MATHEMATICS OF OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0364-765X | EISSN 1526-5471 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Stochastic Zeroth-order Riemannian Derivative Estimation and Optimization

Jiaxiang Li
Department of Mathematics, University of California, Davis, CA, jxjli@ucdavis.edu

Krishnakumar Balasubramanian
Department of Statistics, University of California, Davis, CA, kbala@ucdavis.edu

Shiqian Ma
Department of Mathematics, University of California, Davis, CA, sqma@ucdavis.edu

We consider stochastic zeroth-order optimization over Riemannian submanifolds embedded in Euclidean space, where the task is to solve Riemannian optimization problem with only noisy objective function evaluations. Towards this, our main contribution is to propose estimators of the Riemannian gradient and Hessian from noisy objective function evaluations, based on a Riemannian version of the Gaussian smoothing technique. The proposed estimators overcome the difficulty of non-linearity of the manifold constraint and issues that arise in using Euclidean Gaussian smoothing techniques when the function is defined only over the manifold. We use the proposed estimators to solve Riemannian optimization problems in the following settings for the objective function: (i) stochastic and gradient-Lipschitz (in both nonconvex and geodesic convex settings), (ii) sum of gradient-Lipschitz and non-smooth functions, and (iii) Hessian-Lipschitz. For these settings, we analyze the oracle complexity of our algorithms to obtain appropriately defined notions of ϵ-stationary point or ϵ-approximate local minimizer. Notably, our complexities are independent of the dimension of the ambient Euclidean space and depend only on the intrinsic dimension of the manifold under consideration. We demonstrate the applicability of our algorithms by simulation results and real-world applications on black-box stiffness control for robotics and black-box attacks to neural network

Key words: Riemannian optimization, manifold optimization, zeroth-order optimization, black-box attack
MSC2000 subject classification: 90C30

1. Introduction

Consider the following Riemannian optimization problem:

$$\min\ f(x)+h(x),\ \text{s.t.},\ x\in\mathcal{M}, \tag{1}$$

where $\mathcal{M}$ is a Riemannian submanifold embedded in $\mathbb{R}^n$, $f:\mathcal{M}\to\mathbb{R}$ is a smooth and possibly nonconvex function, and $h:\mathbb{R}^n\to\mathbb{R}$ is a convex and nonsmooth function. Here, convexity and smoothness are interpreted as the function is being considered in the ambient Euclidean space. Iterative algorithms for solving (1) usually require the gradient and Hessian information of the objective function. However, in many applications, the analytical form of the function f (or h) and its gradient are not available, and we can only obtain noisy function evaluations via a zeroth-order oracle. This setting, termed as *stochastic zeroth-order Riemannian optimization*, generalizes stochastic zeroth-order Euclidean optimization (i.e., when $\mathcal{M}\equiv\mathbb{R}^n$ in (1)), a topic which goes back to the early works of [56, 62, 63] in the 1960's; see also [24, 3, 45] for recent books and surveys.

In the Euclidean setting, two popular techniques for estimating the gradient from (noisy) function queries include the finite-differences method [73] and the Gaussian smoothing techniques [63]. Earlier works in this setting focused on using the estimated gradient to obtain asymptotic convergence rates of iterative optimization algorithms. Recently, obtaining non-asymptotic guarantees on the oracle complexity of stochastic zeroth-order optimization has been of great interest. Towards that, [64, 66] analyzed the Gaussian smoothing technique for estimating the Euclidean gradient from noisy function evaluations and proved that for unconstrained convex minimization, one needs $O(n^2/\epsilon^2)$ noisy function evaluations to obtain an ϵ-optimal solution. This complexity was improved by [31] to $O(n/\epsilon^2)$ when the objective function is further assumed to be gradient-smooth. Note that this oracle complexity depends linearly on the problem dimension n and it was proved that the linear dependency on n is unavoidable [37, 28]. Nonconvex and smooth setting was also considered in [31]. In particular, now assuming $h \equiv 0$ and $\mathcal{M} \equiv \mathbb{R}^n$ in (1), it was shown that the number of function evaluations for obtaining an ϵ-stationary point $\bar{x}$ (i.e., $\mathbb{E}\|\nabla f(\bar{x})\| \leq \epsilon$), is $O(n/\epsilon^4)$.

In the Riemannian manifold setting, however, a main challenge in designing and analyzing zeroth-order algorithms is the lack of availability of theoretically sound methods to estimate the Riemannian gradients and Hessians from (noisy) function evaluations. To this end, our main contribution in this work is to construct estimators of the Riemannian gradient and Hessian from noisy function evaluations, based on a modified Gaussian smoothing technique from [66] and [4]. The main difficulty addressed here is that the gradient and Hessian estimator in [66] and [4] respectively, require computing $f(x+\nu u)$, for some parameter $\nu > 0$ and an n-dimensional standard Gaussian vector u. However, the point $x+\nu u$ may not necessarily lie on the manifold $\mathcal{M}$. To resolve this issue, we propose an estimator based on the smoothing technique and sampling Gaussian random vectors on the tangent space of the manifold $\mathcal{M}$. We then use the developed methodology to design stochastic zeroth-order algorithms for solving (1) with oracle complexities that depend only on the manifold dimension d, and independent of the ambient Euclidean dimension n.

1.1. Related Works

As mentioned previously, zeroth-order optimization has a long history; we refer the reader to [24, 3, 45] for more details. The oracle complexity of methods from the above works are at least linear in terms of their dependence on dimensionality. Recent works in this field have been focusing on stochastic zeroth-order optimization in high-dimensions [84, 32, 4, 14]. Assuming a sparse structure (for example, the function being optimized depends only on s of the n coordinates), the above works have shown that the oracle complexity of zeroth-order optimization depends only poly-logarithmically on the dimension n, and it has a linear dependency only on the sparsity parameter s, which is typically small compared to n in several applications. Compared to these works, we assume a manifold structure on the function being optimized and obtain oracle complexities that depend only on the manifold dimension and independent of the ambient Euclidean dimension.

Apart from the above, *Bayesian optimization* is yet another popular class of methods for optimizing functions based on noisy function values. This approach aims at finding the global minimizer by enforcing a Gaussian process prior on the space of function being optimized and using Bayesian sampling techniques. We refer the reader to [59, 60, 71, 30] for an overview of such techniques in the Euclidean settings and their applications to a variety of fields including robotics, recommender systems, preference learning and hyperparameter tuning. A common limitation of the above algorithms is that they are usually not scalable well to solve high-dimensional problems. Recent developments on Bayesian optimization for high-dimensional problems include [48, 85, 61, 70, 83] where people considered zeroth-order optimization with structured functions (for example, sparse or additive functions), and developed Bayesian optimization algorithms and related analysis. Very recently, [67, 39, 38] considered heuristic Bayesian optimization algorithms for function defined over non-Euclidean domains, including Riemannian domains, without any theoretical analysis.

Riemannian optimization in the first or second-order setting has drawn a lot of attention recently due to its applications in various fields, including low-rank matrix completion [11, 81], phase retrieval [5, 76], dictionary learning [23, 75], dimensionality reduction [35, 79, 58] and manifold regression [51, 50]. For smooth Riemannian optimization, i.e., $h \equiv 0$ in (1), it was shown that Riemannian gradient descent method require $\mathcal{O}(1/\epsilon^2)$ iterations to converge to an ϵ-stationary point [12]. Stochastic algorithms were also studied for smooth Riemannian optimization [9, 96, 87, 92, 42, 96, 87]. In particular, using the SPIDER variance reduction technique, [96] proved that $\mathcal{O}(1/\epsilon^3)$ oracle calls are required to obtain an ϵ-stationary point in expectation. When the function f takes a finite-sum structure, the Riemannian SVRG [92] achieves ϵ-stationary solution with $\mathcal{O}(k^{2/3}/\epsilon^2)$ oracle calls where k is number of summands. When the nonsmooth function h presents in (1), Riemannian sub-gradient methods (RSGM) are widely used [10, 49] and they require $\mathcal{O}(1/\epsilon^4)$ iterations. ADMM for solving (1) has also been studied [43, 44], but they usually lack convergence guarantee, while the analysis presented in [94] requires some strong assumptions. The recently proposed manifold proximal gradient method (ManPG) [21] for solving (1) requires $\mathcal{O}(1/\epsilon^2)$ number of iterations to find an ϵ-stationary solution. Variants of ManPG such as ManPPA [20], ManPL [86] and stochastic ManPG [82] have also been studied. Note that none of these works considers the zeroth-order setting. Recently, there are some attempts on stochastic zeroth-order Riemannian optimization [17, 29], but they are mostly heuristics and do not have any rigorous convergence guarantees.

1.2. Motivating Applications

Our motivation for developing a theoretical framework for stochastic zeroth-order Riemannian optimization is due to several important emerging applications; see, e.g., [17, 55, 91, 39, 41]. Below, we discuss two concrete examples, which we will revisit in Section 4.2, to illustrate the applicability of the methods developed in this work. We also briefly discuss a third application in topological data analysis, and numerical experiments on this application will be conducted in a future work, as it is more involved and beyond the scope of this paper.

1.2.1. Black-box Stiffness Control for Robotics

Our first motivating application is from the field of robotics. It has become increasingly common to use zeroth-order optimization techniques to optimize control parameter and policies in robotics [54, 27, 91]. This is because that the cost functions being optimized in robotics are not available in a closed form as a function of the control parameter. Invariably for a given choice of control parameter, the cost function needs to be evaluated through a real-world experiment on a given robot or through simulation. Recently, domain knowledge has been used as constraints on the control parameter space, among which a common choice is the geometry-aware constraint. For example, control parameters like stiffness, inertia and manipulability lie on the positive semidefinite manifold, orthogonal group and unit sphere, respectively. Hence, there is a need to develop zeroth-order optimization methods over the manifolds to optimize the above mentioned control parameters [39].

1.2.2. Zeroth-order Attacks on Deep Neural Networks (DNNs)

Our second motivating application is based on developing black-box attacks to DNNs. Despite the recent success of DNNs, studies have shown that they are vulnerable to adversarial attacks: even a well-trained DNN could completely misclassify a slightly perturbed version of the original image (which is undetectable to the human eyes); see, e.g., [78, 33]. As a result, it is extremely important on the one hand to come up with methods to train DNNs that are robust to adversarial attacks, and on the other hand to develop efficient attacks on DNNs with the goal being to make them misclassify. In practice, as the architecture of the DNN is not known to the attacker, several works, for example, [19, 80, 22], use zeroth-order optimization algorithms for designing adversarial attacks. However, existing works have an inherent drawback– the perturbed training example designed to fool the DNN is not in the same domain as the original training data. For example, despite the fact that natural images typically lie on a manifold [88], the perturbations are not constrained to lie on the same manifold. This naturally motivates us to use zeroth-order Riemannian optimization methods to design adversarial examples to fool DNNs, which at the same time preserves the manifold structures in the training data.

1.2.3. Black-box Methods for Topological Dimension Reduction The third motivating example is from the field of dimension reduction, a popular class of techniques for reducing the dimension of high-dimensional unstructured data for feature extraction and visualization. There exists a variety of methods for this task; we refer the interested reader to [46, 13] for more details. However, a majority of the existing techniques are based on *geometric* motivations. Recently, there has been a growing literature on using *topological* information for performing data analysis [18, 57, 68]. One such method is a dimension reduction technique called Persistent Homology-Based Projection Pursuit [41]. Roughly speaking, given a point-cloud data set with cardinality n and dimension m (i.e., a matrix $X \in \mathbb{R}^{m\times n}$), persistence homology refers to developing a multi-scale characterization of topologically invariant features available in the data. Such information is summarized in terms of the so-called persistence diagram, $D(X)$, which is a multiset of points in a two-dimensional plane. The idea in [41] is to obtain a transformation $P^\top \in \mathbb{R}^{p\times m}$, with $p \ll m$, such that the topological summaries of the original dataset X and the reduced dimensional dataset $P^\top X$ are close to each other; that is, the persistence diagram $D(X)$ and $D(P^\top X)$ are close in the 2-Wasserstein distance. The problem is then formulated as (informally speaking),

$$\min_{\{P\in\mathbb{R}^{m\times p}: P^\top P = I\}} W_2(D(X), D(P^\top X)),$$

which is an optimization problem over the Stiefel manifold (see Section 2.1 for more details). It turns out that calculating the gradient of the above objective function is highly non-trivial and computationally expensive [47]. However, evaluating the objecting function for various value of the matrix P is relatively less expensive. Hence, this serves as yet another problem in which the methodology developed in this work could be applied naturally.

1.3. Main Contributions We now summarize our main contributions.

1. In Section 2, we propose the (stochastic) zeroth-order Riemannian gradient (6) and Hessian (15) estimators, which addresses the infeasibility issue of the sampling for the case of derivative-free optimization over manifolds.
2. In Section 3, we demonstrate the applicability of the developed estimators for stochastic zeroth-order Riemannian optimization, as listed below. A summary of these results is given in Table 1. To the best of our knowledge, our results are the first complexity results for stochastic zeroth-order Riemannian optimization.
   - When $h(x) \equiv 0$ and the exact function evaluations of $f(x)$ are obtainable, we propose a zeroth-order Riemannian gradient descent method (`ZO-RGD`) and provide its oracle complexity for obtaining an ϵ-stationary point of (1) (see Theorem 1).
   - When $h(x) \equiv 0$ and $f(x) = \mathbb{E}_\xi[F(x,\xi)]$, we propose a zeroth-order Riemannian stochastic gradient descent method (`ZO-RSGD`). We analyze its oracle complexity under two different settings (see Theorems 2 and 5).
   - When $h(x)$ is convex and nonsmooth, we propose a zeroth-order stochastic Riemannian proximal gradient method (`ZO-SManPG`) and provide its oracle complexity for obtaining an ϵ-stationary point of (1) (see Theorem 3).
   - When $h(x) \equiv 0$ and $f(x) = \mathbb{E}_\xi[F(x,\xi)]$, where $F(x,\xi)$ satisfies certain Lipschitz Riemannian Hessian property, we propose a zeroth-order Riemannian stochastic cubic regularized Newton's method (`ZO-RSCRN`) that provably converges to an ϵ-approximate local minimizer (see Theorem 4).
3. In Section 4, we provide experimental results on simulated data to quantify the performance of our methods. We then demonstrate the applicability of our methods to the problem of black-box attacks to deep neural networks and robotics.

Algorithm	Structure	Iteration Complexity	Oracle Complexity
ZO-RGD	smooth	$\mathcal{O}(d/\epsilon^2)$	$\mathcal{O}(d/\epsilon^2)$
ZO-RSGD	smooth, stochastic	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}(d/\epsilon^4)$
ZO-RSGD	smooth, stochastic, Geo-convex	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}(d/\epsilon^2)$
ZO-SManPG	nonsmooth stochastic	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}(d/\epsilon^4)$
ZO-RSCRN	Lipschitz Hessian stochastic	$\mathcal{O}(1/\epsilon^{1.5})$	$\mathcal{O}(d/\epsilon^{3.5} + d^4/\epsilon^{2.5})$

Table 1. Summary of the convergence results proved in this paper. For all but the ZO-RSCRN algorithm, the reported complexities correspond to ϵ-stationary solution; for the ZO-RSCRN algorithm the complexities correspond to ϵ-local minimizers. Here, d is the intrinsic dimension of the manifold $\mathcal{M}$. Furthermore, Iteration complexity refers to the number of iterations and oracle complexity refers to the number of calls to the (stochastic) zeroth-order oracle.

2. Preliminaries and Methodology

We start this section with a brief review of basics of Riemannian optimization. We then introduce our stochastic zeroth-order Riemannian gradient and Hessian estimators, and provide bias and moment bounds quantifying the accuracy of the proposed estimators, which will be useful for the convergence analysis later.

2.1. Basics of Riemannian Optimization

Let $\mathcal{M} \subset \mathbb{R}^n$ be a differentiable embedded submanifold. We have the following definition for the tangent space.

Definition 1 (Tangent space). Consider a manifold $\mathcal{M}$ embedded in a Euclidean space. For any $x \in \mathcal{M}$, the tangent space $T_x\mathcal{M}$ at x is a linear subspace that consists of the derivatives of all differentiable curves on $\mathcal{M}$ passing through x:

$$T_x\mathcal{M} = \{\gamma'(0) : \gamma(0) = x, \gamma([-\delta,\delta]) \subset \mathcal{M} \text{ for some } \delta > 0, \gamma \text{ is differentiable}\}. \tag{2}$$

The manifold $\mathcal{M}$ is a Riemannian manifold if it is equipped with an inner product on the tangent space, $\langle \cdot,\cdot\rangle_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$, that varies smoothly on $\mathcal{M}$. We also introduce the concept of the dimension of a manifold.

Definition 2 (Dimension of a manifold [1]). The dimension of the manifold $\mathcal{M}$, denoted as d, is the dimension of the Euclidean space that the manifold is locally homeomorphic to. In particular, the dimension of the tangent space is always equal to the dimension of the manifold.
As an example, consider the Stiefel manifold $\mathcal{M} = \mathrm{St}(n,p) := \{X \in \mathbb{R}^{n\times p} : X^\top X = I_p\}$. The tangent space of $\mathrm{St}(n,p)$ is given by $T_X\mathcal{M} = \{Y \in \mathbb{R}^{n\times p} : X^\top Y + Y^\top X = 0\}$. Hence, the dimension of the Stiefel manifold is $np - \frac{1}{2}p(p+1)$. Note that the dimension of the manifold could be significantly less than the dimension of the ambient Euclidean space. Yet another example is the manifold of low-rank matrices [81]. We now introduce the concept of a Riemannian gradient.

Definition 3 (Riemannian Gradient). Suppose f is a smooth function on $\mathcal{M}$. The Riemannian gradient $\mathrm{grad} f(x)$ is a vector in $T_x\mathcal{M}$ satisfying $\frac{d(f(\gamma(t)))}{dt}\Big|_{t=0} = \langle v, \mathrm{grad} f(x)\rangle_x$ for any $v \in T_x\mathcal{M}$, where $\gamma(t)$ is a curve as described in (2).

Recall that in the Euclidean setting, a function $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth, if it satisfies $|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \le \frac{L}{2}\|x-y\|^2$, for all $x, y \in \mathbb{R}^n$. We now present the Riemannian counterpart of L-smooth functions. To do so, first we need the definition of retraction for a given $x \in \mathcal{M}$.

Definition 4 (Retraction). A retraction mapping R_x is a smooth mapping from $T_x\mathcal{M}$ to $\mathcal{M}$ such that: $R_x(0) = x$, where 0 is the zero element of $T_x\mathcal{M}$, and the differential of R_x at 0 is an identity mapping, i.e., $\frac{dR_x(t\eta)}{dt}\Big|_{t=0} = \eta$, $\forall \eta \in T_x\mathcal{M}$. In particular, the exponential mapping Exp_x is a retraction that generates geodesics.

Assumption 1 (**L-retraction-smoothness**). *There exists $L_g \geq 0$ such that the following inequality holds for function f in* (1)*:*

$$|f(R_x(\eta)) - f(x) - \langle \mathrm{grad} f(x), \eta \rangle_x| \leq \frac{L_g}{2}\|\eta\|^2, \forall x \in \mathcal{M}, \eta \in T_x\mathcal{M}. \tag{3}$$

Assumption 1 is also known as the restricted Lipschitz-type gradient for pullback function $\hat{f}_x(\eta) := f(R_x(\eta))$ [12]. The condition required in [12] is weaker because it only requires Eq. (3) to hold for $\|\eta\|_x \leq \rho_x$, where constant $\rho_x > 0$. In our convergence analysis, we need this assumption to be held for all $\eta \in T_x\mathcal{M}$, i.e., $\rho_x = \infty$. This assumption is satisfied when the manifold $\mathcal{M}$ is a compact submanifold of $\mathbb{R}^n$, the retraction R_x is globally defined[1] and function f is L-smooth in the Euclidean sense; we refer the reader to [12] for more details. We also emphasize that Assumption 1 is weaker than the geodesic smoothness assumption defined in [93]. The geodesic smoothness states that, $\forall \eta \in \mathcal{M}$, $f(\mathrm{Exp}_x(\eta)) \leq f(x) + \langle g_x, \eta \rangle_x + L_g d^2(x, \mathrm{Exp}_x(\eta))/2$, where g_x is a subgradient of f, $d(\cdot,\cdot)$ represents the geodesic distance. Such a condition is stronger than our Assumption 1, in the sense that, if the retraction is the exponential mapping, then geodesic smoothness implies the L-retraction-smoothness with the same parameter L_g [6].

