Riemannian Accelerated Zeroth-order Algorithm: Improved Robustness and Lower Query Complexity

Chang He [1] **Zhaoye Pan** [1] **Xiao Wang** [1 2] **Bo Jiang** [1 2 3]

Abstract

Optimization problems with access to only zeroth-order information of the objective function on Riemannian manifolds arise in various applications, spanning from statistical learning to robot learning. While various zeroth-order algorithms have been proposed in Euclidean space, they are not inherently designed to handle the challenging constraints imposed by Riemannian manifolds. The proper adaptation of zeroth-order techniques to Riemannian manifolds remained unknown until the pioneering work of (Li et al., 2023a). However, zeroth-order algorithms are widely observed to converge slowly and be unstable in practice. To alleviate these issues, we propose a Riemannian accelerated zeroth-order algorithm with improved robustness. Regarding efficiency, our accelerated algorithm has the function query complexity of $\mathcal{O}(\epsilon^{-7/4}d)$ for finding an ϵ-approximate first-order stationary point. By introducing a small perturbation, it exhibits a function query complexity of $\tilde{\mathcal{O}}(\epsilon^{-7/4}d)$ for seeking a second-order stationary point with a high probability, matching state-of-the-art result in Euclidean space. Moreover, we further establish the almost sure convergence in the asymptotic sense through the Stable Manifold Theorem. Regarding robustness, our algorithm requires larger smoothing parameters in the order of $\tilde{\mathcal{O}}(\epsilon^{7/8}d^{-1/2})$, improving the existing result by a factor of $\tilde{\mathcal{O}}(\epsilon^{3/4})$.

[1]School of Information Management and Engineering, Shanghai University of Finance and Economics [2]Key Laboratory of Interdisciplinary Research of Computation and Economics, Shanghai University of Finance and Economics, Ministry of Education [3]Dishui Lake Advanced Finance Institute, Shanghai University of Finance and Economics. Correspondence to: Chang He <ischanghe@gmail.com>, Bo Jiang <isyebojiang@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Many machine learning problems frequently encounter situations where computing function gradients is costly or even infeasible. For instance, the tasks such as optimal linear combination prediction (Das et al., 2022) and Bayesian optimization in robot learning (Jaquier et al., 2018; 2020) involve objective functions, lacking analytical forms, only observable through point-wise evaluations. Furthermore, the design space of interest is also complicated, involving constraints such as the unit sphere, probability simplex, and positive definite matrices. The limited function information and inherent constraints render these problems challenging to solve. One potent strategy for dealing with these constraints is re-expressing them through the lens of *Riemannian manifolds* (Absil et al., 2009; Boumal, 2023). Mathematically, we can formulate the problem in consideration as follows:

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

where $\mathcal{M}$ represents the Riemannian manifold, and $f(\cdot)$ is a *nonconvex* objective function with only zeroth-order information (i.e. function value) available. For ease of discussion, we assume $f(\cdot)$ is lower bounded, i.e. $f(x) \geq f_{\text{low}}$ for all $x \in \mathcal{M}$. Recently, a pioneering work by Li et al. (2023a) introduced several Riemannian *zeroth-order* algorithms to tackle problem (1), relying solely on the query of function values. It is well known that the function query complexity is a key to measure the efficiency of the zeroth-order algorithms, whereas these algorithms only exhibit inferior complexity to the one in Euclidean space. This raises a natural question: *Is it possible to develop a Riemannian accelerated zeroth-order algorithm with lower function query complexity?*

The development of accelerated algorithms is a prominent and active topic within both machine learning and optimization communities. It traces back to the seminal breakthrough by Nesterov (1983), which paved the way for subsequent advancements in acceleration techniques. Since then, numerous fruitful results have emerged in various scenarios, such as accelerated first-order algorithms (Beck & Teboulle, 2009; Lin et al., 2015; Carmon et al., 2017; 2018; Jin et al., 2018; Li & Lin, 2022) and accelerated second-order algorithms (Nesterov, 2008; Bubeck et al., 2019; Jiang et al.,

Table 1. Comparison of zeroth-order algorithms in terms of the ability to handle Riemannian manifolds, value of smoothing parameter, and function query complexity for nonconvex objective function. The symbol † is used to indicate that this algorithm converges to ϵ-approximate first-order stationary points; otherwise, it converges to ϵ-approximate second-order stationary points.

Algorithms	Riemanian Manifolds	Smoothing parameter μ	Function query complexity
PAGD (Vlatakis-Gkaragkounis et al., 2019)	✗	$\mathcal{O}\left(\frac{\epsilon^{3/2}}{\sqrt{d}}\right)$	$\tilde{\mathcal{O}}\left(\frac{d}{\epsilon^2}\right)$
ZO-GD (Bai et al., 2020)	✗	$\mathcal{O}\left(\frac{\epsilon^3}{d^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{d^2}{\epsilon^8}\right)$
ZO-GD-NCF (Zhang et al., 2022)	✗	$\mathcal{O}\left(\frac{\epsilon^{1/2}}{d^{1/4}}\right)$	$\tilde{\mathcal{O}}\left(\frac{d}{\epsilon^2}\right)$
ZO-PAGD (Zhang & Gu, 2022)	✗	$\tilde{\mathcal{O}}\left(\frac{\epsilon^{13/8}}{\sqrt{d}}\right)$	$\tilde{\mathcal{O}}\left(\frac{d}{\epsilon^{7/4}}\right)$
ZOPGD (Ren et al., 2023)	✗	$\tilde{\mathcal{O}}\left(\frac{\epsilon^{1/2}}{d}\right)$	$\tilde{\mathcal{O}}\left(\frac{d}{\epsilon^2}\right)$
ZO-RGD (Li et al., 2023a)	✓	$\mathcal{O}\left(\frac{\epsilon}{d^{3/2}}\right)$	$\mathcal{O}\left(\frac{d}{\epsilon^2}\right)^{\dagger}$
RAZGD with Option I (ours)	✓	$\mathcal{O}\left(\frac{\epsilon^{5/8}}{d^{1/4}}\right)$	$\mathcal{O}\left(\frac{d}{\epsilon^{7/4}}\right)^{\dagger}$
Perturbed RAZGD with Option I (ours)	✓	$\tilde{\mathcal{O}}\left(\frac{\epsilon^{7/8}}{\sqrt{d}}\right)$	$\tilde{\mathcal{O}}\left(\frac{d}{\epsilon^{7/4}}\right)$

2021). Notably, Zhang & Gu (2022) demonstrated that zeroth-order algorithms can also benefit from acceleration and exhibit an improved complexity. Regarding the optimization problem over Riemannian manifolds, there has also been a growing interest in developing accelerated algorithms (Liu et al., 2017; Zhang & Sra, 2018; Criscitiello & Boumal, 2022), to name a few. Due to the space limitation, a detailed discussion is deferred to Appendix A. Despite significant efforts in designing accelerated algorithms, none of them is applicable to problem (1).

To design an algorithm in a gradient-free manner, constructing zeroth-order estimators through function value evaluations becomes necessary. The accuracy of this approximation is tied to the *smoothing parameter* (see Definition 3.1, for example). Although the smaller value of the parameter improves the precision, it may also introduce instability in practical applications (Lian et al., 2016; Liu et al., 2018; 2020). Regrettably, integrating acceleration techniques into zeroth-order algorithms (Zhang & Gu, 2022) requires smaller smoothing parameters compared to the standard ones (Vlatakis-Gkaragkounis et al., 2019; Zhang et al., 2022). In response to these challenges, we introduce a novel *Riemannian accelerated zeroth-order algorithm*. Surprisingly, while maintaining the same function query complexity, our algorithm allows the use of a larger smoothing parameter, compared to the Euclidean counterpart (Zhang & Gu, 2022). This, in turn, ensures the robust and stable performance of our accelerated zeroth-order algorithm.

Contributions. In this paper, we delve into a comprehensive study of Riemannian zeroth-order optimization. Our main contributions are given as follows:

- By leveraging the basis of the tangent space, we extend the classical finite-difference gradient approximation to Riemannian manifolds (Definition 3.1). Based on this estimator, we develop a Riemannian accelerated zeroth-order gradient descent (RAZGD) in Algorithm 1, which alternates between the Riemannian zeroth-order gradient descent step (Subroutine 1) and the tangent space step (Subroutine 2 and 3).

- Under some mild assumptions and by setting the initial point as zero in the tangent space step (Subroutine 2), we prove that the RAZGD with Option I has the function query complexity of $\mathcal{O}(\epsilon^{-7/4}d)$ for finding a Riemannian ϵ-approximate first-order stationary point, which improves the existing result by a factor of $O(\epsilon^{-1/4})$ in (Li et al., 2023a). For a fair comparison, we present selected zeroth-order algorithms in Table 1.

- By introducing a small perturbation to the initial point in the tangent space step (Subroutine 2), the perturbed RAZGD with Option I seeks a second-order stationary point with a high probability under $\tilde{\mathcal{O}}(\epsilon^{-7/4}d)$ query complexity guarantee, matching state-of-the-art complexity in Euclidean zeroth-order optimization (Zhang & Gu, 2022). To get an almost sure convergence result, we further prove that the RAZGD with Option II converges to strict Riemannian second-order stationary points gradually.

- Beyond the function query complexity, the perturbed RAZGD with Option I showcases resilience in choosing the smoothing parameter—an essential factor ensuring the robustness of zeroth-order algorithms. With the same function query complexity guarantee, we estab-

lish that the RAZGD only requires the smoothing parameter $\mu = \tilde{\mathcal{O}}(\epsilon^{7/8}d^{-1/2})$ for seeking ϵ-approximate second-order stationary points, sharpening the existing best result $\tilde{\mathcal{O}}(\epsilon^{13/8}d^{-1/2})$ in Zhang & Gu (2022).

2. Preliminaries: Optimization over manifolds

In this section, we present the basic setup and mathematical tools for optimization over manifolds. For more details, we refer readers to see (Absil et al., 2009; Boumal, 2023). Throughout this paper, we use the convention $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ to denote lower and upper bounds with a universal constant, respectively. $\tilde{\mathcal{O}}(\cdot)$ ignores the polylogarithmic terms. We use d to denote both the dimension of the Riemannian manifold $\mathcal{M}$ (i.e., $\dim(\mathcal{M}) = d$) and the dimension of the Euclidean space $\mathbb{R}^d$.

A d-dimensional manifold $\mathcal{M}$ is a topological space where each point has a neighborhood homomorphic to d-dimensional Euclidean space, as illustrated in Figure 1. A Riemannian manifold $\mathcal{M}$ is a real, smooth manifold equipped with a Riemannian metric. Each $x \in \mathcal{M}$ is associated with a d-dimensional real vector space $\mathrm{T}_x\,\mathcal{M}$, referred to as the tangent space at x. The Riemannian metric defines an inner product $\langle \cdot, \cdot \rangle_x$ on the tangent space $\mathrm{T}_x\,\mathcal{M}$. The inner metric induces a corresponding norm $\|\cdot\|_x$. We denote these by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ when there is no confusion for x from the context. A vector in the tangent space is known as a tangent vector. The set of pairs (x, s_x) for $x \in \mathcal{M}, s_x \in \mathrm{T}_x\,\mathcal{M}$ is called the tangent bundle $\mathrm{T}\,\mathcal{M}$. On the tangent space, we define $\mathbb{B}_{x,r}(s) = \{z \in \mathrm{T}_x\,\mathcal{M} : \|z - s\|_x \le r\}$, representing the closed ball of radius r centered at $s \in \mathrm{T}_x\,\mathcal{M}$. Then we use $\mathrm{Uni}(\mathbb{B}_{x,r}(s))$ to define the uniform distribution over the ball $\mathbb{B}_{x,r}(s)$.

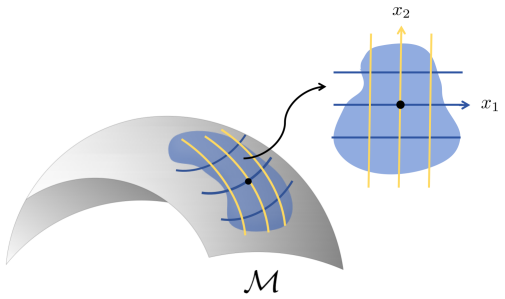


Figure 1. A 2-dimensional manifold

Given a smooth function $f(\cdot)$, the Riemannian gradient $\mathrm{grad}\, f(x)$ of f at $x \in \mathcal{M}$ is the unique vector in $\mathrm{T}_x\,\mathcal{M}$ that satisfies $\mathrm{D}\, f(x)[s] = \langle \mathrm{grad}\, f(x), s \rangle_x$ for all $s \in \mathrm{T}_x\,\mathcal{M}$, where $\mathrm{D}\, f(x)[s]$ is the directional derivative of f at x along s. The Riemannian metric gives rise to a well-defined notion of the derivative of vector fields, known as the Levi–Civita connection ∇. The Riemannian Hessian of f is the derivative of the gradient vector field: $\mathrm{Hess}\, f(x)[u] = \nabla_u \mathrm{grad}\, f(x)$, which is a symmetric linear operator on $\mathrm{T}_x\,\mathcal{M}$. For the smooth curve $\gamma : [0, 1] \to \mathcal{M}$, the velocity of the curve is defined as $\frac{d\gamma}{dt} = \gamma'(t)$. The intrinsic acceleration γ'' of γ is the covariant derivative of the velocity of γ': $\gamma'' = \frac{\mathrm{D}}{dt}\gamma'$ induced by the Levi–Civita connection.

We proceed to introduce the ϵ-approximate stationary point on Riemannian manifolds.

Definition 2.1. *For any $\epsilon > 0$, a point $x \in \mathcal{M}$ is an ϵ-approximate Riemannian first-order stationary point (RFOSP) of the smooth function $f(\cdot)$ if it satisfies $\|\mathrm{grad}\, f(x)\| \le \mathcal{O}(\epsilon)$. Furthermore, if it additionally satisfies $\lambda_{\min}(\mathrm{Hess}\, f(x)) \ge \Omega(-\sqrt{\epsilon})$, then x is an ϵ-approximate Riemannian second-order stationary point (RSOSP), where $\lambda_{\min}(\cdot)$ denotes the the smallest eigenvalue of the symmetric operator.*

We also present the definition of strict Riemannian saddle points and second-order stationary points:

Definition 2.2. *A point $x \in \mathcal{M}$ is a strict Riemannian saddle point of the smooth function $f(\cdot)$ if it satisfies $\mathrm{grad}\, f(x) = 0$ and $\lambda_{\min}(\mathrm{Hess}\, f(x)) < 0$. Otherwise, it is a strict Riemannian second-order stationary point when $\mathrm{grad}\, f(x) = 0$ and $\lambda_{\min}(\mathrm{Hess}\, f(x)) \ge 0$.*

To optimize over Riemannian manifolds, a key concept is the retraction (Figure 2)—a mapping enabling movement along the manifold from a point x in the direction of a tangent vector $s \in \mathrm{T}_x\,\mathcal{M}$. This is formalized as follows:

Definition 2.3. *A retraction mapping $\mathrm{Retr}_x : \mathrm{T}_x\,\mathcal{M} \to \mathcal{M}$ is a smooth mapping satisfies $\mathrm{Retr}_x(0) = x$, where 0 is the zero vector in $\mathrm{T}_x\,\mathcal{M}$. Moreover, for $x \in \mathcal{M}$ and $s \in \mathrm{T}_x\,\mathcal{M}$, let*

$$T_{x,s} = \mathrm{D}\,\mathrm{Retr}_x(s) : \mathrm{T}_x\,\mathcal{M} \to \mathrm{T}_{\mathrm{Retr}_x(s)}\,\mathcal{M}$$

denote the differential of Retr_x at s (a linear operator). The differential of Retr_x at 0, i.e. $T_{x,0}$, is the identity map.

For instance, we employ $\mathrm{Retr}_x(s) = x + s$ when $\mathcal{M} = \mathbb{R}^d$. On the unit sphere $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$, the retraction mapping is typically defined as $\mathrm{Retr}_x(s) = (x + s)/\|x + s\|_2$.

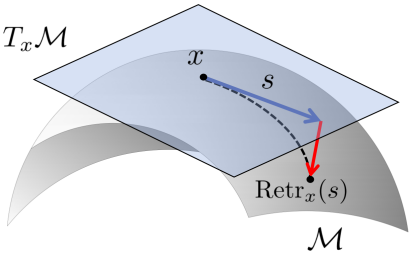


Figure 2. Retraction

Following a similar manner in (Criscitiello & Boumal, 2019; 2022), in this paper, our analysis is based on the pullback function defined as follows.

Definition 2.4. *For any $x \in \mathcal{M}$, the pullback function $\hat{f}_x(\cdot)$ is a composite function of f and the retraction mapping, that is*

$$\hat{f}_x = f \circ \mathrm{Retr}_x \colon \mathrm{T}_x \mathcal{M} \to \mathbb{R}.$$

Specifically, as the differential of Retr_x at 0 is the identity map, it implies that

$$\hat{f}_x(0) = f(x).$$

Note that the pullback function $\hat{f}_x(\cdot)$ is a real function defined on the tangent space $\mathrm{T}_x \mathcal{M}$, which is locally homomorphic to Euclidean space. With a slight abuse of notation, we can define the usual gradient and Hessian of $\hat{f}_x(\cdot)$ as $\nabla \hat{f}_x(\cdot)$ and $\nabla^2 \hat{f}_x(\cdot)$ (mind the overloaded notation of Levi–Civita connection ∇), respectively.

3. Riemannian Accelerated Zeroth-order Gradient Descent Algorithm

3.1. Review of Riemannian gradient descent algorithm

We begin with an ideal situation in which the gradient information is feasible, and consequently the simplest Riemannian gradient descent (Boumal et al., 2019)

$$x_{t+1} = \mathrm{Retr}_{x_t}(-\eta_t \operatorname{grad} f(x_t)),\ t = 0, 1, \ldots.$$

is applicable to problem (1). For the nonconvex objective function, the basic idea behind the convergence analysis of Riemannian gradient descent revolves around a two-case discussion based on the magnitude of the gradient at the current iterate. If the norm of Riemannian gradient satisfies $\|\operatorname{grad} f(x_t)\| \geq \Omega(\epsilon)$, Riemannian gradient descent is shown to result in a decrease in the function value of $O(\epsilon^2)$. On the other hand, if the gradient norm is below this threshold, the current point is already an ϵ-approximate Riemannian first-order stationary point. Thus, the Riemannian gradient descent algorithm requires at most $\mathcal{O}(\epsilon^{-2})$ steps to find a first-order stationary point.

3.2. The algorithm design

Inspired by the Riemannian gradient descent algorithm, we employ an unconventional strategy, aiming for a more aggressive function value decrease at each update—a crucial element in designing a faster Riemannian algorithm. Given the inaccessibility of the Riemannian gradient, we carefully examine the value of the zeroth-order estimator. When the Riemannian zeroth-order estimator at the current iterate x_t exceeds $\Omega(\sqrt{\epsilon})$—deviating from the standard value of $\Omega(\epsilon)$—we choose to proceed with the Riemannian zeroth-order gradient descent step (Subroutine 1), resulting in the function value decrease of $O(\epsilon)$.

For the iterate x_t with a small Riemannian zeroth-order estimator, we choose the tangent space step (Subroutine 2), which involves the accelerated zeroth-order gradient descent update in the tangent space $\mathrm{T}_{x_t} \mathcal{M}$, as depicted in Figure 3. In the tangent space step, we set a similar termination criterion as (Li & Lin, 2022), ensuring that the tangent space step either results in the function value decrease of $O(\epsilon^{3/2})$ or returns a stationary point. Therefore, after a single update from x_t to x_{t+1}, the function value takes a decrease at least $O(\epsilon^{3/2})$, which is larger compared to the standard case. Combining all these components, we introduce the Riemannian accelerated zeroth-order gradient descent in Algorithm 1. By always selecting Option I, it achieves lower query complexity in the non-asymptotic analysis. Moreover, with a slightly modified tangent space step (Subroutine 3), the RAZGD with Option II almost surely avoids strict saddle points asymptotically.

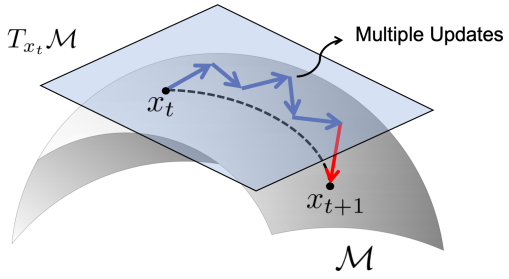


Figure 3. Tangent space step

Algorithm 1 **R**iemannian **A**ccelerated **Z**eroth-order **G**radient **D**escent Algorithm (**RAZGD**)

input: parameters η, θ, B, K and r
initialize: $x_0 \in \mathcal{M}$, $t = 0$
for $t = 0, 1, \cdots, \infty$ **do**
  Compute estimator $g_{x_t}(0; \mu)$
  if $\|g_{x_t}(0; \mu)\| \geq lB$ **then**
    $x_{t+1} = \textbf{RZGDS}(x_t, \eta, g_{x_t}(0; \mu))$
  else
    Option I: $x_{t+1} = \textbf{TSS}(x_t, \eta, \theta, B, K, r)$
    Option II: $x_{t+1} = \textbf{TSSA}(x_t, \eta, \theta, B)$
  end if
end for

As demonstrated in both tangent space steps, multiple zeroth-order updates are performed in the tangent space. To ensure the well-definiteness of the tangent space step,

Subroutine 1 **R**iemannian **Z**eroth-order **G**radient **D**escent **S**tep (**RZGDS**)

input: x, η, and $g_x(0;\mu)$
if $\eta\|g_x(0;\mu)\| \le b$ **then**
  Return $\mathrm{Retr}_x(-\eta g_x(0;\mu))$
else
  Compute $\alpha \in (0,1)$ such that $\alpha\eta\|g_x(0;\mu)\| = b$
  Return $\mathrm{Retr}_x(-\alpha\eta g_x(0;\mu))$
end if

Subroutine 2 **T**angent **S**pace **S**tep (**TSS**)

input: x, η, θ, B, K and r
initialize: $s_x^{-1} = s_x^0 = \xi \sim \mathrm{Uni}(\mathbb{B}_{x,r}(0))$, $k = 0$
while $k < K$ **do**
  $y_x^k = s_x^k + (1-\theta)(s_x^k - s_x^{k-1})$
  Compute estimator $g_x(y_x^k;\mu)$
  $s_x^{k+1} = y_x^k - \eta g_x(y_x^k;\mu)$
  $k = k+1$
  if $k\sum_{j=0}^{k-1}\|s_x^{j+1} - s_x^j\|^2 > B^2$ **then**
    Return $\mathrm{Retr}_x(s_x^k)$ and break
  end if
end while
$K_0 = \mathrm{argmin}_{\lfloor \frac{K}{2} \rfloor \le k \le K-1}\|s_x^{k+1} - s_x^k\|$
$y_x^* = \frac{1}{K_0+1}\sum_{k=0}^{K_0} y_x^k$
Return $\mathrm{Retr}_x(y_x^*)$

we define the zeroth-order estimator for every pair (x, s_x) in the tangent bundle $\mathrm{T}\mathcal{M}$, where s_x is the point in the tangent space $\mathrm{T}_x\mathcal{M}$. In the algorithm and its subroutines, the notation $g_x(s_x;\mu)$ represents the zeroth-order estimator for the gradient of the pullback function $\nabla\hat{f}_x(s_x)$ at the pair $(x, s_x) \in \mathrm{T}\mathcal{M}$, incorporating a smoothing parameter μ. The formal definition is shown as follows, which generalizes the classic finite difference gradient approximation in Euclidean space (Scheinberg, 2022).

