &\leq C_1 \sum_{i=0}^{k-1} \theta_i \|\Delta_{x_{i+1}} - \Delta_{x_i} - \Gamma_{x_k}^{x^*}\Gamma_{x_i}^{x_k}\mathrm{Exp}_{x_i}^{-1}(x_{i+1})\|_{x^*}. \end{aligned}$$

Using Lemma 14, we obtain

$$\begin{aligned} \|\Delta_{x_{i+1}} - \Delta_{x_i} - \Gamma_{x_k}^{x^*}\Gamma_{x_i}^{x_k}\mathrm{Exp}_{x_i}^{-1}(x_{i+1})\|_{x^*} &\leq C_0 d(x_i, x_k) d(x_k, x^*) d(x_i, x_{i+1}) \\ &\quad + C_2 d(x_i, x^*) C_\kappa\big(d(x_i, x^*) + d(x_i, x_{i+1})\big). \end{aligned}$$

Let $e = \Delta_{\bar{x}_{c,x}} - \sum_{i=0}^k c_i \Delta_{x_i}$. Now combining the above results gives

$$\begin{aligned} \|e\|_{x^*} &= \|\Delta_{\bar{x}_{c,x}} - \sum_{i=0}^k c_i \Delta_{x_i}\|_{x^*} \\ &\leq C_2 d\Big(\bar{x}_{c,x}, \mathrm{Exp}_{x^*}\big(\sum_{i=0}^k c_i \Delta_{x_i}\big)\Big) \\ &\leq C_2 d\Big(\bar{x}_{c,x}, \mathrm{Exp}_{x_k}\big(-\Gamma_{x^*}^{x_k}\sum_{i=0}^{k-1}\theta_i(\Delta_{x_{i+1}} - \Delta_{x_i})\big)\Big) \\ &\quad + C_2 d\Big(\mathrm{Exp}_{x^*}\big(\sum_{i=0}^k c_i\Delta_{x_i}\big), \mathrm{Exp}_{x_k}\big(-\Gamma_{x^*}^{x_k}\sum_{i=0}^{k-1}\theta_i(\Delta_{x_{i+1}} - \Delta_{x_i})\big)\Big) \\ &\leq C_2 C_1 \sum_{i=0}^{k-1}\theta_i \Big(C_0 d(x_i, x_k) d(x_k, x^*) d(x_i, x_{i+1}) + C_2 d(x_i, x^*) C_\kappa\big(d(x_i, x^*) + d(x_i, x_{i+1})\big)\Big) \\ &\quad + C_2 d(x_k, x^*) C_\kappa\Big(d(x_k, x^*) + \sum_{i=0}^{k-1}\theta_i (d(x_{i+1}, x^*) + d(x_i, x^*))\Big). \end{aligned}$$

Under Assumption 1 and $C_\kappa(x) = O(x^2)$, we see $\|e\|_{x^*} = O(d^3(x_0, x^*))$. □

E.11 Proof of Lemma 10

Proof of Lemma 10. Let $D(x) \coloneqq \frac{1}{2}\sum_{i=0}^k c_i d^2(x, x_i)$. Then we can show $\mathrm{grad}D(x) = -\sum_{i=0}^k c_i \mathrm{Exp}_x^{-1}(x_i)$. See for example (Alimisis et al., 2020). By the first-order stationarity,

$$\mathrm{grad}D(\bar{x}_{c,x}) = -\sum_{i=0}^k c_i \mathrm{Exp}_{\bar{x}_{c,x}}^{-1}(x_i) = 0$$

and $\mathrm{grad}D(x^*) = -\sum_{i=0}^k c_i \mathrm{Exp}_{x^*}^{-1}(x_i)$.

The first claim that $d(\bar{x}_{c,x}, x^*) \leq \|\sum_{i=0}^k c_i \Delta_{x_i}\|_{x^*}$ follows from Lemma (Tripuraneni et al., 2018, Lemma 10) and we include here for completeness. Define a real-valued function $g(t) \coloneqq D\big(\mathrm{Exp}_{x^*}(t\eta)\big)$ with $\eta = \frac{\Delta_{\bar{x}_{c,x}}}{\|\Delta_{\bar{x}_{c,x}}\|_{x^*}}$. Under the assumption and definition of geodesic μ-strongly convex, we see $g(t)$ is μ-strongly convex in t. Thus, we have $g'(t_0) - g'(0) \geq \mu t_0$ for any t_0. Let $t_0 = \|\Delta_{\bar{x}_{c,x}}\|_{x^*}$ and denote the geodesic $\gamma(t) \coloneqq \mathrm{Exp}_{x^*}(t\eta)$. We derive that $g'(t) = \langle \mathrm{grad}D(\mathrm{Exp}_{x^*}(t\eta)), \gamma'(t)\rangle$ by chain rule. Then we have $g'(t_0) = \langle \mathrm{grad}D(\bar{x}_{c,x}), \gamma'(t_0)\rangle_{\bar{x}_{c,x}} = 0$ and $g'(0) = \langle \mathrm{grad}D(x^*), \eta\rangle$. Finally, we see

$$\|\mathrm{grad}D(x^*)\|_{x^*}^2 \geq (g'(0))^2 = (g'(t_0) - g'(0))^2 \geq \mu^2 t_0^2 = \mu^2 \|\Delta_{\bar{x}_{c,x}}\|_{x^*}^2,$$

where the first inequality is due to Cauchy–Schwarz inequality. The first claim is proved by noticing $\|\mathrm{grad}D(x^*)\|_{x^*} = \|\sum_{i=0}^k c_i \Delta_{x_i}\|_{x^*}$ and $\|\Delta_{\bar{x}_{c,x}}\|_{x^*} = d(\bar{x}_{c,x}, x^*)$.