Throughout this paper, we consider the Riemannian metric on $\mathcal{M}$ that is induced from the Euclidean inner product; i.e. $\langle \cdot,\cdot \rangle_x = \langle \cdot,\cdot \rangle$, $\forall x \in \mathcal{M}$[2]. Using this Riemannian metric, the Riemannian gradient of a function is simply the projection of its Euclidean gradient onto the tangent space:

$$\mathrm{grad} f(x) = \mathrm{Proj}_{T_x\mathcal{M}}(\nabla f(x)). \tag{4}$$

We also present the definition of Riemannian Hessian for embedded submanifolds, which will be used in Section 3.4 about cubic regularized Newton's method.

Definition 5 (Riemannian Hessian [95]). Suppose $\mathcal{M}$ is an embedded submanifold of $\mathbb{R}^n$. The Riemannian Hessian is defined as

$$\mathrm{Hess} f(x)[\eta] = \mathrm{Proj}_{T_x\mathcal{M}}(D\mathrm{grad} f(x)[\eta]), \forall x \in \mathcal{M}, \eta \in T_x\mathcal{M}, \tag{5}$$

where $D\mathrm{grad} f(x)[\eta]$ is the common differential, i.e., $D\mathrm{grad} f(x)[\eta] = (J\mathrm{grad} f(x))[\eta]$, where J is the Jacobian of the gradient mapping.

2.2. The Zeroth-order Riemannian Gradient Estimator

Recall that in the Euclidean setting, Nesterov and Spokoiny [66] analyzed the Gaussian smoothing based zeroth-order gradient estimator. However, as that estimator requires function evaluations outside of the manifold to be well-defined, it is not directly applicable for the Riemannian setting. To address this issue, we introduce our stochastic zeroth-order Riemannian gradient estimator below.

Definition 6 (Zeroth-Order Riemannian Gradient). Generate $u = Pu_0 \in T_x\mathcal{M}$, where $u_0 \sim \mathcal{N}(0, I_n)$ in $\mathbb{R}^n$, and $P \in \mathbb{R}^{n\times n}$ is the orthogonal projection matrix onto $T_x\mathcal{M}$. Therefore u follows the standard normal distribution $\mathcal{N}(0, PP^\top)$ on the tangent space in the sense that, all the eigenvalues of the covariance matrix $PP^\top$ are either 0 (eigenvectors orthogonal to the tangent plane) or 1 (eigenvectors embedded in the tangent plane). The zeroth-order Riemannian gradient estimator is defined as

$$g_\mu(x) = \frac{f(R_x(\mu u)) - f(x)}{\mu} u = \frac{f(R_x(\mu P u_0)) - f(x)}{\mu} P u_0. \tag{6}$$

[1] If the manifold is compact, then the exponential mapping Exp_x is already globally defined. This is known as the Hopf-Rinow theorem [15].

[2] If the manifold is not an embedded submanifold of some Euclidean space, then we cannot have an induced Riemannian metric. In this case, the convergence result is not affected, though it would cause implementation difficulties.

Note that the projection P is easy to compute for commonly used manifolds. For example, for the Stiefel manifold $\mathcal{M}$, the projection is given by $\mathrm{Proj}_{T_X\mathcal{M}}(Y) = (I - XX^\top)Y + X\,\mathrm{skew}(X^\top Y)$, where $\mathrm{skew}(A) := (A - A^\top)/2$ (see [1]).

Remark 1. In this work, we assume that the function f is defined on submanifolds embedded in Euclidean space, so that it is efficient to sample from the associated tangent space, as discussed above; see also [26]. We remark that the above gradient estimation methodology is more generally applicable to other manifolds. However, the generality comes at the cost of practical applicability as it is not an easy task to efficiently sample Gaussian random objects on the tangent space of general manifolds; see [36] for more details.

We now discuss some differences between the zeroth-order gradient estimators in the Euclidean setting [66] and the Riemannian setting (6). In the Euclidean case, the zeroth-order gradient estimator can be viewed as estimating the gradient of the Gaussian smoothed function, $f_\mu(x) = \frac{1}{\kappa}\int_{\mathbb{R}^n} f(x+\mu u)e^{-\frac{1}{2}\|u\|^2}du$, because $\nabla f_\mu(x) = \mathbb{E}_u(g_\mu(x)) = \frac{1}{\kappa}\int_{\mathbb{R}^n} \frac{f(x+\mu u)-f(x)}{\mu} u e^{-\frac{1}{2}\|u\|^2}du$, where κ is the normalization constant for Gaussian. This was also observed as an instantiation of Gaussian Stein's identity [4]. However, this observation is no longer true in the Riemannian setting, as we incorporate the retraction operator when evaluating g_μ, and this forces us to seek for a direct evaluation of $\mathbb{E}_u(g_\mu(x))$, instead of utilizing properties of the smoothed function f_μ.We also remark that, $g_\mu(x)$ is a biased estimator of $\mathrm{grad} f(x)$. The difference between them can be bounded as in Proposition 1. Some intermediate results for this purpose are as follows.

Lemma 1. *Suppose $\mathcal{X}$ is a d-dimensional subspace of $\mathbb{R}^n$, with orthogonal projection matrix $P \in \mathbb{R}^{n\times n}$. u_0 follows a standard norm distribution $\mathcal{N}(0, I_n)$, and $u = Pu_0$ is the orthogonal projection of u_0 onto the subspace $\mathcal{X}$. Then $\forall x \in \mathcal{X}$, we have*

$$x = \frac{1}{\kappa}\int_{\mathbb{R}^n} \langle x, u\rangle u e^{-\frac{1}{2}\|u_0\|^2}du_0, \qquad \text{and} \qquad \|x\|^2 = \frac{1}{\kappa}\int_{\mathbb{R}^n} \langle x, u\rangle^2 e^{-\frac{1}{2}\|u_0\|^2}du_0, \tag{7}$$

where κ is the constant for normal density function: $\kappa := \int_{\mathbb{R}^n} e^{-\frac{1}{2}\|u\|^2}du = (2\pi)^{n/2}$.

Proof of Lemma 1 By the definition of covariance matrix, we have $\frac{1}{\kappa}\int_{\mathbb{R}^n} u_0 u_0^\top e^{-\frac{1}{2}\|u_0\|^2}du_0 = I_n$. Since $\langle x, u\rangle = \langle x, u_0\rangle$, $\forall x \in \mathcal{X}$, we have

$$\frac{1}{\kappa}\int_{\mathbb{R}^n} \langle x, u\rangle u_0 e^{-\frac{1}{2}\|u_0\|^2}du_0 = x, \tag{8}$$

which implies $\frac{1}{\kappa}\int_{\mathbb{R}^n} \langle x, u\rangle u e^{-\frac{1}{2}\|u_0\|^2}du_0 = Px = x$. Similarly, taking inner product with x on both sides of Eq. (8), we have $\|x\|^2 = \frac{1}{\kappa}\int_{\mathbb{R}^n} \langle x, u\rangle^2 e^{-\frac{1}{2}\|u_0\|^2}du_0$.
The following bound for the moments of normal distribution is restated without proof.

Lemma 2. *[66] Suppose $u \sim \mathcal{N}(0, I_n)$ is a standard normal distribution. Then for all integers $p \geq 2$, we have $M_p := \mathbb{E}_u(\|u\|^p) \leq (n+p)^{p/2}$.*

Corollary 1. *For $u_0 \sim \mathcal{N}(0, I_n)$ and $u = Pu_0$, where $P \in \mathbb{R}^{n\times n}$ is the orthogonal projection matrix onto a d dimensional subspace $\mathcal{X}$ of $\mathbb{R}^n$, we have $\mathbb{E}_{u_0}(\|u\|^p) \leq (d+p)^{p/2}$.*

Proof of Corollary 1 Assume the eigen-decomposition of P is $P = Q^\top \Lambda Q$, where Q is an unitary matrix and Λ is a diagonal matrix with the leading d diagonal entries being 1 and other diagonal entries being 0. Denote $\tilde{u} = Qu_0 \sim \mathcal{N}(0, I_n)$, then $\Lambda\tilde{u} = (\tilde{u}_1, ..., \tilde{u}_d, 0, ..., 0)$. Since $u = Q^\top\Lambda\tilde{u}$ has the same distribution as $\Lambda\tilde{u}$, we have $\mathbb{E}\|u\|^p = \mathbb{E}\|(\tilde{u}_1, ..., \tilde{u}_d, 0, ..., 0)\|^p \leq (d+p)^{p/2}$, by Lemma 2.

Now we provide the bounds on the error of our gradient estimator $g_\mu(x)$ (6). Recall that d denotes the dimension of the manifold $\mathcal{M}$.

Proposition 1. *Under Assumption 1, we have*

(a) $\|\mathbb{E}_{u_0}(g_\mu(x)) - \mathrm{grad} f(x)\| \le \frac{\mu L_g}{2}(d+3)^{3/2}$,
(b) $\|\mathrm{grad} f(x)\|^2 \le 2\|\mathbb{E}_{u_0}(g_\mu(x))\|^2 + \frac{\mu^2}{2}L_g(d+6)^3$,
(c) $\mathbb{E}_{u_0}(\|g_\mu(x)\|^2) \le \frac{\mu^2}{2}L_g^2(d+6)^3 + 2(d+4)\|\mathrm{grad} f(x)\|^2$.

Proof of Proposition 1 For part (a), since

$$\mathbb{E}(g_\mu(x)) - \mathrm{grad} f(x) = \frac{1}{\kappa}\int_{\mathbb{R}^n}\left(\frac{f(R_x(\mu u)) - f(x)}{\mu} - \langle \mathrm{grad} f(x), u\rangle\right) u e^{-\frac{1}{2}\|u_0\|^2} du_0,$$

we have

$$\begin{aligned} & \|\mathbb{E}(g_\mu(x)) - \mathrm{grad} f(x)\| \\ = \ & \|\frac{1}{\mu\kappa}\int_{\mathbb{R}^n}\left(f(R_x(\mu u)) - f(x) - \langle \mathrm{grad} f(x), \mu u\rangle\right) u e^{-\frac{1}{2}\|u_0\|^2} du_0\| \\ \le \ & \frac{1}{\mu\kappa}\int_{\mathbb{R}^n}\frac{L_g}{2}\|\mu u\|^2\|u\| e^{-\frac{1}{2}\|u_0\|^2} du_0 = \frac{\mu L_g}{2\kappa}\int_{\mathbb{R}^n}\|u\|^3 e^{-\frac{1}{2}\|u_0\|^2} du_0 \le \frac{\mu L_g}{2}(d+3)^{3/2}, \end{aligned}$$

where the first inequality is by due to (3), and the last inequality is from Corollary 1. This completes the proof of part (a).

To prove part (b), note that

$$\begin{aligned} & \|\mathrm{grad} f(x)\|^2 = \left\|\frac{1}{\kappa}\int_{\mathbb{R}^n}\langle \mathrm{grad} f(x), u\rangle u e^{-\frac{1}{2}\|u_0\|^2} du_0\right\|^2 \\ = & \left\|\frac{1}{\mu\kappa}\int_{\mathbb{R}^n}\left([f(R_x(\mu u)) - f(x)] - [f(R_x(\mu u)) - f(x) - \langle \mathrm{grad} f(x), \mu u\rangle]\right) u e^{-\frac{1}{2}\|u_0\|^2} du_0\right\|^2 \\ \le & 2\|\mathbb{E}(g_\mu(x))\|^2 + \left\|\frac{2}{\mu^2}\int_{\mathbb{R}^n}\left(f(R_x(\mu u)) - f(x) - \langle \mathrm{grad} f(x), \mu u\rangle\right) u e^{-\frac{1}{2}\|u_0\|^2} du_0\right\|^2 \\ \le & 2\|\mathbb{E}(g_\mu(x))\|^2 + \frac{2}{\mu^2}\int_{\mathbb{R}^n}\left(f(R_x(\mu u)) - f(x) - \langle \mathrm{grad} f(x), \mu u\rangle\right)^2\|u\|^2 e^{-\frac{1}{2}\|u_0\|^2} du_0 \\ \le & 2\|\mathbb{E}(g_\mu(x))\|^2 + \frac{\mu^2}{2}L_g(d+6)^3, \end{aligned}$$

where the last inequality is from the same trick as in part (a). This completes the proof of part (b).

Finally, we prove part (c). Since $\mathbb{E}(\|g_\mu(x)\|^2) = \frac{1}{\mu^2}\mathbb{E}_{u_0}\left[(f(R_x(\mu u)) - f(x))^2\|u\|^2\right]$, and $(f(R_x(\mu u)) - f(x))^2 = (f(R_x(\mu u)) - f(x) - \mu\langle \mathrm{grad} f(x), u\rangle + \mu\langle \mathrm{grad} f(x), u\rangle)^2 \le 2(\frac{L_g}{2}\mu^2\|u\|^2)^2 + 2\mu^2\langle \mathrm{grad} f(x), u\rangle^2$, we have

$$\mathbb{E}(\|g_\mu(x)\|^2) \le \frac{\mu^2}{2}L_g^2\mathbb{E}(\|u\|^6) + 2\mathbb{E}(\|\langle \mathrm{grad} f(x), u\rangle u\|^2) \le \frac{\mu^2}{2}L_g^2(d+6)^3 + 2\mathbb{E}(\|\langle \mathrm{grad} f(x), u\rangle u\|^2). \quad (9)$$

Now we bound the term $\mathbb{E}(\|\langle \mathrm{grad} f(x), u\rangle u\|^2)$ using the same trick as in [66]. Without loss of generality, assume $\mathcal{X}$ is the d-dimensional subspace generated by the first d coordinates, i.e., $\forall x \in \mathcal{X}$, the last $n-d$ elements of x are zeros. Also for brevity, denote $g = \mathrm{grad} f(x)$. We have that

$$\begin{aligned} \mathbb{E}(\|\langle \mathrm{grad} f(x), u\rangle u\|^2) &= \frac{1}{\kappa}\int_{\mathbb{R}^n}\langle \mathrm{grad} f(x), u\rangle^2\|u\|^2 e^{-\frac{1}{2}\|u_0\|^2} du_0 \\ &= \frac{1}{\kappa(d)}\int_{\mathbb{R}^d}\left(\sum_{i=1}^d g_i x_i\right)^2\left(\sum_{i=1}^d x_i^2\right) e^{-\frac{1}{2}\sum_{i=1}^d x_i^2} dx_1\cdots dx_d, \end{aligned}$$

where x_i denotes the i-th coordinate of u_0, the last $n-d$ dimensions are integrated to be one, and $\kappa(d)$ is the normalization constant for d-dimensional Gaussian distribution. For simplicity, denote $x=(x_1,...,x_d)$, then

$$\begin{aligned}
&\mathbb{E}(\|\langle \mathrm{grad} f(x), u\rangle u\|^2) = \frac{1}{\kappa(d)}\int_{\mathbb{R}^d} \langle g, x\rangle^2 \|x\|^2 e^{-\frac{1}{2}\|x\|^2} dx \\
\leq & \frac{1}{\kappa(d)}\int_{\mathbb{R}^d} \|x\|^2 e^{-\frac{\tau}{2}\|x\|^2} \langle g,x\rangle^2 e^{-\frac{1-\tau}{2}\|x\|^2} dx \leq \frac{2}{\kappa(d)\tau e}\int_{\mathbb{R}^d} \langle g,x\rangle^2 e^{-\frac{1-\tau}{2}\|x\|^2} dx \\
= & \frac{2}{\kappa(d)\tau(1-\tau)^{1+d/2}e}\int_{\mathbb{R}^d} \langle g,x\rangle^2 e^{-\frac{1}{2}\|x\|^2} dx = \frac{2}{\tau(1-\tau)^{1+d/2}e}\|g\|^2,
\end{aligned} \tag{10}$$

where the second inequality is due to the following fact: $x^p e^{-\frac{\tau}{2}x^2} \leq (\frac{2}{\tau e})^{p/2}$. Taking $\tau = \frac{2}{(d+4)}$ gives the desired result.

2.3. The Zeroth-order Riemannian Hessian Estimator

We now extend the above methodology and propose estimators for the Riemannian Hessian in the stochastic zeroth-order setting. We restrict our discussion to compact submanifolds embedded in Euclidean space, so that the definition of Riemannian Hessian (5) is applied. We assume the following assumption of $F(x,\xi)$:

Assumption 2. *Given any point $x \in \mathcal{M}$ and $\eta \in T_x\mathcal{M}$, we have*

$$\|P_\eta^{-1} \circ \mathrm{Hess} F(R_x(\eta),\xi) \circ P_\eta - \mathrm{Hess} F(x,\xi)\|_{\mathrm{op}} \leq L_H \|\eta\|, \tag{11}$$

almost everywhere for ξ, where $P_\eta : T_x\mathcal{M} \to T_{R_x(\eta)}\mathcal{M}$ denotes the parallel transport [2], an isometry from the tangent space of x to the tangent space of $R_x(\eta)$, and $\circ$ is the function composition. Here $\|\cdot\|_{\mathrm{op}}$ is the operator norm in the ambient Euclidean space.

Assumption 2 is the analogue of the Lipschitz Hessian type assumption from the Euclidean setting, and induces the following equivalent conditions (see, also [2]):

$$\begin{gathered}
\|P_\eta^{-1} \mathrm{grad} F(R_x(\eta),\xi) - \mathrm{grad} f(x) - \mathrm{Hess} F(x,\xi)[\eta]\| \leq \frac{L_H}{2}\|\eta\|^2 \\
\left| F(R_x(\eta),\xi) - \left[F(x,\xi) + \langle \eta, \mathrm{grad} F(x,\xi)\rangle + \frac{1}{2}\langle \eta, \mathrm{Hess} F(x,\xi)[\eta]\rangle \right] \right| \leq \frac{L_H}{6}\|\eta\|^3.
\end{gathered} \tag{12}$$

In the Euclidean setting, P_η reduces to the identity mapping. Throughout this section, we also assume that $F(\cdot,\xi)$ satisfies Assumption 1 and the following assumption, which is used frequently in zeroth-order stochastic optimization [31, 4, 96].

Assumption 3. *We have (with $\mathbb{E} = \mathbb{E}_\xi$) that, $\mathbb{E}[F(x,\xi)] = f(x)$, $\mathbb{E}[\mathrm{grad} F(x,\xi)] = \mathrm{grad} f(x)$ and $\mathbb{E}[\|\mathrm{grad} F(x,\xi) - \mathrm{grad} f(x)\|^2] \leq \sigma^2$, $\forall x \in \mathcal{M}$.*

We first introduce the following identity which follows immediately from the second-order Stein's identity for Gaussian distribution [74].

Lemma 3. *Suppose $\mathcal{X}$ is a d-dimensional subspace of $\mathbb{R}^n$, with orthogonal projection matrix $P \in \mathbb{R}^{n\times n}$, $P = P^2 = P^\top$, and $u_0 \sim \mathcal{N}(0, I_n)$ is a standard normal distribution and $u = Pu_0$ is the orthogonal projection of u_0 onto the subspace. Then $\forall H \in \mathbb{R}^{n\times n}$, $H^\top = H$, and $H = PHP$ (which means that the eigenvectors of H lies all in $\mathcal{X}$), we have*

$$PHP = \frac{1}{2\kappa}\int_{\mathbb{R}^n} \langle u, Hu\rangle (uu^\top - P) e^{-\frac{1}{2}\|u_0\|^2} du_0 = \mathbb{E}\left[\frac{1}{2}\langle u, Hu\rangle(uu^\top - P)\right], \tag{13}$$

where $\|\cdot\|$ here is the Euclidean norm on $\mathbb{R}^n$, and κ is the constant for normal density function given by $\kappa := \int_{\mathbb{R}^n} e^{-\frac{1}{2}\|u\|^2} du = (2\pi)^{n/2}$.