Definition 3.1. *Given a smoothing parameter $\mu > 0$ and a point $x \in \mathcal{M}$, the Riemannian coordinate-wise zeroth-order estimator at the point $s_x \in \mathrm{T}_x\mathcal{M}$ is defined as*

$$g_x(s_x;\mu) = \sum_{i=1}^{d} \frac{\hat{f}_x(s_x + \mu e_i) - \hat{f}_x(s_x - \mu e_i)}{2\mu} e_i,$$

where $\{e_1, e_2, \ldots, e_d\}$ is the basis of the tangent space $\mathrm{T}_x\mathcal{M}$.

For compactness, the approximation error of the Riemannian coordinate zeroth-order estimator is deferred to Appendix D.

Subroutine 3 **T**angent **S**pace **S**tep **A**symptotic (**TSSA**)

input: x, η, θ and B
initialize: $s_x^{-1} = s_x^0 = 0$, $k = 0$, and constant $\beta < 1$, or $\beta_k = 1 - \frac{1}{k+2}$
while $k\sum_{j=0}^{k-1}\|s_x^{j+1} - s_x^j\|^2 \le B^2$ **do**
  $y_x^k = s_x^k + (1-\theta)(s_x^k - s_x^{k-1})$
  Compute estimator $g_x(y_x^k;\mu)$
  $s_x^{k+1} = y_x^k - \eta g_x(y_x^k;\mu)$
  $k = k+1$
  $\mu = \beta\mu$ (or $\mu = \beta_k\mu$)
end while
$K_0 = \mathrm{argmin}_{\lfloor \frac{K}{2} \rfloor \le k \le K-1}\|s_x^{k+1} - s_x^k\|$
$y_x^* = \frac{1}{K_0+1}\sum_{k=0}^{K_0} y_x^k$
Return $\mathrm{Retr}_x(y_x^*)$

4. Convergence Analysis

4.1. Mild assumptions

We start with the following assumptions on the Riemannian manifold and objective function, which will be used throughout our analysis. Due to the space limit, we have left a discussion of these assumptions in Appendix B. Firstly, generalizing from the Euclidean case, we assume the Lipschitz continuity of the gradient and Hessian of the pullback function $\hat{f}_x(\cdot)$. However, it is worth noting that Lipschitz continuity holds only locally due to the nonlinear structure of Riemannian manifolds (Criscitiello & Boumal, 2022).

Assumption 4.1. *There exists constants $b > 0$ and $l > 0$ and $\rho > 0$ such that for all $x \in \mathcal{M}$ and $s, t \in \mathbb{B}_{x,b}(0)$, the pullback function $\hat{f}_x(\cdot)$ satisfies*

$$\|\nabla\hat{f}_x(s) - \nabla\hat{f}_x(t)\| \le l\|s - t\|.$$

Assumption 4.2. *There exists constants $b > 0$ and $\rho > 0$ such that for all $x \in \mathcal{M}$ and and $s, t \in \mathbb{B}_{x,b}(0)$, the pullback function $\hat{f}_x(\cdot)$ satisfies*

$$\|\nabla^2\hat{f}_x(s) - \nabla^2\hat{f}_x(t)\| \le \rho\|s - t\|.$$

The next assumption requires that the retraction is well-behaved.

Assumption 4.3. *For any $x \in \mathcal{M}$ and $s \in \mathrm{T}_x\mathcal{M}$ satisfying $\|s\| \le b$, the singular value of the operator $T_{x,s}$ in Definition 2.3 is bounded, that is, there exists $\sigma_{\max}, \sigma_{\min} > 0$ such that*

$$\sigma_{\min} \le \sigma_{\min}(T_{x,s}) \le \sigma_{\max}(T_{x,s}) \le \sigma_{\max}.$$

Furthermore, there exists $\tau \ge 0$ such that the initial acceleration of the the curve $\gamma_{x,s}(t) = \mathrm{Retr}_x(ts)$ with $\|s\| = 1$ is bounded by τ : $\left\|\gamma''_{x,s}(0)\right\| \le \tau$.

4.2. Non-asymptotic convergence

In this subsection, we aim to find ϵ-approximate stationary points, which is achieved by always choosing Option I in RAZGD. Our theoretical findings yield different convergence results based on the choice of the initial point ξ in the tangent space step 2. When setting ξ to zero, the RAZGD converges to a first-order Riemannian stationary point. Introducing a small perturbation to ξ, the perturbed RAZGD seeks a second-order Riemannian stationary point with high probability. The results are presented below, and the associated proofs are deferred to Appendix E.

Theorem 4.1. *Suppose that Assumption 4.1, 4.2 and 4.3 hold. Set the parameters in Algorithm 1 as follows*

$$\eta = \frac{1}{4l},\ B = \frac{1}{8}\sqrt{\frac{\epsilon}{\rho}},\ \theta = \frac{\rho^{\frac{7}{4}}\epsilon^{\frac{1}{4}}}{l},\ r = 0,\ K = \frac{\rho^{\frac{5}{4}}}{4\epsilon^{\frac{1}{4}}}.$$

For any $x_0 \in \mathcal{M}$ and sufficiently small $\epsilon > 0$, choose $\mu = \mathcal{O}\left(\frac{\epsilon^{1/4}}{d^{1/4}}\right)$ in Lines 3 of Algorithm 1, and $\mu = \mathcal{O}\left(\frac{\epsilon^{5/8}}{d^{1/4}}\right)$ in Line 5 of Subroutine 2. Then Algorithm 1 with Option I outputs an ϵ-approximate first-order stationary point. The total number of function value evaluations is no more than

$$\mathcal{O}\left(\frac{(f(x_0) - f_{\text{low}})d}{\epsilon^{\frac{7}{4}}}\right).$$

Theorem 4.2. *Suppose that Assumption 4.1, 4.2 and 4.3 hold. Set the parameters in Algorithm 1 as follows*

$$\eta = \frac{1}{4l}, \quad \theta = \frac{\rho^{\frac{7}{4}}\epsilon^{\frac{1}{4}}}{l}, \quad \chi = \mathcal{O}\left(\log\frac{d}{\delta\epsilon}\right) \geq 1,$$
$$K = \frac{\chi\rho^{\frac{5}{4}}}{4\epsilon^{\frac{1}{4}}}, \quad B = \frac{1}{8\chi^2}\sqrt{\frac{\epsilon}{\rho}}, \quad r = \frac{\theta B}{6K}.$$

For any $x_0 \in \mathcal{M}$ and sufficiently small $\epsilon > 0$, choose $\mu = \mathcal{O}\left(\frac{\epsilon^{1/4}}{d^{1/4}\chi}\right) = \tilde{\mathcal{O}}\left(\frac{\epsilon^{1/4}}{d^{1/4}}\right)$ in Lines 3 of Algorithm 1, and $\mu = \min\left\{\mathcal{O}\left(\frac{\epsilon^{5/8}}{d^{1/4}\chi^2}\right), \mathcal{O}\left(\frac{\epsilon^{7/8}}{\chi^3\sqrt{d}}\right)\right\} = \tilde{\mathcal{O}}\left(\frac{\epsilon^{7/8}}{\sqrt{d}}\right)$ in Line 5 of Subroutine 2. Then perturbed Algorithm 1 with Option I outputs an ϵ-approximate second-order stationary point with a probability of at least $1-\delta$. The total number of function value evaluations is no more than

$$O\left(\frac{(f(x_0) - f_{\text{low}})d}{\epsilon^{\frac{7}{4}}}\log^6\left(\frac{d}{\delta\epsilon}\right)\right).$$

In dealing with the unavailability of the first-order information, we carefully choose the smoothing parameter μ in the construction of its zeroth-order estimators. On the one hand, a small value of μ reduces the approximation error, yielding a sufficient decrease in the function value. On the other hand, an excessively small μ can cause practical instability. Consequently, a trade-off arises in selecting the smoothing parameter, requiring a careful balance between maintaining the decrease in the function value and ensuring practical robustness. In our proofs, we frequently use Young's inequality to guarantee this balance, leading to a better choice of the smoothing parameter compared to the corresponding choice in the Euclidean counterpart (Zhang & Gu, 2022).

Remark 4.1. *For the special case of $\mathcal{M} = \mathbb{R}^d$ and $\text{Retr}_x(s) = x + s$, Theorem 4.2 reveals that the function query complexity of perturbed RAZGD with Option I matches the state-of-the-art result in Euclidean space (Zhang & Gu, 2022). However, to attain the lower function query complexity, the accelerated zeroth-order algorithms in Zhang & Gu (2022) demand the smoothing parameter $\mu = \tilde{\mathcal{O}}(\epsilon^{13/8}d^{-1/2})$. In contrast, our perturbed RAZGD relaxes this requirement to $\mu = \tilde{\mathcal{O}}(\epsilon^{7/8}d^{-1/2})$, providing a more robust selection guarantee.*

Why smoothing parameter μ is important? Compared to first-order algorithms, the notable distinction of zeroth-order algorithms lies in the necessity to construct zeroth-order estimates through function value evaluations. Among these zeroth-order estimates, the smoothing parameter plays a crucial role as an indicator. The efficiency of zeroth-order algorithms is measured by the total number of function value evaluations, while the value of the smoothing parameter determines its robustness. Generally, a smaller smoothing parameter improves the approximation quality of the zeroth-order estimator, see Lemma D.1, for example. Nevertheless, in practical systems, an excessively small μ might induce the dominance of system noise in function differences, causing the failure to represent the function differential (Lian et al., 2016; Liu et al., 2018; 2020; Nguyen & Balasubramanian, 2023). Therefore, maintaining a relatively large smoothing parameter is paramount to the robustness of zeroth-order algorithms. In the realm of randomized zeroth-order estimators, Ren et al. (2023) improved the value of the smoothing parameter from $\mathcal{O}(\epsilon^3 d^{-2})$ in (Bai et al., 2020) to a more efficient choice $\mathcal{O}(\epsilon^{1/2}d^{-1})$ for finding second-order stationary points in Euclidean space. When dealing with Riemannian manifolds, Wang et al. (2021) demonstrated that choosing random vectors uniformly from the unit sphere enables a less restrictive smoothing parameter. The selection of the smoothing parameter can be improved from $\mathcal{O}(\epsilon(d+3)^{-3/2})$ to $\mathcal{O}(\epsilon d^{-3/2})$ for seeking Riemannian first-order stationary points.

4.3. Asymptotic convergence

Now we turn to investigate the asymptotic convergence of Algorithm 1, which can be proven to avoid Riemannian strict saddle points almost surely by employing Option II, i.e. tangent space step asymptotic. In contrast to Subroutine 2, where the smoothing parameter maintains the same value during the update, in Subroutine 3, we initialize the smoothing parameter μ with an appropriate constant value,

and then make μ gradually decay by multiplying it with the contraction parameter β. Previous results have established non-asymptotic convergence to ϵ-approximate second-order stationary points, indicating that, with high probability, Algorithm 1 with Option I will output a point satisfying the specified threshold. However, there is a gap between high probability and probability 1 for convergencing to second-order stationary points. We close this gap by providing the following result asserting that the set of initial points that can be iterated to Riemannian saddle points has measure (the volume induced by Riemannian metric) zero.

Theorem 4.3. *Suppose that Assumption 4.1, 4.2 and 4.3 hold. For any $x_0 \in \mathcal{M}$ and sufficiently small $\epsilon > 0$, set*

$$\frac{\theta}{(2-\theta)\lambda_*} < \eta \leq \frac{1}{4l}, \quad \theta = \frac{\rho^{\frac{7}{4}}\epsilon^{\frac{1}{4}}}{l} \leq \min\left\{\frac{-2\lambda_*}{4l-\lambda_*}, 1\right\}$$

where λ_* is the negative eigenvalue of Hessian at saddle points with the greatest magnitude. Choose the smoothing parameter μ with a reasonable constant magnitude in both Line 4 of Algorithm 1 and during the initialization of Subroutine 3. Choose constant $\beta < 1$ or a sequence $\beta_k = \left(1 - \frac{1}{k+2}\right)$, and the rest parameters follow the choices of Theorem 4.1. Then Algorithm 1 with Option II avoids strict Riemannian saddle points almost surely. Furthermore, this implies that the Algorithm 1 with Option II asymptotically converges to a strict Riemannian second-order stationary point.

Remark 4.2. *In the tangent space step **TSSA**, the smoothing parameter μ decreases in exponential rate, which makes the zeroth-order method almost identical to a first-order method. Despite this setting being convenient for theoretical analysis, the rapid decaying of smoothing parameters makes the algorithm less attractive from the zeroth-order optimization perspective. To reduce the rate of decaying of the smoothing parameter, we propose an alternating approach with a time-varying contracting factor, i.e., multiplying by a factor of $\left(1 - \frac{1}{k+2}\right)$. It is obvious that the rate of decaying of the smoothing parameter in **TSSA** stage is $\mu_{k+1} = \frac{1}{3(k+2)}\mu$, which is much slower than the exponential decaying given by $\mu_{k+1} = \beta^k\mu$. Fortunately, the asymptotic avoidance of saddle points holds with $\beta_k = 1 - \frac{1}{k+2}$.*

5. Numerical Experiments

In this section, we conduct experiments to demonstrate the robustness and efficiency of RAZGD. Specifically, we implement the tangent space step (Subroutine 2) due to its non-asymptotic complexity guarantee. All experiments are performed on a computer with a 24-core Intel Core i9-13900HX processor.

5.1. Improved robustness

To verify the robust performance, we consider the following quartic function (Lucchi et al., 2021; Zhang & Gu, 2022) on Euclidean space $\mathbb{R}^d$:

$$f(x_1, x_2, \ldots, x_d, y) = \frac{1}{4}\sum_{i=1}^{d} x_i^4 - y\sum_{i=1}^{d} x_i + \frac{d}{2}y^2$$

which has a strict saddle point at $x_0 = (0, \ldots, 0)^\top$ and two global minima at $(1, \ldots, 1)^\top$ and $(-1, \ldots, -1)^\top$.

In this experiment, we test Algorithm 1 with perturbation in the tangent space step (Perturbed-RAZGD) along with two Euclidean accelerated zeroth-order algorithms, ZO-Perturbed-AGD, and ZO-Perturbed-AGD-ANCF in (Zhang & Gu, 2022). We choose the retraction as $\mathrm{Retr}_x(s) = x + s$. The basic parameters for all three algorithms follow respective theorems. The smoothing parameter μ is set to 0.01 for each algorithm, and notably, we run an additional choice of $\mu = 0.3$ for our algorithm. The initial point is set as the saddle point x_0. Due to the inherent randomness in these algorithms, each algorithm is executed 10 times and we report the averaged function value versus the averaged number of function queries in Figure 4. Figure 4 demonstrates that the variance of our algorithm (indicated by the width of the shadow) is smaller than the other two with the same smoothing parameter $\mu = 0.01$. Furthermore, our algorithm still convergences even with a larger smoothing parameter $\mu = 0.3$. The two aspects visually showcase the robustness of our accelerated zeroth-order algorithm.

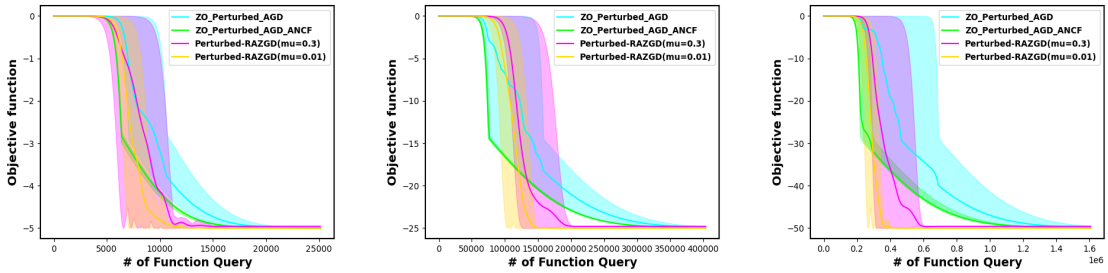
5.2. Lower function queries

In this part, we assess the acceleration effectiveness of the non-perturbed RAZGD with Option I by comparing it with Riemannian zeroth-order gradient descent (RZGD) and Euclidean projected zeroth-order gradient descent (PZGD). The corresponding pseudocodes are left in Appendix G.

We first consider the simplex constrained least-square problem (Li et al., 2023c)

$$\begin{aligned} \min \quad & \|Ax - b\|_2^2 \\ \text{s.t.} \quad & x \in \Delta^{d-1}, \end{aligned}$$

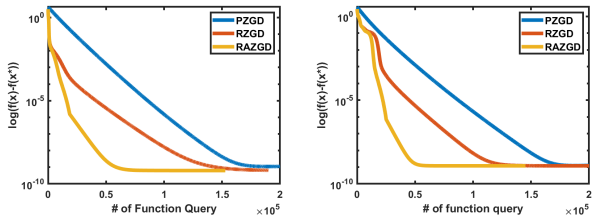
where $\Delta^{d-1} = \{x \in \mathbb{R}^d : \sum_{i=1}^d x_i = 1 \text{ and } x \geq 0\}$, $A \in \mathbb{R}^{m\times d}$ and $b \in \mathbb{R}^m$. Since the positive orthant is a Riemannian manifold with the Shahshahani metric, the simplex has a natural submanifold structure (Shahshahani, 1979). For any point x in the interior of Δ^{d-1}, the tangent space is the hyperplane passing through 0 and parallel to $x \in \Delta^{d-1}$, i.e. $\mathrm{T}_x\,\Delta^{d-1} = \{s \in \mathbb{R}^d : \sum_{j=1}^d s_j = 0\}$. We use the exponential map on the Shahshahani manifold as the retraction (Feng et al., 2022). A detailed discussion of Riemannian geometry of the simplex is deferred to Appendix



(a) d=20 (b) d=100 (c) d=200

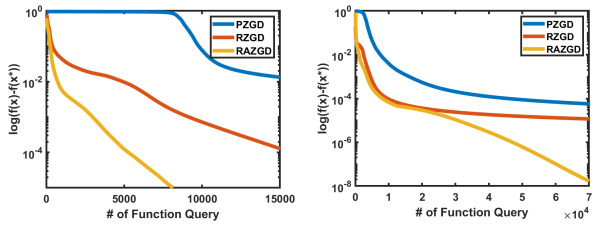
Figure 4. Performance of different zeroth-order accelerated algorithms to minimize the quartic function with growing dimensions. Confidence intervals show mini-max intervals over ten runs.

H. In the experiment, the feature matrix A is drawn from a standard Gaussian distribution, and the label vector b is generated using the expression $A\zeta + \mu$. Here, ζ and μ are randomly sampled from a Gaussian distribution, with the additional constraint that the sum of all elements equals 1 for ζ. For PZGD, we apply the projection in (Chen & Ye, 2011). The results are reported in Figure 5, showing that RAZGD requires lower function queries.



(a) m=200,d=20 (b) m=300,d=30

Figure 5. Performance on linear least-squares over the unit simplex with different problem sizes.



(a) Category:3 Dimension:10 (b) Category:2 Dimension:20

Figure 6. Performance on empirical hypervolume under the sphere manifold with different categories and dimensions.

To further demonstrate the efficiency, we consider a real-world application: the optimal linear combination of continuous predictors in the context of a binary classification problem (Das et al., 2022). For multi-category responses, the optimal predictor combination can be obtained by maximization of the empirical hypervolume under the manifold, with the following form

$$\begin{aligned} \max \quad & f(x) \\ \text{s.t.} \quad & \sum_{i=1}^{d} x_i^2 = 1, x \in \mathbb{R}^d, \end{aligned}$$

where the objective function $f(\cdot)$ takes no analytic form. We test algorithms on both two disease categories and three disease categories, and results are shown in Figure 6. For the unit sphere $\mathbb{S}^{d-1}$, the tangent space is defined as $\mathrm{T}_x\,\mathbb{S}^{d-1} := \{s \in \mathbb{R}^d : \sum_{j=1}^{d} x_j s_j = 0\}$. In the experiment, the process of biomarker data generation is consistent with (Das et al., 2022). The retraction is chosen as $\mathrm{Retr}_x(s) = (x+s)/\|x+s\|_2$. For PZGD, we use $x/\|x\|_2$ as the projection to the unit sphere. It is worth mentioning that in practical scenarios, the lower function queries lead to less running time. Thus, RAZGD reaches the target accuracy within 30 seconds in both cases, while PZGD needs more than 300 seconds to achieve the same accuracy, indicating the effective performance of our accelerated algorithm.

6. Conclusions

In the paper, we introduce a Riemannian accelerated zeroth-order gradient descent based on the deterministic coordinate-wise zeroth-order estimator. Our accelerated algorithm attains the best-known function query complexity for achieving both ϵ-approximate first-order and second-order stationary points respectively. Notably, it allows a larger smoothing parameter and thus demonstrates better robustness. Furthermore, we also establish the asymptotic convergence behavior with probability 1. Experimental results are presented, verifying the superior performance in terms of both function query complexity and robustness.

Acknowledgement

We thank Huikang Liu (Shanghai Jiao Tong University) for several helpful discussions at the early stage of this paper. Xiao Wang acknowledges Grant 202110458 from Shanghai University of Finance and Economics and support from the Shanghai Research Center for Data Science and Decision Technology. This research is partially supported by the National Natural Science Foundation of China (Grants 72394360, 72394364, 72394365, 72171141) and Natural Science Foundation of Shanghai (No. 23ZR1445900).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

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

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

Alimisis, F., Orvieto, A., Becigneul, G., and Lucchi, A. Momentum improves optimization on riemannian manifolds. In *International conference on artificial intelligence and statistics*, pp. 1351–1359. PMLR, 2021.

Bai, Q., Agarwal, M., and Aggarwal, V. Escaping saddle points for zeroth-order non-convex optimization using estimated gradient descent. In *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6. IEEE, 2020.

Balasubramanian, K. and Ghadimi, S. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, pp. 1–42, 2022.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Bishop, R. L. and O'Neill, B. Manifolds of negative curvature. *Transactions of the American Mathematical Society*, 145:1–49, 1969.

Boumal, N. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

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

Bridson, M. R. and Haefliger, A. *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media, 2013.

Bubeck, S., Jiang, Q., Lee, Y. T., Li, Y., and Sidford, A. Near-optimal method for highly smooth convex optimization. In *Conference on Learning Theory*, pp. 492–507. PMLR, 2019.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. "convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *International conference on machine learning*, pp. 654–663. PMLR, 2017.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

Chen, Y. and Ye, X. Projection onto a simplex. *arXiv preprint arXiv:1101.6081*, 2011.

Criscitiello, C. and Boumal, N. Efficiently escaping saddle points on manifolds. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Das, P., De, D., Maiti, R., Kamal, M., Hutcheson, K. A., Fuller, C. D., Chakraborty, B., and Peterson, C. B. Estimating the optimal linear combination of predictors using spherically constrained optimization. *BMC bioinformatics*, 23(3):1–20, 2022.

Fan, X., Gao, Z., Wu, Y., Jia, Y., and Harandi, M. Learning a gradient-free Riemannian optimizer on tangent spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7377–7384, 2021.

Feng, Y., Panageas, I., and Wang, X. Accelerated multiplicative weights update avoids saddle points almost always. *arXiv preprint arXiv:2204.11407*, 2022.