For the second claim, we first observe from the proof of Lemma 14 that

$$\begin{aligned}\|\mathrm{Exp}^{-1}_{\bar{x}_{c,x}}(x_i) - \Gamma_{x^*}^{\bar{x}_{c,x}}\big(\mathrm{Exp}^{-1}_{x^*}(x_i) - \mathrm{Exp}^{-1}_{x^*}(\bar{x}_{c,x})\big)\|_{\bar{x}_{c,x}} &\leq C_2 d(\bar{x}_{c,x}, x^*) C_\kappa \big(d(\bar{x}_{c,x}, x^*) + d(\bar{x}_{c,x}, x_i)\big) \\ &= O(d^3(x_0, x^*)),\end{aligned}$$

where the order can be seen due to that $d(\bar{x}_{c,x}, x^*) \leq \frac{1}{\mu}\sum_{i=0}^k c_i d(x_i, x^*) = O(x_0, x^*)$ from the first claim. Thus let $\bar{\varepsilon} \coloneqq \mathrm{Exp}^{-1}_{\bar{x}_{c,x}}(x_i) - \Gamma_{x^*}^{\bar{x}_{c,x}}\big(\mathrm{Exp}^{-1}_{x^*}(x_i) - \mathrm{Exp}^{-1}_{x^*}(\bar{x}_{c,x})\big)$, we have $\|\bar{\varepsilon}\|_{\bar{x}_{c,x}} = O(d^3(x_0, x^*))$. From the first order stationarity, we see

$$\begin{aligned}0 = \sum_{i=0}^k c_i \mathrm{Exp}^{-1}_{\bar{x}_{c,x}}(x_i) &= \sum_{i=0}^k c_i \Big(\Gamma_{x^*}^{\bar{x}_{c,x}}\big(\mathrm{Exp}^{-1}_{x^*}(x_i) - \mathrm{Exp}^{-1}_{x^*}(\bar{x}_{c,x})\big) + \bar{\varepsilon}\Big) \\ &= \Gamma_{x^*}^{\bar{x}_{c,x}}\Big(\sum_{i=0}^k c_i \Delta_{x_i} - \Delta_{\bar{x}_{c,x}}\Big) + \bar{\varepsilon}.\end{aligned}$$

Taking the norm and using the isometry of parallel transport, we obtain the desired result. □

F PROOFS UNDER GENERAL RETRACTION AND VECTOR TRANSPORT

Discussions on the Assumptions Before we prove the results, we discuss the assumptions made for the general setup. In particular, Assumption 4 is required to bound the deviation from the retraction to the exponential map, which can be considered natural given retraction approximates the exponential map to the first-order. In fact, Assumption 4 has been commonly used in Sato et al. (2019); Kasai et al. (2018); Han and Gao (2021a) for analyzing Riemannian first-order algorithms using retraction and can be satisfied for a sufficiently small neighbourhood (see for example Ring and Wirth (2012); Huang et al. (2015a)). Similarly, Assumption 5 is used to bound the deviation between the vector transport to parallel transport, which is also standard in Huang et al. (2015b); Kasai et al. (2018); Han and Gao (2021a). One can follow the procedures in Huang et al. (2015b) to construct isometric vector transport that satisfies such condition for common manifolds like SPD manifold (Huang et al., 2015b), Stiefel and Grassmann manifold (Huang, 2013).

Here we show that when we use general retraction Retr in place of the exponential map Exp, thus avoiding the lemma on metric distortion (Lemma 1, 2), we can still show a similar result as Lemma 4 but with an error on the order of $O(d^2(x_0, x^*))$ instead of $O(d^3(x_0, x^*))$ as for the case of exponential map. The main idea of proof follows from Tripuraneni et al. (2018). The next proposition formalizes such claim. For this section, we denote $\Delta_x = \mathrm{Retr}^{-1}_{x^*}(x)$ for any $x \in \mathcal{X}$ where the retraction has a smooth inverse. For general retraction, the deviation is on the order of $O(\|\Delta_{x_0}\|^2_{x^*}) = O(d^2(x_0, x^*))$ where we use the fact that retraction approximates the exponential map to the first order.

