Zeroth-order Riemannian Averaging Stochastic Approximation Algorithms

Jiaxiang Li * Krishnakumar Balasubramanian† Shiqian Ma ‡

Abstract

We present Zeroth-order Riemannian Averaging Stochastic Approximation (`Zo-RASA`) algorithms for stochastic optimization on Riemannian manifolds. We show that `Zo-RASA` achieves optimal sample complexities for generating ϵ-approximation first-order stationary solutions using only one-sample or constant-order batches in each iteration. Our approach employs Riemannian moving-average stochastic gradient estimators, and a novel Riemannian-Lyapunov analysis technique for convergence analysis. We improve the algorithm's practicality by using retractions and vector transport, instead of exponential mappings and parallel transports, thereby reducing per-iteration complexity. Additionally, we introduce a novel geometric condition, satisfied by manifolds with bounded second fundamental form, which enables new error bounds for approximating parallel transport with vector transport.

1 Introduction

We consider zeroth-order algorithms for solving the following Riemannian optimization problem,

$$\min_{x\in\mathcal{M}} f(x) := \mathbb{E}_{\xi}[F(x,\xi)], \tag{1.1}$$

where $\mathcal{M}$ is a d-dimensional complete manifold, $f:\mathcal{M}\to\mathbb{R}$ is a smooth function, and we can access only the noisy function evaluations $F(x,\xi)$. A natural zeroth-order algorithm is to estimate the gradients of f and use them in the context of Riemannian stochastic gradient descent. The main difficulty in doing so is the construction of the zeroth-order gradient estimation. Assuming that we have independent samples u_i that are standard normal random vectors supported on $\mathrm{T}_x\,\mathcal{M}$, the tangent space at $x\in\mathcal{M}$, Li et al. (2022) proposed to construct the zeroth-order gradient estimator as

$$G_\mu^{\mathsf{Exp}}(x) = \frac{1}{m}\sum_{i=1}^{m}\frac{F(\mathsf{Exp}_x(\mu u_i),\xi_i)-F(x,\xi_i)}{\mu}u_i \tag{1.2}$$

where $\mu>0$ is a smoothing parameter. Note here that if a retraction is available, then one could also replace the exponential mapping with a retraction based estimator,

$$G_\mu^{\mathsf{Retr}}(x) = \frac{1}{m}\sum_{i=1}^{m}\frac{F(\mathsf{Retr}_x(\mu u_i),\xi_i)-F(x,\xi_i)}{\mu}u_i. \tag{1.3}$$

*Department of Mathematics, University of California, Davis. jxjli@ucdavis.edu
†Department of Statistics, University of California, Davis. kbala@ucdavis.edu
‡Department of Computational Applied Math and Operations Research, Rice University. sqma@rice.edu

Result	Objective	Manifold	Operations	m	N
`Zo-RSGD` (Li et al., 2022, Alg 1)	Smooth, 2MB	general	Retr	$\mathcal{O}(d/\epsilon^2)$	$\Omega(1)$
`Zo-RASA`, Alg 1, Thm 3.1	Smooth, 2MB	General	Exp map, PT	$\mathcal{O}(d)$	$\Omega(1)$
				$\mathcal{O}(1)$	$\Omega(d)$
`Zo-RASA`, Alg 2, Thm 4.2	Smooth, 4MB	Compact, 2nd FF bound	SO-Retr, VT	$\mathcal{O}(d)$	$\Omega(1)$
				$\mathcal{O}(1)$	$\Omega(1)$

Table 1: **Conditions required to establish a sample complexity of $\mathcal{O}(d/\epsilon^4)$ for various algorithms for convergence to stationarity in the sense of Definition 2.2**. For instance, to obtain the $\mathcal{O}(d/\epsilon^4)$ sample complexity for Alg 1, we need to require $m = \mathcal{O}(d)$ and $N = \Omega(1)$, or $m = \mathcal{O}(1)$ and $N = \Omega(d)$. Here, 2MB and 4MB stand for bounded second central moment (i.e., variance) (Assumption 3.2) and fourth central moment (Assumption 4.3) respectively. 2nd FF stands for second fundamental form (Theorem 4.1) (see Section 2 for definition of second fundamental form). SO-RETR stands for second-order retraction (Assumption 4.4). PT and VT stand for parallel and vector transport respectively (see, Definition 2.3). The parameter d is the intrinsic dimension of the manifold $\mathcal{M}$, m is the batch-size, N is the total number of iterations required, and ϵ is the desired precision. Oracle complexity refers to the number of calls to the stochastic zeroth-order oracle. We also remark here that although Li et al. (2022, Algorithm 1) uses retraction, its convergence analysis also assumes retraction-based smoothness. For Zo-RASA, we need the initial batch-size $m_0 = \mathcal{O}(d)$.

The merit of having a Gaussian distribution on the tangent space is that the variance of the constructed estimator $G_\mu(x)$ will only depend on the intrinsic dimension d of the manifold, and is independent of the dimension n of the ambient Euclidean space. We refer to Li et al. (2022) for the details of our zeroth-order estimator and its applications. See also Wang et al. (2021); Wang (2023) for additional follow-up works.

To obtain an ϵ-approximate stationary solution of (1.1) (as in Definition 2.2) using the above approach, Li et al. (2022) established a sample complexity of $\mathcal{O}(d/\epsilon^4)$, with $\mathcal{O}(1/\epsilon^2)$ iteration complexity and $m = \mathcal{O}(d/\epsilon^2)$ per-iteration batch size. Even considering $d = 1$ for simplicity, this suggests for example that to get an accuracy of $\epsilon \approx 10^{-3}$, one needs batch-sizes of order $m \approx 10^6$ resulting in a highly impractical per-iteration complexity. Intriguingly, when implementing these algorithms in practice, favorable results are obtained even when the batch-size is simply set between ten and fifty. Thus, there exists a discrepancy between the current theory and practice of stochastic zeroth-order Riemannian optimization. Furthermore, in online Riemannian optimization problems (Maass et al., 2022; Wang et al., 2023) where the data sequence is observed in a streaming fashion, waiting for very long time-periods in each iteration in order to obtain the required order of batch-sizes is highly undesirable.

The main motivation of the current work stems from the above-mentioned undesirable issues associated with the use of mini-batches in stochastic Riemannian optimization algorithms by Li et al. (2022). We address the problem by getting rid of the use of mini-batches altogether, and by developing batch-free, fully-online algorithm, Zeroth-order Riemannian Averaging Stochastic Approximation (`Zo-RASA`) algorithm, for solving (1.1). We show that to obtain the sample complexity of $\mathcal{O}(d/\epsilon^4)$, `Zo-RASA` only requires $m = 1$ (see the remark after Theorem 3.1), which is a significant improvement compared to Li et al. (2022). The first version of `Zo-RASA` in Algorithm 1 uses exponential mapping and parallel-transports. However, this version is not implementation-friendly. As a case-in-point, consider the Stiefel manifold (see (2.1)) for which we highlight that there is no closed-form expression for the parallel transport $P_{x^k}^{x^{k+1}}$. Indeed, they are only available as solutions to certain ordinary

differential equation, which increases the per-iteration complexity of implementing Algorithm 1. To overcome this issue and to develop a practical version of the RASA framework, we replace the exponential mapping and parallel transport by retraction and vector transport respectively, resulting in the practical version of `Zo-RASA` method in Algorithm 2. As we will discuss in Section 2, in the case of Stiefel manifolds, retractions cost only 1/4 the time of an exponential mapping. Also, while there is no closed-form for parallel transport on Stiefel manifolds, vector transport has an easy closed-form implementation. We establish that Algorithm 2 has the same sample complexity as Algorithm 1, with significantly improved per-iteration complexity. We now highlight two specific novelties that we introduce in this work to establish the above result.

- **Moving-average gradient estimators and Lifting-based Riemannian-Lyapunov analysis.** We introduce a Riemannian moving-average technique (see, Line 4 in Algorithm 1 and Algorithm 2) and a corresponding novel Riemannian-Lyapunov technique for analyzing zeroth-order stochastic Riemannian optimization problems, which works in the lifted space by tracking both the optimization trajectory and the gradient along the trajectory (see (3.4)). For Euclidean problems, these techniques were introduced and extended in Ruszczynski and Syski (1983); Ruszczyński (1987); Ghadimi et al. (2020); Ruszczynski (2021); Balasubramanian et al. (2022). However, those works rely heavily on the Euclidean structure. Non-trivial adaptions are needed to extend such methodology and analyses to the Riemannian settings; see Theorem 3.1 and Theorem 4.2.

- **Approximation error between parallel and vector transports.** A major challenge in analyzing Algorithm 2 is to handle the additional errors introduced by the use of retractions and vector transports. We identify a novel geometric condition on the manifolds under consideration (see Assumption 4.1) under which we provide novel error bounds between parallel and vector transports (see Theorem 4.1). We further show that the proposed condition, which plays a crucial role in our subsequent convergence analysis, is naturally satisfied if the *second fundamental form* of the manifold is bounded. We remark that the obtained error bounds, between parallel and vector transport, are of independent interest and are potentially applicable to a variety of other Riemannian optimization problems.

In Table 1, we summarize the sample complexities of stochastic zeroth-order Riemannian unioptimization algorithms.

1.1 Prior works

We refer to Absil et al. (2008); Boumal (2023) for a discussion on general Riemannian optimization methods. To the best of our knowledge, Li et al. (2022) provided the first oracle complexity results for zeroth-order stochastic Riemannian optimization. Following this, Wang et al. (2021); Wang (2023); Maass et al. (2022) improved and extended the applicability of zeroth-order Riemannian optimization. A central concern in Riemannian optimization is the increased per-iteration complexity caused by the use of exponential mapping and (sometimes) parallel transport. To tackle this, retraction and vector transport are often preferred (Absil et al., 2008; Boumal, 2023). Such replacements have thus far been considered in the deterministic settings, in the context of Riemannian quasi-Newton methods (Huang et al., 2015), Riemannian variance reduction methods (Sato et al., 2019), Riemannian proximal gradient methods (Chen et al., 2020; Huang and Wei, 2022) and Riemannian conjugate gradient methods (Sato, 2022). We discuss precise comparisons to this work later in Section 4.1.1.

Stochastic gradient averaging methods in the Euclidean setting were studied in the several earlier works (Polyak, 1977; Ruszczynski and Syski, 1983; Xiao, 2009). For nonconvex problems, Ghadimi et al. (2020) analyzed the averaging stochastic approximation algorithm and established a sample complexity of $\mathcal{O}(1/\epsilon^4)$ to obtain an ϵ-approximate first-order stationary solution without using mini-batches; see also Ghadimi and Powell (2022) for a zeroth-order extension. For the smooth Riemannian setting, Han and Gao (2020) used a related moving-average technique, and achieve $\mathcal{O}(\epsilon^{-3})$ sample complexity. However, Han and Gao (2020) assumes a Lipschitz smooth-type inequality over $\mathsf{grad}F(x;\xi)$ itself under a given retraction (which is stronger than our assumption) and assume access to the computationally demanding isometric vector transport (see (2.2)). More importantly, they assume an opaque and rather strong condition that all iterates of their algorithm are close to a local optima of the problem to carry out their analysis.

2 Basics of Riemannian optimization

A differentiable manifold $\mathcal{M}$ is a Riemannian manifold if it is equipped with an inner product (called Riemannian metric) on the tangent space, $\langle\cdot,\cdot\rangle_x : \mathrm{T}_x\mathcal{M} \times \mathrm{T}_x\mathcal{M} \to \mathbb{R}$, that varies smoothly on $\mathcal{M}$. The norm of a tangent vector is defined as $\|\xi\|_x := \sqrt{\langle\xi,\xi\rangle_x}$. We drop the subscript x and simply write $\langle\cdot,\cdot\rangle$ (and $\|\xi\|$) if $\mathcal{M}$ is an embedded submanifold with Euclidean metric. Here we use the notion of the tangent space $\mathrm{T}_x\mathcal{M}$ of a differentiable manifold $\mathcal{M}$, whose precise definition can be found in (Tu, 2011, Chapter 8). As an example, consider the Stiefel manifold given by

$$\mathcal{M} = \mathrm{St}(n,p) := \{X \in \mathbb{R}^{n\times p} : X^\top X = I_p\}. \tag{2.1}$$

The tangent space of $\mathrm{St}(n,p)$ is given by $\mathrm{T}_X\mathcal{M} = \{\xi \in \mathbb{R}^{n\times p} : X^\top\xi + \xi^\top X = 0\}$. One could equip the tangent space with common inner product $\langle X, Y\rangle := \mathrm{tr}(X^\top Y)$ to form a Riemannian manifold. For additional examples, see Absil et al. (2008, Chapter 3) or Boumal (2023, Chapter 7) .

We now introduce the concept of a Riemannian gradient and the notion of ϵ-approximate first-order stationary solution for (1.1).

Definition 2.1 (Riemannian Gradient)**.** *Suppose f is a smooth function on Riemannian manifold $\mathcal{M}$. The Riemannian gradient $\mathsf{grad}f(x)$ is a vector in $\mathrm{T}_x\mathcal{M}$ satisfying $\frac{d(f(\gamma(t)))}{dt}\Big|_{t=0} = \langle v, \mathsf{grad}f(x)\rangle_x$ for any $v \in \mathrm{T}_x\mathcal{M}$, where $\gamma(t)$ is a curve satisfying $\gamma(0) = x$ and $\gamma'(0) = v$.*

Definition 2.2 (ϵ-approximate first-order stationary solution for (1.1))**.** *We call a point $\bar{x}$ an ϵ-approximate first-order stationary solution for (1.1) if it satisfies $\mathbb{E}[\|\mathsf{grad}f(\bar{x})\|_{\bar{x}}^2] \le \epsilon^2$, where the expectation is with respect to both the problem and algorithm-based randomness.*

Geodesics, retractions and exponential mappings. Given two tangent vectors $\xi, \eta \in \mathrm{T}\mathcal{M}$, the Levi-Civita connection $\nabla : \mathrm{T}\mathcal{M} \times \mathrm{T}\mathcal{M} \to \mathrm{T}\mathcal{M}$, $(\xi,\eta) \to \nabla_\xi\eta \in \mathrm{T}\mathcal{M}$ is the "directional differential" of η along the direction of ξ, which is determined uniquely by the metric tensor $\langle\cdot,\cdot\rangle_x$. In Euclidean spaces, $\nabla_\xi\eta$ is just calculating the directional derivative of the vector field η along ξ. For a Riemannian manifold $\mathcal{M}$, the geodesic γ is a curve on $\mathcal{M}$ that satisfies $\nabla_{\gamma'}\gamma' = 0$, i.e., the directional derivative along the tangent direction is always zero. Usually we find the geodesic with the initial value condition, $\nabla_{\gamma'}\gamma' = 0,\ \gamma(0) = x,\ \gamma'(0) = v$, whose existence and uniqueness are locally guaranteed by the existence and uniqueness theorem for linear ODEs.

Given any curve $\gamma(t)$ on $\mathcal{M}$, one could calculate the length of the curve and define the distance between the two points $x, y \in \mathcal{M}$ respectively by $L(\gamma) := \int_a^b \|\gamma'(t)\|_{\gamma(t)}dt$ and $\mathsf{d}(x,y) := \min_{\gamma,\gamma(a)=x,\gamma(b)=y} L(\gamma)$. If the manifold is a complete Riemannian manifold, according to (Do Carmo,

1992, Corollary 3.9), there exists a unique minimal geodesic γ satisfying $\gamma(a) = x, \gamma(b) = y$ that minimizes $L(\gamma)$. Therefore, we can always calculate the distance with respect to the minimal geodesic as $\mathsf{d}(x, y) = \int_a^b \|\gamma'(t)\|_{\gamma(t)} dt, \nabla_{\gamma'}\gamma' = 0, \gamma(a) = x, \gamma(b) = y$, which will be utilized in our error analysis in Section 4.

A retraction mapping Retr_x is a smooth mapping from $\mathrm{T}_x\mathcal{M}$ to $\mathcal{M}$ such that: $\mathsf{Retr}_x(0) = x$, where 0 is the zero element of $\mathrm{T}_x\mathcal{M}$, and the differential of Retr_x at 0 is an identity mapping, i.e., $\frac{d\mathsf{Retr}_x(t\eta)}{dt}\Big|_{t=0} = \eta, \forall \eta \in \mathrm{T}_x\mathcal{M}$. In particular, the exponential mapping Exp_x on a Riemannian manifold is a retraction that generated by geodesics, i.e. $\mathsf{Exp}_x(t\xi) \coloneqq \gamma(t)$ where γ is a geodesic with $\gamma(0) = x$ and $\gamma'(0) = \xi$. Notice that the retraction is not always injective from $\mathrm{T}_x\mathcal{M}$ to $\mathcal{M}$ for any point $x \in \mathcal{M}$, thus the existence of the inverse of the retraction function Retr_x^{-1} is not guaranteed. However, when $\mathcal{M}$ is complete, the exponential mapping Exp_x is always defined for every $\xi \in \mathrm{T}_x\mathcal{M}$, and the inverse of the exponential mapping $\mathsf{Exp}_x^{-1}(y) \in \mathrm{T}_x\mathcal{M}$ is always well-defined for any $x, y \in \mathcal{M}$. Also, since $\mathsf{Exp}_x(t\xi)$ generates geodesics, we have $\mathsf{d}(x, \mathsf{Exp}_x(t\xi)) = t\|\xi\|_x$. These are facts that we use in Assumption 3.1 and convergence proofs.