Flokas, L., Vlatakis-Gkaragkounis, E. V., and Piliouras, G. Efficiently avoiding saddle points with zero order methods: No gradients required. In *NeurIPS*, 2019.

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge university press, 2012.

Jaquier, N., Rozo, L. D., Caldwell, D. G., and Calinon, S. Geometry-aware tracking of manipulability ellipsoids. In *Robotics: Science and Systems*, number CONF, 2018.

Jaquier, N., Rozo, L., Calinon, S., and Bürger, M. Bayesian optimization meets riemannian manifolds in robot learning. In *Conference on Robot Learning*, pp. 233–246. PMLR, 2020.

Jiang, B., Wang, H., and Zhang, S. An optimal high-order tensor method for convex optimization. *Mathematics of Operations Research*, 46(4):1390–1412, 2021.

Jin, C., Netrapalli, P., and Jordan, M. I. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference On Learning Theory*, pp. 1042–1085. PMLR, 2018.

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

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

Li, H. and Lin, Z. Restarted nonconvex accelerated gradient descent: No more polylogarithmic factor in the $o(\epsilon^{-7/4})$ complexity. In *International Conference on Machine Learning*, pp. 12901–12916. PMLR, 2022.

Li, J., Balasubramanian, K., and Ma, S. Stochastic zeroth-order Riemannian derivative estimation and optimization. *Mathematics of Operations Research*, 48(2):1183–1211, 2023a.

Li, J., Balasubramanian, K., and Ma, S. Zeroth-order Riemannian averaging stochastic approximation algorithms. *arXiv preprint arXiv:2309.14506*, 2023b.

Li, Q., McKenzie, D., and Yin, W. From the simplex to the sphere: faster constrained optimization using the hadamard parametrization. *Information and Inference: A Journal of the IMA*, 12(3):iaad017, 2023c.

Lian, X., Zhang, H., Hsieh, C.-J., Huang, Y., and Liu, J. A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order. *Advances in Neural Information Processing Systems*, 29, 2016.

Lin, H., Mairal, J., and Harchaoui, Z. A universal catalyst for first-order optimization. *Advances in neural information processing systems*, 28, 2015.

Lin, L., Saparbayeva, B., Zhang, M. M., and Dunson, D. B. Accelerated algorithms for convex and non-convex optimization on manifolds. *arXiv preprint arXiv:2010.08908*, 2020.

Liu, S., Kailkhura, B., Chen, P.-Y., Ting, P., Chang, S., and Amini, L. Zeroth-order stochastic variance reduction for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Liu, S., Chen, P.-Y., Kailkhura, B., Zhang, G., Hero III, A. O., and Varshney, P. K. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54, 2020.

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

Lucchi, A., Orvieto, A., and Solomou, A. On the second-order convergence properties of random search methods. *Advances in Neural Information Processing Systems*, 34: 25633–25645, 2021.

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

Mertikopoulos, P. and Sandholm, W. H. Riemannian game dynamics. *Journal of Economic Theory*, 2018.

Nesterov, Y. Accelerating the cubic regularization of newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

Nesterov, Y. and Spokoiny, V. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

Nesterov, Y. E. A method of solving a convex programming problem with convergence rate o\bigl(k^2\bigr). In *Doklady Akademii Nauk*, volume 269, pp. 543–547. Russian Academy of Sciences, 1983.

Nguyen, A. and Balasubramanian, K. Stochastic zeroth-order functional constrained optimization: Oracle complexity and applications. *INFORMS Journal on Optimization*, 5(3):256–272, 2023.

Ostrowski, A. M. On some metrical properties of operator matrices and matrices partitioned into blocks. *Journal of Mathematical Analysis and Applications*, 2(2):161–209, 1961.

Panageas, I., Piliouras, G., and Wang, X. First-order methods almost always avoid saddle points: The case of vanishing step-sizes. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 6471–6480, 2019.

Ren, Z., Tang, Y., and Li, N. Escaping saddle points in zeroth-order optimization: the power of two-point estimators. In *International Conference on Machine Learning*, pp. 28914–28975. PMLR, 2023.

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

Shahshahani, S. *A New Mathematical Framework for the Study of linage and Selection*, volume 17. American Mathematical Society, 1979.

Shub, M. *Global stability of dynamical systems*. Springer Science & Business Media, 1987.

Vlatakis-Gkaragkounis, E.-V., Flokas, L., and Piliouras, G. Efficiently avoiding saddle points with zero order methods: No gradients required. *Advances in neural information processing systems*, 32, 2019.

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

Wang, T., Huang, Y., and Li, D. From the greene–wu convolution to gradient estimation over Riemannian manifolds. *arXiv preprint arXiv:2108.07406*, 2021.

Zhang, H. and Gu, B. Faster gradient-free methods for escaping saddle points. In *The Eleventh International Conference on Learning Representations*, 2022.

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

Zhang, H. and Sra, S. Towards riemannian accelerated gradient methods. *arXiv preprint arXiv:1806.02812*, 2018.

Zhang, H., Xiong, H., and Gu, B. Zeroth-order negative curvature finding: Escaping saddle points without gradients. *Advances in Neural Information Processing Systems*, 35: 38332–38344, 2022.

A. Further Related work

Zeroth-order optimization on Riemannian manifolds. Riemannian zeroth-order algorithms typically involve two key steps: constructing Riemannian zeroth-order estimators and applying them with standard optimization algorithms, such as Riemannian gradient descent. Through noisy evaluations of the objective function, Li et al. (2023a) studied the randomized zeroth-order estimators for the Riemannian gradient and Hessian, extending the Gaussian smoothing technique (Nesterov & Spokoiny, 2017; Balasubramanian & Ghadimi, 2022) onto the Riemannian manifold. Subsequently, Wang et al. (2021) proposed an alternative zeroth-order gradient estimator based on the Greene–Wu convolution over Riemannian manifolds, demonstrating superior approximation quality compared to the approach by Li et al. (2023a). When it comes to the Riemannian Hessian, Wang (2023) introduced a novel Riemannian zeroth-order estimator that relies solely on constant function evaluations. Turning to Riemannian zeroth-order algorithms, Fan et al. (2021) first proposed a Riemannian meta-optimization method that learns a gradient-free optimizer without theoretical guarantees. Li et al. (2023a) studied several zeroth-order algorithms for stochastic Riemannian optimization, presenting the first complexity results. Subsequently, they improved sample complexities by introducing zeroth-order Riemannian averaging stochastic approximation algorithms in (Li et al., 2023b). Moreover, Maass et al. (2022) studied the exploration of zeroth-order algorithms in the context of Riemannian online learning.

Acceleration on Riemannian manifolds. The main challenge in Riemannian optimization arises from the nonlinear structure of Riemannian manifolds, and two powerful techniques have been developed. The first involves leveraging trigonometric comparison inequalities (Zhang & Sra, 2016; Alimisis et al., 2021), while the second utilizes the tangent space step (Criscitiello & Boumal, 2019; 2022). In cases where the objective function is geodesically convex (Bishop & O'Neill, 1969; Bridson & Haefliger, 2013), a recent line of work (Liu et al., 2017; Zhang & Sra, 2018; Lin et al., 2020; Alimisis et al., 2021; Jin & Sra, 2022; Kim & Yang, 2022) focused on generalizing Nesterov's accelerated update to Riemannian optimization, mirroring the well-known convergence result of accelerated gradient descent on Euclidean convex optimization. Moreover, outside the geodesic convexity, Criscitiello & Boumal (2022) established the extension of Euclidean nonconvex acceleration techniques (Jin et al., 2018; Carmon et al., 2018) to Riemannian manifolds, improving the convergence rate compared to Riemannian gradient descent (Boumal et al., 2019; Criscitiello & Boumal, 2019).

B. Discussion of Assumptions

In the paper, the Lipschitz-type continuity of the pullback function follows from previous works (Boumal et al., 2019; Agarwal et al., 2021; Criscitiello & Boumal, 2019; 2022), and a detailed comparison of the parallel transport based Lipschitz continuity, such as

$$\left\|\operatorname{grad} f(x) - \Gamma_y^x \operatorname{grad} f(y)\right\| \leq O(d_{\mathcal{M}}(x, y))$$

is provided in the textbook (Boumal, 2023), where $\Gamma_y^x : \mathrm{T}_y \mathcal{M} \to \mathrm{T}_x \mathcal{M}$ denotes parallel transport from y to x along any minimizing geodesic, and $d_{\mathcal{M}}(x, y)$ is the Riemannian distance. Since our interest is developing Riemannian zeroth-order algorithms, the Hessian Lipschitz continuity in Assumption 4.2 is stronger compared to those in (Criscitiello & Boumal, 2019) and (Criscitiello & Boumal, 2022). Whereas in the special case where $\mathcal{M} = \mathbb{R}^d$ and $\operatorname{Retr}_x(s) = x + s$, both Assumptions 4.1 and 4.2 reduce to the standard Lipschitz continuity in Euclidean space.

For the well-behaved retraction mapping (Assumption 4.3), when the sectional curvature and the covariant derivative of the Riemann curvature endomorphism are both bounded, exponential mapping ensures it holds (Theorem 2.7 in (Criscitiello & Boumal, 2022)). For more details, readers can refer to (Agarwal et al., 2021; Criscitiello & Boumal, 2022; Boumal, 2023).

C. Auxiliary Lemmas

We first list the concept of the adjoint of a linear operator, which is essential in bridging the differential and Hessian of a function on a manifold and their counterparts obtained by interplay with retraction map.

Definition C.1. *Let E and E' be two Euclidean spaces, with inner products $\langle , \rangle_a$ and $\langle , \rangle_b$ respectively. Let $A : E \to E'$ be a linear operator. The adjoint of A is a linear operator $A^* : E' \to E$ defined by this property:*

$$\forall u \in E, v \in E', \quad \langle A(u), v\rangle_b = \langle u, A^*(v)\rangle_a.$$

In particular, if A maps E to E equipped with an inner product $\langle , \rangle$ and

$$\forall u, v \in E, \quad \langle A(u), v\rangle = \langle u, A(v)\rangle,$$

this is, if $A = A^*$, we say A is self-adjoint.

Several useful lemmas and inequalities are presented below.

Lemma C.1 (Lemma 2.5 in (Criscitiello & Boumal, 2022)). *For $f : \mathcal{M} \to \mathbb{R}$ twice continuously differentiable, $x \in \mathcal{M}$ and $s \in \mathrm{T}_x\mathcal{M}$, with $T^*_{x,s}$ denoting the adjoint of $T_{x,s}$,*

$$\nabla \hat{f}_x(s) = T^*_{x,s} \operatorname{grad} f\left(\operatorname{Retr}_x(s)\right), \quad \nabla^2 \hat{f}_x(s) = T^*_{x,s} \operatorname{Hess} f\left(\operatorname{Retr}_x(s)\right) T_{x,s} + W_s,$$

where $T_{x,s}$ si the differential of Retr_x at s (a linear operator):

$$T_{x,s} = \mathrm{D} \operatorname{Retr}_x(s) : \mathrm{T}_x \mathcal{M} \to \mathrm{T}_{\operatorname{Retr}_x(s)} \mathcal{M},$$

and W_s is a self-adjoint linear operator on $\mathrm{T}_x \mathcal{M}$ defined through polarization by

$$\langle W_s[\dot{s}], \dot{s} \rangle = \left\langle \operatorname{grad} f\left(\operatorname{Retr}_x(s)\right), \gamma''_{x,s}(0) \right\rangle,$$

with $\gamma''_{x,s}(0) \in \mathrm{T}_{\operatorname{Retr}_x(s)}\mathcal{M}$ the intrinsic acceleration on $\mathcal{M}$ of $\gamma(\tau) = \operatorname{Retr}_x(s + \tau\dot{s})$ at $\tau = 0$.

Lemma C.2 (Mechanism in (Li & Lin, 2022)). *For the tangent space step (Subroutine 2), denote $\mathcal{K}$ to be the iteration number when the "if condition" on Line 7 triggers, i.e.*

$$\mathcal{K} = \min_k \left\{ k : k \sum_{j=0}^{k-1} \|s_x^{j+1} - s_x^j\|^2 > B^2 \right\}.$$

Then for each $k = 0, 1, \cdots, \mathcal{K} - 1$, it holds that

$$\|s_x^k - s_x^0\| \le B,$$
$$\|y_x^k - s_x^0\| \le 2B.$$

When the "if condition" does not trigger, for all $k = 0, 1, \cdots, K$, it holds that

$$\|s_x^k - s_x^0\| \le B,$$
$$\|y_x^k - s_x^0\| \le 2B.$$

Lemma C.3 (Young's inequality). *If $a \ge 0$ and $b \ge 0$ are nonnegative real numbers and if $p > 1$ and $q > 1$ are real numbers such that $\frac{1}{p} + \frac{1}{q} = 1$, then*

$$ab \le \frac{a^p}{p} + \frac{b^q}{q}.$$

Equality holds if and only if $a^p = b^q$. Specifically, for any $\epsilon > 0$, it holds that

$$ab \le \frac{a^2}{2\varepsilon} + \frac{\varepsilon b^2}{2}.$$

Lemma C.4 (Minkowski's inequality). *Given $x_1, \ldots, x_n \in \mathbb{R}$ and $y_1, \ldots, y_n \in \mathbb{R}$, for any $p > 0$, it holds that*

$$\left(\sum_{k=1}^n |x_k + y_k|^p \right)^{\frac{1}{p}} \le \left(\sum_{k=1}^n |x_k|^p \right)^{\frac{1}{p}} + \left(\sum_{k=1}^n |y_k|^p \right)^{\frac{1}{p}}.$$

D. Approximation Error of the Estimator

Lemma D.1. *Suppose that Assumption 4.1 and 4.2 holds For any smoothing parameter $\mu \in (0, b)$ and $(x, s_x) \in \mathrm{T}\,\mathcal{M}$ satisfying $\|s_x\| \in \mathbb{B}_{x, b-\mu}(0)$, the Riemannian coordinate-wise zeroth-order estimator in Definition 3.1 satisfies*

$$\left\| g_x(s_x; \mu) - \nabla \hat{f}_x(s_x) \right\| \le \min \left\{ \frac{l\mu\sqrt{d}}{2}, \frac{\rho\mu^2\sqrt{d}}{6} \right\}.$$

Proof. First note that

$$\begin{aligned}
&\left|\hat{f}_x(s_x+\mu e_i)-\hat{f}_x(s_x-\mu e_i)-2\mu\langle\nabla\hat{f}_x(s_x),e_i\rangle\right| \\
=&\left|\left(\hat{f}_x(s_x+\mu e_i)-\hat{f}_x(s_x)-\mu\langle\nabla\hat{f}_x(s_x),e_i\rangle\right)-\left(\hat{f}_x(s_x-\mu e_i)-\hat{f}_x(s_x)+\mu\langle\nabla\hat{f}_x(s_x),e_i\rangle\right)\right| \\
\leq&\left|\hat{f}_x(s_x+\mu e_i)-\hat{f}_x(s_x)-\mu\langle\nabla\hat{f}_x(s_x),e_i\rangle\right|+\left|\hat{f}_x(s_x-\mu e_i)-\hat{f}_x(s_x)+\mu\langle\nabla\hat{f}_x(s_x),e_i\rangle\right| \\
\leq&l\mu^2,
\end{aligned} \tag{2}$$

where the last inequality holds due to the Lipschitz continuity of $\nabla\hat{f}_x(\cdot)$ in Assumption 4.1. Consequently, we have

$$\begin{aligned}
&\left\|g_x(s_x;\mu)-\hat{\nabla}f_x(s)\right\| \\
=&\left\|\sum_{i=1}^{d}\frac{\hat{f}_x(s_x+\mu e_i)-f_x(s_x-\mu e_i)}{2\mu}e_i-\sum_{i=1}^{d}\langle\nabla\hat{f}_x(s_x),e_i\rangle e_i\right\| \\
=&\frac{1}{2\mu}\left\|\sum_{i=1}^{d}\left(\hat{f}_x(s_x+\mu e_i)-\hat{f}_x(s_x-\mu e_i)-2\mu\langle\nabla\hat{f}_x(s_x),e_i\rangle\right)e_i\right\| \\
=&\frac{1}{2\mu}\sqrt{\sum_{i=1}^{d}\left(\hat{f}_x(s_x+\mu e_i)-\hat{f}_x(s_x-\mu e_i)-2\mu\langle\nabla\hat{f}_x(s_x),e_i\rangle\right)^2} \\
\leq&\frac{1}{2\mu}\sqrt{dl^2\mu^4} \\
=&\frac{l\mu\sqrt{d}}{2}.
\end{aligned}$$

Note that it also holds that

$$\begin{aligned}
&\hat{f}_x(s_x+\mu e_i)-\hat{f}_x(s_x-\mu e_i)-2\mu\langle\nabla\hat{f}_x(s_x),e_i\rangle \\
=&\left(\hat{f}_x(s_x+\mu e_i)-\hat{f}_x(s_x)-\mu\langle\nabla\hat{f}_x(s_x),e_i\rangle-\frac{\mu^2}{2}\langle\nabla^2\hat{f}_x(s_x)e_i,e_i\rangle\right) \\
-&\left(\hat{f}_x(s_x-\mu e_i)-\hat{f}_x(s_x)+\mu\langle\nabla\hat{f}_x(s_x),e_i\rangle-\frac{\mu^2}{2}\langle\nabla^2\hat{f}_x(s_x)e_i,e_i\rangle\right),
\end{aligned}$$

and thus the same argument in (2) gives

$$\left|\hat{f}_x(s_x+\mu e_i)-\hat{f}_x(s_x-\mu e_i)-2\mu\langle\nabla\hat{f}_x(s_x),e_i\rangle\right|\leq\frac{\rho\mu^3}{3}.$$

Similarly, we establish

$$\left\|g_x(s_x;\mu)-\hat{\nabla}f_x(s)\right\|\leq\frac{1}{2\mu}\sqrt{d\left(\frac{\rho\mu^3}{3}\right)^2}=\frac{\rho\mu^2\sqrt{d}}{6}.$$

Therefore, we can conclude

$$\left\|g_x(s_x;\mu)-\nabla\hat{f}_x(s_x)\right\|\leq\min\left\{\frac{l\mu\sqrt{d}}{2},\frac{\rho\mu^2\sqrt{d}}{6}\right\}.$$

□

For simplicity, we use $\mathbf{E}(\mu)=\min\left\{\frac{l\mu\sqrt{d}}{2},\frac{\rho\mu^2\sqrt{d}}{6}\right\}$ to represent the upper bound of approximation error of the Riemannian coordinate-wise zeroth-order estimator. This notation is widely used throughout the non-asymptotic convergence analysis.

E. Proofs of Non-asymptotic Convergence Analysis

In the following non-asymptotic analysis, the magnitudes of parameters in RAZGD (Algorithm 1) are set as:

$$\eta = \frac{1}{4l},\ B = \tilde{\mathcal{O}}\left(\epsilon^{\frac{1}{2}}\right),\ \theta = \mathcal{O}\left(\epsilon^{\frac{1}{4}}\right),\ r = \mathcal{O}(\epsilon),\ K = \tilde{\mathcal{O}}\left(\epsilon^{-\frac{1}{4}}\right). \tag{3}$$

E.1. Riemannian zeroth-order gradient descent step

For the iterate x_t with a relatively large zeroth-order estimator, i.e. $\|g_{x_t}(0;\mu)\| \geq lB$, we show that the Riemannian zeroth-order gradient descent step (Subroutine 1) results in the function value decrease of $\mathcal{O}(B^2)$.

Lemma E.1. *Suppose that Assumption 4.1 and 4.2 hold. Under the parameter setting (3), choose a reasonably small μ such that $\mathbf{E}(\mu) \leq \frac{lB}{2}$ in Algorithm 1. Then, for the iterate x_t satisfying $\|g_{x_t}(0;\mu)\| \geq lB$, we have:*

$$f(x_{t+1}) \leq f(x_t) - \min\left\{\frac{lB^2}{16}, lb^2\right\}.$$

Proof. First, consider the scenario where $\|g_{x_t}(0;\mu)\| \leq \frac{b}{\eta}$; thus, $\eta g_{x_t}(0;\mu) \in \mathbb{B}_{x_t,b}(0)$, ensuring that local Lipschitz continuity holds. Based on Assumption 4.1, we have

$$\begin{aligned}
&f(x_{t+1})\\
=&f(\mathrm{Retr}_{x_t}(-\eta g_{x_t}(0;\mu)))\\
=&\hat{f}_{x_t}(-\eta g_{x_t}(0;\mu))\\
\leq&\hat{f}_{x_t}(0) - \eta\langle\nabla\hat{f}_{x_t}(0), g_{x_t}(0;\mu)\rangle + \frac{l\eta^2}{2}\|g_{x_t}(0;\mu)\|^2\\
=&f(x_t) - \frac{\eta}{2}\left(\|\nabla\hat{f}_{x_t}(0)\|^2 + \|g_{x_t}(0;\mu)\|^2 - \|\nabla\hat{f}_{x_t}(0) - g_{x_t}(0;\mu)\|^2\right) + \frac{l\eta^2}{2}\|g_{x_t}(0;\mu)\|^2\\
\leq&f(x_t) - \frac{\eta}{2}\left(1 - l\eta\right)\|g_{x_t}(0;\mu)\|^2 + \frac{\eta}{2}\mathbf{E}(\mu)^2.
\end{aligned}$$

Substituting $\eta = \frac{1}{4l}$, $\mathbf{E}(\mu) \leq \frac{lB}{2}$ and $\|g_{x_t}(0;\mu)\| \geq lB$ gives that

$$f(x_{t+1}) \leq f(x_t) - \frac{lB^2}{16}.$$

For the extremely large estimator $\|g_{x_t}(0;\mu)\| \geq \frac{b}{\eta}$, similarly, it holds that

$$\begin{aligned}
&f(x_{t+1})\\
=&\hat{f}_{x_t}(-\alpha\eta g_{x_t}(0;\mu))\\
\leq&\hat{f}_{x_t}(0) - \alpha\eta\langle\nabla\hat{f}_{x_t}(0), g_{x_t}(0;\mu)\rangle + \frac{l\alpha^2\eta^2}{2}\|g_{x_t}(0;\mu)\|^2\\
=&f(x_t) - \frac{\alpha\eta}{2}\left(\|\nabla\hat{f}_{x_t}(0)\|^2 + \|g_{x_t}(0;\mu)\|^2 - \|\nabla\hat{f}_{x_t}(0) - g_{x_t}(0;\mu)\|^2\right) + \frac{l\alpha^2\eta^2}{2}\|g_{x_t}(0;\mu)\|^2\\
\leq&f(x_t) - \frac{4lb^2}{2\alpha} + \frac{\alpha\eta}{2}\mathbf{E}(\mu)^2 + \frac{lb^2}{2} &\text{(5a)}\\
\leq&f(x_t) - lb^2, &\text{(5b)}
\end{aligned}$$

where we use $\alpha\|g_{x_t}(0;\mu)\| = \frac{b}{\eta}$ and $\eta = \frac{1}{4l}$ in (5a), and (5b) holds because $\alpha < 1$ and $\mathbf{E}(\mu)^2 \leq \frac{l^2B^2}{4} \leq \frac{lb^2}{2}$. Therefore, we conclude

$$f(x_{t+1}) \leq f(x_t) - \min\left\{\frac{lB^2}{16}, lb^2\right\}.$$

□

E.2. Tangent space step: function value decrease

In this subsection, we establish that the tangent space step results in the function value decrease for the case when the "if condition" (Line 8 of Subroutine 2) triggers. According to Lemma C.2, we know that when the "if condition" triggers, for each $k = 0, 1, \ldots, \mathcal{K}-1$, $s_{x_t}^k \in \mathbb{B}_{x_t,b}(0)$ and $y_{x_t}^k \in \mathbb{B}_{x_t,b}(0)$. The following lemma states that $s_{x_t}^{\mathcal{K}}$ stays within the ball $\mathbb{B}_{x_t,b}(0)$ as well, and thus, the local Lipschitz continuity holds for all iterates in the tangent space step.