Proposition 3. *Suppose all iterates $x_i \in \mathcal{X}$, a neighbourhood where retraction has a smooth inverse. Consider the weighted average $\bar{x}_{c,x} = \tilde{x}_k$ given by* (Avg.1) *with retraction. Assume the sequence of iterates is non-divergent in retraction, i.e. $\|\Delta_{x_i}\|_{x^*}, \|\Delta_{\tilde{x}_i}\|_{x^*} = O(\|\Delta_{x_0}\|_{x^*})$. Then we have $\Delta_{\bar{x}_{c,x}} = \sum_{i=0}^k c_i \Delta_{x_i} + e$, with $\|e\|_{x^*} = O(\|\Delta_{x_0}\|^2_{x^*})$,*

Proof. The proof generalize the proof of (Tripuraneni et al., 2018, Lemma 12). First denote $\mathrm{Retr}_x^y \coloneqq \mathrm{Retr}_y^{-1} \circ \mathrm{Retr}_x$ and we notice that

$$\begin{aligned}\Delta_{\tilde{x}_{i+1}} = \mathrm{Retr}^{-1}_{x^*}(\tilde{x}_{i+1}) = \mathrm{Retr}^{-1}_{x^*}\Big(\mathrm{Retr}_{\tilde{x}_i}\big(\gamma_{i+1}\mathrm{Retr}^{-1}_{\tilde{x}_i}(x_{i+1})\big)\Big) &= \mathrm{Retr}_{\tilde{x}_i}^{x^*}\Big(\gamma_{i+1}\mathrm{Retr}^{-1}_{\tilde{x}_i}\big(\mathrm{Retr}_{x^*}(\Delta_{x_{i+1}})\big)\Big) \\ &= \mathrm{Retr}_{\tilde{x}_i}^{x^*}\Big(\gamma_{i+1}\big(\mathrm{Retr}_{\tilde{x}_i}^{x^*}\big)^{-1}(\Delta_{x_{i+1}})\Big) \\ &= F(\Delta_{x_{i+1}}),\end{aligned}$$

where we denote $\gamma_i = \frac{c_i}{\sum_{j=0}^i c_j}$ and $F : T_{x^*}\mathcal{M} \rightarrow T_{x^*}\mathcal{M}$ defined as $F(u) = \mathrm{Retr}_{\tilde{x}_i}^{x^*}\Big(\gamma_{i+1}\big(\mathrm{Retr}_{\tilde{x}_i}^{x^*}\big)^{-1}(u)\Big)$. In addition, it can be verified that $F(\Delta_{\tilde{x}_i}) = \Delta_{\tilde{x}_i}$.

Now by chain rule, we have

$$\begin{aligned}
\mathrm{D}F(u) &= \mathrm{D}\mathrm{Retr}_{\tilde{x}_i}^{x^*}\Big(\gamma_{i+1}(\mathrm{Retr}_{\tilde{x}_i}^{x^*})^{-1}(u)\Big)\Big[\mathrm{D}\gamma_{i+1}(\mathrm{Retr}_{\tilde{x}_i}^{x^*})^{-1}(u)\Big] \\
&= \gamma_{i+1}\mathrm{D}\Big(\frac{1}{\gamma_{i+1}}\mathrm{Retr}_{\tilde{x}_i}^{x^*}\Big)\Big(\gamma_{i+1}(\mathrm{Retr}_{\tilde{x}_i}^{x^*})^{-1}(u)\Big)\Big[\mathrm{D}\gamma_{i+1}(\mathrm{Retr}_{\tilde{x}_i}^{x^*})^{-1}(u)\Big] \\
&= \gamma_{i+1}\Big(\mathrm{D}\gamma_{i+1}(\mathrm{Retr}_{\tilde{x}_i}^{x^*})^{-1}(u)\Big)^{-1}\Big[\mathrm{D}\gamma_{i+1}(\mathrm{Retr}_{\tilde{x}_i}^{x^*})^{-1}(u)\Big] = \gamma_{i+1}\mathrm{id},
\end{aligned}$$

where the third inequality uses the inverse function theorem. Hence the Taylor expansion of F at $\Delta_{\tilde{x}_i}$ up to second order gives

$$\begin{aligned}
\Delta_{\tilde{x}_{i+1}} = F(\Delta_{x_{i+1}}) &= F(\Delta_{\tilde{x}_i}) + \gamma_{i+1}(\Delta_{x_{i+1}} - \Delta_{\tilde{x}_i}) + \tilde{\epsilon}_i \\
&= (1-\gamma_{i+1})\Delta_{\tilde{x}_i} + \gamma_{i+1}\Delta_{x_{i+1}} + \tilde{\epsilon}_i.
\end{aligned}$$

where we let $\tilde{\epsilon}_i = O(\|\Delta_{x_{i+1}} - \Delta_{\tilde{x}_i}\|_{x^*}^2)$. From the expansion, it follows that $\Delta_{\tilde{x}_{i+1}} = \frac{\sum_{j=0}^{i} c_i}{\sum_{j=0}^{i+1} c_j}\Delta_{\tilde{x}_i} + \frac{c_{i+1}}{\sum_{j=0}^{i+1} c_j}\Delta_{x_{i+1}} + \tilde{\epsilon}_i$, which yields