As an example, the retractions on Stiefel manifolds can be defined by the QR decomposition, $R_X(\xi) := Q$ where $X + \xi = QR$. It can also be defined through the Polar decomposition as $R_X(\xi) := UV^\top$, where $X + \xi = U\Sigma V^\top$ is the (thin) singular value decomposition of $X + \xi$. The geodesic on the Stiefel manifold is given by: $X(t) = \begin{bmatrix} X(0) & \dot{X}(0) \end{bmatrix} \exp\left(t \begin{bmatrix} A(0) & -S(0) \\ I & A(0) \end{bmatrix}\right) \begin{bmatrix} I \\ 0 \end{bmatrix} \exp(-A(0)t)$, for $A(t) = X^\top(t)\dot{X}(t)$ and $S(t) = \dot{X}^\top(t)\dot{X}(t)$ with initial point $X(0)$ and initial speed $\dot{X}(0)$. The exponential mapping is thus given by $\mathsf{Exp}_{X(0)}(\dot{X}(0)) = X(1)$. The computation cost of the QR and Polar decomposition retractions are of order $2dk^2 + \mathcal{O}(k^3)$ and $3dk^2 + \mathcal{O}(k^3)$, whereas as shown by Chen et al. (2020, Section 3) the exponential mapping takes $8dk^2 + \mathcal{O}(k^3)$, which illustrates the favorability of retractions in practical computations. We refer to Absil et al. (2008, Chapter 4) and Boumal (2023, Chapter 3) for additional examples and more discussions on retractions and exponential mappings.

Vector and parallel transport. Vector transports are linear mappings from one tangent space to another, which can be formally defined below.

Definition 2.3 (Vector and parallel transport)**.** *A vector transport $\mathcal{T}$ on a smooth manifold $\mathcal{M}$ is a smooth mapping $\mathrm{T}\mathcal{M} \times \mathrm{T}\mathcal{M} \to \mathrm{T}\mathcal{M} : (\eta_x, \xi_x) \to \mathcal{T}_{\eta_x}(\xi_x) \in \mathrm{T}\mathcal{M}$, where the subscript x means that the vector is in $\mathrm{T}_x\mathcal{M}$, such that: (i) There exists a retraction R so that $\mathcal{T}_{\eta_x}(\xi_x) \in \mathrm{T}_{R_x(\eta_x)}\mathcal{M}$, (ii) $\mathcal{T}_{0_x}\xi_x = \xi_x$ for all $\xi_x \in \mathrm{T}_x\mathcal{M}$, and (iii) $\mathcal{T}_{\eta_x}(a\xi_x + b\zeta_x) = a\mathcal{T}_{\eta_x}(\xi_x) + b\mathcal{T}_{\eta_x}(\zeta_x)$, i.e., linearity. Particularly, for a complete Riemannian manifold $(\mathcal{M}, \langle\cdot,\cdot\rangle)$, we can construct a special vector transport, namely the parallel transport P, that can map vectors to another tangent space "parallelly", i.e., $\forall \eta, \xi \in \mathrm{T}_x\mathcal{M}$ and $y \in \mathcal{M}$,*

$$\langle P_{\mathsf{Exp}_x^{-1}(y)}(\eta), P_{\mathsf{Exp}_x^{-1}(y)}(\xi)\rangle_y = \langle \eta, \xi\rangle_x. \tag{2.2}$$

Notice that parallel transport is not the only transport that satisfies (2.2), and we call the vector transport an isometric vector transport if it satisfies (2.2).

We can equivalently view P as a mapping from the tangent space $\mathrm{T}_x\mathcal{M}$ to $\mathrm{T}_y\mathcal{M}$. We hence denote $P_x^y : \mathrm{T}_x\mathcal{M} \to \mathrm{T}_y\mathcal{M}$. Note that parallel transport depends on the curve along which the vectors are moving. If the curve is not specified, it refers to the case when we are considering the minimal geodesic connecting the two points, which exists due to completeness.

As an example, for the Stiefel manifold in (2.1), there is no closed-form expression for the parallel transport, whereas one can always utilize the projection onto the tangent space, given

Algorithm 1: Zo-RASA

Input: Initial point $x^0 \in \mathcal{M}$, $g^0 = G_\mu^{\mathsf{Exp}}(x^0)$, total number of iterations N, parameters $\beta > 0$, $\tau_0 = 1$, $\tau_k = 1/\sqrt{N}$ or $\tau_k = 1/\sqrt{dN}$ when $k \geq 1$, and stepsize $t_k = \tau_k/\beta$.
for $k = 0, 1, 2, \ldots, N-1$ **do**
  $x^{k+1} \leftarrow \mathsf{Exp}_{x^k}(-t_k g^k)$
  $g^{k+1} \leftarrow (1-\tau_k) P_{x^k}^{x^{k+1}} g^k + \tau_k P_{x^k}^{x^{k+1}} G_\mu^k$ where $G_\mu^k = G_\mu^{\mathsf{Exp}}(x^k)$ is given by (1.2) with batch-size $m = m_k$
end for

by $\mathsf{proj}_{\mathrm{T}_X\mathcal{M}}(\xi) = (I - XX^\top)\xi + X\,\mathrm{skew}(X^\top\xi)$, where $\mathrm{skew}(A) := (A - A^\top)/2$, to transport $\xi \in \mathrm{T}_{X_0}\mathrm{St}(d,p)$ to $\mathrm{T}_X\mathrm{St}(d,p)$. We refer to Absil et al. (2008, Chapter 8) and Boumal (2023, Chapter 10) for additional examples and more discussions on vector and parallel transports.

Second fundamental form. We now discuss the notion of second fundamental form, which will be helpful in characterizing a geometric condition used in Section 4 to quantify the error of approximating parallel transports with vector transports. In general, the notion of second fundamental form can be studied for general isometric immersions and we restrict here to the embedding in Euclidean spaces only for brevity.

Definition 2.4 (Second fundamental form)**.** *Suppose $\mathcal{M} \subset \mathbb{R}^D$ is a complete Riemannian manifold equipped with the Euclidean metric. For any $\xi, \eta \in \mathrm{T}\mathcal{M}$, denote the extension of two vector fields to $\mathbb{R}^D$ as $\bar{\xi}, \bar{\eta} \in \mathbb{R}^D$, also the directional derivative of $\bar{\eta}$ along $\bar{\xi}$ as $\bar{\nabla}_{\bar{\xi}}\bar{\eta} \in \mathbb{R}^D$. The second fundamental form refers to the bilinear and symmetric vector, $B(\xi,\eta) = \bar{\nabla}_{\bar{\xi}}\bar{\eta} - \nabla_\xi\eta \in (\mathrm{T}\mathcal{M})^\perp$, which quantifies the deviation of the Riemannian directional derivatives (depicted by Levi-Civita connection ∇) from the Euclidean one (common directional derivative $\bar{\nabla}$).*

Finally, we remark that there are various definitions of second fundamental forms, among which the most common one is a quadratic form related to B; see (Do Carmo, 1992, Chapter 6, Definition 2.2). Here we simply refer to B as the second fundamental form.

3 Zeroth-order RASA for smooth manifold optimization

We now introduce the Zeroth-order Riemannian Average Stochastic Approximation (`Zo-RASA`) algorithm for solving (1.1). The formal procedure is stated in Algorithm 1, where P_x^y is the parallel transport from $\mathrm{T}_x\mathcal{M}$ to $\mathrm{T}_y\mathcal{M}$ along the minimum geodesic connecting x and y. To establish the sample complexity of Algorithm 1, we extend the analysis of Ghadimi et al. (2020), which is in-turn motivated by the lifting-technique introduced in Ruszczynski and Syski (1983); Ruszczyński (1987), to the Riemannian setting. As such works heavily rely on the Euclidean structure, our proofs involve a non-trivial adaption of such techniques.

In our convergence analysis, we always choose $\tau_0 = 1$, and we consider two choices of τ_k when $k \geq 1$:

$$\tau_k = 1/\sqrt{N} \text{ or } \tau_k = 1/\sqrt{dN}, k \geq 1, \tag{3.1}$$

which corresponds to large or single batch, respectively. Moreover, we always choose $t_k = \tau_k/\beta$, where β is a positive constant determined by the smoothness constant in Assumption 3.1 (see Theorem 3.1), so that the step-size and the averaging weights are in the same order. Furthermore,

we define

$$\Gamma_0 = \Gamma_1 = 1, \text{ and } \Gamma_k = \Gamma_1 \prod_{i=1}^{k-1}(1-\tau_i^2). \tag{3.2}$$

This leads to the following inequalities which will be used frequently in our convergence analysis:

$$\sum_{i=k+1}^{N} \tau_i \Gamma_i \le \Gamma_{k+1} \text{ and } \sum_{i=k+1}^{N} \tau_i^2 \Gamma_i \le \tau_k \Gamma_{k+1}. \tag{3.3}$$

To proceed, we construct the following potential function

$$W(x,g) \coloneqq (f(x) - f^*) - \eta(x,g), \quad \text{where} \quad \eta(x,g) \coloneqq -\frac{1}{2\beta}\|g\|_x^2, \ g \in \mathrm{T}_x\mathcal{M}, \tag{3.4}$$

where $f^* = \min_{x\in\mathcal{M}} f(x)$ and $\beta > 0$ is a constant to be determined later. Note that the potential function in (3.4) has the component of both function value and the norm of the (estimated) gradients, also that W is always non-negative. In our analysis, we proceed by bounding the difference of potential function between successive iterates. More specifically, using the convexity of the norm, for any pair (x,g), we have $\|\mathsf{grad} f(x)\|_x^2 \le -2\beta\,\eta(x,g) + 2\|g - \mathsf{grad} f(x)\|_x^2$. This observation will be leveraged in the proof of Theorem 3.1 to obtain the sample complexity of Algorithm 1 for obtaining an ϵ-approximate stationary solution.

We also highlight that our convergence analysis extensively utilizes the isometry property of parallel transport, stated in (2.2), i.e., $\langle P_x^y(\eta), P_x^y(\xi)\rangle_y = \langle \eta, \xi\rangle_x$. This result is a generalization of the isometry in the Euclidean spaces, since the inner product in Euclidean spaces is unchanged if one moves the beginning point of the vectors together. A direct result of this identity is that the length of the vectors is unchanged, namely $\|P_x^y(\xi)\|_y = \|\xi\|_x$, which we will also use extensively.

We now introduce the assumptions needed for our analysis.

Assumption 3.1. *The function $f : \mathcal{M} \to \mathbb{R}$ is L-smooth on $\mathcal{M}$, i.e., $\forall x, y \in \mathcal{M}$, we have $\|P_x^y \mathsf{grad} f(x) - \mathsf{grad} f(y)\|_y \le L\,\mathsf{d}(x,y)$. An immediate consequence (see, for example, Boumal (2023, Proposition 10.53)) of this condition is that we have $|f(y) - f(x) - \langle \mathsf{grad} f(x), \mathsf{Exp}_x^{-1}(y)\rangle_x| \le \frac{L}{2}\|\mathsf{Exp}_x^{-1}(y)\|_x^2$.*

Assumption 3.1 is a generalization of the standard gradient-Lipschitz assumption in Euclidean optimization (Nesterov, 2018; Lan, 2020) to the Riemannian setting, and is made in several works (Boumal, 2023). To generalize it to the Riemannian setting, due to the fact that $\mathsf{grad} f(x)$ and $\mathsf{grad} f(y)$ are not in the same tangent space, we need to utilize parallel transports P_x^y to match the two vectors in the same tangent space.

Throughout the paper, we define $\mathcal{F}_k$ as the σ-algebra generated by all the randomness till iteration k of the algorithms. Namely, for Algorithm 1, we have $\mathcal{F}_k = \sigma(\xi_0, \ldots, \xi_k, x_0, \ldots, x_k, g_0, \ldots, g_k)$.

Assumption 3.2. *Along the trajectory of the algorithm, the stochastic gradients are unbiased and have bounded-variance, i.e., for $k \in \{1, \ldots, N\}$, we have $\mathbb{E}_\xi[\mathsf{grad} F(x^k; \xi_k)|\mathcal{F}_{k-1}] = \mathsf{grad} f(x^k)$ and $\mathbb{E}_\xi[\|\mathsf{grad} F(x^k;\xi_k) - \mathsf{grad} f(x^k)\|_{x^k}^2 |\mathcal{F}_{k-1}] \le \sigma^2$.*

The above assumption is widely used in stochastic Riemannian optimization literature; see, for example, Zhang et al. (2016); Li et al. (2022); Boumal (2023), and generalizes the standard assumption used in Euclidean stochastic optimization (Nesterov, 2018; Lan, 2020).

Now we proceed to the convergence analysis of Algorithm 1. We first state the following standard result characterizing the approximation error of G_μ^{Exp} (given by (1.2)) to the true Riemannian gradient.

Lemma 3.1 (Proposition 1 in Li et al. (2022) with exponential mapping). *Under Assumptions 3.1, 3.2 we have* $\|\mathbb{E}G_\mu^{\mathsf{Exp}}(x) - \mathsf{grad}f(x)\|_x^2 \le \frac{\mu^2L^2}{4}(d+3)^3$, $\mathbb{E}\|G_\mu^{\mathsf{Exp}}(x)\|_x^2 \le \mu^2L^2(d+6)^3 + 2(d+4)\|\mathsf{grad}f(x)\|_x^2$ *and* $\mathbb{E}\|G_\mu^{\mathsf{Exp}}(x) - \mathsf{grad}f(x)\|_x^2 \le \mu^2L^2(d+6)^3 + \frac{8(d+4)}{m}\sigma^2 + \frac{8(d+4)}{m}\|\mathsf{grad}f(x)\|_x^2$, *where the expectation is taken toward all the Gaussian vectors in* G_μ *and the random variable* ξ.

Based on the above result, we have the following Lemma 3.2 which bounds the difference of g^k to the true Riemannian gradient $\mathsf{grad}f(x^k)$, and Lemma 3.3 bounds the difference of two consecutive g^k, where we use parallel transport to make g^k and g^{k+1} in the same tangent space, i.e., $\|P_{x^{k+1}}^{x^k}g^{k+1} - g^k\|_{x^k}^2$.

Lemma 3.2. *Suppose the Assumptions 3.1 and 3.2 hold, and* $\{x^k, g^k\}$ *is generated by Algorithm 1. We have*

$$\begin{aligned}&\mathbb{E}\|g^k - \mathsf{grad}f(x^k)\|_{x^k}^2 \\ \le&\Gamma_k\tilde{\sigma}_0^2 + \Gamma_k\sum_{i=1}^k\Big(\frac{(1+\tau_{i-1})\tau_{i-1}}{\Gamma_i}\frac{L^2\|g^{i-1}\|_{x^{i-1}}^2}{\beta^2} + \frac{\tau_{i-1}^2}{\Gamma_i}\tilde{\sigma}_{i-1}^2 + \tau_k\hat{\sigma}^2\Big),\end{aligned} \tag{3.5}$$

where the expectation $\mathbb{E}$ *is taken with respect to all random variables up to iteration* k*, including the random variables* $\{u_i\}_{i=1}^k$ *used to construct the zeroth-order estimator as in (1.2). Here the notations are defined as:*

$$\begin{aligned}&\hat{\sigma}^2 := \frac{\mu^2L^2}{4}(d+3)^3 \\ &\tilde{\sigma}_k^2 := \sigma_k^2 + \frac{8(d+4)}{m_k}\mathbb{E}\|\mathsf{grad}f(x^k)\|_{x^k}^2 \text{ where } \sigma_k^2 := \mu^2L^2(d+6)^3 + \frac{8(d+4)}{m_k}\sigma^2.\end{aligned} \tag{3.6}$$

Moreover, from (3.3) we have

$$\begin{aligned}\sum_{k=1}^N\tau_k\mathbb{E}\|g^k - \mathsf{grad}f(x^k)\|_{x^k}^2 &\le \sum_{k=0}^{N-1}\left((1+\tau_k)\tau_k\frac{L^2\mathbb{E}\|g^k\|_{x^k}^2}{\beta^2} + \tau_k^2\tilde{\sigma}_k^2 + \tau_k\hat{\sigma}^2\right) + \tilde{\sigma}_0^2, \\ \sum_{k=1}^N\tau_k^2\mathbb{E}\|g^k - \mathsf{grad}f(x^k)\|_{x^k}^2 &\le \sum_{k=0}^{N-1}\left((1+\tau_k)\tau_k^2\frac{L^2\mathbb{E}\|g^k\|_{x^k}^2}{\beta^2} + \tau_k^3\tilde{\sigma}_k^2 + \tau_k^2\hat{\sigma}^2\right) + \sum_{k=1}^N\tau_k^2\tilde{\sigma}_0^2.\end{aligned}$$