Lemma E.2. *Suppose that Assumption 4.1 and 4.2 hold. Under the parameter setting (3), choose a reasonably small μ such that $\mathbf{E}(\mu) \le \frac{lB}{2}$ holds in Algorithm 1. Then, for the tangent space step at iterate x_t, when "if condition" triggers, we have:*

$$\|\nabla \hat{f}_{x_t}(y_{x_t}^{\mathcal{K}-1})\| \le 4lB, \text{ and } \|s_{x_t}^{\mathcal{K}} - s_{x_t}^0\| \le 4B.$$

Proof. By the mechanism of Algorithm 1, we know that the zeroth-order estimator $g_{x_t}(0;\mu)$ satisfies $\|g_{x_t}(0;\mu)\| \le lB$. Recall $s_{x_t}^0 = \xi_t \sim \mathrm{Uni}\left(\mathbb{B}_{x_t,r}(0)\right)$, we have

$$\begin{aligned}
&\|\nabla \hat{f}_{x_t}(s_{x_t}^0)\| \\
\le&\|\nabla \hat{f}_{x_t}(s_{x_t}^0) - \nabla \hat{f}_{x_t}(0)\| + \|\nabla \hat{f}_{x_t}(0) - g_{x_t}(0;\mu)\| + \|g_{x_t}(0;\mu)\| \\
\le& l \cdot \|\xi_t\| + \mathbf{E}(\mu) + lB \\
\le& 2lB,
\end{aligned}$$

where the last inequality uses $\|\xi_t\| = r = \mathcal{O}(\epsilon) \le \frac{B}{2}$. Therefore, we could upper bound $\|\nabla \hat{f}_{x_t}(y_{x_t}^{\mathcal{K}-1})\|$ as

$$\|\nabla \hat{f}_{x_t}(y_{x_t}^{\mathcal{K}-1})\| \le \|\nabla \hat{f}_{x_t}(y_{x_t}^{\mathcal{K}-1}) - \nabla \hat{f}_{x_t}(s_{x_t}^0)\| + \|\nabla \hat{f}_{x_t}(s_{x_t}^0)\| \le l\|y_{x_t}^{\mathcal{K}-1} - s_{x_t}^0\| + 2lB \le 4lB.$$

Since $s_{x_t}^{\mathcal{K}} = y_{x_t}^{\mathcal{K}-1} - \eta g_{x_t}(y_{x_t}^{\mathcal{K}-1};\mu)$, it follows that

$$\begin{aligned}
&\|s_{x_t}^{\mathcal{K}} - s_{x_t}^0\| \\
\le&\|s_{x_t}^{\mathcal{K}} - y_{x_t}^{\mathcal{K}-1}\| + \|y_{x_t}^{\mathcal{K}-1} - s_{x_t}^0\| \\
\le&\eta\|g_{x_t}(y_{x_t}^{\mathcal{K}-1};\mu)\| + 2B \\
\le&\eta\|g_{x_t}(y_{x_t}^{\mathcal{K}-1};\mu) - \nabla \hat{f}_{x_t}(y_{x_t}^{\mathcal{K}-1})\| + \eta\|\nabla \hat{f}_{x_t}(y_{x_t}^{\mathcal{K}-1})\| + 2B \\
\le&\eta\mathbf{E}(\mu) + \eta \cdot 4lB + 2B \\
\le&4B,
\end{aligned}$$

where the last inequality holds due to that $\eta = \frac{1}{4l}$ and $\mathbf{E}(\mu) \le \frac{lB}{2}$. □

To establish the function value decrease in the tangent space step, we mimic the proof strategy in (Li & Lin, 2022). First note that $\nabla^2 \hat{f}_{x_t}(s_t^0)$ is self-adjoint, there exists a basis of eigenvectors $\{u_j\}_{j=1}^d$ satisfying

$$\nabla^2 \hat{f}_{x_t}(s_t^0)u_j = \lambda_j u_j,$$

where $\lambda_1, \cdots, \lambda_d$ are associated eigenvalues. Based on the basis $\{u_j\}_{j=1}^d$ and local coordinate $\{e_j\}_{j=1}^d$ of tangent space $\mathrm{T}_{x_t}\mathcal{M}$, we introduce the following notations for any given $s_{x_t}, \nabla \hat{f}_{x_t}(\cdot) \in \mathrm{T}_{x_t}\mathcal{M}$:

$$\begin{aligned}
&\tilde{s}_{x_t,j} = \langle s_{x_t}, u_j\rangle,\ s_{x_t,j} = \langle s_{x_t}, e_j\rangle,\ j = 1, \cdots, d, \\
&\tilde{\nabla}_j \hat{f}_{x_t}(\cdot) = \langle \nabla \hat{f}_{x_t}(\cdot), u_j\rangle,\ \nabla_j \hat{f}_{x_t}(\cdot) = \langle \nabla \hat{f}_{x_t}(\cdot), e_j\rangle,\ j = 1, \cdots, d.
\end{aligned}$$

Therefore, it holds that

$$
\begin{aligned}
&s_{x_t} = \sum_{j=1}^{d} \tilde{s}_{x_t,j} u_j = \sum_{j=1}^{d} s_{x_t,j} e_j, \\
&\|s_{x_t}\|^2 = \sum_{j=1}^{d} |\tilde{s}_{x_t,j}|^2 = \sum_{j=1}^{d} |s_{x_t,j}|^2, \\
&\hat{f}_{x_t}(\cdot) = \sum_{j=1}^{d} \tilde{\nabla}_j \hat{f}_{x_t}(\cdot) u_j = \sum_{j=1}^{d} \nabla_j \hat{f}_{x_t}(\cdot) e_j, \\
&\|\hat{f}_{x_t}(\cdot)\|^2 = \sum_{j=1}^{d} |\tilde{\nabla}_j \hat{f}_{x_t}(\cdot)|^2 = \sum_{j=1}^{d} |\nabla_j \hat{f}_{x_t}(\cdot)|^2.
\end{aligned}
$$

Lemma E.3. *Suppose that Assumption 4.1 and 4.2 hold. Under the parameter setting (3), choose a reasonably small μ such that $\mathbf{E}(\mu) \leq \frac{lB}{2}$ in Algorithm 1. Then, for the tangent space step at iterate x_t, when the "if condition" triggers, we have*

$$
\hat{f}_{x_t}(s_{x_t}^{\mathcal{K}}) \leq \hat{f}_{x_t}(s_{x_t}^{0}) + \frac{32\rho B^3}{3} + \sum_{j=1}^{d} h_j(\tilde{s}_{x_t,j}^{\mathcal{K}}),
$$

where

$$
h_j(z) = \langle \tilde{\nabla}_j \hat{f}_{x_t}(s_{x_t}^{0}), z - \tilde{s}_{x_t,j}^{0} \rangle + \frac{\lambda_j}{2}(z - \tilde{s}_{x_t,j}^{0})^2,\ j = 1, \ldots, d.
$$

are one-dimensional quadratic functions.

Proof. From the Hessian Lipschitz continuity (Assumption 4.2), we have

$$
\begin{aligned}
&\hat{f}_{x_t}(s_{x_t}^{\mathcal{K}}) \\
\leq &\hat{f}_{x_t}(s_{x_t}^{0}) + \langle \nabla \hat{f}_{x_t}(s_{x_t}^{0}), s_{x_t}^{\mathcal{K}} - s_{x_t}^{0} \rangle + \frac{1}{2}\langle \nabla^2 \hat{f}_{x_t}(s_{x_t}^{0})(s_{x_t}^{\mathcal{K}} - s_{x_t}^{0}), s_{x_t}^{\mathcal{K}} - s_{x_t}^{0} \rangle + \frac{\rho}{6}\|s_{x_t}^{\mathcal{K}} - s_{x_t}^{0}\|^3 \\
\leq &\hat{f}_{x_t}(s_{x_t}^{0}) + \langle \nabla \hat{f}_{x_t}(s_{x_t}^{0}), s_{x_t}^{\mathcal{K}} - s_{x_t}^{0} \rangle + \frac{1}{2}\langle \nabla^2 \hat{f}_{x_t}(s_{x_t}^{0})(s_{x_t}^{\mathcal{K}} - s_{x_t}^{0}), s_{x_t}^{\mathcal{K}} - s_{x_t}^{0} \rangle + \frac{32\rho B^3}{3}, \qquad (8a) \\
= &\hat{f}_{x_t}(s_{x_t}^{0}) + \frac{32\rho B^3}{3} + \sum_{j=1}^{d} \langle \tilde{\nabla}_j \hat{f}_{x_t}(s_{x_t}^{0}), \tilde{s}_{x_t,j}^{\mathcal{K}} - \tilde{s}_{x_t,j}^{0} \rangle + \frac{\lambda_j}{2}(\tilde{s}_{x_t,j}^{\mathcal{K}} - \tilde{s}_{x_t,j}^{0})^2 \qquad (8b) \\
= &\hat{f}_{x_t}(s_{x_t}^{0}) + \frac{32\rho B^3}{3} + \sum_{j=1}^{d} h_j(\tilde{s}_{x_t,j}^{\mathcal{K}}),
\end{aligned}
$$

where (8a) comes from Lemma E.2, and (8b) holds because $\{u_j\}_{j=1}^{d}$ forms a standard basis of the tangent space $\mathrm{T}_{x_t}\mathcal{M}$. □

The above lemma indicates that it is sufficient to analyze the behavior of one-dimensional quadratic functions $h_j(\tilde{s}_{x_t,j}^{\mathcal{K}})$, $j = 1, \ldots, d$. Recall the k-th update in the tangent space step at iterate x_t:

$$
\begin{aligned}
y_{x_t}^{k} &= s_{x_t}^{k} + (1-\theta)(s_{x_t}^{k} - s_{x_t}^{k-1}), \\
s_{x_t}^{k+1} &= y_{x_t}^{k} - \eta g_{x_t}(y_{x_t}^{k}; \mu).
\end{aligned}
$$

It equivalents as

$$
\begin{aligned}
\tilde{y}_{x_t,j}^{k} &= \tilde{s}_{x_t,j}^{k} + (1-\theta)(\tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}), \\
\tilde{s}_{x_t,j}^{k+1} &= \tilde{y}_{x_t,j}^{k} - \eta \nabla h_j(\tilde{y}_{x_t,j}^{k}) - \eta \mathcal{E}_{x_t,j}^{k},
\end{aligned} \qquad (9)
$$

where

$$
\mathcal{E}_{x_t,j}^{k} = \langle g_{x_t}(y_{x_t}^{k}; \mu), u_j \rangle - \nabla h_j(\tilde{y}_{x_t,j}^{k}) := \tilde{g}_{x_t,j}(y_{x_t}^{k}; \mu) - \nabla h_j(\tilde{y}_{x_t,j}^{k}).
$$

for all $j = 1, \cdots, d$. As Li & Lin (2022) pointed out, the k-th update in the tangent space step can be viewed as applying inexact accelerated gradient descent to $h_j(\cdot)$ with the error $\mathcal{E}^k_{x_t,j}$. The following lemma describes the error that could be controlled.

Lemma E.4. *Suppose that Assumption 4.1 and 4.2 hold. Under the parameter setting (3), choose a reasonably small μ such that $\mathbf{E}(\mu) \le \frac{lB}{2}$ in Algorithm 1. Then, for the tangent space step at iterate x_t, the error $\mathcal{E}^k_{x_t,j}$ in update (9) satisfies*

$$\sum_{j=1}^{d} |\mathcal{E}^0_{x_t,j}|^2 \le \mathbf{E}(\mu)^2, \text{ and } \sum_{j=1}^{d} |\mathcal{E}^k_{x_t,j}|^2 \le 2\mathbf{E}(\mu)^2 + 8\rho^2 B^4, \ \forall k \ge 1.$$

Proof. For any $j = 1, \cdots, d$, since $y^0_{x_t} = s^0_{x_t}$, it holds that

$$\mathcal{E}^0_{x_t,j} = \tilde{g}_{x_t,j}(y^0_{x_t};\mu) - \nabla h_j(\tilde{y}^0_{x_t,j}) = \tilde{g}_{x_t,j}(y^0_{x_t};\mu) - \tilde{\nabla}_j \hat{f}_{x_t}(\tilde{y}^0_{x_t}).$$

Summing over j gives

$$\begin{aligned}\sum_{j=1}^{d} |\mathcal{E}^0_{x_t,j}|^2 =& \sum_{j=1}^{d} |\tilde{g}_{x_t,j}(y^0_{x_t};\mu) - \tilde{\nabla}_j \hat{f}_{x_t}(y^0_{x_t})|^2 \\ =& \|g_{x_t}(y^0_{x_t};\mu) - \nabla \hat{f}_{x_t}(y^0_{x_t})\|^2 \\ \le& \mathbf{E}(\mu)^2.\end{aligned}$$

For any $k \ge 1$, by the definition of $\mathcal{E}^k_{x_t,j}$, it follows

$$\begin{aligned}\sum_{j=1}^{d} |\mathcal{E}^k_{x_t,j}|^2 =& \sum_{j=1}^{d} |\tilde{g}_{x_t,j}(y^k_{x_t};\mu) - \nabla h_j(\tilde{y}^k_{x_t,j})|^2 \\ \le& 2\sum_{j=1}^{d} |\tilde{g}_{x_t,j}(y^k_{x_t};\mu) - \tilde{\nabla}_j \hat{f}_{x_t}(y^k_{x_t})|^2 + 2\sum_{j=1}^{d} |\tilde{\nabla}_j \hat{f}_{x_t}(y^k_{x_t}) - \nabla h_j(\tilde{y}^k_{x_t,j})|^2.\end{aligned}$$

For the first term, we have

$$\sum_{j=1}^{d} |\tilde{g}_{x_t,j}(y^k_{x_t};\mu) - \tilde{\nabla}_j \hat{f}_{x_t}(y^k_{x_t})|^2 = \|g_{x_t}(y^k_{x_t};\mu) - \nabla \hat{f}_{x_t}(y^k_{x_t})\|^2 \le \mathbf{E}(\mu)^2.$$

For the second term, since $\nabla^2 \hat{f}_{x_t}(s^0_t) u_j = \lambda_j u_j, \forall j$, we have

$$\begin{aligned}&\sum_{j=1}^{d} |\tilde{\nabla}_j \hat{f}_{x_t}(y^k_{x_t}) - \nabla h_j(\tilde{y}^k_{x_t,j})|^2 \\ =& \sum_{j=1}^{d} |\tilde{\nabla}_j \hat{f}_{x_t}(y^k_{x_t}) - \tilde{\nabla}_j \hat{f}_{x^t}(s^0_{x_t}) - \lambda_j(\tilde{y}^k_{x_t} - \tilde{s}^0_{x_t})|^2 \\ =& \sum_{j=1}^{d} |\langle \nabla \hat{f}_{x_t}(y^k_{x_t}) - \nabla \hat{f}_{x_t}(s^0_{x_t}) - \lambda_j(y^k_{x_t,j} - s^0_{x_t}), u_j\rangle|^2 \\ =& \sum_{j=1}^{d} |\langle \nabla \hat{f}_{x_t}(y^k_{x_t}) - \nabla \hat{f}_{x_t}(s^0_{x_t}) - \nabla^2 \hat{f}_{x_t}(s^0_t)(y^k_{x_t} - s^0_{x_t}), u_j\rangle|^2 \\ =& \|\nabla \hat{f}_{x_t}(y^k_{x_t}) - \nabla \hat{f}_{x_t}(s^0_{x_t}) - \nabla^2 \hat{f}_{x_t}(s^0_t)(y^k_{x_t} - s^0_{x_t})\|^2 \\ \le& \frac{\rho^2}{4}\|y^k_{x_t} - s^0_{x_t}\|^4 \qquad (11a) \\ \le& 4\rho^2 B^4, \qquad (11b)\end{aligned}$$

where (11a) is due to the Lipschitz continuity of $\nabla^2 \hat{f}_{x_t}(\cdot)$, and (11b) follows from the Fact (C.2). Combining all the above inequalities completes the proof. □

Now we proceed to analyze the value of $\sum_{j=1}^{d} h_j(\tilde{s}_{x_t,j}^{\mathcal{K}})$. We split it into the following two cases:

$$\mathcal{S}_1 := \left\{ j : \lambda_j \geq -\frac{\theta}{\eta} \right\} \text{ and } \mathcal{S}_2 := \left\{ j : \lambda_j < -\frac{\theta}{\eta} \right\}.$$

Lemma E.5. *Suppose that Assumption 4.1 and 4.2 hold. Under the parameter setting (3), choose a reasonably small μ such that $\mathbf{E}(\mu) \leq \frac{lB}{2}$ in Algorithm 1. Then, for the tangent space step at iterate x_t, when the "if condition" triggers, we have:*

$$\sum_{j\in\mathcal{S}_1} h_j(\tilde{s}_{x_t,j}^{\mathcal{K}}) \leq -\frac{3\theta}{8\eta} \sum_{j\in\mathcal{S}_1} \sum_{k=0}^{\mathcal{K}-1} |\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 + \frac{4\eta\mathcal{K}}{\theta}\mathbf{E}(\mu)^2 + \frac{16\eta\mathcal{K}\rho^2 B^4}{\theta}.$$

The following proof follows the proof of Lemma 3.2 in (Li & Lin, 2022). We only list the sketch for simplicity.

Proof. For any $k = 0, 1, \cdots, \mathcal{K}-1$ and $j \in \mathcal{S}_1$, as $h_j(\cdot)$ is a one-dimensional quadratic function

$$\begin{aligned}
&h_j(\tilde{s}_{x_t,j}^{k+1}) - h_j(\tilde{s}_{x_t,j}^{k}) \\
=&\langle \nabla h_j(\tilde{s}_{x_t,j}^{k}), \tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k} \rangle + \frac{\lambda_j}{2}|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 \\
=&\langle \nabla h_j(\tilde{s}_{x_t,j}^{k}) - \nabla h_j(\tilde{y}_{x_t,j}^{k}), \tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k} \rangle + \langle \nabla h_j(\tilde{y}_{x_t,j}^{k}), \tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k} \rangle + \frac{\lambda_j}{2}|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 \\
=&\lambda_j \langle \tilde{s}_{x_t,j}^{k} - \tilde{y}_{x_t,j}^{k}, \tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k} \rangle - \frac{1}{\eta}\langle \tilde{s}_{x_t,j}^{k+1} - \tilde{y}_{x_t,j}^{k} + \eta\mathcal{E}_{x_t,j}^{k}, \tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k} \rangle + \frac{\lambda_j}{2}|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 \\
=&\lambda_j \langle \tilde{s}_{x_t,j}^{k} - \tilde{y}_{x_t,j}^{k}, \tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k} \rangle - \frac{1}{\eta}\langle \tilde{s}_{x_t,j}^{k+1} - \tilde{y}_{x_t,j}^{k}, \tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k} \rangle - \langle \mathcal{E}_{x_t,j}^{k}, \tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k} \rangle + \frac{\lambda_j}{2}|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 \\
=&\frac{\lambda_j}{2}(|\tilde{s}_{x_t,j}^{k+1} - \tilde{y}_{x_t,j}^{k}|^2 - |\tilde{s}_{x_t,j}^{k} - \tilde{y}_{x_t,j}^{k}|^2 - |\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2) - \langle \mathcal{E}_{x_t,j}^{k}, \tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k} \rangle \\
&+ \frac{\lambda_j}{2}|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 + \frac{1}{2\eta}(|\tilde{s}_{x_t,j}^{k} - \tilde{y}_{x_t,j}^{k}|^2 - |\tilde{s}_{x_t,j}^{k+1} - \tilde{y}_{x_t,j}^{k}|^2 - |\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2) \\
\leq&\frac{\lambda_j}{2}(|\tilde{s}_{x_t,j}^{k+1} - \tilde{y}_{x_t,j}^{k}|^2 - |\tilde{s}_{x_t,j}^{k} - \tilde{y}_{x_t,j}^{k}|^2) + \frac{2\eta}{\theta}|\mathcal{E}_{x_t,j}^{k}|^2 + \frac{\theta}{8\eta}|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 \\
&+ \frac{1}{2\eta}(|\tilde{s}_{x_t,j}^{k} - \tilde{y}_{x_t,j}^{k}|^2 - |\tilde{s}_{x_t,j}^{k+1} - \tilde{y}_{x_t,j}^{k}|^2 - |\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2),
\end{aligned}$$

where the last inequality is due to $-\langle \mathcal{E}_{x_t,j}^{k}, \tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k} \rangle \leq |\mathcal{E}_{x_t,j}^{k}| \cdot |\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}| \leq \frac{2\eta}{\theta}|\mathcal{E}_{x_t,j}^{k}|^2 + \frac{\theta}{8\eta}|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2$ by applying Young's inequality. Since $l \geq \lambda_j \geq -\frac{\theta}{\eta}$ holds, it follows

$$\begin{aligned}
(-\frac{1}{2\eta} + \frac{\lambda_j}{2})|\tilde{s}_{x_t,j}^{k+1} - \tilde{y}_{x_t,j}^{k}|^2 \leq (-2l + \frac{l}{2})|\tilde{s}_{x_t,j}^{k+1} - \tilde{y}_{x_t,j}^{k}|^2 \leq 0, \\
-\frac{\lambda_j}{2}|\tilde{s}_{x_t,j}^{k} - \tilde{y}_{x_t,j}^{k}|^2 \leq \frac{\theta}{2\eta}|\tilde{s}_{x_t,j}^{k} - \tilde{y}_{x_t,j}^{k}|^2.
\end{aligned}$$

Combing the above inequalities implies that

$$\begin{aligned}
&h_j(\tilde{s}_{x_t,j}^{k+1}) - h_j(\tilde{s}_{x_t,j}^{k}) \\
\leq&\frac{1}{2\eta}(|\tilde{s}_{x_t,j}^{k} - \tilde{y}_{x_t,j}^{k}|^2 - |\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2) + \frac{2\eta}{\theta}|\mathcal{E}_{x_t,j}^{k}|^2 + \frac{\theta}{8\eta}|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 + \frac{\theta}{2\eta}|\tilde{s}_{x_t,j}^{k} - \tilde{y}_{x_t,j}^{k}|^2 \\
=&\frac{(1-\theta)^2}{2\eta}|\tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}|^2 - (\frac{1}{2\eta} - \frac{\theta}{8\eta})|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 + \frac{2\eta}{\theta}|\mathcal{E}_{x_t,j}^{k}|^2 + \frac{\theta(1-\theta)^2}{2\eta}|\tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}|^2 \\
=&\frac{(1+\theta)(1-\theta)^2}{2\eta}|\tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}|^2 - (\frac{1}{2\eta} - \frac{\theta}{8\eta})|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 + \frac{2\eta}{\theta}|\mathcal{E}_{x_t,j}^{k}|^2.
\end{aligned}$$