$$(\sum_{j=0}^{i+1} c_j)\Delta_{\tilde{x}_{i+1}} = (\sum_{j=0}^{i} c_j)\Delta_{\tilde{x}_i} + c_{i+1}\Delta_{x_{i+1}} + (\sum_{j=0}^{i} c_j)\tilde{\epsilon}_i = \sum_{j=0}^{i+1} c_j\Delta_{x_j} + \sum_{j=0}^{i}(\sum_{\ell=0}^{j} c_\ell)\tilde{\epsilon}_j,$$

where the second equality follows by expanding the first equality. Let $i = k-1$, this leads to

$$\Delta_{\bar{x}_{c,x}} = \Delta_{\tilde{x}_k} = \sum_{j=0}^{k} c_j\Delta_{x_j} + e,$$

where we let $e = \sum_{j=0}^{k-1}(\sum_{\ell=0}^{j} c_\ell)\tilde{\epsilon}_j = O\big(\sum_{j=0}^{k-1}(\sum_{\ell=0}^{j} c_\ell)(\|\Delta_{x_{j+1}}\|_{x^*}^2 + \|\Delta_{\tilde{x}_j}\|_{x^*}^2)\big)$. We observe that $\|\Delta_{x_{i+1}}\|_{x^*}^2 = O(\|\Delta_{x_0}\|_{x^*}^2)$ and $\|\Delta_{\tilde{x}_j}\|_{x^*}^2 = O(\|\Delta_{x_0}\|_{x^*}^2)$ due to the non-divergent assumption. The proof is complete. □

F.1 Proof of Theorem 2

Theorem 2 (Restatement). Under Assumption 1, 3, 4 and 5, let $\{x_i\}_{i=0}^k$ be given by Riemannian gradient descent via retraction, i.e., $x_i = \mathrm{Retr}_{x_{i-1}}(-\eta\,\mathrm{grad}f(x_{i-1}))$ and $\{\hat{x}_i\}_{i=0}^k$ be the linearized iterates satisfying $\mathrm{Retr}_{x^*}^{-1}(\hat{x}_i) = G[\mathrm{Retr}_{x^*}^{-1}(\hat{x}_{i-1})]$ with $G = \mathrm{id} - \eta\,\mathrm{Hess}f(x^*)$, satisfying $\|G\|_{x^*} \leq \sigma < 1$. Then, using retraction and vector transport in Algorithm 1 and letting $\bar{x}_{c,x}$ be computed from (5), it satisfies that

$$\begin{aligned}
d(\bar{x}_{c^*,x}, x^*) \leq \|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*}\frac{S_{k,\bar{\lambda}}^{[0,\sigma]}}{1-\sigma}&\sqrt{\frac{1}{a_0^2} + \frac{C_1^2\|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*}^2\big(\frac{2\psi}{1-\sigma}\|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*} + \psi^2\big)^2}{\lambda^3}} \\
&+ C_1\sqrt{\frac{\sum_{i=0}^{k}\|\mathrm{Retr}_{x_i}^{-1}(x_{i+1})\|_{x_i}^2 + \lambda}{(k+1)\lambda}}\Big(\sum_{i=0}^{k}\sum_{j=0}^{i}\|\varepsilon_j\|_{x^*}\Big) + \epsilon_1 + \epsilon_2 + \epsilon_3,
\end{aligned}$$

where $\psi = O(d^2(x_0, x^*))$, $\epsilon_1, \epsilon_2, \epsilon_3 = O(d^2(x_0, x^*))$ and $\varepsilon_i = O(d^2(x_i, x^*))$. Under the same choice of $\lambda = O(d^s(x_0, x^*))$, $s \in (2, \frac{8}{3})$, the same asymptotic optimal convergence rate (Proposition 1) holds.

Proof of Theorem 2. Here we only provide a sketch of proof because the main idea is exactly the same as the case of exponential map.

Under general retraction and vector transport, an analogue of Lemma 3 holds. That is,

$$\mathrm{Retr}_{x^*}^{-1}(x_i) = (\mathrm{id} - \eta\mathrm{Hess}f(x^*))[\mathrm{Retr}_{x^*}^{-1}(x_{i-1})] + \varepsilon_i, \tag{17}$$

where $\|\varepsilon_i\|_{x^*} = O(d^2(x_i, x^*))$. To show (17), we follow the exact same steps as the proof for Lemma 3 where we replace exponential map with retraction. The only difference is that the second order derivative is no longer the Riemann curvature tensor. In addition, we have shown in Proposition 3 that for retraction, we also have

$$\mathrm{Retr}_{x^*}^{-1}(\bar{x}_{c,x}) = \sum_{i=0}^{k} c_i\mathrm{Retr}_{x^*}^{-1}(x_i) + e \tag{18}$$

with $\|e\|_{x^*} = O(d^2(x_0, x^*))$.