The identity in (13) simply follows by applying the second-order Stein's identity, $\mathbb{E}[(xx^\top - I_n)g(x)] = \mathbb{E}[\nabla^2 g(x)]$, directly to the function $g(x) = \frac{1}{2}\langle x, Hx\rangle$ and multiplying the resulting identity by P on both sides.

Lemma 4. *[4] Suppose $\mathcal{X}$ is a d-dimensional subspace of $\mathbb{R}^n$, with orthogonal projection matrix $P \in \mathbb{R}^{n\times n}$, $P = P^2 = P^\top$, and $u_0 \sim \mathcal{N}(0, I_n)$ is a standard normal distribution and $u = Pu_0$ is the orthogonal projection of u_0 onto the subspace. Then*

$$\mathbb{E}[\|u_0u_0^\top - I_n\|_F^8] \le 2(n+16)^8 \quad \textit{and} \quad \mathbb{E}[\|uu^\top - P\|_F^8] \le 2(d+16)^8. \tag{14}$$

Proof of Lemma 4 See [4] for the proof of the first inequality in Eq. (14). We now show how to get the right part from the left. Similar to the proof of Corollary 1, we use an eigen-decomposition of $P = Q^T\Lambda Q$ and get (again $\tilde{u} = Qu$):

$$\mathbb{E}\|uu^\top - P\|_F^8 = \mathbb{E}\|(\tilde{u}_1, ..., \tilde{u}_d)^\top(\tilde{u}_1, ..., \tilde{u}_d) - I_d\|_F^8 \le 2(d+16)^8,$$

which completes the proof.

We now propose our zeroth-order Riemannian Hessian estimator, motivated by the zeroth-order Hessian estimator in the Euclidean setting proposed by [4].

Definition 7 (Zeroth-Order Riemannian Hessian). Generate $u \in T_x\mathcal{M}$ following a standard normal distribution on the tangent space $T_x\mathcal{M}$, by projection $u = P_xu_0$ as described in Section 2.2. Then, the zeroth-order Riemannian Hessian estimator of a function f at the point x is given by

$$H_\mu(x) = \frac{1}{2\mu^2}(uu^\top - P)[F(R_x(\mu u), \xi) + F(R_x(-\mu u), \xi) - 2F(x, \xi)]. \tag{15}$$

Note that our Riemannian Hessian estimator is actually the Hessian estimator of the pullback function $\hat{F}_x(\eta, \xi) = F(R_x(\eta), \xi)$, $\forall x \in \mathcal{M}$ and $\eta \in T_x\mathcal{M}$ projected onto the tangent space $T_x\mathcal{M}$. We immediately have the following bound on the variance of $H_\mu(x)$.

Lemma 5. *Under Assumption 1, the Riemannian Hessian estimator given in Eq. (15) satisfies*

$$\mathbb{E}_{\mathcal{U},\Xi}\|H_\mu(x)\|_F^4 \le \frac{(d+16)^8}{8}L_g^2. \tag{16}$$

Proof of Lemma 5 From Assumption 1 and Corollary 1 we have

$$\begin{aligned}
&\mathbb{E}|F(R_x(\mu u), \xi) + F(R_x(-\mu u), \xi) - 2F(x, \xi)|^8 \\
=&\mathbb{E}|F(R_x(\mu u), \xi) - F(x, \xi) - \langle \mathrm{grad} F(x, \xi), \mu u\rangle + F(R_x(-\mu u), \xi) - F(x, \xi) - \langle \mathrm{grad} F(x, \xi), -\mu u\rangle|^8 \\
\le&\mathbb{E}[\frac{\mu^2 L_g}{2}\|u\|^2 + \frac{\mu^2 L_g}{2}\|u\|^2]^8 = \mathbb{E}[\mu^{16}L_g^8\|u\|^{16}] \le \mu^{16}L_g^8(d+16)^8.
\end{aligned} \tag{17}$$

Moreover, we have

$$\begin{aligned}
\mathbb{E}\|H_\mu(x)\|_F^4 =&\mathbb{E}\left\|\frac{1}{2\mu^2}(uu^\top - P)[F(R_x(\mu u), \xi) + F(R_x(-\mu u), \xi) - 2F(x, \xi)]\right\|_F^4 \\
\le&\frac{1}{16\mu^8}\left(\mathbb{E}|F(R_x(\mu u), \xi) + F(R_x(-\mu u), \xi) - 2F(x, \xi)|^8\mathbb{E}\|uu^\top - P\|^8\right)^{1/2} \\
\le&\frac{(d+16)^4}{8\mu^8}\left(\mathbb{E}|F(R_x(\mu u), \xi) + F(R_x(-\mu u), \xi) - 2F(x, \xi)|^8\right)^{1/2},
\end{aligned} \tag{18}$$

where the first inequality is by Hölder's inequality and the second one is by Lemma 4. Combining (17) and (18) yields the desired result (16).

We will also use the mini-batch multi-sampling technique. For $i = 1, ..., b$, denote each Hessian estimator as

$$H_{\mu,i}(x) = \frac{1}{2\mu^2}(u_i u_i^\top - P)[F(R_x(\mu u_i), \xi_i) + F(R_x(-\mu u_i), \xi_i) - 2F(x, \xi_i)]. \tag{19}$$

The averaged Hessian estimator is given by

$$\bar{H}_{\mu,\xi}(x) = \frac{1}{b}\sum_{i=1}^{b} H_{\mu,i}(x). \tag{20}$$

We now have the following bound of $\bar{H}_{\mu,\xi}(x)$ and $\mathrm{Hess}f(x)$.

Lemma 6. *Under Assumption 1 and Assumption 2, let $\bar{H}_{\mu,\xi}(x)$ be calculated as in Eq. (20), then we have that: $\forall x \in \mathcal{M}$ and $\forall \eta \in T_x\mathcal{M}$,*

$$\mathbb{E}_{\mathcal{U},\Xi}\|\bar{H}_{\mu,\xi}(x) - \mathrm{Hess}f(x)\|_{\mathrm{op}}^2 \le \frac{(d+16)^4}{\sqrt{2}b}L_g + \frac{\mu^2 L_H^2}{18}(d+6)^5, \tag{21}$$

$$\mathbb{E}_{\mathcal{U},\Xi}\|\bar{H}_{\mu,\xi}(x) - \mathrm{Hess}f(x)\|_{\mathrm{op}}^3 \le \tilde{C}\frac{(d+16)^6}{b^{3/2}}L_g^{1.5} + \frac{1}{27}\mu^3 L_H^3(d+6)^{7.5}, \tag{22}$$

where $\|\cdot\|_{op}$ denotes the operator norm and $\tilde{C}$ is some absolute constant.

Proof of Lemma 6 Denote $\mathbb{E} = \mathbb{E}_{\mathcal{U},\Xi}$ as the expectation with respect to all previous random variables. We first show Eq. (21). Denote $X_i = H_{\mu,i} - \mathbb{E}H_{\mu,i}$, then X_i's are iid zero-mean random matrices. Since $\|\cdot\|_{\mathrm{op}} \le \|\cdot\|_F$, we have

$$\begin{aligned}
&\mathbb{E}\|\bar{H}_{\mu,\xi}(x) - \mathbb{E}\bar{H}_{\mu,\xi}(x)\|_{\mathrm{op}}^2 = \mathbb{E}\left\|\frac{1}{b}\sum_{i=1}^{b}X_i\right\|_{\mathrm{op}}^2 \le \mathbb{E}\left\|\frac{1}{b}\sum_{i=1}^{b}X_i\right\|_F^2\\
=&\mathbb{E}\left[\frac{1}{b^2}\sum_{i=1}^{b}\|X_i\|_F^2 + \frac{1}{b^2}\sum_{i\ne j}\langle X_i, X_j\rangle\right] = \mathbb{E}\left[\frac{1}{b^2}\sum_{i=1}^{b}\|X_i\|_F^2\right]\\
=&\mathbb{E}\frac{1}{b^2}b\|X_1\|_F^2 = \mathbb{E}\frac{1}{b}\|H_{\mu,1} - \mathbb{E}H_{\mu,1}\|_F^2 = \frac{1}{b}\mathbb{E}\left[\|H_{\mu,1}\|_F^2 - \|\mathbb{E}H_{\mu,1}\|_F^2\right]\\
\le&\frac{1}{b}\mathbb{E}\|H_{\mu,1}\|_F^2 \le \frac{1}{b}\sqrt{\mathbb{E}\|H_{\mu,1}(x)\|_F^4} \le \frac{(d+16)^4}{2\sqrt{2}b}L_g,
\end{aligned} \tag{23}$$

where the third inequality is from the Jensen's inequality, and the last inequality is due to Eq. (16). Note that (23) immediately implies

$$\begin{aligned}
\mathbb{E}\|\bar{H}_{\mu,\xi}(x) - \mathrm{Hess}f(x)\|_{\mathrm{op}}^2 \le& 2\mathbb{E}\|\bar{H}_{\mu,\xi}(x) - \mathbb{E}\bar{H}_{\mu,\xi}(x)\|_{\mathrm{op}}^2 + 2\|\mathbb{E}\bar{H}_{\mu,\xi}(x) - \mathrm{Hess}f(x)\|_{\mathrm{op}}^2\\
\le&\frac{(d+16)^4}{\sqrt{2}b}L_g + 2\|\mathbb{E}\bar{H}_{\mu,\xi}(x) - \mathrm{Hess}f(x)\|_{\mathrm{op}}^2.
\end{aligned} \tag{24}$$

Now we bound the term $\|\mathbb{E}\bar{H}_{\mu,\xi}(x) - \mathrm{Hess}f(x)\|_{\mathrm{op}}^2$. Note that

$$\begin{aligned}
&|\langle \eta, (\mathbb{E}H_{\mu,i}(x) - \mathrm{Hess}f(x))[\eta]\rangle|\\
=&\left|\langle \eta, \left(\mathbb{E}\left[\frac{1}{2\mu^2}(uu^\top - P)[f(R_x(\mu u)) + f(R_x(-\mu u)) - 2f(x)]\right] - \mathrm{Hess}f(x)\right)[\eta]\rangle\right|\\
=&\left|\langle \eta, \left(\mathbb{E}\left[\frac{1}{2\mu^2}(uu^\top - P)[f(R_x(\mu u)) + f(R_x(-\mu u)) - 2f(x) - \mu^2\langle u, \mathrm{Hess}f(x)[u]\rangle]\right]\right)[\eta]\rangle\right|\\
=&\frac{1}{2\mu^2}\left|\langle \eta, \left(\mathbb{E}\left[[f(R_x(\mu u)) - f(x) - \frac{\mu^2}{2}\langle u, \mathrm{Hess}f(x)[u]\rangle\right.\right.\right.\\
&\left.\left.\left.+ f(R_x(-\mu u)) - f(x) - \frac{\mu^2}{2}\langle u, \mathrm{Hess}f(x)[u]\rangle](uu^\top - P)\right]\right)[\eta]\rangle\right|,
\end{aligned}$$

which together with Assumption 2 yields

$$\begin{aligned} |\langle \eta, (\mathbb{E}H_{\mu,i}(x) - \mathrm{Hess} f(x))[\eta]\rangle| \le& \frac{\mu L_H}{6}\mathbb{E}\left[\|u\|^3\|uu^\top - P\|_{\mathrm{op}}\right]\|\eta\|^2 \\ \text{(Hölder)} \le& \frac{\mu L_H}{6}\sqrt{\mathbb{E}\|u\|^6\mathbb{E}\|uu^\top - P\|_F^2}\|\eta\|^2 \le \frac{\mu L_H}{6}(d+6)^{5/2}\|\eta\|^2, \end{aligned} \tag{25}$$

where the last inequality is by Corollary 1 and Lemma 4. (25) implies

$$\|\mathbb{E}\bar{H}_{\mu,\xi}(x) - \mathrm{Hess} f(x)\|_{\mathrm{op}} \le \frac{\mu L_H}{6}(d+6)^{5/2}. \tag{26}$$

Combining (24) and (26) gives Eq. (21).

Now we show Eq. (22). By a similar analysis we have

$$\begin{aligned} &\mathbb{E}\|\bar{H}_{\mu,\xi}(x) - \mathrm{Hess} f(x)\|_{\mathrm{op}}^3 \\ &\le \mathbb{E}(\|\bar{H}_{\mu,\xi}(x) - \mathbb{E}\bar{H}_{\mu,\xi}(x)\|_{\mathrm{op}} + \|\mathbb{E}\bar{H}_{\mu,\xi}(x) - \mathrm{Hess} f(x)\|_{\mathrm{op}})^3 \\ &\le 8\mathbb{E}\|\bar{H}_{\mu,\xi}(x) - \mathbb{E}\bar{H}_{\mu,\xi}(x)\|_{\mathrm{op}}^3 + 8\|\mathbb{E}\bar{H}_{\mu,\xi}(x) - \mathrm{Hess} f(x)\|_{\mathrm{op}}^3 \\ \text{(Hölder)}\ &\le 8\sqrt{\mathbb{E}\|\bar{H}_{\mu,\xi}(x) - \mathbb{E}\bar{H}_{\mu,\xi}(x)\|_{\mathrm{op}}^2\mathbb{E}\|\bar{H}_{\mu,\xi}(x) - \mathbb{E}\bar{H}_{\mu,\xi}(x)\|_{\mathrm{op}}^4} \\ &\quad + 8\|\mathbb{E}\bar{H}_{\mu,\xi}(x) - \mathrm{Hess} f(x)\|_{\mathrm{op}}^3, \end{aligned} \tag{27}$$

where the second inequality is by the following fact: when $a, b \ge 0$, $(a+b)^3 \le \max\{(2a)^3, (2b)^3\} \le 8a^3 + 8b^3$. Moreover, since $\|\cdot\|_{\mathrm{op}} \le \|\cdot\|_F$, and $X_i = H_{\mu,i} - \mathbb{E}H_{\mu,i}$ are iid zero-mean random matrices, we have

$$\begin{aligned} &\mathbb{E}\|\bar{H}_{\mu,\xi}(x) - \mathbb{E}\bar{H}_{\mu,\xi}(x)\|_{\mathrm{op}}^4 = \mathbb{E}\|\frac{1}{b}\sum_{i=1}^b X_i\|_{\mathrm{op}}^4 \le \frac{C}{b^4}\left(\mathbb{E}\|\sum_{i=1}^b X_i\|_{\mathrm{op}} + (b\mathbb{E}\|X_i\|_{\mathrm{op}}^4)^{1/4}\right)^4 \\ \le& \frac{C}{b^4}\left(\sqrt{\mathbb{E}\|\sum_{i=1}^b X_i\|_F^2} + (b\mathbb{E}\|X_i\|_F^4)^{1/4}\right)^4 = \frac{C}{b^4}\left(\sqrt{\sum_{i=1}^b \mathbb{E}\|X_i\|_F^2} + (b\mathbb{E}\|X_i\|_F^4)^{1/4}\right)^4 \\ =& \frac{C}{b^4}\left(\sqrt{b}\sqrt{\mathbb{E}\|X_1\|_F^2} + (b\mathbb{E}\|X_1\|_F^4)^{1/4}\right)^4 \le \frac{C}{b^4}\left(\sqrt{b}\sqrt[4]{\mathbb{E}\|X_1\|_F^4} + (b\mathbb{E}\|X_1\|_F^4)^{1/4}\right)^4 \\ =& \frac{C}{b^4}(\sqrt{b} + \sqrt[4]{b})^4\mathbb{E}\|H_{\mu,1} - \mathbb{E}H_{\mu,1}\|_F^4 \le \frac{16C}{b^2}\mathbb{E}\|H_{\mu,1} - \mathbb{E}H_{\mu,1}\|_F^4 \\ =& \frac{16C}{b^2}\mathbb{E}(\|H_{\mu,1}\|_F^2 - 2\langle H_{\mu,1}, \mathbb{E}H_{\mu,1}\rangle + \|\mathbb{E}H_{\mu,1}\|_F^2)^2 \\ \le& \frac{16C}{b^2}\mathbb{E}(\|H_{\mu,1}\|_F^2 + 2\|H_{\mu,1}\|_F\|\mathbb{E}H_{\mu,1}\|_F + \|\mathbb{E}H_{\mu,1}\|_F^2)^2 \\ \le& \frac{16C}{b^2}\mathbb{E}(2\|H_{\mu,1}\|_F^2 + 2\|\mathbb{E}H_{\mu,1}\|_F^2)^2 \le \frac{16C}{b^2}\mathbb{E}(2\|H_{\mu,1}\|_F^2 + 2\mathbb{E}\|H_{\mu,1}\|_F^2)^2 \\ \le& \frac{64C}{b^2}(\mathbb{E}\|H_{\mu,1}\|_F^4 + \mathbb{E}\|H_{\mu,1}\|_F^4) \le \frac{128C}{b^2}(d+16)^8 L_g^2, \end{aligned} \tag{28}$$

where the first inequality is due to the Rosenthal inequality [69], C is an absolute constant, the fourth inequality is due to the fact $1 \le \sqrt[4]{b} \le \sqrt{b}$. Plugging Eq. (23), Eq. (26) and Eq. (28) back to Eq. (27) gives the desired result (22).

3. Stochastic Zeroth-order Riemannian Optimization Algorithms

We now demonstrate the applicability of the developed Riemannian derivative estimation methodology in Section 2, for various classes of stochastic zeroth-order Riemannian optimization algorithms.

Algorithm 1 Zeroth-Order Riemannian Gradient Descent (ZO-RGD)

```
Input: Initial point $x_0 \in \mathcal{M}$, smoothing parameter $\mu$, step size $\eta_k$, fixed number of iteration $N$.
for $k = 0$ to $N-1$ do
    Sample a standard Gaussian random vector $u_k \in T_{x_k}\mathcal{M}$ by orthogonal projection in Definition 6.
    Compute the zeroth-order gradient $g_\mu(x_k)$ by Eq. (6).
    Update $x_{k+1} = R_{x_k}(-\eta_k g_\mu(x_k))$.
end for
```

3.1. Zeroth-order Smooth Riemannian Optimization

In this section, we focus on the smooth optimization problem with $h \equiv 0$ and f satisfying Assumption 1. We propose ZO-RGD, the zeroth-order Riemannian gradient descent method and provide its complexity analysis. The algorithm is formally presented in Algorithm 1.

The following theorem gives the iteration and oracle complexities of Algorithm 1 for obtaining an ϵ-stationary point of (1) when $h \equiv 0$.

Theorem 1. *Let f satisfy Assumption 1 and suppose $\{x_k\}$ is the sequence generated by Algorithm 1 with the stepsize $\eta_k = \hat{\eta} = \frac{1}{2(d+4)L_g}$. Then, we have*

$$\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}_{\mathcal{U}_k}\|\mathrm{grad} f(x_k)\|^2 \le \frac{4}{\hat{\eta}}\left(\frac{f(x_0)-f(x^*)}{N+1}+C(\mu)\right), \tag{29}$$

where $\mathcal{U}_k$ denotes the set of all Gaussian random vectors we drew for the first k iterations [3], *and $C(\mu) = \frac{\mu^2 L_g}{16}\frac{(d+3)^3}{(d+4)} + \frac{\mu^2}{16}\frac{(d+6)^3}{(d+4)} + \frac{\mu^2 L_g}{16}\frac{(d+6)^3}{(d+4)^2}$. In order to have*

$$\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}_{\mathcal{U}_k}\|\mathrm{grad} f(x_k)\|^2 \le \epsilon^2, \tag{30}$$

we need the smoothing parameter μ and number of iteration N (which is also the number of calls to the zeroth-order oracle) to be set as $\mu = \mathcal{O}\left(\epsilon/d^{3/2}\right),\ N = \mathcal{O}\left(d/\epsilon^2\right)$.