To proceed, we define the potential function

$$l_{x_t,j}^{k} = h_j(\tilde{s}_{x_t,j}^{k}) + \frac{(1+\theta)(1-\theta)^2}{2\eta}|\tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}|^2,$$

and it gives

$$\begin{aligned}
&l_{x_t,j}^{k+1} - l_{x_t,j}^{k} \\
\leq& -(\frac{1}{2\eta} - \frac{\theta}{8\eta} - \frac{(1+\theta)(1-\theta)^2}{2\eta})|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 + \frac{2\eta}{\theta}|\mathcal{E}_{x_t,j}^{k}|^2 \\
\leq& -\frac{3\theta}{8\eta}|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 + \frac{2\eta}{\theta}|\mathcal{E}_{x_t,j}^{k}|^2.
\end{aligned}$$

Summing over $k = 0, 1, \cdots, \mathcal{K}-1$ and $j \in \mathcal{S}_1$, and using $s_{x_t}^0 = s_{x_t}^{-1}$, we conclude

$$\begin{aligned}
&\sum_{j\in\mathcal{S}_1} h_j(\tilde{s}_{x_t,j}^{\mathcal{K}}) \\
\leq& \sum_{j\in\mathcal{S}_1} l_{x_t,j}^{\mathcal{K}} \\
=& \sum_{j\in\mathcal{S}_1}\sum_{k=0}^{\mathcal{K}-1}(l_{x_t,j}^{k+1} - l_{x_t,j}^{k}) + l_{x_t,j}^{0} \\
=& \sum_{j\in\mathcal{S}_1}\sum_{k=0}^{\mathcal{K}-1}(l_{x_t,j}^{k+1} - l_{x_t,j}^{k}) \\
\leq& -\frac{3\theta}{8\eta}\sum_{j\in\mathcal{S}_1}\sum_{k=0}^{\mathcal{K}-1}|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 + \frac{2\eta}{\theta}\sum_{j\in\mathcal{S}_1}\sum_{k=0}^{\mathcal{K}-1}|\mathcal{E}_{x_t,j}^{k}|^2 \\
\leq& -\frac{3\theta}{8\eta}\sum_{j\in\mathcal{S}_1}\sum_{k=0}^{\mathcal{K}-1}|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 + \frac{4\eta\mathcal{K}}{\theta}\mathbf{E}(\mu)^2 + \frac{16\eta\mathcal{K}\rho^2 B^4}{\theta},
\end{aligned}$$

where the last inequality is due to Lemma (E.4). □

Lemma E.6. *Suppose that Assumption 4.1 and 4.2 hold. Under the parameter setting (3), choose a reasonably small μ such that $\mathbf{E}(\mu) \leq \frac{lB}{2}$ in Algorithm 1. Then, for the tangent space step at iterate x_t, when the "if condition" triggers, we have:*

$$\sum_{j\in\mathcal{S}_2} h_j(\tilde{s}_{x_t,j}^{\mathcal{K}}) \leq -\frac{\theta}{2\eta}\sum_{j\in\mathcal{S}_2}\sum_{k=0}^{\mathcal{K}-1}|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 + \eta\mathcal{K}\mathbf{E}(\mu)^2 + \frac{\mathcal{K}}{2B^{\frac{3}{2}}}\mathbf{E}(\mu)^2 + \frac{\mathcal{K}B^{\frac{7}{2}}}{2} + \frac{\eta\mathcal{K}\mathbf{E}(\mu)^2}{\theta} + \frac{4\eta\mathcal{K}\rho^2 B^4}{\theta}.$$

Proof. Let $v_{x_t,j} = \tilde{s}_{x_t,j}^0 - \frac{1}{\lambda_j}\tilde{\nabla}_j\hat{f}_{x_t}(s_{x_t}^0)$, the one-dimensional quadratic function $h_j(\cdot)$ can be rewritten as

$$h_j(z) = \frac{\lambda_j}{2}(z - v_{x_t,j})^2 - \frac{1}{2\lambda_j}|\tilde{\nabla}_j\hat{f}_{x_t}(s_{x_t}^0)|^2.$$

Consequently, for any $k = 0, 1, \cdots, \mathcal{K}-1$ and $j \in \mathcal{S}_2$, we have

$$\begin{aligned}
&h_j(\tilde{s}_{x_t,j}^{k+1}) - h_j(\tilde{s}_{x_t,j}^{k}) \\
=&\frac{\lambda_j}{2}|\tilde{s}_{j}^{k+1} - v_{x_t,j}|^2 - \frac{\lambda_j}{2}|\tilde{s}_{x_t,j}^{k} - v_{x_t,j}|^2 \\
=&\frac{\lambda_j}{2}|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 + \lambda_j\langle\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}, \tilde{s}_{x_t,j}^{k} - v_{x_t,j}\rangle \\
\leq& -\frac{\theta}{2\eta}|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 + \lambda_j\langle\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}, \tilde{s}_{x_t,j}^{k} - v_{x_t,j}\rangle.
\end{aligned} \tag{16}$$

Recall the update in (9), it holds that

$$\begin{aligned}
&\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k} \\
=&\tilde{y}_{x_t,j}^{k} - \eta\nabla h_j(\tilde{y}_{x_t,j}^{k}) - \eta\mathcal{E}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k} \\
=&(1-\theta)(\tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}) - \eta\nabla h_j(\tilde{y}_{x_t,j}^{k}) - \eta\mathcal{E}_{x_t,j}^{k} \\
=&(1-\theta)(\tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}) - \eta\lambda_j(\tilde{y}_{x_t,j}^{k} - v_j) - \eta\mathcal{E}_{x_t,j}^{k} \\
=&(1-\theta)(\tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}) - \eta\lambda_j(\tilde{s}_{x_t,j}^{k} - v_{x_t,j} + (1-\theta)(\tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1})) - \eta\mathcal{E}_{x_t,j}^{k}.
\end{aligned}$$

Substituting the above equality into the term $\langle \tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}, \tilde{s}_{x_t,j}^{k} - v_{x_t,j}\rangle$ gives

$$\begin{aligned}
&\langle \tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}, \tilde{s}_{x_t,j}^{k} - v_{x_t,j}\rangle \\
=&(1-\theta)\langle \tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}, \tilde{s}_{x_t,j}^{k} - v_{x_t,j}\rangle - \eta\lambda_j|\tilde{s}_{x_t,j}^{k} - v_{x_t,j}|^2 \\
&\underbrace{-\eta\lambda_j(1-\theta)\langle \tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}, \tilde{s}_{x_t,j}^{k} - v_{x_t,j}\rangle - \eta\langle \mathcal{E}_{x_t,j}^{k}, \tilde{s}_{x_t,j}^{k} - v_{x_t,j}\rangle}_{\clubsuit}.
\end{aligned} \tag{18}$$

We first provide a lower bound for the term ♣ in (18):

$$\begin{aligned}
&-\eta\lambda_j(1-\theta)\langle \tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}, \tilde{s}_{x_t,j}^{k} - v_{x_t,j}\rangle - \eta\langle \mathcal{E}_{x_t,j}^{k}, \tilde{s}_{x_t,j}^{k} - v_{x_t,j}\rangle \\
=&\frac{\eta\lambda_j(1-\theta)}{2}(|\tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}|^2 + |v_{x_t,j} - \tilde{s}_{x_t,j}^{k}|^2 - |\tilde{s}_{x_t,j}^{k-1} - v_{x_t,j}|^2) - \eta\langle \mathcal{E}_{x_t,j}^{k}, \tilde{s}_{x_t,j}^{k} - v_{x_t,j}\rangle \\
\geq&\frac{\eta\lambda_j(1-\theta)}{2}(|\tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}|^2 + |\tilde{s}_{x_t,j}^{k} - v_{x_t,j}|^2) + \frac{\eta}{2\lambda_j(1+\theta)}|\mathcal{E}_{x_t,j}^{k}|^2 + \frac{\eta\lambda_j(1+\theta)}{2}|\tilde{s}_{x_t,j}^{k} - v_{x_t,j}|^2 \qquad (19a) \\
=&\frac{\eta\lambda_j(1-\theta)}{2}|\tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}|^2 + \eta\lambda_j|\tilde{s}_{x_t,j}^{k} - v_{x_t,j}|^2 + \frac{\eta}{2\lambda_j(1+\theta)}|\mathcal{E}_{x_t,j}^{k}|^2,
\end{aligned}$$

where (19a) holds because $\lambda_j < 0$ and $-\langle \mathcal{E}_{x_t,j}^{k}, \tilde{s}_{x_t,j}^{k} - v_{x_t,j}\rangle \geq \frac{1}{2\lambda_j(1+\theta)}|\mathcal{E}_{x_t,j}^{k}|^2 + \frac{\lambda_j(1+\theta)}{2}|\tilde{s}_{x_t,j}^{k} - v_{x_t,j}|^2$ due to the Young's inequality. Plugging the above inequality back into (18) gives

$$\begin{aligned}
&\langle \tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}, \tilde{s}_{x_t,j}^{k} - v_{x_t,j}\rangle \\
\geq&(1-\theta)\langle \tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}, \tilde{s}_{x_t,j}^{k} - v_{x_t,j}\rangle + \frac{\eta\lambda_j(1-\theta)}{2}|\tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}|^2 + \frac{\eta}{2\lambda_j(1+\theta)}|\mathcal{E}_{x_t,j}^{k}|^2 \\
=&(1-\theta)\langle \tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}, \tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}\rangle + (1-\theta)\langle \tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}, \tilde{s}_{x_t,j}^{k-1} - v_{x_t,j}\rangle \\
&+ \frac{\eta\lambda_j(1-\theta)}{2}|\tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}|^2 + \frac{\eta}{2\lambda_j(1+\theta)}|\mathcal{E}_{x_t,j}^{k}|^2 \\
=&(1+\frac{\eta\lambda_j}{2})(1-\theta)|\tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}|^2 + (1-\theta)\langle \tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}, \tilde{s}_{x_t,j}^{k-1} - v_{x_t,j}\rangle + \frac{\eta}{2\lambda_j(1+\theta)}|\mathcal{E}_{x_t,j}^{k}|^2 \\
\geq&(1-\theta)\langle \tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}, \tilde{s}_{x_t,j}^{k-1} - v_{x_t,j}\rangle + \frac{\eta}{2\lambda_j}|\mathcal{E}_{x_t,j}^{k}|^2,
\end{aligned}$$

where the last inequality comes from $(1+\frac{\eta\lambda_j}{2})(1-\theta) \geq (1-\frac{\eta l}{2})(1-\theta) = (1-\frac{1}{8})(1-\theta) > 0$. Hence, it implies that

$$
\begin{aligned}
&\langle \tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}, \tilde{s}_{x_t,j}^{k} - v_{x_t,j}\rangle \\
\geq&(1-\theta)\langle \tilde{s}_{x_t,j}^{k} - \tilde{s}_{x_t,j}^{k-1}, \tilde{s}_{x_t,j}^{k-1} - v_{x_t,j}\rangle + \frac{\eta}{2\lambda_j}|\mathcal{E}_{x_t,j}^{k}|^2 \\
\geq&(1-\theta)^k\langle \tilde{s}_{x_t,j}^{1} - \tilde{s}_{x_t,j}^{0}, \tilde{s}_{x_t,j}^{0} - v_{x_t,j}\rangle + \frac{\eta}{2\lambda_j}\sum_{t=1}^{k}(1-\theta)^{k-t}|\mathcal{E}_{x_t,j}^{t}|^2 \\
=&(1-\theta)^k\eta\langle -\nabla h_j(\tilde{y}_{x_t,j}^{0}) - \mathcal{E}_{x_t,j}^{0}, \tilde{y}_{x_t,j}^{0} - v_{x_t,j}\rangle + \frac{\eta}{2\lambda_j}\sum_{t=1}^{k}(1-\theta)^{k-t}|\mathcal{E}_{x_t,j}^{t}|^2 \qquad (20a)\\
=&(1-\theta)^k\eta\langle \lambda_j(v_{x_t,j} - \tilde{y}_{x_t,j}^{0}) - \mathcal{E}_{x_t,j}^{0}, \tilde{y}_{x_t,j}^{0} - v_{x_t,j}\rangle + \frac{\eta}{2\lambda_j}\sum_{t=1}^{k}(1-\theta)^{k-t}|\mathcal{E}_{x_t,j}^{t}|^2 \\
=&-(1-\theta)^k\eta\lambda_j|v_{x_t,j} - \tilde{y}_{x_t,j}^{0}|^2 + (1-\theta)^k\eta\langle \mathcal{E}_{x_t,j}^{0}, v_{x_t,j} - \tilde{y}_{x_t,j}^{0}\rangle + \frac{\eta}{2\lambda_j}\sum_{t=1}^{k}(1-\theta)^{k-t}|\mathcal{E}_{x_t,j}^{t}|^2 \\
\geq&(1-\theta)^k \underbrace{\eta\langle \mathcal{E}_{x_t,j}^{0}, v_{x_t,j} - \tilde{y}_{x_t,j}^{0}\rangle}_{\spadesuit} + \frac{\eta}{2\lambda_j}\sum_{t=1}^{k}(1-\theta)^{k-t}|\mathcal{E}_{x_t,j}^{t}|^2, \qquad (20b)
\end{aligned}
$$

where (20a) follows from the update in (9), and (20b) is implied by $\lambda_j < 0$. Recall $v_{x_t,j} - \tilde{y}_{x_t,j}^{0} = -\frac{1}{\lambda_j}\tilde{\nabla}_j\hat{f}_{x_t}(s_{x_t}^{0})$, and thus the term $\spadesuit$ can be lower bounded as follows:

$$
\begin{aligned}
&\eta\langle \mathcal{E}_{x_t,j}^{0}, v_{x_t,j} - \tilde{y}_{x_t,j}^{0}\rangle \\
=&-\frac{\eta}{\lambda_j}\langle \mathcal{E}_{x_t,j}^{0}, \tilde{\nabla}_j\hat{f}_{x_t}(s_{x_t}^{0})\rangle \\
=&\frac{\eta}{\lambda_j}\langle \mathcal{E}_{x_t,j}^{0}, \mathcal{E}_{x_t,j}^{0}\rangle - \frac{\eta}{\lambda_j}\langle \mathcal{E}_{x_t,j}^{0}, \tilde{g}_{x_t,j}(y_{x_t}^{0};\mu)\rangle \qquad (21a)\\
=&\frac{\eta}{\lambda_j}\langle \mathcal{E}_{x_t,j}^{0}, \mathcal{E}_{x_t,j}^{0}\rangle + \frac{1}{\lambda_j}\langle \mathcal{E}_{x_t,j}^{0}, \tilde{s}_{x_t,j}^{1} - \tilde{y}_{x_t,j}^{0}\rangle \qquad (21b)\\
\geq&\frac{\eta}{\lambda_j}|\mathcal{E}_{x_t,j}^{0}|^2 + \frac{1}{2\lambda_j B^{\frac{3}{2}}}|\mathcal{E}_{x_t,j}^{0}|^2 + \frac{B^{\frac{3}{2}}}{2\lambda_j}|\tilde{s}_{x_t,j}^{1} - \tilde{y}_{x_t,j}^{0}|^2, \qquad (21c)
\end{aligned}
$$

where (21a) is due to $\mathcal{E}_{x_t,j}^{0} = \tilde{g}_{x_t,j}(y_{x_t}^{0};\mu) - \nabla h_j(\tilde{y}_{x_t,j}^{0})$ and $\nabla h_j(\tilde{y}_{x_t,j}^{0}) = \tilde{\nabla}_j\hat{f}_{x_t}(s_{x_t}^{0})$, (21b) follows from $\tilde{s}_{x_t,j}^{1} = \tilde{y}_{x_t,j}^{0} - \tilde{g}_{x_t,j}(y_{x_t}^{0};\mu)$, and (21c) holds because $\lambda_j < 0$ and Young's inequality. Putting all the above inequalities together gives

$$
\begin{aligned}
&h_j(\tilde{s}_{x_t,j}^{k+1}) - h_j(\tilde{s}_{x_t,j}^{k}) \\
\leq&-\frac{\theta}{2\eta}|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 + (1-\theta)^k\left(\eta|\mathcal{E}_{x_t,j}^{0}|^2 + \frac{1}{2B^{\frac{3}{2}}}|\mathcal{E}_{x_t,j}^{0}|^2 + \frac{B^{\frac{3}{2}}}{2}|\tilde{s}_{x_t,j}^{1} - \tilde{y}_{x_t,j}^{0}|^2\right) + \frac{\eta}{2}\sum_{t=1}^{k}(1-\theta)^{k-t}|\mathcal{E}_{x_t,j}^{t}|^2 \\
\leq&-\frac{\theta}{2\eta}|\tilde{s}_{x_t,j}^{k+1} - \tilde{s}_{x_t,j}^{k}|^2 + \eta|\mathcal{E}_{x_t,j}^{0}|^2 + \frac{1}{2B^{\frac{3}{2}}}|\mathcal{E}_{x_t,j}^{0}|^2 + \frac{B^{\frac{3}{2}}}{2}|\tilde{s}_{x_t,j}^{1} - \tilde{y}_{x_t,j}^{0}|^2 + \frac{\eta}{2}\sum_{t=1}^{k}(1-\theta)^{k-t}|\mathcal{E}_{x_t,j}^{t}|^2.
\end{aligned}
$$

Summing over $j \in \mathcal{S}_2$ implies that

$$\begin{aligned}
&\sum_{j\in\mathcal{S}_2} h_j(\tilde{s}^{k+1}_{x_t,j}) - h_j(\tilde{s}^{k}_{x_t,j}) \\
\leq& -\frac{\theta}{2\eta}\sum_{j\in\mathcal{S}_2}|\tilde{s}^{k+1}_{x_t,j}-\tilde{s}^{k}_{x_t,j}|^2 + \eta\sum_{j\in\mathcal{S}_2}|\mathcal{E}^0_{x_t,j}|^2 + \frac{1}{2B^{\frac{3}{2}}}\sum_{j\in\mathcal{S}_2}|\mathcal{E}^0_{x_t,j}|^2 \\
&+ \frac{B^{\frac{3}{2}}}{2}\sum_{j\in\mathcal{S}_2}|\tilde{s}^1_{x_t,j}-\tilde{y}^0_{x_t,j}|^2 + \frac{\eta}{2}\sum_{t=1}^{k}(1-\theta)^{k-t}\sum_{j\in\mathcal{S}_2}|\mathcal{E}^t_{x_t,j}|^2 \\
\leq& -\frac{\theta}{2\eta}\sum_{j\in\mathcal{S}_2}|\tilde{s}^{k+1}_{x_t,j}-\tilde{s}^{k}_{x_t,j}|^2 + \eta\mathbf{E}(\mu)^2 + \frac{\mathbf{E}(\mu)^2}{2B^{\frac{3}{2}}} \\
&+ \frac{B^{\frac{7}{2}}}{2} + \frac{\eta}{2}\sum_{t=1}^{k}(1-\theta)^{k-t}\left(2\mathbf{E}(\mu)^2 + 8\rho^2B^4\right), \qquad (22a) \\
\leq& -\frac{\theta}{2\eta}\sum_{j\in\mathcal{S}_2}|\tilde{s}^{k+1}_{x_t,j}-\tilde{s}^{k}_{x_t,j}|^2 + \eta\mathbf{E}(\mu)^2 + \frac{\mathbf{E}(\mu)^2}{2B^{\frac{3}{2}}} + \frac{B^{\frac{7}{2}}}{2} + \frac{\eta\mathbf{E}(\mu)^2}{\theta} + \frac{4\eta\rho^2B^4}{\theta},
\end{aligned}$$

where we apply $\sum_{j\in\mathcal{S}_2}|\tilde{s}^1_{x_t,j}-\tilde{y}^0_{x_t,j}|^2 \leq \|s^1_{x_t}-y^0_{x_t}\|^2 \leq B^2$ and the result of Lemma E.4 in (22a). Finally, we conclude

$$\begin{aligned}
&\sum_{j\in\mathcal{S}_2} h_j(\tilde{s}^{\mathcal{K}}_{x_t,j}) \\
=&\sum_{j\in\mathcal{S}_2}\sum_{k=0}^{\mathcal{K}-1} h_j(\tilde{s}^{k+1}_{x_t,j}) - h_j(\tilde{s}^{k}_{x_t,j}) + h_j(\tilde{s}^{0}_{x_t,j}) \\
\leq& -\frac{\theta}{2\eta}\sum_{j\in\mathcal{S}_2}\sum_{k=0}^{\mathcal{K}-1}|\tilde{s}^{k+1}_{x_t,j}-\tilde{s}^{k}_{x_t,j}|^2 + \eta\mathcal{K}\mathbf{E}(\mu)^2 + \frac{\mathcal{K}}{2B^{\frac{3}{2}}}\mathbf{E}(\mu)^2 + \frac{\mathcal{K}B^{\frac{7}{2}}}{2} + \frac{\eta\mathcal{K}\mathbf{E}(\mu)^2}{\theta} + \frac{4\eta\mathcal{K}\rho^2B^4}{\theta}.
\end{aligned}$$

□

Corollary E.1. *Suppose that Assumption 4.1 and 4.2 hold. Under the parameter setting (3), choose a reasonably small μ such that $\mathbf{E}(\mu) \leq \frac{lB}{2}$ in Algorithm 1. Then, for the tangent space step at iterate x_t, when the "if condition" triggers, we have:*

$$\hat{f}_{x_t}(s^{\mathcal{K}}_{x_t}) \leq \hat{f}_{x_t}(s^0_{x_t}) - \frac{3\theta B^2}{8\eta K} + \eta K\mathbf{E}(\mu)^2 + \frac{K}{2B^{\frac{3}{2}}}\mathbf{E}(\mu)^2 + \frac{KB^{\frac{7}{2}}}{2} + \frac{5\eta K\mathbf{E}(\mu)^2}{\theta} + \frac{20\eta K\rho^2B^4}{\theta} + \frac{32\rho B^3}{3}.$$

Proof. Putting Lemma E.5 and Lemma E.6 together, we obtain

$$\begin{aligned}
&\sum_{j=1}^{d} h_j(\tilde{s}^{\mathcal{K}}_{x_t,j}) \\
\leq& -\frac{3\theta}{8\eta}\sum_{k=0}^{\mathcal{K}-1}\|\tilde{s}^{k+1}_{x_t}-\tilde{s}^{k}_{x_t}\|^2 + \eta\mathcal{K}\mathbf{E}(\mu)^2 + \frac{\mathcal{K}}{2B^{\frac{3}{2}}}\mathbf{E}(\mu)^2 + \frac{\mathcal{K}B^{\frac{7}{2}}}{2} + \frac{5\eta\mathcal{K}\mathbf{E}(\mu)^2}{\theta} + \frac{20\eta\mathcal{K}\rho^2B^4}{\theta} \\
\leq& -\frac{3\theta B^2}{8\eta\mathcal{K}} + \eta\mathcal{K}\mathbf{E}(\mu)^2 + \frac{\mathcal{K}}{2B^{\frac{3}{2}}}\mathbf{E}(\mu)^2 + \frac{\mathcal{K}B^{\frac{7}{2}}}{2} + \frac{5\eta\mathcal{K}\mathbf{E}(\mu)^2}{\theta} + \frac{20\eta\mathcal{K}\rho^2B^4}{\theta} \\
\leq& -\frac{3\theta B^2}{8\eta K} + \eta K\mathbf{E}(\mu)^2 + \frac{K}{2B^{\frac{3}{2}}}\mathbf{E}(\mu)^2 + \frac{KB^{\frac{7}{2}}}{2} + \frac{5\eta K\mathbf{E}(\mu)^2}{\theta} + \frac{20\eta K\rho^2B^4}{\theta},
\end{aligned}$$

combining with Lemma E.3 completes the proof. □

E.3. Tangent space step: small Riemannian gradient

For the scenario that the "if condition" does not trigger in the tangent space step, we establish that the tangent space step outputs a satisfactory point with a small Riemannian gradient.