Further, we still consider the same error bound decomposition, i.e.,

$$d(\bar{x}_{c^*,x}, x^*) \leq d(\bar{x}_{\hat{c}^*,\hat{x}}, x^*) + d(\bar{x}_{\hat{c}^*,\hat{x}}, \bar{x}_{c^*,\hat{x}}) + d(\bar{x}_{c^*,\hat{x}}, \bar{x}_{c^*,x}).$$

(I). For the linear term $d(\bar{x}_{\hat{c}^*,\hat{x}}, x^*)$, we first see the linearized iterates $\hat{x}_i$ enjoys the same convergence as in Lemma 15 that

$$\| \sum_{i=0}^{k} \hat{c}_i^* \mathrm{Retr}_{x^*}^{-1}(\hat{x}_i) \|_{x^*} \leq \frac{\|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*}}{1-\sigma} \sqrt{(S_{k,\bar{\lambda}}^{[0,\sigma]})^2 - \frac{\lambda}{\|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*}^2} \|\hat{c}^*\|_2^2}, \tag{19}$$

where $\bar{\lambda} \coloneqq \lambda / \|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*}^2$ and we use Assumption 4. Combining (19) with (18) yields

$$\begin{aligned} d(\bar{x}_{\hat{c}^*,\hat{x}}, x^*) \leq \frac{1}{a_0} \|\mathrm{Retr}_{x^*}^{-1}(\bar{x}_{\hat{c}^*,\hat{x}})\|_{x^*} &\leq \| \sum_{i=0}^{k} \hat{c}_i^* \mathrm{Retr}_{x^*}^{-1}(\hat{x}_i) \|_{x^*} + \epsilon_1, \\ &\leq \frac{\|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*}}{a_0(1-\sigma)} \sqrt{(S_{k,\bar{\lambda}}^{[0,\sigma]})^2 - \frac{\lambda}{\|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*}^2} \|\hat{c}^*\|_2^2} + \epsilon_1, \end{aligned}$$

with $\epsilon_1 = O(d^2(x_0, x^*))$.

(II). For the stability term $d(\bar{x}_{\hat{c}^*,\hat{x}}, \bar{x}_{c^*,\hat{x}})$, we first use Assumption 4 to show

$$\begin{aligned} \|\Delta_{\bar{x}_{\hat{c}^*,\hat{x}}} - \Delta_{\bar{x}_{c^*,\hat{x}}} - \big(\mathrm{Retr}_{x^*}^{-1}(\bar{x}_{\hat{c}^*,\hat{x}}) - \mathrm{Retr}_{x^*}^{-1}(\bar{x}_{c^*,\hat{x}})\big)\|_{x^*} &\leq a_2 \|\mathrm{Retr}_{x^*}^{-1}(\bar{x}_{\hat{c}^*,\hat{x}})\|_{x^*}^2 + a_2 \|\mathrm{Retr}_{x^*}^{-1}(\bar{x}_{c^*,\hat{x}})\big)\|_{x^*}^2 \\ &\leq a_2 a_1^2 \big(d^2(\bar{x}_{\hat{c}^*,\hat{x}}, x^*) + d^2(\bar{x}_{c^*,\hat{x}}, x^*)\big) \\ &= O(d^2(x_0, x^*)). \end{aligned}$$

Let $\epsilon_r \coloneqq \Delta_{\bar{x}_{\hat{c}^*,\hat{x}}} - \Delta_{\bar{x}_{c^*,\hat{x}}} - \big(\mathrm{Retr}_{x^*}^{-1}(\bar{x}_{\hat{c}^*,\hat{x}}) - \mathrm{Retr}_{x^*}^{-1}(\bar{x}_{c^*,\hat{x}})\big)$, we have $\|\epsilon_r\|_{x^*} = O(d^2(x_0, x^*))$. In addition, based on Assumption 5, we show

$$\begin{aligned} &\|\mathcal{T}_{x_k}^{x^*} r_i - \big(\mathrm{Retr}_{x^*}^{-1}(x_{i+1}) - \mathrm{Retr}_{x^*}^{-1}(x_i)\big) - \Gamma_{x_k}^{x^*} r_i + \big(\Delta_{x_{i+1}} - \Delta_{x_i}\big)\|_{x^*} \\ &\leq \|\mathcal{T}_{x_k}^{x^*} r_i - \Gamma_{x_k}^{x^*} r_i\|_{x^*} + O(d^2(x_0, x^*)) = O(d^2(x_0, x^*)), \end{aligned}$$

where we use Assumption 4, 5 and notice $\|r_i\|_{x_i} = \|\mathrm{Retr}_{x_i}^{-1}(x_{i+1})\|_{x_i} \leq a_1 d(x_i, x_{i+1}) = O(d(x_0, x^*))$. Let $\epsilon_v \coloneqq \mathcal{T}_{x_k}^{x^*} r_i - \big(\mathrm{Retr}_{x^*}^{-1}(x_{i+1}) - \mathrm{Retr}_{x^*}^{-1}(x_i)\big) - \Gamma_{x_k}^{x^*} r_i + \big(\Delta_{x_{i+1}} - \Delta_{x_i}\big)$, we have $\|\epsilon_v\|_{x^*} = O(d^2(x_0, x^*))$.