Proof. Firstly, note that we have the following: $g^k - \mathsf{grad}f(x^k) = (1-\tau_{k-1})P_{x^{k-1}}^{x^k}g^{k-1} + \tau_{k-1}P_{x^{k-1}}^{x^k}G_\mu^{k-1} - \mathsf{grad}f(x^k) = (1-\tau_{k-1})P_{x^{k-1}}^{x^k}(g^{k-1} - \mathsf{grad}f(x^{k-1})) + (P_{x^{k-1}}^{x^k}\mathsf{grad}f(x^{k-1}) - \mathsf{grad}f(x^k)) + \tau_{k-1}P_{x^{k-1}}^{x^k}(G_\mu^{k-1} - \mathsf{grad}f(x^{k-1})) = (1-\tau_{k-1})P_{x^{k-1}}^{x^k}(g^{k-1} - \mathsf{grad}f(x^{k-1})) + \tau_{k-1}e_{k-1} + \tau_{k-1}\Delta_{k-1}^f$. Hence, we have

$$\begin{aligned}&\|g^k - \mathsf{grad}f(x^k)\|_{x^k}^2 \\ \le&(1-\tau_{k-1})\|g^{k-1} - \mathsf{grad}f(x^{k-1})\|_{x^{k-1}}^2 + \tau_{k-1}\|e_{k-1}\|_{x^k}^2 + \tau_{k-1}^2\|\Delta_{k-1}^f\|_{x^k}^2 \\ &+ 2\tau_{k-1}\langle(1-\tau_{k-1})P_{x^{k-1}}^{x^k}(g^{k-1} - \mathsf{grad}f(x^{k-1})) + \tau_{k-1}e_{k-1}, \Delta_{k-1}^f\rangle_{x^k},\end{aligned} \tag{3.7}$$

where the notation is defined as $e_{k-1} := \frac{1}{\tau_{k-1}}(P_{x^{k-1}}^{x^k}\mathsf{grad}f(x^{k-1}) - \mathsf{grad}f(x^k))$, and $\Delta_{k-1}^f := P_{x^{k-1}}^{x^k}(G_\mu^{k-1} - \mathsf{grad}f(x^{k-1}))$. Denote $\delta_{k-1} = \langle(1-\tau_{k-1})P_{x^{k-1}}^{x^k}(g^{k-1} - \mathsf{grad}f(x^{k-1})) + \tau_{k-1}e_{k-1}, \Delta_{k-1}^f\rangle_{x^k}$. The main novelty in the proof of this lemma is that δ is no longer an unbiased estimator (which is true for the first-order situation). We have by Lemma 3.1 that

$$\begin{aligned}&2\mathbb{E}_{u^k}[\delta_{k-1}] = 2\langle(1-\tau_{k-1})P_{x^{k-1}}^{x^k}(g^{k-1} - \mathsf{grad}f(x^{k-1})) + \tau_{k-1}e_{k-1}, \mathbb{E}_{u^k}[\Delta_{k-1}^f|\mathcal{F}_{k-2}]\rangle_{x^k} \\ \le&\|(1-\tau_{k-1})P_{x^{k-1}}^{x^k}(g^{k-1} - \mathsf{grad}f(x^{k-1})) + \tau_{k-1}e_{k-1}\|_{x^k}^2 + \|\mathbb{E}_{u^k}G_\mu^{k-1} - \mathsf{grad}f(x^{k-1})\|_{x^{k-1}}^2 \\ \le&(1-\tau_{k-1})\|g^{k-1} - \mathsf{grad}f(x^{k-1})\|_{x^{k-1}}^2 + \tau_{k-1}\|e_{k-1}\|_{x^k}^2 + \hat{\sigma}^2.\end{aligned}$$

Notice that in the above computation, the expectation is only taken with respect to the Gaussian random variables that we used to construct $G_\mu(x^{k-1})$. Plugging this back to (3.7), we have $\mathbb{E}_{u^k}\|g^k - \mathsf{grad} f(x^k)\|_{x^k}^2 \leq \tau_{k-1}\hat{\sigma}^2 + (1-\tau_{k-1}^2)\|g^{k-1} - \mathsf{grad} f(x^{k-1})\|_{x^{k-1}}^2 + \tau_{k-1}(1+\tau_{k-1})\|e_{k-1}\|_{x^k}^2 + \tau_{k-1}^2\|\Delta_{k-1}^f\|_{x^k}^2$. Now dividing both sides of this inequality by our new definition of Γ_k, we get $\frac{1}{\Gamma_k}\mathbb{E}_{u^k}\|g^k - \mathsf{grad} f(x^k)\|_{x^k}^2 \leq \frac{1}{\Gamma_{k-1}}\|g^{k-1} - \mathsf{grad} f(x^{k-1})\|_{x^{k-1}}^2 + \frac{(1+\tau_{k-1})\tau_{k-1}}{\Gamma_k}\|e_{k-1}\|_{x^k}^2 + \frac{\tau_{k-1}^2}{\Gamma_k}\|\Delta_{k-1}^f\|_{x^k}^2 + \frac{\tau_{k-1}}{\Gamma_k}\hat{\sigma}^2$.

By Assumptions 3.1, 3.2 and Lemma 3.1, we have that $\|e_i\|_{x^{i+1}}^2 \leq \frac{L^2}{\tau_i^2}\mathsf{d}(x^i, x^{i+1})^2 \leq \frac{L^2 t_i^2 \|g^i\|_{x^i}^2}{\tau_i^2} = \frac{L^2\|g^i\|_{x^i}^2}{\beta^2}$, and $\mathbb{E}[\|\Delta_i^f\|_{x^{i+1}}^2|\mathcal{F}_{i-1}] \leq \sigma_i^2 + \frac{8(d+4)}{m_i}\mathbb{E}[\|\mathsf{grad} f(x^i)\|_{x^i}^2|\mathcal{F}_{i-1}]$. Hence, by applying law of total expectation (to take the expectation over all random variables), we have $\frac{1}{\Gamma_k}\mathbb{E}\|g^k - \mathsf{grad} f(x^k)\|_{x^k}^2 \leq \frac{1}{\Gamma_{k-1}}\mathbb{E}\|g^{k-1} - \mathsf{grad} f(x^{k-1})\|_{x^{k-1}}^2 + \frac{(1+\tau_{k-1})\tau_{k-1}}{\Gamma_k}\frac{L^2\mathbb{E}\|g^{k-1}\|_{x^{k-1}}^2}{\beta^2} + \frac{\tau_{k-1}^2}{\Gamma_k}\tilde{\sigma}_{k-1}^2 + \frac{\tau_{k-1}}{\Gamma_k}\hat{\sigma}^2$. Now by telescoping the sum in the above equation, we get (note that we take $g^0 = G_\mu(x^0)$)

$$
\begin{aligned}
&\mathbb{E}\|g^k - \mathsf{grad} f(x^k)\|_{x^k}^2 \leq \Gamma_k\mathbb{E}\|G_\mu(x^0) - \mathsf{grad} f(x^0)\|_{x^0}^2 \\
&+ \Gamma_k\sum_{i=1}^k\left(\frac{(1+\tau_{i-1})\tau_{i-1}}{\Gamma_i}\frac{L^2\mathbb{E}\|g^{i-1}\|_{x^{i-1}}^2}{\beta^2} + \frac{\tau_{i-1}^2}{\Gamma_i}\tilde{\sigma}_{i-1}^2 + \frac{\tau_{i-1}}{\Gamma_i}\hat{\sigma}^2\right) \\
&\leq \Gamma_k\tilde{\sigma}_0^2 + \Gamma_k\sum_{i=1}^k\left(\frac{(1+\tau_{i-1})\tau_{i-1}}{\Gamma_i}\frac{L^2\mathbb{E}\|g^{i-1}\|_{x^{i-1}}^2}{\beta^2} + \frac{\tau_{i-1}^2}{\Gamma_i}\tilde{\sigma}_{i-1}^2 + \frac{\tau_{i-1}}{\Gamma_i}\hat{\sigma}^2\right).
\end{aligned}
$$

This proves (3.5). From (3.3) we have

$$
\begin{aligned}
&\sum_{k=1}^N \tau_k\mathbb{E}\|g^k - \mathsf{grad} f(x^k)\|_{x^k}^2 \\
&\leq \sum_{k=1}^N \tau_k\Gamma_k\sum_{i=1}^k\left(\frac{(1+\tau_{i-1})\tau_{i-1}}{\Gamma_i}\frac{L^2\mathbb{E}\|g^{i-1}\|_{x^{i-1}}^2}{\beta^2} + \frac{\tau_{i-1}^2}{\Gamma_i}\tilde{\sigma}_{i-1}^2 + \frac{\tau_{i-1}}{\Gamma_i}\hat{\sigma}^2\right) + \tilde{\sigma}_0^2 \\
&= \sum_{k=0}^{N-1}\left(\sum_{i=k+1}^N \tau_i\Gamma_i\right)\frac{1}{\Gamma_{k+1}}\left((1+\tau_k)\tau_k\frac{L^2\mathbb{E}\|g^k\|_{x^k}^2}{\beta^2} + \tau_k^2\tilde{\sigma}_k^2 + \tau_k\hat{\sigma}^2\right) + \tilde{\sigma}_0^2 \\
&\leq \sum_{k=0}^{N-1}\left((1+\tau_k)\tau_k\frac{L^2\mathbb{E}\|g^k\|_{x^k}^2}{\beta^2} + \tau_k^2\tilde{\sigma}_k^2 + \tau_k\hat{\sigma}^2\right) + \tilde{\sigma}_0^2,
\end{aligned}
$$

where we used $\sum_{k=1}^N \tau_k\Gamma_k \leq \Gamma_1 = 1$ due to (3.3), so that the last term is simply $\tilde{\sigma}_0^2$.

By using similar calculations, we have that

$$
\begin{aligned}
&\sum_{k=1}^N \tau_k^2\mathbb{E}\|g^k - \mathsf{grad} f(x^k)\|_{x^k}^2 \leq \\
&\sum_{k=1}^N \tau_k^2\Gamma_k\sum_{i=1}^k\left(\frac{(1+\tau_{i-1})\tau_{i-1}}{\Gamma_i}\frac{L^2\mathbb{E}\|g^{i-1}\|_{x^{i-1}}^2}{\beta^2} + \frac{\tau_{i-1}^2}{\Gamma_i}\tilde{\sigma}_{i-1}^2 + \frac{\tau_{i-1}}{\Gamma_i}\hat{\sigma}^2\right) + \sum_{k=1}^N \tau_k^2\tilde{\sigma}_0^2 \\
&= \sum_{k=0}^{N-1}\left(\sum_{i=k+1}^N \tau_i^2\Gamma_i\right)\frac{1}{\Gamma_{k+1}}\left((1+\tau_k)\tau_k\frac{L^2\mathbb{E}\|g^k\|_{x^k}^2}{\beta^2} + \tau_k^2\tilde{\sigma}_k^2 + \tau_k\hat{\sigma}^2\right) + \sum_{k=1}^N \tau_k^2\tilde{\sigma}_0^2
\end{aligned}
$$

$$\leq \sum_{k=0}^{N-1} \tau_k \left((1+\tau_k)\tau_k \frac{L^2 \mathbb{E}\|g^k\|_{x^k}^2}{\beta^2} + \tau_k^2 \tilde{\sigma}_k^2 + \tau_k \hat{\sigma}^2 \right) + \sum_{k=1}^{N} \tau_k^2 \tilde{\sigma}_0^2,$$

which completes the proof. □

Lemma 3.3. *Suppose Assumptions 3.1 and 3.2 hold. We have*

$$\begin{aligned} \sum_{k=1}^{N} \mathbb{E}\|P_{x^{k+1}}^{x^k} g^{k+1} - g^k\|_{x^k}^2 &\leq 2\sum_{k=0}^{N} \tau_k^2 \hat{\sigma}^2 + 2\sum_{k=0}^{N} \left(\tau_k^2 + \tau_k^3\right) \sigma_k^2 + 2\sum_{k=0}^{N} \tau_k^2 \tilde{\sigma}_0^2 \\ &+ 2\sum_{k=0}^{N} (1+\tau_k)\tau_k^2 \frac{L^2 \mathbb{E}\|g^k\|_{x^k}^2}{\beta^2} + 2\sum_{k=0}^{N} \left(\tau_k^2 + \tau_k^3\right) \frac{8(d+4)}{m_k} \mathbb{E}\|\mathsf{grad} f(x^k)\|_{x^k}^2 \end{aligned} \tag{3.8}$$

where the expectation $\mathbb{E}$ is taken with respect to all random variables up to iteration k, which includes the Gaussian variables u in the zeroth-order estimator as in (1.2).

Proof. First note that $\|P_{x^{k+1}}^{x^k} g^{k+1} - g^k\|_{x^k}^2 = \tau_k^2 \|G_\mu^k - g^k\|_{x^k}^2 = \tau_k^2 \|G_\mu^k - \mathsf{grad} f(x^k) + \mathsf{grad} f(x^k) - g^k\|_{x^k}^2 \leq 2\tau_k^2 \|G_\mu^k - \mathsf{grad} f(x^k)\|_{x^k}^2 + 2\tau_k^2 \|\mathsf{grad} f(x^k) - g^k\|_{x^k}^2$. Taking the expectation conditioned on $\mathcal{F}_{k-1}$, we get

$$\begin{aligned} &\frac{1}{2}\mathbb{E}[\|P_{x^{k+1}}^{x^k} g^{k+1} - g^k\|_{x^k}^2 | \mathcal{F}_{k-1}] \\ \leq &\tau_k^2 \mathbb{E}[\|G_\mu^k - \mathsf{grad} f(x^k)\|_{x^k}^2 | \mathcal{F}_{k-1}] + \tau_k^2 \mathbb{E}[\|\mathsf{grad} f(x^k) - g^k\|_{x^k}^2 | \mathcal{F}_{k-1}] \\ \leq &\tau_k^2 \left(\sigma_k^2 + \frac{8(d+4)}{m_k} \mathbb{E}[\|\mathsf{grad} f(x^k)\|_{x^k}^2 | \mathcal{F}_{k-1}] \right) + \tau_k^2 \mathbb{E}[\|\mathsf{grad} f(x^k) - g^k\|_{x^k}^2 | \mathcal{F}_{k-1}], \end{aligned}$$

where last inequality is by Lemma 3.1. Now using law of total expectation to take the expectation for all random variables and summing up over $k = 0, ..., N-1$, we have

$$\begin{aligned} &\frac{1}{2}\sum_{k=1}^{N} \mathbb{E}\|P_{x^{k+1}}^{x^k} g^{k+1} - g^k\|_{x^k}^2 \\ \leq &\sum_{k=1}^{N} \tau_k^2 \sigma_k^2 + \sum_{k=1}^{N} \tau_k^2 \frac{8(d+4)}{m_k} \mathbb{E}\|\mathsf{grad} f(x^k)\|_{x^k}^2 + \sum_{k=1}^{N} \tau_k^2 \mathbb{E}\|\mathsf{grad} f(x^k) - g^k\|_{x^k}^2 \\ \leq &\sum_{k=0}^{N} \tau_k^2 \hat{\sigma}^2 + \sum_{k=0}^{N} \left(\tau_k^2 + \tau_k^3\right) \sigma_k^2 + \sum_{k=0}^{N} \tau_k^2 \tilde{\sigma}_0^2 \\ &+ \sum_{k=0}^{N} (1+\tau_k)\tau_k^2 \frac{L^2 \mathbb{E}\|g^k\|_{x^k}^2}{\beta^2} + \sum_{k=0}^{N} \left(\tau_k^2 + \tau_k^3\right) \frac{8(d+4)}{m_k} \mathbb{E}\|\mathsf{grad} f(x^k)\|_{x^k}^2, \end{aligned}$$

where the second inequality is by Lemma 3.2. □

Now we are ready to present our main result.

Theorem 3.1. *Suppose Assumptions 3.1 and 3.2 hold. In Algorithm 1, we set $\mu = \mathcal{O}\left(\frac{1}{Ld^{3/2}N^{1/4}}\right)$, and $\beta \geq 4L$. Then the following holds.*

(i) If we choose $\tau_0 = 1$, $\tau_k = 1/\sqrt{N}$, $k \geq 1$ *and* $m_k \equiv 8(d+4)$, $k \geq 0$, *then we have* $\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}\|\mathsf{grad} f(x^k)\|_{x^k}^2 \leq \mathcal{O}(1/\sqrt{N})$.

(ii) If we choose $\tau_0 = 1$, $\tau_k = 1/\sqrt{dN}$, $k \geq 1$, $m_0 = d$ *and* $m_k = 1$ *for* $k \geq 1$, *then we have* $\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}\|\mathsf{grad} f(x^k)\|_{x^k}^2 \leq \mathcal{O}(\sqrt{d/N})$, *for all* $N = \Omega(d)$.

Here the expectation $\mathbb{E}$ *is taken with respect to all random variables up to iteration* k, *which includes the random variables* u *in zeroth-order estimator* (1.2).

Before we proceed to the proof of Theorem 3.1, we have the following Lemma 3.4 which will be utilized in the proof.

Lemma 3.4. *Suppose we take parameters the same as Theorem 3.1, then we have*

$$\frac{\tau_k}{2\beta} - \frac{\tau_k^2 L}{2\beta^2} - \frac{(1+\tau_k)\tau_k^2}{\beta}\frac{L^2}{\beta^2} \geq \frac{\tau_k}{4\beta}, \tag{3.9a}$$

$$\frac{\tau_k}{2} - \left(4\left(\frac{2L^2}{\beta^2}+1\right)(1+\tau_k)+1\right)\frac{8(d+4)}{m_k}\tau_k^2 \geq \frac{\tau_k}{4}. \tag{3.9b}$$

Proof. To show (3.9a), using $\beta \geq 4L$, we just need to show that $\tau_k/8 + (1+\tau_k)\tau_k/16 \leq 1/4$, which holds naturally in both cases (i) and (ii).