Lemma E.7. *Suppose that Assumption 4.1, 4.2 and 4.3 hold. Under the parameter setting (3), choose a reasonably small μ such that $\mathbf{E}(\mu) \leq \frac{lB}{2}$ in Algorithm 1.Then, for the tangent space step at iterate x_t, when the "if condition" does not trigger, we have:*

$$\|\operatorname{grad} f(x_{t+1})\| \leq \frac{1}{\sigma_{\min}} \cdot (2\rho B^2 + \frac{2\sqrt{2}B}{K^2\eta} + \frac{2\theta B}{K\eta} + \big(2\mathbf{E}(\mu)^2 + 8\rho^2 B^4\big)^{\frac{1}{2}}).$$

Proof. According to Subroutine 2, when the "if condition" does not trigger, the tangent space step outputs $x_{t+1} = \operatorname{Retr}_{x_t}(y^*_{x_t})$, and thus, it holds that:

$$\operatorname{grad} f(x_{t+1}) = \operatorname{grad} f(\operatorname{Retr}_{x_t}(y^*_{x_t})) = (T^*_{x_t, y^*_{x_t}})^{-1} \nabla \hat{f}_{x_t}(y^*_{x_t}).$$

Therefore, it is sufficient to upper bound the term $\|\nabla \hat{f}_{x_t}(y^*_{x_t})\|$. By applying Minkowski's inequality, we have

$$\|\nabla \hat{f}_{x_t}(y^*_{x_t})\| = \left(\sum_{j=1}^{d} |\tilde{\nabla}_j \hat{f}_{x_t}(y^*_{x_t})|^2\right)^{\frac{1}{2}} \leq \left(\sum_{j=1}^{d} |\tilde{\nabla}_j \hat{f}_{x_t}(y^*_{x_t}) - \nabla h_j(\tilde{y}^*_{x_t,j})|^2\right)^{\frac{1}{2}} + \left(\sum_{j=1}^{d} |\nabla h_j(\tilde{y}^*_{x_t,j})|^2\right)^{\frac{1}{2}}.$$

Note that $\nabla h_j(\tilde{y}^*_{x_t,j}) = \tilde{\nabla}_j \hat{f}_{x_t}(s^0_{x_t}) + \lambda_j(\tilde{y}^*_{x_t,j} - \tilde{s}_{x_t,j})$, using the same argument in the proof of Lemma E.4 gives

$$\begin{aligned}
&\left(\sum_{j=1}^{d} |\tilde{\nabla}_j \hat{f}_{x_t}(y^*_{x_t}) - \nabla h_j(\tilde{y}^*_{x_t,j})|^2\right)^{\frac{1}{2}} \\
=&\left(\sum_{j=1}^{d} |\tilde{\nabla}_j \hat{f}_{x_t}(y^*_{x_t}) - \tilde{\nabla}_j \hat{f}_{x_t}(s^0_{x_t}) - \lambda_j(\tilde{y}^*_{x_t,j} - \tilde{s}_{x_t,j})|^2\right)^{\frac{1}{2}} \\
=&\|\nabla \hat{f}_{x_t}(y^*_{x_t}) - \nabla \hat{f}_{x_t}(s^0_{x_t}) - \nabla^2 \hat{f}_{x_t}(s^0_{x_t})(y^*_{x_t} - s^0_{x_t})\| \\
\leq&\frac{\rho}{2}\|y^*_{x_t} - s^0_{x_t}\|^2 \qquad (24a) \\
\leq&2\rho B^2, \qquad (24b)
\end{aligned}$$

where (24a) is due to the Lipschitz continuity of $\nabla^2 \hat{f}_{x_t}(\cdot)$ and the (24b) comes from the following result

$$\|y^*_{x_t} - s^0_{x_t}\| \leq \frac{1}{K_0+1} \sum_{k=0}^{K_0} \|y^k_{x_t} - s^0_{x_t}\| \leq 2B.$$

For the term $\left(\sum_{j=1}^{d} |\nabla h_j(\tilde{y}^*_{x_t,j})|^2\right)^{\frac{1}{2}}$, note that $\tilde{y}^*_{x_t,j} = \frac{1}{K_0+1}\sum_{k=0}^{K_0} \tilde{y}^k_{x_t,j}$ and $\nabla h_j(\cdot)$ is a one-dimensional linear function, it equivalents as

$$\left(\sum_{j=1}^{d} |\nabla h_j(\tilde{y}^*_{x_t,j})|^2\right)^{\frac{1}{2}} = \left(\sum_{j=1}^{d} \left|\nabla h_j\left(\frac{1}{K_0+1}\sum_{k=0}^{K_0} \tilde{y}^k_{x_t,j}\right)\right|^2\right)^{\frac{1}{2}} = \frac{1}{K_0+1}\left(\sum_{j=1}^{d} \left|\sum_{k=0}^{K_0} \nabla h_j(\tilde{y}^k_{x_t,j})\right|^2\right)^{\frac{1}{2}}.$$

Recall the update (9), we have

$$\begin{aligned}
\eta \nabla h_j(\tilde{y}^k_{x_t,j}) &= \tilde{s}^{k+1}_{x_t,j} - \tilde{y}^k_{x_t,j} + \eta \mathcal{E}^k_{x_t,j} \\
&= \tilde{s}^{k+1}_{x_t,j} - \tilde{s}^k_{x_t,j} - (1-\theta)(\tilde{s}^k_{x_t,j} - \tilde{s}^{k-1}_{x_t,j}) + \eta \mathcal{E}^k_{x_t,j}.
\end{aligned}$$

Consequently, it holds that

$$
\begin{aligned}
&\left(\sum_{j=1}^{d}|\nabla h_j(\tilde{y}^*_{x_t,j})|^2\right)^{\frac{1}{2}}\\
=&\frac{1}{(K_0+1)\eta}\left(\sum_{j=1}^{d}\left|s^{K_0}_{x_t,j}-\tilde{s}^0_{x_t,j}+\theta(\tilde{s}^{K_0}_{x_t,j}-\tilde{s}^0_{x_t,j})+\eta\sum_{k=0}^{K_0}\mathcal{E}^k_{x_t,j}\right|^2\right)^{\frac{1}{2}}\\
\le&\frac{1}{(K_0+1)\eta}\left(\sum_{j=1}^{d}\left|\tilde{s}^{K_0+1}_{x_t,j}-\tilde{s}^{K_0}_{x_t,j}\right|^2\right)^{\frac{1}{2}}+\frac{\theta}{(K_0+1)\eta}\left(\sum_{j=1}^{d}\left|\tilde{s}^{K_0}_{x_t,j}-\tilde{s}^0_{x_t,j}\right|^2\right)^{\frac{1}{2}}+\frac{1}{K_0+1}\left(\sum_{j=1}^{d}\left|\sum_{k=0}^{K_0}\mathcal{E}^k_{x_t,j}\right|^2\right)^{\frac{1}{2}}\\
\le&\frac{1}{(K_0+1)\eta}\left\|s^{K_0+1}_{x_t}-s^{K_0}_{x_t}\right\|+\frac{\theta}{(K_0+1)\eta}\left\|s^{K_0}_{x_t}-s^0_{x_t}\right\|+\frac{1}{\sqrt{K_0+1}}\left(\sum_{k=0}^{K_0}\sum_{j=1}^{d}|\mathcal{E}^k_{x_t,j}|^2\right)^{\frac{1}{2}} \qquad (25a)\\
\le&\frac{2}{K\eta}\left\|s^{K_0+1}_{x_t}-s^{K_0}_{x_t}\right\|+\frac{2\theta B}{K\eta}+\left(2\mathbf{E}(\mu)^2+8\rho^2B^4\right)^{\frac{1}{2}} \qquad (25b)
\end{aligned}
$$

where (25a) holds because $\left|\sum_{k=0}^{K_0}\mathcal{E}^k_{x_t,j}\right|^2\le(K_0+1)\sum_{k=0}^{K_0}|\mathcal{E}^k_{x_t,j}|^2$, and (25b) follows from the $K_0+1\ge\frac{K}{2}$, $\theta\le1$, Lemma C.2 and E.4. Recall the definition of K_0, we obtain

$$
\begin{aligned}
&\left\|s^{K_0+1}_{x_t}-s^{K_0}_{x_t}\right\|^2\\
\le&\frac{1}{K-\lfloor K/2\rfloor}\sum_{k=\lfloor K/2\rfloor}^{K-1}\|s^{k+1}_{x_t}-s^k_{x_t}\|^2\\
\le&\frac{1}{K-\lfloor K/2\rfloor}\sum_{k=0}^{K-1}\|s^{k+1}_{x_t}-s^k_{x_t}\|^2\\
\le&\frac{2B^2}{K^2}.
\end{aligned}
$$

Therefore, combining all the above inequalities gives that

$$
\begin{aligned}
&\|\operatorname{grad}f(x_{t+1})\|\\
\le&\|(T^*_{x_t,y^*_{x_t}})^{-1}\|\cdot(2\rho B^2+\frac{2\sqrt{2}B}{K^2\eta}+\frac{2B}{K\eta}+\left(2\mathbf{E}(\mu)^2+8\rho^2B^4\right)^{\frac{1}{2}})\\
\le&\frac{1}{\sigma_{\min}}\cdot(2\rho B^2+\frac{2\sqrt{2}B}{K^2\eta}+\frac{2\theta B}{K\eta}+\left(2\mathbf{E}(\mu)^2+8\rho^2B^4\right)^{\frac{1}{2}}),
\end{aligned}
$$

where the last inequality comes from Assumption 4.3. □

E.4. Proof of Theorem 4.1

Theorem E.1 (Theorem 4.1 restated)**.** *Suppose that Assumption 4.1, 4.2 and 4.3 hold. Set the parameters in Algorithm 1 as follows*

$$
\eta=\frac{1}{4l},\ B=\frac{1}{8}\sqrt{\frac{\epsilon}{\rho}},\ \theta=\frac{\rho^{\frac{7}{4}}\epsilon^{\frac{1}{4}}}{l},\ r=0,\ K=\frac{\rho^{\frac{5}{4}}}{4\epsilon^{\frac{1}{4}}}. \qquad (26)
$$

For any $x_0\in\mathcal{M}$ and sufficiently small $\epsilon>0$, choose $\mu=\mathcal{O}\left(\frac{\epsilon^{1/4}}{d^{1/4}}\right)$ in Lines 3 of Algorithm 1, and $\mu=\mathcal{O}\left(\frac{\epsilon^{5/8}}{d^{1/4}}\right)$ in Line 5 of Subroutine 2. Then Algorithm 1 with Option I outputs an ϵ-approximate first-order stationary point. The total number of function value evaluations is no more than

$$
\mathcal{O}\left(\frac{(f(x_0)-f_{\text{low}})d}{\epsilon^{\frac{7}{4}}}\right).
$$

Proof. Given an iterate x_t, in the large estimator scenario where $\|g_{x_t}(0;\mu)\| \geq lB$, we choose $\mu = \mathcal{O}\left(\frac{\epsilon^{1/4}}{d^{1/4}}\right)$; combining this with Lemma D.1 results in $\mathbf{E}(\mu) \leq \mathcal{O}\left(\sqrt{\epsilon}\right)$. Consequently, Lemma E.1 yields:

$$f(x_{t+1}) - f(x_t) \leq -\min\left\{\frac{lB^2}{16}, lb^2\right\} = -\frac{l\epsilon}{1024\rho}.$$

For the small estimator scenario where $\|g_{x_t}(0;\mu)\| \leq lB$, the Algorithm 1 switches to Subroutine 2, i.e. the tangent space step. Note that $r = 0$, it implies

$$f(x_t) = \hat{f}_{x_t}(s^0_{x_t}).$$

Moreover, we choose $\mu = \mathcal{O}\left(\frac{\epsilon^{5/8}}{d^{1/4}}\right)$ in Subroutine 2, resulting in $\mathbf{E}(\mu) \leq \mathcal{O}\left(\epsilon^{5/4}\right)$. Thus, when the "if condition" triggers, from Corollary E.1, we have

$$\begin{aligned}
&f(x_{t+1}) - f(x_t)\\
=&\hat{f}_{x_t}(s^{\mathcal{K}}_{x_t}) - \hat{f}_{x_t}(s^0_{x_t})\\
\leq& -\frac{3\theta B^2}{8\eta K} + \eta K\mathbf{E}(\mu)^2 + \frac{K}{2B^{\frac{3}{2}}}\mathbf{E}(\mu)^2 + \frac{KB^{\frac{7}{2}}}{2} + \frac{5\eta K\mathbf{E}(\mu)^2}{\theta} + \frac{20\eta K\rho^2B^4}{\theta} + \frac{32\rho B^3}{3}\\
\leq& -\frac{\epsilon^{\frac{3}{2}}}{24\sqrt{\rho}} + \eta K\mathbf{E}(\mu)^2 + \frac{K}{2B^{\frac{3}{2}}}\mathbf{E}(\mu)^2 + \frac{5\eta K\mathbf{E}(\mu)^2}{\theta}\\
\leq& -\frac{\epsilon^{\frac{3}{2}}}{24\sqrt{\rho}} + \mathcal{O}\left(\epsilon^{\frac{9}{4}}\right) + \mathcal{O}\left(\epsilon^{\frac{3}{2}}\right) + \mathcal{O}\left(\epsilon^2\right)\\
\leq& -\frac{\epsilon^{\frac{3}{2}}}{32\sqrt{\rho}}.
\end{aligned}$$

When the "if condition" does not trigger, Lemma E.7 tells

$$\begin{aligned}
&\|\operatorname{grad} f(x_{t+1})\|\\
\leq&\frac{1}{\sigma_{\min}}\cdot\left(2\rho B^2 + \frac{2\sqrt{2}B}{K^2\eta} + \frac{2\theta B}{K\eta} + \left(2\mathbf{E}(\mu)^2 + 8\rho^2B^4\right)^{\frac{1}{2}}\right)\\
\leq&\frac{1}{\sigma_{\min}}\cdot\left(2\rho B^2 + \frac{2\sqrt{2}B}{K^2\eta} + \frac{2\theta B}{K\eta} + 4\rho B^2\right) &\text{(28a)}\\
\leq&\mathcal{O}(\epsilon), &\text{(28b)}
\end{aligned}$$

where (28a) holds because $\mathbf{E}(\mu)^2 \leq \mathcal{O}\left(\epsilon^{5/2}\right)$ and $8\rho^2B^4 = \mathcal{O}\left(\epsilon^2\right)$, and (28b) comes from the parameter setting (26). Therefore, at each iteration t, once $\|g_{x_t}(0;\mu)\| \geq lB$ holds or the "if condition" does not trigger in the tangent space step, we observe the following function value decrease:

$$f(x_{t+1}) - f(x_t) \leq -\min\left\{\frac{l\epsilon}{1024\rho}, \frac{\epsilon^{\frac{3}{2}}}{32\sqrt{\rho}}\right\} = -\frac{\epsilon^{\frac{3}{2}}}{32\sqrt{\rho}}.$$

Otherwise, if the 'if condition' does not trigger, x_{t+1} is already an ϵ-approximate first-order stationary point. As the tangent space step requires at most $K = \mathcal{O}\left(\epsilon^{-1/4}\right)$ iterations, and each iterate needs $2d$ function value evaluations to construct the zeroth-order estimator, the total number of function value evaluations must be less than

$$O\left(\frac{(f(x_0) - f_{\text{low}})d}{\epsilon^{\frac{7}{4}}}\right).$$

□

E.5. Proof of Theorem 4.2

Theorem E.2 (Theorem 4.2 restated)**.** *Suppose that Assumption 4.1, 4.2 and 4.3 hold. Set the parameters in Algorithm 1 as follows*

$$\eta = \frac{1}{4l},\ \chi = \mathcal{O}\left(\log\frac{d}{\delta\epsilon}\right) \geq 1,\ B = \frac{1}{8\chi^2}\sqrt{\frac{\epsilon}{\rho}},\ \theta = \frac{\rho^{\frac{7}{4}}\epsilon^{\frac{1}{4}}}{l},\ r = \frac{\theta B}{6K},\ K = \frac{\chi\rho^{\frac{5}{4}}}{4\epsilon^{\frac{1}{4}}}. \tag{29}$$

For any $x_0 \in \mathcal{M}$ and sufficiently small $\epsilon > 0$, choose $\mu = \mathcal{O}\left(\frac{\epsilon^{1/4}}{d^{1/4}\chi}\right) = \tilde{\mathcal{O}}\left(\frac{\epsilon^{1/4}}{d^{1/4}}\right)$ in Lines 3 of Algorithm 1, and $\mu = \min\left\{\mathcal{O}\left(\frac{\epsilon^{5/8}}{d^{1/4}\chi^2}\right), \mathcal{O}\left(\frac{\epsilon^{7/8}}{\chi^3\sqrt{d}}\right)\right\} = \tilde{\mathcal{O}}\left(\frac{\epsilon^{7/8}}{\sqrt{d}}\right)$ in Line 5 of Subroutine 2. Then perturbed Algorithm 1 with Option I outputs an ϵ-approximate second-order stationary point with a probability of at least $1-\delta$. The total number of function value evaluations is no more than

$$O\left(\frac{(f(x_0)-f_{\text{low}})d}{\epsilon^{\frac{7}{4}}}\log^6\left(\frac{d}{\delta\epsilon}\right)\right).$$

Proof. By a similar argument, for the scenario $\|g_{x_t}(0;\mu)\| \geq lB$ at iterate x_t, we have

$$f(x_{t+1}) - f(x_t) \leq -\min\left\{\frac{lB^2}{16}, lb^2\right\} = -\frac{l\epsilon}{1024\chi^4\rho}.$$

For the tangent space step at iterate x_t, we start with a perturbed point in $\mathrm{T}_{x_t}\mathcal{M}$, that is $s^0_{x_t} = \xi_t \sim \mathrm{Uni}(\mathbb{B}_{x_t,r}(0))$, it follows

$$f(x_t) - \hat{f}_{x_t}(s^0_{x_t}) = \hat{f}_{x_t}(0) - \hat{f}_{x_t}(\xi_t) \leq \langle\nabla\hat{f}_{x_t}(\xi_t), -\xi_t\rangle + \frac{l}{2}\|\xi_t\|^2 \leq r\cdot\|\nabla\hat{f}_{x_t}(\xi_t)\| + \frac{lr^2}{2}.$$

Recall we choose $\mu = \mathcal{O}\left(\frac{\epsilon^{1/4}}{d^{1/4}\chi}\right)$ in Line 3 of Algorithm 1, and thus $\mathbf{E}(\mu) \leq \frac{lB}{2}$ holds. Consequently, the term $\|\nabla\hat{f}_{x_t}(\xi_t)\|$ can be upper bounded as

$$\|\nabla\hat{f}_{x_t}(\xi_t)\| \leq \|\nabla\hat{f}_{x_t}(\xi_t) - \nabla\hat{f}_{x_t}(0)\| + \|\nabla\hat{f}_{x_t}(0) - g_{x_t}(0;\mu)\| + \|g_{x_t}(0;\mu)\| \leq l\cdot\|\xi_t\| + \mathbf{E}(\mu) + lB \leq lr + \frac{3lB}{2}.$$

Substituting $r = \frac{\theta B}{6K} = \tilde{\mathcal{O}}(\epsilon)$ gives

$$f(x_t) - \hat{f}_{x_t}(s^0_{x_t}) \leq \frac{3lr^2}{2} + \frac{3lrB}{2} \leq \tilde{\mathcal{O}}\left(\epsilon^2\right) + \frac{\theta B^2}{16\eta K} \leq \frac{\theta B^2}{8\eta K},$$

combining with the choice of $\mu \leq \mathcal{O}\left(\frac{\epsilon^{5/8}}{d^{1/4}\chi^2}\right)$ in Line 5 of Subroutine 2 leads to

$$\begin{aligned}
&f(x_{t+1}) - f(x_t)\\
=&\hat{f}_{x_t}(s^{\mathcal{K}}_{x_t}) - \hat{f}_{x_t}(s^0_{x_t}) + \hat{f}_{x_t}(\xi_t) - f(x_t)\\
\leq& -\frac{\theta B^2}{4\eta K} + \eta K\mathbf{E}(\mu)^2 + \frac{K}{2B^{\frac{3}{2}}}\mathbf{E}(\mu)^2 + \frac{KB^{\frac{7}{2}}}{2} + \frac{5\eta K\mathbf{E}(\mu)^2}{\theta} + \frac{20\eta K\rho^2B^4}{\theta} + \frac{32\rho B^3}{3}\\
\leq& -\frac{\epsilon^{\frac{3}{2}}}{96\chi^5\sqrt{\rho}} + \eta K\mathbf{E}(\mu)^2 + \frac{K}{2B^{\frac{3}{2}}}\mathbf{E}(\mu)^2 + \frac{5\eta K\mathbf{E}(\mu)^2}{\theta}\\
\leq& -\frac{\epsilon^{\frac{3}{2}}}{96\chi^5\sqrt{\rho}} + \mathcal{O}\left(\frac{\epsilon^{\frac{9}{4}}}{\chi^7}\right) + \mathcal{O}\left(\frac{\epsilon^{\frac{3}{2}}}{\chi^9}\right) + \mathcal{O}\left(\frac{\epsilon^2}{\chi^7}\right)\\
\leq& -\frac{\epsilon^{\frac{3}{2}}}{192\chi^5\sqrt{\rho}}.
\end{aligned}$$

when the "if condition" triggers. Therefore, in the case of the function value decrease, we have

$$f(x_{t+1}) - f(x_t) \leq -\min\left\{\frac{l\epsilon}{1024\chi^4\rho}, \frac{\epsilon^{\frac{3}{2}}}{192\chi^5\sqrt{\rho}}\right\} = -\frac{\epsilon^{\frac{3}{2}}}{192\chi^5\sqrt{\rho}}. \tag{30}$$

Similarly, since the tangent space step requires at most $K = \mathcal{O}(\frac{\chi}{\epsilon^{1/4}})$ iterations and each iterate needs $2d$ function value evaluations to construct the zeroth-order estimator, the total number of function value evaluations does not exceed

$$O\left(\frac{(f(x_0)-f_{\text{low}})d}{\epsilon^{\frac{7}{4}}}\log^6\left(\frac{d}{\delta\epsilon}\right)\right).$$