Using Lemma 2, we then obtain

$$\begin{aligned} d(\bar{x}_{\hat{c}^*,\hat{x}}, \bar{x}_{c^*,\hat{x}}) \leq C_1 \|\Delta_{\bar{x}_{\hat{c}^*,\hat{x}}} - \Delta_{\bar{x}_{c^*,\hat{x}}}\|_{x^*} &\leq C_1 \|\mathrm{Retr}_{x^*}^{-1}(\bar{x}_{\hat{c}^*,\hat{x}}) - \mathrm{Retr}_{x^*}^{-1}(\bar{x}_{c^*,\hat{x}})\|_{x^*} + C_1 \|\epsilon_r\|_{x^*} \\ &\leq \frac{C_1 \|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*}}{1-\sigma} \|c^* - \hat{c}^*\|_2 + O(d^2(x_0, x^*)), \end{aligned}$$

where we apply (18). Now we proceed to bound $\|c^* - \hat{c}^*\|_2 \leq \frac{\|P\|_2}{\lambda} \|\hat{c}^*\|_2$ in a similar manner as Lemma 6 where $P = R - \hat{R}$. From the proof of Lemma 6, we have

$$\|P\|_2 \leq \frac{2}{1-\sigma} \|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*} \|E\|_2 + \|E\|_2^2,$$

where $\|E\|_2 \leq \sum_{i=0}^{k} \|\mathcal{T}_{x_k}^{x^*} r_i - \hat{r}_i\|_{x^*}$. Thus it remains to bound $\|\mathcal{T}_{x_k}^{x^*} r_i - \hat{r}_i\|_{x^*}$. Similarly, we can show

$$\begin{aligned} \|\mathcal{T}_{x_k}^{x^*} r_i - \hat{r}_i\|_{x^*} &\leq \|\mathcal{T}_{x_k}^{x^*} r_i - \big(\mathrm{Retr}_{x^*}^{-1}(x_{i+1}) - \mathrm{Retr}_{x^*}^{-1}(x_i)\big)\|_{x^*} + \sum_{j=1}^{i+1} \|\varepsilon_j\|_{x^*} \\ &\leq \|\Gamma_{x_k}^{x^*} r_i - \big(\Delta_{x_{i+1}} - \Delta_{x_i}\big)\|_{x^*} + \|\epsilon_v\|_{x^*} + \sum_{j=1}^{i+1} \|\varepsilon_j\|_{x^*} = O(d^2(x_0, x^*)), \end{aligned}$$

where ε_j is defined in (17) and we use Lemma 14 for the exponential map. Thus $\|P\|_2 \leq 2\psi \frac{a_1}{1-\sigma} d(x_0, x^*) + \psi^2$ where $\psi = O(d^2(x_0, x^*))$. This leads to

$$d(\bar{x}_{\hat{c}^*,\hat{x}}, \bar{x}_{c^*,\hat{x}}) \leq \frac{C_1 \|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*}}{\lambda(1-\sigma)} \Big(\frac{2\psi}{1-\sigma} \|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*} + \psi^2 \Big) \|\hat{c}^*\|_2 + \epsilon_2,$$

where $\epsilon_2 = O(d^2(x_0, x^*))$.

(III). Finally for the nonlinearity term $d(\bar{x}_{c^*,\hat{x}}, \bar{x}_{c^*,x})$, we show

$$\begin{aligned} d(\bar{x}_{c^*,\hat{x}}, \bar{x}_{c^*,x}) &\leq C_1 \|\Delta_{\bar{x}_{c^*,\hat{x}}} - \Delta_{\bar{x}_{c^*,x}}\|_{x^*} \leq C_1 \|\mathrm{Retr}_{x^*}^{-1}(\bar{x}_{c^*,\hat{x}}) - \mathrm{Retr}_{x^*}^{-1}(\bar{x}_{c^*,x})\|_{x^*} + O(d^2(x_0, x^*)) \\ &\leq C_1 \|c^*\|_2 (\sum_{i=0}^k \|\mathrm{Retr}_{x^*}^{-1}(x_i) - \mathrm{Retr}_{x^*}^{-1}(\hat{x}_i)\|_{x^*}) + O(d^2(x_0, x^*)) \\ &\leq C_1 \sqrt{\frac{\sum_{i=0}^k \|\mathrm{Retr}_{x_i}^{-1}(x_{i+1})\|_{x_i}^2 + \lambda}{(k+1)\lambda}} \Big(\sum_{i=0}^k \sum_{j=0}^i \|\varepsilon_j\|_{x^*} \Big) + \epsilon_3, \end{aligned}$$

where $\epsilon_3 = O(d^2(x_0, x^*))$ and we follow similar steps as in Lemma 6.