Proof of Theorem 1 From Assumption 1 we have

$$f(x_{k+1}) \le f(x_k) - \eta_k\langle g_\mu(x_k), \mathrm{grad} f(x_k)\rangle + \frac{\eta_k^2 L_g}{2}\|g_\mu(x_k)\|^2.$$

Taking the expectation w.r.t. u_k on both sides, we have

$$\begin{aligned}\mathbb{E}_{u_k}[f(x_{k+1})] &\le f(x_k) - \eta_k\langle \mathbb{E}_{u_k}(g_\mu(x_k)), \mathrm{grad} f(x_k)\rangle + \frac{\eta_k^2 L_g}{2}\mathbb{E}_{u_k}(\|g_\mu(x_k)\|^2)\\ &\le f(x_k) - \eta_k\langle \mathbb{E}_{u_k}(g_\mu(x_k)), \mathrm{grad} f(x_k)\rangle + \frac{\eta_k^2 L_g}{2}\left(\frac{\mu^2}{2}L_g^2(d+6)^3 + 2(d+4)\|\mathrm{grad} f(x_k)\|^2\right),\end{aligned}$$

where the last inequality is by Proposition 1. Now Take $\eta_k = \hat{\eta} = \frac{1}{2(d+4)L_g}$, we have

$$\begin{aligned}&\mathbb{E}_{u_k}[f(x_{k+1})]\\ \le& f(x_k) + \frac{\hat{\eta}}{2}(\|\mathrm{grad} f(x_k)\|^2 - 2\langle \mathbb{E}_{u_k}(g_\mu(x_k)), \mathrm{grad} f(x_k)\rangle) + \frac{\mu^2 L_g}{16}\frac{(d+6)^3}{(d+4)^2}\end{aligned}$$

[3] The notation of taking the expectation w.r.t. a set, is to take the expectation for each of the elements in the set.

$$
\begin{aligned}
=&f(x_k)+\frac{\hat{\eta}}{2}(\|\mathrm{grad}f(x_k)-\mathbb{E}_{u_k}(g_\mu(x_k))\|^2-\|\mathbb{E}_{u_k}(g_\mu(x_k))\|^2)+\frac{\mu^2 L_g}{16}\frac{(d+6)^3}{(d+4)^2}\\
\le&f(x_k)+\frac{\hat{\eta}}{2}\left(\frac{\mu^2L_g^2}{4}(d+3)^3-\frac{1}{2}\|\mathrm{grad}f(x_k)\|^2+\frac{\mu^2}{4}L_g(d+6)^3\right)+\frac{\mu^2 L_g}{16}\frac{(d+6)^3}{(d+4)^2}\\
=&f(x_k)-\frac{\hat{\eta}}{4}\|\mathrm{grad}f(x_k)\|^2+C(\mu),
\end{aligned}
$$

where the second inequality is from Proposition 1. Define $\phi_k := f(x_k)-f(x^*)$. Now take the expectation w.r.t. $\mathcal{U}_k=\{u_0,u_1,\ldots,u_{k-1}\}$, we have

$$\phi_{k+1}\le\phi_k-\frac{\hat{\eta}}{4}\mathbb{E}_{\mathcal{U}_k}\|\mathrm{grad}f(x_k)\|^2+C(\mu).$$

Summing the above inequality over $k=0,\ldots,N$ yields (29).

Therefore with $\mu=\mathcal{O}(\epsilon/d^{3/2})$ we have $C(\mu)\le\hat{\eta}\epsilon^2/4$. Taking $N\ge 8(d+4)L_g(f(x_0)-f(x^*))/\epsilon^2$ yields (30). In summary, the number of iterations for obtaining an ϵ-stationary solution is $\mathcal{O}(d/\epsilon^2)$, and hence the total zeroth-order oracle complexity is also $\mathcal{O}(d/\epsilon^2)$.

Remark 2. Note that in Algorithm 1, we only sample one Gaussian vector in each iteration of the algorithm. In practice, one can also sample multiple Gaussian random vectors in each iteration and obtain an averaged gradient estimator. Suppose we sample m i.i.d. Gaussian random vectors in each iteration and use the average $\bar{g}_\mu(x)=\frac{1}{m}\sum_{i=1}^m g_{\mu,i}(x)$, then the bound for our zeroth-order estimator becomes

$$\mathbb{E}(\|\bar{g}_\mu(x)-\mathrm{grad}f(x)\|^2)\le\mu^2L_g^2(d+6)^3+\frac{2(d+4)}{m}\|\mathrm{grad}f(x)\|^2. \tag{31}$$

Hence, the final result in Theorem 1 can be improved to

$$\frac{1}{N+1}\sum_{k=0}^N\mathbb{E}_{\mathcal{U}_k}\|\mathrm{grad}f(x_k)\|^2\le 4L_g\frac{f(x_0)-f(x^*)}{N+1}+\mu^2L_g^2(d+6)^3, \tag{32}$$

with $\hat{\eta}=1/L_g$ and $C(\mu)=\mu^2L_g(d+6)^3/2$. Therefore the number of iterations required is improved to $N=\mathcal{O}(1/\epsilon^2)$ when we set $\mu=\mathcal{O}(\epsilon/d^{3/2})$ and $m=\mathcal{O}(d)$. However, the zeroth-order oracle complexity is still $\mathcal{O}(d/\epsilon^2)$. The proof of (31) and (32) is given in the appendix. This multi-sampling technique will play a key role in our stochastic and non-smooth case analyses.

3.2. Zeroth-Order Stochastic Riemannian Optimization for Nonconvex Problem

In this section, we focus on the following nonconvex smooth problem:

$$\min_{x\in\mathcal{M}} f(x):=\int_\xi F(x,\xi)dP(\xi), \tag{33}$$

where P is a random distribution, F is a function satisfying Assumption 1, in variable x, almost surely. Note that f automatically satisfies Assumption 1 by the Jensen's inequality.

In the stochastic case, sampling multiple times in every iteration can improve the convergence rate. Our zeroth-order Riemannian gradient estimator is given by

$$\bar{g}_{\mu,\xi}(x)=\frac{1}{m}\sum_{i=1}^m g_{\mu,\xi_i}(x),\ \text{ where }\ g_{\mu,\xi_i}(x)=\frac{F(R_x(\mu u_i),\xi_i)-F(x,\xi_i)}{\mu}u_i, \tag{34}$$

and u_i is a standard normal random vector on $T_x\mathcal{M}$. We also immediately have that

$$\mathbb{E}_{\xi_i}g_{\mu,\xi_i}(x)=\frac{f(R_x(\mu u))-f(x)}{\mu}u=g_\mu(x). \tag{35}$$

The multi-sampling technique enables us to obtain the following bound on $\mathbb{E}\|\bar{g}_{\mu,\xi}(x)-\mathrm{grad}f(x)\|^2$, the proof of which is given in the Appendix C.

Lemma 7. *For the Riemannian gradient estimator in (34), under Assumptions 1 and 3, we have*

$$\mathbb{E}\|\bar{g}_{\mu,\xi}(x)-\mathrm{grad}f(x)\|^2 \le \mu^2 L_g^2 (d+6)^3 + \frac{8(d+4)}{m}\sigma^2 + \frac{8(d+4)}{m}\|\mathrm{grad}f(x)\|^2, \tag{36}$$

where the expectation $\mathbb{E}$ is taken for both Gaussian vectors $\mathcal{U}=\{u_1,...,u_m\}$ and ξ.

Our zeroth-order Riemannian stochastic gradient descent algorithm (`ZO-RSGD`) for solving (33), is presented in Algorithm 2.

Algorithm 2 Zeroth-order Riemannian Stochastic Gradient Descent (`ZO-RSGD`)

```
Input: Initial point $x_0 \in \mathcal{M}$, smoothing parameter $\mu$, multi-sample constant $m$, step size $\eta_k$,
    fixed number of iteration $N$.
for $k = 0$ to $N-1$ do
    Sample the standard Gaussian random vectors $u_i^k$ on $T_{x_k}\mathcal{M}$ by orthogonal projection in
        Definition 6, and sample $\xi_i^k$, $i = 1,...,m$.
    Compute the zeroth-order gradient $\bar{g}_{\mu,\xi}(x_k)$ by Eq. (34).
    Update $x_{k+1} = R_{x_k}(-\eta_k \bar{g}_{\mu,\xi}(x_k))$.
end for
```

Now we present convergence analysis for obtaining an ϵ-stationary point of (33).

Theorem 2. *Let F satisfy Assumption 1, w.r.t. variable x almost surely. Suppose $\{x_k\}$ is the sequence generated by Algorithm 2 with the stepsize $\eta_k = \hat{\eta} = \frac{1}{L_g}$. Under Assumption 3, we have*

$$\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}_{\mathcal{U}_k,\Xi_k}\|\mathrm{grad}f(x_k)\|^2 \le 4L_g\frac{f(x_0)-f(x^*)}{N+1} + C(\mu), \tag{37}$$

where $C(\mu) = 2\mu^2 L_g^2(d+6)^3 + \frac{16(d+4)}{m}\sigma^2$, $\mathcal{U}_k$ denotes the set of all Gaussian random vectors and Ξ_k denotes the set of all random variable ξ_k in the first k iterations. In order to have $\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}_{\mathcal{U}_k,\Xi_k}\|\mathrm{grad}f(x_k)\|^2 \le \epsilon^2$, we need the smoothing parameter μ, number of sampling m in each iteration and number of iterations N to be

$$\mu = \mathcal{O}\left(\epsilon/d^{3/2}\right),\ m = \mathcal{O}\left(d\sigma^2/\epsilon^2\right),\ N = \mathcal{O}\left(1/\epsilon^2\right). \tag{38}$$

Hence, the number of calls to the zeroth-order oracle is $mN = \mathcal{O}(d/\epsilon^4)$.

Proof of Theorem 2 From Assumption 1, we have:

$$f(x_{k+1}) \le f(x_k) - \eta_k\langle \bar{g}_{\mu,\xi}(x), \mathrm{grad}f(x_k)\rangle + \frac{\eta_k^2 L_g}{2}\|\bar{g}_{\mu,\xi}(x)\|^2$$

Take $\eta_k = \hat{\eta} = \frac{1}{L_g}$, we have

$$\begin{aligned} f(x_{k+1}) &\le f(x_k) - \eta_k\langle \bar{g}_{\mu,\xi}(x), \mathrm{grad}f(x_k)\rangle + \frac{\eta_k^2 L_g}{2}\|\bar{g}_{\mu,\xi}(x)\|^2 \\ &= f(x_k) + \frac{1}{2L_g}\left(\|\bar{g}_{\mu,\xi}(x)-\mathrm{grad}f(x)\|^2 - \|\mathrm{grad}f(x)\|^2\right). \end{aligned}$$

Take the expectation for the random variables at iteration k on both sides, we have

$$\begin{aligned}\mathbb{E}_k f(x_{k+1}) &\le f(x_k) + \frac{1}{2L_g}\left(\mathbb{E}_k\|\bar{g}_{\mu,\xi}(x) - \mathrm{grad} f(x)\|^2 - \|\mathrm{grad} f(x)\|^2\right)\\ Eq.\ (31) &\le f(x_k) + \frac{1}{2L_g}\left(\mu^2 L_g^2 (d+6)^3 + \frac{8(d+4)}{m}\sigma^2 + \left(\frac{8(d+4)}{m} - 1\right)\|\mathrm{grad} f(x)\|^2\right).\end{aligned}$$

Summing up over $k = 0, ..., N$ (assuming that $m \ge 16(d+4)$) yields (37). In summary, the total number of iterations for obtaining an ϵ-stationary solution of (33) is $\mathcal{O}(1/\epsilon^2)$, and the stochastic zeroth-order oracle complexity is $\mathcal{O}(d/\epsilon^4)$.

In Appendix A, we present the oracle complexity of Algorithm 2 when f is geodesically convex and $\mathcal{M}$ is the Hadamard manifold.

3.3. Zeroth-order Stochastic Riemannian Proximal Gradient Method We now consider the general optimization problem of the form in Eq. (1). For the sake of notation, we denote $p(x) := f(x) + h(x)$. We assume that $\mathcal{M}$ is a compact submanifold, h is convex in the embedded space $\mathbb{R}^n$ and is also Lipschitz continuous with parameter L_h, and $f(x) := \int_\xi F(x,\xi)dP(\xi)$ satisfying Assumption 3.

The non-differentiability of h prohibits Riemannian gradient methods to be applied directly. In [21], by assuming that the exact gradient of f is available, a manifold proximal gradient method (ManPG) is proposed for solving (1). One typical iteration of ManPG is as follows:

$$\begin{aligned} v_k &:= \arg\min\ \langle \mathrm{grad} f(x_k), v\rangle + \frac{1}{2t}\|v\|^2 + h(x_k + v),\ \text{s.t.},\ v \in T_{x_k}\mathcal{M}\\ x_{k+1} &:= R_{x_k}(\eta_k v_k),\end{aligned} \tag{39}$$

where $t > 0$ and $\eta_k > 0$ are step sizes. In this section, we develop a zeroth-order counterpart of ManPG (`ZO-ManPG`), where we assume that only noisy function evaluations of f are available. The following lemma from [21] provides a notion of stationary point that is useful for our analysis.

LEMMA 8. *Let $\bar{v}_k$ be the minimizer of the v-subproblem in (39). If $\bar{v}_k = 0$, then x_k is a stationary point of problem (1). We say x_k is an ϵ-stationary point of (1) with $t = \frac{1}{L_g}$, if $\|\bar{v}_k\| \le \epsilon/L_g$.*

Our `ZO-ManPG` iterates as:

$$\begin{aligned} v_k &:= \arg\min\ \langle \bar{g}_{\mu,\xi}(x_k), v\rangle + \frac{1}{2t}\|v\|^2 + h(x_k + v),\ \text{s.t.},\ v \in T_{x_k}\mathcal{M},\\ x_{k+1} &:= R_{x_k}(\eta_k v_k),\end{aligned} \tag{40}$$

where $\bar{g}_{\mu,\xi}(x_k)$ is defined in Eq. (34). Note that the only difference between `ZO-ManPG` (40) and ManPG (39) is that in (40) we use $\bar{g}_{\mu,\xi}(x)$ to replace the Riemannian gradient grad$f$ in (39). A more complete description of the algorithm is given in Algorithm 3. Now we provide some useful lemmas for analyzing the iteration complexity of Algorithm 3.

LEMMA 9. *(Non-expansiveness) Suppose $v := \arg\min_{v\in T_x\mathcal{M}}\langle g_1, v\rangle + \frac{1}{2t}\|v\|^2 + h(x+v)$ and $w := \arg\min_{w\in T_x\mathcal{M}}\langle g_2, w\rangle + \frac{1}{2t}\|w\|^2 + h(x+w)$. Then we have*

$$\|v - w\| \le t\|g_1 - g_2\|. \tag{41}$$

Proof of Lemma 9 By the first order optimality condition [90], we have $0 \in \frac{1}{t}v + g_1 + \mathrm{Proj}_{T_x\mathcal{M}}\partial h(x+v)$ and $0 \in \frac{1}{t}w + g_2 + \mathrm{Proj}_{T_x\mathcal{M}}\partial h(x+w)$, i.e. $\exists p_1 \in \partial h(x+v)$ and $p_2 \in \partial h(x+w)$ such that $v = -t(g_1 + \mathrm{Proj}_{T_x\mathcal{M}}(p_1))$ and $w = -t(g_2 + \mathrm{Proj}_{T_x\mathcal{M}}(p_2))$. Therefore we have

$$\begin{aligned}\langle v, w - v\rangle &= t\langle g_1 + \mathrm{Proj}_{T_x\mathcal{M}}(p_1), v - w\rangle\\ \langle w, v - w\rangle &= t\langle g_2 + \mathrm{Proj}_{T_x\mathcal{M}}(p_2), w - v\rangle.\end{aligned} \tag{42}$$

Algorithm 3 Zeroth-Order Riemannian Proximal Gradient Descent (`ZO-ManPG`)

```
Input: Initial point $x_0$ on $\mathcal{M}$, smoothing parameter $\mu$, number of multi-sample $m$, step size $\eta_k$, fixed number of iteration $N$.
for $k=0$ to $N-1$ do
    Sample $m$ standard Gaussian random vector $u_i$ on $T_{x_k}\mathcal{M}$ by by orthogonal projection in Definition 6, $i=1,...,m$.
    Compute the zeroth-order gradient the random oracle $\bar{g}_\mu(x_k)$ by Eq. (34).
    Solve $v_k$ from Eq. (40).
    Update $x_{k+1}=R_{x_k}(\eta_k v_k)$.
end for
```

Now since $v,w\in T_x\mathcal{M}$, and using the convxity of h, we have

$$\langle \mathrm{Proj}_{T_x\mathcal{M}}(p_1), v-w\rangle = \langle p_1, v-w\rangle = \langle p_1, (v+x)-(w+x)\rangle \geq h(v+x)-h(w+x). \tag{43}$$

Substituting Eq. (42) and into (43) yields,

$$\begin{aligned}\langle v, w-v\rangle &\geq t\langle g_1, v-w\rangle + h(v+x)-h(w+x)\\ \langle w, v-w\rangle &\geq t\langle g_2, w-v\rangle + h(w+x)-h(v+x).\end{aligned}$$

Summing these two inequalities gives $\langle v-w, v-w\rangle \leq t\langle g_2-g_1, v-w\rangle$, and Eq. (41) follows by applying the Cauchy-Schwarz inequality.

Corollary 2. *Suppose v_k is given by (40), and $\bar{v}_k$ is solution of the v-subproblem in Eq. (39), then we have*

$$\mathbb{E}_{\mathcal{U}_k,\Xi_k}\|v_k-\bar{v}_k\|_F^2 \leq t^2\left(\mu^2 L_g^2(d+6)^3 + \frac{8(d+4)}{m}\sigma^2 + \frac{8(d+4)}{m}\|\mathrm{grad} f(x_k)\|^2\right).$$

Proof of Corollary 2 By Lemma 9, we have

$$\mathbb{E}_{\mathcal{U}_k,\Xi_k}\|v_k-\bar{v}_k\|_F^2 \leq t^2\mathbb{E}_{\mathcal{U}_k,\Xi_k}\|\bar{g}_{\mu,\xi}(x_k)-\mathrm{grad} f(x_k)\|_F^2.$$

From Lemma 7,

$$\begin{aligned}&\mathbb{E}_{\mathcal{U}_k,\Xi_k}\|\bar{g}_{\mu,\xi}(x_k)-\mathrm{grad} f(x_k)\|_F^2\\ \leq&\mu^2 L_g^2(d+6)^3 + \frac{8(d+4)}{m}\sigma^2 + \frac{8(d+4)}{m}\|\mathrm{grad} f(x_k)\|^2.\end{aligned}$$

The desired result hence follows by combining these two inequalities.

The following lemma shows the sufficient decrease property for one iteration of `ZO-ManPG`.

Lemma 10. *For any $t>0$, there exists a constant $\bar{\eta}>0$ such that for any $0\leq\eta_k\leq\min\{1,\bar{\eta}\}$, the (x_k,v_k) generated by Algorithm 3 satisfies*

$$p(x_{k+1})-p(x_k)\leq -\left(\frac{\eta_k}{2t}-\tilde{C}\right)\|v_k\|^2, \tag{44}$$

where $\tilde{C}=\mu^2L_g^2(d+6)^3+\frac{8(d+4)}{m}\sigma^2+\frac{8(d+4)}{m}G^2$ and G is the upper bound of the Riemannian gradient $\mathrm{grad} f(x)$ *(existence by the compactness of $\mathcal{M}$).*

Proof of Lemma 10 Notice that

$$\begin{aligned}f(x_{k+1})-f(x_k)&\leq\langle \mathrm{grad} f(x_k), R_{x_k}(\eta_k v_k)-x_k\rangle + \frac{L_g}{2}\|R_{x_k}(\eta_k v_k)-x_k\|^2\\ &=\langle \mathrm{grad} f(x_k)-\bar{g}_{\mu,\xi}(x), R_{x_k}(\eta_k v_k)-x_k\rangle + \langle \bar{g}_{\mu,\xi}(x), R_{x_k}(\eta_k v_k)-x_k\rangle + \frac{L_g}{2}\|R_{x_k}(\eta_k v_k)-x_k\|^2,\end{aligned}$$

where the inequality follows from Assumption 1. Moreover, by Lemma 7 and the Fact 3.6 of [21], we have

$$\begin{aligned}&\langle \mathrm{grad} f(x_k) - \bar{g}_{\mu,\xi}(x), R_{x_k}(\eta_k v_k) - x_k\rangle \le \|\mathrm{grad} f(x_k) - \bar{g}_{\mu,\xi}(x)\| \|R_{x_k}(\eta_k v_k) - x_k\| \\ &\le M_1^2 \eta_k^2 \left[\mu^2 L_g^2 (d+6)^3 + \frac{8(d+4)}{m}\sigma^2 + \frac{8(d+4)}{m}\|\mathrm{grad} f(x)\|^2\right] \|v_k\|^2.\end{aligned}$$

The rest of the proof of bounding $\langle \bar{g}_{\mu,\xi}(x), R_{x_k}(\eta_k v_k) - x_k\rangle + \frac{L_g}{2}\|R_{x_k}(\eta_k v_k) - x_k\|^2$ follows from exactly the same process as in ([21], Lemma 5.2). We omit the details for brevity.