For the scenario that the "if condition" does not trigger in the tangent space step at iterate x_t, from the proof of Theorem 4.1, it holds that

$$\|\operatorname{grad} f(x_{t+1})\| \leq \mathcal{O}(\epsilon).$$

To achieve the ϵ-approximate second-order stationary points, it remains to analyze the value of $\lambda_{\min}(\operatorname{Hess} f(x_{t+1}))$. Suppose $\lambda_{\min}(\nabla^2 \hat{f}_{x_t}(0)) \geq -\sqrt{\rho\epsilon}$, then it holds that

$$\begin{aligned}
&\lambda_{\min}(\nabla^2 \hat{f}_{x_t}(y^*_{x_t})) \\
\geq&\lambda_{\min}(\nabla^2 \hat{f}_{x_t}(0)) - |\lambda_{\min}(\nabla^2 \hat{f}_{x_t}(y^*_{x_t})) - \lambda_{\min}(\nabla^2 \hat{f}_{x_t}(0))| \\
\geq&\lambda_{\min}(\nabla^2 \hat{f}_{x_t}(0)) - \|\nabla^2 \hat{f}_{x_t}(y^*_{x_t}) - \nabla^2 \hat{f}_{x_t}(0)\| \\
\geq&\lambda_{\min}(\nabla^2 \hat{f}_{x_t}(0)) - \rho\|y^*_{x_t} - s^0_{x_t}\| - \rho\|s^0_{x_t}\| \\
\geq&\lambda_{\min}(\nabla^2 \hat{f}_{x_t}(0)) - 2\rho B - \rho r \qquad (31a) \\
\geq&-2\sqrt{\rho\epsilon},
\end{aligned}$$

where (31a) follows from $\|y^*_{x_t} - s^0_{x_t}\| \leq 2B$ and $\|s^0_{x_t}\| = \|\xi_t\| \leq r$. From Lemma C.1, we have

$$\nabla^2 \hat{f}_{x_t}(y^*_{x_t}) = T^*_{x_t, y^*_{x_t}} \operatorname{Hess} f(x_{t+1}) T_{x_t, y^*_{x_t}} + W_{y^*_{x_t}},$$

and it implies that

$$\begin{aligned}
&\lambda_{\min}(\operatorname{Hess} f(x_{t+1})) \\
\geq&\frac{\lambda_{\min}\left(T^*_{x_t, y^*_{x_t}} \operatorname{Hess} f(x_{t+1}) T_{x_t, y^*_{x_t}}\right)}{\lambda_{\max}(T^*_{x_t, y^*_{x_t}} T_{x_t, y^*_{x_t}})} \qquad (32a) \\
\geq&\frac{\lambda_{\min}\left(\nabla^2 \hat{f}_{x_t}(y^*_{x_t}) - W_{y^*_{x_t}}\right)}{\sigma^2_{\max}} \qquad (32b) \\
\geq&\frac{\lambda_{\min}\left(\nabla^2 \hat{f}_{x_t}(y^*_{x_t})\right) + \lambda_{\min}\left(-W_{y^*_{x_t}}\right)}{\sigma^2_{\max}} \qquad (32c) \\
\geq&-\frac{2\sqrt{\rho\epsilon}}{\sigma^2_{\max}} - \frac{25\tau}{\sigma^2_{\max}\sigma_{\min}}\epsilon \qquad (32d) \\
\geq&-\frac{4\sqrt{\rho\epsilon}}{\sigma^2_{\max}}
\end{aligned}$$

where (32a) is due to the Ostrowski's Theorem (Ostrowski, 1961), (32b) comes from Assumption 4.3, (32c) uses Wely's inequality (Horn & Johnson, 2012), and (32d) comes from the following inequality

$$\|W_{y^*_{x_t}}\| = \max_{\dot{s}_{x_t} \in \mathrm{T}_{x_t}\mathcal{M}, \|\dot{s}_{x_t}\|=1} \langle W_{y^*_{x_t}} \dot{s}_{x_t}, \dot{s}_{x_t} \rangle \leq \|\gamma''_{x_t, \dot{s}_{x_t}}(0)\| \|\operatorname{grad} f(\operatorname{Retr}_{x_t}(y^*_{x_t}))\| \leq \frac{25\tau}{\sigma_{\min}}\epsilon.$$

Consider the case $\lambda_{\min}(\nabla^2 \hat{f}_{x_t}(0)) < -\sqrt{\rho\epsilon}$, we define the following stuck region in the tangent space step at iterate x_t:

$$\mathcal{X}_t^{\text{stuck}} = \begin{cases} \left\{ s_{x_t} \in \mathbb{B}_{x_t, r}(0) : \{s^k_{x_t}\}_{k=1}^K \text{ satisfies } s^0_{x_t} = s_{x_t} \text{ and } K\sum_{k=0}^{K-1} \|s^{k+1}_{x_t} - s^k_{x_t}\| \leq B^2 \right\}, & \text{if } \lambda_{\min}(\nabla^2 \hat{f}_{x_t}(0)) < -\sqrt{\rho\epsilon}, \\ \emptyset, \text{ otherwise.} \end{cases}$$

From Lemma E.8, we know that the probability of $s^0_{x_t} = \xi_t \in \mathcal{X}_t^{\text{stuck}}$ satisfies $\Pr\left\{\xi_t \in \mathcal{X}_t^{\text{stuck}}\right\} \leq \delta$. Therefore, once the 'if condition' does not trigger in the tangent space step, with a probability of at least $1-\delta$, x_{t+1} is an ϵ-approximate second-order stationary point. □

Lemma E.8. *Suppose that Assumption 4.1, 4.2 and 4.3 hold. Under the parameter settings in Theorem 4.2, let* $r_0 = \frac{\delta r}{\sqrt{d}}$. *In cases where* $\lambda_{\min}(\nabla^2 \hat{f}_{x_t}(0)) < -\sqrt{\rho\epsilon}$, *given* $s'^0_{x_t}, s''^0_{x_t} \in \mathbb{B}x_t, r(0)$ *with* $s{x_t}'^0 - s''^0_{x_t} = r_0 v_1$, *where* v_1 *is the minimum*

eigen-direction of $\nabla^2 \hat{f}_{x_t}(0)$*, choose* $\mu = \mathcal{O}\left(\frac{\epsilon^{7/8}}{\chi^3\sqrt{d}}\right) = \tilde{\mathcal{O}}\left(\frac{\epsilon^{7/8}}{\sqrt{d}}\right)$ *such that* $\mathbf{E}(\mu) \leq \frac{\rho B\theta r_0}{2}$*. By running the tangent space step starting at* $s_{x_t}^{\prime 0}$ *and* $s_{x_t}^{\prime\prime 0}$ *respectively we have*

$$\max\left\{K\sum_{k=0}^{K-1}\left\|s_{x_t}^{\prime k+1}-s_{x_t}^{\prime k}\right\|^2, K\sum_{k=0}^{K-1}\left\|s_{x_t}^{\prime\prime k+1}-s_{x_t}^{\prime\prime k}\right\|^2\right\} > B^2.$$

that is, at least one of the iterates triggers the "if condition".

The proof of this lemma follows from Lemma 18 in (Jin et al., 2018) and Lemma B.2 in (Li & Lin, 2022), and thus we list the sketch. The details can be found in (Li & Lin, 2022) and (Jin et al., 2018)

Proof. For any point $s_{x_t} \in \mathrm{T}_{x_t}\mathcal{M}$, we introduce the notation $e_{x_t}(s_{x_t};\mu) := \nabla \hat{f}_{x_t}(s_{x_t}) - g_{x_t}(s_{x_t};\mu)$, and thus, the update in tangent space step at iterate x_t can be rewritten as

$$\begin{aligned} y_{x_t}^k &= s_{x_t}^k + (1-\theta)(s_{x_t}^k - s_{x_t}^{k-1}) \\ s_{x_t}^{k+1} &= y_{x_t}^k - \eta\nabla\hat{f}_{x_t}(y_{x_t}^k) + \eta e_{x_t}(y_{x_t}^k;\mu). \end{aligned}$$

Denoting $\mathbf{w}_{x_t}^k := s_{x_t}^{\prime k} - s_{x_t}^{\prime\prime k}$, from the above update, we obtain

$$\begin{aligned} \begin{bmatrix}\mathbf{w}_{x_t}^{k+1}\\ \mathbf{w}_{x_t}^k\end{bmatrix} &= \begin{bmatrix}(2-\theta)(I-\eta\nabla^2\hat{f}_{x_t}(0)) & -(1-\theta)(I-\eta\nabla^2\hat{f}_{x_t}(0))\\ I & 0\end{bmatrix}\begin{bmatrix}\mathbf{w}_{x_t}^k\\ \mathbf{w}_{x_t}^{k-1}\end{bmatrix} \\ &\quad - \eta\begin{bmatrix}(2-\theta)\Delta_{x_t}^k\mathbf{w}_{x_t}^k - (1-\theta)\Delta_{x_t}^k\mathbf{w}_{x_t}^{k-1} + e_{x_t}(y_{x_t}^{\prime\prime k};\mu) - e_{x_t}(y_{x_t}^{\prime k};\mu)\\ 0\end{bmatrix}, \end{aligned}$$

where $\Delta_{x_t}^k = \int_0^1 \left(\nabla^2\hat{f}\left(\tau y_{x_t}^{\prime k} + (1-\tau)y_{x_t}^{\prime\prime k}\right) - \nabla^2\hat{f}_{x_t}(0)\right)\mathrm{d}\tau$. For simplicity, let

$$A_{x_t} = \begin{bmatrix}(2-\theta)(I-\eta\nabla^2\hat{f}_{x_t}(0)) & -(1-\theta)(I-\eta\nabla^2\hat{f}_{x_t}(0))\\ I & 0\end{bmatrix}$$

and $\phi_{x_t}^k = (2-\theta)\Delta_{x_t}^k\mathbf{w}_{x_t}^k - (1-\theta)\Delta_{x_t}^k\mathbf{w}_{x_t}^{k-1} + e_{x_t}(y_{x_t}^{\prime\prime k};\mu) - e_{x_t}(y_{x_t}^{\prime k};\mu)$, we further have

$$\begin{bmatrix}\mathbf{w}_{x_t}^{k+1}\\ \mathbf{w}_{x_t}^k\end{bmatrix} = A_{x_t}\begin{bmatrix}\mathbf{w}_{x_t}^k\\ \mathbf{w}_{x_t}^{k-1}\end{bmatrix} - \eta\begin{bmatrix}\phi_{x_t}^k\\ 0\end{bmatrix} = A_{x_t}^{k+1}\begin{bmatrix}\mathbf{w}_{x_t}^0\\ \mathbf{w}_{x_t}^0\end{bmatrix} - \eta\sum_{r=0}^{k}A_{x_t}^{k-r}\begin{bmatrix}\phi_{x_t}^r\\ 0\end{bmatrix}, \tag{33}$$

To proceed, we prove this lemma by contradiction. Assume that none of the iterates $s_{x_t}^{\prime 0}, s_{x_t}^{\prime 1}, \ldots, s_{x_t}^{\prime K}$ and $s_{x_t}^{\prime\prime 0}, s_{x_t}^{\prime\prime 1}, \ldots, s_{x_t}^{\prime\prime K}$ trigger the "if condition", which implies that

$$\begin{aligned} &\|s_{x_t}^{\prime k} - s_{x_t}^{\prime 0}\| \leq B,\ \|y_{x_t}^{\prime k} - s_{x_t}^{\prime 0}\| \leq 2B,\ k = 1,\ldots,K, \\ &\|s_{x_t}^{\prime\prime k} - s_{x_t}^{\prime\prime 0}\| \leq B,\ \|y_{x_t}^{\prime\prime k} - s_{x_t}^{\prime\prime 0}\| \leq 2B,\ k = 1,\ldots,K. \end{aligned}$$

Combining with the fact that $\|s_{x_t}^{\prime 0}\| \leq r$, $\|s_{x_t}^{\prime\prime 0}\| \leq r$ and $r \leq B$, we have

$$\|\Delta_{x_t}^k\| \leq \rho\max\left\{\|y_{x_t}^{\prime k}\|, \|y_{x_t}^{\prime\prime k}\|\right\} \leq \rho\max\left\{\|y_{x_t}^{\prime k} - s_{x_t}^{\prime 0}\| + \|s_{x_t}^{\prime 0}\|, \|y_{x_t}^{\prime\prime k} - s_{x_t}^{\prime\prime 0}\| + \|s_{x_t}^{\prime\prime 0}\|\right\} \leq 3\rho B.$$

Consequently, the term $\|\phi_{x_t}^k\|$ can be upper bounded as

$$\|\phi_{x_t}^k\| \leq 2\|\Delta_{x_t}^k\|\|\mathbf{w}_{x_t}^k\| + \|\Delta_{x_t}^k\|\|\mathbf{w}_{x_t}^{k-1}\| + 2\mathbf{E}(\mu) \leq 6\rho B(\|\mathbf{w}_{x_t}^k\| + \|\mathbf{w}_{x_t}^{k-1}\|) + 2\mathbf{E}(\mu).$$

From the update (33), we see

$$\mathbf{w}_{x_t}^k = [I \quad 0]A_{x_t}^k\begin{bmatrix}\mathbf{w}_{x_t}^0\\ \mathbf{w}_{x_t}^0\end{bmatrix} - \eta[I \quad 0]\sum_{r=0}^{k-1}A_{x_t}^{k-1-r}\begin{bmatrix}\phi_{x_t}^r\\ 0\end{bmatrix}.$$

Next, we set up an induction on k to show:

$$\left\|\eta[I \quad 0]\sum_{r=0}^{k-1} A_{x_t}^{k-1-r}\begin{bmatrix}\phi_{x_t}^r\\ 0\end{bmatrix}\right\| \le \frac{1}{2}\left\|[I \quad 0]A_{x_t}^k\begin{bmatrix}\mathbf{w}_{x_t}^0\\ \mathbf{w}_{x_t}^0\end{bmatrix}\right\|.$$

It is easy to check the base case holds for $k=0$ since $\mathbf{E}(\mu)=\frac{\rho B\theta r_0}{2}\le 2\eta\rho B r_0$. Then, assume that for all iterations less than or equal to k, the induction assumption holds. We have

$$\left\|\mathbf{w}_{x_t}^k\right\| = \left\|[I \quad 0]A_{x_t}^k\begin{bmatrix}\mathbf{w}_{x_t}^0\\ \mathbf{w}_{x_t}^0\end{bmatrix} - \eta[I \quad 0]\sum_{r=0}^{k-1} A_{x_t}^{k-1-r}\begin{bmatrix}\phi_{x_t}^r\\ 0\end{bmatrix}\right\| \le 2\left\|[I \quad 0]A_{x_t}^k\begin{bmatrix}\mathbf{w}_{x_t}^0\\ \mathbf{w}_{x_t}^0\end{bmatrix}\right\|,$$

and it further implies that

$$\left\|\phi_{x_t}^k\right\| \le 12\rho B\left(\left\|[I \quad 0]A_{x_t}^k\begin{bmatrix}\mathbf{w}_{x_t}^0\\ \mathbf{w}_{x_t}^0\end{bmatrix}\right\| + \left\|[I \quad 0]A_{x_t}^{k-1}\begin{bmatrix}\mathbf{w}_{x_t}^0\\ \mathbf{w}_{x_t}^0\end{bmatrix}\right\|\right) + 2\mathbf{E}(\mu) \le 24\rho B\left\|[I \quad 0]A_{x_t}^k\begin{bmatrix}\mathbf{w}_{x_t}^0\\ \mathbf{w}_{x_t}^0\end{bmatrix}\right\| + 2\mathbf{E}(\mu),$$

where the last inequality is due to the monotonicity of $\left\|[I \quad 0]A_{x_t}^k\begin{bmatrix}\mathbf{w}_{x_t}^0\\ \mathbf{w}_{x_t}^0\end{bmatrix}\right\|$ in k (Lemma 33 in (Jin et al., 2018)). For the case $k+1$, we have

$$\begin{aligned}
&\left\|\eta[I \quad 0]\sum_{r=0}^{k} A_{x_t}^{k-r}\begin{bmatrix}\phi_{x_t}^r\\ 0\end{bmatrix}\right\| \\
\le&\eta\sum_{r=0}^{k}\left\|[I \quad 0]A_{x_t}^{k-r}\begin{bmatrix}I\\ 0\end{bmatrix}\right\|\left\|\phi_{x_t}^r\right\| \\
\le&\eta\sum_{r=0}^{k}\left\|[I \quad 0]A_{x_t}^{k-r}\begin{bmatrix}I\\ 0\end{bmatrix}\right\|\left(24\rho B\left\|[I \quad 0]A_{x_t}^{r}\begin{bmatrix}\mathbf{w}_{x_t}^0\\ \mathbf{w}_{x_t}^0\end{bmatrix}\right\| + 2\mathbf{E}(\mu)\right) \\
=&\eta\sum_{r=0}^{k}|a_{x_t}^{k-r}|\left(24\rho B|a_{x_t}^r - b_{x_t}^r|r_0 + 2\mathbf{E}(\mu)\right) &\text{(34a)}\\
\le&26\eta\rho B\sum_{r=0}^{k}|a_{x_t}^{k-r}||a_{x_t}^r - b_{x_t}^r|r_0 &\text{(34b)}\\
\le&26\eta\rho B\sum_{r=0}^{k}\left(\frac{2}{\theta}+k+1\right)|a_{x_t}^{k+1} - b_{x_t}^{k+1}|r_0 &\text{(34c)}\\
\le&26\eta\rho BK\left(\frac{2}{\theta}+K\right)\left\|[I \quad 0]A_{x_t}^{k+1}\begin{bmatrix}\mathbf{w}_{x_t}^0\\ \mathbf{w}_{x_t}^0\end{bmatrix}\right\| \\
\le&\frac{1}{2}\left\|[I \quad 0]A_{x_t}^{k+1}\begin{bmatrix}\mathbf{w}_{x_t}^0\\ \mathbf{w}_{x_t}^0\end{bmatrix}\right\|, &\text{(34d)}
\end{aligned}$$

where we define $[a_{x_t}^k \quad -b_{x_t}^k] = [1 \quad 0]A_{x_t,\min}^k$ and

$$A_{x_t,\min} = \begin{bmatrix}(2-\theta)(1-\eta\lambda_{\min}(\nabla^2\hat{f}_{x_t}(0))) & -(1-\theta)(I-\eta\lambda_{\min}(\nabla^2\hat{f}_{x_t}(0)))\\ I & 0\end{bmatrix}.$$

Then, we apply the same argument in the proof of Lemma B.2 in (Li & Lin, 2022) to the inequality (34a), (34b) comes from $|a_{x_t}^r - b_{x_t}^r| \ge \frac{\theta}{2}$ (Lemma 38 in (Jin et al., 2018)) and $\mathbf{E}(\mu)\le\frac{\rho B\theta r_0}{2}$, and (34c) uses Lemma 31 in (Jin et al., 2018). From the parameter settings, we have $26\eta\rho BK\left(\frac{2}{\theta}+K\right)\le\frac{1}{2}$ in (34d). Therefore, the introduction is established, which yields

$$\|\mathbf{w}_{x_t}^K\| \ge \frac{1}{2}\left\|[I \quad 0]A_{x_t}^K\begin{bmatrix}\mathbf{w}_{x_t}^0\\ \mathbf{w}_{x_t}^0\end{bmatrix}\right\| = \frac{r_0}{2}|a_{x_t}^K - b_{x_t}^K| \ge \frac{\theta r_0}{4}\left(1+\frac{\theta}{2}\right)^K \ge 5B,$$

where we use Lemma 33 in (Jin et al., 2018), $\eta\lambda_{\min}(\nabla^2 \hat{f}_{x_t}(0))) \leq -\theta^2$ and $K \geq \frac{2}{\theta}\log\frac{20B}{\theta r_0}$. However, for the term $\|\mathbf{w}_{x_t}^K\|$, it also holds that

$$\|\mathbf{w}_{x_t}^K\| \leq \|s_{x_t}'^K - s_{x_t}'^0\| + \|s_{x_t}'^0 - s_{x_t}''^0\| + \|s_{x_t}''^K - s_{x_t}''^0\| \leq 4B,$$

which leads to a contradiction. Therefore, at least one of the iterates $s_{x_t}'^0, s_{x_t}'^1, \ldots, s_{x_t}'^K$ and $s_{x_t}''^0, s_{x_t}''^1, \ldots, s_{x_t}''^K$ trigger the "if condition".

□

F. Proofs of Non-asymptotic Convergence Analysis

In this section, we prove that non-perturbed RAZGD with Option II converges to second-order stationary points asymptotically. It follows from that the tangent space step TSS locally avoids saddle points. To prove the local saddle avoidance, it is helpful to use the augmentation method to extend the update rule in the tangent space to a dynamical system of s^{k+1} and w^{k+1} that only depend on s^k and w^k, i.e., regard y^k as an intermediate variable. Despite we are interested in the zeroth-order method, the stability analysis of the zeroth-order algorithm heavily depends on the structure of its first-order counterpart. Therefore, we start with the analysis of the first-order tangent space step, which provides the second-order convergence immediately.

F.1. First-order tangent space step

We use the augmentation method to re-write the tangent space step in the following way,

$$y^k = s^k + (1-\theta)(s^k - w^k) \tag{35}$$
$$s^{k+1} = y^k - \eta g(y^k;\mu) \tag{36}$$
$$w^{k+1} = s^k \tag{37}$$

where $g(y;\mu)$ is the zeroth order approximation of the gradient $\nabla f(y)$ with smoothing parameter μ. We will not emphasize that g is performed at the point x at this stage, just to reduce the complexity of notations. The three steps of the updating rule can be denoted by three mappings that consist of the mapping of the algorithm that updates s^k, w^k to s^{k+1}, w^{k+1}. We denote

$$\begin{aligned} F_1(s,w) &= s + (1-\theta)(s-w) \\ F_2(y) &= y - \eta g(y;\mu) \\ F_3(s) &= s, \end{aligned}$$

and then the algorithm can be written compactly as

$$\psi(s,w) = (F_2 \circ F_1(s,w), F_3(s))$$

which is a mapping from $T_xM \times T_xM$ onto itself. The fixed point of the first order accelerated (s^*, w^*) is necessarily a point such that $s^* = w^*$ and the gradient

$$\nabla f(y^*) = \nabla f(s^* + (1-\theta)(s^* - w^*)) = 0.$$

We will investigate the local structure of the zeroth order variant at the point (s^*, w^*). The differential $D\psi(s^*, w^*)$ equals to

$$D\psi(s^*, w^*) = \begin{bmatrix} DF_2 \circ DF_1(s^*, w^*) \\ DF_3(s^*) \end{bmatrix} \tag{38}$$

As an immediate result and important argument bridging first-order and zeroth-order accelerated gradient descent in the tangent space, we first prove that the first-order tangent space step avoids saddle points. The following classic result of the stable manifold theorem will be used to complete the proof for the first-order method.