As for (3.9b), again by $\beta \geq 4L$ we just need to show that $\big(4(1/8+1)(1+\tau_k)+1\big)(8(d+4)/m_k)\tau_k \leq 1/4$. In case (i), this is equivalent to $18\tau_k^2 + 22\tau_k - 1 \leq 0$, which is guaranteed when $N \geq 520$. For case (ii), similar calculation shows that we need $\tau_k \leq (\sqrt{22^2 + 9/(d+4)} - 22)/36$, which is guaranteed when $N \geq 3.2 \cdot 10^4 \cdot (d+4)^2/d$. □

Proof. [Proof of Theorem 3.1] By the isometry property of parallel transport,

$$\begin{aligned}
&\eta(x^k, g^k) - \eta(x^{k+1}, g^{k+1}) = \frac{1}{2\beta}\|g^{k+1}\|_{x^{k+1}}^2 - \frac{1}{2\beta}\|g^k\|_{x^k}^2 \\
=&\frac{1}{2\beta}\|P_{x^{k+1}}^{x^k} g^{k+1}\|_{x^k}^2 - \frac{1}{2\beta}\|g^k\|_{x^k}^2 \\
=&-\langle -\frac{1}{\beta}g^k, P_{x^{k+1}}^{x^k} g^{k+1} - g^k\rangle_{x^k} + \frac{1}{2\beta}\|P_{x^{k+1}}^{x^k} g^{k+1} - g^k\|_{x^k}^2.
\end{aligned}$$

By combining this and Assumption 3.1, we have the following bound for the difference of the merit function (defined in (3.4)), evaluated at successive iterates:

$$\begin{aligned}
&W(x^{k+1}, g^{k+1}) - W(x^k, g^k) \\
\leq& -t_k\langle \mathsf{grad} f(x^k), g^k\rangle_{x^k} + \frac{t_k^2 L}{2}\|g^k\|_{x^k}^2 + \frac{1}{\beta}\langle g^k, P_{x^{k+1}}^{x^k} g^{k+1} - g^k\rangle_{x^k} + \frac{1}{2\beta}\|P_{x^{k+1}}^{x^k} g^{k+1} - g^k\|_{x^k}^2 \\
=&\left(\frac{t_k^2 L}{2} - t_k\right)\|g^k\|_{x^k}^2 + t_k\langle g^k, G_\mu^k - \mathsf{grad} f(x^k)\rangle_{x^k} + \frac{1}{2\beta}\|P_{x^{k+1}}^{x^k} g^{k+1} - g^k\|_{x^k}^2.
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
&\mathbb{E}_{u^k}[\langle g^k, G_\mu(x^k) - \mathsf{grad} f(x^k)\rangle_{x^k}] = \langle g^k, \mathbb{E}_{u^k} G_\mu(x^k) - \mathsf{grad} f(x^k)\rangle_{x^k} \\
\leq&\frac{1}{2}\|g^k\|_{x^k}^2 + \frac{1}{2}\|\mathbb{E}_{u^k} G_\mu(x^k) - \mathsf{grad} f(x^k)\|_{x^k}^2 \leq \frac{1}{2}\|g^k\|_{x^k}^2 + \frac{1}{2}\hat{\sigma}^2,
\end{aligned}$$

where the expectation is only taken with respect to the Gaussian random variables that we used to construct $G_\mu(x^k)$. Therefore, by using the law of total expectation, we have $\mathbb{E}W(x^{k+1},g^{k+1})-\mathbb{E}W(x^k,g^k)\leq\frac{1}{\beta}\left(\frac{\tau_k^2L}{2\beta}-\frac{\tau_k}{2}\right)\mathbb{E}\|g^k\|_{x^k}^2+\frac{\tau_k}{2\beta}\hat{\sigma}^2+\frac{1}{2\beta}\mathbb{E}\|P_{x^{k+1}}^{x^k}g^{k+1}-g^k\|_{x^k}^2$, and we thus have (by summing up the above inequality over $k=0,...,N$):

$$\begin{aligned}&\sum_{k=0}^{N}\left(\mathbb{E}W(x^{k+1},g^{k+1})-\mathbb{E}W(x^k,g^k)\right)\\&\leq\sum_{k=0}^{N}\frac{1}{2\beta}\left(\frac{\tau_k^2L}{\beta}-\tau_k\right)\mathbb{E}\|g^k\|_{x^k}^2+\sum_{k=0}^{N}\frac{\tau_k}{2\beta}\hat{\sigma}^2+\frac{1}{2\beta}\sum_{k=1}^{N}\mathbb{E}\|P_{x^{k+1}}^{x^k}g^{k+1}-g^k\|_{x^k}^2,\end{aligned}\tag{3.10}$$

where the last term sums from 1 since $g^1-P_{x^0}^{x^1}g^0=\tau_0(G_\mu^0-g^0)=0$.

Utilizing (3.8) and (3.10), we have (note that $W\geq 0$)

$$\begin{aligned}&\sum_{k=0}^{N}\frac{1}{2\beta}\left(\tau_k-\frac{\tau_k^2L}{\beta}\right)\mathbb{E}\|g^k\|_{x^k}^2\leq W(x^0,g^0)+\sum_{k=0}^{N}\frac{\tau_k}{2\beta}\hat{\sigma}^2+\frac{1}{2\beta}\sum_{k=1}^{N}\mathbb{E}\|P_{x^{k+1}}^{x^k}g^{k+1}-g^k\|_{x^k}^2\\\leq&W(x^0,g^0)+\frac{1}{2\beta}\sum_{k=0}^{N}(\tau_k+2\tau_k^2)\hat{\sigma}^2+\frac{1}{\beta}\sum_{k=0}^{N}\left(\tau_k^2+\tau_k^3\right)\sigma_k^2+\frac{1}{\beta}\sum_{k=0}^{N}\tau_k^2\tilde{\sigma}_0^2\\&+\frac{1}{\beta}\sum_{k=0}^{N}(1+\tau_k)\tau_k^2\frac{L^2\mathbb{E}\|g^k\|_{x^k}^2}{\beta^2}+\frac{1}{\beta}\sum_{k=0}^{N}\left(\tau_k^2+\tau_k^3\right)\frac{8(d+4)}{m_k}\mathbb{E}\|\mathsf{grad}f(x^k)\|_{x^k}^2.\end{aligned}$$

Combining this with (3.9a) we have

$$\begin{aligned}\sum_{k=0}^{N}\tau_k\mathbb{E}\|g^k\|_{x^k}^2\leq&\,4\beta W(x^0,g^0)+2\sum_{k=0}^{N}(\tau_k+2\tau_k^2)\hat{\sigma}^2+4\sum_{k=0}^{N}\left(\tau_k^2+\tau_k^3\right)\sigma_k^2\\&+4\sum_{k=0}^{N}\tau_k^2\tilde{\sigma}_0^2+4\sum_{k=0}^{N}\left(\tau_k^2+\tau_k^3\right)\frac{8(d+4)}{m_k}\mathbb{E}\|\mathsf{grad}f(x^k)\|_{x^k}^2.\end{aligned}\tag{3.11}$$

By Lemma 3.2 and (3.11), we get (also by $\tau_k\leq 1$)

$$\begin{aligned}&\frac{1}{2}\sum_{k=0}^{N}\tau_k\mathbb{E}\|\mathsf{grad}f(x^k)\|_{x^k}^2\leq\sum_{k=0}^{N}\tau_k\mathbb{E}\|g^k-\mathsf{grad}f(x^k)\|_{x^k}^2+\sum_{k=0}^{N}\tau_k\mathbb{E}\|g^k\|_{x^k}^2\\\leq&\sum_{k=0}^{N-1}\tau_k^2\tilde{\sigma}_k^2+\sum_{k=0}^{N-1}\tau_k\hat{\sigma}^2+\left(\frac{2L^2}{\beta^2}+1\right)\sum_{k=0}^{N}\tau_k\mathbb{E}\|g^k\|_{x^k}^2+2\tilde{\sigma}_0^2\\\leq&\left(\frac{8L^2}{\beta}+4\beta\right)W(x^0,g^0)+\sum_{k=0}^{N}\left[\tau_k+2\left(\frac{2L^2}{\beta^2}+1\right)(\tau_k+2\tau_k^2)\right]\hat{\sigma}^2\\&+\sum_{k=0}^{N}\left[\tau_k^2+4\left(\frac{2L^2}{\beta^2}+1\right)(\tau_k^2+\tau_k^3)\right]\sigma_k^2+\left[4\left(\frac{2L^2}{\beta^2}+1\right)\sum_{k=0}^{N}\tau_k^2+2\right]\tilde{\sigma}_0^2\\&+\sum_{k=0}^{N}\left[4\left(\frac{2L^2}{\beta^2}+1\right)(\tau_k^2+\tau_k^3)+\tau_k^2\right]\frac{8(d+4)}{m_k}\mathbb{E}\|\mathsf{grad}f(x^k)\|_{x^k}^2,\end{aligned}\tag{3.12}$$

where $\tau_0 \mathbb{E}\|g^0 - \mathsf{grad} f(x^0)\|_{x^0}^2 \leq \tilde{\sigma}_0^2$ is used in the last term on the second line. By combining (3.12) and (3.9b) we get

$$\begin{aligned}
&\sum_{k=0}^{N} \frac{\tau_k}{4} \mathbb{E}\|\mathsf{grad} f(x^k)\|_{x^k}^2 \\
\leq &\sum_{k=0}^{N} \left[\frac{\tau_k}{2} - \left(4\left(\frac{2L^2}{\beta^2} + 1 \right)(\tau_k^2 + \tau_k^3) + \tau_k^2 \right) \frac{8(d+4)}{m_k} \right] \mathbb{E}\|\mathsf{grad} f(x^k)\|_{x^k}^2 \\
\leq &\left(\frac{8L^2}{\beta} + 4\beta \right) W(x^0, g^0) + \sum_{k=0}^{N} \left[\tau_k + 2\left(\frac{2L^2}{\beta^2} + 1 \right)(\tau_k + 2\tau_k^2) \right] \hat{\sigma}^2 \\
&+ \sum_{k=0}^{N} \left[\tau_k^2 + 4\left(\frac{2L^2}{\beta^2} + 1 \right)(\tau_k^2 + \tau_k^3) \right] \sigma_k^2 + \left[4\left(\frac{2L^2}{\beta^2} + 1 \right) \sum_{k=0}^{N} \tau_k^2 + 2 \right] \tilde{\sigma}_0^2.
\end{aligned} \tag{3.13}$$

For case (i) in Theorem 3.1, (3.13) can be rewritten as

$$\frac{1}{N+1} \sum_{k=0}^{N} \mathbb{E}\|\mathsf{grad} f(x^k)\|_{x^k}^2 \leq \frac{c_1 W(x^0, g^0)}{\sqrt{N}} + c_2 \hat{\sigma}^2 + \frac{c_3 \frac{1}{N} \sum_{k=0}^{N} \sigma_k^2}{\sqrt{N}} + \frac{c_4}{\sqrt{N}} \tilde{\sigma}_0^2,$$

for some absolute positive constants c_1, c_2, c_3 and c_4. The proof for case (i) is completed by noting that (see (3.6)) $\hat{\sigma}^2 = \mathcal{O}(1/\sqrt{N})$, $\frac{1}{N} \sum_{k=0}^{N} \sigma_k^2 = \mathcal{O}(1)$ and $\tilde{\sigma}_0^2 = \mathcal{O}(1)$.

For case (ii) in Theorem 3.1, (3.13) can be rewritten as

$$\frac{1}{N+1} \sum_{k=0}^{N} \mathbb{E}\|\mathsf{grad} f(x^k)\|_{x^k}^2 \leq c_1' W(x^0, g^0) \sqrt{\frac{d}{N}} + c_2' \hat{\sigma}^2 + \frac{c_3' \frac{1}{N} \sum_{k=0}^{N} \sigma_k^2}{\sqrt{dN}} + c_4' \sqrt{\frac{d}{N}} \tilde{\sigma}_0^2,$$

for some positive constants c_1', c_2', c_3' and c_4'. The proof of case (ii) is completed by noting that $\tilde{\sigma}_0^2 = \mathcal{O}(1)$, $\hat{\sigma}^2 = \mathcal{O}(1/\sqrt{N})$ and $\frac{1}{N} \sum_{k=0}^{N} \sigma_k^2 = \mathcal{O}(d)$. □

Remark 3.1. *If we sample $R \in \{0, 1, 2, ..., N\}$ with $\mathbb{P}(R = k) = \tau_k / (\sum_{k=0}^{N} \tau_k)$, then the left hand side of the inequalities in Theorem 3.1, i.e., $\frac{1}{N+1} \sum_{k=0}^{N} \mathbb{E}\|\mathsf{grad} f(x^k)\|_{x^k}^2$, becomes $\mathbb{E}\|\mathsf{grad} f(x^R)\|_{x^R}^2$. If we use this sampling in case (i) of Theorem 3.1, then to get an ϵ-approximate stationary solution as in Definition 2.2, we require an iteration complexity of $N = \mathcal{O}(1/\epsilon^4)$ and so an oracle complexity of $Nm = \mathcal{O}(d/\epsilon^4)$. Case (i) requires $m = \mathcal{O}(d)$ per-iteration, which might be inconvenient in practice. Case (ii) of Theorem 3.1 avoids this, as in case (ii) both the iteration complexity and the oracle complexity are $N = \mathcal{O}(d/\epsilon^4)$, with batch size $m = \mathcal{O}(1)$. This makes case (ii) more convenient to use in practice, from a streaming or online perspective. For the simulations in Section 5, we thus choose $m = \mathcal{O}(1)$ and apply the result from case (ii). We also remark that the above results provide concrete solutions to the question raised by Scheinberg (2022), namely, on the need for mini-batches (and its order per-iteration) in zeroth-order stochastic optimization*[1].

Remark 3.2. *Notice that to prove (3.9b), we need $N = \Omega(d)$ for case (ii) in Theorem 3.1. We can remove this condition if in addition we have that $\mathsf{grad} f(x)$ is uniformly upper bounded:*

[1] Although Scheinberg (2022) focuses on the Euclidean case, the discussion there also holds in the Riemannian setting.

$\|\mathsf{grad} f(x)\|_x \leq G,\ \forall x \in \mathcal{M}$; *see also Assumption 4.2 which we utilize in the next section. Under this condition,* (3.12) *directly gives:*

$$\frac{1}{2}\sum_{k=0}^{N}\tau_k\mathbb{E}\|\mathsf{grad} f(x^k)\|_{x^k}^2 \leq \sum_{k=0}^{N}\left[\tau_k + 2\left(\frac{2L^2}{\beta^2}+1\right)(\tau_k+2\tau_k^2)\right]\hat{\sigma}^2$$
$$+\sum_{k=0}^{N}\left[\tau_k^2 + 4\left(\frac{2L^2}{\beta^2}+1\right)(\tau_k^2+\tau_k^3)\right]\sigma_k^2 + \left[4\left(\frac{2L^2}{\beta^2}+1\right)\sum_{k=0}^{N}\tau_k^2+2\right]\tilde{\sigma}_0^2$$
$$+\sum_{k=0}^{N}\left[4\left(\frac{2L^2}{\beta^2}+1\right)(\tau_k^2+\tau_k^3)+\tau_k^2\right]\frac{8(d+4)}{m_k}G^2 + \left(\frac{8L^2}{\beta}+4\beta\right)W(x^0,g^0),$$

whose right hand side has the same order as (3.13)*. Therefore in this case we do not need* $N=\Omega(d)$ *for case (ii) to achieve the same rates of convergence as in Theorem 3.1.*

4 RASA with retractions and vector transports

Algorithm 1 is based on exponential mapping and parallel transport, which has a high per-iteration complexity for various manifold choices $\mathcal{M}$. In this section, we focus on reducing the per-iteration complexity of the `Zo-RASA` algorithm. The approach is based on replacing the exponential mapping and parallel transport with retractions and vector transports, respectively, which leads to practically efficient implementations and improved per-iteration complexity.

The convergence analysis of algorithms with retractions and vector transports are sharply different and much harder than the one we presented in Section 3. Recall that the analysis in Section 3 relied on the isometry property (2.2) of the parallel transports, which is no longer available for vector transports. We hence assume explicit global error bounds between the difference of retraction to exponential mapping, as well as vector transport to parallel transport in Assumption 4.1. In Section 4.1.2 we provide conditions on the manifold under which such assumptions are naturally satisfied and provide explicit examples. Based on this, we establish that under a bounded fourth (instead of the second) central moment condition, the same sample complexity result as in the previous section could be obtained for the practical versions of `Zo-RASA` algorithm on compact manifolds.

4.1 Approximation error of retractions and vector transports

We start with the following condition on the vector transport used; recall the notation from Definition 2.3.

Assumption 4.1. *If* $x^+ = \mathsf{Retr}_x(g)$, $g \in \mathrm{T}_x\mathcal{M}$, *then with* d *denoting the geodesic distance, the vector transport* $\mathcal{T}_g$ *satisfies the following inequalities:*

$$\|\mathcal{T}_g(v)\|_{x^+} \leq \|v\|_x,\ \mathsf{d}(x,x^+) \leq \|g\|_x,\ \|\mathcal{T}_g(v) - P_x^{x^+}(v)\|_{x^+} \leq C\|v\|_x\mathsf{d}(x,x^+) \tag{4.1}$$

for any vector $v \in \mathrm{T}_x\mathcal{M}$.

An intuitive explanation of the first inequality in (4.1) is that our retraction and vector transport are "conservative" so that their length/magnitude is not longer than the exact operation of exponential mapping and parallel transport. As for the last inequality in (4.1), we are essentially positing that the vector transport would not "twist" the vector too much so that its difference from the parallel-transported vector is not large. In general, conditions in (4.1) require the vector transport not to be very far from the parallel-transported vectors on the new tangent space.

4.1.1 Comparison to prior works

We now provide a detailed comparison to similar type of conditions proposed in two prior works, Huang et al. (2015) and Sato (2022), and highlight the differences and advantages of our proposal. According to the definition of vector transport in Definition 2.3, we need to specify a retraction associated with the transport so that $\mathcal{T}_{\eta_x}(\xi_x) \in \mathrm{T}_{R_x(\eta_x)}\mathcal{M}$. In this section, we consider the projection retraction, denoted simply as R.