THEOREM 3. *Under Assumption 3 and Assumption 1, the sequence generated by Algorithm 3, with $\eta_k = \hat{\eta} < \min\{1, \bar{\eta}\}$ and $t = 1/L_g$, satisfies:*

$$\frac{1}{N}\sum_{k=0}^{N-1} \mathbb{E}_{\mathcal{U}_k,\Xi_k}\|\bar{v}_k\|^2 \le \frac{4t(p(x_0) - p(x^*))}{(\hat{\eta} - 8\tilde{C})tN} + \frac{\hat{\eta} N t^2}{\hat{\eta} - 8\tilde{C}t}\tilde{C} + \frac{8t^3}{\hat{\eta} - 8\tilde{C}t}\tilde{C}^2, \tag{45}$$

where $\tilde{C} = \mu^2 L_g^2 (d+6)^3 + \frac{8(d+4)}{m}\sigma^2 + \frac{8(d+4)}{m}G^2$ and G is the upper bound of the Riemannian gradient $\mathrm{grad} f(x)$ *over the manifold $\mathcal{M}$. To guarantee*

$$\min_{k=0,\ldots,N-1} \mathbb{E}_{\mathcal{U}_k,\Xi_k}\|\bar{v}_k\|_F^2 \le \epsilon^2 / L_g^2,$$

the parameters need to be set as: $\mu = \mathcal{O}\left(\epsilon/d^{3/2}\right)$, $m = \mathcal{O}\left(dG^2/\epsilon^2\right)$, $N = \mathcal{O}\left(1/\epsilon^2\right)$. Hence, the number of calls to the stochastic zeroth-order oracle is $\mathcal{O}(d/\epsilon^4)$.

Proof of Theorem 3 Summing up (44) over $k = 0, \ldots, N-1$ and using Corollary 2, we have:

$$\begin{aligned}&p(x_0) - \mathbb{E}_{\mathcal{U}_k,\Xi_k} p(x_k) \ge \sum_{k=0}^{N-1}[\frac{\eta_k}{2t} - \tilde{C}]\mathbb{E}_{\mathcal{U}_k}\|v_k\|_F^2 \ge [\frac{\hat{\eta}}{4t} - 2\tilde{C}]\sum_{k=0}^{N-1} 2\mathbb{E}_{\mathcal{U}_k,\Xi_k}\|v_k\|_F^2 \\ &\ge [\frac{\hat{\eta}}{4t} - 2\tilde{C}]\sum_{k=0}^{N-1}\left[\mathbb{E}_{\mathcal{U}_k,\Xi_k}\|\bar{v}_k\|_F^2 - t^2\left(\mu^2 L_g^2 (d+6)^3 + \frac{8(d+4)}{m}\sigma^2\right.\right. \\ &\left.\left. + \frac{8(d+4)}{m}\|\mathrm{grad} f(x_k)\|^2\right)\right] \\ &\ge [\frac{\hat{\eta}}{4t} - 2\tilde{C}]\sum_{k=0}^{N-1}\mathbb{E}_{\mathcal{U}_k,\Xi_k}\|\bar{v}_k\|_F^2 - \frac{\hat{\eta} N t}{4}\left(\mu^2 L_g^2 (d+6)^3 + \frac{8(d+4)}{m}\sigma^2 + \frac{8(d+4)}{m}G^2\right) \\ &+ 2t^2\left(\mu^2 L_g^2 (d+6)^3 + \frac{8(d+4)}{m}\sigma^2 + \frac{8(d+4)}{m}G^2\right)^2,\end{aligned}$$

which immediately implies the desired result (45).

REMARK 3. The subproblem Eq. (40) is the main computational effort in Algorithm 3. Fortunately, this subproblem can be efficiently solved by a regularized semi-smooth Newton's method when $\mathcal{M}$ takes certain forms. We refer the reader to [89, 21] for more details.

3.4. Escaping saddle points: Zeroth-order stochastic cubic regularized Newton's method over Riemannian manifolds In this section, we consider the problem of escaping saddle-points and converging to local minimizers in a stochastic zeroth-order Riemannian setting. Towards that, we leverage the Hessian estimator methodology developed in Section 2.3 and analyze a zeroth-order Riemannian stochastic cubic regularized Newton's method (`ZO-RSCRN`) for solving (33), which provably escapes the saddle points. Our approach is motivated by [95], where the authors proposed the minimization of function $m_{x,\sigma}(\eta) = f(x) + \langle \mathrm{grad} f(x), \eta\rangle + \frac{1}{2}\langle P_x \circ \mathrm{Hess} f(x) \circ P_x[\eta], \eta\rangle + \frac{\alpha_k}{6}\|\eta\|^3$ at each iteration. The zeroth-order counterpart replaces the Riemannian gradient

Algorithm 4 Zeroth-Order Riemannian Stochastic Cubic Regularized Newton's Method (`ZO-RSCRN`)

```
Input: Initial point $x_0$ on $\mathcal{M}$, smoothing parameter $\mu$, multi-sample parameter $m$ and $b$, cubic regularization parameter $\alpha$, number of iteration $N$.
for $k=0$ to $N-1$ do
    Compute $\bar{g}_{\mu,\xi}(x_k)$ and $\bar{H}_{\mu,\xi}(x_k)$ based on (34) and (20) respectively.
    Solve $\eta_k = \mathrm{argmin}_\eta\, \hat{m}_{x_k,\alpha}(\eta)$, where $\hat{m}_{x,\alpha}(\eta)$ is defined in (46).
    Update $x_{k+1} = R_{x_k}(P_x(\eta_k))$.
end for
```

and Hessian with the corresponding zeroth-order estimators. The proposed `ZO-RSCRN` algorithm is described in Algorithm 4. In `ZO-RSCRN`, the function in the cubic regularized subproblem is

$$\hat{m}_{x,\alpha}(\eta) = f(x) + \langle \bar{g}_{\mu,\xi}(x), \eta\rangle + \frac{1}{2}\langle \bar{H}_{\mu,\xi}(x)[\eta], \eta\rangle + \frac{\alpha}{6}\|\eta\|^3. \tag{46}$$

Note that if $\hat{\eta} = \mathrm{argmin}_\eta\, \hat{m}_{x,\alpha}(\eta)$, then the projection $P_x(\hat{\eta})$ is also a minimizer, because $\bar{g}_{\mu,\xi}(x)$ and $\bar{H}_{\mu,\xi}(x)$ only take effect on the component that is in $T_x\mathcal{M}$.

Theorem 4. *For manifold $\mathcal{M}$ and function $f:\mathcal{M}\to\mathbb{R}$ under Assumptions 1, 2 and 3, define $k_{\min} := \arg\min_k \mathbb{E}_{\mathcal{U}_k,\Xi_k}\|\eta_k\|$, then the update in Algorithm 4 with $\alpha \geq L_H$ satisfies:*

$$\mathbb{E}\|g_{k_{\min}+1}\| \leq \mathcal{O}(\epsilon), \ \textit{and}\ \mathbb{E}[\lambda_{\min}(\mathrm{Hess} f_{k_{\min}+1})] \geq -\mathcal{O}(\sqrt{\epsilon}), \tag{47}$$

given that the parameters satisfy:

$$N = \mathcal{O}\left(1/\epsilon^{3/2}\right), \ \mu = \mathcal{O}\left(\min\left\{\frac{\epsilon}{d^{3/2}}, \sqrt{\frac{\epsilon}{d^5}}\right\}\right), \ m = \mathcal{O}(d/\epsilon^2), \ b = \mathcal{O}(d^4/\epsilon), \tag{48}$$

where $\lambda_{\min}$ denotes the smallest eigenvalue. Hence, the zeroth-order oracle complexity is $\mathcal{O}(d/\epsilon^{7/2} + d^4/\epsilon^{5/2})$.

Proof of Theorem 4 Denote $f_k = f(x_k)$, $g_k = \mathrm{grad} f(x_k)$ and $\mathbb{E} = \mathbb{E}_{\mathcal{U}_k,\Xi_k}$ for ease of notation. We first provide the global optimality conditions of subproblem Eq. (46) following [65]:

$$(\bar{H}_{\mu,\xi}(x) + \lambda^* I)\eta + \bar{g}_{\mu,\xi}(x) = 0, \ \lambda^* = \frac{\alpha}{2}\|\eta\|, \ \bar{H}_{\mu,\xi}(x) + \lambda^* I \succeq 0. \tag{49}$$

Since the parallel transport P_η is an isometry, we have

$$\begin{aligned}
&\|g_{k+1}\| = \|P_{\eta_k}^{-1} g_{k+1}\| \\
&= \|(P_{\eta_k}^{-1} g_{k+1} - g_k - \mathrm{Hess} f_k[\eta_k]) + (g_k - \bar{g}_{\mu,\xi}(x_k)) \\
&\quad + (\mathrm{Hess} f_k[\eta_k] - \bar{H}_{\mu,\xi}(x_k)[\eta_k]) + (\bar{g}_{\mu,\xi}(x_k) + \bar{H}_{\mu,\xi}(x_k)[\eta_k])\| \\
&\leq \|P_{\eta_k}^{-1} g_{k+1} - g_k - \mathrm{Hess} f_k[\eta_k]\| + \|g_k - \bar{g}_{\mu,\xi}(x_k)\| \\
&\quad + \|\mathrm{Hess} f_k[\eta_k] - \bar{H}_{\mu,\xi}(x_k)[\eta_k]\| + \|\bar{g}_{\mu,\xi}(x_k) + \bar{H}_{\mu,\xi}(x_k)[\eta_k]\| \\
\textit{Eq. } (12)\ &\leq \frac{L_H}{2}\|\eta_k\|^2 + \|g_k - \bar{g}_{\mu,\xi}(x_k)\| \\
&\quad + \|\mathrm{Hess} f_k[\eta_k] - \bar{H}_{\mu,\xi}(x_k)[\eta_k]\| + \|\bar{g}_{\mu,\xi}(x_k) + \bar{H}_{\mu,\xi}(x_k)[\eta_k]\| \\
\textit{Eq. } (49)\ &= \frac{L_H}{2}\|\eta_k\|^2 + \|g_k - \bar{g}_{\mu,\xi}(x_k)\| + \|\mathrm{Hess} f_k[\eta_k] - \bar{H}_{\mu,\xi}(x_k)[\eta_k]\| + \lambda^*\|\eta_k\| \\
\textit{Eq. } (49)\ &\leq \frac{L_H}{2}\|\eta_k\|^2 + \|g_k - \bar{g}_{\mu,\xi}(x_k)\| + \|\mathrm{Hess} f_k[\eta_k] - \bar{H}_{\mu,\xi}(x_k)\|_{\mathrm{op}}\|\eta_k\| + \frac{\alpha}{2}\|\eta_k\|^2 \\
&\leq \frac{L_H}{2}\|\eta_k\|^2 + \|g_k - \bar{g}_{\mu,\xi}(x_k)\| + \frac{1}{2}\|\mathrm{Hess} f_k - \bar{H}_{\mu,\xi}(x_k)\|_{\mathrm{op}}^2 + \frac{1}{2}\|\eta_k\|^2 + \frac{\alpha}{2}\|\eta_k\|^2.
\end{aligned}$$

Taking expectation on both sides of the above inequality gives (by Eq. (36) and Eq. (21))

$$\mathbb{E}\|g_{k+1}\| - \sqrt{\delta_g} - \delta_H \le \frac{1}{2}(L_H + \alpha + 1 + 2L_2\|g_k\|)\mathbb{E}\|\eta_k\|^2, \tag{50}$$

where $\delta_g = \mu^2 L_g^2 (d+6)^3 + \frac{8(d+4)}{m}(G^2+\sigma^2)$, G is the upper bound of $\|\mathrm{grad} f\|$ over $\mathcal{M}$, and $\delta_H = \frac{(d+16)^4}{b} L_g + \frac{\mu^2 L_H^2}{18}(d+6)^5$. Since $P_{\eta_k}^{-1}$ is an isometry, we have:

$$\begin{aligned}
\lambda_{\min}(\mathrm{Hess} f_{k+1}) =& \lambda_{\min}(P_{\eta_k}^{-1} \circ \mathrm{Hess} f_{k+1} \circ P_{\eta_k}) \\
\ge& \lambda_{\min}(P_{\eta_k}^{-1} \circ \mathrm{Hess} f_{k+1} \circ P_{\eta_k} - \mathrm{Hess} f_k) \\
&+ \lambda_{\min}(\mathrm{Hess} f_k - \bar{H}_{\mu,\xi}(x_k)) + \lambda_{\min}(\bar{H}_{\mu,\xi}(x_k)) \\
Eq.\ (11)\ \ge& - L_H\|\eta_k\| + \lambda_{\min}(\mathrm{Hess} f_k - \bar{H}_{\mu,\xi}(x_k)) + \lambda_{\min}(\bar{H}_{\mu,\xi}(x_k)) \\
=& \lambda_{\min}(\mathrm{Hess} f_k - \bar{H}_{\mu,\xi}(x_k)) + \lambda_{\min}(\bar{H}_{\mu,\xi}(x_k) - L_H\|\eta_k\| I) \\
Eq.\ (49)\ \ge& \lambda_{\min}(\mathrm{Hess} f_k - \bar{H}_{\mu,\xi}(x_k)) - \frac{\alpha + 2L_H}{2}\|\eta_k\|.
\end{aligned}$$

Taking expectation, we obtain (by Eq. (21))

$$\frac{\alpha + 2L_H}{2}\mathbb{E}\|\eta_k\| \ge -(\sqrt{\delta_H} + \mathbb{E}\lambda_{\min}(\mathrm{Hess} f_{k+1})). \tag{51}$$

Now we will upper bound $\mathbb{E}\|\eta_k\|$. From Assumption 2, we have

$$\begin{aligned}
\hat{f}_{x_k}(\eta_k) \le& f(x_k) + g_k^\top \eta_k + \frac{1}{2}\eta_k^\top H_k \eta_k + \frac{L_H}{6}\|\eta_k\|^3 \\
=& \left(f(x_k) + \bar{g}_\mu(x_k)^\top \eta_k + \frac{1}{2}\eta_k^\top \bar{H}_\mu(x_k)\eta_k + \frac{L_H}{6}\|\eta_k\|^3 \right) \\
&+ \left((g_k - \bar{g}_\mu(x_k))^\top \eta_k + \frac{1}{2}\eta_k^\top (H_k - \bar{H}_\mu(x_k))\eta_k \right).
\end{aligned} \tag{52}$$

Using Eq. (49) we have

$$\begin{aligned}
& f(x_k) + \bar{g}_\mu(x_k)^\top \eta_k + \frac{1}{2}\eta_k^\top \bar{H}_\mu(x_k)\eta_k + \frac{L_H}{6}\|\eta_k\|^3 \\
=& f(x_k) - \frac{1}{2}\eta_k^\top \bar{H}_\mu(x_k)\eta_k + (\frac{L_H}{6} - \frac{\alpha}{2})\|\eta_k\|^3 \\
=& f(x_k) - \frac{1}{2}\eta_k^\top (\bar{H}_\mu(x_k) + \frac{\alpha}{2}\|\eta_k\| I)\eta_k - (\frac{\alpha}{4} - \frac{L_H}{6})\|\eta_k\|^3 \\
\le& f(x_k) - (\frac{\alpha}{4} - \frac{L_H}{6})\|\eta_k\|^3 \le f(x_k) - \frac{\alpha}{12}\|\eta_k\|^3,
\end{aligned} \tag{53}$$

where the last inequality is due to $\alpha \ge L_H$. Moreover, by Cauchy-Schwarz inequality and Young's inequality, we have

$$\begin{aligned}
& \mathbb{E}\left[(g_k - \bar{g}_\mu(x_k))^\top \eta_k + \frac{1}{2}\eta_k^\top (H_k - \bar{H}_\mu(x_k))\eta_k \right] \\
\le& \mathbb{E}\|g_k - \bar{g}_\mu(x_k)\|\|\eta_k\| + \frac{1}{2}\mathbb{E}\|H_k - \bar{H}_\mu(x_k)\|_{\mathrm{op}}\|\eta_k\|^2 \\
\le& \frac{32}{3\alpha}\mathbb{E}\|g_k - \bar{g}_\mu(x_k)\|^{3/2} + \frac{12}{\alpha}\mathbb{E}\|H_k - \bar{H}_\mu(x_k)\|_{\mathrm{op}}^3 + \frac{\alpha}{24}\mathbb{E}\|\eta_k\|^3.
\end{aligned} \tag{54}$$

Plugging (53) and (54) to Eq. (52), we have

$$\mathbb{E} f_{k+1} \le f_k - \frac{\alpha}{24}\mathbb{E}\|\eta_k\|^3 + \frac{32}{3L_H}\delta_g^{3/4} + \frac{12}{L_H}\tilde{\delta}_H, \tag{55}$$

Dimension	ϵ	Stepsize	No. iter. `ZO-RGD`	Aver. No. iter. RGD
15×5	10^{-3}	10^{-2}	460 ± 137	442
25×15	10^{-3}	10^{-2}	892 ± 99	852
50×20	10^{-2}	5×10^{-3}	255 ± 26	236

Table 2. Comparison of `ZO-RGD` and RGD on the Procrustes problem.

where $\tilde{\delta}_H = \tilde{C}\frac{(d+16)^6}{b^{3/2}} L_g^{1.5} + \frac{1}{27}\mu^3 L_H^3 (d+6)^{7.5}$. Taking the sum for (55) over $k = 0, \ldots, N-1$, we have

$$\frac{1}{N}\sum_{k=0}^{N} \mathbb{E}\|\eta_k\|^3 \leq \frac{24}{L_H}\left(\frac{f_0 - f^*}{N} + \frac{32}{3L_H}\delta_g^{3/4} + \frac{12}{L_H}\tilde{\delta}_H\right),$$

which together with (48) yields

$$\mathbb{E}\|\eta_{k_{\min}}\|^3 \leq \mathcal{O}(\epsilon^{3/2}), \text{ and } \mathbb{E}\|\eta_{k_{\min}}\|^2 \leq \mathcal{O}(\epsilon). \tag{56}$$

Combining Eq. (56), Eq. (50) and Eq. (51) yields (47).

Remark 4. To solve the subproblem, we implement the same Krylov subspace method as in [2], where the Riemannian Hessian and vector multiplication is approximated by Lanczos iterations. Note also that in our setting, we only require vector-vector multiplications due to the structure of our Hessian estimator in Eq. (15). For the purpose of brevity, we refer to [16, 2] for a comprehensive study of this method.

4. Numerical Experiments and Applications

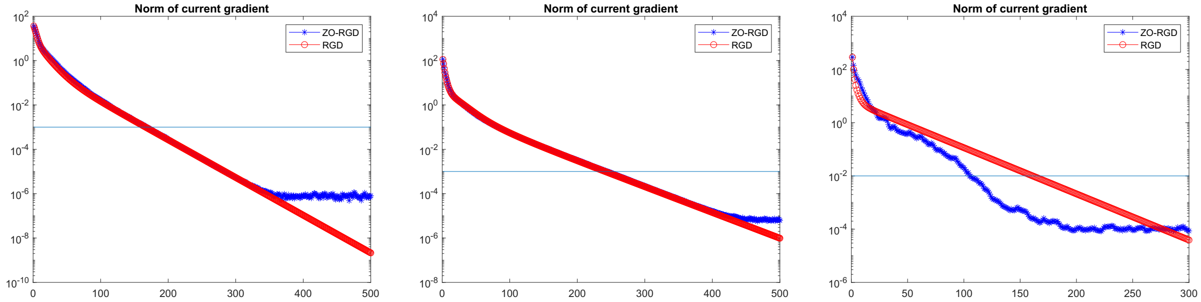
We now explore the performance of the proposed algorithms on various simulation experiments. Finally, we demonstrate the applicability of stochastic zeroth-order Riemannian optimization for the problems of zeroth-order attacks on deep neural networks and controlling stiffness matrix in robotics. We conducted our experiments on a desktop with Intel Core 9600K CPU and NVIDIA GeForce RTX 2070 GPU.