Finally, combining results from (I), (II), (III), we have

$$\begin{aligned} d(\bar{x}_{c^*,x}, x^*) \leq\ & \frac{\|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*}}{a_0(1-\sigma)} \sqrt{(S_{k,\bar{\lambda}}^{[0,\sigma]})^2 - \frac{\lambda}{\|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*}^2} \|\hat{c}^*\|_2^2} \\ &+ \frac{C_1 \|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*}}{\lambda(1-\sigma)} \Big(\frac{2\psi}{1-\sigma} \|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*} + \psi^2 \Big) \|\hat{c}^*\|_2 \\ &+ C_1 \sqrt{\frac{\sum_{i=0}^k \|\mathrm{Retr}_{x_i}^{-1}(x_{i+1})\|_{x_i}^2 + \lambda}{(k+1)\lambda}} \Big(\sum_{i=0}^k \sum_{j=0}^i \|\varepsilon_j\|_{x^*} \Big) + \epsilon_1 + \epsilon_2 + \epsilon_3. \end{aligned}$$

Maximizing the bound over $\|\hat{c}^*\|_2$ yields

$$\begin{aligned} d(\bar{x}_{c^*,x}, x^*) \leq\ & S_{k,\bar{\lambda}}^{[0,\sigma]} \sqrt{\frac{\|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*}^2}{a_0^2(1-\sigma)^2} + \frac{C_1^2 \|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*}^4 \big(\frac{2\psi}{1-\sigma} \|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*} + \psi^2 \big)^2}{\lambda^3(1-\sigma)^2}} \\ &+ C_1 \sqrt{\frac{\sum_{i=0}^k \|\mathrm{Retr}_{x_i}^{-1}(x_{i+1})\|_{x_i}^2 + \lambda}{(k+1)\lambda}} \Big(\sum_{i=0}^k \sum_{j=0}^i \|\varepsilon_j\|_{x^*} \Big) + \epsilon_1 + \epsilon_2 + \epsilon_3. \end{aligned}$$

Finally, to see the asymptotic convergence rate, we notice that $\|\mathrm{Retr}_{x^*}^{-1}(x_0)\|_{x^*} = O(d(x_0, x^*))$ and $\lim_{d(x_0,x^*) \to 0} \frac{1}{d(x_0,x^*)} (\epsilon_1 + \epsilon_2 + \epsilon_3) = 0$. □

G EXTENSIONS

In this section, we consider various extensions to the proposed nonlinear acceleration on manifolds.

G.1 Online Riemannian Nonlinear Acceleration

Following Scieur et al. (2018); Bollapragada et al. (2022), we can extend Algorithm 1 to the online setting, where the extrapolated point $\bar{x}_{c,x}$ is used to update the iterate sequence. The idea is to add a mixing step by updating $\bar{x}_{c,x}$ in the direction of the weighted average of the gradients, i.e., $\overline{\mathrm{grad}} f(\bar{x}_{c,x}) = \sum_{i=0}^k c_i \Gamma_{x_i}^{\bar{x}_{c,x}} \mathrm{grad} f(x_i)$. For the averaging schemes (Avg.1), (Avg.3), the next iteration starts with $\mathrm{Exp}_{\bar{x}_{c,x}}(-\delta\, \overline{\mathrm{grad}} f(\bar{x}_{c,x}))$ for some mixing parameter $\delta > 0$. Particularly for the tangent space averaging scheme (Avg.2), we show a more efficient strategy of mixing, which we focus in this paper. The averaging and mixing steps are both performed on the same tangent space. Specifically, let $x_{-1} = x_0$, we define the

Algorithm 8 Riemannian nonlinear acceleration (RiemNA-online)

```
Input: Initialization $x_0$. Regularization parameter $\lambda$. Mixing parameter $\delta$.
for $k = 0, ..., K-1$ do
    Compute $r_i = \Gamma_{x_i}^{x_k} \mathrm{Exp}_{x_i}^{-1}(x_{i+1}) \in T_{x_k}\mathcal{M}, i = 0, ..., k$
    Solve $c^* = \arg\min_{c \in \mathbb{R}^{k+1}: c^\top 1 = 1} \| \sum_{i=0}^k c_i r_i \|_{x_k}^2 + \lambda \|c\|_2^2$.
    Compute $x_{k+1} = \mathrm{Exp}_{x_k}\big( -\delta c_k^* \mathrm{grad} f(x_k) - \sum_{i=0}^{k-1} \Gamma_{x_i}^{x_k} \big( \theta_i^* \mathrm{Exp}_{x_i}^{-1}(x_{i+1}) + \delta c_i^* \mathrm{grad} f(x_i) \big) \big)$, where $\theta_i^* = \sum_{j=0}^i c_j^*$.
end for
Output: $x_K$.
```