Given two transports, $\mathcal{T}_S$ and $\mathcal{T}_R$, Huang et al. (2015) propose certain conditions on approximating one with the other. First they require that $\mathcal{T}_S$ is isometric, i.e., $\langle \mathcal{T}_{S_\eta}(\xi), \mathcal{T}_{S_\eta}(\zeta)\rangle_{R_x(\eta)} = \langle \xi, \zeta\rangle_x$, hence we can basically regard $\mathcal{T}_S$ as parallel transport for comparison. Let $\mathcal{T}_R$ denote the differential of the retraction, given by $\mathcal{T}_{R_\eta}(\xi) = DR_x(\eta)[\xi] = \frac{d}{dt}R_x(\eta + t\xi) \in \mathrm{T}_{R_x(\eta)}\mathcal{M}$. Now the conditions stated in Equations (2.5) and (2.6) in Huang et al. (2015) are as follows: there exists a *neighborhood* $\mathcal{U}$ of x, such that $\forall y \in \mathcal{U}$ we have $\|\mathcal{T}_{S_\eta} - \mathcal{T}_{R_\eta}\|_{\mathrm{op}} \leq c_0\|\eta\|_x$ and $\|\mathcal{T}_{S_\eta}^{-1} - \mathcal{T}_{R_\eta}^{-1}\|_{\mathrm{op}} \leq c_0\|\eta\|_x$, where $\eta = R_x^{-1}(y)$ and $\|\cdot\|_{\mathrm{op}}$ is the operator norm. These assumptions are essentially local results, and as a result, Huang et al. (2015) needs to impose an additional stringent condition (see, their Assumption 3.2) that all the updates in their algorithms are already sufficiently close to the (local) optimal value to prove their convergence results. With the above conditions (in particular for a $\mathcal{T}_{1_\eta}$ satisfying their conditions in (2.5) and (2.6)), Huang et al. (2015) shows in Lemma 3.5 that *locally* we have $\|\mathcal{T}_{1_\eta}(\xi) - \mathcal{T}_{2_\eta}(\xi)\|_y \leq c_0\|\eta\|_x\|\xi\|_x$. The proof of their Lemma 3.5 relies on the smoothness of the local coordinate form of the vector transports, which could hold only when we have a coordinate chart covering the local neighborhood we consider. Hence, the assumptions in Huang et al. (2015) are in a different flavor from ours. In particular, our assumptions are global, and we show in Theorem 4.1 that they are satisfied by a certain (global) assumption on the second fundamental form of the manifold $\mathcal{M}$.

The existing work Huang et al. (2015) also assumes the so-called locking condition $\mathcal{T}_{S_\eta}(\xi) = \beta\mathcal{T}_{R_\eta}(\xi)$, where $\beta = \|\xi\|_x/\|\mathcal{T}_{R_\xi}(\xi)\|_{R_\xi(x)}$, which means that the approximating transport keeps the same direction as the parallel transport $\mathcal{T}_S$. In our analysis, we avoid such a condition since we are trying to transport two vectors g^k and G_μ^k (see Algorithm 2), and not just one previous gradient as in the Riemannian quasi-Newton method (Huang et al., 2015). Another existing work Sato (2022) requires algorithm-specific conditions in their Assumption 3.1. To elaborate, we recall that the *deterministic* Riemannian conjugate gradient iterates (Algorithm 1 in Sato (2022)) are given by $x_{k+1} \leftarrow R_{x_k}(t_k\eta_k)$ and $\eta_{k+1} \leftarrow -\mathsf{grad}f(x_{k+1}) + \beta_{k+1}s_k\mathscr{T}^k(\eta_k)$, where t_k, β_k and s_k are parameters and $\mathscr{T}^k$ is a transport map from $\mathrm{T}_{x_k}\mathcal{M}$ to $\mathrm{T}_{x_{k+1}}\mathcal{M}$. Given this, their Assumption 3.1 requires that there exist $C \geq 0$ and index sets $K_1 \subset \mathbb{N}$ and $K_2 = \mathbb{N} - K_1$ such that $\left\|\mathscr{T}^{(k)}(\eta_k) - \mathrm{D}R_{x_k}(t_k\eta_k)[\eta_k]\right\|_{x_{k+1}} \leq Ct_k\|\eta_k\|_{x_k}^2, k \in K_1$ and $\left\|\mathscr{T}^{(k)}(\eta_k) - \mathrm{D}R_{x_k}(t_k\eta_k)[\eta_k]\right\|_{x_{k+1}} \leq C\left(t_k + t_k^2\right)\|\eta_k\|_{x_k}^2, k \in K_2$.

Our assumption differs from the above in three aspects: (i) we do not make algorithm-specific assumptions, where each inequality depends on the iterate number k; (ii) we are not only comparing transporting η_k (which is the direction along which we update x^k), but also the zeroth-order estimator G_μ^k (see Algorithm 2), i.e., we assume a more general inequality by replacing $\mathrm{D}R_x(t_k\eta)[\eta]$ with $\mathrm{D}R_x(t_k\eta)[\xi]$, where ξ can be different from η; (iii) we *derive* the last inequality in (4.1) using global assumption of second fundamental form of the manifold $\mathcal{M}$ in Theorem 4.1, instead of *assuming* it.

4.1.2 Illustrative Examples

We now further inspect Assumption 4.1 by checking the conditions under which (4.1) holds in general, and also verifying it for various matrix-manifolds arising in applications.

We start with the first inequality in (4.1). It holds naturally if the manifold is a submanifold and the vector transport is the orthogonal projection, due to the non-expansiveness of orthogonal projections. The second inequality in (4.1) is much trickier. For the scope of this work, we show that the second equation in (4.1) holds for projectional retractions and projectional vector transports on Stiefel manifold, which also includes spheres and orthogonal groups as special cases. If the inverse of the retraction in Assumption 4.1 is well-defined, the second inequality in (4.1) could equivalently be stated as $\|\mathsf{Exp}_x^{-1}(x^+)\|_x \leq \|\mathsf{Retr}_x^{-1}(x^+)\|_x$, which may hold for a larger class of manifolds and retractions. We leave a detailed study of this as future work.

Stiefel manifold. Consider the Stiefel manifold $\mathrm{St}(d,p)$ defined in (2.1), with the tangent space $\mathrm{T}_X \mathrm{St}(d,p) = \{\xi | X^\top \xi + \xi^\top X = 0\}$ and Euclidean inner product $\langle X, Y\rangle := \mathrm{tr}(X^\top Y)$. We consider the projectional retraction (Absil and Malick, 2012) given by $X^+ = R_X(\xi) := UV^\top$, where $X + \xi = U\Sigma V^\top$ is the (thin) singular value decomposition of $X + \xi$. Also, the projectional vector transport $\mathcal{T}$ is simply projecting a tangent vector $\xi \in \mathrm{T}_{X_0}\mathrm{St}(d,p)$ to $\mathrm{T}_X \mathrm{St}(d,p)$. It is clear that $\|\mathcal{T}(\xi)\| \leq \|\xi\|$ due to the non-expansiveness of orthogonal projections (note that $\mathrm{T}_X \mathrm{St}(d,p)$ is simply a linear subspace). To show $\mathsf{d}(X, X^+) \leq \|\xi\|$, denote $\gamma(t)$ the minimal geodesic connecting X and X^+ with $\gamma(0) = X$ and $\gamma(1) = X^+$, so that $\mathsf{d}(X, X^+) = \int_0^1 \|\gamma'(t)\| dt$. Notice that we can define another curve $c(t) = U(t)V^\top(t)$, where $X + t\xi = U(t)\Sigma(t)V^\top(t)$ is the singular value decomposition. The curve $c(t) = \mathsf{Retr}_X(t\xi)$ is the parameterized curve of projectional retraction. Now using the distance with respect to the minimal geodesic, we have $\mathsf{d}(X, X^+) = \int_0^1 \|\gamma'(t)\| dt \leq \int_0^1 \|c'(t)\| dt \leq \int_0^1 \|\xi\| dt = \|\xi\|$, where $\|c'(t)\| \leq \|\xi\|$ is due to the non-expansiveness of orthogonal projections, namely, $\|c(t_1) - c(t_2)\| \leq \|X + t_1\xi - (X + t_2\xi)\|$. Indeed, although $\mathrm{St}(d,p)$ is not a convex set, the non-expansiveness condition still holds (Gallivan and Absil, 2010), because $(X+\xi)^\top(X+\xi) = I_p + \xi^\top\xi \succeq I_p$, and the projection of $X + \xi$ onto the Stiefel manifold is the same as projection onto its convex hull $\{X \in \mathbb{R}^{d\times p} | \|X\|_2 \leq 1\}$. Now we turn to the last inequality in (4.1). Given a complete embedded submanifold, we can show that the last inequality in (4.1) holds under the boundedness of the second fundamental form in Theorem 4.1, given that the vector transport is the orthogonal projection to the new tangent space.

Theorem 4.1. *Suppose $\mathcal{M}$ is an embedded complete Riemannian submanifold of Euclidean space. Suppose for all unit vector $\xi, \eta \in \mathrm{T}\mathcal{M}$, $\|\xi\| = \|\eta\| = 1$, the norm of the second fundamental form $B(\xi, \eta)$ is bounded by constant C. Consider the parallel transport P_x^y along the minimal geodesic from $x \in \mathcal{M}$ to $y \in \mathcal{M}$, we have $\|\mathsf{proj}_{\mathrm{T}_y\mathcal{M}}(v) - P_x^y(v)\| \leq C\|v\|\mathsf{d}(x,y)$, for any $v \in \mathrm{T}_x\mathcal{M}$. That is, the last inequality in (4.1) holds with constant C.*

Proof. Without loss of generality, we assume $\|v\| = 1$, otherwise conduct the proof for $v/\|v\|$. Denote the minimum geodesic γ with unit speed connecting x and y, parameterized by variable t, also denote the parallel transported vector of v along γ as $v(t)$, i.e. $v(0) = v$. Now for the extrinsic geometry, we denote $v = v^\top(t) + v^\perp(t)$, where $v^\top(t) \in \mathrm{T}_{\gamma(t)}\mathcal{M}$ and $v^\perp(t)$ is orthogonal to $\mathrm{T}_{\gamma(t)}\mathcal{M}$. Note that the left-hand side of the inequality we want to prove is now parameterized as $\|v(t) - v^\top(t)\|$.

Now since $v(t)$ is a parallel transport of v, the tangent component must be zero, i.e., $(v'(t))^\top = 0$. Now consider any parallel unit vector $z(t) \in \mathrm{T}_{\gamma(t)}\mathcal{M}$ along γ, then $\langle (v^\perp)'(t), z(t)\rangle = -\langle v^\perp(t), z'(t)\rangle = -\langle v^\perp(t), B(\gamma'(t), z(t))\rangle$, where B is the second fundamental form. Along with the fact that $(v^\top)' = -(v^\perp)'$ we get $\langle (v^\top)'(t), z(t)\rangle = \langle v^\perp(t), B(\gamma'(t), z(t))\rangle$. Now the right-hand side has a uniform upper bound of C, and by the arbitrarily chosen $z(t) \in \mathrm{T}_{\gamma(t)}\mathcal{M}$, we get $\|((v^\top)'(t))^\top\| \leq C$.

We can now bound the derivative of $\|v(t) - v^\top(t)\|$ as $(\|v(t) - v^\top(t)\|^2)' = (1 - 2\langle v(t), v^\top(t)\rangle + \|v^\top(t)\|^2)' = -2\langle v(t), (v^\top(t))'\rangle + 2\langle v^\top(t), (v^\top(t))'\rangle = 2\langle v^\top(t) - v(t), ((v^\top(t))')^\top\rangle \leq 2C\|v^\top(t) - v(t)\|$. Therefore, we get $\|v(t) - v^\top(t)\|' \leq C$. Now integrating the above inequality from x to y along

the minimal geodesic γ (i.e., with respect to t) and using the distance with respect to the minimal geodesic, we obtain $\|\mathsf{proj}_{\mathrm{T}_y\mathcal{M}}(v) - P_x^y(v)\| \leq C\mathsf{d}(x,y)$, which completes the proof. □

Theorem 4.1 connects extrinsic and intrinsic geometry by measuring the difference of orthogonal projection (extrinsic operation) and parallel transport (intrinsic operation), which might be of independent interest for studying embedded submanifolds. The condition in Theorem 4.1 is stronger than the bounded sectional curvature condition since if the second fundamental form is bounded, the sectional curvature is also bounded by the Gauss formula (see Chapter 6, Theorem 2.5 in Do Carmo (1992)). We point out that the condition of Theorem 4.1 is still satisfied by all the embedded submanifold applications we consider, namely the sphere, the orthogonal group and the Stiefel manifold. In particular, we have the following observation.

Proposition 4.1. *Suppose $\mathcal{M}$ is a compact complete embedded Riemannian submanifold of Euclidean space (i.e. satisfying Assumption 4.2), then the norm of the second fundamental form $\|B(\xi,\eta)\|$ is uniformly bounded for all unit vector $\xi,\eta \in \mathrm{T}\mathcal{M}$, $\|\xi\| = \|\eta\| = 1$.*

The proof is immediate, since for all unit vector $\xi,\eta \in \mathrm{T}\mathcal{M}$, $\|B(\xi,\eta)\| \in \mathbb{R}$ is a smooth function defined over a compact domain, and therefore it is upper bounded. As a result, Assumption 4.1 holds for all the embedded submanifold applications we consider, namely the sphere, the orthogonal group and the Stiefel manifold.

Remark 4.1. *We remind the readers that Theorem 4.1 requires the embedded submanifold assumption, yet Assumption 4.1 does not, as long as (4.1) hold. This is also the main reason why we summarize our assumption as in Assumption 4.1, and not present Theorem 4.1 directly.*

Example: Grassmann manifold. Above, we have shown that Assumption 4.1 holds for a class of embedded matrix submanifolds. Yet another setting is that of quotient manifolds (e.g., the Grassmann manifold) which arises in applications of Riemannian optimization. Such manifolds are not naturally embedded submanifolds of a Euclidean space. As a result, we can inspect Assumption 4.1 directly for such manifolds. Taking the Grassmann manifold as an example, we next verify Assumption 4.1. To proceed, we utilize the following result.

Lemma 4.1. *Suppose $X \in \mathrm{St}(d,p)$, $G \in \mathbb{R}^{d\times p}$ with $X^\top G = 0$, and the QR decomposition of $X+G = QR$ where $Q \in \mathrm{St}(d,p)$ and $R \in \mathbb{R}^{p\times p}$ is upper triangular. The principal angle between the subspace spanned by X and Q is given by $\|\Theta\|_F$, where $\Theta := \arccos(\Sigma)$ where Σ is the singular value matrix of $X^\top Q$, i.e., $X^\top Q = U\Sigma V^\top$; see, for example Edelman et al. (1998, Section 4.3). We have that $\|\Theta\|_F \leq \|G\|_F$.*

Proof. Since $R^\top R = (X+G)^\top(X+G) = I_p + \|G\|_F^2$, we know that all the singular values of R are greater than or equal to 1. Denote $\Sigma = \mathrm{diag}([\sigma_1,...,\sigma_p])$. Since $X^\top Q = X^\top(X+G)R^{-1} = R^{-1}$, we know that the singular value decomposition of $R = V\Sigma^{-1}U^\top$ (which implies that $\sigma_i \leq 1$, $\forall i = 1,2,....,p$) and $\|R\|_F^2 = \|\Sigma^{-1}\|_F^2 = \sum_{i=1}^p \frac{1}{\sigma_i^2}$. Also, as $\|R\|_F^2 = \|X+G\|_F^2 = \mathrm{tr}((X+G)^\top(X+G)) = p + \|G\|_F^2$, we get $\|G\|_F^2 = \sum_{i=1}^p \frac{1}{\sigma_i^2} - p$. Thus, $\|\Theta\|_F^2 = \|\arccos(\Sigma)\|_F^2 = \sum_{i=1}^p (\arccos(\sigma_i))^2 \leq \sum_{i=1}^p (\frac{1}{\sigma_i^2} - 1) = \|G\|_F^2$, where we use the fact that $(\arccos(t))^2 \leq \frac{1}{t^2} - 1$, $\forall t \in (0,1]$. □

Now we can inspect the Grassmann manifold. The Grassmann manifold $\mathrm{Gr}(d,p)$ is the set of all p-dimensional subspace of $\mathbb{R}^d$; see, for example, (Absil et al., 2008, Section 2.1). A quotient formulation writes $\mathrm{Gr}(d,p) = \mathrm{St}(d,p)/\mathcal{O}(p)$ with $\mathcal{O}(p) = \{Q \in \mathbb{R}^{p\times p} | Q^\top Q = I_p\}$ being the orthogonal group. The elements of the Grassmann manifold can be expressed as $[X] \in \mathrm{Gr}(d,p)$ with $[X] := \{XQ | Q \in \mathcal{O}(p)\}$

and $X \in \mathrm{St}(d,p)$. The element $\bar{\xi}$ on the tangent space $\mathrm{T}_{[X]}\,\mathrm{Gr}(d,p)$ can be shown with a one-to-one mapping (called the horizontal lift) to the set $[\xi]$ with $\xi \in \mathrm{T}_X\,\mathrm{St}(d,p)$ and $X^\top \xi = 0$.