4.1. Simulation Experiments

For all the simulation experiments listed below, we plot the average result over 100 runs.

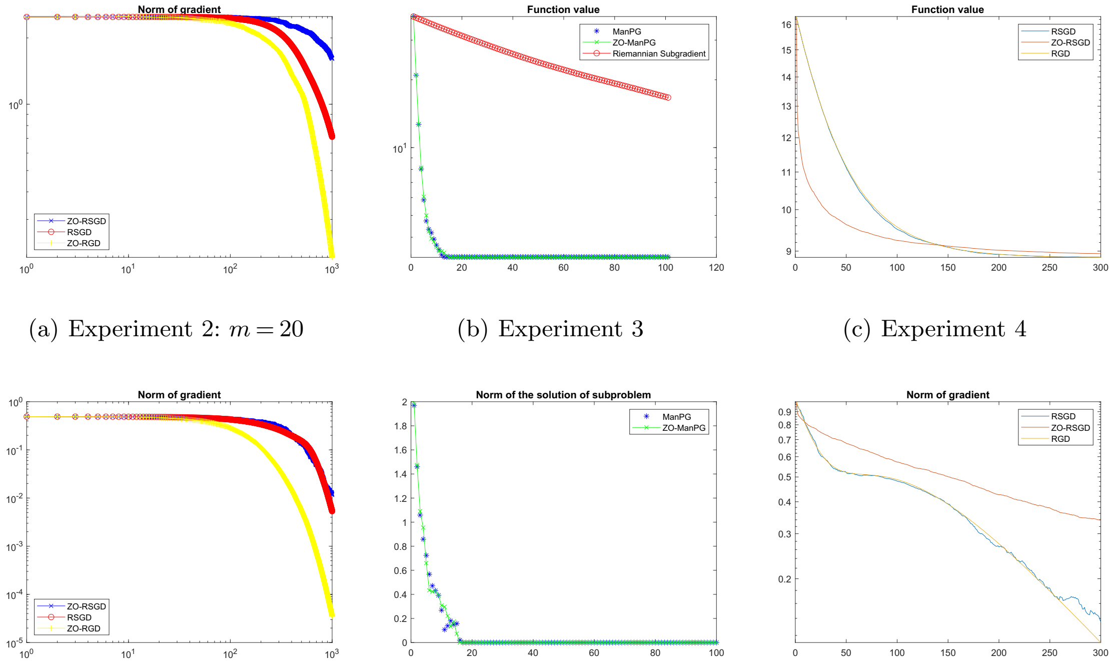
Experiment 1: Procrustes problem [1]. This is a matrix linear regression problem on a given manifold: $\min_{X \in \mathcal{M}} \|AX - B\|_F^2$, where $X \in \mathbb{R}^{n \times p}$, $A \in \mathbb{R}^{l \times n}$ and $B \in \mathbb{R}^{l \times p}$. The manifold we use is the Stiefel manifold $\mathcal{M} = \mathrm{St}(n,p)$. In our experiment, we pick up different dimension $n \times p$ and record the time cost to achieve prescribed precision ϵ. The entries of matrix A are generated by standard Gaussian distribution. We compare our `ZO-RGD` (Algorithm 1) with the first-order Riemannian gradient method (RGD) on this problem. The results are shown in Table 2. For each run, we sample $m = n \times p$ Gaussian samples for each iteration. The multi-sample version of `ZO-RGD` closely resembles the convergence rate of RGD, as shown in Fig. 1. These results indicate our zeroth-order method `ZO-RGD` is comparable with its first-order counterpart RGD, though the former one only uses zeroth-order information.

Experiment 2: k-PCA [92, 79, 96]. k-PCA on Grassmann manifold is a Rayleigh quotient minimization problem. Given a symmetric positive definite matrix $H \in \mathbb{R}^{n \times n}$, we need to solve $\min_{X \in \mathrm{Gr}(n,p)} -\frac{1}{2}\mathrm{Tr}(X^\top H X)$. The Grassmann manifold $\mathrm{Gr}(n,p)$ is the set of p-dimensional subspaces in $\mathbb{R}^n$. We refer the reader to [1] for more details about the Grassmann quotient manifold. This problem can be written as a finite sum problem: $\min_{X \in \mathrm{Gr}(n,p)} \sum_{i=1}^n -\frac{1}{2}\mathrm{Tr}(X^\top h_i h_i^\top X)$, where $h_i \in \mathbb{R}^n$ and $H = \sum_{i=1}^n h_i h_i^\top$. We compare our `ZO-RSGD` algorithm (Algorithm 2) and its first-order counterpart RSGD on this problem. The results are shown in Fig. 2 (a) and (d). In our experiment, we set $n = 100$, $p = 50$, and the matrix H is generated by $H = AA^\top$, where $A \in \mathbb{R}^{n \times p}$ is a normalized randomly generated data matrix. From Fig. 2 (a) and (d), we see that the performance of `ZO-RSGD` is similar to its first-order counterpart RSGD.



(a) $(n,p)=(15,5)$ (b) $(n,p)=(25,15)$ (c) $(n,p)=(50,20)$

Figure 1. The convergence curve of ZO-RGD v.s. RGD. x-axis is the number of iterations and y-axis is the norm of Riemannian gradient at corresponding points. Note that our zeroth-order algorithm does not use gradient information in updates, while the graph still shows the norm of gradient to show the effectiveness of our method. The horizontal lines are the prescribed precisions.



(a) Experiment 2: $m=20$ (b) Experiment 3 (c) Experiment 4

(d) Experiment 2: $m=40$ (e) Experiment 3 (f) Experiment 4

Figure 2. The convergence of three numerical experiments. The x-axis always denotes the number of iterations. Figures (a) and (d) are results for k-PCA (Experiment 2). Here three algorithms are compared: `ZO-RSGD` (Algorithm 2), RSGD, and `ZO-RGD` (Algorithm 1). Figures (b) and (e) are results for sparse PCA (Experiment 3) in which the y-axis of Figure (e) denotes the norm of v_k in (39) (for ManPG) and (40) (for `ZO-ManPG`), which actually measures the optimality of the problem. Here three algorithms are compared: `ZO-ManPG` (Algorithm 3), ManPG and Riemannian subgradient method. Figures (c) and (f) are results for Karcher mean of PSD matrices problem (Experiment 4). Here three algorithms are compared: RSGD, `ZO-RSGD` (Algorithm 2), and RGD.

Experiment 3: Sparse PCA [40, 97, 98]. The sparse PCA problem, arising in statistics, is a Riemannian optimization problem over the Stiefel manifold with nonsmooth objective: $\min_{X\in \mathrm{St}(n,p)} -\frac{1}{2}\mathrm{Tr}(X^\top A^\top AX)+\lambda\|X\|_1$. Here, $A\in\mathbb{R}^{m\times n}$ is the normalized data matrix. We compare our ZO-ManPG (Algorithm 3) with ManPG [21] and Riemannian subgradient method [49]. In our numerical experiments, we chose $(m, n, p) = (50, 100, 10)$, and entries of A are drawn from Gaussian distribution and rows of A are then normalized. The comparison results are shown in Fig. 2 (b) and (e). These results show that our ZO-ManPG is comparable to its first-order counterpart ManPG and they are both much better than the Riemannian subgradient method.

Experiment 4: Karcher mean of given PSD matrices [7, 93, 42]. Given a set of positive semidefinite (PSD) matrices $\{A_i\}_{i=1}^n$ where $A_i \in \mathbb{R}^{d\times d}$ and $A_i \succeq 0$, we want to calculate their Karcher mean: $\min_{X\in\mathcal{S}_{++}^d} \frac{1}{2n}\sum_{i=1}^n (\mathrm{dist}\,(X, A_i))^2$, where $\mathrm{dist}\,(X, Y) = \|\,\mathrm{logm}(X^{-1/2} Y X^{-1/2})\|_F$ (logm stands for matrix logarithm) represents the distance along the corresponding geodesic between the two points $X, Y \in \mathcal{S}_{++}^d$. This experiment serves as an example of optimizing geodesically convex functions over Hadamard manifolds, with ZO-RSGD (Algorithm 2). In our numerical experiment, we take $d = 3$ and $n = 500$. We compare our ZO-RSGD algorithm with its first-order counterpart RSGD and RGD. The results are shown in Fig. 2 (c) and (f), and from these results we see that ZO-RSGD is comparable to its first-order counterpart RSGD in terms of function value, though it is inferior to RSGD and RGD in terms of the size of the gradient.

Experiment 5: Procrustes problem with ZO-RSCRN. Here, we consider the Procrustes problem in Experiment 1 and use the ZO-RSCRN with both estimated gradients and Hessians. Following [2], we use the gradient norm as a performance measure (although the algorithm converges to local-minimzers). We use the Lanczos method (specifically Algorithm 2 from [2]) for solving the sub-problem in Step 4. Furthermore, as we are estimating the second order information, we set $n = 6$ and $p = 4$ and consider $\epsilon = 10^{-3}$. In Figure 3, (a), we plot the gradient norm versus iterations for Riemannian Stochastic Cubic-Regularized Newton method in the zeroth order and second-order setting. We notice that the zeroth-order method compares favourably to the second-order counterpart in terms of iteration complexity. Admittedly, scaling up the ZO-RSCRN method to work in higher-dimensions, based on variance reduction techniques, is an interesting problem that we plan to tackle as future work.

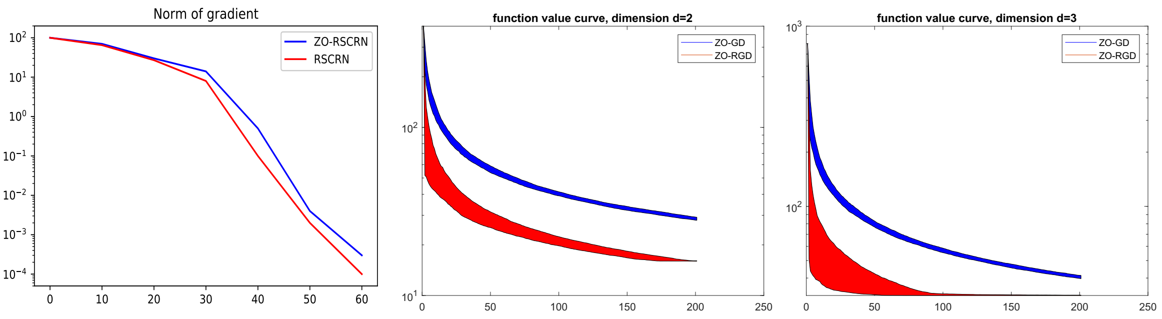
4.2. Real world applications

Black-box stiffness control for robotics. We now study the first motivating example discussed in Section 1.2.1 on the control of robotics with the policy parameter being the stiffness matrix $\boldsymbol{K}^{\mathcal{P}} \in \mathcal{S}_{++}^d$, see [39] for more engineering details. Mathematically, given the current position of robot $\hat{\boldsymbol{p}}$ and current speed $\dot{\boldsymbol{p}}$, the task is to minimize

$$f(\boldsymbol{K}^{\mathcal{P}}) = w_{\boldsymbol{p}}\|\hat{\boldsymbol{p}} - \boldsymbol{p}\|^2 + w_d \det(\boldsymbol{K}^{\mathcal{P}}) + w_c \,\mathrm{cond}(\boldsymbol{K}^{\mathcal{P}}) \tag{57}$$

with $\boldsymbol{p}$ being the new position, and cond is the condition number. With a constant external force $\boldsymbol{f}^e$ applied to the system, we have the following identity which solves $\boldsymbol{p}$ by $\boldsymbol{K}^{\mathcal{P}}$: $\boldsymbol{f}^e = \boldsymbol{K}^{\mathcal{P}}(\hat{\boldsymbol{p}} - \boldsymbol{p}) - \boldsymbol{K}^{\mathcal{D}}\dot{\boldsymbol{p}}$, where the damping matrix $\boldsymbol{K}^{\mathcal{D}} = \boldsymbol{K}^{\mathcal{P}}$ for critical damped case. As the stiffness matrix is a positive definite matrix, the above optimization problem is a Riemannian optimization problem over the positive definite manifold (where the manifold structure is the same as the Karcher mean problem). The function f is not known analytically and following [39], we use a simulated setting for a robot (7-DOF Franka Emika Panda robot) to evaluate the function f for a given value of $\boldsymbol{K}^{\mathcal{P}}$, with the same parameters as in [39]. We compare our ZO-RGD method with Euclidean Zeroth-order gradient descent (ZO-GD) method [4]. We test the cases when $d = 2$ and $d = 3$ for minimizing function f w.r.t $\boldsymbol{K}^{\mathcal{P}}$, and the results are shown in Figure 3, (b) and (c). In our experiments, the stepsize of ZO-GD is 3×10^{-4} and ZO-RGD is 10^{-3}. Note that for ZO-GD method, one has to project the matrix back to the positive definite set, whereas the ZO-RGD method intrinsically guarantees that the iterates are positive definite, thus is much more stable. Also, due to the fact that ZO-RGD is more stable, the stepsize of ZO-RGD can be larger than ZO-GD, which results in faster convergence.

Zeroth-order black-box attack on Deep Neural Networks (DNNs). We now return to the motivating example described in Section 1.2.2 and propose our black-box attack algorithm, as



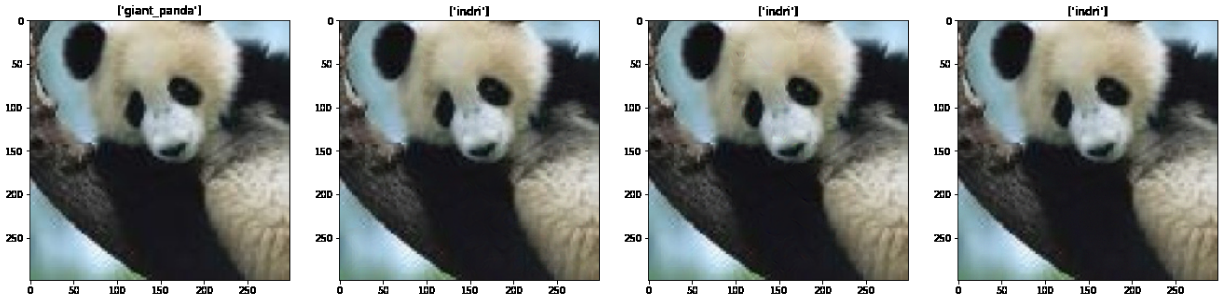
(a) SCRN (b) 2d case (c) 3d case

FIGURE 3. Figure (a) corresponds to Experiment 5. Figures (b) and (c) correspond to the experiments on the robotic minimization function in (57). The x-axis in all figures correspond to iteration number.

stated in Algorithm 6 (in Appendix D). For the sake of comparison, we also assume the architecture of the DNN is known and use "white-box" attacks based on first-order Riemannian optimization methods (Algorithm 5) and compare against the PGD attack [53], which does not explicitly enforce any constraints on the perturbed training data. For simplicity, we assume the manifold is a sphere. That is, we assume that the perturbation set S is given by $S(R) = \{\delta : \|\delta\|_2 = R\}$, where R is the radius of the sphere. This is consistent with the optimal ℓ_2-norm attack studied in the literature [52]. Furthermore, the sphere constraint guarantees that the perturbed image is always in a certain distance from the original image. We start our zeroth-order attack from a perturbation and maximize the loss function on the sphere. For the black-box method, to accelerate the convergence, we use Euclidean zeroth-order optimization to find an appropriate initial perturbation (Algorithm 7). It is worth noting that the zeroth-order attack in [19, 80] has a non-smooth objective function, which has $\mathcal{O}(n^3/\epsilon^3)$ complexity to guarantee convergence [66], whereas the complexity needed for our method is $\mathcal{O}(d/\epsilon^2)$.

We first tested our method on the giant panda picture in the Imagenet data set [25], with the network structure the Inception v3 network structure [77]. The attack radius in our algorithm is proportional to the ℓ_2 norm of the original image. Both white-box and black-box Riemannian attacks are successful, which means that they both converge to images that lie in a different image class (i.e. with a different label), see Figure 4. We also tested Algorithms 5 and 6 on the CIFAR10 dataset, and the network structure we used is the VGG net [72]. The corresponding results are provided in Appendix D.

5. Conclusions In this paper, we proposed zeroth-order algorithms for solving Riemannian optimization over submanifolds embedded in Euclidean space in which only noisy function evaluations are available for the objective. These algorithms adopt new estimators of the Riemannian gradient and Hessian from noisy objective function evaluations, based on a Riemannian version of the Gaussian smoothing technique. The proposed estimators overcome the difficulty of the non-linearity of the manifold constraint and the issues that arise in using Euclidean Gaussian smoothing techniques when the function is defined only over the manifold. The iteration complexity and oracle complexity of the proposed algorithms are analyzed for obtaining an appropriately defined ϵ-stationary point or ϵ-approximate local minimum. The established complexities are independent of the dimension of the ambient Euclidean space and only depend on the intrinsic dimension of the manifold. Numerical experiments demonstrated that the proposed zeroth-order algorithms are comparable to their first-order counterparts.



(a) Original image (b) PGD attack (c) First-order attack on the sphere (d) Zeroth-order attack on the sphere

Figure 4. The attack on giant panda picture [25]. (a): the original image; (b): the PGD attack with a small diameter; (c) Riemannian attack (Algorithm 5) on the sphere with the same diameter; (d): Riemannian zeroth-order attack (Algorithm 6). 'Indri' refers to the class to which the original image is misclassified to.

Appendix A: Geodesically Convex Problem In this section we consider the smooth problem (33) where f is geodesically convex. The definition of geodesic convexity is given below (see, e.g., [93]).

Definition 8. A function $f:\mathcal{M}\to\mathbb{R}$ is geodesically convex if for all $x,y\in\mathcal{M}$, there exists a geodesic γ such that $\gamma(0)=x$, $\gamma(1)=y$ and $\forall t\in[0,1]$ we have $f(\gamma(t))\le(1-t)f(x)+tf(y)$.
It can be shown that this definition is equivalent to, $f(\mathrm{Exp}_x(\eta))\ge f(x)+\langle g_x,\eta\rangle_x$, $\forall\eta\in T_x\mathcal{M}$, where g_x is a subgradient of f at x, Exp is the exponential mapping, and $\langle\cdot,\cdot\rangle_x$ is the inner product in $T_x\mathcal{M}$ induced by Riemannian metric $d(\cdot,\cdot)$. When f is smooth, we have $g_x=\mathrm{grad}f(x)$, the Riemannian gradient at x. It is known that geodesically convex function is a constant on compact manifolds. Therefore, in this subsection, we assume that $\mathcal{M}$ is an Hadamard manifold [8, 34], and $\mathcal{X}$ is a bounded and geodesically convex subset of $\mathcal{M}$.

Assumption 4. *The subset $\mathcal{X}$ of Hadamard manifold $\mathcal{M}$ is bounded by diameter D, and the sectional curvature is lower bounded by ϱ. The function $F(x,\xi)$ is geodesically convex w.r.t. $x\in\mathcal{M}$, almost everywhere for ξ (and hence f is geodesically convex).*

The following lemma from [93] is useful for our subsequent analysis. Here $\mathcal{P}_{\mathcal{X}}$ denotes the projection onto $\mathcal{X}$, i.e., $\mathcal{P}_{\mathcal{X}}(x):=\{y\in\mathcal{X}:d(x,y)=\inf_{z\in\mathcal{X}}d(x,z)\}$.