Algorithm 9 Adaptive regularized Riemannian nonlinear acceleration (AdaRiemNA)

```
Input: A sequence of iterates $x_0, ..., x_{k+1}$. Tentative regularization parameters $\{\lambda_j\}_{j=1}^k$.
Compute $r_i = \Gamma_{x_i}^{x_k} \mathrm{Exp}_{x_i}^{-1}(x_{i+1}) \in T_{x_k}\mathcal{M}, i = 0, ..., k$
for $j = 1, ..., k$ do
    Solve $c^*(\lambda_j) = \arg\min_{c \in \mathbb{R}^{k+1}: c^\top 1 = 1} \| \sum_{i=0}^k c_i r_i \|_{x_k}^2 + \lambda_j \|c\|_2^2$.
    Compute $\bar{x}(\lambda_j) = \bar{x}_{c,x}$ using $c^*(\lambda_j)$.
end for
Set $\bar{x}^* = \arg\min_{j=1,...,k} f(\bar{x}(\lambda_j))$.
Compute $u = \mathrm{Exp}_{x_0}^{-1}(\bar{x}^*)$ and set $t = 1$.
while $f(\mathrm{Exp}_{x_0}(2tu)) < f(\mathrm{Exp}_{x_0}(tu))$ do
    Update $t = 2t$.
end while
Output: $\mathrm{Exp}_{x_0}(tu)$.
```

following progression of the online nonlinear acceleration on manifolds.

$$
\begin{aligned}
x_{k+1} &= \mathrm{Exp}_{x_k}\Big(- \sum_{i=0}^{k-1} \theta_i \Gamma_{x_i}^{x_k} \mathrm{Exp}_{x_i}^{-1}(x_{i+1}) - \delta \sum_{i=0}^{k} c_i \Gamma_{x_i}^{x_k} \mathrm{grad} f(x_i) \Big) \\
&= \mathrm{Exp}_{x_k}\Big(- \delta c_k \mathrm{grad} f(x_k) - \sum_{i=0}^{k-1} \Gamma_{x_i}^{x_k} \big(\theta_i \mathrm{Exp}_{x_i}^{-1}(x_{i+1}) + \delta c_i \mathrm{grad} f(x_i) \big) \Big).
\end{aligned}
$$

The complete procedures are presented in Algorithm 8.

G.2 Practical Considerations

Here are some practical considerations to use nonlinear acceleration on manifolds.

Iterates From Riemannian Gradient Descent with Line-search Suppose the iterates $\{x_i\}_{i=0}^k$ are generated from $x_i = \mathrm{Exp}_{x_i}(-\eta_i \mathrm{grad} f(x_{i-1}))$ where the stepsize is determined from a line-search procedure (such as backtracking line-search (Boumal et al., 2019)) and thus varies across iterations. Nevertheless, Lemma 3 still holds with $G_i = \mathrm{id} - \eta_i \mathrm{Hess} f(x^*)$. Suppose the stepsize is chosen such that $\|G_i\| \leq \sigma < 1$. Then the analysis still holds under this setting.

Safeguarding Decrease Due to the curved geometry of the manifold and nonlinearity of the objective function, it is not guaranteed that $f(\bar{x}_{c,x})$ will decrease. In the main text, we only show local convergence of the acceleration strategy. A typical globalization technique is to only keep the extrapolated point if it shows sufficient decrease compared to previous iterates, i.e., $f(\bar{x}_{c,x}) \leq \tau \min_{i=0,...,k} f(x_i)$ for some $\tau < 1$. In Scieur et al. (2020), an adaptive regularization strategy has been proposed to select regularization parameter λ. Here we adapt the same strategy on manifolds, which we show in Algorithm 9. As noticed in Scieur et al. (2020), a higher value of λ pushes the weights close to uniform and thus stays closer to x_0. Thus the line-search over t tries to enhance the progress compared to the initialization. In addition, for online Riemannian nonlinear acceleration specifically, we may consider performing a line-search over the parameter δ to ensure a sufficient descent condition is met.

Limited-memory and Extrapolation Frequency Rather than keeping all the previous iterates for extrapolation, we can set a memory depth of m and using only the most recent m iterates to compute the extrapolated point. In practice, m is usually set to be less than 10. In addition, we notice that compared to the Euclidean version, the computational cost for the Riemannian nonlinear acceleration can be high due to the use of parallel transport. Hence to mitigate this issue, we may only compute the extrapolated point every m iteration.

Efficient Update of the Residual Matrix R Recall for each application of Riemannian nonlinear acceleration, we need to compute $R = [\langle r_i, r_j \rangle_{x_k}]_{0 \leq i,j \leq k}$, where $r_i = \mathcal{T}_{x_i}^{x_k} \mathrm{Retr}_{x_i}^{-1}(x_{i+1})$, where we write using (isometric) vector transport and general retraction. This includes parallel transport and exponential map as special cases. By isometry, in the next iteration when we receive r_{k+1}, the update of R only requires computing $\langle \Gamma_{x_k}^{x_{k+1}} r_i, r_{k+1} \rangle_{x_{k+1}}$, $i = 0, ..., k+1$. Denote the vector $r_+ := [\langle \Gamma_{x_k}^{x_{k+1}} r_i, r_{k+1} \rangle_{x_{k+1}}]_{0 \leq i \leq k}$. Then the updated residual matrix is

$$R_+ = \begin{bmatrix} R & r_+ \\ r_+^\top & \|r_{k+1}\|_{x_{k+1}}^2 \end{bmatrix}.$$