Suppose we start from an element $[X] \in \mathrm{Gr}(d,p)$ with $X \in \mathrm{St}(d,p)$ and the initial speed $\bar{G} \in \mathrm{T}_{[X]}\,\mathrm{Gr}(d,p)$, where $G \in \mathrm{T}_X\,\mathrm{St}(d,p)$ and $X^\top G = 0$. We denote the singular value decomposition of $G = U\Sigma V^\top$ with $U \in \mathbb{R}^{d\times p}$ and $\Sigma, V \in \mathbb{R}^{p\times p}$. Then the exponential mapping is given by $Y := \mathsf{Exp}_{[X]}(\bar{G}) = [XV\cos(\Sigma) + U\sin(\Sigma)]$, where sin and cos are matrix trigonometric functions; see (Absil et al., 2008, Example 5.4.3). Also, the parallel transport is given by: $\bar{\xi_1} = P^{[Y]}_{[X]}(\bar{\xi})$ with $\xi_1 = -XV\sin(\Sigma)U^\top\xi + U\cos(\Sigma)U^\top\xi + (I - UU^\top)\xi$. See (Absil et al., 2008, Example 8.1.3). Hence, the projectional retraction is given by $Y' := \mathsf{Retr}_{[X]}(\bar{G}) = [X+G] = [Q]$, where $X + G = QR$ is the QR decomposition of $X+G$; see (Absil et al., 2008, Example 4.1.5). Furthermore, the projectional vector transport is given by $\bar{\xi_2} = \mathcal{T}_{\bar{G}}(\bar{\xi})$ with $\xi_2 = (I - YY^\top)\xi$. See (Absil et al., 2008, Example 8.1.10).

Now we show that (4.1) is satisfied. It is obvious that $\|\mathcal{T}_{\bar{G}}(\bar{\xi})\| = \|(I - YY^\top)\xi\| \le \|\xi\|$. The geodesic distance of $[X]$ and the projectional retraction $[Q]$ is exactly the principal angle between the subspace spanned by X and Q, see (Edelman et al., 1998, Section 4.3). Following Lemma 4.1, we can hence conclude that $\mathsf{d}([X],[Q]) = \|\Theta\|_F \le \|G\|_F$. Now we inspect the last equation in (4.1). We can directly check that $\|\xi_1 - \xi_2\|_F = \|A\xi\|_F \le \|A\|_F\|\xi\|_F$, with

$$\begin{aligned} A :=& -XV\sin(\Sigma)U^\top + U\cos(\Sigma)U^\top + YY^\top - UU^\top \\ =& -XV\sin(\Sigma)U^\top + U\cos(\Sigma)U^\top - U\cos^2(\Sigma)U^\top + XV\cos^2(\Sigma)V^\top X^\top \\ &+ U\sin(\Sigma)\cos(\Sigma)V^\top X^\top + XV\cos(\Sigma)\sin(\Sigma)U^\top. \end{aligned}$$

Note also that we have the bound

$$\begin{aligned} \|A\| =& \| - XV\sin(\Sigma)U^\top + U\cos(\Sigma)U^\top - U\cos^2(\Sigma)U^\top + XV\cos^2(\Sigma)V^\top X^\top \\ &+ U\sin(\Sigma)\cos(\Sigma)V^\top X^\top + XV\cos(\Sigma)\sin(\Sigma)U^\top\| \\ \le& \|\sin(\Sigma)\| + \|\cos(\Sigma)(I - \cos(\Sigma))\| + 2\|\sin(\Sigma)\cos(\Sigma)\| \le 4\|\sin(\Sigma)\| \le 4\|G\|, \end{aligned}$$

where we use the fact that $X^\top X = U^\top U = V^\top V = I_p$ and all norms are the Frobenius norm. Therefore, we see that the last equation in (4.1) is satisfied with $C = 4$.

4.2 Convergence of retraction and vector transport based `Zo-RASA`

We now proceed to the convergence analysis of `Zo-RASA` algorithm with retraction and vector transports. Algorithm 2 is the analog of Algorithm 1, using retraction and vector transport. Notice that the zeroth-order estimator used in Algorithm 2 is as defined in (1.3), which is with respect to the retraction in contrast to (1.2). Also $\mathcal{T}$ is the vector transport where we write $\mathcal{T}^k := \mathcal{T}_{-t_k g^k}$ for brevity. The vector transport we use in experiments is simply the orthogonal projection onto the target tangent space.

For our analysis, apart from the smoothness condition in Assumption 3.1, we also need to assume that the manifold is compact.

Assumption 4.2. *The manifold $\mathcal{M}$ is compact with diameter D, and the Riemannian gradient satisfies $\|\mathsf{grad} f(x)\|_x \le G$.*

Here, G could potentially be a function of D and the constant L from Assumption 3.1, due to compactness and smoothness. We remark that this compactness assumption is satisfied by various

Algorithm 2: Zo-RASA with retraction and vector transport

1: Change the updates of x^{k+1} and g^{k+1} in Algorithm 1 respectively to

$$x^{k+1} \leftarrow \mathsf{Retr}_{x^k}(-t_k g^k) \quad \text{and} \quad g^{k+1} \leftarrow (1-\tau_k)\mathcal{T}^k(g^k) + \tau_k \mathcal{T}^k(G_\mu^k),$$

where $G_\mu^k = G_\mu^{\mathsf{Retr}}(x^k)$ is given by (1.3) with batch-size $m = m_k$.

matrix manifolds like the Stiefel manifold and the Grassmann manifold (see, for example, Lemma 5.1 in Milnor and Stasheff (1974)).

Turning to the stochastic gradient oracles, the bounded second moment condition in Assumption 3.2 is now replaced by the following condition of bounded fourth central moment. Such a condition is needed to conduct our convergence analysis. It is interesting to relax this assumption or show this condition is necessary and sufficient to design batch-free, fully-online algorithms with vector transports and retractions.

Assumption 4.3. *Along the trajectory of the algorithm, we have that the stochastic gradients are unbiased and have bounded fourth central moment, i.e., for each $k \in \{1, \dots, N\}$, we have $\mathbb{E}_\xi[\mathsf{grad} F(x^k; \xi_k) | \mathcal{F}_{k-1}] = \mathsf{grad} f(x^k)$ and $\mathbb{E}_\xi[\|\mathsf{grad} F(x^k; \xi_k) - \mathsf{grad} f(x^k)\|_{x^k}^4 | \mathcal{F}_{k-1}] \le \sigma^4$.*

Note that Assumption 4.3 implies Assumption 3.2. To proceed with the convergence analysis of Algorithm 2, we also need to assume that the retraction we use in Algorithm 2 is a second-order retraction, as in Assumption 4.4.

Assumption 4.4. *The retraction we use in Algorithm 2 is a second order retraction, i.e. $\forall \xi \in \mathrm{T}_x \mathcal{M}$, we have $\mathsf{d}(\mathsf{Retr}_x(\xi), \mathsf{Exp}_x(\xi)) \le C\|\xi\|_x^2$.*

Note that the notion of second order retraction is only a local property, i.e., the above inequality only holds when $\|\xi\|$ is not too large. We refer to second order retraction without this locality restriction, since we assume the compactness of $\mathcal{M}$ in Assumption 4.2 and thus the condition in Assumption 4.4 also holds for large $\|\xi\|$ and the constant C will globally depend on the curvature of the manifold. We also point out that the condition in Assumption 4.4 is satisfied by projectional retractions; see, for example, (Absil and Malick, 2012, Proposition 2.2). The study of higher-order (better) approximation to the exponential mapping by the retractions is still an on-going research topic Gawlik and Leok (2018), while here we only need a second-order retraction.

The following result in Lemma 4.2, which is a standard comparison-type result, will be utilized in the subsequent proof.

Lemma 4.2 (Theorem 6.5.6 in Burago et al. (2022))**.** *Suppose the sectional curvature of $\mathcal{M}$ is upper bounded, then $\forall \xi, \eta \in \mathrm{T}_x \mathcal{M}$, we have $\|\xi - \eta\|_x \le C\, \mathsf{d}(\mathsf{Exp}_x(\xi), \mathsf{Exp}_x(\xi))$, without loss of generality we assume the constant to be $C = 1$ for the rest of the paper.*

The following result shows that with a second-order retraction, the smoothness with respect to exponential mapping implies the smoothness with respect to retractions.

Lemma 4.3. *Suppose Assumption 3.1, 4.1 and 4.2 hold, if the retraction we use in Algorithm 2 and (1.3) satisfy Assumption 4.4, then there exists a parameter $L' > 0$, such that f is also L'-smooth with the retraction, i.e., $|f(\mathsf{Retr}_x(\eta)) - f(x) - \langle \mathsf{grad} f(x), \eta \rangle_x| \le \frac{L'}{2}\|\eta\|_x^2$, $\forall \eta \in \mathrm{T}_x \mathcal{M}$. From now on, we denote L as the parameter that satisfies both Assumption 3.1 and Lemma 4.3 for brevity.*

Proof. Denote $y = \mathsf{Retr}_x(\eta)$. Note that we have $|f(y) - f(x) - \langle \mathsf{grad} f(x), \eta\rangle_x| \le |f(y) - f(x) - \langle \mathsf{grad} f(x), \mathsf{Exp}_x^{-1}(y)\rangle_x| + |\langle \mathsf{grad} f(x), \mathsf{Exp}_x^{-1}(y) - \eta\rangle_x| \le L\|\mathsf{Exp}_x^{-1}(y)\|_x^2 + \|\mathsf{grad} f(x)\|_x \|\eta - \mathsf{Exp}_x^{-1}(y)\|_x \le L\|\eta\|_x^2 + G\mathsf{d}(\mathsf{Exp}_x(\eta), y) \le (L + GC)\|\eta\|_x^2 =: L'\|\eta\|_x^2$, where the first inequality is by Assumption 3.2, the second is by Assumption 4.1 and Lemma 4.2, and the last inequality is by Assumption 4.4. □

We remind the readers that Lemma 4.3 can guarantee that the retraction-based zeroth-order estimator (1.3) still satisfies Lemma 3.1. In addition, we have the following bound on the fourth moment of G_μ^{Retr}.

Lemma 4.4. *Consider G_μ given by (1.3). Under Assumptions 3.1, 4.1, 4.2 and 4.3, we have $\mathbb{E}\|G_\mu^{\mathsf{Retr}}(x)\|_x^4 \le \frac{\mu^4 L^4}{2}(d+12)^6 + 3d^2\|\mathsf{grad} f(x)\|_x^4$, where the expectation is taken toward the Gaussian vectors when constructing G_μ and the random variable ξ.*

Proof. Since $\mathbb{E}\|G_\mu^{\mathsf{Retr}}(x)\|_x^4 = \frac{1}{\mu^4}\mathbb{E}_u[(f(\mathsf{Retr}_x(\mu u)) - f(x))^4 \|u\|_x^4]$ and

$$
\begin{aligned}
&(f(\mathsf{Retr}_x(\mu u)) - f(x))^4 \\
=&(f(\mathsf{Retr}_x(\mu u)) - f(x) - \langle \mathsf{grad} f(x), \mu u\rangle_x + \langle \mathsf{grad} f(x), \mu u\rangle_x)^4 \\
\le& 8(f(\mathsf{Retr}_x(\mu u)) - f(x) - \langle \mathsf{grad} f(x), \mu u\rangle_x)^4 + 8(\langle \mathsf{grad} f(x), \mu u\rangle_x)^4 \\
\le& 8\left(\frac{L}{2}\|\mu u\|_x^2\right)^4 + 8(\langle \mathsf{grad} f(x), \mu u\rangle_x)^4,
\end{aligned}
$$

where the last inequality is by Lemma 4.3. Therefore we have

$$
\begin{aligned}
\mathbb{E}\|G_\mu^{\mathsf{Retr}}(x)\|_x^4 &\le \frac{\mu^4 L^4}{2}\mathbb{E}\|u\|_x^{12} + 8\mathbb{E}[\langle \mathsf{grad} f(x), u\rangle_x^4 \|u\|_x^4] \\
&\le \frac{\mu^4 L^4}{2}(d+12)^6 + 8\mathbb{E}[\langle \mathsf{grad} f(x), u\rangle_x^4 \|u\|_x^4],
\end{aligned}
$$

where the last inequality is by Lemma 2 in Li et al. (2022). It remains to bound the last term on the right hand side, and we apply the same trick as in Proposition 1 in Li et al. (2022) here. Since u is an Gaussian vector on the tangent space $\mathrm{T}_x\mathcal{M}$ (dimension is d), we can calculate the expectation using the integral directly (denote $g = \mathsf{grad} f(x)$ and omit the subscript x for simplicity):

$$
\begin{aligned}
&\mathbb{E}(\|\langle \mathsf{grad} f(x), u\rangle u\|^4) = \frac{1}{\kappa(d)}\int_{\mathbb{R}^d} \langle g, x\rangle^4 \|x\|^4 e^{-\frac{1}{2}\|x\|^2} dx \\
\le& \frac{1}{\kappa(d)}\int_{\mathbb{R}^d} \|x\|^4 e^{-\frac{\tau}{2}\|x\|^2} \langle g, x\rangle^4 e^{-\frac{1-\tau}{2}\|x\|^2} dx \le \frac{1}{\kappa(d)}\left(\frac{4}{\tau e}\right)^2 \int_{\mathbb{R}^d} \langle g, x\rangle^4 e^{-\frac{1-\tau}{2}\|x\|^2} dx \\
=& \frac{1}{\kappa(d)}\left(\frac{4}{\tau e}\right)^2 \left(\frac{1}{1-\tau}\right)^{d/2-2} \int_{\mathbb{R}^d} \langle g, x\rangle^4 e^{-\frac{1}{2}\|x\|^2} dx = 48\left(\frac{1}{\tau e}\right)^2 \left(\frac{1}{1-\tau}\right)^{d/2-2} \|g\|^4,
\end{aligned}
$$

where $\kappa(d) := \int_{\mathbb{R}^d} e^{-\frac{1}{2}\|x\|^2} dx$ is the constant that normalizes Gaussian distribution, the second inequality is by the following fact: $x^p e^{-\frac{\tau}{2}x^2} \le (\frac{p}{\tau e})^{p/2}$, the second equality is by change of variables and the last equality is by $\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\langle g, x\rangle^4 = 3\|g\|^4$. Taking $\tau = 4/d$ gives the desired result. □

We now provide the convergence result for `Zo-RASA` (Algorithm 2). We remind the readers that we assume $C = 1$ in both Assumptions 4.1 and 4.4. We would first need to utilize the following Lemma 4.5, which is an analog to Lemma 3.2.

Lemma 4.5. *Suppose Assumptions 3.1, 4.1, 4.2, 4.3 and 4.4 hold, and $\{x^k, g^k\}$ is generated by Algorithm 1. We have*

$$\mathbb{E}\|g^k - \mathsf{grad} f(x^k)\|_{x^k}^2 \le \Gamma_k \tilde{\sigma}_0^2 + \Gamma_k \sum_{i=1}^{k} \Big(\frac{(1+\tau_{i-1})\tau_{i-1}}{\Gamma_i} \frac{L^2 \|g^{i-1}\|_{x^{i-1}}^2}{\beta^2} + \frac{\tau_{i-1}^2}{\Gamma_i} \tilde{\sigma}_{i-1}^2 + \tau_k \hat{\sigma}^2 \Big),$$

where the expectation $\mathbb{E}$ is taken with respect to all random variables up to iteration k, including the Gaussian variables $\{u_i\}_{i=1}^k$ in the zeroth-order estimator (1.2), and $\tilde{\sigma}_k^2$ is defined in (3.6). Further, from the definition of τ_k in (3.1), we have

$$\sum_{k=1}^{N} \tau_k \mathbb{E}\|g^k - \mathsf{grad} f(x^k)\|_{x^k}^2 \le \sum_{k=0}^{N-1} \left((1+\tau_k)\tau_k \frac{L^2 \mathbb{E}\|g^k\|_{x^k}^2}{\beta^2} + \tau_k^2 \tilde{\sigma}_k^2 + \tau_k \hat{\sigma}^2 \right) + \tilde{\sigma}_0^2,$$

$$\sum_{k=1}^{N} \tau_k^2 \mathbb{E}\|g^k - \mathsf{grad} f(x^k)\|_{x^k}^2 \le \sum_{k=0}^{N-1} \left((1+\tau_k)\tau_k^2 \frac{L^2 \mathbb{E}\|g^k\|_{x^k}^2}{\beta^2} + \tau_k^3 \tilde{\sigma}_k^2 + \tau_k^2 \hat{\sigma}^2 \right) + \sum_{k=1}^{N} \tau_k^2 \tilde{\sigma}_0^2.$$

Proof. The proof is almost identical to the proof of Lemma 3.2, and we thus omit the details. Note that here we need to utilize Assumption 4.1 to show $\mathsf{d}(x^i, x^{i+1})^2 \le t_i^2 \|g^i\|_{x^i}^2$. □

To show the bound for the term $\mathbb{E}\|P_{x^{k+1}}^{x^k} g^{k+1} - g^k\|_{x^k}^2$, we further need to utilize the following bound for $\|g^k\|_{x^k}$ first.