Lemma 11 (**[93]**). *For any Riemannian manifold $\mathcal{M}$ where the sectional curvature is lower bounded by ϱ and any points x, $x_s\in\mathcal{M}$, the update*

$$x_{s+1}=\mathcal{P}_{\mathcal{X}}(\mathrm{Exp}_{x_s}(-\eta_s g_s))$$

satisfies: $\langle -g_s,x-x_s\rangle\le\frac{1}{2\eta_s}(d^2(x_s,x)-d^2(x_{s+1},x))+\frac{\zeta(\varrho,d(x_s,x))\eta_s}{2}\|g_s\|^2$, where $d(\cdot,\cdot)$ is the Riemannian metric defined globally on $\mathcal{M}$, and $\zeta(\varrho,c):=c\sqrt{|\varrho|}/\tanh(c\sqrt{|\varrho|})$.

In this subsection, we consider the following algorithm, which is a special case of Algorithm 2.

$$x_{k+1}=\mathcal{P}_{\mathcal{X}}(\mathrm{Exp}_{x_k}(-\eta_k\bar{g}_{\mu,\xi}(x_k))). \tag{58}$$

We now present our result for obtaining an ϵ-optimal solution of (33).

Theorem 5. *Let the manifold $\mathcal{M}$ and the function $f:\mathcal{M}\to\mathbb{R}$ satisfy Assumptions 1, 3, and 4. Suppose Algorithm 2 is run with the update in Eq. (58) and with $\eta_k=1/L_g$. Denote $\Delta_k=\mathbb{E}_{\mathcal{U}_k,\Xi_k}(f(x_k)-f^*)$. To have $\min_{1\le k\le t}\Delta_k\le\epsilon$, we need the smoothing parameter μ, number of sampling at each iteration m and the number of iteration t to be respectively of order:*

$$\mu=\mathcal{O}(\sqrt{\epsilon}/d^{3/2}),\ m=\mathcal{O}(d/\epsilon),\ t=\mathcal{O}(1/\epsilon). \tag{59}$$

Hence, the zeroth-order oracle complexity is $N = mt = \mathcal{O}(d/\epsilon^2)$.

Proof of Theorem 5 From Assumption 1 we have that:

$$f(x_{k+1}) - f(x_k) \le -\eta_k \langle \mathrm{grad} f(x_k), \bar{g}_{\mu,\xi}(x_k)\rangle + \frac{L_g}{2}\eta_k^2 \|\bar{g}_{\mu,\xi}(x_k)\|^2.$$

Taking $\eta_k = \frac{1}{L_g}$, we have

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\le \frac{1}{2L_g}\left(-2\langle \mathrm{grad} f(x_k), \bar{g}_{\mu,\xi}(x_k)\rangle + \|\bar{g}_{\mu,\xi}(x_k)\|^2\right) \\ &= \frac{1}{2L_g}\left(\|\bar{g}_{\mu,\xi}(x_k) - \mathrm{grad} f(x_k)\|^2 - \|\mathrm{grad} f(x_k)\|^2\right). \end{aligned}$$

Taking expectation with respect to u_k on both sides of the inequality above and taking $m \ge 16(d+4)$, we have (by Eq. (36))

$$\begin{aligned} &\mathbb{E}_{u_k} f(x_{k+1}) - f(x_k) \\ \le& \frac{1}{2L_g}\left(\mu^2 L_g^2 (d+6)^3 + \frac{8(d+4)}{m}\sigma^2 + \left(\frac{8(d+4)}{m} - 1\right)\|\mathrm{grad} f(x_k)\|^2\right) \\ \le& \frac{\mu^2 L_g (d+6)^3}{2} + \frac{4(d+4)}{mL_g}\sigma^2 - \frac{1}{4L_g}\|\mathrm{grad} f(x_k)\|^2. \end{aligned} \tag{60}$$

Now considering the geodesic convexity and Lemma 11, we have

$$f(x_{k+1}) - f^* \le \langle -\bar{g}_{\mu,\xi}(x_k), \mathrm{Exp}_{x_k}^{-1}(x^*)\rangle \le \frac{L_g}{2}(d^2(x_k, x^*) - d^2(x_{k+1}, x^*)) + \frac{\zeta(\varrho, D)\|\bar{g}_{\mu,\xi}(x_k)\|^2}{2L_g}. \tag{61}$$

From Lemma 7 we have

$$\begin{aligned} &\mathbb{E}\|\bar{g}_{\mu,\xi}(x_k)\|^2 \le 2\mathbb{E}\|\bar{g}_{\mu,\xi}(x_k) - \mathrm{grad} f(x_k)\|^2 + 2\mathbb{E}\|\mathrm{grad} f(x_k)\|^2 \\ \le& 2\mu^2 L_g^2 (d+6)^3 + \frac{16(d+4)}{m}\sigma^2 + \left(\frac{16(d+4)}{m} + 2\right)\|\mathrm{grad} f(x_k)\|^2. \end{aligned} \tag{62}$$

Now take the expectation w.r.t. u_k for both sides of (61), and combine with (62), we have

$$\Delta_{k+1} \le \frac{L_g}{2}(d^2(x_k, x^*) - d^2(x_{k+1}, x^*)) + \frac{\zeta(\varrho, D)}{2L_g}\left(2\mu^2 L_g^2 (d+6)^3 + \frac{16(d+4)}{m}\sigma^2 + 3\|\mathrm{grad} f(x_k)\|^2\right). \tag{63}$$

Multiplying (63) with $\frac{1}{6\zeta(\varrho, D)}$, and sum up with Eq. (60), we have

$$\left(1 + \frac{1}{6\zeta}\right)\Delta_{k+1} - \Delta_k \le \frac{L_g}{12\zeta}(d^2(x_k, x^*) - d^2(x_{k+1}, x^*)) + \mu^2 L_g (d+6)^3 + \frac{16(d+4)}{3mL_g}\sigma^2.$$

Summing it over $k = 0, \ldots, t-1$ we have

$$\Delta_t - \Delta_0 + \frac{1}{6\zeta}\sum_{k=1}^{t}\Delta_k \le \frac{L_g}{12\zeta}d^2(x_0, x^*) + (\mu^2 L_g (d+6)^3 + \frac{16(d+4)}{3mL_g}\sigma^2)t.$$

Equivalently, we have

$$\frac{1}{t}\sum_{k=1}^{t}\Delta_k \le \frac{L_g}{2t}d^2(x_0, x^*) + 6\zeta(\mu^2 L_g (d+6)^3 + \frac{16(d+4)}{3mL_g}\sigma^2) + \frac{6\zeta}{t}\Delta_0,$$

which together with (59) yields $\min_{1 \le k \le t}\Delta_k \le \epsilon$.

Appendix B: Proof of Remark 2

Proof of the improved bound Eq. (31) Since $\bar{g}_\mu(x) = \frac{1}{m}\sum_{i=1}^m g_{\mu,i}(x)$, we have (denote $\mathcal{U} = \{u_1,...,u_m\}$):

$$\begin{aligned}
&\mathbb{E}_{\mathcal{U}}\|\bar{g}_\mu(x)-\mathrm{grad}f(x)\|^2\\
\leq&2\mathbb{E}_{\mathcal{U}}\|\bar{g}_\mu(x)-\mathbb{E}_{\mathcal{U}}\bar{g}_\mu(x)\|^2+2\|\mathbb{E}_{\mathcal{U}}\bar{g}_\mu(x)-\mathrm{grad}f(x)\|^2\\
=&2\mathbb{E}_{\mathcal{U}}\left\|\frac{1}{m}\sum_{i=1}^m[g_{\mu,i}(x)-\mathbb{E}_{\mathcal{U}}g_{\mu,i}(x)]\right\|^2+2\left\|\frac{1}{m}\sum_{i=1}^m[\mathbb{E}_{\mathcal{U}}g_{\mu,i}(x)-\mathrm{grad}f(x)]\right\|^2\\
=&\frac{2}{m^2}\mathbb{E}_{\mathcal{U}}\sum_{i=1}^m\|g_{\mu,i}(x)-\mathbb{E}_{\mathcal{U}}g_{\mu,i}(x)\|^2+\frac{2}{m^2}\left\|\sum_{i=1}^m[\mathbb{E}_{\mathcal{U}}g_{\mu,i}(x)-\mathrm{grad}f(x)]\right\|^2\\
\leq&\frac{2}{m}\mathbb{E}_{u_1}\|g_{\mu,1}(x)-\mathbb{E}_{\mathcal{U}}g_{\mu,1}(x)\|^2+2\|\mathbb{E}_{u_1}g_{\mu,1}(x)-\mathrm{grad}f(x)\|^2\\
\leq&\frac{2}{m}\mathbb{E}_{u_1}\|g_{\mu,1}(x)\|^2+\frac{\mu^2L_g^2}{2}(d+3)^3\\
\leq&\frac{\mu^2}{m}L_g^2(d+6)^3+\frac{4(d+4)}{m}\|\mathrm{grad}f(x)\|^2+\frac{\mu^2L_g^2}{2}(d+3)^3\\
\leq&\mu^2L_g^2(d+6)^3+\frac{4(d+4)}{m}\|\mathrm{grad}f(x)\|^2,
\end{aligned}$$

where the second equality is from the fact that u_i and u_j are independent when $i \neq j$.

Proof of (32) Following the L_g-retraction-smooth, we have: $f(x_{k+1}) \leq f(x_k) - \eta_k\langle\bar{g}_\mu(x), \mathrm{grad}f(x_k)\rangle + \frac{\eta_k^2 L_g}{2}\|\bar{g}_\mu(x)\|^2$. Taking $\eta_k = \hat{\eta} = 1/L_g$, we have

$$\begin{aligned}
f(x_{k+1})&\leq f(x_k)-\eta_k\langle\bar{g}_\mu(x),\mathrm{grad}f(x_k)\rangle+\frac{\eta_k^2L_g}{2}\|\bar{g}_\mu(x)\|^2\\
&=f(x_k)+\frac{1}{2L_g}\left(\|\bar{g}_\mu(x)-\mathrm{grad}f(x)\|^2-\|\mathrm{grad}f(x)\|^2\right).
\end{aligned}$$

Now take the expectation for the random variables at the iteration k on both sides, we have

$$\begin{aligned}
\mathbb{E}_kf(x_{k+1})&\leq f(x_k)+\frac{1}{2L_g}\left(\mathbb{E}_k\|\bar{g}_\mu(x)-\mathrm{grad}f(x)\|^2-\|\mathrm{grad}f(x)\|^2\right)\\
Eq.\ (36)&\leq f(x_k)+\frac{1}{2L_g}\left(\mu^2L_g^2(d+6)^3+\left(\frac{4(d+4)}{m}-1\right)\|\mathrm{grad}f(x)\|^2\right).
\end{aligned}$$

By choosing $m \geq 8(d+4)$, summing the above inequality over $k = 0,\dots,N$ gives (32).

Appendix C: Proof of Lemma 7

For the sake of notation, here we denote $\mathbb{E} = \mathbb{E}_{u_0}$. From (34) we have

$$\mathbb{E}(\|g_{\mu,\xi}(x)\|^2)=\frac{1}{\mu^2}\mathbb{E}\left[(F(R_x(\mu u,\xi))-F(x,\xi))^2\|u\|^2\right]. \tag{64}$$

From Assumption 1 we have

$$\begin{aligned}
&(F(R_x(\mu u,\xi))-F(x,\xi))^2\\
=&(F(R_x(\mu u,\xi))-F(x,\xi)-\mu\langle\mathrm{grad}F(x,\xi),u\rangle+\mu\langle\mathrm{grad}F(x,\xi),u\rangle)^2\\
\leq&2\left(\frac{L_g}{2}\mu^2\|u\|^2\right)^2+2\mu^2\langle\mathrm{grad}F(x,\xi),u\rangle^2.
\end{aligned}\tag{65}$$

Combining (64) and (65) yields

$$\begin{aligned}
\mathbb{E}(\|g_{\mu,\xi}(x)\|^2)&\leq\frac{\mu^2}{2}L_g^2\mathbb{E}(\|u\|^6)+2\mathbb{E}(\|\langle\mathrm{grad}F(x,\xi),u\rangle u\|^2)\\
(Corollary\ 1)&\leq\frac{\mu^2}{2}L_g^2(d+6)^3+2\mathbb{E}(\|\langle\mathrm{grad}F(x,\xi),u\rangle u\|^2).
\end{aligned}\tag{66}$$

Denote our d-dimensional tangent space as $\mathcal{X}$. Without loss of generality, suppose $\mathcal{X}$ is the subspace generated by projecting onto the first d coordinates, i.e., $\forall x \in \mathcal{X}$, the last $(n-d)$ elements of x are zeros. Also for brevity, denote $g = \mathrm{grad}F(x,\xi)$. Use x_i to denote the i-th coordinate of u_0, and $\kappa(d)$ denote the normalization constant for d-dimensional Gaussian distribution. For simplicity, denote $x = (x_1, ..., x_d)$. We have

$$\begin{aligned}
&\mathbb{E}(\|\langle \mathrm{grad}F(x,\xi), u\rangle u\|^2) = \frac{1}{\kappa}\int_{\mathbb{R}^n} \langle \mathrm{grad}F(x,\xi), u\rangle^2 \|u\|^2 e^{-\frac{1}{2}\|u_0\|^2} du_0 \\
=&\frac{1}{\kappa(d)}\int_{\mathbb{R}^d}\left(\sum_{i=1}^d g_i x_i\right)^2\left(\sum_{i=1}^d x_i^2\right) e^{-\frac{1}{2}\sum_{i=1}^d x_i^2} dx_1 \cdots dx_d, \\
=&\frac{1}{\kappa(d)}\int_{\mathbb{R}^d}\langle g, x\rangle^2 \|x\|^2 e^{-\frac{1}{2}\|x\|^2} dx = \frac{1}{\kappa(d)}\int_{\mathbb{R}^d}\|x\|^2 e^{-\frac{\tau}{2}\|x\|^2}\langle g, x\rangle^2 e^{-\frac{1-\tau}{2}\|x\|^2} dx \\
\le&\frac{2}{\kappa(d)\tau e}\int_{\mathbb{R}^d}\langle g, x\rangle^2 e^{-\frac{1-\tau}{2}\|x\|^2} dx = \frac{2}{\kappa(d)\tau(1-\tau)^{1+d/2} e}\int_{\mathbb{R}^d}\langle g, x\rangle^2 e^{-\frac{1}{2}\|x\|^2} dx \\
=&\frac{2}{\tau(1-\tau)^{1+d/2} e}\|g\|^2 \le (d+4)\|g\|^2,
\end{aligned} \tag{67}$$

where the last $n-d$ dimensions of u_0 are integrated to be one, the first inequality is due to the following fact: $x^p e^{-\frac{\tau}{2}x^2} \le (\frac{2}{\tau e})^{p/2}$, and the second inequality follows by setting $\tau = \frac{2}{(d+4)}$. From Assumption 3, we have

$$\mathbb{E}_\xi\|\mathrm{grad}F(x,\xi)\|^2 \le 2\mathbb{E}_\xi\|\mathrm{grad}F(x,\xi) - \mathrm{grad}f(x)\|^2 + 2\|\mathrm{grad}f(x)\|^2 \le 2\sigma^2 + 2\|\mathrm{grad}f(x)\|^2. \tag{68}$$

Combining (66), (67) and (68) yields

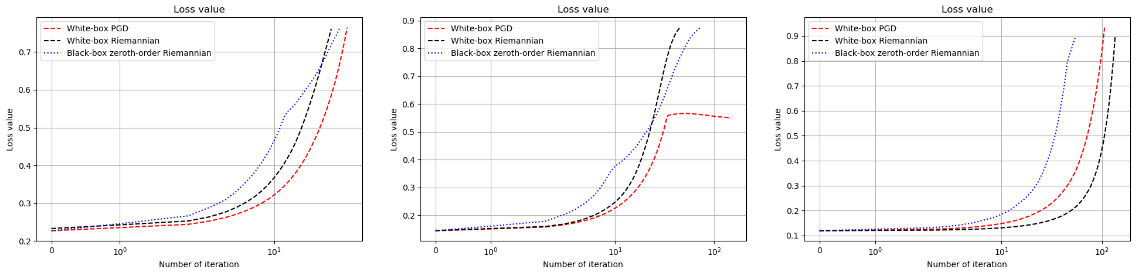
$$\begin{aligned}
\mathbb{E}_\xi\left[\mathbb{E}_{u_0}(\|g_{\mu,\xi}(x)\|^2)\right] &\le \mathbb{E}_\xi\left[\frac{\mu^2}{2}L_g^2(d+6)^3 + 2(d+4)\|\mathrm{grad}F(x,\xi)\|^2\right] \\
&\le \frac{\mu^2}{2}L_g^2(d+6)^3 + 4(d+4)(\sigma^2 + \|\mathrm{grad}f(x)\|^2).
\end{aligned} \tag{69}$$

Finally, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{U},\xi}\|\bar{g}_{\mu,\xi}(x) - \mathrm{grad}f(x)\|^2 \\
\le&2\mathbb{E}_{\mathcal{U},\xi}\|\bar{g}_{\mu,\xi}(x) - \mathbb{E}_{\mathcal{U},\xi}g_\mu(x)\|^2 + 2\|\mathbb{E}_{\mathcal{U},\xi}g_\mu(x) - \mathrm{grad}f(x)\|^2 \\
\le&\frac{2}{m}\mathbb{E}_{u_1,\xi_1}\|g_{\mu_1,\xi_1}(x) - \mathbb{E}_{u_1,\xi_1}g_{\mu_1,\xi_1}(x)\|^2 + \frac{\mu^2 L_g^2}{2}(d+3)^3 \\
\le&\frac{2}{m}\mathbb{E}_{u_1,\xi_1}\|g_{\mu_1,\xi_1}(x)\|^2 + \frac{\mu^2 L_g^2}{2}(d+3)^3 \\
\le&\frac{2}{m}\left(\frac{\mu^2 L_g^2}{2}(d+6)^3 + 4(d+4)[\|\mathrm{grad}f(x)\|^2 + \sigma^2]\right) + \frac{\mu^2 L_g^2}{2}(d+3)^3 \\
\le&\mu^2 L_g^2(d+6)^3 + \frac{8(d+4)}{m}\sigma^2 + \frac{8(d+4)}{m}\|\mathrm{grad}f(x)\|^2,
\end{aligned}$$

where the second inequality is from Proposition 1 and the fourth inequality is from (69).

Appendix D: Implementation Details of Black-Box Attacks Here we provide our white and black-box Riemannian attack algorithm in Algorithms 5 and 6, respectively. For the black-box attack, to accelerate convergence, we introduce a pre-attack step to search for a sufficiently large loss value on the prescribed sphere, still in a black-box manner (only use the function value). For further acceleration of the black-box attack, the hierarchical attack [19] and the auto-encoder technique [80] might be applicable. The attack results on CIFAR-10 images are shown in Figure 6. We also provide the loss function curve in Figure 5. Again the network structure we used is the VGG net [72]. It



(a) (b) (c)

FIGURE 5. Loss function versus the iteration numbers. We observe that the loss function increases while performing our attacks. The three figures correspond to the last three rows in Figure 6. For the failed the PGD attack, we notice that the function value stuck in the middle and then decreased, while white and black-box Riemannian attack increased loss value successfully.

can be seen from Figure 6 that our black-box attack yields similar attack result as the PGD attack, however PGD failed in one of the five images, and the loss function curve in Figure 5 indicates that the loss function curve of PGD attack stagnated. This may be due to an inappropriate choice of parameters for the PGD attack. Nevertheless this experiment shows the ability of our Riemannian attack method. We also mention here that the Riemannian attack is in general slower than the PGD attack due to the multi-sampling technique discussed in Remark 2.