Theorem F.1 ((Shub, 1987)). *Let p be a fixed point for the C^r local diffeomorphism $h : U \to \mathbb{R}^n$ where $U \subset \mathbb{R}^n$ is an open neighborhood of p in $\mathbb{R}^n$ and $r \geq 1$. Let $E^s \oplus E^c \oplus E^u$ be the invariant splitting of $\mathbb{R}^n$ into generalized eigenspaces of $Dh(p)$ corresponding to eigenvalues of absolute value less than one, equal to one, and greater than one. To the $Dh(p)$ invariant subspace $E^s \oplus E^c$ there is an associated local h invariant embedded disc W_{sc}^{loc} which is the graph of a C^r function $r : E^s \oplus E^c \to E^u$, and ball B around p such that: $h(W_{sc}^{loc}) \cap B \subset W_{sc}^{loc}$. If $h^n(x) \in B$ for all $n \geq 0$, then $x \in W_{sc}^{loc}$.*

Lemma F.1. *Suppose that* 0 *is a strict saddle point of the pullback function in the tangent space, then the measure of the local initial points that converge to* 0 *is zero.*

Proof. The structure of ψ gives the expression of its differential. Since

$$DF_2 \circ DF_1 = (I - \eta\nabla^2 f(y^*))((2-\theta)I, -(1-\theta)I) \tag{39}$$
$$= \big((2-\theta)(I - \eta\nabla^2 f(y^*)), -(1-\theta)(I - \eta\nabla^2 f(y^*))\big) \tag{40}$$

and

$$DF_3 = (I, 0)\,,$$

we have that

$$D\psi(s^*, w^*) = \begin{bmatrix} (2-\theta)(I - \eta\nabla^2 f(y^*)) & -(1-\theta)(I - \eta\nabla^2 f(y^*)) \\ I & 0 \end{bmatrix}.$$

Note that $D\psi$ is similar to

$$\begin{bmatrix} (2-\theta)(I - \eta H) & -(1-\theta)(I - \eta H) \\ I & 0 \end{bmatrix}$$

provided $\nabla^2 f(y^*)$ is diagonalizable where H is the diagonal matrix consisting of eigenvalues of $\nabla^2 f(y^*)$. We can abuse the notation by

$$\det(D\psi - \lambda I) = \det\left(\begin{bmatrix} (2-\theta)(I - \eta H) - \lambda I & -(1-\theta)(I - \eta H) \\ I & -\lambda I \end{bmatrix}\right) \tag{41}$$
$$= \det\left(((2-\theta)(I - \eta H) - \lambda I) + (1-\theta)(I - \eta H)\left(-\frac{1}{\lambda} I\right)\right)(-\lambda)^n \tag{42}$$
$$= \det\left(-\lambda((2-\theta)(I - \eta H) - \lambda I) + (1-\theta)(I - \eta H)\right) \tag{43}$$
$$= \det\left(\lambda^2 I - \lambda(2-\theta)(I - \eta H) + (1-\theta)(I - \eta H)\right) \tag{44}$$

Since all matrices involved above are all diagonal, the determinant is nothing but the product of the following polynomials for $i \in [n]$:

$$\lambda^2 - (2-\theta)(1 - \eta\lambda_i)\lambda + (1-\theta)(1 - \eta\lambda_i).$$

Suppose λ_i is a negative eigenvalue (existence is guaranteed by assuming y^* is a saddle point), the eigenvalue of $D\psi$ must contain the following one

$$\lambda = \frac{(2-\theta)(1 - \eta\lambda_i) + \sqrt{(2-\theta)^2(1 - \eta\lambda_i)^2 - 4(1-\theta)(1 - \eta\lambda_i)}}{2}.$$

Since we can choose θ and η so that

$$\eta > \frac{\frac{2}{2-\theta} - 1}{-\lambda_i}$$

which guarantees that

$$(2-\theta)(1 - \eta\lambda_i) > 2,$$

thus $\lambda > 1$ (unstable fixed point). The step η can be arbitrarily small (so that ψ is a diffeomorphism) by taking θ as small as possible. Applying the center-stable manifold theorem F.1, we complete the proof.

□

F.2. Zeroth-order tangent space step with constant contraction

In this subsection, we show the asymptotic convergence for the zeroth-order tangent space step with constant contracting parameter β. The zeroth order tangent space step can be extended with the smoothing parameter to a new mapping $\tilde{\psi}(s, w, \mu)$ with a contraction factor β as follows,

$$\tilde{\psi}(s, w, \mu) = (\psi(s, w, \mu), \beta\mu) \tag{45}$$

where $\psi(s, w, \mu)$ considers the smoothing parameter as a proper variable such that ψ is a mapping defined on $T_xM \times T_xM \times \mathbb{R} \to T_xM \times T_xM$. Note that the zeroth order approximation $g(y; \mu)$ may not provide a fixed point of the gradient descent, in order to asymptotically output a fixed point the gradient descent, it is necessary to contract the smoothing parameter so that the zeroth order approximation algorithm has the same set of fixed points as the gradient descent. Motivated by the zeroth order approximation scheme of (Flokas et al., 2019). The tangent space mapping requires a contracting smoothing parameter $\beta\mu$ for the whole tangent space step **TSSA**. Another observation on the tangent space step **TSSA** from the asymptotic perspective, is the condition in the **while** loop. Since the asymptotic convergence empirically works well and is more convenient in the parameter settings, there is no need to use finite step K in **TSS** mapping, but the condition $k\sum_{j=0}^{k} \|s_k^{j+1} - s_x^j\|^2 > B^2$ suffices to control the process of the **while** loop. The next lemma shows that the **TSSA** step is almost impossible to converge to a saddle point.

Lemma F.2. *Consider mapping $\tilde{\psi}$ is defined as (45). The set of initial condition in the tangent space that converges to saddle point, i.e., 0 in this setting, has measure zero.*

Proof. The differential of $\tilde{\psi}$ can be computed in the following way,

$$D\tilde{\psi} = \begin{bmatrix} D_s\psi & D_w\psi & D_\mu\psi \\ 0 & 0 & \beta \end{bmatrix}.$$

Recall that in the zeroth order approximation, ψ is a mapping consisting of the approximated gradient $g(y; \mu)$, which is different from the first order method. The differential of $g(y; \mu)$ gives the differential of $\tilde{\psi}$ and ψ, so we compute $Dg(y; \mu)$ concretely, since 0 is the only fixed point for the μ component, we need to compute the Taylor expansion at $(y^*, 0)$ where y^* is the fixed point of the first order counterpart of the algorithm. Thus, we have $D_{s,w}\psi(y^*, 0)$ coincide with the differential computed in the first order method, and $D_\mu\psi$ is $(-\eta D_\mu g(y; \mu), 0)^\top$, where

$$D_\mu g(y; \mu) = \begin{bmatrix} \frac{\partial g_1(y;\mu)}{\partial \mu} \\ \vdots \\ \frac{\partial g_d(y;\mu)}{\partial \mu} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \mu}\left(\frac{f(y+\mu e_1)-f(y)}{\mu}\right) \\ \vdots \\ \frac{\partial}{\partial \mu}\left(\frac{f(y+\mu e_d)-f(y)}{\mu}\right) \end{bmatrix}. \tag{46}$$

Since the block matrix $[D_s\psi, D_w\psi]$ computed at $(y^*, 0)$ is the same as that has been computed in the first order method, and we have shown that the determinant of the block matrix is not zero, therefore, we are ready to obtain the determinant of $D\tilde{\psi}$ at the fixed point $(y^*, 0)$. It is obvious that

$$\det\left(D\tilde{\psi}(s^*, w^*, 0)\right) = \det\left(D\psi(s^*, w^*, 0)\right) \cdot \beta$$

and

$$\det\left(D\tilde{\psi}(s^*, w^*, 0) - \lambda I\right) = \det\left(D\psi(s^*, w^*, 0) - \lambda I\right)(\beta - \lambda).$$

Based on the spectral analysis of the first order tangent step, we conclude that the escaping direction the zeroth order approximation tangent space step is provided by the unstable direction of the first order method. In the end, applying the stable manifold theorem (Shub, 1987), we conclude that the set of initial points that converge to saddle point in the asymptotic variant of tangent space step is of measure zero since these initial points belong to a lower dimensional manifold. □

F.3. Zeroth-order tangent space step with time-varying contraction

We next prove the asymptotic saddle avoidance of the tangent space step **TSSA** when the smoothing parameter reduces in a slower rate, which is more practical from a zeroth-order perspective.

Lemma F.3. *Suppose **TSSA** is executed with the update rule on the smoothing parameter μ given by*

$$\mu_{k+1} = \left(1 - \frac{1}{k+2}\right)\mu_k$$

*, then the probability of **TSSA** converging to a saddle point is zero.*

Proof. The dynamical system augmented by μ is the following,

$$y^k = s^k + (1-\theta)(s^k - w^k) \tag{47}$$
$$s^{k+1} = y^k - \eta g(y^k; \mu_k) \tag{48}$$
$$w^{k+1} = s^k \tag{49}$$
$$\mu_{k+1} = (1 - \frac{1}{k+2})\mu_k \tag{50}$$

which is an augmentation with the smoothing parameter μ Following the previous arguments, we can write the mapping on the parameters (s, w, μ) as follows,

$$\tilde{\psi}_k(s, w, \mu) = \left(\psi(s, w, \mu), \left(1 - \frac{1}{k+2}\right)\mu\right)$$

and the differential of $\tilde{\psi}$ is

$$D\tilde{\psi}_k = \begin{bmatrix} D_s\psi & D_w\psi & D_\mu\psi \\ 0 & 0 & 1 - \frac{1}{k+2} \end{bmatrix}$$

Since the zeroth order method with contraction factor converges to stationary points of the corresponding first-order method, we can investigate the same Taylor expansion and differential of the algorithm expanded at the stationary point, especially at saddle point. The eigenvalues of the operator $D\tilde{\psi}(s^*, w^*, 0)$ can be analyzed by the characteristic polynomial

$$\det\left(D\tilde{\psi}_k(s^*, w^*, 0) - \lambda I\right) = \det\left(D\psi(s^*, w^*, 0) - \lambda I\right)\left(1 - \frac{1}{k+2} - \lambda\right).$$

Note that except for the eigenvalue $1 - \frac{1}{k+2}$, all the other eigenvalues are the same as the case whose the contraction parameter is a constant β. Therefore, there is an $(2d+1) \times (2d+1)$ invertible matrix C_k for each k such that

$$A_k = C_k^{-1} D\tilde{\psi}_k(s^*, w^*, 0) C_k = \begin{bmatrix} P_k & \\ & Q_k \end{bmatrix}$$

where the eigenvalues $\lambda_1, ..., \lambda_s$ of P_k have magnitude less than 1, and the eigenvalues $\lambda_{s+1}, ..., \lambda_{2d+1}$ of the matrix Q_k have magnitude greater than 1 (guaranteed by the property of a saddle point). Since the algorithm now is time dependent, i.e., the update rule $\tilde{\psi}_k$ contains a time dependent term $1 - \frac{1}{k+2}$ and thus the Jordan block is also time dependent, the time independent argument that directly follows the stable manifold theorem is not valid in this time dependent setting. To show the same result as what holds for constant contraction case, we need to investigate the structure of the dynamical system in detail. Denote $A(m, n)$ the successive product of the nth till the mth matrices, i.e., $A(m, n) = A_m \cdot ... \cdot A_n$. With the help of this notation, we can express the product

$$A(m, n) = \begin{bmatrix} P_m ... P_n & \\ & Q_m ... Q_n \end{bmatrix} = \begin{bmatrix} P(m, n) & \\ & Q(m, n) \end{bmatrix}.$$

Recall that the dynamical system induced by the tangent space step is

$$\begin{bmatrix} s^{k+1} \\ w^{k+1} \\ \mu_{k+1} \end{bmatrix} = \tilde{\psi}(s^k, w^k, \mu_k).$$

Assuming that the saddle point is $(s^*, w^*, 0) = (0, 0, 0)$, and the above dynamical system has the following expression obtained from the Taylor expansion around $(0, 0, 0)$,

$$\begin{bmatrix} s^{k+1} \\ w^{k+1} \\ \mu_{k+1} \end{bmatrix} = D\tilde{\psi}_k(0,0,0) \begin{bmatrix} s^k \\ w^k \\ \mu_k \end{bmatrix} + \xi_k(s^k, w^k, \mu_k)$$

where $\xi_k(\cdot,\cdot,\cdot)$ is the remainder of $\tilde{\psi}_k$. Starting from the initial condition (s^0, w^0, μ_0), the dynamical system can be represented by

$$z_{k+1} = \begin{bmatrix} P(k,0) & \\ & Q(k,0) \end{bmatrix} z_0 + \sum_{i=0}^{k} \begin{bmatrix} P(k,i+1) & \\ & Q(k,i+1) \end{bmatrix} \xi_i(z_i),$$

where z_k is the dynamical system topologically conjugated to (s^k, w^k, μ_k). Splitting z_k and $\xi_i(z_i)$ into contracting and expanding components according to $P(k,0)$ and $Q(k,0)$, i.e., this decomposition is actually based on the magnitudes of the eigenvalues of $D\tilde{\psi}_k$ which is determined by the Hessian of the objective function f at saddle points. Further information of the Jordan matrix P_k and Q_k can be inferred. The expanding matrix Q_k contains only constant eigenvalues with magnitude greater than 1. The contracting matrix P_k contains constant eigenvalues and one eigenvalue that is exactly $1 - \frac{1}{k+2}$. The stable-unstable decomposition of z_k can be further refined into stable with constant eigenvalues less than 1, stable with eigenvalue $1 - \frac{1}{k+2}$, and unstable with constant eigenvalues greater than 1. Specifically, we decompose z_k into

$$z_k = \begin{bmatrix} z_k^+ \\ z_k^\mu \\ z_k^- \end{bmatrix},$$

and $\xi_i(z_i)$ into

$$\xi_i(z_i) = \begin{bmatrix} \xi_i^+(z_i) \\ 0 \\ \xi_i^-(z_i) \end{bmatrix}$$

where the remainder with respect to z_k^μ is zero because the update rule of μ is a linear function. Based on this decomposition, we can refine the formulation of the dynamical system of z_k in the following way,

$$\begin{aligned} z_{k+1}^+ &= P(k,0) z_0^+ + \sum_{i=0}^{k} P(k,i+1)\xi_i^+(z_i) \\ z_{k+1}^\mu &= \left(1 - \frac{1}{k+2}\right) z_k^\mu \\ z_{k+1}^- &= Q(k,0) z_0^- + \sum_{i=0}^{k} Q(k,i+1)\xi_i^-(z_i) \end{aligned}$$

where we still use P as the Jordan block of stable component without distinguishing from the one containing $1 - \frac{1}{k+2}$. Letting $k \to \infty$, we have formally the unstable component z_0^- of the initial condition z_0 satisfying

$$z_0^- = -\sum_{i=1}^{\infty} Q(i-1,0)^{-1} \xi_{i-1}^-(z_{i-1}),$$

and then the updated term z_{k+1} can be written as

$$\begin{aligned} z_{k+1} &= z_{k+1}^+ \oplus z_{k+1}^\mu \oplus z_{k+1}^- \\ &= \left(P(k,0)z_0^+ + \sum_{i=0}^{k} P(k,i+1)\xi_i^+(z_i)\right) \oplus \left(1 - \frac{1}{k+2}\right) z_k^\mu \oplus \left(Q(k,0)z_0^- + \sum_{i=0}^{k} Q(k,i+1)\xi_i^-(z_i)\right) \end{aligned}$$

where the last summand can be further written as

$$-\sum_{i=0}^{\infty} Q(k+1+i,k+1)^{-1}\xi^{-}_{k+1+i}(z_{k+1+i}).$$

The update rule can be understood as an operator acting on the space of bounded sequences converging to zero. Since $P(k,0)$ and $Q(k,0)$ are matrices only involving constant eigenvalues, there exists constants $K_1, K_2 < 1$ such that

$$\|P(m,n)\|_2 \leq K_1^{m-n+1} \tag{51}$$

$$\|Q(m,n)^{-1}\|_2 \leq K_2^{m-n+1}. \tag{52}$$

The Lyapunov-Perron argument (Panageas et al., 2019) asserts that there exists a small neighborhood around the saddle point, such that T is an contraction map on the space of sequences converging to zero, and consequently, the initial point that can be carried to the saddle point (the zero) by the algorithm must lie on a lower dimensional manifold. To make this point precise, we investigate the norm of the difference of two sequences T acting on. Let $u = \{u_n\}_{n\in\mathbb{N}}$ and $v = \{v_n\}_{n\in\mathbb{N}}$,

$$(Tu - Tv)_{k+1} = (Tu)_{k+1} - (Tv)_{k+1} \tag{53}$$

$$= \left(Q(k,0)(u_0^{+} - v_0^{+}) + \sum_{i=0}^{k} P(t,i+1)(\xi_i^{+}(u_i) - \xi_i^{+}(v_i))\right) \tag{54}$$

$$\oplus \frac{1}{3(k+2)}(u_0^{\mu} - v_0^{\mu}) \tag{55}$$

$$\oplus \left(-\sum_{i=0}^{\infty} Q(k+1+i,k+1)^{-1}(\xi^{-}_{k+1+i}(u_{k+1+i}) - \xi^{-}_{k+1+i}(v_{k+1+i}))\right) \tag{56}$$

where the coefficient of the middle component comes from the product

$$\frac{1}{3(k+2)} = \prod_{i=0}^{k}\left(1 - \frac{1}{i+2}\right).$$

Let $d(u,v)$ be the metric defined by the supremum norm of the sequence $\{u_i - v_i\}_{i\in\mathbb{N}}$. Since it has been proven by (Feng et al., 2022) that T is a contracting map without the component of $\frac{1}{3(k+2)}(u_0^{\mu} - v_0^{\mu})$, i.e., there exists a constant $K < 1$ such that

$$d(Tu, Tv) \leq K d(u,v),$$

and $\frac{1}{3(k+2)} \leq \frac{1}{6} < 1$, it guarantees a new constant $K' < 1$, so that T acting on the space of the considered sequence with μ-component is an contracting map. Thus, the existence and uniqueness of the stable manifold in a neighborhood of the saddle point follow from the existence and uniqueness of the fixed point of T. So the probability of the initial condition lying on such lower dimensional manifold so that the iterates converge to saddle point is zeor. □

Now we are able to finalize the proof of Theorem 4.3.

proof of Theorem 4.3. It has been established that the probability of **TSSA** staying in a neighborhood of a saddle point is zero, for any **TSSA** stage.

$$\Pr\left\{\lim_{k\to\infty} k\sum_{j=0}^{k-1}\|s_x^{j+1} - s_x^{j}\|^2 \leq B^2\right\} = 0$$

and then the probability for the iterations to stay in a neighborhood of a second-order stationary point is 1. Since the zeroth-order acceleration with contracting parameter $\beta < 1$ converges to stationary point, it follows that the probability for **TSSA** output a second-order stationary point is 1. Together with the above Lemma F.3 for the case when the contracting parameter decreases in a slower manner (which slows the decreasing of smoothing parameter μ in the **TSSA** stage), we complete the proof of the theorem.

□

G. Implementation of RZGD and PZGD

For Riemannian zeroth-order gradient descent (RZGD), it iteratively utilizes the Riemannian zeroth-order gradient descent step (Subroutine 1) until convergence. In the case of Euclidean projected zeroth-order gradient descent (PZGD), we first compute the Euclidean zeroth-order estimator (denoted by $g_{\mathrm{E}}(\cdot)$), take a Euclidean zeroth-order gradient descent step, and then project onto the Riemannian manifold. The pseudocodes of both algorithms are presented below.

Algorithm 2 **R**iemannian **Z**eroth-order **G**radient **D**escent Algorithm (**RZGD**)

1: **input:** parameters η, and B
2: **initialize:** $x_0 \in \mathcal{M}$, $t = 0$
3: **for** $t = 0, 1, \cdots, \infty$ **do**
4: Compute estimator $g_{x_t}(0; \mu)$
5: **if** $\|g_{x_t}(0; \mu)\| \geq lB$ **then**
6: $x_{t+1} = \mathbf{RZGDS}(x_t, \eta, g_{x_t}(0; \mu))$
7: **else**
8: Terminate with x_t
9: **end if**
10: **end for**

Algorithm 3 Euclidean **P**rojected **Z**eroth-order **G**radient **D**escent Algorithm (**PZGD**)

1: **input:** parameters η_t
2: **initialize:** $x_0 \in \mathcal{M}$, $t = 0$
3: **for** $t = 0, 1, \cdots, \infty$ **do**
4: Compute Euclidean estimator $g_{\mathrm{E}}(x_t)$
5: $x_{t+1} = \mathrm{proj}_{\mathcal{M}}\left(x_t - \eta_t g_{\mathrm{E}}(x_t)\right)$
6: **end for**

For completeness, we establish the function query complexity of RZGD, which serves as a benchmark for demonstrating the acceleration achieved by our RAZGD.

Theorem G.1. *Suppose that Assumptions 4.1, 4.2 and 4.3 hold. Set parameters in Algorithm 2 as follows*

$$\eta = \frac{1}{4l},\ B = \frac{\epsilon}{2l}.$$

For any $x_0 \in \mathcal{M}$ and sufficiently small $\epsilon > 0$, choose $\mu = \mathcal{O}\left(\frac{\sqrt{\epsilon}}{d^{1/4}}\right)$ in Line 4 of Algorithm 2. Then Algorithm 2 outputs an ϵ-approximate first-order stationary point. The total number of function value evaluations is no more than

$$O\left(\frac{(f(x_0) - f_{\mathrm{low}})d}{\epsilon^2}\right).$$

Proof. Recall the approximation error of the Riemannian coordinate-wise zeroth-order estimator (Lemma D.1), it holds that

$$\|g_{x_t}(0; \mu) - \nabla \hat{f}_{x_t}(0)\| \leq \frac{\epsilon}{4} = \frac{lB}{2}$$

by setting $\mu = \mathcal{O}\left(\frac{\sqrt{\epsilon}}{d^{1/4}}\right)$. For the scenario where $\|g_{x_t}(0; \mu)\| \geq lB$ holds, Lemma E.1 gives

$$f(x_{t+1}) - f(x_t) \leq -\min\left\{\frac{lB^2}{16}, lb^2\right\} = -\frac{\epsilon^2}{64l}.$$

Otherwise, we have

$$\|\operatorname{grad} f(x_t)\| = \|\nabla \hat{f}_{x_t}(0)\| \leq \|g_{x_t}(0; \mu) - \nabla \hat{f}_{x_t}(0)\| + \|g_{x_t}(0; \mu)\| \leq \frac{3}{4}\epsilon,$$

where the first equality holds as $T_{x_t,0}$ is identity. Therefore, as computing the zeroth-order estimator once requires $2d$ function value evaluations, the total number of function value evaluations must be less than

$$O\left(\frac{(f(x_0) - f_{\mathrm{low}})d}{\epsilon^2}\right).$$

□

H. Riemannian Geometry of the Simplex

The Riemannian geometry of the positive orthant $\mathbb{R}^d_+ = \{x : x_i > 0 \text{ for all } i \in [d]\}$ was studied by researchers from mathematical biology and evolutionary game theory (Shahshahani, 1979; Mertikopoulos & Sandholm, 2018). For completeness, this section provides missing details of calculation based on the Riemannian geometry of positive orthant and simplex in the experiment. Formally the positive orthant is $\mathbb{R}^d_+$ is endowed with a Riemannian metric whose metric matrix $\{g_{ij}(x)\}$ is diagonal with $g_{ii}(x) = \frac{|x|}{x_i}$ where $|x| = \sum_{j=1}^d x_j$, i.e.,

$$g(x) = \begin{bmatrix} \frac{|x|}{x_1} & & 0 \\ & \ddots & \\ 0 & & \frac{|x|}{x_d} \end{bmatrix}$$

$\mathbb{R}^d_+$ is a single chart manifold with a non-Euclidean structure. To compute the pullback function $\hat{f}_x = f \circ \mathrm{Retr}_x$ on the unit simplex, we introduce the exponential map on the Shahshahani manifold as the retraction. Given a point $x \in \Delta^{d-1}$ and a vector $s \in T_x\Delta^{d-1}$, the exponential map is

$$\mathrm{Exp}_x(s) = \left(\frac{x_1 e^{s_1}}{\sum_j x_j e^{s_j}}, ..., \frac{x_d e^{s_d}}{\sum_j x_j e^{s_j}} \right)^\top \in \mathbb{R}^d.$$