Lemma 4.6. *Consider g^k given by Algorithm 2. Suppose Assumption 3.1, 4.1, 4.2, 4.3 and 4.4 hold. Then, we have $\mathbb{E}\|g^k\|_{x^k}^2 \le \mu^2 L^2 (d+6)^3 + 2(d+4)G^2$ and $\mathbb{E}\|g^k\|_{x^k}^4 \le \frac{\mu^4 L^4}{2}(d+12)^6 + 3d^2 G^4$, where the expectation $\mathbb{E}$ is taken with respect to all random variables up to iteration k.*

Proof. Note that we have

$$\begin{aligned} \|g^k\|_{x^k}^2 &= \|(1-\tau_{k-1})\mathcal{T}^{k-1}(g^{k-1}) + \tau_{k-1}\mathcal{T}^{k-1}(G_\mu^{k-1})\|_{x^k}^2 \\ &\le (1-\tau_{k-1})\|g^{k-1}\|_{x^{k-1}}^2 + \tau_{k-1}\|G_\mu^{k-1}\|_{x^{k-1}}^2. \end{aligned}$$

Taking expectation conditioned on $\mathcal{F}_{k-1}$, we have by Lemma 3.1 that $\mathbb{E}[\|g^k\|_{x^k}^2 | \mathcal{F}_{k-1}] \le (1-\tau_{k-1})\mathbb{E}\|g^{k-1}\|_{x^{k-1}}^2 + \tau_{k-1}(\mu^2 L^2 (d+6)^3 + 2(d+4)\|\mathsf{grad} f(x^{k-1})\|_{x^{k-1}}^2)$. We remove the conditional expectation by law of total expectation, also by Assumption 4.2 we have that

$$\mathbb{E}\|g^k\|_{x^k}^2 \le (1-\tau_{k-1})\mathbb{E}\|g^{k-1}\|_{x^{k-1}}^2 + \tau_{k-1}(\mu^2 L^2 (d+6)^3 + 2(d+4)G^2).$$

Denote $A_k = \mathbb{E}\|g^k\|_{x^k}^2$, note that we have $A_k \le (1-\tau_{k-1})A_{k-1} + \tau_{k-1}(\mu^2 L^2 (d+6)^3 + 2(d+4)G^2)$. Again from Lemma 3.1 we have $A_0 \le \mu^2 L^2 (d+6)^3 + 2(d+4)G^2$, from which and using induction, we conclude that $A_k = \mathbb{E}\|g^k\|_{x^k}^2 \le \mu^2 L^2 (d+6)^3 + 2(d+4)G^2$. As for the fourth moment, note that

$$\begin{aligned} &\mathbb{E}(\|g^k\|_{x^k}^2)^2 \le \mathbb{E}\left((1-\tau_{k-1})\|g^{k-1}\|_{x^{k-1}}^2 + \tau_{k-1}\|G_\mu^{k-1}\|_{x^{k-1}}^2 \right)^2 \\ \le &(1-\tau_{k-1})\mathbb{E}\|g^{k-1}\|_{x^{k-1}}^4 + \tau_{k-1}\mathbb{E}\|G_\mu^{k-1}\|_{x^{k-1}}^4, \\ \le &(1-\tau_{k-1})\mathbb{E}\|g^{k-1}\|_{x^{k-1}}^4 + \tau_{k-1}\left(\frac{\mu^4 L^4}{2}(d+12)^6 + 3d^2 \|\mathsf{grad} f(x^k)\|_{x^k}^4 \right) \end{aligned}$$

where the last inequality is by Lemma 4.4. The final result follows similarly to the second moment case. □

Now we are ready to study the difference between g^k and g^{k+1}.

Lemma 4.7. *Suppose Assumptions 3.1, 4.1, 4.2, 4.3 and 4.4 hold, and take τ_k as in (3.1). Then, we have*

$$\begin{aligned}\sum_{k=1}^{N}\mathbb{E}\|P_{x^{k+1}}^{x^k}g^{k+1}-g^k\|_{x^k}^2 \le \frac{4L^2}{\beta^2}\sum_{k=0}^{N-1}(1+\tau_k)\tau_k^2\mathbb{E}\|g^k\|_{x^k}^2 + 4\sum_{k=0}^{N}(\tau_k^2+\tau_k^3)\tilde{\sigma}_k^2 \\ + \left[4\tilde{\sigma}_0^2+4\hat{\sigma}^2+\frac{8}{\beta^2}\left(\frac{\mu^4L^4}{2}(d+12)^6+3d^2G^4\right)\right]\sum_{k=0}^{N}\tau_k^2,\end{aligned} \tag{4.2}$$

where the expectation $\mathbb{E}$ is taken with respect to all random variables up to iteration k, which includes the random variables u in the zeroth-order estimator (1.3).

Proof. Since

$$\begin{aligned}&\|P_{x^{k+1}}^{x^k}g^{k+1}-g^k\|_{x^k}^2 = \|g^{k+1}-P_{x^k}^{x^{k+1}}g^k\|_{x^{k+1}}^2\\ \le&2\|g^{k+1}-\mathcal{T}^kg^k\|_{x^{k+1}}^2+2\|\mathcal{T}^kg^k-P_{x^k}^{x^{k+1}}g^k\|_{x^{k+1}}^2\\ \le&2\tau_k^2\|G_\mu^k-g^k\|_{x^k}^2+2\mathsf{d}(x^{k+1},x^k)^2\|g^k\|_{x^k}^2\\ \le&4\tau_k^2\|G_\mu^k-\mathsf{grad}f(x^k)\|_{x^k}^2+4\tau_k^2\|\mathsf{grad}f(x^k)-g^k\|_{x^k}^2+2\frac{\tau_k^2}{\beta^2}\|g^k\|_{x^k}^4,\end{aligned}$$

where the second inequality is by the update and Assumption 4.1, and the last inequality is by Assumption 4.1. Now taking the expectation conditioned on $\mathcal{F}_{k-1}$ we get:

$$\begin{aligned}\mathbb{E}[\|P_{x^{k+1}}^{x^k}g^{k+1}-g^k\|_{x^k}^2|\mathcal{F}_{k-1}] \le 4\tau_k^2\mathbb{E}[\|G_\mu^k-\mathsf{grad}f(x^k)\|_{x^k}^2|\mathcal{F}_{k-1}]\\ +4\tau_k^2\mathbb{E}[\|\mathsf{grad}f(x^k)-g^k\|_{x^k}^2|\mathcal{F}_{k-1}]+2\frac{\tau_k^2}{\beta^2}\mathbb{E}[\|g^k\|_{x^k}^4|\mathcal{F}_{k-1}].\end{aligned}$$

Thus we have (by law of total expectation):

$$\begin{aligned}&\sum_{k=1}^{N}\mathbb{E}\|P_{x^{k+1}}^{x^k}g^{k+1}-g^k\|_{x^k}^2\\ \le&4\sum_{k=1}^{N}\tau_k^2\mathbb{E}\|G_\mu^k-\mathsf{grad}f(x^k)\|_{x^k}^2+4\sum_{k=1}^{N}\tau_k^2\mathbb{E}\|\mathsf{grad}f(x^k)-g^k\|_{x^k}^2+\frac{2}{\beta^2}\sum_{k=1}^{N}\tau_k^2\mathbb{E}\|g^k\|_{x^k}^4\\ \le&4\sum_{k=1}^{N}\tau_k^2\tilde{\sigma}_k^2+4\sum_{k=1}^{N}\tau_k^2\mathbb{E}\|\mathsf{grad}f(x^k)-g^k\|_{x^k}^2+\frac{8}{\beta^2}\left(\frac{\mu^4L^4}{2}(d+12)^6+3d^2G^4\right)\sum_{k=1}^{N}\tau_k^2\end{aligned}$$

where the second inequality is by Lemmas 3.1 and 4.6. The desired result follows by applying Lemma 4.5 to the above inequality. □

We now state the main result in Theorem 4.2, as an analog to Theorem 3.1. Notice that different from Theorem 3.1, we do not need $N=\Omega(d)$ in case (ii), in view of Remark 3.2 and Assumption 4.2.

Theorem 4.2. *Suppose Assumptions 3.1, 4.1, 4.2, 4.3 and 4.4 hold. In Algorithm 2, we set $\mu=\mathcal{O}\big(\frac{1}{Ld^{3/2}N^{1/4}}\big)$ and $\beta\ge\sqrt{d}L$. Then the following holds.*

(i) If we choose $\tau_0=1$, $\tau_k=1/\sqrt{N}$, $k\ge1$ and $m_k\equiv 8(d+4)$, $k\ge0$, then we have $\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}\|\mathsf{grad}f(x^k)\|_{x^k}^2\le\mathcal{O}(1/\sqrt{N})$.

(ii) If we choose $\tau_0 = 1$, $\tau_k = 1/\sqrt{dN}$, $k \geq 1$, $m_0 = d$ *and* $m_k = 1$ *for* $k \geq 1$, *then we have* $\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}\|\mathsf{grad} f(x^k)\|_{x^k}^2 \leq \mathcal{O}(\sqrt{d/N})$.

Here the expectation $\mathbb{E}$ *is taken with respect to all random variables up to iteration* k, *which includes the random variables* u *in zeroth-order estimator* (1.3).

Proof. [Proof of Theorem 4.2] The proof is very similar to the proof of Theorem 3.1. We first will have the following inequality analogue to (3.11):

$$\begin{aligned}\frac{1}{8\beta^2}\sum_{k=0}^{N}\tau_k\mathbb{E}\|g^k\|_{x^k}^2 \leq & W^0 + \frac{1}{2\beta}\sum_{k=0}^{N}\tau_k\hat{\sigma}^2 + \frac{2}{\beta}\sum_{k=0}^{N}(\tau_k^2+\tau_k^3)\tilde{\sigma}_k^2 \\ & + \frac{1}{2\beta}[4\tilde{\sigma}_0^2 + 4\hat{\sigma}^2 + \frac{8}{\beta^2}(\frac{\mu^2L^2}{2}(d+12)^6 + 3d^2G^4)]\sum_{k=0}^{N}\tau_k^2\end{aligned}$$

Note that we still need (3.9a) to show the above inequality.

We then directly provide the result corresponding to (3.13):

$$\begin{aligned}\sum_{k=1}^{N}\frac{\tau_k}{2}\mathbb{E}\|\mathsf{grad} f(x^k)\|_{x^k}^2 \leq & (8\beta^2+16L^2)\Bigg(W^0 + \frac{1}{2\beta}\sum_{k=0}^{N}\tau_k\hat{\sigma}^2 + \frac{2}{\beta}\sum_{k=0}^{N}(\tau_k^2+\tau_k^3)\tilde{\sigma}_k^2 \\ & + \frac{1}{2\beta}[4\tilde{\sigma}_0^2 + 4\hat{\sigma}^2 + \frac{8}{\beta^2}(\frac{\mu^2L^2}{2}(d+12)^6 + 3d^2G^4)]\sum_{k=0}^{N}\tau_k^2\Bigg) + \sum_{k=0}^{N-1}\tau_k^2\tilde{\sigma}_k^2 + \sum_{k=0}^{N-1}\tau_k^2\hat{\sigma}^2 + \tilde{\sigma}_0^2\end{aligned} \tag{4.3}$$

Now by Assumption 4.2, we have $\tilde{\sigma}_k^2 \leq \sigma_k^2 + \frac{8(d+4)}{m_k}G^2$, which is exactly the reason we don't need to show an inequality similar to (3.9b).

For case (i) in Theorem 4.2, (4.3) can be rewritten as

$$\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}\|\mathsf{grad} f(x^k)\|_{x^k}^2 \leq \frac{c_1 W(x^0,g^0)}{\sqrt{N}} + c_2\hat{\sigma}^2 + \frac{c_3\frac{1}{N}\sum_{k=0}^{N}\tilde{\sigma}_k^2}{\sqrt{N}} + \frac{c_4}{\sqrt{N}}\tilde{\sigma}_0^2,$$

for some absolute positive constants c_1, c_2, c_3 and c_4. The proof for case (i) is completed by noting that (see (3.6)) $\hat{\sigma}^2 = \mathcal{O}(1/\sqrt{N})$, $\frac{1}{N}\sum_{k=0}^{N}\tilde{\sigma}_k^2 = \mathcal{O}(1)$ and $\tilde{\sigma}_0^2 = \mathcal{O}(1)$.

For case (ii) in Theorem 4.2, (4.3) can be rewritten as

$$\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}\|\mathsf{grad} f(x^k)\|_{x^k}^2 \leq c_1' W(x^0,g^0)\sqrt{\frac{d}{N}} + c_2'\hat{\sigma}^2 + \frac{c_3'\frac{1}{N}\sum_{k=0}^{N}\tilde{\sigma}_k^2}{\sqrt{dN}} + c_4'\sqrt{\frac{d}{N}}\tilde{\sigma}_0^2,$$

for some positive constants c_1', c_2', c_3' and c_4'. The proof of case (ii) is completed by noting that $\tilde{\sigma}_0^2 = \mathcal{O}(1)$, $\hat{\sigma}^2 = \mathcal{O}(1/\sqrt{N})$ and $\frac{1}{N}\sum_{k=0}^{N}\tilde{\sigma}_k^2 = \mathcal{O}(d)$. □

Remark 4.2. *By the technique discussed in Remark 3.1, to obtain an* ϵ*-approximate stationary point in Definition 2.2 we need an oracle complexity of* $\mathcal{O}(d/\epsilon^4)$.

5 Numerical experiments

5.1 k-PCA

We now provide numerical results on the k-PCA problem to demonstrate the effectiveness of the Zo-RASA algorithms. For a given centered random vector $\mathbf{z} \in \mathbb{R}^n$, the k-PCA problem corresponds

to finding the subspace spanned by the top-k eigenvectors of its positive definite covariance matrix $\Sigma = \mathbb{E}[\mathbf{z}\mathbf{z}^\top]$. Formally, we have the following problem on the Stiefel manifold:

$$\min_{X \in \mathrm{St}(n,r)} f(X) := -\frac{1}{2} \operatorname{tr}(X^\top \mathbb{E}[\mathbf{z}\mathbf{z}^\top] X). \tag{5.1}$$

Note that the dimension of the Stiefel is given by $d = nr - r(r+1)/2$.

For any $Y = XQ$ where $Q \in \mathbb{R}^{r \times r}$, and $Q^\top Q = QQ^\top = I_r$, we have $f(X) = f(Y)$. Hence, we can equivalently view (5.1) as the following minimization problem on the Grassmann manifold:

$$\min_{[X] \in \mathrm{Gr}(n,r)} f([X]) := -\frac{1}{2} \operatorname{tr}(X^\top \mathbb{E}[\mathbf{z}\mathbf{z}^\top] X).$$

Note that the dimension of the Grassmannian is given by $d = r(n-r)$.

We solve (5.1) using Algorithm 2 and compare it with the zeroth-order Riemannian SGD method from Li et al. (2022). In all the experiments, we used projecting vector transport rather than parallel transport for Stiefel manifolds, due to the aforementioned facts that parallel transport is time-consuming to numerically compute on Stiefel manifold, and has no closed form. In the stochastic zeroth-order setting, for each query point X_k, the stochastic oracle returns a noise estimate of $f(x)$ based on a single observation $\mathbf{z}_k$, i.e. $F(X^k; \mathbf{z}_k) = -1/2 \operatorname{tr}((X^k)^\top \mathbf{z}_k \mathbf{z}_k^\top X^k)$. For our experiments, we assume $\mathbf{z}_k$ is sampled from a centered Gaussian distribution with covariance matrix given by $\Sigma = \sum_{i=1}^{r} \lambda_i v_i v_i^\top + \sum_{i=r+1}^{n} \lambda_i v_i v_i^\top$, where $V = [v_1, ..., v_n]$ is an orthogonal matrix. The first r λ_is are uniform random numbers in $[100, 200]$ and the last $n-r$ are uniform random numbers in $[1, 50]$. For our experiments, we fix r and try different n (reflected in different rows in Figure 1).

We set $N = 50000 \times n$ for `Zo-RASA` and one-batch Zo-RSGD (`Zo-RSGD-1`) algorithms, while $N = 50000$ for our mini-batch Zo-RSGD algorithm (`Zo-RSGD-m`). The reason here is that for `Zo-RSGD-m`, we take $m = n = \mathcal{O}(d)$ since we fix r and change n. While the theoretical result in Li et al. (2022) requires the batch-size m to be $\mathcal{O}(d/\epsilon^2)$, they empirically observed reasonable-order batch-sizes suffices. For `Zo-RASA`, according to our theory, we again take $\tau_k = 0.01/\sqrt{N}$ and $\beta = 100$. For `Zo-RSGD-1` and `Zo-RSGD-m`, we set t_k as $t_k = 10^{-4}/\sqrt{N}$ and $t_k = 5 \times 10^{-4}/\sqrt{N}$ respectively.

For all algorithms, we again compare the function value, norm of the Riemannian gradient and the principal angles between the current iterate and the optimal subspace. Figures 1 plots the results. The experimental results provide support for the proposed algorithms (and the established theory), demonstrating that the proposed Zo-RASA algorithm is more efficient in terms of decreasing the Riemannian gradient and principal angles compared to conventional zeroth-order Riemannian stochastic gradient descent methods that utilize mini-batches.