Algorithm 5 White-box attack via Riemannian optimization

```
Input: Original image $\tilde{x}$, original label $y_0$, radius of the attack region $R$, step size $\eta_k$, some convergence criterion.
Randomly sample $\delta$ s.t. $\|\delta\| = R$.
Set the initial point $x_0 = \tilde{x} + \delta$, $k = 0$.
repeat
    Update $x_{k+1} = R_{x_k}(\eta_k \mathrm{grad}_x L(\theta, x, y))$, $k = k + 1$.
until Convergence criterion is met.
```

Algorithm 6 Black-box attack via Riemannian zeroth-order optimization

```
Input: Original image $\tilde{x}$, original label $y_0$, radius of the attack region $R$, step size $\eta_k$, smoothing parameter $\mu$, number of multi-sample $m$, some convergence criterion.
Obtain $\delta$ (initial perturbation) by pre-training steps (Algorithm 7).
Set the initial point $x_0 = \tilde{x} + \delta$, $k = 0$.
repeat
    Sample $m$ standard Gaussian random matrix $u_i$ on $T_{x_k}\mathcal{M}$.
    Set the random oracle $\bar{g}_\mu(x_k)$ by (34).
    Update $x_{k+1} = R_{x_k}(\eta_k \bar{g}_\mu(x_k))$, $k = k + 1$.
until Convergence criterion is met.
```

Acknowledgments. JL and SM acknowledge the support by NSF grants DMS-1953210 and CCF-2007797. KB and SM acknowledge the support by UC Davis CeDAR (Center for Data Science and Artificial Intelligence Research) Innovative Data Science Seed Funding Program. JL also wants to thank Tesi Xiao for helpful discussions on black-box attacks.

Algorithm 7 Pre-training step for black-box attack

```
Input: Original image $\tilde{x}$, original label $y_0$, radius of the attack region $R$, step size $\eta_k$, smoothing parameter $\mu$, number of multi-sample $m$.
$x_0 = \tilde{x}$.
repeat
    Sample $m$ standard Gaussian random matrix $u_i$.
    Set the random oracle $\bar{g}_\mu(x_k)$ by (34), with $g_{\mu_i}(x) = \frac{f(x+\mu u_i)-f(x)}{\mu} u_i$
    Update $x_{k+1} = x_k + \eta_k \bar{g}_\mu(x_k)$.
    Update $\delta = x_{k+1} - \tilde{x}$, $k = k+1$.
until $\|\delta\| \geq R$
$\delta = \frac{\delta}{\|\delta\|} R$
```

References

[1] Absil PA, Mahony R, Sepulchre R (2009) *Optimization algorithms on matrix manifolds* (Princeton University Press).

[2] Agarwal N, Boumal N, Bullins B, Cartis C (2020) Adaptive regularization with cubics on manifolds. *Mathematical Programming (to appear)* .

[3] Audet C, Hare W (2017) *Derivative-free and blackbox optimization* (Springer).

[4] Balasubramanian K, Ghadimi S (2019) Zeroth-order nonconvex stochastic optimization: Handling constraints, high-dimensionality, and saddle-points. *arXiv preprint arXiv:1809.06474* 651–676.

[5] Bendory T, Eldar YC, Boumal N (2017) Non-convex phase retrieval from STFT measurements. *IEEE Transactions on Information Theory* 64(1):467–484.

[6] Bento GC, Ferreira OP, Melo JG (2017) Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds. *Journal of Optimization Theory and Applications* 173(2):548–562.

[7] Bini DA, Iannazzo B (2013) Computing the Karcher mean of symmetric positive definite matrices. *Linear Algebra and its Applications* 438(4):1700–1710.

[8] Bishop RL, O'Neill B (1969) Manifolds of negative curvature. *Transactions of the American Mathematical Society* 145:1–49.

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

[10] Borckmans PB, Selvan SE, Boumal N, Absil PA (2014) A Riemannian subgradient algorithm for economic dispatch with valve-point effect. *Journal of Computational and Applied Mathematics* 255:848–866.

[11] Boumal N, Absil P (2011) RTRMC: A Riemannian trust-region method for low-rank matrix completion. *Advances in neural information processing systems*, 406–414.

[12] Boumal N, Absil PA, Cartis C (2018) Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis* 39(1):1–33.

[13] Burges CJ (2010) *Dimension reduction: A guided tour* (Now Publishers Inc).

[14] Cai H, Mckenzie D, Yin W, Zhang Z (2020) Zeroth-order regularized optimization (zoro): Approximately sparse gradients and adaptive sampling. *arXiv preprint arXiv:2003.13001* .

[15] Carmo MPd (1992) *Riemannian Geometry* (Birkhäuser).

[16] Carmon Y, Duchi JC (2018) Analysis of Krylov subspace solutions of regularized non-convex quadratic problems. *Advances in Neural Information Processing Systems*, 10705–10715.

[17] Chattopadhyay A, Selvan SE, Amato U (2015) A derivative-free Riemannian Powell's method, minimizing hartley-entropy-based ICA contrast. *IEEE transactions on neural networks and learning systems* 27(9):1983–1990.

[18] Chazal F, Michel B (2017) An introduction to topological data analysis: fundamental and practical aspects for data scientists. *arXiv preprint arXiv:1710.04019* .

[19] Chen PY, Zhang H, Sharma Y, Yi J, Hsieh CJ (2017) ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 15–26 (ACM).

[20] Chen S, Deng Z, Ma S, So AMC (2020) Manifold proximal point algorithms for dual principal component pursuit and orthogonal dictionary learning. *https://arxiv.org/abs/2005.02356* .

[21] Chen S, Ma S, Man-Cho So A, Zhang T (2020) Proximal gradient method for nonsmooth optimization over the Stiefel manifold. *SIAM Journal on Optimization* 30(1):210–239.

[22] Cheng M, Le T, Chen PY, Yi J, Zhang H, Hsieh CJ (2018) Query-efficient hard-label black-box attack: An optimization-based approach. *arXiv preprint arXiv:1807.04457* .

[23] Cherian A, Sra S (2016) Riemannian dictionary learning and sparse coding for positive definite matrices. *IEEE transactions on neural networks and learning systems* 28(12):2859–2871.

[24] Conn A, Scheinberg K, Vicente L (2009) *Introduction to Derivative-Free Optimization*, volume 8 (SIAM).

[25] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: A Large-Scale Hierarchical Image Database. *CVPR09*.

[26] Diaconis P, Holmes S, Shahshahani M (2013) Sampling from a manifold. *Advances in modern statistical theory and applications: a Festschrift in honor of Morris L. Eaton*, 102–125 (Institute of Mathematical Statistics).

[27] Drieß D, Englert P, Toussaint M (2017) Constrained bayesian optimization of combined interaction force/task space controllers for manipulations. *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 902–907 (IEEE).

[28] Duchi JC, Jordan MI, Wainwright MJ, Wibisono A (2015) Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory* 61(5):2788–2806.

[29] Fong RS, Tino P (2019) Stochastic derivative-free optimization on Riemannian manifolds. *arXiv preprint arXiv:1908.06783* .

[30] Frazier PI (2018) A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811* .

[31] Ghadimi S, Lan G (2013) Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization* 23(4):2341–2368.

[32] Golovin D, Karro J, Kochanski G, Lee C, Song X (2019) Gradientless descent: High-dimensional zeroth-order optimization. *arXiv preprint arXiv:1911.06317* .

[33] Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* .

[34] Gromov M (1978) Manifolds of negative curvature. *Journal of Differential Geometry* 13(2):223–230.

[35] Harandi M, Salzmann M, Hartley R (2017) Dimensionality reduction on SPD manifolds: The emergence of geometry-aware methods. *IEEE transactions on pattern analysis and machine intelligence* 40(1):48–62.

[36] Hsu EP (2002) *Stochastic Analysis on Manifolds*, volume 38 (American Mathematical Soc.).

[37] Jamieson KG, Nowak R, Recht B (2012) Query complexity of derivative-free optimization. *Advances in Neural Information Processing Systems*, 2672–2680.

[38] Jaquier N, Rozo L (2020) High-dimensional Bayesian optimization via nested Riemannian manifolds. *Advances in Neural Information Processing Systems* 33.

[39] Jaquier N, Rozo L, Calinon S, Bürger M (2020) Bayesian optimization meets Riemannian manifolds in robot learning. *Conference on Robot Learning*, 233–246 (PMLR).

[40] Jolliffe I, NTrendafilov, Uddin M (2003) A modified principal component technique based on the LASSO. *Journal of computational and Graphical Statistics* 12(3):531–547.

[41] Kachan O (2020) Persistent homology-based projection pursuit. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 856–857.

[42] Kasai H, Sato H, Mishra B (2018) Riemannian stochastic recursive gradient algorithm. *International Conference on Machine Learning*, 2516–2524.

[43] Kovnatsky A, Glashoff K, Bronstein MM (2016) MADMM: a generic algorithm for non-smooth optimization on manifolds. *European Conference on Computer Vision*, 680–696 (Springer).

[44] Lai R, Osher S (2014) A splitting method for orthogonality constrained problems. *Journal of Scientific Computing* 58(2):431–449.

[45] Larson J, Menickelly M, Wild SM (2019) Derivative-free optimization methods. *Acta Numerica* 28:287–404.

[46] Lee JA, Verleysen M (2007) *Nonlinear Dimensionality Reduction* (Springer Science & Business Media).

[47] Leygonie J, Oudot S, Tillmann U (2019) A framework for differential calculus on persistence barcodes. *arXiv preprint arXiv:1910.00960* .

[48] Li CL, Kandasamy K, Póczos B, Schneider J (2016) High dimensional bayesian optimization via restricted projection pursuit models. *Artificial Intelligence and Statistics*, 884–892.

[49] Li X, Chen S, Deng Z, Qu Q, Zhu Z, So AMC (2019) Nonsmooth optimization over Stiefel manifold: Riemannian subgradient methods. *arXiv preprint arXiv:1911.05047* .

[50] Lin L, Lazar D, Sarpabayeva B, Dunson DB (2020) Robust optimization and inference on manifolds. *arXiv preprint arXiv:2006.06843* .

[51] Lin L, St Thomas B, Zhu H, Dunson DB (2017) Extrinsic local regression on manifold-valued data. *Journal of the American Statistical Association* 112(519):1261–1273.

[52] Lyu C, Huang K, Liang HN (2015) A unified gradient regularization family for adversarial examples. *2015 IEEE International Conference on Data Mining*, 301–309 (IEEE).

[53] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2017) Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* .

[54] Marco A, Hennig P, Bohg J, Schaal S, Trimpe S (2016) Automatic LQR tuning based on Gaussian process global optimization. *2016 IEEE international conference on robotics and automation (ICRA)*, 270–277 (IEEE).

[55] Marco A, Hennig P, Schaal S, Trimpe S (2017) On the design of LQR kernels for efficient controller learning. *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, 5193–5200 (IEEE).

[56] Matyas J (1965) Random optimization. *Automation and Remote control* 26(2):246–253.

[57] McInnes L, Healy J, Melville J (2018) UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* .

[58] Mishra B, Kasai H, Jawanpuria P, Saroop A (2019) A Riemannian gossip approach to subspace learning on Grassmann manifold. *Machine Learning* 1–21.

[59] Mockus J (1994) Application of Bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization* 4(4):347–365.

[60] Mockus J (2012) *Bayesian approach to global optimization: theory and applications*, volume 37 (Springer Science & Business Media).

[61] Mutny M, Krause A (2018) Efficient high dimensional Bayesian optimization with additivity and quadrature fourier features. *Advances in Neural Information Processing Systems*, 9005–9016.

[62] Nelder JA, Mead R (1965) A simplex method for function minimization. *The computer journal* 7(4):308–313.

[63] Nemirovski AS, Yudin DB (1983) Problem complexity and method efficiency in optimization. *Wiley & Sons* .

[64] Nesterov Y (2011) Random gradient-free minimization of convex functions. *Technical Report. Center for Operations Research and Econometrics (CORE), Catholic University of Louvain* .

[65] Nesterov Y, Polyak BT (2006) Cubic regularization of Newton method and its global performance. *Mathematical Programming* 108(1):177–205.

[66] Nesterov Y, Spokoiny V (2017) Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics* 17(2):527–566.

[67] Oh C, Gavves E, Welling M (2018) BOCK: Bayesian optimization with cylindrical kernels. *International Conference on Machine Learning*, 3868–3877.

[68] Rabadán R, Blumberg AJ (2019) *Topological Data Analysis for Genomics and Evolution: Topology in Biology* (Cambridge University Press).

[69] Rio E (2009) Moment inequalities for sums of dependent random variables under projective conditions. *Journal of Theoretical Probability* 22(1):146–163.

[70] Rolland P, Scarlett J, Bogunovic I, Cevher V (2018) High-dimensional Bayesian optimization via additive models with overlapping groups. *International Conference on Artificial Intelligence and Statistics*, 298–307.

[71] Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N (2015) Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE* 104(1):148–175.

[72] Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

[73] Spall JC (2005) *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, volume 65 (John Wiley & Sons).

[74] Stein C (1972) A bound for the error in the Normal approximation to the distribution of a sum of dependent random variables. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory* (The Regents of the University of California).

[75] Sun J, Qu Q, Wright J (2016) Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory* 63(2):853–884.

[76] Sun J, Qu Q, Wright J (2018) A geometric analysis of phase retrieval. *Foundations of Computational Mathematics* 18(5):1131–1198.

[77] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

[78] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* .

[79] Tripuraneni N, Flammarion N, Bach F, Jordan MI (2018) Averaging stochastic gradient descent on Riemannian manifolds. *Conference On Learning Theory*, 650–687.

[80] Tu CC, Ting P, Chen PY, Liu S, Zhang H, Yi J, Hsieh CJ, Cheng SM (2019) AutoZOOM: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 742–749.

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

[82] Wang B, Ma S, Xue L (2020) Riemannian stochastic proximal gradient methods for nonsmooth optimization over the Stiefel manifold. *https://arxiv.org/pdf/2005.01209.pdf* .

[83] Wang L, Fonseca R, Tian Y (2020) Learning search space partition for black-box optimization using monte carlo tree search. *Advances in Neural Information Processing Systems* 33.

[84] Wang Y, Du S, Balakrishnan S, Singh A (2018) Stochastic zeroth-order optimization in high dimensions. *International Conference on Artificial Intelligence and Statistics*, 1356–1365.

[85] Wang Z, Hutter F, Zoghi M, Matheson D, de Feitas N (2016) Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research* 55:361–387.

[86] Wang Z, Liu B, Chen S, Ma S, Xue L, Zhao H (2020) A manifold proximal linear method for sparse spectral clustering with application to single-cell RNA sequencing data analysis. *https://arxiv.org/abs/2007.09524* .

[87] Weber M, Sra S (2019) Nonconvex stochastic optimization on manifolds via Riemannian Frank-Wolfe methods. *arXiv preprint arXiv:1910.04194* .

[88] Weinberger KQ, Saul LK (2004) Unsupervised learning of image manifolds by semidenite programming. *CVPR*.

[89] Xiao X, Li Y, Wen Z, Zhang L (2018) A regularized semi-smooth Newton method with projection steps for composite convex programs. *Journal of Scientific Computing* 76(1):364–389.

[90] Yang WH, Zhang LH, Song R (2014) Optimality conditions for the nonlinear programming problems on Riemannian manifolds. *Pacific Journal of Optimization* 10(2):415–434.

[91] Yuan K, Chatzinikolaidis I, Li Z (2019) Bayesian optimization for whole-body control of high-degree-of-freedom robots through reduction of dimensionality. *IEEE Robotics and Automation Letters* 4(3):2268–2275.

[92] Zhang H, Reddi SJ, Sra S (2016) Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. *Advances in Neural Information Processing Systems*, 4592–4600.

[93] Zhang H, Sra S (2016) First-order methods for geodesically convex optimization. *Conference on Learning Theory*, 1617–1638.

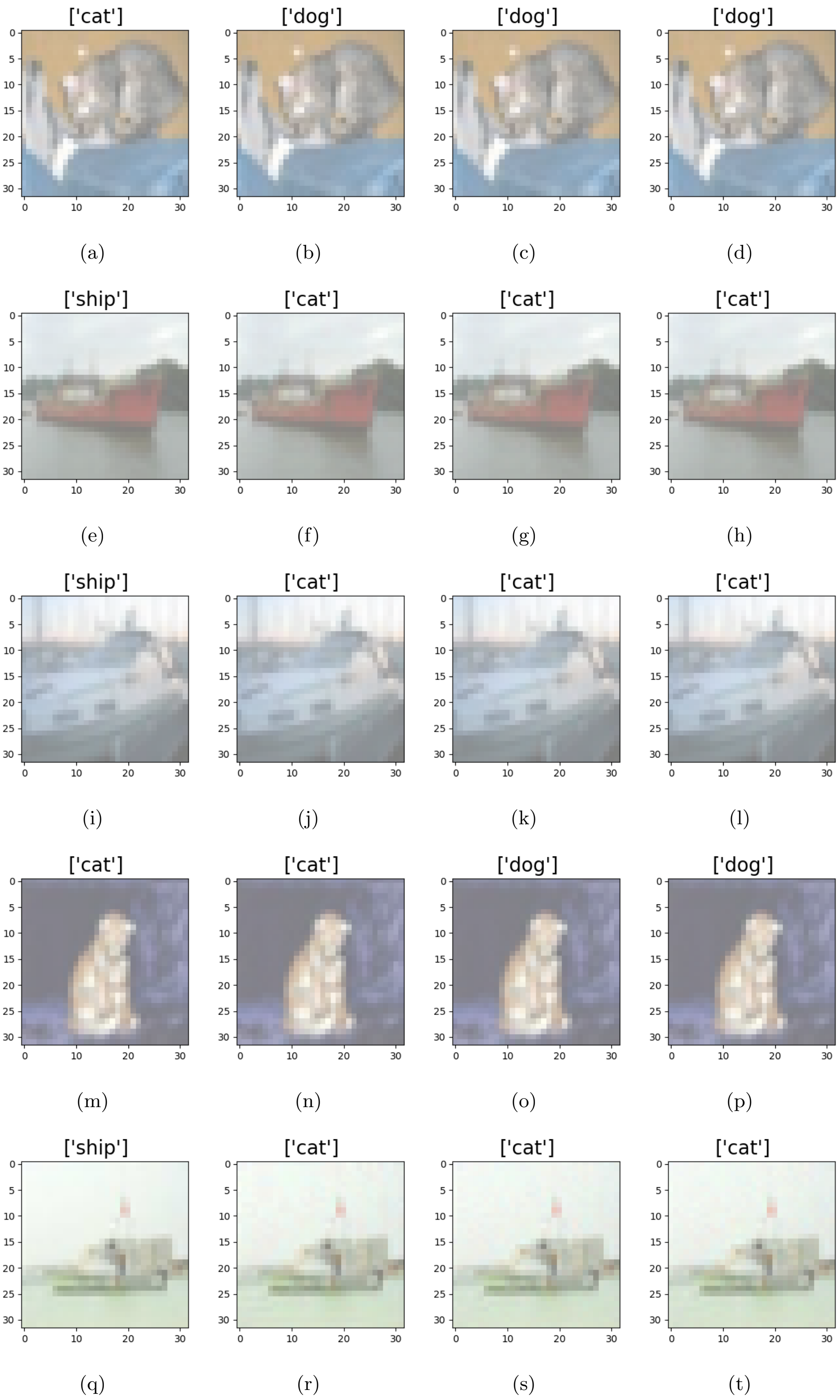
[94] Zhang J, Ma S, Zhang S (2020) Primal-dual optimization algorithms over Riemannian manifolds: an iteration complexity analysis. *Mathematical Programming Series A* 184:445–490.

[95] Zhang J, Zhang S (2018) A cubic regularized Newton's method over Riemannian manifolds. *arXiv preprint arXiv:1805.05565* .

[96] Zhou P, Yuan X, Yan S, Feng J (2019) Faster first-order methods for stochastic non-convex optimization on Riemannian manifolds. *IEEE transactions on pattern analysis and machine intelligence* .

[97] Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. *J. Comput. Graph. Stat.* 15(2):265–286.

[98] Zou H, Xue L (2018) A selective overview of sparse principal component analysis. *Proceedings of the IEEE* 106(8):1311–1320.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o) (p)

(q) (r) (s) (t)

Figure 6. The attack on CIFAR10 picture. From left to right columns: the original image; the PGD attack with a small diameter; white box Riemannian attack on the sphere with the same diameter; black box Riemannian attack on the sphere with the same diameter. Notice that for the figure in the fourth row, the PGD attack failed while the Riemannian attacks succeeded. The diameter is set to be 0.01 times the norm of the original images.