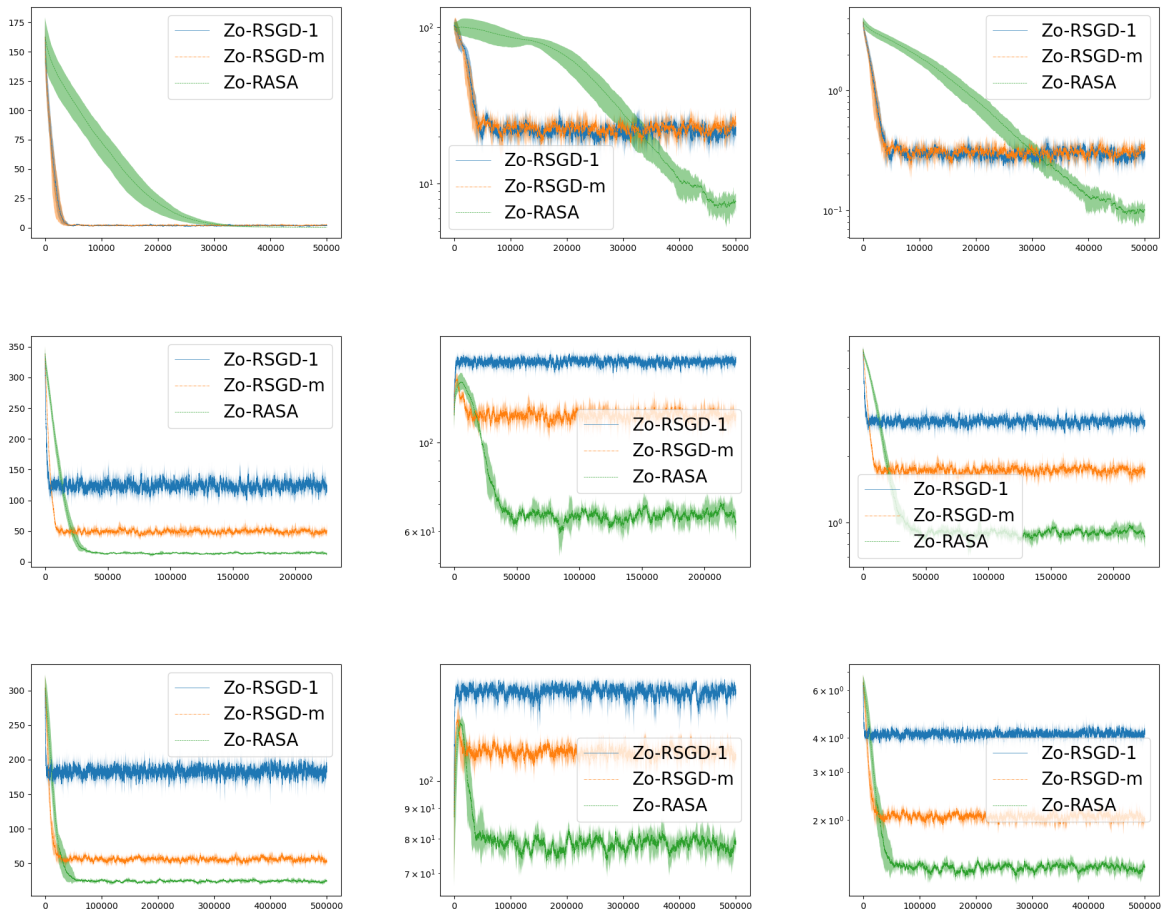
5.2 Identification of a fixed rank symmetric positive semi-definite matrix

We now provide another numerical example from Bonnabel (2013). Consider a matrix-version linear model as in Tsuda et al. (2005):

$$y_t = \operatorname{tr}(W \mathbf{x}_t \mathbf{x}_t^\top) = \mathbf{x}_t^\top W \mathbf{x}_t$$

where $\mathbf{x}_t \in \mathbb{R}^n$ is the input and $y_t \in \mathbb{R}$ is the output, and the unknown matrix $W \in \mathbb{R}^{n \times n}$ is a positive semi-definite matrix with a fixed rank r ($r \leq n$). Denote the set

$$S_+(n,r) = \{W \in \mathbb{R}^{n \times n} | W = W^\top, \operatorname{rank}(W) = r\} \tag{5.2}$$



(a) Optimality gap (b) $\|\mathsf{grad} f(X^t)\|$ (c) Principal angles

Figure 1: Results for kPCA (5.1) with $n \in \{10, 30, 50\}$ (corresponding to three rows) and $r = 5$. The resulting manifold (Stiefel) dimensions are $d = \{35, 135, 235\}$. The x-axis is the number of zeroth-order oracle calls (i.e. number of function value calls).

which is the set of positive definite matrices with rank r. The problem is thus formulated as a matrix least square problem

$$\min_{W \in S_+(n,r)} f(W) := \frac{1}{2}\mathbb{E}_{\mathbf{x},y}(\mathbf{x}^\top W \mathbf{x} - y)^2 \tag{5.3}$$

Notice that W can be represented as $W = GG^\top$ where $G \in \mathbb{R}^{n,r}$ is a matrix with full column rank. Also notice that for any orthogonal matrix $O \in \mathbb{R}^{r\times r}$ we have $W = GOO^\top G^\top = GG^\top$, we have the following quotient representation of the set of fixed rank positive definite matrices $S_+(n,r) \simeq \mathbb{R}_*^{n\times r}/\mathcal{O}(r)$, where the right hand side represents the set of equivalent classes:

$$[G] = \{GO | O \in \mathcal{O}(r)\}.$$

We could thus conduct our experiment on the quotient manifold $\mathbb{R}_*^{n\times r}/\mathcal{O}(r)$, with the following

re-formulated problem:

$$\min_{[G]\in\mathbb{R}_*^{n\times r}/\mathcal{O}(r)} f(G) := \frac{1}{2}\mathbb{E}_{\mathbf{x},y}(\mathbf{x}^\top GG^\top\mathbf{x} - y)^2 \tag{5.4}$$

The manifold $S_+(n,r)$ has dimension $d = nr - r(r-1)/2$ and is not a compact manifold. We test (5.4) to show the efficiency of our proposed algorithm even without the compactness assumption (Assumption 4.2) which we need to conduct our theoretical analysis.

We solve (5.4) using Algorithm 2 and compare it with the zeroth-order Riemannian SGD method from Li et al. (2022). In all the experiments, we used again retraction and projecting vector transport rather than exponential mapping and parallel transport. The ground-truth $G^\star \in \mathbb{R}^{n\times r}$ is sampled randomly with standard Gaussian entries. For our experiments, we sample $\mathbf{x} \sim \mathcal{N}(0, I_d)$ and construct $y = \mathbf{x}^\top W\mathbf{x}$ noiselessly. Specifically, given a query point G^t and a Gaussian sample $\mathbf{x}_t$ with $y_t = \mathbf{x}_t^\top G^\star(G^\star)^\top\mathbf{x}_t$, the stochastic zeroth-order oracle gives the value $\frac{1}{2}(\mathbf{x}_t^\top G^t(G^t)^\top\mathbf{x}_t - y_t)^2$. For our experiments, we fix r and test with different n (reflected in different rows in Figure 2).

We set $N = 5000 \times n$ for `Zo-RASA` and one-batch Zo-RSGD (`Zo-RSGD-1`) algorithms, while $N = 5000$ for our mini-batch Zo-RSGD algorithm (`Zo-RSGD-m`) for the same reason as the kPCA experiments. For `Zo-RASA`, according again to our theory, we again take $\tau_k = 10^{-3}/\sqrt{N}$ and $\beta = 100$. For `Zo-RSGD-1` and `Zo-RSGD-m`, we set $t_k = 10^{-5}/\sqrt{N}$.

For all algorithms, we again compare the function value, norm of the Riemannian gradient and the quantity $\|G^t(G^t)^\top - G^\star(G^\star)^\top\|$ which measures the error to the ground truth positive semi-definite matrix. Figures 2 plots the results. It's worth noticing here that mini-batch Zo-RSGD seems to work the worst in the plots, which is due to the fact that we take the step sizes the same for `Zo-RSGD-1` and `Zo-RSGD-m`. The reason we cannot enlarge the step size for `Zo-RSGD-m` is that the projectional retraction and projectional vector transport requires solving a Sylvester equation which leads to numerical stability issues if the step sizes become large (see Boumal et al. (2014) for details). The experimental results provide support for the proposed algorithms (and the established theory), demonstrating that the proposed Zo-RASA algorithm is more efficient in terms of decreasing the Riemannian gradient and function values compared to conventional zeroth-order Riemannian stochastic gradient descent methods that utilize mini-batches.

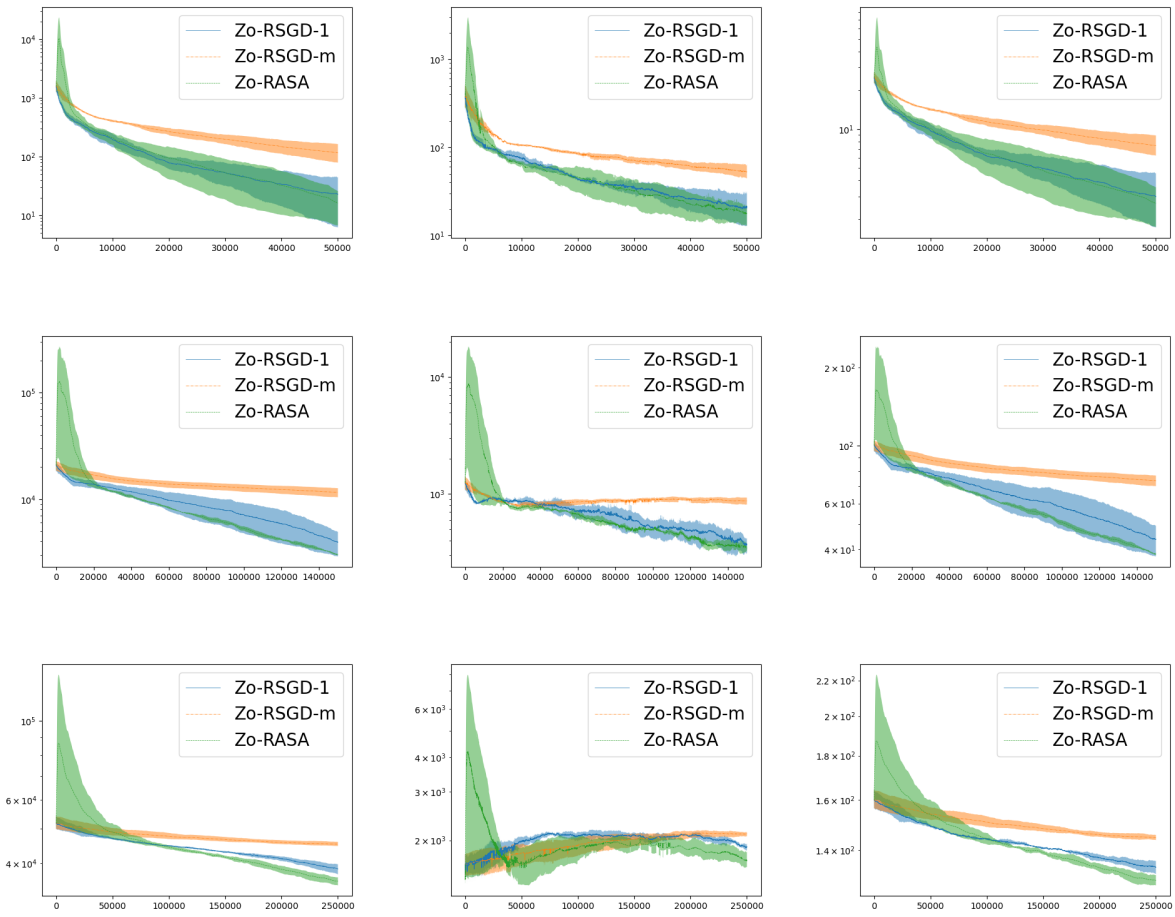
Acknowledgements

We thank Prof. Otis Chodosh (Stanford) for several helpful discussions and clarifications regarding several differential geometric concepts. JL thanks Xuxing Chen for helpful discussions. KB was supported in part by National Science Foundation (NSF) grant DMS-2053918. SM was supported in part by NSF grants DMS-2243650, CCF-2308597, CCF-2311275 and ECCS-2326591, UC Davis CeDAR (Center for Data Science and Artificial Intelligence Research) Innovative Data Science Seed Funding Program, and a startup fund from Rice University.

References

P.-A. Absil and J. Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012. (Cited on pages 16 and 19.)

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008. (Cited on pages 3, 4, 5, 6, 17, and 18.)



(a) Optimality gap (b) $\|\mathrm{grad} f(G^t)\|$ (c) $\|G^t(G^t)^\top - G^\star(G^\star)^\top\|$

Figure 2: Results for (5.4) with $n \in \{10, 30, 50\}$ (corresponding to three rows) and $r = 5$. The resulting manifold as defined in (5.2) are $d = \{40, 140, 240\}$ dimensional, respectively. The x-axis is the number of zeroth-order oracle calls (i.e. number of function value calls).

K. Balasubramanian, S. Ghadimi, and A. Nguyen. Stochastic multilevel composition optimization algorithms with level-independent convergence rates. *SIAM Journal on Optimization*, 32(2): 519–544, 2022. (Cited on page 3.)

S. Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013. (Cited on page 24.)

N. Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023. (Cited on pages 3, 4, 5, 6, and 7.)

N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15(42):1455–1459, 2014. URL https://www.manopt.org. (Cited on page 26.)

D. Burago, Y. Burago, and S. Ivanov. *A course in metric geometry*, volume 33. American Mathematical Society, 2022. (Cited on page 19.)

S. Chen, S. Ma, A. M.-C. So, and T. Zhang. Proximal gradient method for nonsmooth optimization over the Stiefel manifold. *SIAM Journal on Optimization*, 30(1):210–239, 2020. (Cited on pages 3 and 5.)

M. P. Do Carmo. *Riemannian geometry*, volume 6. Springer, 1992. (Cited on pages 4, 6, and 17.)

A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. (Cited on pages 17 and 18.)

K. A. Gallivan and P. Absil. Note on the convex hull of the Stiefel manifold. *Technical note*, 2010. (Cited on page 16.)

E. S. Gawlik and M. Leok. High-order retractions on matrix manifolds using projected polynomials. *SIAM Journal on Matrix Analysis and Applications*, 39(2):801–828, 2018. (Cited on page 19.)

S. Ghadimi and W. B. Powell. Stochastic search for a parametric cost function approximation: Energy storage with rolling forecasts. *arXiv preprint arXiv:2204.07317*, 2022. (Cited on page 4.)

S. Ghadimi, A. Ruszczynski, and M. Wang. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM Journal on Optimization*, 30(1):960–979, 2020. (Cited on pages 3, 4, and 6.)

A. Han and J. Gao. Riemannian stochastic recursive momentum method for non-convex optimization. In *IJCAI*, 2020. (Cited on page 4.)

W. Huang and K. Wei. Riemannian proximal gradient methods. *Mathematical Programming*, 194 (1-2):371–413, 2022. (Cited on page 3.)

W. Huang, K. A. Gallivan, and P.-A. Absil. A Broyden class of quasi-Newton methods for Riemannian optimization. *SIAM Journal on Optimization*, 25(3):1660–1685, 2015. (Cited on pages 3 and 15.)

G. Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer Nature, 2020. (Cited on page 7.)

J. Li, K. Balasubramanian, and S. Ma. Stochastic zeroth-order Riemannian derivative estimation and optimization. *Mathematics of Operations Research*, 48(2):1183–1211, 2022. (Cited on pages 1, 2, 3, 7, 8, 20, 24, and 26.)

A. I. Maass, C. Manzie, D. Nesic, J. H. Manton, and I. Shames. Tracking and regret bounds for online zeroth-order Euclidean and Riemannian optimization. *SIAM Journal on Optimization*, 32 (2):445–469, 2022. (Cited on pages 2 and 3.)

J. W. Milnor and J. D. Stasheff. *Characteristic classes*. Number 76. Princeton university press, 1974. (Cited on page 19.)

Y. Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018. (Cited on page 7.)

B. T. Polyak. Comparison of the convergence rates for single-step and multi-step optimization algorithms in the presence of noise. *Engineering Cybernetics*, 15(1):6–10, 1977. (Cited on page 4.)

A. Ruszczyński. A linearization method for nonsmooth stochastic programming problems. *Mathematics of Operations Research*, 12(1):32–49, 1987. (Cited on pages 3 and 6.)

A. Ruszczynski. A stochastic subgradient method for nonsmooth nonconvex multilevel composition optimization. *SIAM Journal on Control and Optimization*, 59(3):2301–2320, 2021. (Cited on page 3.)

A. Ruszczynski and W. Syski. Stochastic approximation method with gradient averaging for unconstrained problems. *IEEE Transactions on Automatic Control*, 28(12):1097–1105, 1983. (Cited on pages 3, 4, and 6.)

H. Sato. Riemannian conjugate gradient methods: General framework and specific algorithms with convergence analyses. *SIAM Journal on Optimization*, 32(4):2690–2717, 2022. (Cited on pages 3 and 15.)

H. Sato, H. Kasai, and B. Mishra. Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport. *SIAM Journal on Optimization*, 29(2):1444–1472, 2019. (Cited on page 3.)

K. Scheinberg. Finite difference gradient approximation: To randomize or not? *INFORMS Journal on Computing*, 34(5):2384–2388, 2022. (Cited on page 13.)

K. Tsuda, G. Rätsch, and M. K. Warmuth. Matrix exponentiated gradient updates for on-line learning and Bregman projection. *Journal of Machine Learning Research*, 6(Jun):995–1018, 2005. (Cited on page 24.)

L. W. Tu. *An Introduction to Manifolds*. Springer, 2011. (Cited on page 4.)

T. Wang. On sharp stochastic zeroth-order Hessian estimators over Riemannian manifolds. *Information and Inference: A Journal of the IMA*, 12(2):787–813, 2023. (Cited on pages 2 and 3.)

T. Wang, Y. Huang, and D. Li. From the Greene–Wu convolution to gradient estimation over Riemannian manifolds. *arXiv:2108.07406*, 2021. (Cited on pages 2 and 3.)

X. Wang, Z. Tu, Y. Hong, Y. Wu, and G. Shi. Online optimization over Riemannian manifolds. *Journal of Machine Learning Research*, 24(84):1–67, 2023. (Cited on page 2.)

L. Xiao. Dual averaging method for regularized stochastic learning and online optimization. *Advances in Neural Information Processing Systems*, 22, 2009. (Cited on page 4.)

H. Zhang, S. J Reddi, and S. Sra. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. *NeurIPS*, 29, 2016. (Cited on page 7.)